rounding errors
band systems
Gaussian elimination
pivotal growth

Z V O N I M I R B O H T E

A N A L I Z A Z A O K R O Ž I T V E N I H N A P A K P R I R E Š E V A N J U P A S O V N I H S I S T E M O V L I N E A R N I H E N A Č B P O G A U S S O V I M E T O D I

DISERTACIJA

LJUBLJANA, 1971

Wilkinson (1961) je izdelal podrobno analizo zaokrožitvenih napak pri reševanju sistemov linearnih enačb po Gaussovi metodi za aritmetiko s premično vejico.

V tem delu sem obravnaval kot poseben primer pasovne sisteme linearnih enačb in dobil za Gaussovo metodo z delnim pivotiranjem izboljšane ocene za zaokrožitvene napake. Posebej sem obravnaval pivotno rast in dobil novo oceno za maksimalni pivot.

Ob tej priložnosti čutim prijetno dolžnost zahvaliti se svojemu učitelju, profesorju dr. I. Vidavu za vso matematično vzgojo, docentu dr. A. Suhadolcu, ki je vedno z zanimanjem spremljal moje delo in dr. J. H. Wilkinsonu, ki me je s svojimi deli navdušil za numerično linearno algebro in posebej za analizo zaokrožitvenih napak. Prvoimenovana sta tudi skrbno prebrala rokopis in mi dala vrsto dragocenih nasvetov. Veliko zahvalo sem dolžan tudi študentki A. Pavličevi za trud pri tipkanju tega dela.

Z. B.

Ljubljana, januar 1971

K A Z A L O

I. U V O D

1. Osnovne predpostavke in pomožne ocene

Pri analizi zaokrožitvenih napak je potrebno upoštevati način izvajanja aritmetičnih operacij in navadno narediti še nekaj dodatnih predpostavk, ki omogočajo poenostavitev sicer kompliciranih ocen.

Vzemimo, da imamo opravka z računalnikom, ki računa s števili v b-narnem sistemu. Ta števila naj bodo zapisana v obliki s premično vejico (floating point) s t b-narnimi ciframi. Obseg dopustnih števil naj bo tako velik, da pri obravnavanih računih nikdar ne pride do prekoračitve tega obsega.

Množica števil P, ki se dajo eksaktno upodobiti v računalniku, je definirana takole:

$$P = \{ fl(x),\ fl(x) = \pm m \cdot b^{e} \}$$

kjer imenujemo m mantiso in e eksponent števila fl(x) pri številski osnovi b. Pri tem naj veljajo naslednje omejitve:

(i) $\quad b^{-1} \leqq m < 1$

Mantisa je vedno tako normalizirana, da je manjša od 1 in ni manjša od b^{-1}.

(ii) $\quad -M_1 \leqq e \leqq M_2$

Eksponent je celo število, ki je po absolutni vrednosti sicer omejeno, a ta omejitev za analizo zaokrožitvenih napak ni bistvena.

(iii) $\quad m = 0,d_1d_2\ldots d_t =$

$$= d_1b^{-1} + d_2b^{-2} + \ldots + d_tb^{-t}$$

Za upodobitev mantise je na razpolago t b-narnih mest. Cifre d_i so cela števila med 0 in b-1, le d_1 je najmanj 1. Izjema je število 0, ki ima mantiso 0. Za znak števila vzemimo, da je poskrbljeno posebej.

Množica P je torej končna množica racionalnih števil, ki ležijo na dovolj velikem intervalu, in ki so med dvema zaporednima potencama osnove b razporejena ekvidistantno. Vsako

število iz množice P ima natanko t b-narnih cifer, od katerih prva ni nič, položaj vejice pa določa eksponent e. Zato pravimo, da so to števila s premično vejico. Števila s premično vejico se pri avtomatičnem računanju pretežno uporabljajo pri simboličnih jezikih (ALGOL, FORTRAN in pd.).

Vsa druga števila, ki ne spadajo v množico P, je treba aproksimirati s števili iz množice P. Vzemimo, da pri tem upoštevamo pravilo pravega zaokrožanja, to je, da k vsakemu realnemu številu x poiščemo kot aproksimacijo v množici P k x najbližje število. Če označimo to število s fl(x), potem velja (Bohte (1970))

$$fl(x) = x(1+\varepsilon), \qquad |\varepsilon| \leq \frac{b}{2}.b^{-t}$$

če le leži $|x|$ na intervalu med absolutno najmanjšim (razen 0) in absolutno največjim številom iz množice P. Torej se dajo števila v tej obliki aproksimirati z majhno relativno napako.

Število

$$a = \frac{b}{2}.b^{-t} \qquad (1.1)$$

bomo imenovali osnovna zaokrožitvena napaka.

Nadalje privzemimo, da je aritmetična enota računalnika tako zgrajena, da omogoča izvajanje osnovnih štirih aritmetičnih operacij s števili iz množice P tako, da je rezultat spet število iz te množice. Ker eksaktni rezultat take operacije v splošnem ni število iz množice P, pride pri tem do zaokrožitvene napake. Vzemimo, da dobimo vedno tak rezultat, kot bi ga dobili, če bi eksaktni rezultat zaokrožili na t mest.

Naj bosta x in y dve števili iz množice P in naj * pomeni enega od štirih aritmetičnih operatorjev +, -, ·, /. Izračunani rezultat imenujmo fl(x*y). Predpostavljamo torej, da pri vseh dopustnih aritmetičnih operacijah velja zveza

$$fl(x*y) = (x*y)(1+\varepsilon), \quad |\varepsilon| \leq a \qquad (1.2)$$

To prav gotovo velja za vse računalnike, ki imajo akumulator z dvojno dolžino (Wilkinson (1963)).

Analiza zaokrožitvenih napak vodi često do bolj ali manj kompliciranih izrazov, ki vsebujejo osnovno zaokro-

žitveno napako a. Za boljšo preglednost ocen se izplača te izraze poenostaviti, čeprav gre poenostavitev delno na račun rahlega poslabšanja ocene.

Za strogo upoštevanje členov drugega velikostnega reda bomo privzeli dve predpostavki, ki naj bosta v nadaljnjem vedno izpolnjeni:

$$\text{(i)} \qquad a \leqq 9 \cdot 10^{-3} \tag{1.3}$$

Za osnovno zaokrožitveno napako a bomo vedno predpostavili, da ni večja od $9 \cdot 10^{-3}$. To je zelo majhna omejitev, saj je navadno pri avtomatičnem računanju a reda 10^{-8} ali vsaj 10^{-5}. Ta predpostavka je izpolnjena celo v primeru $b = 10$, $t = 3$, ko je a po (1.1) enak $5 \cdot 10^{-3}$, to je takrat, ko računamo s trimestnimi dekadičnimi števili s premično vejico in pravim zaokrožanjem.

$$\text{(ii)} \qquad na \leqq 0{\cdot}1 \tag{1.4}$$

Kjerkoli bo celo število n pomenilo dimenzijo računskega postopka, to je število zaporednih operacij istega tipa, bomo predpostavili veljavnost omejitve (1.4). To tudi ni huda omejitev, kot se pokaže v tistih primerih, kjer bomo to oceno izkoristili. Na primer, pri največjem dopustnem $a = 9 \cdot 10^{-3}$, bodo naše ocene veljavne pri vsakem n, ki je manjši od 12. Pri a reda 10^{-5} pa je največji dopustni n že reda 10^{4}, kar zadošča za vse praktične potrebe.

Pri nadaljnjih analizah bomo uporabljali še naslednje pomožne ocene, ki sledijo iz binomskega izreka in ocen (1.3) in (1.4) (Bohte (1970)):

$$\text{(iii)} \qquad (1 - a)^{-1} \leqq 1 + 1{\cdot}01a \tag{1.5}$$

$$\text{(iv)} \qquad (1 + a)^{n} \leqq 1 + 1{\cdot}06na \tag{1.6}$$

$$\text{(v)} \qquad (1 - a)^{-n} \leqq 1 + 1{\cdot}12na \tag{1.7}$$

$$\text{(vi)} \qquad (1 + a)(1 - a)^{-n} \leqq 1 + 1{\cdot}12(n+1)a \tag{1.8}$$

Kot zgled za uporabo enačbe (1.2) in omenjenih ocen si oglejmo analizo zaokrožitvenih napak pri izračunu skalarnega produkta dveh vektorjev. Podobne izraze bomo pozneje večkrat srečali.

Naj bosta x in y dana vektorja:

$$x^T = (x_1,\ldots,x_n), \qquad y^T = (y_1,\ldots,y_n)$$

in naj bo

$$s_n = fl(x^T y) = fl\Big(\sum_{i=1}^{n} x_i y_i\Big)$$

Računanje skalarnega produkta s_n poteka takole:

$$p_i = fl(x_i y_i) , \quad i=1,\ldots,n \tag{1.9}$$

$$s_1 = p_1$$

$$s_{i+1} = fl(s_i + p_{i+1}) , \quad i=1,\ldots,n-1 \tag{1.10}$$

Če upoštevamo zaokrožitvene napake, moremo na osnovi zveze (1.2) enačbi (1.9) in (1.10) zapisati takole:

$$p_i = (x_i y_i)(1 + \varepsilon_i) , \quad |\varepsilon_i| \leqq a$$

$$s_{i+1} = (s_i + p_{i+1})(1 + \eta_i) , \quad |\eta_i| \leqq a$$

Če te enačbe povežemo skupaj, dobimo rezultat

$$s_n = x_1 y_1 (1 + \zeta_1) + \ldots + x_n y_n (1 + \zeta_n)$$

kjer je

$$1 + \zeta_1 = (1 + \varepsilon_1)(1 + \eta_1) \ldots (1 + \eta_{n-1})$$

$$1 + \zeta_{k+1} = (1 + \varepsilon_{k+1})(1 + \eta_k) \ldots (1 + \eta_{n-1}) \qquad k=1,2,\ldots,n-1$$

Za števila $1 + \zeta_k$ veljajo očitno ocene

$$(1 - a)^n \leqq 1 + \zeta_1 \leqq (1 + a)^n$$

$$(1 - a)^{n-k+1} \leqq 1 + \zeta_{k+1} \leqq (1 + a)^{n-k+1} \qquad k=1,2,\ldots,n-1$$

Če upoštevamo poenostavitev (1.6), dobimo za števila $|\zeta_k|$ naslednje ocene:

$$|\zeta_1| \leqq 1{\cdot}06\ na \tag{1.11}$$

$$|\zeta_{k+1}| \leqq 1{\cdot}06\ (n-k+1)a , \quad k=1,\ldots,n-1 \tag{1.12}$$

Rezultat analize zaokrožitvenih napak moremo v tem primeru izraziti takole:

Izračunani skalarni produkt dveh vektorjev je enak eksaktnemu skalarnemu produktu vektorja x z nekoliko spremenjenim vektorjem y:

$$fl(x^T y) = x^T z$$

kjer je

$$z_i = y_i(1 + \zeta_i), \quad i=1,\ldots,n$$

in veljajo za relativne spremembe v komponentah ocene (1.11) in (1.12).

Relativne napake v izračunanem skalarnem produktu brez dodatnih predpostavk ni mogoče oceniti, saj je ta lahko poljubno velika.

2. Gaussova metoda za reševanje sistemov linearnih enačb

Na kratko si oglejmo Gaussovo metodo za reševanje sistemov linearnih enačb in sicer osnovno varianto in varianto z delnim pivotiranjem. Čeprav je metoda dobro znana, je le potrebno vsaj glede označb nekaj pojasnil.

Dani sistem linearnih enačb zapišemo v obliki

$$Ax = b \qquad (2.1)$$

kjer je A dana nesingularna kvadratna matrika reda n, b dani vektor desnih strani, x pa iskani vektor.

Pri Gaussovi eliminacijski metodi tvorimo postopoma ekvivalentne sisteme

$$A^{(r)}x = b^{(r)}, \quad r=1,2,\ldots,n \qquad (2.2)$$

kjer je $A^{(1)} = A$ in $b^{(1)} = b$, tako, da je končna matrika $A^{(n)}$ zgornja trikotna matrika.

Matrika $A^{(r)}$ je že zgornja trikotna matrika v prvih r vrsticah in v prvih r-1 stolpcih. V desnem spodnjem vogalu imamo še kvadratno matriko reda n-r+1. Na r-tem koraku eliminacije želimo eliminirati neznanko x_r iz zadnjih n-r enačb, to je, doseči v r-tem stolpcu pod diagonalo same ničle. Matriko $A^{(r+1)}$ dobimo torej iz matrike $A^{(r)}$ tako, da

množimo r-to vrstico s primerno izbranim faktorjem m_{ir} in jo odštejemo od i-te vrstice za i=r+1,...,n.

Očitno je treba vzeti

$$m_{ir} = a_{ir}^{(r)}/a_{rr}^{(r)} , \quad i=r+1,\ldots,n \tag{2.3}$$

in novi elementi so enaki

$$a_{ik}^{(r+1)} = a_{ik}^{(r)} - m_{ir}a_{rk}^{(r)} , \quad i,k=r+1,\ldots,n \tag{2.4}$$

Pri tej eliminaciji se desne strani enačb transformirajo na podoben način:

$$b_i^{(r+1)} = b_i^{(r)} - m_{ir}b_r^{(r)} , \quad i=r+1,\ldots,n \tag{2.5}$$

Rešitev sistema x dobimo nato zelo enostavno iz zgornjega trikotnega sistema

$$Ux = y \tag{2.6}$$

kjer je $U = A^{(n)}$ in $y = b^{(n)}$, po formulah

$$x_i = (y_i - \sum_{k=i+1}^{n} u_{ik}x_k)/u_{ii} , \quad i=n,n-1,\ldots,1 \tag{2.7}$$

Elemente zgornje trikotne matrike $A^{(n)}$ smo označili z

$$u_{ik} = a_{ik}^{(n)} = a_{ik}^{(i)} , \quad k \geq i \tag{2.8}$$

Če definiramo spodnjo trikotno matriko L takole:

$$\begin{aligned} \ell_{ii} &= 1 , \quad i=1,\ldots,n \\ \ell_{ik} &= m_{ik} , \quad i > k \end{aligned} \tag{2.9}$$

potem se izkaže, da velja (gl. npr. Bohte (1970))

$$A = L\cdot U , \quad b = Ly \tag{2.10}$$

če so le vsi $a_{rr}^{(r)}$ različni od nič. Matriko A smo torej razcepili na produkt spodnje in zgornje trikotne matrike.

Število $a_{rr}^{(r)}$ imenujemo pivot na r-tem koraku eliminacije. Opisani računski postopek odpove, brž ko je kak pivot enak nič. Pri nesingularni matriki A se da deljenju z nič izogniti s pivotiranjem, to je s posebnim načinom izbire pivota na vsakem koraku eliminacije. Izkaže se tudi, da je pivotiranje potrebno zaradi zaokrožitvenih napak, ki so sicer lahko poljubno velike.

Na kratko si oglejmo le delno pivotiranje. Pri tem načinu izbire pivotov si na vsakem koraku eliminacije izberemo za pivot absolutno največji element v prvem stolpcu preostale kvadratne matrike. Pri tem moramo po potrebi zamenjati dve enačbi v sistemu, to je zamenjati dve vrstici v preostali matriki in ustrezna elementa v vektorju desnih strani.

Naj velja na r-tem koraku eliminacije

$$|a_{sr}^{(r)}| = \max_{r \leq i \leq n} |a_{ir}^{(r)}|$$

Tedaj zamenjamo v sistemu (2.2) r-to in s-to enačbo ter izvedemo eliminacijo. Pri nesingularni matriki A so tako izbrani pivoti vedno različni od nič.

Pri delnem pivotiranju seveda ne veljata enačbi (2.10), pač pa velja (gl. npr. Bohte (1970))

$$PA = L \cdot U \ , \quad Ly = Pb$$

kjer je P primerna permutacijska matrika, ki je

$$P = I_{n-1,(n-1)'} \cdots I_{1,1'}$$

če na r-tem koraku zamenjamo vrstici z indeksoma r in r' ($r' \geq r$). Matrike I_{ij} so elementarne permutacijske matrike. Pri tem so vsi elementi spodnje trikotne matrike L absolutno omejeni z 1. Rešitev x dobimo kot prej iz zgornjega trikotnega sistema (2.6).

3. Analiza zaokrožitvenih napak pri Gaussovi metodi z delnim pivotiranjem

Oglejmo si nekoliko podrobneje analizo zaokrožitvenih napak pri Gaussovi metodi z delnim pivotiranjem v splošnem primeru, da bomo laže navezali ustrezno analizo za primer pasovnih matrik.

Wilkinson (1963) je dokazal, da je izračunana rešitev x sistema linearnih enačb (2.1) pri tej metodi enaka eksaktni rešitvi nekega perturbiranega sistema

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

in podal oceno za normo perturbacijske matrike δA.

Analizirajmo najprej razcep matrike A na produkt dveh trikotnih matrik.

Teoretični formuli (2.3) in (2.4) se z upoštevanjem zaokrožitvenih napak prevedeta v formuli:

$$m_{ir} = fl(a_{ir}^{(r)}/a_{rr}^{(r)}) = (a_{ir}^{(r)}/a_{rr}^{(r)})(1 + \varepsilon_1) \qquad (3.1)$$
$$i=r+1,\ldots,n$$

$$a_{ik}^{(r+1)} = fl(a_{ik}^{(r)} - m_{ir}a_{rk}^{(r)}) =$$
$$= (a_{ik}^{(r)} - m_{ir}a_{rk}^{(r)}(1 + \varepsilon_2))(1 + \varepsilon_3) \qquad (3.2)$$
$$i,k=r+1,\ldots,n$$

kjer velja

$$|\varepsilon_i| \leq a\ , \quad i=1,2,3 \qquad (3.3)$$

Če upoštevamo definicijo matrik L in U, moremo enačbo (3.1) zapisati v obliki

$$0 = a_{ir}^{(r)} - m_{ir}a_{rr}^{(r)} + \varepsilon_{ir}^{(r)} =$$
$$= a_{ir}^{(r)} - \ell_{ir}u_{rr} + \varepsilon_{ir}^{(r)} \qquad (3.4)$$

kjer je

$$\varepsilon_{ir}^{(r)} = \varepsilon_1 a_{ir}^{(r)} \qquad (3.5)$$

Enačbo (3.2) pa zapišimo takole:

$$a_{ik}^{(r+1)} = a_{ik}^{(r)} - m_{ir}a_{rk}^{(r)} + \varepsilon_{ik}^{(r)} =$$
$$= a_{ik}^{(r)} - \ell_{ir}u_{rk} + \varepsilon_{ik}^{(r)} \qquad (3.6)$$

kjer je

$$\varepsilon_{ik}^{(r)} = \frac{\varepsilon_3}{1+\varepsilon_3} a_{ik}^{(r+1)} - \ell_{ir}u_{rk}\varepsilon_2 \qquad (3.7)$$

Oglejmo si sedaj, kako se spreminja element $a_{ik}^{(r)}$, ko r teče od 1 do n-1. Ločiti moramo dva primera.

(i) $k \geq i$ (zgornji trikotnik)

Element v zgornjem trikotniku se spreminja po formuli (3.6), dokler r ne doseže vrednosti i-1, nakar ostane pri vseh nadaljnjih korakih nespremenjen. Torej velja enačba (3.6) za $r = 1,2,\ldots,i-1$. Če vse te enačbe seštejemo in upoštevamo, da

je $A = A^{(1)}$ in $a_{ik}^{(i)} = u_{ik}$, dobimo

$$a_{ik} = u_{ik} + \sum_{r=1}^{i-1} \ell_{ir} u_{rk} - e_{ik}$$

ali

$$a_{ik} + e_{ik} = \sum_{r=1}^{i} \ell_{ir} u_{rk} \qquad (3.8)$$

kjer je

$$e_{ik} = \sum_{r=1}^{i-1} \varepsilon_{ik}^{(r)} \qquad (3.9)$$

(ii) $k < i$ (spodnji trikotnik)

Element v spodnjem trikotniku se tudi spreminja po formuli (3.6) za $r = 1,2,\ldots,k-1$, nakar ga uporabimo v formuli (3.4), potem pa ga zamenjamo s številom 0. Pri nadaljnjih korakih se ne spreminja več.

Torej smemo podobno kot prej sešteti enačbe (3.6) za $r = 1,\ldots,k-1$ in še enačbo (3.4) za $r = k$. Tako dobimo

$$a_{ik} + e_{ik} = \sum_{r=1}^{k} \ell_{ir} u_{rk} \qquad (3.10)$$

kjer je

$$e_{ik} = \sum_{r=1}^{k} \varepsilon_{ik}^{(r)} \qquad (3.11)$$

Enačbi (3.8) in (3.10) moremo zapisati v matrični obliki

$$A + E = L \cdot U \qquad (3.12)$$

kjer so elementi matrike E definirani s formulama (3.9) in (3.11).

Izračunani trikotni matriki L in U sta torej taki, da predstavljata eksaktni razcep neke perturbirane matrike A + E.

Če hočemo dobiti ocene za elemente matrike E, moramo narediti nekaj predpostavk, kajti potrebujemo ocene za števila m_{ir} in $a_{ik}^{(r)}$.

Najprej je očitno, da je m_{ir} lahko poljubno veliko število, če izvajamo eliminacijo brez pivotiranja, saj je pivot na kakem koraku lahko enak nič. Tedaj je tudi matrika E lahko poljubno velika.

Zato vzemimo, da izvajamo eliminacijo z delnim pivotiranjem in zaradi enostavnejšega zapisa vzemimo, da ima prvot-

na matrika A že tako permutirane vrstice, da velja pri naravnem vrstnem redu eliminacij ocena

$$|m_{ir}| \leq 1 \tag{3.13}$$

Glede velikosti elementov $a_{ik}^{(r)}$ pa zaenkrat predpostavimo le to, da so omejeni in naj velja

$$g = \max_{i,k,r} |a_{ik}^{(r)}| \tag{3.14}$$

O ocenah za število g bomo govorili v zadnjem poglavju.

Pri teh dveh predpostavkah moremo dobiti za elemente perturbacijske matrike dokaj ugodne ocene.

Najprej ocenimo števila $\varepsilon_{ik}^{(r)}$. Iz (3.3) in (3.7) sledi

$$|\varepsilon_{ik}^{(r)}| \leq \frac{a}{1-a} |a_{ik}^{(r+1)}| + |m_{ir}| \, |a_{rk}^{(r)}| a$$

Če upoštevamo še predpostavki (3.13) in (3.14), dobimo od tod

$$|\varepsilon_{ik}^{(r)}| \leq (a/(1 - a) + a)g , \quad r < i,k \tag{3.15}$$

Iz (3.5) pa sledi

$$|\varepsilon_{ir}^{(r)}| \leq ag$$

Oceno (3.15) lahko poenostavimo, če upoštevamo oceni (1.3) in (1.5). Tedaj je

$$|\varepsilon_{ik}^{(r)}| \leq 2{\cdot}01 \, ag \tag{3.16}$$

ki očitno velja tudi v primeru k = r.

Končno dobimo za elemente matrike E iz (3.9) in (3.11) oceni:

$$|e_{ik}| \leq 2{\cdot}01(i-1)ga , \quad k \geq i$$

$$|e_{ik}| \leq 2{\cdot}01 \, kga , \quad k < i$$

Brez posebnih težav dobimo od tod naslednje ocene za norme matrike E:

$$\| E \|_1 \leq 2{\cdot}01 \, ga \, \frac{1}{2} n(n-1)$$

$$\| E \|_\infty \leq 2{\cdot}01 \, ga \, \frac{1}{2} (n-1)(n+2)$$

$$\| E \|_E \leq 2{\cdot}01 \, ga \left(\frac{1}{6} n^2 (n^2-1) \right)^{1/2}$$

Tu pomenijo

$$\| A \|_1 = \max_k \sum_{i=1}^{n} |a_{ik}|$$

$$\| A \|_\infty = \max_i \sum_{k=1}^{n} |a_{ik}|$$

$$\| A \|_E^2 = \sum_{i=1}^{n} \sum_{k=1}^{n} |a_{ik}|^2$$

Naslednji del računa je transformacija desnih strani ali rešitev spodnjega trikotnega sistema

$$Ly = b$$

Formule za rešitev tega sistema so:

$$y_i = b_i - \sum_{k=1}^{i-1} m_{ik} y_k , \quad i=1,\ldots,n \tag{3.17}$$

Števila $b_i^{(r)}$ v formulah (2.5) so namreč ravno delne vsote v enačbi (3.17) in $b_i^{(i)} = y_i$.

V resnici seveda izračunamo število

$$y_i = fl(b_i - \sum_{k=1}^{i-1} m_{ik} y_k) \tag{3.18}$$

Podobno kot pri skalarnem produktu, ki smo ga obravnavali v 1. razdelku, dobimo z upoštevanjem zaokrožitvenih napak zvezo:

$$y_i = b_i(1 + \zeta_i) - \sum_{k=1}^{i-1} m_{ik} y_k (1 + \zeta_k) \tag{3.19}$$

kjer je

$$1 + \zeta_i = (1 + \eta_1)(1 + \eta_2) \ldots (1 + \eta_{i-1})$$

$$1 + \zeta_k = (1 + \varepsilon_k)(1 + \eta_k) \ldots (1 + \eta_{i-1})$$

in veljajo ocene

$$|\varepsilon_k| \leq a , \quad |\eta_j| \leq a \tag{3.20}$$

Faktorji $(1 + \varepsilon_k)$ nastanejo pri množenju, faktorji $(1 + \eta_j)$ pa pri seštevanju.

Enačbo (3,19) zapišimo rajši v obliki

$$b_i = y_i(1 + \xi_i) + \sum_{k=1}^{i-1} m_{ik} y_k (1 + \xi_k) =$$

$$= \sum_{k=1}^{i} \ell_{ik} y_k (1 + \xi_k)$$

kjer velja

$$1 + \xi_i = (1 + \zeta_i)^{-1} =$$
$$= (1 + \eta_1)^{-1}(1 + \eta_2)^{-1} \ldots (1 + \eta_{i-1})^{-1} \qquad (3.21)$$
$$1 + \xi_k = (1 + \zeta_k)(1 + \zeta_i)^{-1} =$$
$$= (1 + \varepsilon_k)(1 + \eta_1)^{-1} \ldots (1 + \eta_{k-1})^{-1}$$

Izračunani vektor y je torej eksaktna rešitev sistema

$$(L + \delta L)y = b \qquad (3.22)$$

kjer je perturbacijska matrika δL tudi spodnja trikotna matrika in velja

$$\delta \ell_{ik} = \ell_{ik}\xi_k , \quad i \geq k \qquad (3.23)$$

Za števila ξ_k dobimo iz (3.21) in (3.20) ocene:

$$(1 + a)^{-(i-1)} \leq 1 + \xi_i \leq (1 - a)^{-(i-1)}$$
$$(1 - a)(1 + a)^{-(k-1)} \leq 1 + \xi_k \leq (1 + a)(1 - a)^{-(k-1)}$$

ki jih moremo poenostaviti s pomočjo ocen (1.7) in (1.8) v ocene

$$|\xi_i| \leq 1{\cdot}12(i-1)a$$
$$|\xi_k| \leq 1{\cdot}12\ ka$$

Ker je $|\ell_{ik}| \leq 1$, dobimo iz (3.23) ocene:

$$|\delta \ell_{ii}| \leq 1{\cdot}12(i-1)a$$
$$|\delta \ell_{ik}| \leq 1{\cdot}12\ ka , \quad i > k$$

Od tod dobimo hitro ocene za norme matrike δL:

$$\|\delta L\|_1 \leq 1{\cdot}12\ a\ \tfrac{1}{4}(n-1)(n+3)$$

$$\|\delta L\|_\infty \leq 1{\cdot}12\ a\ \tfrac{1}{2}(n-1)(n+2)$$

$$\|\delta L\|_E \leq 1{\cdot}12\ a\left(\tfrac{1}{12}\, n\,(n-1)(n^2 + 5n - 2)\right)^{1/2} \qquad (3.24)$$

Zadnji del reševanja prvotnega sistema je reševanje zgornjega trikotnega sistema

$$Ux = y$$

Formule (2.7) nam pri upoštevanju zaokrožitvenih napak dajo za komponente rešitve naslednje vrednosti:

$$x_i = fl((y_i - \sum_{k=i+1}^{n} u_{ik}x_k)/u_{ii}) =$$

$$= (y_i(1+\zeta_i) - \sum_{k=i+1}^{n} u_{ik}x_k(1+\zeta_k))(1+\varepsilon_i)/u_{ii} \qquad (3.25)$$

kjer je

$$1 + \zeta_i = (1 + \eta_{i+1})(1 + \eta_{i+2}) \ldots (1 + \eta_n)$$

$$1 + \zeta_k = (1 + \varepsilon_k)(1 + \eta_k) \ldots (1 + \eta_n)$$

Faktorji $(1 + \eta_j)$ nastanejo pri seštevanju, faktorji $(1 + \varepsilon_j)$ pa pri množenju ali deljenju. Pri tem veljajo ocene

$$|\varepsilon_j| \leqq a \ , \quad |\eta_j| \leqq a$$

Enačbo (3.25) moremo zapisati v obliki

$$y_i = \sum_{k=i}^{n} u_{ik}x_k(1 + \xi_k)$$

kjer je

$$1 + \xi_i = (1+\varepsilon_i)^{-1}(1+\zeta_i)^{-1} =$$

$$= (1+\varepsilon_i)^{-1}(1+\eta_{i+1})^{-1} \ldots (1+\eta_n)^{-1}$$

$$1 + \xi_k = (1+\zeta_k)(1+\zeta_i)^{-1} =$$

$$= (1+\varepsilon_k)(1+\eta_{i+1})^{-1} \ldots (1+\eta_{k-1})^{-1}$$

Izračunana rešitev x je torej eksaktna rešitev sistema

$$(U + \delta U)x = y \qquad (3.26)$$

kjer je perturbacijska matrika δU tudi zgornja trikotna matrika in velja

$$\delta u_{ik} = u_{ik}\xi_k \ , \quad k \geqq i$$

Števila u_{ik}, ki so enaka $a_{ik}^{(i)}$, moremo oceniti s številom g (3.14):

$$|u_{ik}| \leqq g$$

Za števila ξ_k pa veljajo ocene

$$(1+a)^{-(n-i+1)} \leqq 1 + \xi_i \leqq (1-a)^{-(n-i+1)}$$

$$(1-a)(1+a)^{-(k-i-1)} \leqq 1 + \xi_k \leqq (1+a)(1-a)^{-(k-i-1)}$$

Te ocene moremo poenostaviti z uporabo pomožnih ocen (1.7) in (1.8) takole:

$$|\xi_i| \leq 1{,}12\, a(n-i+1)$$

$$|\xi_k| \leq 1{,}12\, a(k-i)$$

Od tod direktno sledijo ocene za elemente matrike δU:

$$|\delta u_{ii}| \leq 1{,}12(n-i+1)ga$$

$$|\delta u_{ik}| \leq 1{,}12(k-i)ga \ , \ k > i$$

Brez težav dobimo od tod ocene za norme matrike δU:

$$\| \delta U \|_1 \leq 1{,}12\ ga\ \tfrac{1}{2}(n^2 - n + 2)$$

$$\| \delta U \|_\infty \leq 1{,}12\ ga\ \tfrac{1}{2}\, n\,(n+1)$$

$$\| \delta U \|_E \leq 1{,}12\ ga\,(\tfrac{1}{12}\, n(n+1)^2(n+2))^{1/2} \qquad (3.27)$$

Sedaj moremo združiti analize posameznih korakov v končni rezultat.

Iz (3.12), (3.22) in (3.26) sledi, da je izračunana rešitev x eksaktna rešitev sistema

$$(L + \delta L)(U + \delta U)x = b$$

ali sistema

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

kjer je perturbacijska matrika δA enaka

$$\delta A = E + \delta L \cdot U + L \cdot \delta U + \delta L \cdot \delta U$$

Za katerokoli normo velja potem ocena

$$\|\delta A\| \leq \|E\| + \|\delta L\|\ \|U\| + \|L\|\ \|\delta U\| + \|\delta L\|\ \|\delta U\|$$

Ocene za $\|E\|_p$, $\|\delta L\|_p$ in $\|\delta U\|_p$ pri $p = 1, \infty, E$ že poznamo. Norme matrik L in U pa ni težko oceniti s predpostavkama (3.13) in (3.14):

$$\| L \|_1 \leq n$$

$$\| L \|_\infty \leq n$$

$$\| L \|_E \leq \left(\tfrac{1}{2} n(n+1)\right)^{1/2}$$

$$\| U \|_1 \leq gn$$

$$\| U \|_\infty \leq gn$$

$$\| U \|_E = g\left(\tfrac{1}{2} n(n+1)\right)^{1/2}$$

Z elementarnim računom, primerno poenostavitvijo in upoštevanjem predpostavke (1.4) dobimo od tod naslednje ocene:

$$\| \delta A \|_1 \leq 0{\cdot}86(n^3 + 2n^2)ga$$

$$\| \delta A \|_\infty \leq 1{\cdot}16(n^3 + 2n^2)ga$$

$$\| \delta A \|_E \leq 0{\cdot}46(n^3 + 5n^2)ga$$

Vse tri ocene imajo skupno obliko

$$\| \delta A \|_p \leq f_p(n)ga \ , \ p = 1,\infty,E \qquad (3.29)$$

kjer je $f_p(n)$ polinom tretje stopnje v n z vodilnim koeficientom reda 1. Ta koeficient je najmanjši pri p = E. Z bolj natančnim ocenjevanjem elementov $|\delta a_{ik}|$ bi lahko ta koeficient rahlo izboljšali. Pri p = E pride ta koeficient spet najmanjši in sicer 0·36, kar pa ne da bistveno boljše ocene.

Z oceno (3.28) primerljive ocene dobimo v Wilkinson (1963), Isaacson and Keller (1966) in Forsythe and Moler (1967).

Omeniti velja, da se da koeficient $f_p(n)$ v oceni (3.29) v splošnem bistveno izboljšati le, če računamo skalarne produkte v vseh formulah, kjer nastopajo, z akumulacijo v dvojni dolžini. Wilkinson (1965) je pokazal, da je tedaj $f_p(n)$ približno enak n. Na žalost pa je tako računanje lahko izvedljivo le pri majhnem številu računalnikov.

Oceno (3.29) moremo uporabiti za apriorno oceno napake v rešitvi po formuli (Wilkinson (1965))

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

kjer je na levi relativna sprememba v rešitvi sistema, če matriko spremenimo za δA. Ta ocena velja, če je

$$\|A^{-1}\| \quad \|\delta A\| < 1$$

Uporabna je seveda le, če znamo oceniti $\|A^{-1}\|$, kar pa navadno ne gre brez dodatnega računanja.

Splošno oceno (3.29) moremo v nekaterih posebnih primerih matrik izboljšati.

II. P A S O V N E M A T R I K E

V praksi pogosto nastopajo matrike, ki imajo od nič različne elemente zbrane vzdolž glavne diagonale in ob njej, vsi drugi elementi pa so enaki nič. S takimi matrikami imamo največkrat opraviti pri numeričnem reševanju diferencialnih enačb.

V tem poglavju bomo izdelali podrobno analizo zaokrožitvenih napak pri reševanju pasovnega sistema linearnih enačb po Gaussovi metodi z delnim pivotiranjem. Izkaže se, da se da splošna ocena (3.29) v tem posebnem primeru bistveno izboljšati.

4. Pasovni sistem linearnih enačb

Matriko A imenujemo pasovno matriko s širino pasu 2p+1, če velja

$$a_{ik} = 0 , \quad |i-k| > p \tag{4.1}$$

in če je

$$n \geq 2p+1 \tag{4.2}$$

Tako matriko imenujemo včasih tudi (2p+1)-diagonalna matrika. Za elemente a_{ik} pri $|i-k| \leq p$ bomo predpostavljali, da so v splošnem od nič različni. Če je pri dani pasovni matriki kak tak element enak nič, tega v naši analizi ne bomo upoštevali.

Omejitev (4.2) ni bistvena, saj v praksi obravnavamo matriko A kot pasovno le, če je $n >> 2p+1$. Če upoštevamo posebno strukturo pasovnih matrik, se največji red, ki ga dopušča spomin računalnika pri polni matriki, znatno poveča.

Naj bo sedaj A nesingularna pasovna matrika, ki ustreza pogojema (4.1) in (4.2).

Pri reševanju pasovnega sistema enačb

$$Ax = b \tag{4.3}$$

z Gaussovo metodo bi kot v splošnem primeru brez kakršnegakoli

pivotiranja lahko dobili rešitev s poljubno veliko napako. Delno pivotiranje se izkaže za zelo primerno. Ne samo, da moremo pri takem reševanju pasovnega sistema enačb skoraj v vsej splošnosti upoštevati pasovno strukturo matrike in s tem varčevati s spominom v računalniku, ampak moremo s podrobno analizo napak tudi pokazati, da je ta strategija s stališča natančnosti računanja popolnoma zadovoljiva.

Pri delnem pivotiranju zamenjujemo med eliminacijami vrstice matrike s ciljem, da dobimo pri tekoči matriki na ustreznem mestu na diagonali absolutno največji element od vseh, ki so pod njim. Zaradi tega pridejo na r-tem koraku eliminacije za zamenjavo z r-to vrstico v poštev le vrstice r+1, r+2, ... , r+p, saj so v vseh naslednjih vrsticah v r-tem stolpcu same ničle.

Vzemimo, da smo na r-tem koraku eliminacije zamenjali r-to vrstico z r'-to. Tej zamenjavi ustreza elementarna permutacijska matrika $I_{r,r'}$.

Če označimo s P permutacijsko matriko

$$P = I_{n-1,(n-1)'} \cdots I_{1,1'} \tag{4.4}$$

potem ima matrika PA, kot smo omenili v prejšnjem poglavju, lastnost, da nam da pri Gaussovi eliminaciji brez pivotiranja razcep

$$PA = L \cdot U$$

pri čemer so vsi elementi spodnje trikotne matrike L absolutno omejeni z 1.

V (4.4) velja pogoj

$$r \leq r' \leq \min(r+p,n) \tag{4.5}$$

Matrika PA, ki ima iste vrstice kot matrika A, le da so primerno permutirane med seboj, v splošnem ni več (2p+1)-diagonalna, pač pa ima v svoji strukturi vseeno posebnosti, ki izvirajo iz pasovne strukture matrike A in ki se dajo izkoristiti pri podrobni analizi napak.

Dokazali bomo, da je izračunana rešitev x sistema (4.3) eksaktna rešitev sistema

$$(PA + \delta A)x = Pb \tag{4.6}$$

in podali ocene za tri norme matrike δA.

Iz (4.6) in dejstva, da je P ortogonalna matrika, sledi, da je x tudi eksaktna rešitev sistema

$$(A + P^T \delta A)x = b$$

pri čemer očitno velja

$$\|P^T \delta A\|_p = \|\delta A\|_p , \quad p = 1,\infty,E$$

saj je P permutacijska matrika.

5. Struktura matrike PA

Najprej si podrobno oglejmo strukturo matrike PA, kajti analiza napak je enostavnejša, če si mislimo, da je prvotna matrika tako permutirana, da v toku eliminacij ni potrebno narediti nobene zamenjave vrstic.

Naj bo sedaj A poljubna (2p+1)-diagonalna matrika, P pa poljubna permutacijska matrika oblike (4.4), pri čemer veljajo pogoji (4.5). Vseh možnih takih matrik je $p!(p+1)^{n-p}$.

Če je P = I, ali, kar je isto, če je $r' = r$, $r=1,..,n-1$, potem je PA = A, torej je PA (2p+1)-diagonalna matrika. Primer matrike, ki ima to lastnost, je diagonalno dominantna pasovna matrika. Ta primer bomo obravnavali posebej.

Pri katerikoli drugi dopustni permutaciji P pride vsaj en element, ki leži znotraj (2p+1)-diagonalnega pasu, izven njega in zato matrika PA ni več (2p+1)-diagonalna. Zanima nas predvsem, kako ležijo ničle v matriki PA, če smatramo vse elemente znotraj pasu za različne od nič.

Matrika P (4.4), ki je produkt elementarnih permutacijskih matrik $I_{r,r'}$, je permutacijska matrika. Ima natanko n enojk, vsi drugi elementi pa so enaki nič. Te enojke ležijo tako, da je v vsaki vrstici in v vsakem stolpcu natanko ena enojka. Položaj enojk fiksirajmo z zapisom. Naj bo v i-ti vrstici enojka v s_i-tem stolpcu:

$$p_{is_i} = 1 , \quad p_{ik} = 0 , \quad k \neq s_i$$

Tedaj matriko P na kratko označimo s $P_{s_1,\ldots,s_n}$. Števila $s_1,\ldots,s_n$ so neka permutacija števil $1,2,\ldots,n$.

Matrika PA ima potemtakem permutirane vse vrstice matrike A in sicer je i-ta vrstica matrike PA enaka s_i-ti vrstici matrike A.

Zanima nas pa tudi obratna zveza. Na katero mesto v matriki PA pride i-ta vrstica matrike A ?

Očitno je zaradi (4.4)

$$P^{-1} = P_{s_1,\ldots,s_n}^{-1} = P_{s_1,\ldots,s_n}^{T}$$

Transponirana matrika permutacijske matrike je tudi permutacijska matrika in naj velja

$$P_{s_1,\ldots,s_n}^{T} = P_{t_1,\ldots,t_n} \qquad (5.1)$$

To pa pomeni, da velja

$$A = P_{t_1,\ldots,t_n}(PA) \qquad (5.2)$$

Torej i-ta vrstica matrike A preide v t_i-to vrstico matrike PA, kjer so števila t_i definirana s (5.1).

Lahko se je prepričati bodisi iz zveze (5.1), bodisi iz (5.2), da velja

$$t_{s_i} = i\ ,\quad s_{t_i} = i\ ,\quad i=1,2,\ldots,n \qquad (5.3)$$

Sedaj pa upoštevajmo pogoje (4.5) in jih prenesimo na števila s_i in t_i. Matriko P zapišimo v obliki

$$P = P_{s_1,\ldots,s_n} = I_{n-1,(n-1)'} \cdots I_{1,1'} I$$

To pomeni, da dobimo matriko P tako, da v enotni matriki I po vrsti izvršujemo zamenjave vrstic $(1,1')$, $(2,2')$, ... , $(n-1,(n-1)')$. Poglejmo, kaj se utegne zgoditi pri teh zamenjavah (r,r'), $r=1,\ldots,n-1$ z s_i-to vrstico matrike I.

Zaradi pogoja $r' \leq r+p$ ostane s_i-ta vrstica na svojem mestu, dokler je $r < s_i-p$. Od $r = s_i-p$ do $r = s_i-1$ se utegne zgoditi zamenjava (r,s_i), tedaj pride s_i-ta vrstica matrike I na r-to mesto in tam tudi ostane do konca vseh zamenjav, saj vedno velja $r' \geq r$. V tem primeru velja $i = r$ in $i \geq s_i-p$ ali

$$s_i \leq i + p \qquad (5.4)$$

Če se taka zamenjava ne zgodi, je nadalje možno, da je pri $r = s_i$ ustrezni $r' = r$, torej, da s_i-ta vrstica ostane na svojem mestu. Tedaj je $i = s_i$ in ocena (5.4) tudi velja. Če pa se pri $r = s_i$ zgodi zamenjava (r,r'), $r' > s_i$, se v tem primeru s_i-ta vrstica pomakne navzdol in velja kvečjemu $i > s_i$ in ocena (5.4) tudi velja.

Ugotovili smo torej, da pogoj (4.5) pomeni, da je

$$s_i \leq i + p , \quad i=1,\ldots,n \tag{5.5}$$

Analogen pogoj za števila t_i pa dobimo, če vstavimo namesto i v (5.5) t_i in upoštevamo (5.3):

$$t_i \geq i - p , \quad i=1,\ldots,n \tag{5.6}$$

Ta rezultat moremo z besedami tolmačiti takole: Nobena vrstica matrike A se v PA ne pomakne za več kot p mest navzgor.

Sedaj, ko dobro poznamo permutacijsko matriko P, lahko tudi podrobneje opišemo strukturo matrike PA, kjer je A (2p+1)-diagonalna matrika. Dogovorili smo se že, da bomo pod ničlami matrike A razumeli samo elemente izven pasu.

Dokažimo najprej nekaj pomožnih izrekov, ki jih bomo uporabili pri nadaljnji obravnavi.

Vse ničle matrike A, *ki ležijo v spodnjem trikotniku* (i > k), *ostanejo v spodnjem trikotniku matrike* PA. (T1)

Dokaz. Matrika A ima v spodnjem trikotniku ničle na mestih

$$a_{ik} = 0 , \quad i=p+2,\ldots,n , \quad k=1,\ldots,i-p-1$$

V i-ti vrstici je skrajna desna ničla v (i-p-1)-tem stolpcu. Ker je i-ta vrstica matrike A enaka t_i-ti vrstici matrike PA, je ta skrajna desna ničla še vedno v spodnjem trikotniku, saj zaradi (5.6) velja

$$i - p - 1 \leq t_i - 1$$

kar pomeni, da ostane levo od diagonale.

V spodnjem trikotniku matrike PA se torej ohrani vseh $\frac{1}{2}(n-p-1)(n-p)$ ničel iz spodnjega trikotnika matrike A.

Število teh ničel v vsakem stolpcu je enako kot pri matriki A (saj so le vrstice permutirane med seboj), v k-tem stolpcu jih je torej natanko

$$\max (0, n-p-k)$$

Izkaže se, da so za analizo napak bistvene le tiste ničle v spodnjem trikotniku matrike PA, ki izvirajo iz ničel v spodnjem trikotniku matrike A. Te ničle so skupaj na začetku vsake vrstice.

Definirajmo števila u_i, ki naj povedo, koliko zaporednih ničel, šteto od prvega stolpca dalje, je v i-ti vrstici matrike PA. Naj torej velja za vsak i

$$(PA)_{ik} = 0 \ , \quad k=1,\ldots,u_i$$

$$(PA)_{ik} \neq 0 \ , \quad k=u_i+1$$

Ker preide v i-to vrstico matrike PA s_i-ta vrstica matrike A, očitno velja

$$u_i = \max (0, s_i-p-1) \ , \quad i=1,\ldots,n \qquad (5.7)$$

Števila u_i imajo lastnost, da je med njimi p+1 enakih nič, vsa druga pa so neka permutacija števil $1,2,\ldots,n-p-1$.

Ker je zaradi (5.5)

$$s_i - p \leqq i$$

je od tod in iz (5.7) razvidno, da je

$$u_i \leqq i - 1 \ , \quad i=1,\ldots,n$$

kar je le druga oblika trditve (T1).

Ničle v matriki PA, ki izvirajo iz zgornjega trikotnika matrike A, se lahko bolj premešajo z elementi, ki so različni od nič. Nekatere od teh ničel morejo preiti tudi v spodnji trikotnik matrike PA, a so desno od elementov, ki niso enaki nič, zato ne vplivajo na definicijo števil u_i, ki štejejo le začetne zaporedne ničle v vrstici.

Ničle iz zgornjega trikotnika matrike A se pri vsaki permutaciji vrstic razdelijo v matriki PA v dve skupini. Prva skupina se ohrani ves čas v toku Gaussove eliminacije in te moremo s pridom upoštevati pri analizi napak. Druga skupina pa se v toku eliminacij sčasoma izgubi in teh pri analizi ne bomo upoštevali. Z upoštevanjem teh ničel bi se analiza zaokrožitvenih napak nepotrebno skomplicirala. Za posamezne ele-

mente matrike δA bi sicer lahko dobili rahlo ugodnejšo oceno, norme te matrike pa ne bi mogli na preprost način bolje oceniti.

Prvo skupino ničel opišimo takole: Naj bo v_k število zaporednih ničel matrike PA v k-tem stolpcu, šteto od prve vrstice dalje. Torej, za vsak k naj bo

$$(PA)_{ik} = 0 \ , \quad i=1,\ldots,v_k$$

$$(PA)_{ik} \neq 0 \ , \quad i=v_k+1$$

Kaj lahko povemo o številih v_k ?

Pri poljubni dopustni permutaciji vrstic matrike A *velja ocena* (T2)

$$v_k \geq \max(0,\ k-2p-1) \tag{5.8}$$

Dokaz. Naj bo $i = t_j$, to je, naj j-ta vrstica matrike A postane i-ta vrstica matrike PA.

Če je $i = j$, $j = 1, \ldots, n$, potem je $PA = A$ in je očitno

$$v_k = \max(0,\ k-p-1) \ , \quad k=1,\ldots,n \tag{5.9}$$

saj je v k-tem stolpcu matrike A pri $k=p+1,\ldots,n$ natanko $k-p-1$ začetnih zaporednih ničel. Ker pa je pri poljubni dopustni permutaciji vrstic zaradi (5.6) $i \geq j-p$, $j=1,\ldots,n$, se nobena vrstica matrike A ne pomakne za več kot p mest navzgor. To pa pomeni, da se število začetnih zaporednih ničel matrike v nobenem stolpcu ne more zmanjšati od prvotnih (5.9) za več kot p. Torej velja ocena (5.8) za poljubno dopustno permutacijo.

Mimogrede lahko omenimo, da tvorijo števila v_k nepadajoče zaporedje:

$$v_{k+1} \geq v_k \tag{5.10}$$

To je posledica dejstva, da v nobeni vrstici od nič različni elementi niso prepleteni z ničlami. Ničle, ki so levo od elementov, ki so različni od nič, pa po (T1) ostanejo v spodnjem trikotniku.

Z drugimi besedami lahko trditev (T2) formuliramo takole: Od prvotnih $\frac{1}{2}(n-p-1)(n-p)$ ničel v zgornjem trikotniku matrike A, se v matriki PA ohrani najmanj $\frac{1}{2}(n-2p-1)(n-2p)$ ničel,

ki so razporejene tako, da so nad vsako in desno od vsake same ničle.

Te ničle, ki so opisane s števili v_k, torej štete po stolpcih, moremo opisati tudi drugače.

Definirajmo števila

$$z_i = k - i \ , \quad i=1,\ldots,n$$

kjer je k najmanjši indeks, za katerega velja, da so vsi v_j, $j \geq k$ najmanj enaki i. Torej

$$v_j \geq i \ , \quad j \geq k$$

$$v_{k-1} < i$$

Z drugimi besedami: Elementi matrike PA v i-ti vrstici so pričenši z $(i+z_i)$-tim elementom vsi enaki nič:

$$(PA)_{ik} = 0 \ , \quad k=i+z_i,\ldots,n$$

Dodatno definirajmo

$$z_i = n - i + 1$$

če je $(PA)_{in} \neq 0$.

Pri tem je lahko $(PA)_{ik} = 0$ tudi pri $k < i+z_i$, toda nad temi ničlami so od nič različni elementi, ker je pri takem k $v_k < i$. Torej so s števili z_i popisane natanko iste ničle kot s števili v_k.

Pri P = I imamo očitno

$$z_i = \min(p+1, n-i+1)$$

in ker se pri nobeni dopustni permutaciji nobena vrstica matrike A ne pomakne za več kot p mest navzgor, se tudi nobeno število z_i ne more povečati za več kot p. Torej velja ocena

$$z_i \leq 2p + 1 \ , \quad i=1,\ldots,n$$

Ničle druge skupine v matriki PA, ki izvirajo iz zgornjega trikotnika matrike A, so torej take ničle, nad katerimi je vsaj po en od nič različen element, to je tak element, ki je bil v matriki A znotraj pasu.

Ničle druge skupine se v toku eliminacij izgubijo. (T3)

Dokaz. Naj bo prva ničla iz druge skupine v i-ti vrstici na k-tem mestu matrike PA. Tedaj je nad njo vsaj en od nič

različen element in naj bo (j,k)-ti tisti izmed njih, ki ima najmanjši prvi indeks. To pomeni, da je

$$v_k = j - 1 , \quad j < i \tag{5.11}$$

Naj bo B = PA, tedaj je po predpostavki

$$\begin{aligned} b_{ik} &= 0 , \quad b_{jk} \neq 0 \\ b_{rk} &= 0 , \quad r=1,\ldots,j-1 \end{aligned} \tag{5.12}$$

Dokazali bomo, da se na j-tem koraku eliminacije, ko eliminiramo j-to neznanko iz i-te enačbe, ničla b_{ik} pokvari, to je, postane od nič različno število, če ne upoštevamo eventualnih ničel iz notranjosti pasu ali ničel, ki nastanejo slučajno med računom na mestih, kjer so bili v začetku ali pozneje od nič različni elementi.

Iz formul (2.3) in (2.4) dobimo

$$b_{ik}^{(j+1)} = b_{ik}^{(j)} - \frac{b_{ij}^{(j)}}{b_{jj}^{(j)}} b_{jk}^{(j)}$$

Število $b_{ik}^{(j)}$ se ohrani kot ničla na vseh prejšnjih korakih eliminacije zaradi (5.12), toda $b_{ik}^{(j+1)}$ je v splošnem različen od nič, ker je različen od nič $b_{jk}^{(j)}$, ki se tudi ohrani na vseh prejšnjih korakih, pa tudi $b_{ij}^{(j)}$ je v splošnem od nič različen element. Kajti, če je b_{ik} prva ničla iz druge skupine v i-ti vrstici, potem je levo od te ničle min(2p+1, k-1) od nič različnih elementov in zato je ustrezni

$$u_i = \max(0, k-2p-2) \leq v_k = j - 1 \tag{5.13}$$

pri čemer smo upoštevali (5.8) in (5.11). Torej b_{ij} ne more biti ničla iz spodnjega trikotnika matrike A in je zato od nič različen element. Število $b_{ij}^{(j)}$ torej lahko postane nič samo slučajno.

Z analognim sklepanjem ugotovimo, da se tudi vse nadaljnje ničle druge skupine v i-ti vrstici izgubijo na poznejših korakih, saj velja (5.10). Le v (5.13) namesto prvega enačaja velja znak $\leq$.

S tem je trditev (T3) dokazana.

Oglejmo si sedaj nekaj primerov.

(i) Če je $r' = r$, $r=1,\ldots,n$, je $P = I$ in matrika $PA = A$, torej $(2p+1)$-diagonalna matrika. Pri $n = 11$, $p = 3$ ima naslednjo obliko:

$$A = \begin{bmatrix} * & * & * & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & 0 & 0 & 0 & 0 \\ 0 & * & * & * & * & * & * & * & 0 & 0 & 0 \\ 0 & 0 & * & * & * & * & * & * & * & 0 & 0 \\ 0 & 0 & 0 & * & * & * & * & * & * & * & 0 \\ 0 & 0 & 0 & 0 & * & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * \end{bmatrix} = A_1$$

Tu smo z znakom * označili v splošnem od nič različen element. V tem primeru veljajo naslednje enačbe:

$$\begin{aligned} u_r &= \max(0,\ r-p-1) \\ v_r &= \max(0,\ r-p-1)\ ,\ r=1,\ldots,n \\ z_r &= \min(p+1,\ n-r+1) \end{aligned} \qquad (5.14)$$

(ii) Vzemimo sedaj primer z $n = 11$, $p = 3$ in naslednjimi vrednostmi:

r	r'	s_r	t_r	u_r	v_r	z_r
1	1	1	1	0	0	4
2	5	5	5	1	0	7
3	4	4	7	0	0	6
4	6	6	3	2	0	6
5	5	2	2	0	1	5
6	9	9	4	5	1	6
7	9	3	8	0	1	5
8	9	7	10	3	1	4
9	10	10	6	6	3	3
10	10	8	9	4	5	2
11	11	11	11	7	5	1

Matrika PA ima sedaj naslednjo obliko:

$$PA = \begin{bmatrix} * & * & * & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & * & * & * & * & * & * & * & 0 & 0 & 0 \\ * & * & * & * & * & * & * & \emptyset & 0 & 0 & 0 \\ 0 & 0 & * & * & * & * & * & * & * & 0 & 0 \\ * & * & * & * & * & \emptyset & \emptyset & \emptyset & \emptyset & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & * & * & * & * & * & * \\ * & * & * & * & * & * & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \\ 0 & 0 & 0 & * & * & * & * & * & * & * & \emptyset \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * & * \\ 0 & 0 & 0 & 0 & * & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * \end{bmatrix} = A_2$$

V spodnjem trikotniku so vse ničle iz spodnjega trikotnika matrike A in sicer so tako razporejene, da ležijo skupaj, pričenši s prvim stolpcem. Ničle iz zgornjega trikotnika matrike A pa se razdelijo v dve skupini. Prvo smo tudi označili z ničlami in jih opisujejo števila v_r ali z_r. Tudi te ničle ležijo skupaj na začetku vsakega stolpca in na koncu vsake vrstice. Drugo skupino, ki se v toku računa izgubi, pa smo označili s prečrtano ničlo.

(iii) Oglejmo si še skrajni primer. Naj se v toku eliminacij izvršijo naslednje zamenjave (r,r'), r=1,...,n:

$$r' = r + p \ , \quad \text{če je} \quad r+p \leq n$$

$$r' = r \ , \quad \text{če je} \quad r+p > n$$

Pri tem se izkaže, da veljajo naslednje formule:

$$\begin{aligned} u_r &= r - 1 \ , \quad r=1,\ldots,n-p \\ u_r &= 0 \ , \quad r=n-p+1,\ldots,n \\ v_r &= \max(0,\ r-2p-1) \ , \quad r=1,\ldots,n \\ z_r &= \min(2p+1,\ n-r+1), \quad r=1,\ldots,n \end{aligned} \tag{5.15}$$

V primeru n = 11, p = 3 imamo naslednje vrednosti:

r	r'	s_r	t_r	u_r	v_r	z_r
1	4	4	10	0	0	7
2	5	5	11	1	0	7
3	6	6	9	2	0	7
4	7	7	1	3	0	7
5	8	8	2	4	0	7
6	9	9	3	5	0	6
7	10	10	4	6	0	5
8	11	11	5	7	1	4
9	9	3	6	0	2	3
10	10	1	7	0	3	2
11	11	2	8	0	4	1

Matrika PA ima torej naslednjo obliko:

$$PA = \begin{bmatrix} * & * & * & * & * & * & * & 0 & 0 & 0 & 0 \\ 0 & * & * & * & * & * & * & * & 0 & 0 & 0 \\ 0 & 0 & * & * & * & * & * & * & * & 0 & 0 \\ 0 & 0 & 0 & * & * & * & * & * & * & * & 0 \\ 0 & 0 & 0 & 0 & * & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * \\ * & * & * & * & * & * & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \\ * & * & * & * & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \\ * & * & * & * & * & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \end{bmatrix} = A_3 \qquad (5.16)$$

V tem primeru imamo najmanj ničel iz prve skupine v zgornjem trikotniku in največ iz druge skupine.

Preden nadaljujemo z analizo reševanja pasovnega sistema enačb, strnimo dosedanje ugotovitve.

Naj bo A poljubna matrika, ki more nastati iz nesingularne (2p+1)-diagonalne matrike z dopustnimi permutacijami vrstic (4.4) pri pogoju (4.5). Položaj ničel v matriki A, ki jih bomo upoštevali pri analizi, moremo opisati takole:

(i) spodnji trikotnik (i > k)

$a_{ik} = 0,\ k=1,\ldots,u_i\ ,\ i=1,\ldots,n$

(ii) zgornji trikotnik (i < k)

$a_{ik} = 0,\ i=1,\ldots,v_k\ ,\ k=1,\ldots,n$

ali

$$a_{ik} = 0,\ k=i+z_i,\ldots,n\ ,\ i=1,\ldots,n$$

Pri tem veljajo naslednje enačbe ali ocene:

$$\text{(i)}\quad u_i = \max(0,\ s_i-p-1)\ ,\quad i=1,\ldots,n \tag{5.17}$$

kjer so s_i , $i=1,\ldots,n$ neka permutacija števil $1,\ldots,n$, pri čemer velja

$$s_i \leqq i + p \tag{5.18}$$

in

$$u_i \leqq i - 1 \tag{5.19}$$

$$\text{(ii)}\quad v_k \geqq \max(0,\ k-2p-1)\ ,\quad k=1,\ldots,n \tag{5.20}$$

$$\text{(iii)}\quad z_i \leqq \min(2p+1,\ n-i+1)\ ,\quad i=1,\ldots,n \tag{5.21}$$

Matrika A ima torej v spodnjem trikotniku natanko $\frac{1}{2}(n-p-1)(n-p)$ ničel, ki so razporejene tako, da ležijo v vsaki vrstici skupaj na začetku in jih je v k-tem stolpcu natanko $\max(0,\ n-p-k)$.

V zgornjem trikotniku pa je najmanj $\frac{1}{2}(n-2p-1)(n-2p)$ ničel, ki so razporejene tako, da ležijo skupaj na začetku vsakega stolpca in skupaj na koncu vsake vrstice.

6. Struktura matrik L in U

Oglejmo si sedaj razcep matrike A na produkt spodnje trikotne matrike L z enojkami na diagonali in zgornje trikotne matrike U. Tudi matriki L in U imata posebno strukturo glede položaja ničel, ki izvira iz posebne strukture matrike A.

Vse ničle iz spodnjega trikotnika matrike A *se ohranijo v matriki* L. (T4)

Dokaz. Matrika L je bila v splošnem primeru definirana s formulami (2.9) in (2.3). Naj velja pri nekem i

$$m_{ij} = 0\ ,\quad j=1,\ldots,r\ ,\quad r < u_i$$

Potem sledi iz formul (2.4), da je

$$a_{ik}^{(j+1)} = a_{ik}^{(j)} , \quad j=1,\ldots,r , \quad k=j+1,\ldots,n \qquad (6.1)$$

Ker pa je $a_{ik}^{(1)} = a_{ik} = 0$ pri $k \leq u_i$, sledi iz (6.1), da je

$$a_{ik}^{(r+1)} = 0 , \quad k=r+1,\ldots,n$$

Tedaj pa je po formuli (2.3) tudi

$$m_{i,r+1} = a_{i,r+1}^{(r+1)} / a_{r+1,r+1}^{(r+1)} = 0$$

Pri r = 0 je trditev trivialna, saj je $m_{i1} = a_{i1}/a_{11} = 0$, če je $u_i > 0$.

Torej pri vsakem i velja

$$\ell_{ik} = 0 , \quad k=1,\ldots,u_i \qquad (6.2)$$

Vse ničle iz zgornjega trikotnika matrike A *se ohranijo v matriki* U. (T5)

Dokaz. Matrika U je bila v splošnem primeru definirana kot $A^{(n)}$ z elementi (2.8).

Naj velja pri nekem k

$$u_{jk} = a_{jk}^{(j)} = 0 , \quad j=1,\ldots,r , \quad r < v_k$$

Potem sledi iz formul (2.4)

$$a_{ik}^{(j+1)} = a_{ik}^{(j)} , \quad j=1,\ldots,r , \quad i=j+1,\ldots,n$$

in ker je

$$a_{ik}^{(1)} = 0 , \quad i=1,\ldots,v_k$$

je

$$a_{ik}^{(j+1)} = 0 , \quad i=j+1,\ldots,v_k$$

Torej je tudi pri j = r in i = j+1

$$u_{r+1,k} = a_{r+1,k}^{(r+1)} = 0$$

Pri r = 0 je trditev trivialna, saj je $u_{1k} = a_{1k}^{(1)} = a_{1k} = 0$, če je $v_k > 0$.

Torej je

$$u_{ik} = 0 , \quad i=1,\ldots,v_k$$

ali

$$u_{ik} = 0 \ , \quad k=i+z_i,\ldots,n \tag{6.3}$$

7. Analiza zaokrožitvenih napak

Najprej na kratko ponovimo glavne korake analize zaokrožitvenih napak pri reševanju sistema linearnih enačb v splošnem primeru (razdelek 3), nato pa upoštevajmo posebnosti obravnavanega pasovnega primera.

Izračunana rešitev x sistema enačb

$$Ax = b$$

je eksaktna rešitev sistema

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

kjer je perturbacijska matrika podana z izrazom

$$\delta A = E + \delta L \cdot U + L \cdot \delta U + \delta L \cdot \delta U \tag{7.1}$$

in veljajo eksaktno enačbe

$$A + E = L \cdot U$$

$$(L + \delta L)y = b \tag{7.2}$$

$$(U + \delta U)x = y \tag{7.3}$$

Za norme nastopajočih matrik smo dobili ocene:

$$\|E\|_p \leq 2{\cdot}01 \text{ ga } \|E'\|_p \tag{7.4}$$

$$\|L\|_p \leq 1{\cdot}12 \text{ a } \|\delta L'\|_p \tag{7.5}$$

$$\|U\|_p \leq 1{\cdot}12 \text{ ga } \|\delta U'\|_p \tag{7.6}$$

$$\|L\|_p \leq \|L'\|_p \tag{7.7}$$

$$\|U\|_p \leq g \|U'\|_p \tag{7.8}$$

za $p = 1,\infty$, E. Tu je a osnovna zaokrožitvena napaka, g maksimalni nastopajoči element, matrike E', δL', δU', L' in U' pa so matrike s celoštevilčnimi elementi, ki dajo pri danem načinu ocenjevanja skupaj z ustreznim faktorjem na desni strani ocene za ustrezne elemente matrike na levi.

V splošnem primeru smo dobili za te elemente naslednje vrednosti:

$$e'_{ik} = i - 1 \ , \quad k \geq i$$
$$e'_{ik} = k \ , \quad k < i$$
$$\delta\ell'_{ii} = i - 1$$
$$\delta\ell'_{ik} = k \ , \quad k < i$$
$$\delta u'_{ik} = k - i \ , \quad k > i$$
$$\delta u'_{ii} = n - i + 1$$
$$\ell'_{ik} = 1 \ , \quad k \leq i$$
$$u'_{ik} = 1 \ , \quad k \geq i$$

Nedefinirani elementi so vsi enaki 0.

Preden na kratko ponovimo analizo posameznih korakov z upoštevanjem ničel, to je števil u_i, v_k in z_i, dodatno definirajmo še števila

$$w_{ik} = \max(u_i, v_k) \ , \quad i,k=1,\ldots,n \tag{7.9}$$

a) Matrika E

Pri razcepu matrike A na produkt LU smo v splošnem primeru imeli formule (3.1) - (3.11) za r=1,...,n-1. Ponovimo samo najpomembnejše:

$$m_{ir} = (a_{ir}^{(r)}/a_{rr}^{(r)})(1 + \varepsilon_1)$$
$$\varepsilon_{ir}^{(r)} = \varepsilon_1 a_{ir}^{(r)} \ , \quad i=r+1,\ldots,n \tag{7.10}$$
$$a_{ik}^{(r+1)} = (a_{ik}^{(r)} - m_{ir}a_{rk}^{(r)}(1+\varepsilon_2))(1+\varepsilon_3)$$
$$\varepsilon_{ik}^{(r)} = \frac{\varepsilon_3}{1+\varepsilon_3} a_{ik}^{(r+1)} - m_{ir}a_{rk}^{(r)}\varepsilon_2 \tag{7.11}$$
$$i,k=r+1,\ldots,n$$
$$e_{ik} = \sum_{r=1}^{i-1} \varepsilon_{ik}^{(r)} \ , \quad k \geq i \tag{7.12}$$
$$e_{ik} = \sum_{r=1}^{k} \varepsilon_{ik}^{(r)} \ , \quad k < i \tag{7.13}$$

Ker smo po (3.16) vse $\varepsilon_{ik}^{(r)}$ ocenili z istim številom 2·01 ga, potem pomeni število e'_{ik} število od nič različnih

sumandov v vsotah (7.12) in (7.13).

Sedaj pa upoštevajmo, da pri trivialni operaciji, to je pri množenju z 0 ali pri prištetju števila 0, ne nastane nobena zaokrožitvena napaka.

Napaka ε_1 je nič, če je $a_{ir}^{(r)} = 0$, ali, kar je isto, če je $m_{ir} = \ell_{ir} = 0$. Ugotovili smo (T4), da je to pri danem i za $r=1,\ldots,u_i$.

Napaka ε_2 pa je nič in hkrati z njo tudi ε_3 takrat, ko je produkt $m_{ir}a_{rk}^{(r)} = \ell_{ir}u_{rk} = 0$. To pa je takrat, ko je vsaj eden od obeh faktorjev enak nič. Prvi faktor je nič po (T4) za $r=1,\ldots,u_i$, drugi pa po (T5) za $r=1,\ldots,v_k$. Torej sta glede na definicijo (7.9) ε_2 in ε_3 enaka nič za $r=1,\ldots,w_{ik}$.

Na tem mestu je razvidno, da bi bilo možno, da je samo ε_3 enak nič, če je $a_{ik}^{(r)} = 0$, produkt $\ell_{ir}u_{rk}$ pa ne. To so ničle iz druge skupine, o katerih smo rekli, da jih ne bomo upoštevali. V tem primeru bi za ustrezni $\varepsilon_{ik}^{(r)}$ iz (7.11) dobili oceno ga namesto 2·01 ga. Z upoštevanjem takih ničel bi ne mogli bistveno izboljšati ocene za normo matrike E.

Namesto enačbe (7.10) imamo torej za vsak i:

$$\varepsilon_{ir}^{(r)} = 0 \ , \quad r=1,\ldots,u_i$$

$$\varepsilon_{ir}^{(r)} = \varepsilon_1 a_{ir}^{(r)} \ , \quad r=u_i+1,\ldots,i-1$$

in namesto (7.11) za vsak i in k

$$\varepsilon_{ik}^{(r)} = 0 \ , \quad r=1,\ldots,w_{ik}$$

$$\varepsilon_{ik}^{(r)} = \frac{\varepsilon_3}{1+\varepsilon_3} a_{ik}^{(r+1)} - m_{ir}a_{rk}^{(r)}\varepsilon_2$$

$$r=w_{ik}+1,\ldots,\min(i-1,k-1)$$

Torej veljajo v našem primeru namesto (7.12) in (7.13) formule:

$$e_{ik} = \sum_{r=w_{ik}+1}^{i-1} \varepsilon_{ik}^{(r)} \ , \quad e'_{ik} = \max(0,i-1-w_{ik}) \ , \ k \geq i \qquad (7.14)$$

$$e_{ik} = \sum_{r=w_{ik}+1}^{k-1} \varepsilon_{ik}^{(r)} + \varepsilon_{ik}^{(k)}$$

$$e'_{ik} = \max(0,\ k-w_{ik}) \ ,\ k < i \qquad (7.15)$$

Pri (7.15) je treba pojasniti, da so vsi $\varepsilon_{ik}^{(r)}$, $r < k$ enaki nič, če je $\varepsilon_{ik}^{(k)} = 0$, kajti iz $\ell_{ik} = 0$ sledi, da je $\ell_{ir} = 0$, $r < k$. Ocena (3.16) seveda ostane pri tem v veljavi, zato je

$$|e_{ik}| \leq 2{\cdot}01\ ga\ e'_{ik}$$

Za ilustracijo navedimo matrike E' za naše tri primere matrik, ki smo si jih ogledali v razdelku 5.

Iz (7.14) in (7.15) dobimo za te tri primere naslednje matrike:

$$E'_1 = \begin{bmatrix} 0&0&0&0&0&0&0&0&0&0&0\\ 1&1&1&1&0&0&0&0&0&0&0\\ 1&2&2&2&1&0&0&0&0&0&0\\ 1&2&3&3&2&1&0&0&0&0&0\\ 0&1&2&3&3&2&1&0&0&0&0\\ 0&0&1&2&3&3&2&1&0&0&0\\ 0&0&0&1&2&3&3&2&1&0&0\\ 0&0&0&0&1&2&3&3&2&1&0\\ 0&0&0&0&0&1&2&3&3&2&1\\ 0&0&0&0&0&0&1&2&3&3&2\\ 0&0&0&0&0&0&0&1&2&3&3 \end{bmatrix}$$

$$E'_2 = \begin{bmatrix} 0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 1&2&2&2&1&1&1&1&0&0&0\\ 0&0&1&1&1&1&1&1&0&0&0\\ 1&2&3&4&3&3&3&3&1&0&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 1&2&3&4&4&5&5&5&3&1&1\\ 0&0&0&1&2&3&4&4&4&2&2\\ 0&0&0&0&0&0&1&2&2&2&2\\ 0&0&0&0&1&2&3&4&5&4&4\\ 0&0&0&0&0&0&0&1&2&3&3 \end{bmatrix} \qquad E'_3 = \begin{bmatrix} 0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 1&2&3&4&5&6&7&7&6&5&4\\ 1&2&3&4&5&6&7&7&7&6&5\\ 1&2&3&4&5&6&7&7&7&7&6 \end{bmatrix}$$

b) Matriki L in δL

Za matriko L smo že v (T4) ugotovili, kje ima ničle. Za druge elemente pa velja

$$|\ell_{ik}| \leq \ell'_{ik} \ ,\ \ell'_{ik} = 1 \ ,\ k=u_i+1,\ldots,i \ ,\ i \geq k \qquad (7.16)$$

Pri analizi reševanja spodnjega trikotnega sistema bomo upoštevali ničle v matriki L.

V formuli (3.19) upoštevajmo (6.2). Tako dobimo, če sledimo splošni analizi (3.18) - (3.24):

$$y_i = b_i(1 + \zeta_i) - \sum_{k=u_i+1}^{i-1} \ell_{ik} y_k (1 + \zeta_k) \qquad (7.17)$$

kjer je

$$1 + \zeta_i = (1 + \eta_{u_i+1}) \ldots (1 + \eta_{i-1})$$

$$1 + \zeta_k = (1 + \varepsilon_k)(1 + \eta_k) \ldots (1 + \eta_{i-1})$$

in veljajo ocene

$$|\varepsilon_k| \leq a \ , \quad |\eta_j| \leq a$$

Enačbo (7.17) zapišimo v obliki

$$b_i = \sum_{k=u_i+1}^{i} \ell_{ik} y_k (1 + \xi_k)$$

kjer je

$$1 + \xi_i = (1+\zeta_i)^{-1} = (1+\eta_{u_i+1})^{-1} \ldots (1+\eta_{i-1})^{-1}$$

$$1 + \xi_k = (1+\zeta_k)(1+\zeta_i)^{-1} = (1+\varepsilon_k)(1+\eta_{u_i+1})^{-1} \ldots (1+\eta_{k-1})^{-1}$$

Od tod sledi, da je izračunani vektor y eksaktna rešitev sistema (7.2), pri čemer veljajo ocene

$$|\delta\ell_{ii}| \leq 1{,}12\ a\ \delta\ell'_{ii} \ , \quad \delta\ell'_{ii} = i - 1 - u_i$$

$$|\delta\ell_{ik}| \leq 1{,}12\ a\ \delta\ell'_{ik} \ , \quad \delta\ell'_{ik} = \max(0, k-u_i), k<i \qquad (7.18)$$

Matriki L' in δL' sta v naših treh zgledih enaki:

$$L'_1 = \begin{bmatrix}
1 & & & & & & & & & & \\
1 & 1 & & & & & & & & & \\
1 & 1 & 1 & & & & & & & & \\
1 & 1 & 1 & 1 & & & & & & & \\
0 & 1 & 1 & 1 & 1 & & & & & & \\
0 & 0 & 1 & 1 & 1 & 1 & & & & & \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & & & & \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & & & \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1
\end{bmatrix} \qquad \delta L'_1 = \begin{bmatrix}
0 & & & & & & & & & & \\
1 & 1 & & & & & & & & & \\
1 & 2 & 2 & & & & & & & & \\
1 & 2 & 3 & 3 & & & & & & & \\
0 & 1 & 2 & 3 & 3 & & & & & & \\
0 & 0 & 1 & 2 & 3 & 3 & & & & & \\
0 & 0 & 0 & 1 & 2 & 3 & 3 & & & & \\
0 & 0 & 0 & 0 & 1 & 2 & 3 & 3 & & & \\
0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 3 & & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 3 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 3
\end{bmatrix}$$

$$L_2' = \begin{bmatrix} 1 & & & & & & & & & & \\ 0 & 1 & & & & & & & & & \\ 1 & 1 & 1 & & & & & & & & \\ 0 & 0 & 1 & 1 & & & & & & & \\ 1 & 1 & 1 & 1 & 1 & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 1 & & & & & \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & & & & \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & & \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad \delta L_2' = \begin{bmatrix} 0 & & & & & & & & & & \\ 0 & 0 & & & & & & & & & \\ 1 & 2 & 2 & & & & & & & & \\ 0 & 0 & 1 & 1 & & & & & & & \\ 1 & 2 & 3 & 4 & 4 & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & & & \\ 1 & 2 & 3 & 4 & 5 & 6 & 6 & & & & \\ 0 & 0 & 0 & 1 & 2 & 3 & 4 & 4 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 & & \\ 0 & 0 & 0 & 0 & 1 & 2 & 3 & 4 & 5 & 5 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 3 \end{bmatrix}$$

$$L_3' = \begin{bmatrix} 1 & & & & & & & & & & \\ 0 & 1 & & & & & & & & & \\ 0 & 0 & 1 & & & & & & & & \\ 0 & 0 & 0 & 1 & & & & & & & \\ 0 & 0 & 0 & 0 & 1 & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 1 & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & & & \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & & \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad \delta L_3' = \begin{bmatrix} 0 & & & & & & & & & & \\ 0 & 0 & & & & & & & & & \\ 0 & 0 & 0 & & & & & & & & \\ 0 & 0 & 0 & 0 & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 8 & & \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 9 & \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 10 \end{bmatrix}$$

c) Matriki U in δU

V (T5) smo ugotovili, kje ima matrika U ničle. Za druge elemente veljajo ocene:

$$|u_{ik}| \leq g\, u'_{ik} \quad , \quad u'_{ik} = 1 \ , \ k=i,\ldots,i+z_i-1 \qquad (7.19)$$

$$\text{ali} \qquad u'_{ik} = 1 \ , \ i=v_k+1,\ldots,k$$

Pri analizi reševanja zgornjega trikotnega sistema bomo upoštevali ničle matrike U, določene s (6.3).

Tako dobimo, če sledimo splošni analizi po formulah (3.25) - (3.27):

$$x_i = (y_i(1+\zeta_i) - \sum_{k=i+1}^{i+z_i-1} u_{ik}x_k(1+\zeta_k))(1+\varepsilon_i)/u_{ii} \qquad (7.20)$$

kjer je

$$1 + \zeta_i = (1 + \eta_{i+1}) \ldots (1 + \eta_{i+z_i-1})$$

$$1 + \zeta_k = (1 + \varepsilon_k)(1 + \eta_k) \ldots (1 + \eta_{i+z_i-1})$$

in veljajo ocene

$$|\varepsilon_j| \leq a \ , \quad |\eta_j| \leq a$$

Enačbo (7.20) moremo zapisati v obliki

$$y_i = \sum_{k=i}^{i+z_i-1} u_{ik} x_k (1 + \xi_k)$$

kjer je

$$1+\xi_i = (1+\varepsilon_i)^{-1}(1+\zeta_i)^{-1} = (1+\varepsilon_i)^{-1}(1+\eta_{i+1})^{-1}\ldots(1+\eta_{i+z_i-1})^{-1}$$

$$1+\xi_k = (1+\zeta_k)(1+\zeta_i)^{-1} = (1+\varepsilon_k)(1+\eta_{i+1})^{-1}\ldots(1+\eta_{k-1})^{-1}$$

Izračunana rešitev x je torej eksaktna rešitev sistema (7.3), pri čemer veljajo ocene

$$\begin{aligned} |\delta u_{ii}| &\leq 1{\cdot}12 \text{ ga } \delta u'_{ii}, \quad \delta u'_{ii} = z_i \\ |\delta u_{ik}| &\leq 1{\cdot}12 \text{ ga } \delta u'_{ik}, \quad \delta u'_{ik} = k-i,\ i<k<i+z_i \end{aligned} \tag{7.21}$$

Vsi drugi elementi so enaki nič.

Matriki U' in δU' sta v naših treh primerih enaki:

$$U'_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ & & & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ & & & & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ & & & & & 1 & 1 & 1 & 1 & 0 & 0 \\ & & & & & & 1 & 1 & 1 & 1 & 0 \\ & & & & & & & 1 & 1 & 1 & 1 \\ & & & & & & & & 1 & 1 & 1 \\ & & & & & & & & & 1 & 1 \\ & & & & & & & & & & 1 \end{bmatrix} \qquad \delta U'_1 = \begin{bmatrix} 4 & 1 & 2 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 4 & 1 & 2 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 4 & 1 & 2 & 3 & 0 & 0 & 0 & 0 & 0 \\ & & & 4 & 1 & 2 & 3 & 0 & 0 & 0 & 0 \\ & & & & 4 & 1 & 2 & 3 & 0 & 0 & 0 \\ & & & & & 4 & 1 & 2 & 3 & 0 & 0 \\ & & & & & & 4 & 1 & 2 & 3 & 0 \\ & & & & & & & 4 & 1 & 2 & 3 \\ & & & & & & & & 3 & 1 & 2 \\ & & & & & & & & & 2 & 1 \\ & & & & & & & & & & 1 \end{bmatrix}$$

$$U'_2 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ & & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ & & & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ & & & & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ & & & & & 1 & 1 & 1 & 1 & 1 & 1 \\ & & & & & & 1 & 1 & 1 & 1 & 1 \\ & & & & & & & 1 & 1 & 1 & 1 \\ & & & & & & & & 1 & 1 & 1 \\ & & & & & & & & & 1 & 1 \\ & & & & & & & & & & 1 \end{bmatrix} \qquad \delta U'_2 = \begin{bmatrix} 4 & 1 & 2 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 7 & 1 & 2 & 3 & 4 & 5 & 6 & 0 & 0 & 0 \\ & & 6 & 1 & 2 & 3 & 4 & 5 & 0 & 0 & 0 \\ & & & 6 & 1 & 2 & 3 & 4 & 5 & 0 & 0 \\ & & & & 5 & 1 & 2 & 3 & 4 & 0 & 0 \\ & & & & & 6 & 1 & 2 & 3 & 4 & 5 \\ & & & & & & 5 & 1 & 2 & 3 & 4 \\ & & & & & & & 4 & 1 & 2 & 3 \\ & & & & & & & & 3 & 1 & 2 \\ & & & & & & & & & 2 & 1 \\ & & & & & & & & & & 1 \end{bmatrix}$$

$$U_3' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ & & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ & & & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ & & & & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & & & & & 1 & 1 & 1 & 1 & 1 & 1 \\ & & & & & & 1 & 1 & 1 & 1 & 1 \\ & & & & & & & 1 & 1 & 1 & 1 \\ & & & & & & & & 1 & 1 & 1 \\ & & & & & & & & & 1 & 1 \\ & & & & & & & & & & 1 \end{bmatrix} \qquad \delta U_3' = \begin{bmatrix} 7 & 1 & 2 & 3 & 4 & 5 & 6 & 0 & 0 & 0 & 0 \\ & 7 & 1 & 2 & 3 & 4 & 5 & 6 & 0 & 0 & 0 \\ & & 7 & 1 & 2 & 3 & 4 & 5 & 6 & 0 & 0 \\ & & & 7 & 1 & 2 & 3 & 4 & 5 & 6 & 0 \\ & & & & 7 & 1 & 2 & 3 & 4 & 5 & 6 \\ & & & & & 6 & 1 & 2 & 3 & 4 & 5 \\ & & & & & & 5 & 1 & 2 & 3 & 4 \\ & & & & & & & 4 & 1 & 2 & 3 \\ & & & & & & & & 3 & 1 & 2 \\ & & & & & & & & & 2 & 1 \\ & & & & & & & & & & 1 \end{bmatrix}$$

8. Ocene norm nastopajočih matrik

Poskusimo sedaj čimnatančneje oceniti vse tri norme petih nastopajočih matrik E', $\delta L'$, $\delta U'$, L' in U'. Začnimo pri enostavnejših računih.

a) Matrika L'

Od nič različni elementi matrike L' so podani s (7.16), poleg tega pa vemo po (T4) in (T1), da je v spodnjem trikotniku matrike L' natanko $\frac{1}{2}(n-p-1)(n-p)$ ničel in sicer jih je v k-tem stolpcu natanko $\max(0,n-p-k)$.

Izračun vseh treh norm matrike L' je preprost.

$$\text{(i)} \qquad \|L'\|_1 = \max_k \sum_{i=1}^{n} \ell'_{ik} = \max_k \sum_{i=k}^{n} \ell'_{ik} =$$

$$= \max_k \left((n-k+1) - \max(0,n-p-k)\right) =$$

$$= \max_k \min(n-k+1, p+1) = p + 1$$

Zadnji enačaj velja, če je $n > p$, kar je vedno zaradi omejitve (4.2), ki je ne bomo več povdarjali.

Iz (7.7) potem sledi, da velja

$$\|L\|_1 \leq p + 1$$

$$\text{(ii)} \qquad \|L'\|_\infty = \max_i \sum_{k=1}^{n} \ell'_{ik} = \max_i \sum_{k=u_i+1}^{i} \ell'_{ik} =$$

$$= \max_i (i - u_i) \leq n$$

Iz (7.7) potem sledi

$$\|L\|_\infty \leq n$$

(iii) $$\|L'\|_E^2 = \sum_{i=1}^{n} \sum_{k=1}^{n} \ell_{ik}'^2 = \frac{1}{2}n(n+1) - \frac{1}{2}(n-p-1)(n-p) =$$

$$= \frac{1}{2}(p+1)(2n-p)$$

Zato velja

$$\|L\|_E \leq \left(\frac{1}{2}(p+1)(2n-p)\right)^{1/2}$$

b) Matrika U'

Od nič različni elementi matrike U' so podani s (7.18).

(i) $$\|U'\|_1 = \max_k \sum_{i=1}^{n} u_{ik}' = \max_k \sum_{i=v_k+1}^{k} u_{ik}' =$$

$$= \max_k (k - v_k)$$

Ker velja (5.20) za vsak k, je

$$k - v_k \leq 2p + 1 \qquad (8.1)$$

Zato je

$$\|U'\|_1 \leq 2p + 1$$

in iz (7.8) sledi

$$\|U\|_1 \leq g(2p + 1)$$

(ii) $$\|U'\|_\infty = \max_i \sum_{k=1}^{n} u_{ik}' = \max_i \sum_{k=i}^{i+z_i-1} u_{ik}' = \max_i z_i$$

Iz (5.21) potem sledi

$$\|U'\|_\infty \leq 2p + 1$$

in

$$\|U\|_\infty \leq g(2p + 1)$$

(iii) $$\|U'\|_E^2 = \sum_{i=1}^{n} \sum_{k=1}^{n} u_{ik}'^2 = \sum_{i=1}^{n} \sum_{k=i}^{i+z_i-1} u_{ik}'^2 = \sum_{i=1}^{n} z_i$$

Iz ocene (5.21) dobimo

$$\|U'\|_E^2 \leq \sum_{i=1}^{n} \min(2p+1, n-i+1) = (2p+1)(n-p)$$

Torej velja splošno

$$\|U\|_E \leq g((2p+1)(n-p))^{1/2}$$

c) Matrika $\delta U'$

Od nič različni elementi matrike $\delta U'$ so podani s (7.21).

$$\text{(i)} \quad \|\delta U'\|_1 = \max_k \sum_{i=1}^{n} \delta u'_{ik} = \max_k \Big(\sum_{i=v_k+1}^{k-1} (k-i) + z_k \Big) =$$

$$= \max_k \left(\tfrac{1}{2}(k-v_k-1)(k-v_k) + z_k \right)$$

Če upoštevamo oceni (8.1) in (5.21), dobimo od tod

$$\|\delta U'\|_1 \leq (p+1)(2p+1)$$

Torej je zaradi (7.6)

$$\|\delta U\|_1 \leq 1{,}12\ ga\ (p+1)(2p+1)$$

$$\text{(ii)} \quad \|\delta U'\|_\infty = \max_i \sum_{k=1}^{n} \delta u'_{ik} = \max_i \sum_{k=i}^{i+z_i-1} \delta u'_{ik} =$$

$$= \max_i \Big(z_i + \sum_{k=i+1}^{i+z_i-1} (k-i) \Big) =$$

$$= \max_i \tfrac{1}{2} z_i (z_i + 1)$$

Iz ocene (5.21) potem sledi

$$\|\delta U'\|_\infty \leq (p+1)(2p+1)$$

in

$$\|\delta U\|_\infty \leq 1{,}12\ ga\ (p+1)(2p+1)$$

$$\text{(iii)} \quad \|\delta U'\|_E^2 = \sum_{i=1}^{n} \sum_{k=1}^{n} \delta u'^2_{ik} =$$

$$= \sum_{i=1}^{n} \Big(z_i^2 + \sum_{k=i+1}^{i+z_i-1} (k-i)^2 \Big) =$$

$$= \sum_{i=1}^{n} \tfrac{1}{6} z_i (z_i+1)(2z_i+1) \qquad (8.2)$$

Spet upoštevajmo oceno (5.21) in vsoto v (8.2) primerno raz-
delimo:

$$\|\delta U'\|_E^2 \leq \sum_{i=1}^{n-2p} \frac{1}{6}(2p+1)(2p+2)(4p+3) +$$

$$+ \sum_{i=n-2p+1}^{n} \frac{1}{6}(n-i+1)(n-i+2)(2n-2i+3) =$$

$$= \frac{1}{3}(p+1)(2p+1)((4p+3)n - p(6p+5))$$

Od tod sledi

$$\|\delta U\|_E \leq 1{,}12\ ga\left(\frac{1}{3}(p+1)(2p+1)((4p+3)n - p(6p+5))\right)^{1/2}$$

d) Matrika $\delta L'$

Od nič različni elementi matrike $\delta L'$ so podani s (7.18).

$$\text{(i)} \qquad \|\delta L'\|_1 = \max_k \sum_{i=1}^{n} \delta \ell'_{ik} =$$

$$= \max_k \left(\sum_{i=k+1}^{n} \max(0, k-u_i) + k - u_k - 1 \right) =$$

$$= \max_k \left(\sum_{i=k}^{n} \max(0, k-u_i) - 1 \right) \qquad (8.3)$$

Naj bo najprej $k > n-p$. Tedaj je

$$\max_{k>n-p} \left(\sum_{i=k}^{n} \max(0, k-u_i) - 1 \right) \leq \max_{k>n-p} (k(n-k+1) - 1) =$$

$$= np - p^2 + p - 1 \qquad (8.4)$$

Namreč funkcija $-k^2 + (n+1)k - 1$ je nenaraščajoča za $k > n-p$, saj je odvod $-2k + n+1$ manjši od $-n + 2p+1$, kar je po predpostavki (4.2) nepozitivno število. Zato je maksimum dosežen pri $k = n-p+1$.

Sedaj pa naj bo $k \leq n-p$. Najprej ocenimo vsoto v oklepaju v (8.3):

$$\sum_{i=k}^{n} \max(0, k-u_i) = \sum_{i=1}^{n} \max(0, k-u_i) - \sum_{i=1}^{k-1} \max(0, k-u_i)$$

V prvi vsoti pišimo namesto u_i izraz (5.17) in preuredimo sumacijo po indeksu $j = s_i$ (j pomeni indeks vrstice prvotne pasovne matrike, ki preide v i-to vrstico obravnavane matrike). V drugi vsoti pa upoštevajmo oceno (5.19). Tako dobimo

$$\sum_{i=k}^{n} \max(0,k-u_i) \leq \sum_{j=1}^{n} \max(0,k-\max(0,j-p-1)) - \sum_{i=1}^{k-1} \max(0,k-i+1) = pk + 1$$

Torej je

$$\max_{k\leq n-p} \sum_{i=1}^{n} \delta l'_{ik} \leq \max_{k\leq n-p} pk = p(n-p) \leq$$

$$\leq pn - p^2 + p - 1 \qquad (8.5)$$

Oceni (8.4) in (8.5) skupaj nam dasta

$$\|\delta L'\|_1 \leq pn - p^2 + p - 1$$

Iz (7.5) potem sledi

$$\|\delta L\|_1 \leq 1{\cdot}12\ a\ (pn - p^2 + p - 1)$$

(ii)
$$\|\delta L'\|_\infty = \max_i \sum_{k=1}^{n} \delta l'_{ik} =$$

$$= \max_i \Big(\sum_{k=1}^{i-1} \max(0,k-u_i) + i-1-u_i \Big) =$$

$$= \max_i \Big(\sum_{k=u_i+1}^{i-1} (k-u_i) + i-1-u_i \Big) =$$

$$= \max_i \tfrac{1}{2}(i-u_i-1)(i-u_i+2) \leq$$

$$\leq \max_i \tfrac{1}{2}(i-1)(i+2) = \tfrac{1}{2}(n-1)(n+2)$$

Torej velja tudi

$$\|\delta L\|_\infty \leq 1{\cdot}12\ a\ \tfrac{1}{2}(n-1)(n+2)$$

(iii)
$$\|\delta L'\|_E^2 = \sum_{i=1}^{n} \sum_{k=1}^{n} \delta l'^2_{ik} =$$

$$= \sum_{i=1}^{n} \Big(\sum_{k=u_i+1}^{i-1} (k-u_i)^2 + (i-u_i-1)^2 \Big) =$$

$$= \tfrac{1}{6} \sum_{i=1}^{n} (2i^3+3i^2-11i+6-2u_i^3+3u_i^2+11u_i) -$$

$$- \sum_{i=1}^{n} iu_i(i-u_i+1) \qquad (8.6)$$

Če upoštevamo enačbo (5.17) in pišemo $j = s_i$, dobimo

$$\sum_{i=1}^{n} u_i = \sum_{j=1}^{n} \max(0,j-p-1) = \frac{1}{2}(n-p-1)(n-p) \qquad (8.7)$$

$$\sum_{i=1}^{n} u_i^2 = \sum_{j=1}^{n} \max(0,j-p-1)^2 = \frac{1}{6}(n-p-1)(n-p)(2n-2p-1) \qquad (8.8)$$

$$\sum_{i=1}^{n} u_i^3 = \sum_{j=1}^{n} \max(0,j-p-1)^3 = \frac{1}{4}(n-p-1)^2(n-p)^2 \qquad (8.9)$$

Prva vsota v (8.6) se da torej izračunati, drugo pa bomo ocenili z upoštevanjem enačbe (5.17) in ocene (5.18) pri $j=s_i$:

$$\sum_{i=1}^{n} iu_i(i-u_i+1) = \sum_{j=1}^{n} i\max(0,j-p-1)(i-\max(0,j-p-1)+1) =$$

$$= \sum_{j=p+2}^{n} i(j-p-1)(i-(j-p-1)+1) \geq \sum_{j=p+2}^{n} 2(j-p)(j-p-1) =$$

$$= 2\left(\frac{1}{6}(n-p-1)(n-p)(2n-2p-1) + \frac{1}{2}(n-p-1)(n-p)\right) \qquad (8.10)$$

Če združimo (8.6) z izračuni (8.7) - (8.9) in z oceno (8.10), dobimo

$$\|\delta L'\|_E^2 \leq \frac{p}{12}(4n^3-6(p-2)n^2+2(2p^2-6p-7)n-(p^3-4p^2-7p-2))$$

Torej je

$$\|\delta L\|_E \leq 1{\cdot}12\ a\left(\frac{p}{12}(4n^3-6(p-2)n^2+2(2p^2-6p-7)n-(p^3-4p^2-7p-2))\right)^{1/2}$$

e) Matrika E'

Elementi matrike E' so bili definirani s formulama (7.14) in (7.15). Pri tem se spomnimo na definicijo (7.9).

(i) $$\|E'\|_1 = \max_k \sum_{i=1}^{n} e'_{ik} =$$

$$= \max_k \left(\sum_{i=1}^{k} \max(0,i-1-w_{ik}) + \sum_{i=k+1}^{n} \max(0,k-w_{ik}) \right)$$

Zaradi ocene (5.20) velja

$$w_{ik} = \max(u_i,v_k) \geq \max(u_i,k-2p-1) \qquad (8.11)$$

Od tod potem sledi

$$\|E'\|_1 \leq \max_k \Big(\sum_{i=1}^{k} \max(0,i-1-\max(u_i,k-2p-1)) + \sum_{i=k+1}^{n} \max(0,k-\max(u_i,k-2p-1)) \Big) \qquad (8.12)$$

Najprej vzemimo, da je

$$n \geq 3p + 1 \qquad (8.13)$$

in razdelimo množico $1 \leq k \leq n$ na tri dele: $1 \leq k \leq 2p+1$, $2p+1 < k \leq n-p$, $n-p < k \leq n$ in poiščimo maksimum (8.12) na vsakem delu posebej.

Na prvi množici velja $\max(u_i,k-2p-1) = u_i$ in zato imamo, ker je $u_i \geq 0$ in $u_i \leq i-1$:

$$\max_{1\leq k\leq 2p+1} \Big(\sum_{i=1}^{k} (i-1-u_i) + \sum_{i=k+1}^{n} \max(0,k-u_i) \Big) =$$

$$= \max_{1\leq k\leq 2p+1} \Big(\sum_{i=1}^{k} (i-1-u_i - \max(0,k-u_i)) + \sum_{i=1}^{n} \max(0,k-u_i) \Big) \qquad (8.14)$$

V drugi vsoti upoštevajmo zvezo (5.17) in preuredimo sumacijo po indeksu $j = s_i$:

$$\sum_{i=1}^{n} \max(0,k-u_i) = \sum_{j=1}^{n} \max(0,k-\max(0,j-p-1)) =$$

$$= \sum_{j=1}^{p+1} k + \sum_{j=p+2}^{n} \max(0,k-j+p+1)$$

Ker je na prvi množici zaradi (8.13) $k+p \leq n$, dobimo od tod

$$\sum_{i=1}^{n} \max(0,k-u_i) = (p+1)k + \sum_{j=p+2}^{k+p} (k-j+p+1) =$$

$$= (p+1)k + \frac{1}{2}\, k(k-1) \qquad (8.15)$$

Prva vsota v (8.14) pa je enaka:

$$\sum_{i=1}^{k} ((i-1-u_i) - \max(0,k-u_i)) = \sum_{i=1}^{k} \min(i-1-u_i,i-k-1) =$$

$$= \sum_{i=1}^{k} (i-k-1) = -\frac{1}{2} k(k+1)$$

saj je $u_i \leq i-1 \leq k-1$.

Zato velja na prvi množici

$$\max_{1\leq k\leq 2p+1} \sum_{i=1}^{n} e'_{ik} \leq \max_{1\leq k\leq 2p+1} (-\frac{1}{2} k(k+1) + (p+1)k + \frac{1}{2} k(k-1)) =$$

$$= \max_{1\leq k\leq 2p+1} pk = p(2p+1) \qquad (8.16)$$

Na drugi in tretji množici je izraz k-2p-1 pozitiven. Najprej izraz na desni v (8.12) prepišimo v obliko

$$\max_{2p+1<k\leq n} \sum_{i=1}^{n} e'_{ik} \leq \max_{2p+1<k\leq n} \Big(\sum_{i=1}^{k} (\max(0,i-1-\max(u_i,k-2p-1)) -$$

$$- \max(0,k-\max(u_i,k-2p-1))) + \sum_{i=1}^{n} \max(0,k-\max(u_i,k-2p-1)) \Big) \qquad (8.17)$$

V drugi vsoti upoštevajmo (5.17) in preuredimo sumacijo po indeksu $j = s_i$:

$$\sum_{i=1}^{n} \max(0,k-\max(u_i,k-2p-1)) =$$

$$= \sum_{j=1}^{n} \max(0,k-\max(\max(0,j-p-1),k-2p-1)) =$$

$$= \sum_{j=1}^{k-p} (2p+1) + \sum_{j=k-p+1}^{n} \max(0,k-j+p+1) \qquad (8.18)$$

Prvo vsoto v (8.17) pa razdelimo na dva dela. Prvi del je enak

$$\sum_{i=1}^{k-2p-1} (\max(0,\min(i-1-u_i,i-k+2p)) - \max(0,\min(k-u_i,2p+1))) =$$

$$= \sum_{i=1}^{k-2p-1} -(2p+1) = -(2p+1)(k-2p-1) \qquad (8.19)$$

ker je

$$u_i \leq i-1 \leq k-2p-2$$

torej

$$i-1-u_i \geq i-k+2p+1$$

in

$$k-u_i \geqq 2p+2$$

Drugi del prve vsote v (8.17) pa je enak

$$\sum_{i=k-2p}^{k} (\max(0,\min(i-1-u_i, i-k+2p)) - \max(0,\min(k-u_i, 2p+1))) =$$

$$= \sum_{i=k-2p}^{k} (i-k-1) = -(p+1)(2p+1) \qquad (8.20)$$

ker so sedaj zaradi

$$u_i \leqq i-1 \leqq k-1$$

in

$$k - u_i \geqq 1$$

vsi štirje izrazi $i-1-u_i$, $i-k+2p$, $k-u_i$ in $2p+1$ nenegativni in hkrati nastopita prvi in tretji ali drugi in četrti. Velja namreč

$$(i-1-u_i) - (i-k+2p) = k-u_i-2p-1$$

$$(k-u_i) - (2p+1) = k-u_i-2p-1$$

Zato je izraz v zunanjem oklepaju v (8.20) enak

$$(i-1-u_i) - (k-u_i) = (i-k+2p) - (2p+1) = i-k-1$$

Prva vsota v (8.17) je torej na drugi in tretji množici enaka vsoti izrazov (8.19) in (8.20):

$$-(2p+1)(k-2p-1) - (p+1)(2p+1) = -(2p+1)(k-p) \qquad (8.21)$$

Iz (8.18) in (8.21) sledi, da velja na drugi množici

$$\max_{2p+1<k\leqq n-p} \sum_{i=1}^{n} e'_{ik} \leqq \max_{2p+1<k\leqq n-p} \left((2p+1)(k-p) + \sum_{j=k-p+1}^{k+p} \max(0,k-j+p+1) - (2p+1)(k-p) \right) = p(2p+1) \qquad (8.22)$$

Na tretji množici pa imamo

$$\max_{n-p<k\leq n} \sum_{i=1}^{n} e'_{ik} \leq \max_{n-p<k\leq n} \sum_{j=k-p+1}^{n} (k-j+p+1) =$$

$$= \max_{n-p<k\leq n} (p(2p+1) - \tfrac{1}{2}(k-n+p+1)(k-n+p)) = p(2p+1) - 1 \qquad (8.23)$$

kajti izraz v oklepaju zavzame maksimum pri $k = n-p+1$.

Torej sledi iz (8.16), (8.22) in (8.23) pri pogoju (8.13), da velja

$$\|E'\|_1 \leq p(2p+1) \qquad (8.24)$$

Ugotovimo še, da je ta ocena veljavna tudi v primeru

$$2p+1 \leq n < 3p+1 \qquad (8.25)$$

Izkaže se, da velja tedaj stroga neenakost.

Množico $k = 1,\ldots,n$ razdelimo na tri dele takole: $1 \leq k \leq n-p$, $n-p < k \leq 2p+1$, $2p+1 < k \leq n$.

Če upoštevamo (8.16), dobimo za prvo množico, ker je $n-p < 2p+1$, oceno:

$$\max_{1\leq k\leq n-p} \sum_{i=1}^{n} e'_{ik} \leq \max_{1\leq k\leq n-p} pk = p(n-p) =$$

$$= p(2p+1) - p(3p+1 - n) < p(2p+1)$$

Na drugi množici pa upoštevajmo (8.16) s korekturo v formuli (8.15), ker je sedaj $n < k+p$. Tako dobimo

$$\max_{n-p<k\leq 2p+1} \sum_{i=1}^{n} e'_{ik} \leq \max_{n-p<k\leq 2p+1} (pk - \tfrac{1}{2}(k-n+p+1)(k-n+p)) =$$

$$= p(2p+1) - \tfrac{1}{2}(3p+1-n)(3p+2-n) < p(2p+1)$$

saj je maximum dosežen pri $k = 2p+1$.

Na tretji množici pa velja kot v (8.23):

$$\max_{2p+1<k\leq n} \sum_{i=1}^{n} e'_{ik} \leq \max_{2p+1<k\leq n} (p(2p+1) - \tfrac{1}{2}(k-n+p+1) \cdot$$

$$\cdot (k-n+p)) = p(2p+1) - \tfrac{1}{2}(3p+2-n)(3p+3-n) < p(2p+1) \quad -$$

Sedaj je maksimum dosežen pri k = 2p+2.

Torej pri pogoju (8.25) velja

$$\|E'\|_1 < p(2p+1)$$

in zato velja ocena (8.24) splošno za vsak $n \geq 2p+1$.

Iz ocene (8.24) in iz (7.4) potem sledi

$$\|E\|_1 \leq 2{\cdot}01 \text{ ga } p(2p+1)$$

(ii) $$\|E'\|_\infty = \max_i \sum_{k=1}^{n} e'_{ik} =$$

$$= \max_i \left(\sum_{k=1}^{i-1} \max(0,k-w_{ik}) + \sum_{k=i}^{n} \max(0,i-1-w_{ik}) \right)$$

Iz definicije (7.9) in ocene (5.20) sledi, da velja za vsak i ocena

$$w_{ik} = \max(u_i,v_k) \geq v_k \geq \max(0,k-2p-1)$$

Zato imamo oceno

$$\|E'\|_\infty \leq \max_i \left(\sum_{k=1}^{i-1} \max(0,k-\max(0,k-2p-1)) + \right.$$

$$\left. + \sum_{k=i}^{n} \max(0,i-1-\max(0,k-2p-1)) \right) \qquad (8.26)$$

Vzemimo najprej, da velja

$$n \geq 4p + 1 \qquad (8.27)$$

V tem primeru razdelimo množico i = 1,...,n na tri dele takole: $1 \leq i \leq 2p+1$, $2p+1 < i \leq n-2p$, $n-2p < i \leq n$.

Na prvem delu imamo

$$\max_{1\leq i\leq 2p+1} \sum_{k=1}^{n} e'_{ik} \leq \max_{1\leq i\leq 2p+1} \left(\sum_{k=1}^{i-1} k + \sum_{k=i}^{2p+1} (i-1) + \right.$$

$$\left. + \sum_{k=2p+1}^{i+2p-1} (i+2p-k) \right) = \max_{1\leq i\leq 2p+1} (2p+1)(i-1) = 2p(2p+1) \qquad (8.28)$$

Na drugem delu dobimo

$$\max_{2p+1<i\leq n-2p} \sum_{k=1}^{n} e'_{ik} \leq \max_{2p+1<i\leq n-2p} \left(\sum_{k=1}^{2p} k + \right.$$

$$+ \sum_{k=2p+1}^{i-1} (2p+1) + \sum_{k=i}^{i+2p-1} (i+2p-k)) = \max_{2p+1<i\leq n-2p} (2p+1)(i-1) =$$

$$= (2p+1)n - 4p^2 - 4p - 1 \qquad (8.29)$$

Na tretjem delu pa velja

$$\max_{n-2p<i\leq n} \sum_{k=1}^{n} e'_{ik} \leq \max_{n-2p<i\leq n} (\sum_{k=1}^{2p} k + \sum_{k=2p+1}^{i-1} (2p+1) +$$

$$+ \sum_{k=i}^{n} (i+2p-k)) = \max_{n-2p<i\leq n} ((2p+1)(i-1) -$$

$$- \frac{1}{2}(i+2p-n)(i+2p-n-1)) = (2p+1)n - 2p^2 - p - 1 \qquad (8.30)$$

kajti maksimum je dosežen pri i = n.

Ker je izraz (8.30) večji od izrazov (8.29) in (8.28), je zato pri pogoju (8.27)

$$\|E'\|_\infty \leq (2p+1)n - 2p^2 - p - 1 \qquad (8.31)$$

Prepričajmo se še, da je tudi pri pogoju

$$2p+1 \leq n < 4p+1$$

ocena (8.31) veljavna. Sedaj pa razdelimo množico i = 1,...,n na tri dele takole: $1 \leq i \leq n-2p$, $n-2p < i \leq 2p+1$, $2p+1 < i \leq n$. Brez težav dobimo iz (8.26) naslednje ocene.

Na prvem delu je

$$\max_{1\leq i\leq n-2p} \sum_{k=1}^{n} e'_{ik} \leq \max_{1\leq i\leq n-2p} (\sum_{k=1}^{i-1} k + \sum_{k=i}^{2p+1} (i-1) +$$

$$+ \sum_{k=2p+2}^{i+2p-1} (i+2p-k)) = \max_{1\leq i\leq n-2p} (2p+1)(i-1) = (2p+1)(n-2p-1) =$$

$$= (2p+1)n - 4p^2 - 4p - 1 < (2p+1)n - 2p^2 - p - 1$$

Na drugem delu dobimo

$$\max_{n-2p<i\leq 2p+1} \sum_{k=1}^{n} e'_{ik} \leq \max_{n-2p<i\leq 2p+1} (\sum_{k=1}^{i-1} k + \sum_{k=i}^{2p+1} (i-1) +$$

$$+ \sum_{k=2p+2}^{n} (i+2p-k)) = \max_{n-2p<i\leq 2p+1} ((2p+1)(i-1) -$$

$$- \tfrac{1}{2}(i+2p-n)(i+2p-n-1)) = 2p(2p+1) - \tfrac{1}{2}(4p+1-n)(4p-n) \tag{8.32}$$

Maksimum je namreč dosežen pri $i = 2p+1$. Izraz (8.32) ni večji od izraza (8.31), enak mu je lahko le pri $n = 2p+1$.

Na tretjem delu pa velja

$$\max_{2p+1<i\leq n} \sum_{k=1}^{n} e'_{ik} \leq \max_{2p+1<i\leq n} \Big(\sum_{k=1}^{2p} k + \sum_{k=2p+1}^{i-1} (2p+1) + \sum_{k=i}^{n} (i+2p-k) \Big) = \max_{2p+1<i\leq n} \big((2p+1)(i-1) - \tfrac{1}{2}(i+2p-n)(i+2p-n-1)\big) =$$

$$= (2p+1)n - 2p^2 - p - 1$$

Tu je spet maksimum dosežen pri $i = n$. Torej velja ocena (8.31) za vsak $n \geq 2p+1$.

Iz (8.31) torej sledi

$$\|E\|_\infty \leq 2{\cdot}01\ ga((2p+1)n - 2p^2 - p - 1)$$

(iii)
$$\|E'\|_E = \sum_{k=1}^{n} \sum_{i=1}^{n} e'^2_{ik} =$$

$$= \sum_{k=1}^{n} \Big(\sum_{i=1}^{k} \max(0, i-1-w_{ik})^2 + \sum_{i=k+1}^{n} \max(0, k-w_{ik})^2 \Big)$$

Tu spet uporabimo oceno (8.11) in dobimo

$$\|E'\|_E^2 \leq \sum_{k=1}^{n} \Big(\sum_{i=1}^{k} \max(0, i-1-\max(u_i, k-2p-1))^2 + \sum_{i=k+1}^{n} \max(0, k-\max(u_i, k-2p-1))^2 \Big) =$$

$$= \sum_{k=1}^{n} \Big(\sum_{i=1}^{k} \big(\max(0, i-1-\max(u_i, k-2p-1))^2 - \max(0, k-\max(u_i, k-2p-1))^2 \big) + \sum_{i=1}^{n} \max(0, k-\max(u_i, k-2p-1))^2 \Big) \tag{8.33}$$

Vsoto na $k = 1$ do $k = n$ razdelimo na tri dele: $1 \leq k \leq 2p+1$, $2p+2 \leq k \leq n-p$, $n-p+1 \leq k \leq n$. Če je $2p+1 \leq n < 3p+1$, potem

ima srednja vsota zgornjo mejo manjšo od spodnje, kar pri seštevanju prav nič ne moti.

Pri prvi vsoti je $k-2p-1 \leq 0$, $\max(u_i, k-2p-1) = u_i$ in pri $i \leq k$ je tudi $u_i \leq i-1 < k$. V vsoti od $i = 1$ do $i = n$ pa uporabimo že večkrat uporabljen prijem: upoštevamo formulo (5.17) in preuredimo sumacijo po indeksu $j = s_i$. Tako dobimo

$$\sum_{k=1}^{2p+1} \sum_{i=1}^{n} e'^2_{ik} \leq \sum_{k=1}^{2p+1} \Big(\sum_{i=1}^{k} ((i-1-u_i)^2 - (k-u_i)^2) + \sum_{j=1}^{n} \max(0, k-\max(0, j-p-1))^2 \Big) =$$

$$= \sum_{k=1}^{2p+1} \Big(\sum_{i=1}^{k} ((i-1)^2 - k^2 + 2(k+1-i)u_i) + \sum_{j=1}^{p+1} k^2 + \sum_{j=p+2}^{k+p} (k-j+p+1)^2 \Big)$$

Tu upoštevajmo še, da je $u_i \leq i-1$. Tako dobimo

$$\sum_{k=1}^{2p+1} \sum_{i=1}^{n} e'^2_{ik} \leq \sum_{k=1}^{2p+1} \Big(\sum_{i=1}^{k} -(i-k-1)^2 + (p+1)k^2 + \frac{1}{6} k(k-1)(2k-1) \Big) = \sum_{k=1}^{2p+1} pk^2 =$$

$$= \frac{1}{3} p(p+1)(2p+1)(4p+3) \qquad (8.3^*)$$

Drugo vsoto od $k=2p+2$ do $k=n-p$ v (8.33) najprej nekoliko predelajmo:

$$\sum_{k=2p+2}^{n-p} \sum_{i=1}^{n} e'^2_{ik} \leq \sum_{k=2p+2}^{n-p} \Big(\sum_{i=1}^{k} c_{ik} + \sum_{j=1}^{n} \max(0, k-\max(\max(0, j-p-1), k-2p-1))^2 \Big) =$$

$$= \sum_{k=2p+2}^{n-p} \Big(\sum_{i=1}^{k-2p-1} c_{ik} + \sum_{i=k-2p}^{k} c_{ik} + \sum_{j=1}^{k-p} (2p+1)^2 + \sum_{j=k-p+1}^{k+p} (k-j+p+1)^2 \Big) \qquad (8.34)$$

Tu pomeni

$$c_{ik} = \max(0, \min(i-1-u_i, i-k+2p))^2 - \max(0, \min(k-u_i, 2p+1))^2$$

Ko je $i \leqq k-2p-1$, je $u_i \leqq i-1 \leqq k-2p-2$ in zato je $k-u_i \geqq 2p+2$ in $i-k+2p \leqq -1$. Torej je

$$\sum_{i=1}^{k-2p-1} c_{ik} = \sum_{i=1}^{k-2p-1} -(2p+1)^2 = -(k-2p-1)(2p+1)^2 \qquad (8.35)$$

Ko pa je $k-2p \leqq i \leqq k$, je $i-1-u_i \geqq 0$, $i-k+2p \geqq 0$, $k-u_i \geqq k-i+1 \geqq 1$ in ker je

$$(i-1-u_i) - (i-k+2p) = (k-u_i) - (2p+1)$$

je zato bodisi

$$c_{ik} = (i-1-u_i)^2 - (k-u_i)^2 = (i-1)^2 - k^2 + 2(k-i+1)u_i$$

bodisi

$$c_{ik} = (i-k+2p)^2 - (2p+1)^2 = -(k-i+1)(i-k+4p+1)$$

Ker pa je $u_i \leqq i-1$, je

$$(i-1)^2 - k^2 + 2(k-i+1)u_i \leqq -(k-i+1)^2$$

Obenem pa velja tudi

$$-(k-i+1)(i-k+4p+1) \leqq -(k-i+1)^2$$

saj je

$$-(k-i+1)^2 + (k-i+1)(i-k+4p+1) = 2(k-i+1)(i-k+2p) \geqq 0$$

Zato moremo oceniti drugo vsoto takole:

$$\sum_{i=k-2p}^{k} c_{ik} \leqq \sum_{i=k-2p}^{k} -(k-i+1)^2 = -\tfrac{1}{3}(p+1)(2p+1)(4p+3) \qquad (8.36)$$

Iz (8.34), (8.35) in (8.36) dobimo torej

$$\sum_{k=2p+2}^{n-p} \sum_{i=1}^{n} e'^2_{ik} \leqq \sum_{k=2p+2}^{n-p} \left(-(k-2p-1)(2p+1)^2 - \right.$$
$$\left. - \tfrac{1}{3}(p+1)(2p+1)(4p+3) + (k-p)(2p+1)^2 + \tfrac{1}{3}p(2p+1)(4p+1)\right) =$$
$$= \sum_{k=2p+2}^{n-p} p(2p+1)^2 = p(2p+1)^2(n-3p-1) \qquad (8.37)$$

Ocenimo še tretji del vsote v (8.33) in sicer vsoto od $k = n-p+1$ do $k = n$. Ta del obdelamo podobno kot prejšnji

del, le da sedaj upoštevamo, da je k+p > n in zato moramo v vsoti (8.34) zgornjo mejo k+p nadomestiti z n. Upoštevajoč prejšnji rezultat tako dobimo:

$$\sum_{k=n-p+1}^{n} \sum_{i=1}^{n} e'^{2}_{ik} \leq \sum_{k=n-p+1}^{n} (p(2p+1)^2 - \sum_{j=n+1}^{k+p} (k-j+p+1)^2) =$$

$$= p^2(2p+1)^2 - \frac{1}{12} p(p+1)^2(p+2) =$$

$$= \frac{p}{12}(47p^3 + 44p^2 + 7p - 2) \qquad (8.38)$$

Če seštejemo ocene (8.3*), (8.37) in (8.38), dobimo

$$\|E'\|_E^2 \leq p(2p+1)^2 n - \frac{p}{12}(65p^3 + 76p^2 + 25p + 2) \qquad (8.39)$$

Torej velja

$$\|E\|_E \leq 2{\cdot}01\ ga(p(2p+1)^2 n - \frac{p}{12}(65p^3 + 76p^2 + 25p + 2))^{1/2}$$

9. Končne ocene

Zdaj lahko izpeljemo končne ocene za norme perturbacijske matrike δA.

Iz zveze (7.1) in ocen za norme nastopajočih matrik iz prejšnjega razdelka dobimo

$$\|\delta A\|_1 \leq \|E\|_1 + \|\delta L\|_1 \|U\|_1 + \|L\|_1 \|\delta U\|_1 + \|\delta L\|_1 \|\delta U\|_1$$

in

$$\frac{1}{ga}\|\delta A\|_1 \leq 2{\cdot}01\ p(2p+1) + 1{\cdot}12(2p+1)(pn-p^2+p-1) +$$
$$+ 1{\cdot}12(p+1)^2(2p+1) + 1{\cdot}12^2(p+1)(2p+1)(pn-p^2+p-1)a$$

V zadnjem členu upoštevajmo pogoj (1.4):

$$(pn-p^2+p-1)a < pna \leq o{\cdot}1\ p$$

Tako dobimo

$$\frac{1}{ga}\|\delta A\|_1 \leq 1{\cdot}12\ p(2p+1)(n + 0{\cdot}12p + 4{\cdot}92)$$

Če to oceno še malo polepšamo, dobimo

$$\|\delta A\|_1 \leq 1{\cdot}12\ p(2p+1)(n+p+5)ga \qquad (9.1)$$

Za $\|\delta A\|_\infty$ dobimo po podobni poti naslednje ocene:

$$\|\delta A\|_\infty \leq \|E\|_\infty + \|\delta L\|_\infty \|U\|_\infty + \|L\|_\infty \|\delta U\|_\infty + \|\delta L\|_\infty \|\delta U\|_\infty$$

$$\frac{1}{ga}\|\delta A\|_\infty \leq 2{\cdot}01((2p+1)n-2p^2-p-1) + 0{\cdot}56(2p+1)(n^2+n-2) +$$

$$+ 1{\cdot}12(2p+1)(p+1)n + \frac{1}{2}1{\cdot}12^2(p+1)(2p+1)(n-1)(n+2)a$$

V zadnjem členu upoštevajmo pogoj (1.4):

$$(n-1)a < na \leq 0{\cdot}1$$

in tako dobimo

$$\frac{1}{ga}\|\delta A\|_\infty \leq (2p+1)(0{\cdot}56n^2 + (1{\cdot}19p+3{\cdot}76)n - (1{\cdot}88p+0{\cdot}99))$$

in od tod po primerni zaokrožitvi navzgor

$$\|\delta A\|_\infty \leq 0{\cdot}56(2p+1)n(n+3p+6)ga \qquad (9.2)$$

Ta ocena je asimptotično za faktor n/2p slabša od ocene za $\|\delta A\|_1$.

Končno ocenimo še $\|\delta A\|_E$. Pri tem moramo zaradi enostavnejšega računa nekaj žrtvovati na strogosti ocene, toda le pri manj pomembnih členih. Znebiti se moramo namreč kompliciranih izrazov pod koreni.

Z elementarnim računom poenostavimo ocene za evklidske norme nastopajočih matrik takole:

$$\|E\|_E \leq 2{\cdot}01\ ga(2p+1)\sqrt{pn} \qquad (9.3)$$

$$\|\delta L\|_E \leq o{\cdot}56\ an\sqrt{2pn}$$

$$\|\delta U\|_E \leq 0{\cdot}65\ ga(2p+1)\sqrt{5pn}$$

$$\|L\|_E \leq \sqrt{2pn}$$

$$\|U\|_E \leq g\sqrt{3pn}$$

Pri nekaterih od teh ocen smo predpostavili, da je $n \geq 4$.

Iz ocene

$$\|\delta A\|_E \leq \|E\|_E + \|\delta L\|_E \|U\|_E + \|L\|_E \|\delta U\|_E + \|\delta L\|_E \|\delta U\|_E$$

sledi

$$\frac{1}{ga}\|\delta A\|_E \leq 2{\cdot}01(2p+1)\sqrt{pn} + 0{\cdot}56\sqrt{6}pn^2 + 0{\cdot}65\sqrt{10}(2p+1)pn +$$

$$+ 0{\cdot}56\cdot 0{\cdot}65\sqrt{10}(2p+1)pn\cdot na$$

Ker je $na \leq 0{\cdot}1$, dobimo po elementarnih poenostavitvah

$$\|\delta A\|_E \leq 1{,}38\ pn(n+5p+3)ga \qquad (9.4)$$

Tudi ta ocena je slabša od ocene za $\|\delta A\|_1$.

Da se ocene ne dajo izboljšati, se moremo prepričati na primeru. Pri tem gre seveda le za preskus strogosti ocen za posamezne norme nastopajočih matrik.

Vzemimo matriko tipa A_3 iz razdelka 5 (5.16). Iz formul (5.15) in definicij matrik E', δL', L', δU' in U' dobimo z direktnim računom, če vzamemo, da je

$$n \geq 4p + 1$$

za vse norme teh matrik enake vrednosti kot so dobljene ocene, razen pri evklidski normi matrike E', kjer dobimo naslednjo vrednost:

$$\|E_3'\|_E^2 = p(2p+1)^2 n - \frac{p}{12}(71p^3 + 96p^2 + 43p + 6)$$

Torej je ocena (8.39) kvečjemu malo pregroba, saj se v glavnem členu ujemata. V poenostavljeni oceni (9.3) se ta razlika nič ne pozna.

Končne ocene (9.1), (9.2) in (9.4) so torej kar se da natančno izračunane.

10. Diagonalno dominantne pasovne matrike

Še ugodnejše ocene za normo perturbacijske matrike dobimo, če je dana pasovna matrika diagonalno dominantna. Tudi take matrike v praksi niso redke.

Vzemimo sedaj, da je dana nesingularna matrika A takšna, da velja poleg

$$a_{ik} = 0 , \quad |i-k| > p$$

še pogoj

$$|a_{kk}| \geq \sum_{i=k-p}^{k-1} |a_{ik}| + \sum_{i=k+1}^{k+p} |a_{ik}| , \quad k=1,\ldots,n$$

Vsak diagonalni element je večji ali kvečjemu enak vsoti absolutnih vrednosti izvendiagonalnih elementov v istem stolpcu.

Izkaže se, da v primeru, ko je matrika diagonalno dominantna, pri delnem pivotiranju ni potrebna nobena zamenjava vrstic. Diagonalna dominantnost se namreč ohranja v toku Gaussove eliminacije (Wilkinson (1961)). Za tako matriko veljajo torej formule (5.14) (primer A_1). Tedaj imamo

$$u_r = \max(0, r-p-1)$$
$$v_r = \max(0, r-p-1) \qquad (10.1)$$
$$z_r = \min(p+1, n-r+1)$$

Pri oceni norm posameznih perturbacijskih matrik moremo torej izkoristiti splošne formule (7.4) - (7.8) in definicije matrik E', $\delta L'$, $\delta U'$, L' in U'. Pojdimo kar v istem vrstnem redu kot v razdelku 8.

a) Matrika L'

Iz (10.1) in definicije (7.16) sledi, da so od nič različni elementi matrike L' podani takole:

$$\ell'_{ik} = 1 \ , \ \max(0, i-p-1) < k \leq i$$

Od tod dobimo

$$\|L'\|_1 = p + 1 \qquad (10.2)$$
$$\|L'\|_\infty = p + 1$$
$$\|L'\|_E^2 = \tfrac{1}{2}(p+1)(2n-p) \qquad (10.3)$$

b) Matrika U'

Iz (10.1) in definicije (7.19) sledi, da so od nič različni elementi matrike U' podani takole:

$$u'_{ik} = 1 \ , \quad i \leq k \leq \min(i+p, n)$$

Tudi tu dobimo brez težav isti rezultat kot prej.

$$\|U'\|_1 = p + 1 \qquad (10.4)$$
$$\|U'\|_\infty = p + 1$$
$$\|U'\|_E^2 = \tfrac{1}{2}(p+1)(2n-p) \qquad (10.5)$$

c) Matrika $\delta U'$

Iz (10.1) in definicije (7.21) dobimo, da so od nič različni elementi matrike $\delta U'$ takile:

$$\delta u'_{ii} = \min(p+1, n-i+1)$$

$$\delta u'_{ik} = k-i \ , \quad i < k \leq \min(i+p, n)$$

Od tod izračunamo

$$\|\delta U'\|_1 = \frac{1}{2}(p+1)(p+2) \tag{10.6}$$

$$\|\delta U'\|_\infty = \frac{1}{2}(p+1)(p+2)$$

$$\|\delta U'\|_E^2 = \frac{1}{12}(p+1)(p+2)(2(2p+3)n - p(3p+5)) \tag{10.7}$$

d) Matrika $\delta L'$

Iz definicije (7.18) in iz (10.1) sledi, da so od nič različni elementi matrike $\delta L'$ podani takole:

$$\delta \ell'_{ii} = \min(i-1, p)$$

$$\delta \ell'_{ik} = \max(0, k-\max(0, i-p-1)), \quad k < i$$

Od tod izračunamo

$$\|\delta L'\|_1 = \frac{1}{2}p(p+3) \tag{10.8}$$

$$\|\delta L'\|_\infty = \frac{1}{2}p(p+3)$$

$$\|\delta L'\|_E^2 = \frac{p}{12}(2(2p^2+9p+1)n - (p+1)(3p^2+11p-2)) \tag{10.9}$$

e) Matrika E'

Iz (10.1) in definicije (7.14) sledi, da je matrika E' podana takole:

$$e'_{ik} = \max(0, \min(i-1, i+p-k)) \ , \quad k \geq i$$

$$e'_{ik} = \max(0, \min(k, k+p+1-i)) \ , \quad k < i$$

saj je pri $k \geq i$, $w_{ik} = v_k$ in pri $k < i$, $w_{ik} = u_i$.

Od tod izračunamo

$$\|E'\|_1 = p(p+1) \tag{10.10}$$

$$\|E'\|_\infty = p(p+1)$$

$$\|E'\|_E^2 = \frac{p}{6}(p+1)(2(2p+1)n - 3p(p+1)) \qquad (10.11)$$

Zdaj pa izpeljimo še končne ocene za perturbacijsko matriko δA. Pri tem bomo uporabili poleg zveze (7.1) še ocene (7.4) - (7.8).

Iz (10.2), (10.4), (10.6), (10.8) in (10.10) sledi:

$$\frac{1}{ga}\|\delta A\|_1 \leqq 2{\cdot}01\, p(p+1) + 1{\cdot}12\, \frac{1}{2}\, p(p+3)(p+1) +$$

$$+ 1{\cdot}12\, \frac{1}{2}(p+1)^2(p+2) + 1{\cdot}12^2\, \frac{1}{4}\, p(p+1)(p+2)(p+3)a$$

Tu upoštevajmo, da je

$$2pa < na \leqq 0{\cdot}1$$

in po elementarnih poenostavitvah dobimo oceno

$$\|\delta A\|_1 \leqq 1{\cdot}14(p+1)(p^2+5p+1)ga$$

Prav tako velja

$$\|\delta A\|_\infty \leqq 1{\cdot}14(p+1)(p^2+5p+1)ga$$

saj so ocene za posamezne matrike iste.

Ocenimo še evklidsko normo. Najprej posamezne norme (10.3), (10.5), (10.7), (10.9) in (10.11) ocenimo tako, da si poenostavimo nadaljnji račun:

$$\|L'\|_E < \sqrt{(p+1)n}$$

$$\|U'\|_E < \sqrt{(p+1)n}$$

$$\|\delta U'\|_E < (p+2)\sqrt{(p+1)n/3}$$

$$\|\delta L'\|_E < p\sqrt{(p+1)n}$$

$$\|E'\|_E < (p+1)\sqrt{2pn/3}$$

Od tod dobimo

$$\frac{1}{ga}\|\delta A\|_E \leqq 2{\cdot}01(p+1)\sqrt{2pn/3} + 1{\cdot}12\, p(p+1)n +$$

$$+ 1{\cdot}12(p+2)\sqrt{1/3}(p+1)n + 1{\cdot}12^2 p(p+2)\sqrt{1/3}(p+1)na$$

Tu upoštevajmo, da je $na \leqq 0{\cdot}1$. Tako dobimo po poenostavitvi

$$\|\delta A\|_E \leqq 1{\cdot}77(p+1)^2(n + \sqrt{n} + p)ga$$

11. Povzetek rezultatov

Pri ocenjevanju norm perturbacijske matrike δA smo prišli do naslednjih zaključkov (pri fiksiranih pogojih glede aritmetike):

Pri splošni matriki so vse tri norme ocenjene s (3.29), kjer je $f_p(n) = O(n^3)$, pri splošni pasovni $(2p+1)$-diagonalni matriki je $f_1(n) = O(p^2 n)$, druga dva faktorja $f_\infty(n)$ in $f_E(n)$ pa sta reda $O(pn^2)$. Pri diagonalno dominantni pasovni matriki pa velja $f_1(n) = f_\infty(n) = O(p^3)$ in $f_E(n) = O(p^2 n)$.

Iz tega moremo sklepati, da je Gaussova metoda z delnim pivotiranjem za pasovne matrike tudi s stališča analize napak zelo uspešna in posebno priporočljiva, če so te matrike diagonalno dominantne.

Za še boljšo potrditev tega sklepa pa si moramo ogledati podrobno še ocene za pivotno rast, to je ocene za število g.

III. P I V O T N A R A S T

12. Znane ocene

V oceni (3.29) za normo perturbacijske matrike nastopa tudi faktor g, ki je bil definiran kot

$$g = \max_{i,k,r} | a_{ik}^{(r)} |$$

Število g nam pove, kako veliki morejo postati elementi matrik $A^{(r)}$ v toku Gaussove eliminacije.

Vpeljimo ustrezno relativno število:

$$R = g/\max_{i,k} | a_{ik} |$$

To število nam pove, za kakšen faktor morejo kvečjemu narasti elementi posameznih matrik $A^{(r)}$ glede na absolutno največji element prvotne matrike.

Če R poznamo, je tedaj

$$g = R \max_{i,k} | a_{ik} |$$

Pri ocenjevanju faktorja R bomo suponirali, da je matrika A nesingularna in tako normirana, da je

$$\max_{i,k} | a_{ik} | = 1 \qquad (12.1)$$

Tedaj je očitno g = R. Študirali bomo torej, kako naraščajo elementi v toku eliminacij pri pogoju (12.1).

Če izvajamo eliminacije brez pivotiranja, se v splošnem lahko zgodi, da je R poljubno veliko število. Torej brez dodatnih predpostavk ni mogoče dobiti nobene ocene za R. Obenem se spomnimo tudi, da je bila ocena (3.29) izpeljana pri pogoju, da so elementi spodnje trikotne matrike L absolutno omejeni z 1, torej pri pogoju, da izvajamo eliminacije npr. z delnim pivotiranjem.

Oglejmo si nekaj ocen za število R, ki jih je izpeljal Wilkinson (1961).

a) Pri delnem pivotiranju velja ocena

$$R \leq 2^{n-1} \tag{12.2}$$

Iz formul (2.4)

$$a_{ik}^{(r+1)} = a_{ik}^{(r)} - m_{ir} a_{rk}^{(r)}$$

in pogoja

$$|m_{ir}| \leq 1$$

sledi

$$|a_{ik}^{(r+1)}| \leq 2 \max_{i,k} |a_{ik}^{(r)}|$$

Torej velja tudi

$$\max_{i,k} |a_{ik}^{(r+1)}| \leq 2 \max_{i,k} |a_{ik}^{(r)}| \tag{12.3}$$

Na vsakem koraku eliminacije je kvečjemu možen porast elementov za faktor 2, torej velja

$$\max_{i,k} |a_{ik}^{(n)}| \leq 2^{n-1} \max_{i,k} |a_{ik}| = 2^{n-1} \tag{12.4}$$

Iz (12.3) in (12.4) takoj sledi veljavnost ocene (12.2).

Enačaj je v tej oceni možen. Vzemimo primer take matrike pri n = 6.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix} \qquad A^{(6)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 0 & 4 \\ 0 & 0 & 0 & 1 & 0 & 8 \\ 0 & 0 & 0 & 0 & 1 & 16 \\ 0 & 0 & 0 & 0 & 0 & 32 \end{bmatrix}$$

Toda to je zelo specialen primer. V praksi je znaten porast elementov zelo redek.

b) Če je A diagonalno dominantna matrika, je

$$R \leq 2$$

brez pivotiranja.

c) Če je A tridiagonalna matrika, velja ocena

$$R \leq 2 \tag{12.5}$$

pri delnem pivotiranju.

Oglejmo si dokaz za oceno (12.5), ker bomo v zadnjem razdelku ta rezultat posplošili na poljubne pasovne matrike.

Naj bo torej A tridiagonalna matrika:

$$a_{ik} = 0 \ , \quad |i-k| > 1 \ , \quad |a_{ik}| \leq 1$$

Pri Gaussovi eliminaciji z delnim pivotiranjem prideta v poštev pri izbiri pivota vedno samo dva elementa. Oglejmo si r-ti korak eliminacije in si ponazorimo samo tiste elemente matrike $A^{(r)}$, ki so na tem koraku zanimivi:

$$\begin{matrix} a_{rr}^{(r)} & a_{r,r+1}^{(r)} & 0 \\ a_{r+1,r} & a_{r+1,r+1} & a_{r+1,r+2} \\ 0 & a_{r+2,r+1} & a_{r+2,r+2} \end{matrix}$$

Elementi brez zgornjega indeksa so elementi prvotne matrike.

Vzemimo nadalje, da veljajo ocene:

$$\begin{aligned} &|a_{rr}^{(r)}| \leq 2 \ , \quad |a_{r,r+1}^{(r)}| \leq 1 \\ &|a_{ik}| \leq 1 \ , \quad i \geq r+1, \ k \geq r \end{aligned} \tag{12.6}$$

kar je pri r = 1 prav gotovo izpolnjeno.

Oglejmo si sedaj en korak eliminacije. Imamo dve možnosti:

(i) $$|a_{rr}^{(r)}| \geq |a_{r+1,r}| \tag{12.7}$$

V tem primeru vrstic r in r+1 ne zamenjamo in r-ta vrstica zgornje trikotne matrike U postane

$$0, \ \ldots \ ,0, \ a_{rr}^{(r)}, \ a_{r,r+1}^{(r)}, \ 0, \ \ldots \ , \ 0$$

Nova elementa (r+1)-te vrstice izračunamo po formulah:

$$a_{r+1,r+1}^{(r+1)} = a_{r+1,r+1} - \frac{a_{r+1,r}}{a_{rr}^{(r)}} \ a_{r,r+1}^{(r)}$$

$$a^{(r+1)}_{r+1,r+2} = a_{r+1,r+2}$$

Vsi drugi elementi ostanejo nespremenjeni. Če upoštevamo (12.6) in (12.7), dobimo od tod naslednje ocene:

$$|a^{(r+1)}_{r+1,r+1}| \leq |a_{r+1,r+1}| + |a^{(r)}_{r,r+1}| \leq 2$$

$$|a^{(r+1)}_{r+1,r+2}| = |a_{r+1,r+2}| \leq 1$$

$$|a_{ik}| \leq 1 \ , \quad i \geq r+2, \quad k \geq r+1$$

$$\text{(ii)} \qquad |a^{(r)}_{rr}| < |a_{r+1,r}| \tag{12.8}$$

V tem primeru najprej zamenjamo r-to in (r+1)-to vrstico tako, da dobimo na ustreznih mestih situacijo:

$$\begin{matrix} a_{r+1,r} & a_{r+1,r+1} & a_{r+1,r+2} \\ a^{(r)}_{rr} & a^{(r)}_{r,r+1} & 0 \\ 0 & a_{r+2,r+1} & a_{r+2,r+2} \end{matrix}$$

Tako postane r-ta vrstica matrike U

$$0, \ \ldots \ , 0, a_{r+1,r}, a_{r+1,r+1}, a_{r+1,r+2}, 0, \ \ldots \ , 0$$

Nato izvedemo eliminacijo po formulah

$$a^{(r+1)}_{r+1,r+1} = a^{(r)}_{r,r+1} - \frac{a^{(r)}_{rr}}{a_{r+1,r}} a_{r+1,r+1}$$

$$a^{(r+1)}_{r+1,r+2} = - \frac{a^{(r)}_{rr}}{a_{r+1,r}} a_{r+1,r+2}$$

Vsi drugi elementi ostanejo nespremenjeni. Iz (12.6) in (12.8) potem sledi:

$$|a^{(r+1)}_{r+1,r+1}| \leq |a^{(r)}_{r,r+1}| + |a_{r+1,r+1}| \leq 2$$

$$|a^{(r+1)}_{r+1,r+2}| \leq |a_{r+1,r+2}| \leq 1$$

$$|a_{ik}| \leq 1 \ , \quad i \geq r+2 \ , \quad k \geq r+1$$

V obeh primerih imamo torej spet situacijo kot pred tem korakom z ocenami (12.6) pri r, ki je za 1 večji. Na osnovi indukcije torej sklepamo, da velja ocena (12.5).

Preden preidemo na izpeljavo ocene za R pri poljubni (2p+1)-diagonalni matriki, si pripravimo potrebne pripomočke.

13. Pomožni izrek

Pri danem naravnem številu p konstruirajmo pravokotno matriko B, ki ima p+1 vrstic in 2p stolpcev po pravilu:

$$b_{i,2p} = 1 \ , \quad i=1,2,\ldots,p+1 \tag{13.1}$$

$$b_{p+1,k} = 1 \ , \quad k=1,2,\ldots,2p \tag{13.2}$$

$$b_{ik} = b_{i+1,k+1} + b_{1,k+1} \tag{13.3}$$

za $k=2p-1,\ldots,1$ in $i=1,2,\ldots,p$.

S temi predpisi je matrika B natanko določena. Ima naslednje lastnosti:

(i) $$b_{ik} > 1 \ , \ i=1,\ldots,p \ , \ k=1,\ldots,2p-1 \tag{13.4}$$

(ii) $$b_{1k} \geq b_{ik} \ , \ k=1,\ldots,2p \ , \ i=2,\ldots,p+1 \tag{13.5}$$

(iii) $$b_{1k} > b_{1,k+1} \ , \ k=1,\ldots,2p-1 \tag{13.6}$$

(iv) $$\max_{i,k} b_{ik} = b_{11} \tag{13.7}$$

(v) $$b_{ik} = 2^{2p-k} - 2^{p-k+i-1} + 1 - (p-k-2)2^{p-k-1} + (i-k-3)2^{i-k-2} \tag{13.8}$$

za $i=2,3,\ldots,p$ in $k=1,2,\ldots,i-1$

$$b_{ik} = 2^{2p-k} - 2^{p-k+i-1} - (p-k-2)2^{p-k-1} \tag{13.9}$$

za $i=1,2,\ldots,p$ in $k=i,i+1,\ldots,p$

$$b_{ik} = 2^{2p-k} - 2^{p-k+i-1} + 1 \tag{13.10}$$

za $i=2,3,\ldots,p$ in $k=p+1,\ldots,p+i-1$

$$b_{ik} = 2^{2p-k} \qquad (13.11)$$

za $i=1,2,\ldots,p$ in $k=p+i,\ldots,2p$

Dokaz.

(i) Ocena (13.4) sledi neposredno iz definicije elementov matrike B (13.1) - (13.3), saj je zaradi (13.3) b_{ik} vedno vsota dveh pozitivnih celih števil.

(ii) Pri dokazovanju neenačbe (13.5) ločimo dva primera:

a) $k \geq p + i - 1$ (13.12)

Tedaj sledi iz (13.3), da je

$$b_{1k} - b_{ik} = b_{2,k+1} - b_{1,k+1} - b_{i+1,k+1} + b_{1,k+1} =$$
$$= b_{2,k+1} - b_{i+1,k+1} =$$
$$= - - - - - - - - - =$$
$$= b_{2p-k+1,2p} - b_{2p-k+i,2p} = 1 - 1 = 0$$

Pri pogoju (13.12) je namreč

$$2p - k + i \leq p + 1$$

in zato smemo uporabiti definicijo (13.1). Torej pri pogoju (13.12) je

$$b_{1k} = b_{ik}$$

b) $k < p + i - 1$ (13.12)

V tem primeru pa na osnovi (13.3) velja:

$$b_{1k} - b_{ik} = b_{2,k+1} - b_{i+1,k+1} =$$
$$= - - - - - - - - - =$$
$$= b_{p-i+2,p+k-i+1} - b_{p+1,p+k-i+1} =$$
$$= b_{p-i+2,p+k-i+1} - 1 > 0$$

saj je sedaj

$$p+k-i+1 < 2p$$

in moremo uporabiti oceno (13.4).

Pri pogoju (13.12) velja torej

$$b_{1k} > b_{ik}$$

Neenačba (13.5) je torej dokazana.

(iii) Veljavnost neenačbe (13.6) sledi neposredno iz (13.3) in (13.4) ali (13.1):

$$b_{1k} = b_{2,k+1} + b_{1,k+1} > b_{1,k+1}$$

(iv) Enostavno se je prepričati, da je maksimalni element matrike B ravno b_{11}. Iz (13.5) in (13.6) sledi

$$\max_{i,k} b_{ik} = \max_{k} b_{1k} = b_{11}$$

(v) Nekoliko več računanja je pri izpeljavi eksplicitnih formul za elemente matrike B. Ločiti je treba štiri trikotne dele matrike B.

V vseh štirih primerih bomo uporabljali formulo, ki sledi iz aditivne lastnosti (13.3):

$$\begin{aligned} b_{ik} &= b_{i+1,k+1} + b_{1,k+1} = \\ &= b_{i+2,k+2} + b_{2,k+2} + 2b_{1,k+1} = \\ &= b_{i+3,k+3} + b_{3,k+3} + 2b_{2,k+3} + 4b_{1,k+3} = \\ &= - - - - - - - - - = \\ &= b_{i+r,k+r} + \sum_{s=0}^{r-1} 2^s b_{r-s,k+r} \end{aligned} \qquad (13.13)$$

V tej formuli bomo vzeli čimvečji r, to se pravi, da bo bodisi $i+r = p+1$ in $k+r \leq 2p$, bodisi $k+r = 2p$ in $i+r \leq p+1$.

Oglejmo si sedaj elemente matrike B v štirih trikotnikih.

a) $\quad i=1,2,\ldots,p \ , \quad k=p+i,\ldots,2p$

V tem primeru je

$$k \geq p + i$$

V enačbi (13.13) smemo torej vzeti $r = 2p-k$, saj je $i+r = 2p-k+i \leq p$. Pri tem upoštevajmo definicijo (13.1). Tako dobimo

$$b_{ik} = b_{2p-k+i,2p} + \sum_{s=0}^{2p-k-1} 2^s b_{2p-k-s,2p} =$$

$$= 1 + \sum_{s=0}^{2p-k-1} 2^s = 2^{2p-k} \qquad (13.14)$$

S tem je dokazana formula (13.11).

b) $i=2,3,\ldots,p$, $k=p+1,\ldots,p+i-1$

V tem primeru je

$$p < k < p + i \qquad (13.15)$$

V formuli (13.13) moremo sedaj vzeti $r = p-i+1$, saj je $k+r = p+k-i+1 < 2p+1$. Pri tem upoštevajmo definicijo (13.2). Tako dobimo

$$b_{ik} = b_{p+1,p+k-i+1} + \sum_{s=0}^{p-i} 2^s b_{p-i-s+1,p+k-i+1} =$$

$$= 1 + \sum_{s=0}^{p-i} 2^s b_{p-i-s+1,p+k-i+1} \qquad (13.16)$$

Ker je sedaj na osnovi neenačbe (13.15)

$$(p+k-i+1) - (p-i-s+1) = k+s \geq k > p$$

moremo v (13.16) uporabiti formulo (13.14) in zato je

$$b_{p-i-s+1,p+k-i+1} = 2^{p-k+i-1}$$

Tako dobimo iz (13.16)

$$b_{ik} = 1 + \sum_{s=0}^{p-i} 2^s \cdot 2^{p-k+i-1} = 2^{2p-k} - 2^{p-k+i-1} + 1 \qquad (13.17)$$

S tem smo dokazali formulo (13.10).

c) $i=1,2,\ldots,p$, $k=i,i+1,\ldots,p$

Sedaj je

$$i \leq k \leq p \qquad (13.18)$$

Spet uporabimo formulo (13.13) z $r = p-i+1$, pri tem pa razdelimo vsoto na dva dela:

$$b_{ik} = 1 + \sum_{s=0}^{p-k-1} 2^s b_{p-i-s+1,p+k-i+1} +$$

$$+ \sum_{s=p-k}^{p-i} 2^s b_{p-i-s+1,p+k-i+1}$$

Prva vsota je pri $k = p$ enaka nič, kot je navadno pri tem zapisu. V prvi vsoti velja zaradi (13.18)

$$p+k-i+1 \geq p+1 > p$$

in

$$(p+k-i+1) - (p-i-s+1) = k+s \leq p-1 < p$$

Zato moremo v prvi vsoti uporabiti formulo (13.17).

V drugi vsoti pa velja

$$(p+k-i+1) - (p-i-s+1) = k+s \geq p$$

Zato moremo uporabiti formulo (13.14).

Tako dobimo

$$b_{ik} = 1 + \sum_{s=0}^{p-k-1} 2^s (2^{p-k+i-1} - 2^{p-k-s+1} + 1) +$$

$$+ \sum_{s=p-k}^{p-i} 2^s \cdot 2^{p-k+i-1} =$$

$$= 2^{2p-k} - 2^{p-k+i-1} - (p-k-2)2^{p-k-1} \qquad (13.19)$$

S tem je dokazana formula (13.9).

Končno si oglejmo še zadnji trikotnik.

d) $\quad i=2,3,\ldots,p \quad , \quad k=1,2,\ldots,i-1$

Sedaj pa je

$$k < i$$

in spet uporabimo formulo (13.13) z $r = p-i+1$:

$$b_{ik} = 1 + \sum_{s=0}^{p-i} 2^s \, b_{p-i-s+1,\,p+k-i+1} \qquad (13.20)$$

Ker sedaj velja

$$(p+k-i+1) - (p-i-s+1) = k+s \geq k > 0$$

in

$$p+k-i+1 < p+1 \leq p$$

moremo v (13.20) uporabiti formulo (13.19). Tako dobimo

$$b_{ik} = 1 + \sum_{s=0}^{p-i} 2^s (2^{p-k+i-1} - 2^{p-k-s-1} - (i-k-3)2^{i-k-2}) =$$

$$= 2^{2p-k} - 2^{p-k+i-1} + 1 - (p-k-2)2^{p-k-1} +$$

$$+ (i-k-3)2^{i-k-2}$$

Tako je dokazana tudi formula (13.8) in s tem ves pomožni izrek.

14. Ocena za R pri pasovni matriki

Izrek. Če je A *nesingularna* (2p+1)*-diagonalna matrika in če izvajamo na njej Gaussovo eliminacijo z delnim pivotiranjem, velja pri eksaktni aritmetiki*

$$R \leq 2^{2p-1} - (p-1)2^{p-2} \qquad (14.1)$$

Ocena je stroga.

Dokaz. Izrek bomo dokazali z indukcijo. Na vsakem koraku pride v poštev za eliminacijo samo p+1 vrstic matrike. Vzemimo situacijo na r-tem koraku. Napišimo samo tiste elemente matrike $A^{(r)}$ $(A^{(1)} = A)$, ki nastopijo pri izračunu matrike $A^{(r+1)}$:

$$\begin{matrix}
a^{(r)}_{rr} & a^{(r)}_{r,r+1} & \cdots & a^{(r)}_{r,r+2p-1} & 0 \\
a^{(r)}_{r+1,r} & a^{(r)}_{r+1,r+1} & \cdots & a^{(r)}_{r+1,r+2p-1} & 0 \\
- & - & - & - & - \\
a^{(r)}_{r+p-1,r} & a^{(r)}_{r+p-1,r+1} & \cdots & a^{(r)}_{r+p-1,r+2p-1} & 0 \\
a^{(r)}_{r+p,r} & a^{(r)}_{r+p,r+1} & \cdots & a^{(r)}_{r+p,r+2p-1} & a^{(r)}_{r+p,r+2p}
\end{matrix}$$

Pri tem so elementi zadnje vrstice še nespremenjeni elementi prvotne matrike.

$$a^{(r)}_{r+p,k} = a_{r+p,k} \quad , \quad k=r,\ldots,r+2p$$

Vzemimo, da veljajo ocene

$$|a^{(r)}_{r+i-1,r+k-1}| \leq b_{ik} \ , \ i=1,\ldots,p+1 \ , \ k=1,\ldots,2p \quad (14.2)$$

$$|a^{(r)}_{r+p,r+2p}| \leq 1$$

kjer so b_{ik} elementi matrike B, ki je bila definirana v prejšnjem razdelku. Pri r = 1 je situacija taka, da tem pogojem prav gotovo ustreza, saj je $b_{ik} \geq 1$.

Ugotovili bomo, da dobimo po enem koraku eliminacije analogno situacijo z enakimi ocenami.

Pri delnem pivotiranju si ogledamo elemente v r-tem stolpcu na glavni diagonali in pod njo ter poiščemo tistega, ki ima največjo absolutno vrednost. Če je takih več, je vseeno, katerega vzamemo.

Naj bo v našem primeru pivot v (r+j)-ti vrstici, kjer je j neko število med 0 in p. Velja naj torej

$$|a^{(r)}_{r+j,r}| \geq |a^{(r)}_{r+i,r}| \quad , \quad i=0,1,\ldots,p \quad (14.3)$$

Tedaj zamenjamo r-to vrstico z (r+j)-to in elementi zgornje trikotne matrike v r-ti vrstici so:

$$u_{r,r+k} = a^{(r)}_{r+j,r+k} \ , \quad k=0,1,\ldots,2p \quad (14.4)$$

Nato izvedemo eliminacijo po formulah

$$a^{(r+1)}_{r+i,r+k} = a^{(r)}_{r+i,r+k} - \frac{a^{(r)}_{r+i,r}}{u_{rr}} u_{r,r+k} \quad (14.5)$$

za $i=1,2,\ldots,p$, $i \neq j$ in $k=1,2,\ldots,2p$

in posebej

$$a^{(r+1)}_{r+j,r+k} = a^{(r)}_{r,r+k} - \frac{a^{(r)}_{rr}}{u_{rr}} u_{r,r+k} \ , \quad k=1,\ldots,2p \quad (14.6)$$

Če je j = 0, ta enačba ni potrebna. Vsi drugi elementi ostanejo nespremenjeni in velja torej tudi

$$a^{(r+1)}_{r+p+1,r+k} = a_{r+p+1,r+k} \ , \quad k=1,\ldots,2p+1$$

Ocenimo sedaj absolutne vrednosti novih elementov.

Najprej dobimo iz (14.4) in (14.2) pri k=1,2,...,2p-1 oceno

$$|u_{r,r+k}| \leq b_{j+1,k+1} \tag{14.7}$$

in iz (14.5), (14.2), (14.3) in (14.7)

$$|a^{(r+1)}_{r+i,r+k}| \leq |a^{(r)}_{r+i,r+k}| + |u_{r,r+k}| \leq$$

$$\leq b_{i+1,k+1} + b_{j+1,k+1}$$

Sedaj pa uporabimo lastnosti (13.5) in (13.3) matrike B in dobimo

$$|a^{(r+1)}_{r+i,r+k}| \leq b_{i+1,k+1} + b_{1,k+1} = b_{ik}$$

Posebej pri k = 2p pa imamo iz (14.5) za i=1,2,...,p-1 , če $i \neq j$ in iz (13.1)

$$|a^{(r+1)}_{r+i,r+2p}| \leq |u_{r,r+2p}| \leq 1 = b_{i,2p}$$

in pri i = p, kajti, če je $j \neq p$, je $u_{r,r+2p} = 0$:

$$|a^{(r+1)}_{r+p,r+2p}| = |a^{(r)}_{r+p,r+2p}| \leq 1 = b_{p,2p}$$

Torej velja za i=1,2,...,p , $i \neq j$ in k=1,2,...,2p ocena

$$|a^{(r+1)}_{r+1+i-1,r+1+k-1}| \leq b_{ik} \tag{14.8}$$

torej analogna ocena kot (14.2) pri r+1.

To oceno moramo potrditi še pri i = j. Tedaj pa uporabimo (14.6) in dobimo za k=1,...,2p-1

$$|a^{(r+1)}_{r+j,r+k}| \leq |a^{(r)}_{r,r+k}| + |u_{r,r+k}| \leq$$

$$\leq b_{1,k+1} + b_{j+1,k+1} = b_{jk}$$

Pri k = 2p pa imamo

$$|a^{(r+1)}_{r+j,r+2p}| = |u_{r,r+2p}| = 0 < 1 = b_{j,2p}$$

če je $j \leq p$. Če pa je j = p+1, je

$$|a^{(r+1)}_{r+p+1,r+2p}| = |u_{r,r+2p}| \leq 1 = b_{p+1,2p}$$

Torej velja ocena (14.8) za vse indekse $i=1,\ldots,p+1$ in $k=1,\ldots,2p$.

Po enem koraku eliminacije se torej situacija popolnoma ponovi.

Zato velja

$$R = \max_{i,k,r} |a_{ik}^{(r)}| \leq \max_{i,k} b_{ik} = b_{11}$$

Ker pa je iz (13.9)

$$b_{11} = 2^{2p-1} - (p-1)2^{p-2}$$

je s tem izrek dokazan.

Pokazati moramo le še, da je ocena stroga.

Konstruirajmo matriko A reda $n = 2p+1$ s predpisom:

$$a_{ii} = 1 , \quad i=1,\ldots,n$$

$$a_{ik} = -1 , \quad i-k=1,\ldots,p$$

$$a_{in} = 1 , \quad i=1,p+2,p+3,\ldots,n$$

Vsi drugi elementi naj bodo enaki nič. Tako matriko lahko dobimo iz (2p+1)-diagonalne matrike, ki zadošča pogojem $|a_{ik}| \leq 1$, če najprej zamenjamo prvo in (p+1)-to vrstico.

Izkaže se, da elementi zadnjega stolpca matrik $A^{(r)}$ nastajajo po istih pravilih kot smo tvorili elemente matrike B. Velja namreč

$$a_{in}^{(r)} = b_{i-r+1,n-r+1} , \; r=2,\ldots,n , \; i=r,\ldots,r+p$$

Torej je v tem primeru

$$a_{nn}^{(n)} = b_{11}$$

Kot zgled vzemimo tako matriko pri p = 5, n = 11:

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

V tem primeru je

$$A^{(11)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ & & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ & & & 1 & 0 & 0 & 0 & 0 & 0 & 4 \\ & & & & 1 & 0 & 0 & 0 & 0 & 8 \\ & & & & & 1 & 0 & 0 & 0 & 16 \\ & & & & & & 1 & 0 & 0 & 32 \\ & & & & & & & 1 & 0 & 63 \\ & & & & & & & & 1 & 124 \\ & & & & & & & & & 244 \\ & & & & & & & & & 480 \end{bmatrix}$$

Ustrezna matrika B pa je pri p = 5 enaka:

$$B = \begin{bmatrix} 480 & 244 & 124 & 63 & 32 & 16 & 8 & 4 & 2 & 1 \\ 464 & 236 & 120 & 61 & 31 & 16 & 8 & 4 & 2 & 1 \\ 432 & 220 & 112 & 57 & 29 & 15 & 8 & 4 & 2 & 1 \\ 369 & 188 & 96 & 49 & 25 & 13 & 7 & 4 & 2 & 1 \\ 245 & 125 & 64 & 33 & 17 & 9 & 5 & 3 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Ugotovili smo torej, da je pri pasovnih matrikah tudi ocena za faktor g znatno boljša od splošne ocene (12.2). Ocena (14.1) je odvisna samo od p, kar je zelo ugodno, saj je pri pasovnih matrikah navadno n znatno večji od p.

Če strnemo vse ugotovitve zadnjih dveh poglavij, moremo trditi, da je Gaussova eliminacijska metoda z delnim pivotiranjem tudi s stališča analize napak zelo priporočljiva za pasovne sisteme linearnih enačb.

L I T E R A T U R A

1. Bohte, Z., 1970, Analiza zaokrožitvenih napak pri numeričnih metodah linearne algebre, Elaborat za SBK.

2. Forsythe, G. and Moler, C.B., 1967, Computer solution of linear algebraic systems. Prentice-Hall, Englewood Cliffs.

3. Isaacson, E. and Keller B.K., 1966, Analysis of numerical methods. John Wiley.

4. Wilkinson, J.H., 1961, Error analysis of direct methods of matrix inversion. J. Ass. Comp. Mach. 8, 281 - 330.

5. Wilkinson, J.H., 1963, Rounding errors in algebraic processes. Her Majesty's Stationery Office, London.

6. Wilkinson, J.H., 1965, The algebraic eigenvalue problem. Clarendon Press, Oxford.