Utjecaj pogreške zaokruživanja na točnosti proračuna konstrukcije

Jaguljnjak-Lazarević, Antonia; Dvornik, Josip; Frgić, Lidija

UDK 624.046.001.8:519.28 *Primljeno 23. 7. 2011*

Utjecaj pogreške zaokruživanja na točnosti proračuna konstrukcije

Antonia Jaguljnjak-Lazarević, Josip Dvornik, Lidija Frgić

Ključne riječi

konstrukcija, točnost proračuna, pogreška zaokruživanja, Lagrangeov trikubični element, kvadar, ploča, štap

A. Jaguljnjak-Lazarević, J. Dvornik, L. Frgić *Izvorni znanstveni rad*

Utjecaj pogreške zaokruživanja na točnosti proračuna konstrukcije

U članku je, radi ocjene točnosti proračuna konstrukcije, prikazan sustavni niz primjera s analitičkim rješenjima koja se metodom konačnih elemenata mogu prikazati u obliku cijelih brojeva ili razlomaka. Kao primjer odabran je Lagrangeov trikubični element u obliku kvadra. Primjeri su riješeni numerički uporabom standardnog broja značajnih znamenaka. Na kraju je iz razlike točnog i numeričkog rješenja određen točan iznos pogreške. Cijeli je postupak proveden programom Mathematica.

Key words

structure, accuracy of analysis, rounding error, three-cube Lagrange element, cuboid, plate, member

A. Jaguljnjak-Lazarević, J. Dvornik, L. Frgić *Original scientific paper*

Effect of rounding error on the accuracy of structural analysis

In order to estimate accuracy of structural analyses, the authors introduce a systemic set of examples with analytic solutions that can be presented, through the finite element method, in form of whole numbers and fractions. The three-cube Lagrange element in form of cuboid is selected as an example. Examples are solved numerically using a standard number of significant digits. In the end, an accurate calculation of error is derived from the difference between the exact and numerical solutions. The entire procedure is realized using the Mathematica software.

Mots clés

structure, précision d'analyse, erreur d'arrondi, élément Lagreange à trois cubes, pavé droit, plaque, membre

A. Jaguljnjak-Lazarević, J. Dvornik, L. Frgić *Ouvrage scientifique original*

L'effet d'erreur d'arrondi sur la précision de l'analyse structurelle

Afin d'estimer la précision des analyses structurelles, les auteurs introduisent une série systémique des exemples avec les solutions analytiques qui peuvent être présentées, en utilisant la méthode des éléments finis, en forme des nombres entiers et des fractions. L'élément Lagrange à trois cubes et en forme de pavé droit est utilisé comme exemple. Les exemples sont résolus de manière numérique en utilisant un nombre standard des digits significatifs. A la fin, un calcul précis d'erreur est dérivé de la différence entre solutions exactes et numériques. La procédure entière est réalisée en utilisant le logiciel Mathematica.

Ключевые слова

конструкция, точность расчета, ошибка округления, трикубический элемент Лагранжа, квадр, плита, стержень

А.Ягульняк-Лазаревич, Й. Дворник, Л. Фргич *Оригинальная научная работа*

Влияние ошибки округления на точность расчета конструкции

В статье в целях оценки точности расчета конструкции приведен систематический ряд примеров с аналитическими решениями, которые с помощью метода конечных элементов могут быть представлены в виде целых чисел или дробей. В качестве примера выбран трикубический элемент Лагранжа в форме квадра. Примеры решены числовым путем с использованием стандартного числа значащих цифр. В конце по разнице между точным и числовым решениями определен точный размер ошибки. Вся процедура проведена с использованием программы Mathematica.

Schlüsselworte

Konstruktion, Berechnungsgenauigkeit, Abrundungsfehler, Lagrange's dreikubisches Element, Quader, Platte, Stab

A. Jaguljnjak-Lazarević, J. Dvornik, L. Frgić *Wissenschaftlicher Originalbeitrag*

Einfluss des Fehlers der Abrundung auf die Genauigkeit der Konstruktionsberechnung

Im Artikel ist, im Ziel der Bewertung der Genauigkeit der Konstruktionsberechnung eine systematische Reihe von Beispielen dargestellt, mit analytischen Lösungen die man mit der Methode der endlichen Elemente als ganze Zahl oder Bruch darstellen kann. Als Beispiel wählte man Lagrange's dreikubisches Element in Form eines Quaders. Die Bespiele löste man numerisch mit Hilfe der standarden Anzahl kennzeichnender Ziffern. Zum Ende ist aus der Differenz der genauen und numerischen Lösung die genaue Grösse des Fehlers errechnet. Das Verfahren wurde mit dem Programm Mathematica durchgeführt.

Autori: Doc. dr. sc. **Antonia Jaguljnjak-Lazarević**, dipl. ing. građ., Rudarsko-geološko-naftni fakultet, Zagreb; prof. emer. dr. sc. **Josip Dvornik**, dipl. ing. građ., Građevinski fakultet; Zagreb; prof. dr. sc. **Lidija Frgić**, dipl. ing. građ., Rudarsko-geološko-naftni fakultet, Zagreb

1 Uvod

Suvremeni proračuni inženjerskih konstrukcija gotovo su nezamislivi bez uporabe računala, a sustavi jednadžbi koji proizlaze iz numeričkih modela često sadrže više od 10^5 nepoznanica. Takvi se sustavi učinkovito rješavaju različitim metodama (izravnim, iteracijskim, gradijentnim ili njihovom kombinacijom), koje sa sobom nose *neizbježnu* pogrešku zaokruživanja. Pogreška ovoga tipa posljedica je konačnog broja znamenaka kojima se pri zapisu nekog, ponajprije realnog broja služi računalo. Primijetimo: već sami strojni zapis ulaznih podataka tvori (početnu) pogrešku. Daljnjim, brojnim zaokruživanjima tijekom realizacije numeričkih operacija nekog algoritma pogreške se dijelom pribrajaju, a dijelom poništavaju. Ipak, ukupni je porast početne pogreške neizbježan i (ovisno o svojstvima sustava) često vrlo brz.

Primjerice, sustav jednadžbi temeljen na metodi konačnih elemenata, kojemu pripada model s izrazitim razlikama u krutostima, brže akumulira pogreške od sustava iste veličine s izjednačenim krutostima.

Iako uobičajeni računalni zapis broja na približno 15 dekadskih znamenaka izgleda vrlo pouzdano, gubitak točnih znamenaka[1], čak i kod dobrih modela, s prosječnim brojem nepoznanica, riješenih stabilnom metodom za proračun sustava, može biti zapanjujuće velik.

Pod tvorbom dobrog modela smatramo izbor prikladnih i pravilnih konačnih elemenata, ispravno postavljene rubne uvjete, dobro aproksimirana opterećenja, izbor odgovarajućeg postupka proračuna (linearnog ili nelinearnog) i općenito uklanjanje svih pogrešaka koje znanjem i iskustvom možemo izbjeći.

Prema [1] pri proračunu dobrog modela od samo desetak tisuća nepoznanica, procijenjeno je da šest do sedam znamenaka možemo smatrati netočnim. Za veće, nelinearne modele, kod kojih broj operacija pri proračunu izrazito raste, gubitak od devet značajnih znamenaka nije rijetkost. Zbog toga suvremeni računalni programi pri dekompoziciji globalne matrice krutosti prognoziraju broj netočnih znamenaka m pomoću izraza

$$m = \log(k_{ii}) - \log(\bar{k}_{ii}), \tag{1}$$

gdje je k_{ii} dijagonalni član matrice krutosti, a $\bar{k}_{ii}$ njegova promjena pri dekompoziciji.
Čim je $m > 6$ program nagovještava manje pouzdano rješenje jer prognozira gubitak preostalih (približno) devet znamenaka prilikom supstitucije naprijed odnosno natrag. Ako je $m > 11$ u modelu postoje ozbiljni problemi poput neispravnih rubnih uvjeta (globalne nestabilnosti), lokalnog mehanizma (unutar modela), prevelikih razlika u krutostima i slično. Tada nakon završetka proračuna ne možemo očekivati nikakvu točnost rezultata jer su preostale četiri znamenke premalo za kompenzaciju pogreške zaokruživanja. O ovome više u odjeljku 8.2.

2 Motivacija

Iz literature o numeričkoj analizi [2, 3] poznate su teorijske gornje granice pogrešaka zaokruživanja (time i gubitak točnosti značajnih znamenaka) nastalih pri rješavanju sustava linearnih jednadžbi. One su uvijek prikazane u obliku dokazanih nejednakosti. Međutim, prevladava mišljenje da su tako određene granice previše „pesimistične“, a mogu se preciznije odrediti samo u primjerima s poznatim teorijskim rješenjem. Prema tome, postavlja se zapravo pitanje: Kolika je točnost naših numeričkih proračuna? Neki autori [4] tvrde da pogreške zaokruživanja ne mogu ugroziti dobivene rezultate. Ipak, u ovome je radu pokazano da najveći gubitak značajnih znamenaka, pri proračunu relativno malih modela (s nešto manje od 35 000 nepoznanica), iznosi 6 za dobro koncipirane i čak 11 do 12 znamenaka za loše koncipirane modele. Dobri modeli formirani su tako da gubitak znamenaka *isključivo* ovisi o propagaciji pogreške zaokruživanja, dok je kod loših modela uključen i utjecaj nepovoljnog oblika elementa (odjeljak 7.4). U nastavku ćemo opisati izvorište smanjene točnosti proračuna: pogrešku zaokruživanja.

3 Pogreška zaokruživanja

Svaki proračun inženjerske konstrukcije je približan, odnosno opterećen je pogreškama. Od svih pogrešaka koje postoje u postupku projektiranja i izvedbe [5] ovim je radom obuhvaćena samo pogreška zaokruživanja. Ona nastaje pri zapisu realnog broja u ograničenom obliku tzv. *strojnog broja* – broja pohranjenog u memoriji računala. Dovoljna je jedna računska operacija dvaju takvih brojeva da rezultat bude opterećen pogreškom zaokruživanja. Tako dobivena vrijednost je ulazni podatak za nastavak proračuna, pa govorimo o propagaciji pogreške kroz numerički postupak [6, 7]. Ako se radi o samo nekoliko koraka jednostavnog algoritma, pogreška je vrlo mala. Razlog tome je velika preciznost zapisa strojnog broja (4. poglavlje). Međutim, kod složenih algoritama i velikog broja računskih operacija postavlja se pitanje utjecaja takve pogreške na konačnu točnost rezultata. Prema gruboj procjeni, konačne pogreške nastale računalnim proračunom moraju se ograničiti na najviše 1%, odnosno proračunska pogreška mora biti manja za barem jedan red veličine od očekivane ukupne pogreške.

Prema [5] pod ukupnom pogreškom smatramo razliku između rezultata proračuna i ponašanja gotove konstrukcije. Ako isključimo nepouzdana djelovanja potresa ili vjetra ta je razlika

[1] U ovome smislu gubitak ne treba shvatiti doslovno. Radi se zapravo o gubitku (smanjenju) točnosti zapisa.

oko 10% u slučaju dobrih, a 30% u slučaju nešto nepreciznijih modela (primjerice s lošijim oblicima elemenata).
Na taj se način ukupna pogreška neće nepotrebno povećati zbog postupka proračuna. Posvetimo se sada začetku pogreške zaokruživanja: strojnom zapisu realnih brojeva.

4 Prikaz realnih brojeva u memoriji računala

Postoji nekoliko načina zapisivanja brojeva u memoriji računala od kojih je, u numeričkim inženjerskim problemima, zapis s pomičnim zarezom (engl. *floating–point*) u najčešćoj uporabi. Osim tehnike pomičnog zareza u uporabi su još neke tehnike poput nepomičnog zareza (engl. *fixed point*), racionalnog broja (engl. *floating–slash*) i logaritamskog oblika (engl. *signed logarithm*). Svakome je zapisu pridružen način postupanja s brojevima, odnosno nad njima su definirane logičke i aritmetičke operacije. Format broja prikazanog pomičnim zarezom sastoji se od: predznaka $+$ ili $-$, baze β, eksponenta E i preciznosti p koja predstavlja broj znamenaka mantise $\pm d_0, d_1 d_2 d_3 \ldots d_{p-1}$. Tako definirani parametri određuju *strojni broj* koji možemo zapisati u obliku:

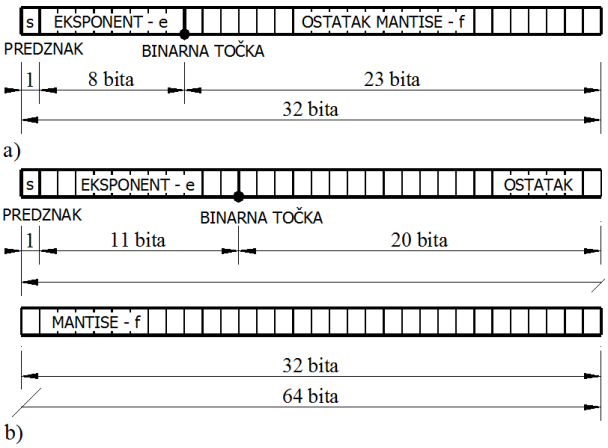
$$\pm d_0, d_1 d_2 d_3 \ldots d_{p-1} \times \beta^{\mathsf{E}}, \tag{2}$$

gdje oznaka $\times$ nije računska operacija, a za znamenke mantise vrijedi $0 \leq d_i < \beta$.
Teorijski, baza, preciznost i eksponent mogu poprimiti različite vrijednosti. Ipak, one su određene raznim normama. Trenutačno najpopularnija norma koja definira format brojeva s pomičnim zarezom i računske operacije među njima jest IEEE 754 (engl. IEEE 754 *Standard for Binary Floating–Point Arithmetic*, ANSI/IEEE Std. 754–198). Ova je norma ugrađena u sve SPARC, Intel i PowerPC procesore. Baza norme IEEE 754 uvijek je dva ($\beta = 2$), a preciznost p može poprimiti dvije vrijednosti: 24 za brojeve jednostruke preciznosti (engl. *single precision*) i 53 za brojeve dvostruke preciznosti (engl. *double precision*). Uz ovu osnovnu podjelu IEEE 754 nudi još i produženu jednostruku i dvostruku preciznost (engl. *single* (*double*) *extended precision*). Vrijednosti parametara koje pripadaju ovim formatima prikazani su u tablici 1.

Tablica 1. Vrijednosti parametara prema IEEE 754

Formati	**Baza** β	**Preciznost** p	**Eksponent**	
			E_{max}	E_{min}
jednostruki	2	24	127	-126
produženi jednostruki	2	32	1 023	-1 022
dvostruki	2	53	1 023	-1 022
produženi dvostruki	2	64	16 383	-16 383

Navedeni formati brojeva strojno se zapisuju kao jedna ili dvije 32–bitne memorijske riječi podijeljene na tri cjeline, pri čemu u binarnom sustavu bitovi poprimaju vrijednosti 0 ili 1 (slika 1.).

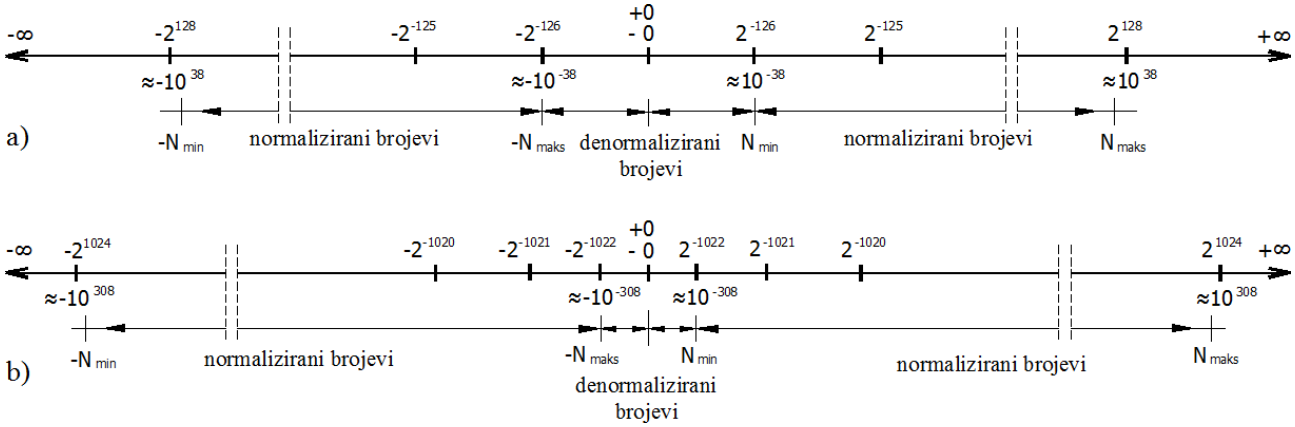


Slika 1. Raspored bitova unutar memorijske riječi (prema IEEE 754**): a) brojevi jednostruke preciznosti, b) brojevi dvostruke preciznosti**

Vrijednost prvoga bita određuje predznak broja: 0 ako je broj pozitivan ili 1 ako je negativan. Polje za prikaz eksponenta ima 8 odnosno 11 bitova i počinje s pomakom od 127 za jednostruku odnosno 1023 za dvostruku preciznost. Time se izbjegava negativna vrijednost eksponenta e. Na taj način minimalnu vrijednost eksponenta u jednostrukom formatu $\mathsf{E}_{min} = -126$ pohranjujemo kao $e = -126 + 127 = 1$. Preostala 23 ili 52 bita ne sadrže cijelu mantisu. Na slici 1. dio memorijske riječi označen je kao ostatak mantise jer se izostavlja znamenka odnosno bit lijevo od binarne točke (engl. *hidden bit*). Naime, ako je moguće, brojevi s pomičnim zarezom pohranjuju se u normaliziranom obliku; obliku s vodećom znamenkom kao jedinicom. Normalizirani brojevi (engl. *normalized numbers*) imaju dvije prednosti [7]:

- svaki normalizirani broj ima jedinstveni prikaz i
- nije potrebno izravno pamtiti vodeću znamenku.

Vodeća znamenka uvijek je jednaka jedinici pa se pripadni bit dodaje radi točnijeg prikaza ostatka. Brojevi koje nije moguće prikazati u normaliziranom obliku, imaju nulu kao vodeću znamenku i nazivaju se denormaliziranim (engl. *subnormal numbers*). Na slici 2. prikazani su brojevni pravci s označenim ekstremima normaliziranih i denormaliziranih brojeva jednostruke i dvostruke preciznosti prema IEEE 754. Zbog uvođenja normalizacije, nulu je potrebno posebno definirati. Brojevni pravci na slici 2. sadrže dva različita zapisa nule, tzv. nulu s predznakom (engl. *signed zero*). Predznaci nule nemaju posebno numeričko značenje jer se uvijek radi o nuli. Oni se pojavljuju samo kao posljedica strojnog zapisa.



Slika 2. Skica brojevnog pravca prema IEEE 745 **za brojeve: a) jednostruke preciznosti, b) dvostruke preciznosti**

Osim nule s predznakom, norma IEEE 754 uvodi i posebne znakove koji se odnose na beskonačne vrijednosti: $-\texttt{Inf}(-\infty)$ i $+\texttt{Inf}(+\infty)$ (engl. *infinite numbers*) i nedefinirani iznos `NaN` (engl. ***Not a Number***). Očekivana pojava nule ili beskonačnih vrijednosti u računskim operacijama ne mora dovesti do prekida proračuna jer IEEE aritmetika podržava poznata računska pravila poput:

$$x \pm \infty = \pm\infty, \qquad x \cdot 0 = 0, \qquad x/0 = \infty,$$
$$+\infty + \infty = \infty, \qquad x \cdot \infty = \infty, \qquad x/\infty = 0.$$

Međutim, računskim operacijama koje nisu dobro definirane:

$$\infty - \infty, \qquad 0/0, \qquad x\,\texttt{REM}\,0, \qquad \sqrt{-x},$$
$$0 \cdot \infty, \qquad \infty/\infty, \qquad \infty\,\texttt{REM}\,x,$$

gdje je `REM` ostatak dijeljenja (engl. *reminder after division*), IEEE standard pridružuje vrijednost `NaN`. Pojava `NaN` u nekome izrazu koji sadrži aritmetičku ili logičku operaciju također rezultira s `NaN`. Prema tome i sve izlazne varijable poprimaju vrijednost `NaN` pa se proračun, najčešće uz upozorenje, prekida.

Iz opisa strojnog broja možemo zaključiti da svaka od mogućih kombinacija nula i jedinica smještenih u 32, odnosno 64 bita predstavlja po jedan broj s pomičnim zarezom. Očito je da možemo zapisati samo konačan odnosno ograničen skup brojeva. Drugim riječima, postoje najmanji i najveći broj skupa, a između dvaju susjednih brojeva ne možemo definirati novi. Prema tome, nije moguće svaki realni, pa čak niti svaki racionalni broj točno prikazati u formatu broja s pomičnim zarezom i konačnim brojem znamenaka. Primjeri su brojevi koji su manji (veći) od najmanjeg (najvećeg) broja s pomičnim zarezom, ili (što je najčešći slučaj) brojevi koji leže između dva broja s pomičnim zarezom. Primjerice, decimalni broj 0,1 ne može se u binarnome formatu bilo koje točnost prikazati egzaktno, nego približno kao $1,10011001100110011\ldots \times 2^{-4}$. Zbog toga će broj s pomičnim zarezom i prije proračuna biti aproksimativan odnosno zaokružen.

4.1 Ispravnost i točnost postupka zaokruživanja

Uobičajeno pravilo koje rabi IEEE 745 jest zaokruživanje prema najbližem broju, uz iznimke koje se odnose na granične vrijednosti normaliziranih strojnih brojeva:

$$\begin{aligned} x > N_{\max}, &\qquad \texttt{round}\,(x) = \infty, \\ x < -N_{\min}, &\qquad \texttt{round}\,(x) = -\infty, \end{aligned} \tag{3}$$

gdje su $N_{\max}$ i $N_{\min}$ najveći i najmanji normalizirani brojevi, a `round` oznaka za funkciju zaokruživanja.
Ako se broj nalazi na sredini između dvaju strojnih brojeva tada se bira onaj broj s pomičnim zarezom koji ima nulu u najdaljem bitu, odnosno na najmanje značajnom mjestu, a to znači da se bira parni broj (engl. *rounding to nearest even*). Općenito, zaokružena vrijednost prema bilo kojem od navedenih modela jest približna i iznosi

$$\texttt{round}\,(x) = x(1+\delta), \tag{4}$$

gdje je δ *relativna pogreška zaokruživanja*, čija apsolutna vrijednost mora biti manja od strojne preciznosti ε:

$$|\delta| = \frac{|\texttt{round}\,(x) - x|}{|x|} \leq \varepsilon. \tag{5}$$

To je temeljni izraz koji se često rabi pri analizama pogreške zaokruživanja [8]. Vrijednost strojne preciznosti ε (engl. *machine precision*, *machine epsilon*, *macheps*) sadržana je u razlici dvaju susjednih strojnih brojeva $(x_+ - x_-)$ koja iznosi:

$$\texttt{ulp} = x_+ - x_- = (0,00\ldots01)_2 \times 2^e \tag{6}$$

i predstavlja razliku brojeva s jedinicom na zadnjem mjestu (kratica `ulp` od engl. ***u**nit in the **l**ast **p**lace*). Prvi dio izraza (6) sadrži konstantu za određenu preciznost p i određuje strojnu preciznost kao:

$$\varepsilon = (0,00\ldots01)_2 = 2^{-(p-1)}. \tag{7}$$

Prema [9] strojnu preciznost možemo definirati i kao najmanji pozitivni broj za koji vrijedi:

$$1 + \varepsilon \neq 1. \tag{8}$$

Vrijednost strojne preciznosti prema (7) i (8) za jednostruki format iznosi $\varepsilon = 2^{-23} \approx 10^{-7}$, što približno odgovara vrijednosti od sedam značajnih decimalnih znamenaka. Kada se upotrebljava format dvostruke preciznosti vrijednost strojne preciznosti raste na $\varepsilon = 2^{-52} \approx 10^{-16}$, odnosno kriterij za relativnu pogrešku postaje stroži, a broj značajnih decimalnih znamenaka penje se na približno šesnaest. Za model zaokruživanja koji podržava IEEE 745 (zaokruživanje prema najbližem broju), gornja granica relativne pogreške (5) smanjuje se na pola:

$$|\delta| \leq \frac{1}{2}\varepsilon = 2^{-p}. \tag{9}$$

Osim pogrešaka zaokruživanja koje nastaju prikazom realnoga broja strojnim i sama aritmetika pomične točke stvara pogreške. Vrlo često rezultat neke računske operacije provedene nad brojevima s pomičnim zarezom ne daje rezultat u zadanom formatu. Kao primjer možemo podijeliti broj 1 s brojem 10. Oba broja imaju točan prikaz kao strojni brojevi, dok njihov omjer $1/10 = 0,1$, kako smo pokazali, prije pohrane mora biti zaokružen.
U postupcima numeričkog proračuna IEEE norma zahtijeva da rezultat osnovnih računskih operacija zbrajanja, oduzimanja, množenja i dijeljenja bude točno određen i tek onda zaokružen. Konačna aritmetika koja zadovoljava navedeni uvjet naziva se aritmetikom s korektnim zaokruživanjem (engl. *correctly* (*exactly*) *rounded operations*). U tome slučaju relativna pogreška rezultata jedne operacije ne prelazi vrijednost određenu izrazom (9).
Osim za osnovne operacije, IEEE norma ovaj zahtjev proširuje i na operaciju drugog korijena, ostatka pri dijeljenju i pretvorbe između različitih formata brojeva. Radi obuhvaćanja svih slučajeva koji se mogu dogoditi primjenom aritmetike s korektnim zaokruživanjem, ponekad su potrebni tzv. sigurnosni (zaštitni, rezervni) bitovi smješteni unutar registara u kojima se provode osnovne operacije. Najčešće se radi o tri dodatna bita. Pri tome pohrana samoga broja ostaje u IEEE formatu.
Može se dogoditi da dobiveni rezultat nekog proračuna ne pripada intervalima normaliziranih brojeva s pomičnim zarezom. Ako je ta vrijednost konačna, ali veća (manja) od $+N_{\max}$ ($-N_{\min}$), govorimo o prekoračenju (engl. *overflow*). U IEEE aritmetici takav se broj zaokružuje prema modelu (3) bez prekida proračuna. Pojava rezultata različitog od nule koji je po svojoj apsolutnoj vrijednosti manji od najmanjeg normaliziranog broja naziva se potkoračenje (engl. *underflow*). Ako u IEEE aritmetici zaokružimo takav broj na najbližu strojnu vrijednost, možemo se naći u intervalu denormaliziranih brojeva. Pri tome uvodimo relativnu pogrešku zaokruživanja koja je veća od uobičajene (4) s tendencijom porasta ako se denormalizirani broj smanjuje. Prema [6] vrijednost porasta relativne pogreške nije veća od $2^{-150} \approx 10^{-45}$ za jednostruku odnosno $2^{-1075} \approx 10^{-324}$ za dvostruku preciznost. To je u većini slučajeva prihvatljivije od potpunog gubitka relativne točnosti izbjegavanjem denormaliziranih brojeva i zaokruživanjem na nulu.
Uvođenje aritmetike denormaliziranih brojeva naziva se postupnim potkoračenjem (engl. *gradual underflow*) i predstavlja sporni dio IEEE 745 norme. Iako može izazvati znatno produljenje trajanja proračuna (ako nema strojnu podršku), postupno potkoračenje odnosno aritmetika denormaliziranih brojeva danas je prihvaćena kao praktično rješenje za popunjavanje relativno velike praznine između nule i najvećeg negativnog ($-N_{\max}$) odnosno najmanjega pozitivnoga normaliziranog broja ($N_{\min}$) [6, 9] (slika 2.).

Napomenimo da je u nekim programskim paketima moguće proizvoljno povećati broj značajnih znamenaka. Ovo alternativno rješenje za povećanje točnosti provodi se programski (nema strojnu podršku), pa se proračuni najčešće odvijaju uz zamjetno produljenje trajanja (i do 400 puta) i znatan utrošak memorije računala.

5 Procjena pogreške zaokruživanja

Neka je $\mathbf{Ku} = \mathbf{f}$ osnovni sustav n linearnih jednadžbi pri kojem su zadani podaci: članovi matrice sustava $\mathbf{K}$ ($\mathbf{K} \in \mathbb{Q}^n \times \mathbb{Q}^n$) i vektora $\mathbf{f}$ ($\mathbf{f} \in \mathbb{Q}^n$), elementi skupa racionalnih brojeva $\mathbb{Q}$ (zapisani u obliku cijelih brojeva ili razlomaka), pa ne sadrže ulaznu pogrešku. Utjecaj pogreške zaokruživanja na točno rješenje $\mathbf{u}$ ($\mathbf{u} \in \mathbb{Q}^n$) procjenjuje se analizom uvjetovanosti sustava. Tako pri dobro uvjetovanom sustavu mala promjena (perturbacija) ulaznih podataka daje malu promjenu rezultata, dok pri loše uvjetovanom sustavu mala promjena ulaznih vrijednosti može prouzročiti iznenađujuće veliku pogrešku u rezultatu. Ako perturbiramo osnovni sustav možemo pisati

$$(\mathbf{K} + \delta\mathbf{K})(\mathbf{u} + \delta\mathbf{u}) = \mathbf{f} + \delta\mathbf{f}, \tag{10}$$

gdje su $\delta\mathbf{K}$ i $\delta\mathbf{f}$ male pogreške ulaznih podataka, a $\delta\mathbf{u}$ pogreška rezultata (ne nužno mala). Razlika perturbiranog i osnovnog sustava daje početni izraz za pogrešku rezultata:

$$\delta\mathbf{u} = \mathbf{K}^{-1}(\delta\mathbf{f} - \delta\mathbf{K}\mathbf{u} - \delta\mathbf{K}\delta\mathbf{u}). \tag{11}$$

Gornja granica ove pogreške dobiva se primjenom neke norme na članove prethodnog izraza koji tada (zbog poznate nejednakosti trokuta) prelazi u nejednadžbu

$$\|\delta\mathbf{u}\| \leq \|\mathbf{K}^{-1}\|\left(\|\delta\mathbf{K}\|\,\|\delta\mathbf{u}\| + \|\delta\mathbf{K}\|\,\|\mathbf{u}\| + \|\delta\mathbf{f}\|\right). \tag{12}$$

Sređivanjem ovog izraza uz uvjet da je perturbacija mala, čime je zadovoljen uvjet

$$\|\mathbf{K}^{-1}\|\,\|\delta\mathbf{K}\| < 1, \tag{13}$$

dobivamo gornju granicu relativne pogreške:

$$\frac{\|\delta\mathbf{u}\|}{\|\mathbf{u}\|} \leq \frac{\|\mathbf{K}^{-1}\|\,\|\mathbf{K}\|}{1-\|\mathbf{K}^{-1}\|\,\|\mathbf{K}\|\dfrac{\|\delta\mathbf{K}\|}{\|\mathbf{K}\|}}\left(\frac{\|\delta\mathbf{K}\|}{\|\mathbf{K}\|}+\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|}\right). \quad (14)$$

Vrijednost umnoška

$$\kappa(\mathbf{K}) = \|\mathbf{K}^{-1}\|\,\|\mathbf{K}\| \quad (15)$$

naziva se brojem uvjetovanosti (engl. *condition number*) matrice $\mathbf{K}$ i temelji se na inverzu matrice.[2] Uvrštavanjem (15) u (14) dobivamo izraz

$$\frac{\|\delta\mathbf{u}\|}{\|\mathbf{u}\|} \leq \frac{\kappa(\mathbf{K})}{1-\kappa(\mathbf{K})\dfrac{\|\delta\mathbf{K}\|}{\|\mathbf{K}\|}}\left(\frac{\|\delta\mathbf{K}\|}{\|\mathbf{K}\|}+\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|}\right) \quad (16)$$

koji povezuje relativnu pogrešku rezultata $\|\delta\mathbf{u}\|/\|\mathbf{u}\|$ s relativnom pogreškom matrice sustava i vektora desne strane $\|\delta\mathbf{K}\|/\|\mathbf{K}\| + \|\delta\mathbf{f}\|/\|\mathbf{f}\|$ kao ulaznih podataka. Koeficijent koji ih povezuje približno je jednak broju uvjetovanosti $\kappa(\mathbf{K})$ ako je norma pogreške (perturbacije) $\|\delta\mathbf{K}\|$ dovoljno mala [6]. Tada je $\kappa(\mathbf{K})\,\|\delta\mathbf{K}\|/\|\mathbf{K}\| \approx 0$. Uz navedenu pretpostavku izraz (16) prelazi u oblik

$$\frac{\|\delta\mathbf{u}\|}{\|\mathbf{u}\|} \leq \kappa(\mathbf{K})\left(\frac{\|\delta\mathbf{K}\|}{\|\mathbf{K}\|}+\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|}\right) \quad (17)$$

koji najčešće srećemo u literaturi i prema kojemu je gornja granica relativne pogreške rezultata proporcionalna zbroju relativnih pogrešaka ulaznih podataka. Koeficijent proporcionalnosti jest broj uvjetovanosti $\kappa(\mathbf{K})$. Za proračun prethodnih izraza najčešće se upotrebljavaju sljedeće vektorske i matrične norme [2]:

$$\|\mathbf{u}\|_1 = \sum_{i=1}^{n}|u_i|, \qquad \text{norma sume}$$

$$\|\mathbf{u}\|_2 = \sqrt{\sum_{i=1}^{n}|u_i|^2}, \qquad \text{euklidska norma}$$

$$\|\mathbf{u}\|_\infty = \max_i |u_i|, \qquad \text{norma beskonačno}$$

$$\|\mathbf{K}\|_1 = \max_{1\leq j\leq n}\sum_{i=1}^{n}|a_{ij}|, \qquad \text{norma sume po stupcima}$$

$$\|\mathbf{K}\|_2 = \sqrt{\lambda_{\max}(\mathbf{K}^{\mathrm{T}}\mathbf{K})}, \qquad \text{norma dva}$$

$$\|\mathbf{K}\|_\infty = \max_{1\leq i\leq n}\sum_{j=1}^{n}|a_{ij}|, \qquad \text{norma sume po redcima i}$$

$$\|\mathbf{K}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\mathrm{tr}\,(\mathbf{K}\mathbf{K}^{\mathrm{T}})}, \text{ Schurova norma}$$

gdje je $|\cdot|$ oznaka za apsolutnu vrijednost komponente matrice ili vektora, $\lambda_{\max}$ najveća vlastita vrijednost matrice, a $\mathrm{tr}\,(\cdot)$ oznaka za trag matrice. Normu dva često zovemo spektralnom, a Schurovu normu Frobeniusovom.

Osim za procjenu gornje granice relativne pogreške rezultata odnosno uvjetovanosti sustava jednadžbi, broj uvjetovanosti možemo povezati s očekivanim brojem izgubljenih (pogrešnih) decimalnih mjesta dobivenog rješenja. Ako relativna pogreška ulaznih podataka (članova $\mathbf{K}$ i $\mathbf{f}$) ne prelazi vrijednost strojne preciznosti 10^{-7} odnosno 10^{-16}, tada prema [10] relativnu pogrešku rješenja (17) možemo zapisati kao

$$\frac{\|\delta\mathbf{u}\|}{\|\mathbf{u}\|} = 10^{-s} \leq \kappa(\mathbf{K})10^{-p}, \quad (18)$$

gdje je p iznos jednostruke odnosno dvostruke preciznosti određen prema (9), a s očekivana točnost rezultata. Logaritmiranjem prethodnog izraza dobivamo

$$s \leq p - \log\kappa(\mathbf{K}) \quad (19)$$

što znači da logaritam broja uvjetovanosti predstavlja maksimalni očekivani broj izgubljenih značajnih znamenaka. U skladu s (1) možemo pisati $m = \log\kappa(\mathbf{K})$ iako se ne mora raditi o istoj brojki. Izraz (1) prirodan je pri dekompoziciji, često u uporabi, i trivijalan prema određivanju $\log\kappa(\mathbf{K})$ kod kojega, prema (15), treba invertirati matricu krutosti, a to je za velike sustave spor i memorijski zahtjevan postupak.

6 Temeljna zamisao postupka za izoliranu procjenu pogreške zaokruživanja

Cilj je odrediti sve članove $\mathbf{K}$, $\mathbf{u}$ i $\mathbf{f}$ osnovnog sustava u obliku razlomaka, čime isključujemo utjecaj pogreške zaokruživanja. Tako dobiveni sustav i njegovo rješenje zovemo *etalonom*. Njega uvijek možemo odrediti obratnom metodom: ako je $\mathbf{K}$ prikazana racionalnim brojevima, na lijevu stranu uvedemo rješenje $\mathbf{u}$ prikazano na isti način i običnim množenjem $\mathbf{Ku}$ odredimo i $\mathbf{f}$ u obliku racionalnih brojeva. Pri tome je ipak poželjno da sustav opisuje neki realni problem.

Osnovni sustav bismo mogli odrediti i izravnom metodom u cjelobrojnoj aritmetici (izračunati $\mathbf{u} = \mathbf{K}^{-1}\mathbf{f}$ u cjelobrojnom obliku). Međutim, već kod malog broja nepoznanica dolazi do vrtoglavog povećanja brojnika i nazivnika (čak i uz kraćenje razlomaka) pa nastaju veliki zahtjevi za brzinom i memorijom računala kojima nismo mogli raspolagati.

Radi numeričkog pristupa rješenju osnovnog sustava potrebno je zapisati članove $\mathbf{K}$ i $\mathbf{f}$ u obliku strojnih brojeva čime ulazni podaci dobivaju (unesenu, početnu) pogrešku zaokruživanja $\delta\mathbf{K}$ i $\delta\mathbf{f}$. Time dobivamo perturbirani sustav:

$$(\mathbf{K}+\delta\mathbf{K})\overline{\mathbf{u}} = \mathbf{f}+\delta\mathbf{f}, \quad (20)$$

[2] Ovo nije jedini način određivanja broja uvjetovanosti. Primjerice, problemu vlastitih vrijednosti matrice $\mathbf{K}$ pripada izraz $\kappa(\mathbf{K}) = \lambda_{\max}(\mathbf{K})/\lambda_{\min}(\mathbf{K})$, gdje su $\lambda_{\max}(\mathbf{K})$ i $\lambda_{\min}(\mathbf{K})$ najveća i najmanja vlastita vrijednost od $\mathbf{K}$ [6].

gdje $\bar{\mathbf{u}}$ označava približno rješenje koje, osim početne pogreške, sadrži i njezin rast zbog provedbe algoritma za proračun sustava.

Poznavajući točno i približno rješenje i pazeći na veličinu početnih pogrešaka možemo prema (17) postaviti nejednakost

$$\frac{\|\mathbf{u}-\bar{\mathbf{u}}\|}{\|\mathbf{u}\|} \leq \kappa(\mathbf{K}) \left(\frac{\|\delta\mathbf{K}\|}{\|\mathbf{K}\|} + \frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|} \right), \tag{21}$$

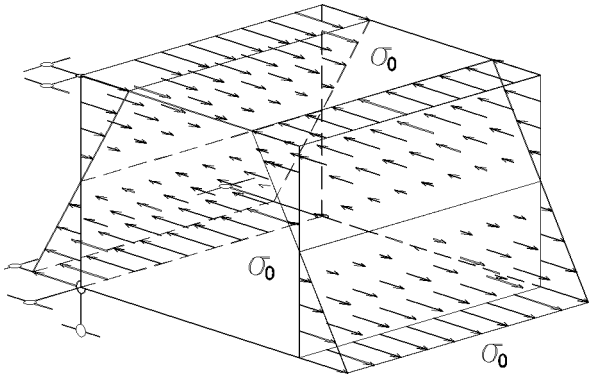
kod koje su, uz izbor prikladne norme, poznati svi članovi. Najmanji broj uvjetovanosti dobiven je primjenom norme dva [11] pa je, zbog „najmanje pesimistične“ ocjene pogreške, primijenjena u svim izrazima. Na temelju posljednje nejednakosti možemo usporediti gornju granicu relativne pogreške (desna strana u (21)) s *točnom* pogreškom (lijeva strana u (21)) i na taj način utvrditi korektnost tako određene granice, ponajprije zbog uvriježenog mišljenja o njezinoj pretjeranoj sigurnosti. Štoviše, možemo usporediti i procjenu gubitka značajnih znamenaka prema (19) sa stvarnim gubitkom određenim razlikom $(\mathbf{u}-\bar{\mathbf{u}})$.

7 Realizacija postupka za procjenu pogreške zaokruživanja

Da bismo proveli opisani postupak treba odabrati problem kojemu pripada osnovni sustav zapisan racionalnim brojevima. Uz današnje programe koji posjeduju simboličku algebru takav je *svaki* primjer kod kojega ulazne podatke **K** i **f** možemo zapisati cijelim brojevima i razlomcima jer primjenom cjelobrojne aritmetike na izravnu metodu rješavanja sustava dobivamo i cjelobrojno rješenje. Ako nemamo snažno računalo za provedbu postupka, uvijek možemo odabrati obratnu metodu ili problem s analitičkim rješenjem kojemu pripada osnovni sustav s racionalnim brojevima. U potonjem slučaju obratna metoda služi samo za provjeru desne strane, a izravna metoda (ako je možemo provesti) za provjeru rješenja. Pokažimo u nastavku realizaciju postupka primjenom klasičnog problema linearne teorije elastičnosti s analitičkim rješenjem.

7.1 Matematički model etalona: čisto savijanje kvadra

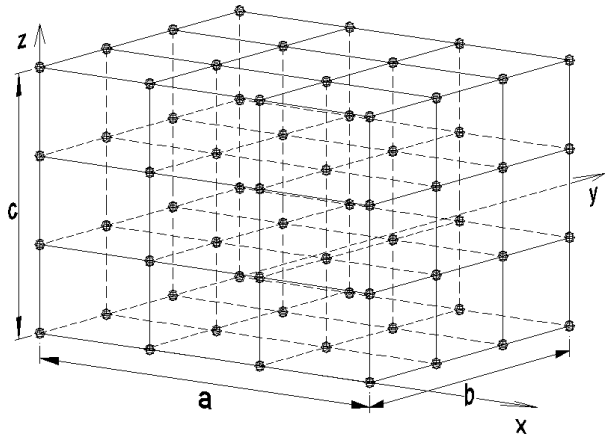
Odabran je matematički model kvadra s nehomogenim stanjem naprezanja σ_0 po rubu, koje se linearno mijenja po visini (djelovanje čistog savijanja; slika 3.). Minimalni broj i raspored ležajnih veza koje sprječavaju pomake modela kao krutog tijela prikazani su na slici. Opisanom problemu pripada analitičko rješenje za komponente polja pomaka [12], koje se uz izbor ulaznih podataka (o geometriji, materijalu i opterećenju) u obliku racionalnih brojeva također mogu prikazati racionalnim brojevima.



Slika 3. Matematički model etalona

7.2 Zapis etalona u cjelobrojnom obliku

Za tvorbu osnovnog sustava upotrijebljena je metoda konačnih elemenata, odnosno Lagrangeov trikubični konačni element u obliku kvadra [13]. Element je C^0 kompatibilan, s koordinatnim funkcijama u obliku Lagrangeovog polinoma trećeg stupnja. Sadrži 64 pravilno raspoređena čvora (slika 4.), svaki s tri translacijska stupnja slobode, što ukupno daje 192 stupnja slobode.



Slika 4. Lagrangeov trikubični konačni element

Uz cjelobrojne ulazne podatke koeficijenti Lagrangeovog polinoma su razlomci, pa integrale za tvorbu lokalne matrice krutosti možemo riješiti analitički, s rezultatom u obliku razlomaka. Time izbjegavamo postupak Gaußove integracije čiji je rezultat gotovo uvijek realan broj. Na taj način dobivamo članove matrice u obliku racionalnih brojeva. Tvorba lokalne i globalne matrice krutosti, vektora pomaka i opterećenja realizirani su programom Mathematica [14] koja podržava rad s racionalnim brojevima.

Uobičajeni postupak statičke kondenzacije unutarnjih čvorova nije proveden jer kondenziranu matricu krutosti nije moguće odrediti u obliku razlomaka ili su iznosi brojnika i nazivnika preveliki. Jasno, nekondenzirana matrica ne utječe na točnost elementa; samo sadrži veći broj članova što je sa stajališta potrebe za tvorbom velikog sustava jednadžbi čak povoljno.

Element je provjeren na nekim elementarnim rješenjima linearne teorije elastičnosti i problemu sinusnog opterećenja pravokutne, zglobno oslonjenje ploče [15]. Dodatno je, za potrebe ovoga rada, za nepridržani element određena determinanta i vlastite vrijednosti matrice krutosti. Iznos determinante i prvih šest vlastitih vrijednosti (kojima odgovaraju pomaci elementa kao krutoga tijela) zbog cjelobrojnog je zapisa matrice jednako točnoj, cjelobrojnoj nuli. (U Mathematici postoji i nula s ostatkom, odnosno nula kao realan broj.) Slično tome, rezidual $\mathbf{f} - \mathbf{K}\mathbf{u}$ i pogreška rješenja $\mathbf{u} - \mathbf{K}^{-1}\mathbf{f}$ jednaki su cjelobrojnoj nuli. (Umnožak **Ku** ovdje simbolizira obratnu, a $\mathbf{K}^{-1}\mathbf{f}$ izravnu metodu rješavanja sustava.)

7.3 Numerička realizacija etalona

Zapisom ulaznih podataka (članova **K** i **f**) u strojnom obliku jednostruke ili dvostruke preciznosti unosimo u osnovni sustav početnu pogrešku čime počinje numerička realizacija etalona. Nastaje, dakle, perturbirani sustav (20) koji je potom riješen primjenom dvostruke preciznosti, bez obzira na broj decimalnih mjesta na koja su zaokruženi ulazni podaci. To je uobičajeni način realizacije standardnih numeričkih postupaka s realnim brojevima. Time smo dobili rješenje perturbiranog sustava $\bar{\mathbf{u}}$.

Proračunom elemenata **K** i **f** s jednostrukom, a rješavanjem sustava s dvostrukom preciznosti možemo dodatno ispitati osjetljivost (stabilnost) sustava na poremećaj ulaznih podataka. Ipak, takav postupak nije dosljedan niti uobičajen u standardnim numeričkim proračunima.

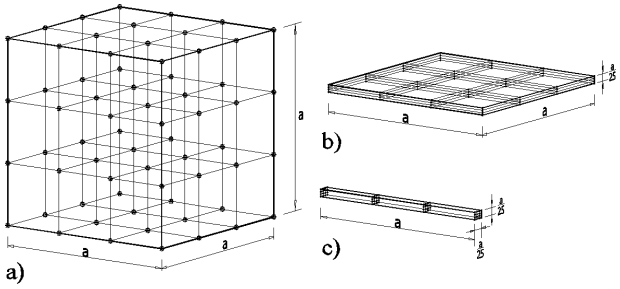
Zbog velikog broja nula u globalnoj matrici i znatnog utroška memorije, upotrijebljen je štedni zapis matrice temeljen na metodi potpunog knjiženja (engl. *full bookkeeping method*) kojom se spremaju samo položaji i iznos članova različitih od nule. Štedno spremljeni sustav riješen je izravnom, multifrontalnom metodom temeljenom na Gaußovoj eliminaciji. Radi se o provjerenom i stabilnom algoritmu programa Mathematica koji ne uzrokuje dodatne inducirane pogreške (engl. *induced numerical errors*) nastale primjenom lošeg postupka proračuna.

7.4 Odabrani primjeri

Za provedbu postupka odabrana su tri osnovna numerička primjera etalona: kocka, ploča i štap (slika 5.). Broj konačnih elemenata, odnosno nepoznatih stupnjeva slobode jednak je za sve modele. Podaci su priloženi u tablici 2. Spomenimo usput da širina poluvrpce iznosi

$$\max_{k_{ij}\neq 0} |i - j|.$$

Zbog potrebe za velikom količinom memorije pri simboličkim operacijama s velikim brojem razlomaka nije bilo moguće analizirati veće modele.



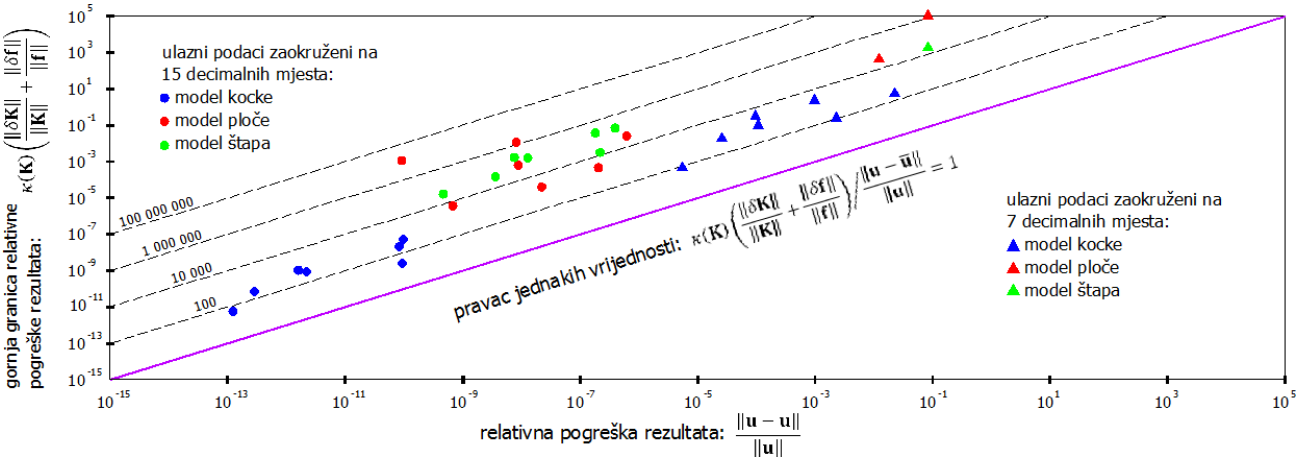
Slika 5. Numerički primjeri etalona (diskretizacija jednim elementom): a) kocka, b) ploča, c) štap

Primjer dobrog modela: kocka. U ovome modelu pogreške počinju i propagiraju samo zbog zaokruživanja realnih brojeva. Prilikom tvorbe modela eliminirani su svi ostali učinici koji utječu na točnost proračuna. Tako je, osim prije spomenutih mjera, odabrana kocka kao najbolji oblik elementa, s jednakim rasporedom čvorova duž koordinatnih smjerova, pa i izotropija modela ostaje sačuvana.

Loši primjeri: ploča i štap. Ploča i štap diskretizirani su volumnim elementima vrlo nepovoljnog oblika jer smo htjeli analizirati utjecaj takve pojave na točnost proraču-

Tablica 2. Osnovni podaci o diskretizaciji. Tablica vrijedi za sve modele

Stupanj diskretizacije	Broj elemenata	Broj nepoznanica	Broj članova matrice K	Broj nenultih članova matrice K	Širina poluvrpce	Postotak popunjenosti
$1\times1\times1$	1	186	34 596	28 836	192	83
$2\times2\times2$	8	1 023	1 046 529	212 819	526	20
$3\times3\times3$	27	2 994	8 964 036	688 262	1018	7,7
$4\times4\times4$	64	6 585	43 362 225	1 593 653	1846	3,7
$5\times5\times5$	125	12 282	150 847 524	3 057 502	2468	2,0
$6\times6\times6$	216	20 571	423 166 041	5 248 319	3466	1,2
$7\times7\times7$	343	31 938	957 159 844	8 274 614	5400	0,86



Slika 6. Ovisnost gornje granice o točnoj vrijednosti relativne pogreške rezultata

na. Zbog toga je ploča modelirana izrazito pločastim elementima, omjera stranica 25 : 25 : 1 (slika 5.b), a štap vrlo izduljenim elementima, omjera 25 : 1 : 1 (slika 5.c). Naravno, za praktične bismo potrebe ove probleme učinkovitije aproksimirali plošnim odnosno štapnim elementima.

8 Rezultati proračuna

Proračuni su realizirani primjenom 64–bitnog operativnog sustava Linux Debian sa 16 GB RAM memorije. Rezultati su, ponajprije, poslužili da bismo odgovorili na dva pitanja. Prvo: Koliko je „pesimistična“ primjena broja uvjetovanosti na procjenu pogreške? I drugo: Koliko je točno značajnih znamenaka izgubljeno tijekom proračuna?

8.1 Usporedba točne i procijenjene relativne pogreške

Usporedba teorijske gornje granice i točne vrijednosti relativne pogreške primjenom (21) provedena je samo za numeričke modele koji zadovoljavaju (13). Taj uvjet nije ispunilo nekoliko loših primjera (ploče i štapa) s podacima zaokruženim na sedam decimalnih mjesta. Loša uvjetovanost takvih primjera čini normu $\|\mathbf{K}^{-1}\|$ vrlo velikom, a sustav izrazito osjetljivim na perturbaciju ulaznih podataka.

Tada nije moguće, primjenom (21), procijeniti gornju granicu pogreške. Za ostale primjere rezultati očekivano leže iznad pravca jednakih vrijednosti točne i procijenjene pogreške (slika 6.). Omjer pogrešaka je za točke toga pravca jednak jedinici. Točne su vrijednosti za dobre primjere (kocke) najbliže procijenjenim iznosima, dakle najbliže spomenutom pravcu. Za ulazne podatke zaokružene na sedam decimala najveći omjer procijenjene i točne vrijednosti pogreške iznosi 3000 (za sustav od 6585 nepoznanica), a za početni zapis na 15 decimala samo 600 (sustav od 2994 nepoznanice). Međutim, za loše primjere (ploče i štapovi) omjer postaje puno veći. Kod zapisa ulaznih podataka na sedam decimala taj omjer iznosi 10^6 (samo 1023 nepoznanice), a pri zapisu na 15 decimala 10^7 (12 282 nepoznanice). Možemo zaključiti da gornja granica relativne pogreške rezultata prema (17) daje prihvatljive vrijednosti za sve dobre modele. U tim je primjerima produkt relativne pogreške ulaznih podataka i broja uvjetovanosti dobra procjena relativne pogreške rezultata. S druge strane, iznos broja uvjetovanosti loših modela tako je velik (tablica 3.) da spomenuti produkt daleko nadmašuje stvarnu vrijednost pogreške. U tome je slučaju „pesimizam“ o primjeni broja uvjetovanosti očito opravdan.

8.2 Usporedba točnog i procijenjenog broja izgubljenih značajnih znamenaka

Prema (19) gubitak značajnih znamenaka procjenjuje se primjenom dekadskog logaritma broja uvjetovanosti (tablica 3.). Ta je vrijednost uspoređena s točnim brojem izgubljenih znamenaka određenim najvećom razlikom između komponente pomaka točke etalonskog rješenja i njegove numeričke inačice (slika 7.). Istaknimo odmah: bez obzira na preciznost zapisa ulaznih podataka (članova **K** i **f**), kod istih dimenzija sustava nije zabilježena značajna promjena broja uvjetovanosti. Međutim, početni zapis u jednostrukoj preciznosti znači trenutačni gubitak sedam (netočnih) decimalnih mjesta. Drugim riječima, takvi primjeri odmah raspolažu manjom „rezervom“ (od približno osam) značajnih znamenaka. Premda je proračun i takvih primjera proveden u dvostrukoj preciznosti, manju točnost ulaznih podataka nije moguće nadoknaditi. Zbog toga se gubitak znamenaka prouzročen postupkom proračuna pomiče od osmog prema prvom decimalnom mjestu. Tako kod dobrih modela (kocka) u najgorem slučaju preostaju dvije točne znamenke, a kod loših modela (ploča i štap) već kod prvog stupnja diskretizacije gubimo sve preostale znamenke, odnosno dobivamo pogrešan rezultat. Prema tome, „grublje“ zaokruživanje ulaznih podataka unosi

Tablica 3. Teorijska procjena broja izgubljenih značajnih znamenaka

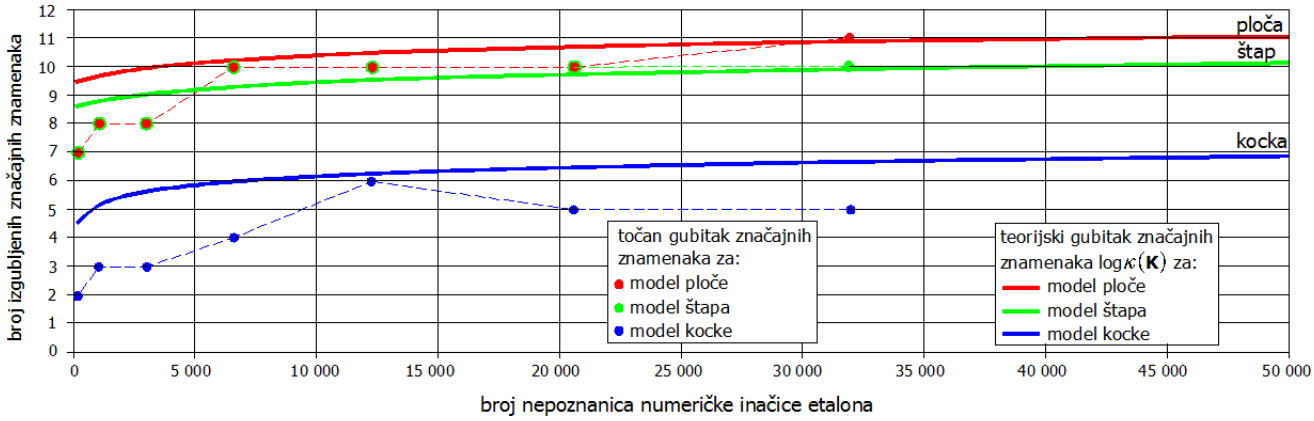
Broj nepoznanica n	Kocka		Ploča		Štap	
	Broj uvjetovanosti $\kappa(\mathbf{K})$	Broj izgubljenih znamenaka $\log\kappa(\mathbf{K})$	Broj uvjetovanosti $\kappa(\mathbf{K})$	Broj izgubljenih znamenaka $\log\kappa(\mathbf{K})$	Broj uvjetovanosti $\kappa(\mathbf{K})$	Broj izgubljenih znamenaka $\log\kappa(\mathbf{K})$
186	$3,43\cdot10^{4}$	4,5	$3,19\cdot10^{9}$	9,5	$4,51\cdot10^{8}$	8,7
1 023	$1,54\cdot10^{5}$	5,2	$5,07\cdot10^{9}$	9,7	$6,61\cdot10^{8}$	8,8
2 994	$4,35\cdot10^{5}$	5,6	$9,48\cdot10^{9}$	10,0	$1,15\cdot10^{9}$	9,1
6 585	$9,48\cdot10^{5}$	6,0	$1,75\cdot10^{10}$	10,2	$2,05\cdot10^{9}$	9,3
12 282	$1,76\cdot10^{6}$	6,2	$3,03\cdot10^{10}$	10,5	$3,48\cdot10^{9}$	9,5
20 571	$2,95\cdot10^{6}$	6,5	$4,89\cdot10^{10}$	10,7	$5,55\cdot10^{9}$	9,7
31 938	$4,57\cdot10^{6}$	6,7	$7,43\cdot10^{10}$	10,9	$8,39\cdot10^{9}$	9,9
100 000	$1,43\cdot10^{7}$	7,2	$2,27\cdot10^{11}$	11,0	$2,54\cdot10^{10}$	10,4

početnu pogrešku koja uzrokuje neprihvatljiva rješenja svih loših primjera. Zbog toga je u nastavku prikazana usporedba teorijskog i stvarnog broja izgubljenih znamenaka samo za zapis ulaznih podataka na 15 decimala (slika 7.).

Iz slike možemo zaključiti da je najveće odstupanje zabilježeno kod numeričkih modela s malim brojem nepoznanica. S porastom nepoznatih stupnjeva slobode odstupanja su za dobre modele na sigurnu stranu, odnosno broj uvjetovanosti precjenjuje točan broj izgubljenih znamenaka. S druge strane, s porastom diskretizacije loših modela odstupanja se smanjuju pa je broj uvjetovanosti dobra procjena gubitka znamenaka. Pri tome je već kod, za današnje prilike, malog broja nepoznanica (oko 32 000) izgubljeno 10 odnosno 11 znamenaka. Zamijetimo čak „optimističnu“ procjenu gubitka znamenaka temeljenu na broju uvjetovanosti za štapne primjere između 5 000 i 20 000 nepoznanica. Zbog nedostatka snažnog računala teorijska je procjena broja $\log\kappa(\mathbf{K})$ za $n > 32\,000$ dobivena primjenom ekstrapolacije temeljene na metodi najmanjih kvadrata. Istaknimo konačno da gubitak značajnih znamenaka nema monotoni rast jer je učinak međusobnog poništavanja pogrešaka u nekim fazama proračuna prilično izražen pa usporava akumulaciju pogreške. Ipak, globalna je tendencija porasta pogreške prilično jasna.

9 Zaključak

Prema prikazanim je rezultatima očito da (neizbježna) pogreška zaokruživanja *sigurno* smanjuje točnost naših numeričkih proračuna, čak i onda kada je sve napravljeno najbolje što možemo. To pokazuju dobri primjeri kod kojih je ipak zabilježen gubitak od šest značajnih znamenaka. Čim u model unesemo poremećaj, primjerice elemente lošeg oblika, gubitak znamenaka penje se na



Slika 7. Ovisnost teorijske procjene i točnog gubitka značajnih znamenaka o broju nepoznanica

čak 11. Pri tome teorijska procjena prema normi dva dobro prognozira gubitak ako su ulazni podaci zapisani u dvostrukoj preciznosti (što je u suvremenim proračunima uobičajeno).
Međutim, u našem se slučaju radilo o jednostavnom problemu čistog savijanja diskretiziranom sa samo 32 000 nepoznanica. Pitanje je kako se ponašaju puno složeniji, praktični primjeri od 10^5 i više nepoznanica, složenog oblika i opterećenja, nejasnih rubnih uvjeta, modelirani nepravilnom mrežom različitih tipova elemenata s mogućim izraženim promjenama krutosti! Što ako je model k tome izrazito nelinearan pa broj numeričkih operacija značajno raste zbog iteracijskog rješavanja sustava? Iako, zbog skromnih mogućnosti računala, nismo dali odgovore na ova pitanja, jednostavnim smo primjerima dali barem naslutiti da problem nije bezazlen. Smatramo ipak da će s napretkom računala i algoritama za proračune cjelobrojnom aritmetikom biti moguće, primjenom ovoga pristupa, realno ocijeniti točnost numeričkog proračuna većeg broja praktičnih inženjerskih zadaća.

LITERATURA

[1] Mihanović, A.; Marović, P.; Dvornik, J.: *Nelinearni proračuni armiranobetonskih konstrukcija*, Split, 1993.

[2] Stoer, J.; Bulirsch, R.: *Introduction to Numerical Analysis*, Springer–Verlag, New York, 1992.

[3] Varga, R. S.: *Matrix Iterative Analysis*, Springer–Verlag, Berlin Heidelberg, 2000.

[4] Bischoff, M.; Wall, W. A.; Bletzinger, K.–U.; Ramm, E.: *Models and Finite Elements for Thin–walled Structures*, Encyclopedia of Computational Mechanics, Volume 2, Solids and Structures, Chapter 3, John Wiley & Sons, Ltd., West Sussex, 2004.

[5] Dvornik, J.; Lazarević, D.: *Manjkavosti proračunskih modela inženjerskih konstrukcija*, Građevinar **57** (2005)4, 227–326.

[6] Demell, J. W.: *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[7] Drmač, Z.; Hari, V.; Marušić, M.; Rogina, M.; Singer, S.; Singer, S.: *Numerička analiza*, elektronički udžbenik, `www.math.hr/znanost/iprojekti/numat`, Zagreb, 2003.

[8] Davis, T. A.: *Algorithm 832: UMFPACK*, an unsymmetric pattern multifrontal method, ACM Transactions on Mathematical Software, vol. 30, no. 2, June 2004, str. 196–199.

[9] Overton, M. L.: *Numerical Computing with IEEE Floating Point Arithmetic*, SIAM, Philadelphia, 2001.

[10] Bathe, K. J.: *Finite Element Procedures*, Prentice Hall, New Jersey, 1996.

[11] Jaguljnjak–Lazarević, A.: *Ocjena točnosti numeričkog proračuna inženjerskih konstrukcija*, doktorski rad, Sveučilište u Zagrebu, Građevinski fakultet, Zagreb, 2008.

[12] Timošenko, S. S.; Gudier, J. N.: *Teorija elastičnosti*, Građevinska knjiga, Beograd, 1962.

[13] Zienkiewicz, O. C.; Taylor, R. L.; Zhu, J.Z.: *The Finite Element Method: Its Basis and Fundamentals*, Sixth edition, Elsevier Butterworth–Heinemann, Oxford, 2006.

[14] Wellin, P.; Gaylord, R.; Kamin, S.: *An introduction to Programming with Mathematica*, Cambridge University Press, Cambridge, 2008.

[15] Jaguljnjak–Lazarević, A.: *Usporedba analitičkih i numeričkih rješenja prostorne rubne zadaće za pravokutne ploče*, magistarski rad, Sveučilište u Zagrebu, Građevinski fakultet, Zagreb, 2005.