

Numerička matematika

2. predavanje

Saša Singer

`singer@math.hr`

`web.math.pmf.unizg.hr/~singer`

PMF – Matematički odsjek, Zagreb

Sadržaj predavanja

- Uvodna priča o greškama:
 - Vrste grešaka (ponavljanje).
 - Analiza pojedinih vrsta grešaka.
 - Greške metode — teorija aproksimacija.
 - Greške u podacima — teorija perturbacija.
 - Uvjetovanost problema.
 - Širenje grešaka u aritmetici — uvjetovanost osnovnih operacija.
 - Približno računanje i perturbacije podataka.
 - Mjerenje grešaka — razne norme.
 - Stabilnost algoritma.
 - Primjeri stabilnih i nestabilnih algoritama.
 - Primjeri grešaka zaokruživanja.

Informacije

Moja web stranica za **Numeričku matematiku** je

http://web.math.hr/~singer/num_mat/

Tamo su kompletna **predavanja** od prošlih **5** godina, a stizat će i **nova** (kako nastaju).

Skraćena verzija **skripte** — **1. dio** (prvih **7** tjedana):

http://web.math.hr/~singer/num_mat/num_mat1.pdf

Skraćena verzija **skripte** — **2. dio** (drugih **6** tjedana):

http://web.math.hr/~singer/num_mat/num_mat2.pdf

Informacije — demonstratori

Kolegij “Numerička matematika” ima dva demonstratora:

- Mario Berljafa

- termini: utorak, 9–10 i petak, 12–13.

- e-mail: mberljaf@student.math.hr

- Marin Bužančić

- termin: utorak, 16–18.

- e-mail: buzancic@student.math.hr

Demosi lijepo mole da im se najavite mailom bar dan ranije!

- Sastanak za demonstrature je pred oglasnom pločom (bar zasad).

Informacije — kolokviji

Numerička matematika je u kolokvijskom razredu **A2**.

Službeni termini svih **kolokvija** su:

- **Prvi** kolokvij: **utorak, 2. 4. 2013.**, u **12** sati.
- **Drugi** kolokvij: **ponedjeljak, 27. 5. 2013.**, u **12** sati.
- **Popravni** kolokvij: **ponedjeljak, 10. 6. 2013.**, u **12** sati.

Uputa: “**izbjegnite**” popravni — obavite to **ranije!**

Greške i uvjetovanost

Greške — ponavljanje

Pri **numeričkom** rješavanju nekog problema javljaju se različiti tipovi **grešaka**:

- greške **modela** — svođenje **realnog** problema na neki “**matematički**” problem,
- greške u **ulaznim podacima** (mjerjenja i sl.),
- greške **numeričkih metoda** za rješavanje “**matematičkog**” problema,
- greške “**približnog**” **računanja** — obično su to
 - greške **zaokruživanja** u **aritmetici računala**.

Greške **modela** su “**izvan**” dosega **numeričke matematike**.

- Spadaju u fiziku, kemiju, biologiju, tehniku, ekonomiju, ...

Greške (nastavak)

Sljedeće tri kategorije (**podaci**, **metoda**, **računanje**) su vezane za “matematički” problem, i

- spadaju u domenu **numeričke matematike**!

O njima nešto “moramo reći”.

Skica **numeričkog** rješavanja nekog problema slič **algoritmu**:



Posebno, ako dozvolimo da, umjesto riječi “**algoritam**”,

- piše i riječ “**metoda**”.

Zamislite da pojam “**algoritam**” uključuje

- metodu** i stvarno **računanje** rezultata!

Greške (nastavak)



Sve tri vrste grešaka — podaci, metoda, računanje,
• rezultiraju nekom greškom u konačnom rezultatu!

Ta greška nas “zanima”.

Uočite da greške u ulaznim podacima možemo gledati

- neovisno o metodi za rješenje problema,
- i tako dolazimo do pojma uvjetovanosti problema.

Za razliku od toga, greške metode i računanja, naravno,

- ovise o metodi, odnosno, algoritmu za rješenje problema.

Analiza grešaka

Greška metode

Gruba podjela **numeričkih metoda** — prema greškama:

Egzaktne metode

- ☛ daju **egzaktno** rješenje u **konačnom** broju “koraka”, odnosno, računskih operacija.

Primjer:

- ☛ Gaussove eliminacije ili LR faktorizacija za linearne sustave.

Greška takvih metoda je **nula**, uz **egzaktno** računanje.

Približne ili **neegzaktne** metode

- ☛ daju **približno** rješenje problema, u **konačnom** broju “koraka” (računskih operacija).

Greška metode — približne metode

Mogu biti **egzaktne** — na nekom limesu!

Primjeri:

- zamjena kompliciranog modela jednostavnijim,
- greške diskretizacije (numerička integracija),
- greške odbacivanja/rezanja, konačne iteracije (rješavanje nelinearnih jednažbi)

Analiza ovih grešaka spada u **teoriju aproksimacija**.

Pošteno, to je **standardni** predmet proučavanja **numeričke matematike**, u **širem** smislu,

- numerička analiza, funkcionalna analiza, itd.

Time se bavimo **veći** dio kolegija!

Greške u podacima

Ključno svojstvo **problema** je

- ovisnost **rješenja** o **greškama** ili **perturbacijama** ulaznih podataka.

To spada u **teoriju perturbacije**.

Da bi problem uopće **imao smisla**, očekujemo

- neku vrstu **neprekidnosti** rješenja,
- ili barem **ograničenu** osjetljivost rješenja na perturbacije.

Inače imamo “**loše**” postavljen problem!

Osjetljivost se obično mjeri tzv. **brojem uvjetovanosti** problema (engl. “condition number”). Može ih biti i **više**.

Uvjetovanost problema

Neformalno rečeno, **uvjetovanost problema** mjeri

- **osjetljivost** problema na **greške** u **podacima**.

Osnovno svojstvo **uvjetovanosti**:

- **Ne ovisi** o konkretnoj **numeričkoj metodi** za rješenje problema, već samo o **problemu**.

Svrha **uvjetovanosti** = daje odgovor na pitanje:

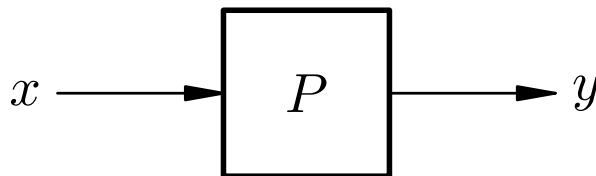
- Koju **točnost rezultata** možemo očekivati
- pri **točnom računanju**
- s (malo) **pomaknutim** — **netočnim podacima**?

Model problema

Matematički model **problema**, zovimo ga P :

- za zadani **ulaz** — podatak $x \in \mathcal{X}$,
- dobivamo **izlaz** — rezultat $y \in \mathcal{Y}$.

Slikica modela je



Problem P interpretiramo kao računanje vrijednosti **funkcije**

$$f : \mathcal{X} \rightarrow \mathcal{Y},$$

gdje su \mathcal{X} i \mathcal{Y} odgovarajući matematički **objekti**. Na primjer, **vektorski** prostori, a vrlo često su i **normirani** prostori (treba nam mjera za grešku).

Uvjetovanost problema (nastavak)

Ideja **uvjetovanosti**:

greška u rezultatu \approx **uvjetovanost** · greška u podacima

Ovisi o **obje** vrijednosti: točnoj x i približnoj \hat{x} .

Napomene:

- Obično nas **uvjetovanost** posebno zanima za **male** perturbacije (greške, smetnje) podataka.
- Ako je f dovoljno glatka funkcija, možemo koristiti **Taylorov** razvoj u okolini **točnog** ulaznog podatka x
- i dobiti procjenu **uvjetovanosti** preko **derivacija**!

Više detalja malo kasnije, kad “sredimo” greške računanja!

Primjeri problema (nastavak)

Primjer 1. Računanje **sume** dva **realna** broja $x_1, x_2 \in \mathbb{R}$.
Tada je

$$f(x_1, x_2) = x_1 + x_2,$$

s tim da je $\mathcal{X} = \mathbb{R}^2$ i $\mathcal{Y} = \mathbb{R}$.

Primjer 2. Računanje **produkta** dva **realna** broja $x_1, x_2 \in \mathbb{R}$.
Tada je

$$f(x_1, x_2) = x_1 \cdot x_2,$$

s tim da je opet $\mathcal{X} = \mathbb{R}^2$ i $\mathcal{Y} = \mathbb{R}$.

Analogno, možemo promatrati i ostale (binarne) aritmetičke operacije — **oduzimanje** i **dijeljenje**.

Primjeri problema (nastavak)

Primjer 3. Računanje **sjecišta** pravaca

$$P_1 = \{ (y_1, y_2) \in \mathbb{R}^2 \mid a_{11}y_1 + a_{12}y_2 = x_1 \},$$

$$P_2 = \{ (y_1, y_2) \in \mathbb{R}^2 \mid a_{21}y_1 + a_{22}y_2 = x_2 \}.$$

Smatramo da su koeficijenti a_{ij} i x_i , za $i, j = 1, 2$, **ulazni** podaci, tj. imamo **6** ulaznih podataka.

Ovaj problem pišemo u matričnom zapisu, kao **linearni** sustav od **dvije** jednačbe, oblika $Ay = x$, gdje je

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \in \mathbb{R}^{2 \times 2}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \in \mathbb{R}^2.$$

Traženo **sjecište** y je **rješenje** linearnog sustava $Ay = x$.

Primjeri problema (nastavak)

Pretpostavimo da je matrica A sustava **regularna**, tj. da vrijedi $\det A \neq 0$. Onda je $y = A^{-1}x$, a funkcija f ima “opći” oblik

$$f(A, x) = A^{-1}x,$$

s tim da je $\mathcal{X} = \mathbb{R}^{2 \times 2} \times \mathbb{R}^2 = \mathbb{R}^6$ i $\mathcal{Y} = \mathbb{R}^2$.

Ako uzmemo da je A **fiksna** matrica, a samo x **varira**, onda je

$$f_A(x) = A^{-1}x, \quad \text{uz} \quad \mathcal{X} = \mathcal{Y} = \mathbb{R}^2.$$

Analogno, ako je x **fiksna**, a samo matrica A **varira**, onda je

$$f_x(A) = A^{-1}x, \quad \text{uz} \quad \mathcal{X} = \mathbb{R}^{2 \times 2} = \mathbb{R}^4, \quad \mathcal{Y} = \mathbb{R}^2.$$

Ako A **nije** regularna, onda imamo **loše postavljen** problem!

Širenje grešaka u aritmetici (uvjetovanost operacija)

Širenje grešaka u aritmetici

Za analizu širenja grešaka u aritmetici, treba pogledati

- što se događa s greškama u rezultatu,
- kad imamo greške u operandima.

Prvo u egzaktnoj aritmetici, a onda i u aritmetici računala.

Pretpostavimo onda da su polazni podaci (ili operandi) x i y malo perturbirani, s pripadnim relativnim greškama ε_x i ε_y .

Koje su operacije opasne (ako takvih ima), ako nam je aritmetika egzaktna, a operandi su $x(1 + \varepsilon_x)$ i $y(1 + \varepsilon_y)$?

Treba ocijeniti relativnu grešku ε_o rezultata operacije \circ

$$(x \circ y)(1 + \varepsilon_o) := [x(1 + \varepsilon_x)] \circ [y(1 + \varepsilon_y)].$$

Širenje grešaka u aritmetici (nastavak)

Naravno, za početak, moramo nešto **pretpostaviti** o ε_x i ε_y .

Što smatramo **malom** relativnom perturbacijom?

- Svakako **mora** biti $|\varepsilon_x|, |\varepsilon_y| < 1$, inače perturbacijom **gubimo predznak** operanda.

Međutim, to nije dovoljno za neki razuman rezultat.

- Stvarno **očekujemo** $|\varepsilon_x|, |\varepsilon_y| \leq c \ll 1$, tako da imamo barem **nekoliko točnih znamenki** u perturbiranim operandima. Na pr., $c = 10^{-1}$ (jedna točna znamenka).
- **Idealno**, u računalu je $|\varepsilon_x|, |\varepsilon_y| \leq u$, tj. kao da smo oba operanda **samo spremili u memoriju** računala (jedna greška zaokruživanja).

Širenje grešaka kod množenja

Množenje je bezopasno (benigno), jer vrijedi

$$\begin{aligned}(x * y) (1 + \varepsilon_*) &:= [x (1 + \varepsilon_x)] * [y (1 + \varepsilon_y)] \\ &= xy (1 + \varepsilon_x + \varepsilon_y + \varepsilon_x \varepsilon_y),\end{aligned}$$

kad stvar napišemo bez nepotrebnih zagrada i *. Onda je

$$\varepsilon_* = \varepsilon_x + \varepsilon_y + \varepsilon_x \varepsilon_y \approx \varepsilon_x + \varepsilon_y,$$

ako su $|\varepsilon_x|$ i $|\varepsilon_y|$ dovoljno mali da $\varepsilon_x \varepsilon_y$ možemo zanemariti.

Dakle, relativna greška se samo zbraja.

U idealnom slučaju $|\varepsilon_x|, |\varepsilon_y| \leq u$, dobivamo približnu ocjenu relativne greške $|\varepsilon_*| \leq 2u$ (do na u^2), ili, na pr., $|\varepsilon_*| \leq 2.01u$.

Širenje grešaka kod dijeljenja

Dijeljenje je, također, bezopasno (benigno), samo je zaključak malo dulji. Na početku je

$$(x / y) (1 + \varepsilon_{/}) := [x (1 + \varepsilon_x)] / [y (1 + \varepsilon_y)] = \frac{x (1 + \varepsilon_x)}{y (1 + \varepsilon_y)}.$$

Ako su $|\varepsilon_x|$ i $|\varepsilon_y|$ dovoljno mali da sve možemo linearizirati (tj. zanemariti “kvadratne” i više potencije epsilon), onda je

$$\frac{1}{1 + \varepsilon_y} = 1 - \varepsilon_y + \sum_{n=2}^{\infty} (-1)^n \varepsilon_y^n \approx 1 - \varepsilon_y$$

i

$$(1 + \varepsilon_x) (1 - \varepsilon_y) = 1 + \varepsilon_x - \varepsilon_y - \varepsilon_x \varepsilon_y \approx 1 + \varepsilon_x - \varepsilon_y.$$

Širenje grešaka kod dijeljenja (nastavak)

Kad to uvrstimo u prvi izraz, dobivamo

$$(x / y) (1 + \varepsilon_{/}) \approx \frac{x}{y} (1 + \varepsilon_x) (1 - \varepsilon_y) \approx \frac{x}{y} (1 + \varepsilon_x - \varepsilon_y).$$

Za relativnu grešku (približno) vrijedi

$$\varepsilon_{/} \approx \varepsilon_x - \varepsilon_y, \quad |\varepsilon_{/}| \approx |\varepsilon_x| + |\varepsilon_y|.$$

Dakle, relativne greške se oduzimaju, a ocjene zbrajaju.

U idealnom slučaju $|\varepsilon_x|, |\varepsilon_y| \leq u$, opet dobivamo približnu ocjenu relativne greške $|\varepsilon_{/}| \leq 2u$.

Vidimo da su i množenje i dijeljenje bezopasne operacije za širenje grešaka zaokruživanja.

Širenje grešaka kod zbrajanja i oduzimanja

Zbrajanje i oduzimanje. Ovdje rezultat ključno ovisi o predznacima od x i y .

Sasvim općenito, neka su x i y proizvoljnih predznaka. Za zbrajanje i oduzimanje (oznaka \pm) vrijedi

$$(x \pm y) (1 + \varepsilon_{\pm}) := [x (1 + \varepsilon_x)] \pm [y (1 + \varepsilon_y)].$$

Pogledajmo prvo trivijalne slučajeve. Ako je egzaktan rezultat $x \pm y = 0$, onda imamo dvije mogućnosti.

- Ako je $x (1 + \varepsilon_x) \pm y (1 + \varepsilon_y) = 0$, relativna greška ε_{\pm} može biti koji broj (nije određena), a prirodno je uzeti $\varepsilon_{\pm} = 0$.
- U protivnom, za $x (1 + \varepsilon_x) \pm y (1 + \varepsilon_y) \neq 0$, gornja jednakost je nemoguća, pa stavljamo $\varepsilon_{\pm} = \pm\infty$.

Širenje grešaka kod zbrajanja i oduzimanja

Pretpostavimo nadalje da je $x \pm y \neq 0$. Onda je

$$\begin{aligned}(x \pm y)(1 + \varepsilon_{\pm}) &= x(1 + \varepsilon_x) \pm y(1 + \varepsilon_y) \\ &= (x \pm y) + (x\varepsilon_x \pm y\varepsilon_y) \\ &= (x \pm y) \left(1 + \frac{x\varepsilon_x \pm y\varepsilon_y}{x \pm y} \right).\end{aligned}$$

Relativnu grešku ε_{\pm} možemo napisati u obliku **linearne kombinacije** polaznih grešaka ε_x i ε_y

$$\varepsilon_{\pm} = \frac{x\varepsilon_x \pm y\varepsilon_y}{x \pm y} = \frac{x}{x \pm y} \varepsilon_x \pm \frac{y}{x \pm y} \varepsilon_y.$$

Širenje grešaka kod zbrajanja i oduzimanja

Naravno, za nastavak rasprave **ključno** je pitanje

• koliko su **veliki faktori** uz polazne greške, tj. da li “**prigušuju**” ili “**napuhavaju**” greške.

Da ne bismo stalno pisali hrpu oznaka \pm (nepregledno), pogledajmo što se zbiva kad

• x i y imaju **isti** predznak, a

• **posebno** gledamo operacije $+$ i $-$.

Ako su x i y **različitih** predznaka, zamijenimo operaciju u suprotnu ($+$ \mapsto $-$, $-$ \mapsto $+$), pa će vrijediti isti zaključci.

Nadalje, zbrajamo i oduzimamo brojeve **istih** predznaka.

Širenje grešaka kod zbrajanja

Zbrajanje brojeva istog predznaka je bezopasno (benigno). To izlazi ovako.

Zbog istih predznaka od x i y , vrijedi $|x|, |y| \leq |x + y|$, pa je

$$\left| \frac{x}{x + y} \right|, \left| \frac{y}{x + y} \right| \leq 1.$$

To vrijedi i kad je $x = 0$ ili $y = 0$. Odavde odmah slijedi

$$|\varepsilon_+| \leq |\varepsilon_x| + |\varepsilon_y|.$$

Dakle, relativna greška se, u najgorem slučaju, zbraja.

U idealnom slučaju $|\varepsilon_x|, |\varepsilon_y| \leq u$, opet dobivamo ocjenu relativne greške $|\varepsilon_+| \leq 2u$.

Širenje grešaka kod zbrajanja (nastavak)

Uz malo truda, dobivamo i **bolju** ocjenu. Prvo uočimo da za faktore vrijedi

$$\left| \frac{x}{x+y} \right| + \left| \frac{y}{x+y} \right| = 1,$$

i još iskoristimo $|\varepsilon_x|, |\varepsilon_y| \leq \max\{|\varepsilon_x|, |\varepsilon_y|\}$. Onda je

$$\begin{aligned} |\varepsilon_+| &\leq \left| \frac{x}{x+y} \right| |\varepsilon_x| + \left| \frac{y}{x+y} \right| |\varepsilon_y| \\ &\leq \left(\left| \frac{x}{x+y} \right| + \left| \frac{y}{x+y} \right| \right) \max\{|\varepsilon_x|, |\varepsilon_y|\} \\ &= \max\{|\varepsilon_x|, |\varepsilon_y|\}. \end{aligned}$$

Širenje grešaka kod zbrajanja (nastavak)

Dakle, relativna greška zbrajanja je, u najgorem slučaju,

• **maksimum** polaznih grešaka (ne treba ih zbrajati).

U idealnom slučaju $|\varepsilon_x|, |\varepsilon_y| \leq u$, sada dobivamo ocjenu relativne greške $|\varepsilon_+| \leq u$. Bolje ne može!

Naravno, isto vrijedi i za **oduzimanje** brojeva **različitih** predznaka. I to je **bezopasno**.

Širenje grešaka kod oduzimanja

Oduzimanje brojeva istog predznaka može biti opasno, čak katastrofalno loše.

● Točnije, ne mora uvijek biti opasno, ali može!

Zašto i kada je opasno?

Zbog različitih predznaka od x i y , uz $x \neq 0$ i $y \neq 0$, sigurno vrijedi

$$|x - y| < \max\{|x|, |y|\},$$

pa je barem jedan od faktora veći od 1, tj.

$$\max\left\{\left|\frac{x}{x-y}\right|, \left|\frac{y}{x-y}\right|\right\} > 1.$$

Širenje grešaka kod oduzimanja (nastavak)

Odavde odmah slijedi da u ocjeni relativne greške

$$|\varepsilon_-| \leq \left| \frac{x}{x-y} \right| |\varepsilon_x| + \left| \frac{y}{x-y} \right| |\varepsilon_y|$$

na barem **jednom** mjestu imamo **rast** greške, a to se može dogoditi i na **oba** mjesta.

Kad je to **zaista opasno**? Ako je $|x-y| \ll |x|, |y|$, ovi faktori

$$\left| \frac{x}{x-y} \right|, \quad \left| \frac{y}{x-y} \right|,$$

mogu biti **proizvoljno veliki**, pa i relativna greška $|\varepsilon_-|$ rezultata može biti **proizvoljno velika**!

Opasno oduzimanje ili kraćenje

Opasna situacija $|x - y| \ll |x|, |y|$ znači da je

- rezultat oduzimanja brojeva istog predznaka =
- broj koji je po apsolutnoj vrijednosti mnogo manji od polaznih podataka (oba operanda),

a to znači da operandi x i y moraju biti bliski, tako da dolazi do kraćenja. Zato se ovaj fenomen obično zove

Opasno ili katastrofalno kraćenje.

Dosad smo govorili da relativna greška u tom slučaju može biti velika, ali da li se to zaista događa?

- Naime, ovdje je ipak riječ o ocjeni greške, pa se možda događa da je ocjena vrlo loša, a prava greška ipak mala!

Nažalost, nije tako! To se itekako događa u praksi!

Primjer:
“Katastrofalno” kraćenje

Primjer katastrofalnog kraćenja

Zakruživanjem ulaznih podataka dolazi do male relativne greške. Kako ona može utjecati na konačan rezultat?

Primjer. Uzmimo realnu aritmetiku “računala” u bazi 10. Za mantisu (značajni dio broja) imamo $p = 4$ dekadске znamenke, a za eksponent 2 znamenke (što nije bitno). Neka je

$$\begin{aligned}x &= 8.8866 = 8.8866 \times 10^0, \\y &= 8.8844 = 8.8844 \times 10^0.\end{aligned}$$

Umjesto brojeva x i y , koji nisu prikazivi, u “memoriju” spremamo brojeve $fl(x)$ i $fl(y)$, pravilno zaokružene na $p = 4$ znamenke

$$\begin{aligned}fl(x) &= 8.887 \times 10^0, \\fl(y) &= 8.884 \times 10^0.\end{aligned}$$

Primjer katastrofalnog kraćenja (nastavak)

Ovim zaokruživanjima napravili smo **malu** relativnu grešku u x i y (ovdje je $u = \frac{1}{2} b^{-p} = 5 \times 10^{-5}$).

Razliku $fl(x) - fl(y)$ računamo tako da **izjednačimo eksponente** (što već jesu), **oduzmemo** značajne dijelove (mantise), pa **normaliziramo**

$$\begin{aligned} fl(x) - fl(y) &= 8.887 \times 10^0 - 8.884 \times 10^0 \\ &= 0.003 \times 10^0 = 3.??? \times 10^{-3}. \end{aligned}$$

Kod normalizacije, zbog pomaka “**ulijevo**”, pojavljuju se

● **?** = znamenke koje više **ne možemo** restaurirati (ta informacija se **izgubila**).

Što sad?

Primjer katastrofalnog kraćenja (nastavak)

Računalo radi **isto** što bismo i mi napravili:

👉 na ta mjesta ? upisuje 0.

Razlog: da rezultat bude **točan**, ako su **polazni** brojevi **točni**. Dakle, ovo oduzimanje je **egzaktno** i u aritmetici računala.

Konačni **izračunati** rezultat je $fl(x) - fl(y) = 3.000 \times 10^{-3}$.

Pravi rezultat je

$$\begin{aligned}x - y &= 8.8866 \times 10^0 - 8.8844 \times 10^0 \\ &= 0.0022 \times 10^0 = 2.2 \times 10^{-3}.\end{aligned}$$

Već **prva** značajna znamenka u $fl(x) - fl(y)$ je **pogrešna**, a relativna greška je **ogromna**! Uočite da je ta znamenka (**3**), ujedno, i **jedina** koja nam je ostala — sve ostalo se **skratilo**!

Primjer katastrofalnog kraćenja (nastavak)

Prava **katastrofa** se događa ako $3.??? \times 10^{-3}$ uđe u naredna zbrajanja (oduzimanja), a onda se **skrati** i ta **trojka!**

Uočite da je **oduzimanje** $fl(x) - fl(y)$ bilo **egzaktno** i u aritmetici našeg “**računala**”, ali **rezultat je**, svejedno, **pogrešan**.

Krivac, očito, **nije oduzimanje** (kad je egzaktno).

- Uzrok su **polazne greške** u operandima.

Ako njih **nema**, tj. ako su polazni operandi **egzaktni**,

- i dalje, naravno, dolazi do **kraćenja**,

- ali je **rezultat** (uglavnom, a po IEEE standardu **sigurno**) **egzaktan**,

pa se ovo kraćenje onda zove **benigno kraćenje**.

Relativna uvjetovanost osnovnih operacija — sažetak

Oznake i pretpostavke na ulaz

Egzaktne ulazne vrijednosti:

- Neka su x , y skalari (realni ili kompleksni brojevi) takvi da je $x \neq 0$ i $y \neq 0$.

Približne ulazne vrijednosti:

- Neka su \hat{x} , \hat{y} njihove **aproksimacije** s **relativnim** greškama

$$\varepsilon_x := \frac{\hat{x} - x}{x}, \quad \varepsilon_y := \frac{\hat{y} - y}{y}.$$

Dodatno, neka je

$$\varepsilon := \max\{|\varepsilon_x|, |\varepsilon_y|\}$$

maksimum relativnih grešaka aproksimacija na **ulazu** u aritmetičku operaciju.

Oznake i pretpostavke na izlaz

Egzaktni rezultat operacije \circ je $x \circ y$. Uz pretpostavku da je $x \circ y \neq 0$, relativna greška približnog izlaza $\hat{x} \circ \hat{y}$ je

$$\varepsilon_{\circ} := \frac{(\hat{x} \circ \hat{y}) - (x \circ y)}{x \circ y}.$$

Ovu grešku ε_{\circ} želimo ocijeniti u obliku

$$\left\{ \begin{array}{l} \text{relativna greška} \\ \text{na izlazu} \end{array} \right\} \leq \text{uvjetovanost} \cdot \left\{ \begin{array}{l} \text{relativna greška} \\ \text{na ulazu} \end{array} \right\},$$

uz neke razumne pretpostavke na ulazne veličine i grešku ε .
Skraćeni zapis je

$$|\varepsilon_{\circ}| \leq \kappa_{\circ} \cdot \varepsilon,$$

gdje je

 $\kappa_{\circ} =$ relativni broj uvjetovanosti za operaciju \circ .

Relativna uvjetovanost zbrajanja i oduzimanja

Pretpostavimo za egzaktni rezultat vrijedi $x \pm y \neq 0$. Onda je

$$|\varepsilon_{\pm}| \leq \kappa_{\pm} \cdot \varepsilon, \quad \kappa_{\pm} = \frac{|x| + |y|}{|x \pm y|}.$$

Za zbrajanje brojeva istog predznaka (oznaka $+$) vrijedi

$$\kappa_{+} = \frac{|x| + |y|}{|x + y|} = 1.$$

Za zbrajanje brojeva različitog predznaka (oznaka $-$) vrijedi

$$\kappa_{-} = \frac{|x| + |y|}{|x - y|} > 1.$$

Ako je $\kappa_{-} \gg 1$, oduzimanje je loše uvjetovano — katastrofalno kraćenje!

Relativna uvjetovanost množenja

Pretpostavimo da je $\hat{x} \neq 0$ i $\hat{y} \neq 0$, što osigurava i $\hat{x} * \hat{y} \neq 0$.
Onda je

$$|\varepsilon_*| \leq \kappa_* \cdot \varepsilon, \quad \kappa_* = 2 + \varepsilon.$$

Dodatno, ako još pretpostavimo da je $\varepsilon \leq 1$, onda je

$$\kappa_* \leq 3.$$

Dakle, ako približne vrijednosti imaju malu relativnu grešku, onda je množenje dobro uvjetovano u relativnom smislu.

Pretpostavka $\varepsilon \leq 1$ osigurava da približne vrijednosti nemaju suprotan predznak od pravih (što je razumno).

Relativna uvjetovanost dijeljenja

Pretpostavimo da je $\hat{x} \neq 0$ i $\hat{y} \neq 0$, što osigurava i $\hat{x}/\hat{y} \neq 0$.

Za iole **razumnu** ocjenu, ovdje još moramo **pretpostaviti**

- da približni nazivnik \hat{y} **nije** pretjerano pogrešan, tako da ima **korektan** predznak, tj. da je $\varepsilon_y > -1$.

Uz malo **jaču** pretpostavku $\varepsilon < 1$, izlazi

$$|\varepsilon_{/}| \leq \kappa_{/} \cdot \varepsilon, \quad \kappa_{/} = \frac{2}{1 - \varepsilon}.$$

Dodatno, ako još **pretpostavimo** da je $\varepsilon \leq 1/2$, onda je

$$\kappa_{/} \leq 4.$$

Dakle, ako približne vrijednosti imaju **malu** relativnu grešku, onda je **dijeljenje dobro** uvjetovano u relativnom smislu.

Približno računanje i perturbacije podataka

Interpretacija grešaka zaokruživanja

Kod **približnog** računanja — na pr. u aritmetici računala, imamo greške **zaokruživanja**. One nastaju

- **spremanjem** ulaznih podataka u algoritam,
- kao rezultat **svake** pojedine aritmetičke operacije.

Ključna stvar za **analizu** tih grešaka je

- svođenje na **teoriju perturbacija**, u smislu
- **egzaktnog** računanja s **perturbiranim** polaznim podacima!

Kako to ide? Ilustracija na IEEE standardu.

Greške prikaza i aritmetike

Ako je ulazni podatak $x \in \mathbb{R}$

• unutar raspona brojeva prikazivih u računalu, onda se, umjesto x , sprema zaokruženi prikazivi broj $fl(x)$, tako da vrijedi

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

gdje je

- ε relativna greška napravljena tim zaokruživanjem,
- a u je jedinična greška zaokruživanja.

Imamo malu relativnu grešku, a računalo dalje računa

- s perturbiranim polaznim podatkom $fl(x)$.

Slična stvar vrijedi i za aritmetičke operacije.

Zaokruživanje u aritmetici

Osnovna pretpostavka za realnu aritmetiku u računalu:

- za sve četiri osnovne aritmetičke operacije vrijedi ista ocjena greške zaokruživanja kao i za prikaz brojeva.

Isto vrijedi i za neke matematičke funkcije, poput $\sqrt{\quad}$, ali ne vrijedi za sve funkcije (na pr. za \cos i \sin u okolini nule).

Preciznije: Neka \circ označava bilo koju operaciju $+$, $-$, $*$, $/$. Za prikazive brojeve u dozvoljenom rasponu $x, y \in \mathcal{F}$, takve da je i egzaktni rezultat $x \circ y$ u dozvoljenom rasponu (tj. u \mathcal{F}), vrijedi ocjena relativne greške

$$fl(x \circ y) = (1 + \varepsilon)(x \circ y), \quad |\varepsilon| \leq u.$$

Broj ε ovisi o x , y , operaciji \circ i aritmetici računala.

Zaokruživanje u aritmetici (nastavak)

Ova ocjena je **ekvivalentna** **idealnom** izvođenju aritmetičkih operacija:

- **egzaktno** izračunaj rezultat operacije $x \circ y$,
- **zaokruži** ga, pri spremanju rezultata u memoriju.

To **ne znači** da računalo **zaista** tako i računa. Naime,

- za $+$, $-$, $*$ to bi se još i moglo napraviti (egzaktne rezultati imaju konačan binarni prikaz),
- ali kod **dijeljenja** to sigurno **ne ide** (egzaktan kvocijent može imati beskonačan binarni prikaz).

Dakle, **važno** je samo da dobijemo istu **ocjenu greške** kao u “idealnom” računanju, a **nije važno** kako se stvarno računa!

Širenje grešaka zaokruživanja

Kad imamo puno operacija — nastaje problem:

- greške se šire i
- treba procijeniti grešku u rezultatu.

Kako to napraviti?

Ponovimo, za aritmetiku računala ne vrijedi:

- asocijativnost zbrajanja i množenja,
- distributivnost množenja prema zbrajanju.

Posljedica: poredak izvršavanja operacija je bitan!

Jedino što vrijedi je:

- komutativnost za zbrajanje i množenje.

Širenje grešaka zaokruživanja (nastavak)

Za analizu grešaka zaokruživanja ne možemo koristiti nikakva “normalna” pravila za aritmetičke operacije u računalu, jer ti zakoni naprosto ne vrijede.

Stvarna algebarska struktura je izrazito komplicirana i postoje debele knjige na tu temu.

- Vrijede neka “zamjenska” pravila, ali su neupotrebljiva za analizu iole većih proračuna.

Međutim, analiza pojedinih operacija postaje bitno lakša, ako uočimo da:

- greške zaokruživanja u aritmetici računala možemo interpretirati i kao egzaktne operacije, ali na “malo” pogrešnim podacima!

Širenje grešaka zaokruživanja (nastavak)

Kako? Dovoljno je faktor $(1 + \varepsilon)$ u ocjeni greške

$$fl(x \circ y) = (1 + \varepsilon)(x \circ y), \quad |\varepsilon| \leq u,$$

“zalijepiti” na x i/ili y . To je isto kao da operand(i) ima(ju) neku relativnu grešku na ulazu u operaciju, a operacija \circ je egzaktna. Dakle,

- izračunati (ili “zaokruženi”) rezultat jednak je egzaktnom rezultatu, ali za malo promijenjene (tj. perturbirane) podatke (u relativnom smislu).

Što dobivamo ovom interpretacijom?

- Onda možemo koristiti “normalna” pravila egzaktne aritmetike za analizu grešaka.

Širenje grešaka zaokruživanja (nastavak)

Ne zaboravimo još da ε ovdje ovisi o x , y , i operaciji \circ . Kad takvih operacija ima više, pripadne greške obično označavamo nekim indeksom u ε .

Na primjer, ako je \circ zbrajanje (+), onda je

$$\begin{aligned} fl(x + y) &= (1 + \varepsilon_{x+y}) (x + y) \\ &= [(1 + \varepsilon_{x+y}) x] + [(1 + \varepsilon_{x+y}) y], \end{aligned}$$

uz $|\varepsilon_{x+y}| \leq u$, ako su x , y i $x + y$ u prikazivom rasponu.

Potpuno ista stvar vrijedi i za oduzimanje.

Kod množenja i dijeljenja možemo birati kojem ulaznom podatku ćemo “zalijepiti” faktor $(1 + \varepsilon)$.

Širenje grešaka zaokruživanja (nastavak)

Za množenje možemo pisati

$$\begin{aligned} fl(x * y) &= (1 + \varepsilon_{x*y}) (x * y) \\ &= [(1 + \varepsilon_{x*y}) x] * y = x * [(1 + \varepsilon_{x*y}) y], \end{aligned}$$

a za dijeljenje

$$\begin{aligned} fl(x / y) &= (1 + \varepsilon_{x/y}) (x / y) \\ &= [(1 + \varepsilon_{x/y}) x] / y = x / [y / (1 + \varepsilon_{x/y})]. \end{aligned}$$

Postoje i druge varijante. Na primjer, da svakom operandu “zalijepimo” $\sqrt{1 + \varepsilon}$ (odnosno $1/\sqrt{1 + \varepsilon}$), ali to nije naročito važno. Bitno je samo da je izračunati rezultat egzaktan za malo perturbirane podatke.

Širenje grešaka (bilo kojih)

Zasad **nije vidljivo** koja je točno **korist** od ove interpretacije. Stvar se **bolje** vidi tek kad imamo **više operacija zaredom**.

Međutim, ova ideja s “**malo pogrešnim podacima**” je

- baš ono što nam **treba** za **analizu širenja grešaka**,
- i to bez obzira na uzrok grešaka, čim se sjetimo da
- rezultati **ranijih** operacija
- s nekom **greškom** **ulaze** u **nove** operacije.

Naime, **uzroka** grešaka može biti mnogo, ovisno o tome što računamo. Od grešaka **modela** i **metode**, preko grešaka **mjerenja** (u ulaznim podacima), do grešaka **zaokruživanja**.

Širenje grešaka u aritmetici računala

Dosad smo gledali širenje grešaka u egzaktnoj aritmetici.

U aritmetici računala postupamo na potpuno isti način. Samo treba zgodno iskoristiti onu raniju interpretaciju da je

- izračunati (ili “zaokruženi”) rezultat jednak egzaktnom, ali za malo perturbirane podatke (u relativnom smislu).

A širenje grešaka u egzaktnoj aritmetici znamo.

Ukratko, bez dokaza:

Svaka pojedina aritmetička operacija u računalu samo

- povećava perturbaciju svojih ulaznih podataka za jedan faktor oblika $(1 + \varepsilon)$, uz ocjenu $|\varepsilon| \leq u$,

ovisno o tome kojim operandima “zalijepimo” taj faktor.

Širenje grešaka u aritmetici računala — primjer

Primjer. Računamo zbroj $\hat{x} + \hat{y}$, gdje su \hat{x} i \hat{y} spremljeni u računalu i već **imaju** neku **grešku**, nastalu spremanjem pravih vrijednosti x i y u memoriju računala i eventualnim ranijim računom. Znamo da za izračunati rezultat vrijedi

$$fl(\hat{x} + \hat{y}) = (1 + \varepsilon_+) (\hat{x} + \hat{y}) = (1 + \varepsilon_+) \hat{x} + (1 + \varepsilon_+) \hat{y},$$

uz $|\varepsilon_+| \leq u$, ako su \hat{x} , \hat{y} i $\hat{x} + \hat{y}$ u **prikazivom** rasponu.

No, \hat{x} i \hat{y} već imaju neke **relativne** greške obzirom na prave egzaktnosti x i y

$$\hat{x} = (1 + \varepsilon_x)x, \quad \hat{y} = (1 + \varepsilon_y)y,$$

i to treba uvrstiti u gornju formulu.

Širenje grešaka u aritmetici računala — primjer

Dobivamo da je

$$\begin{aligned} fl(\hat{x} + \hat{y}) &= (1 + \varepsilon_+) (\hat{x} + \hat{y}) \\ &= (1 + \varepsilon_+)(1 + \varepsilon_x) x + (1 + \varepsilon_+)(1 + \varepsilon_y) y. \end{aligned}$$

Drugim riječima, **izračunati** rezultat $fl(\hat{x} + \hat{y})$ se opet može interpretirati kao **egzaktni** rezultat,

• ali na “**malo više**” perturbiranim polaznim podacima

$$\begin{aligned} \tilde{x} &= (1 + \eta_x)x = (1 + \varepsilon_+)(1 + \varepsilon_x)x, \\ \tilde{y} &= (1 + \eta_y)y = (1 + \varepsilon_+)(1 + \varepsilon_y)y. \end{aligned}$$

Relativne greške η_x i η_y su tzv. “greške **unatrag**” ili “**obratne greške**” — jer se odnose na perturbaciju **ulaza** (uz **egzaktno** računanje).

Širenje grešaka u aritmetici računala — primjer

Nažalost, prethodne formule **ne daju** neku informaciju o tome

• koliko je **izračunati** rezultat “daleko” od **pravog** rezultata, a upravo to je ono što nas **stvarno zanima**.

• Ovo je tzv. “greška **unaprijed**”, jer mjeri “perturbaciju” rezultata, tj. **izlaza** (uz **približno** računanje).

Zadatak. Nađite relativnu grešku **unaprijed** η_+ u izračunatom rezultatu

$$fl(\hat{x} + \hat{y}) = (1 + \eta_+)(x + y),$$

u terminima **ulaznih** podataka x i y , te grešaka ε_x , ε_y i ε_+ .

Zaključak. Greška **unaprijed** se **puno teže** računa od grešaka **unatrag**! Zato se obično nalazi “zaobilaznim” putem.

Natuknice o analizi grešaka

Bilo koji **algoritam** gledamo kao **preslikavanje**:

ulaz (domena) \rightarrow izlaz (kodomena).

Naravno, zanima nas

- **greška** u izračunatom **rezultatu** — u kodomeni,
- uz **približno** računanje aritmetikom računala.

Ova greška zove se greška **unaprijed** (engl. forward error).

Nažalost, postupak “**direktne**” analize grešaka je **težak**,

- relativno **rijetko** “ide” i često daje **loše** ocjene greške.

Primjer. Obična **norma** vektora u \mathbb{R}^2 (i još dodaj “scaling”).
(v. **Z. Drmač**, članak u **MFL-u**).

Natuknice o analizi grešaka (nastavak)

U praksi se puno češće koristi tzv. “obratna” analiza grešaka. Osnovna ideja je **ista** kao i za **pojedine** operacije:

- izračunati **rezultati** algoritma mogu se dobiti **egzaktnim** računanjem,
- ali na **perturbiranim ulaznim** podacima — u domeni.

Ova greška u domeni zove se greška **unatrag** ili **obratna** greška (engl. backward error).

Prednost: ocjena tih perturbacija u **domeni** je bitno **lakša**,

- jer se **akumulacija** onih faktora oblika $(1 + \varepsilon)$ **prirodno** radi **unatrag** — od **rezultata** prema **polaznim podacima**.

U protivnom, moramo **znati** grešku za **operande**, a to je greška **unaprijed** za prethodni dio algoritma.

Natuknice o analizi grešaka (nastavak)

Postupak “unatrag” za nalaženje grešaka u izračunatim rezultatima ide u dva koraka.

- Prvo se obratnom analizom naprave ocjene perturbacija polaznih podataka u domeni,
- a zatim se koristi matematička teorija perturbacije, koja daje ocjene grešaka rezultata u kodomeni. Ovaj izvod ide za egzaktni račun, pa vrijede sva normalna pravila.

Tako stižemo do pojmova:

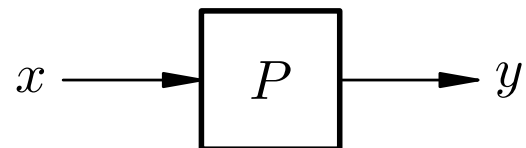
- stabilno i nestabilno računanje ili algoritam = “prigušivač” ili “pojačalo” grešaka.

Slikice (skripta NA, Higham) — su na sljedećoj stranici.

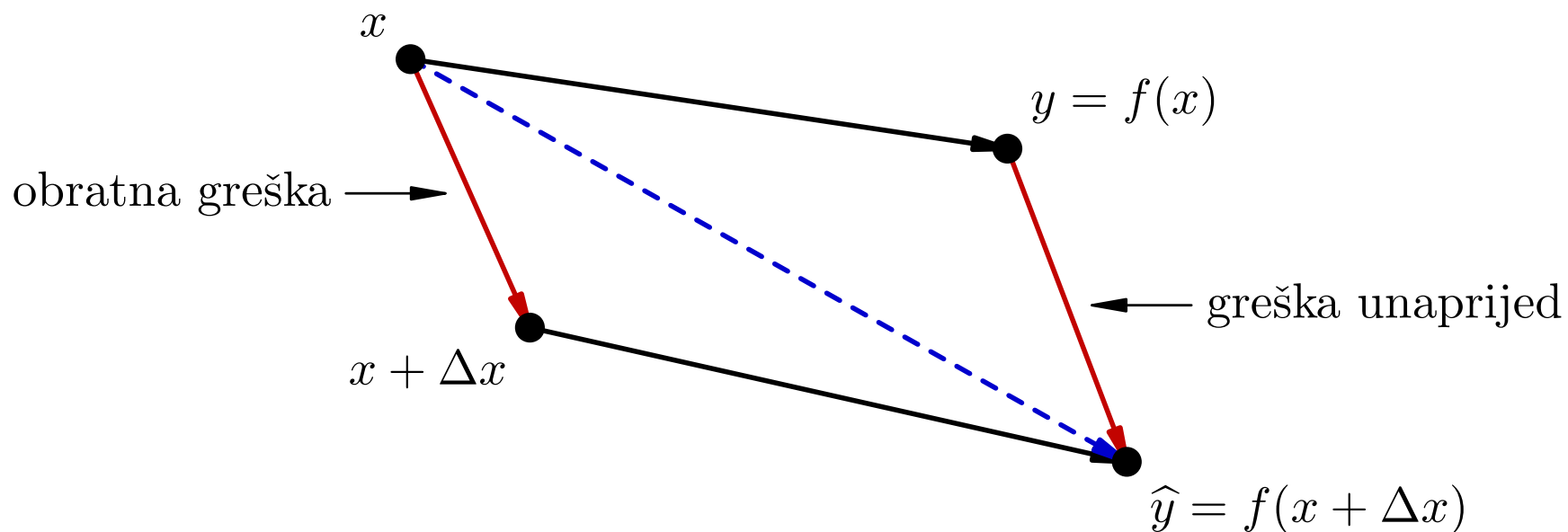
- Primjeri nestabilnosti — uklonjivi i NEuklonjivi.

Greška unaprijed i obratna greška

Uzmimo da **algoritam** “rješava” problem P .



Ako problem P intepretiramo kao **računanje** funkcije f , onda grešku **unaprijed** i **obratnu** grešku možemo prikazati ovako:

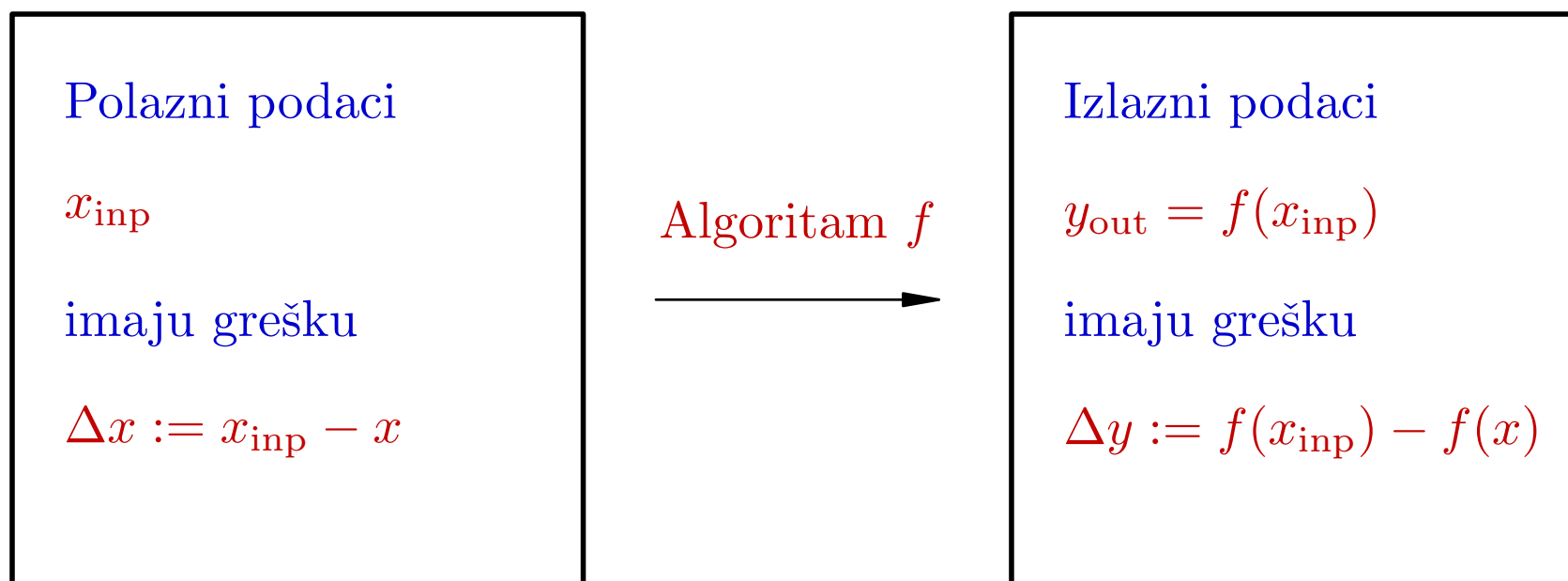


Norme i uvjetovanost

Greška na ulazu — što na izlazu?

Jedan od zadataka **numeričke analize** je

- odrediti **vezu** između greške na **ulazu** i greške na **izlazu**.



Uzimamo da su \mathcal{X} i \mathcal{Y} (barem) **vektorski** prostori.

Kako mjeriti grešku?

Kad x_{inp} i $f(x_{\text{inp}})$ nisu brojevi, nego vektori ili matrice, grešku možemo mjeriti na više načina.

- Po svakoj od komponenata vektora/matrica,
 - što je “vrlo precizno”,
 - međutim, to je malo previše brojeva.
- Kao neku “ukupnu ili najveću” grešku,
 - što je samo jedan broj — pa se lakše nalazi,
 - iako može biti “neprecizno” (sažeta informacija).

Ovo se radi korištenjem vektorskih i/ili matričnih normi.

Prisjetite se: vektorski prostor na kojem je definirana norma zove se normirani prostor.

Vektorske norme

“Vektorska” norma na vektorskom prostoru V (nad poljem F , gdje je $F = \mathbb{R}$ ili $F = \mathbb{C}$) je

• svaka funkcija $\| \cdot \| : V \rightarrow \mathbb{R}$

koja zadovoljava sljedeća svojstva:

1. $\|x\| \geq 0, \quad \forall x \in V,$

a jednakost vrijedi ako i samo ako je $x = 0$,

2. $\|\alpha x\| = |\alpha| \|x\|, \quad \forall \alpha \in F, \quad \forall x \in V,$

3. $\|x + y\| \leq \|x\| + \|y\|, \quad \forall x, y \in V$

(nejednakost poznata pod imenom **nejednakost trokuta**).

Najpoznatije vektorske norme

Kad je vektorski prostor **konačnodimenzionalan**, $V = \mathbb{R}^n$ ili $V = \mathbb{C}^n$, najčešće se koriste sljedeće tri norme:

● **1-norma** ili ℓ_1 norma $\|x\|_1 = \sum_{i=1}^n |x_i|$,

● **2-norma** ili ℓ_2 norma ili **euklidska** norma

$$\|x\|_2 = (x^* x)^{1/2} = \sqrt{\sum_{i=1}^n |x_i|^2},$$

● **∞ -norma** ili ℓ_∞ norma $\|x\|_\infty = \max_{i=1, \dots, n} |x_i|$.

Samo je **2-norma** izvedena iz **skalarnog produkta**.

Norme na prostoru funkcija

Vektorski prostor V **ne mora** biti konačnodimenzionalan.

Na primjer, norme definirane na vektorskom prostoru $C[a, b]$ **neprekidnih funkcija** f na segmentu $[a, b]$, definiraju se slično normama na \mathbb{R}^n (suma \mapsto integral):

● L_1 norma $\|f\|_1 = \int_a^b |f(t)| dt,$

● L_2 norma $\|f\|_2 = \left(\int_a^b |f(t)|^2 dt \right)^{1/2},$

● L_∞ norma $\|f\|_\infty = \max\{ |f(x)| \mid x \in [a, b] \}.$

Ekvivalentnost normi

Može se pokazati da vrijedi sljedeći teorem.

Teorem. Na svakom **konačnodimenzionalnom** vektorskom prostoru V sve su norme **ekvivalentne**, tj. za svake dvije norme $\|\cdot\|_a$ i $\|\cdot\|_b$, postoje konstante c i C , takve da za sve $v \in V$ vrijedi

$$c\|v\|_a \leq \|v\|_b \leq C\|v\|_a.$$

Na primjer,

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty,$$

za sve $x \in \mathbb{R}^n$.

Razlika između teorije i prakse — kad je n **ogroman**.

Matrične norme

Zamijenimo li u definiciji vektorske norme, čisto formalno, vektor x matricom A , dobivamo **matričnu normu**.

Matrična norma je svaka funkcija $\| \cdot \| : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$ koja zadovoljava sljedeća svojstva:

1. $\|A\| \geq 0$, $\forall A \in \mathbb{C}^{m \times n}$,
a jednakost vrijedi ako i samo ako je $A = 0$,
2. $\|\alpha A\| = |\alpha| \|A\|$, $\forall \alpha \in \mathbb{C}$, $\forall A \in \mathbb{C}^{m \times n}$,
3. $\|A + B\| \leq \|A\| + \|B\|$, $\forall A, B \in \mathbb{C}^{m \times n}$.

Tome se često dodaje i zahtjev **konzistentnosti**

4. $\|AB\| \leq \|A\| \|B\|$, kad god je produkt AB definiran.

Matrične norme (nastavak)

Matrične norme nastaju na dva načina.

- Matricu A promatramo kao **vektor** s $m \times n$ elemenata i za taj vektor koristimo odgovarajuću vektorsku normu.

Najpoznatija takva norma odgovara vektorskoj **2-normi** i zove se **euklidska**, **Frobeniusova**, **Hilbert–Schmidtova**, ili **Schurova** norma

$$\|A\|_F = (\operatorname{tr}(A^* A))^{1/2} = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}.$$

- **Operatorske norme:**

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \quad \text{ili} \quad \|A\| = \max_{\|x\|=1} \|Ax\|.$$

Najpoznatije operatorske matične norme

Uvrštavanjem odgovarajućih vektorskih normi, dobivamo

- matična **1-norma**, “maksimalna stupčana norma”

$$\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^m |a_{ij}|,$$

- matična **2-norma**, spektralna norma

$$\|A\|_2 = (\rho(A^* A))^{1/2} = \sigma_{\max}(A),$$

ρ je spektralni radijus, a σ singularna vrijednost matrice,

- matična **∞ -norma**, “maksimalna retčana norma”

$$\|A\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}|.$$

Matrične norme (nastavak)

Svojstva:

- Za matrične norme, također, vrijedi **ekvivalentnost**.
- Matrična **2-norma** se **teško računa** u praksi — uobičajeno se **procjenjuje** korištenjem ostalih normi.
- Za svaku **operatorsku normu** vrijedi

$$\|Ay\| \leq \|A\| \|y\|,$$

za svaki vektor y . To se često koristi kod ocjena. Ova formula direktno izlazi iz definicije operatorske norme.

- **Unitarna** invarijantnost **spektralne** i **Frobeniusove** norme: za bilo koje **unitarne** matrice U (reda m) i V (reda n) vrijedi $\|UAV\|_2 = \|A\|_2$, $\|UAV\|_F = \|A\|_F$.

Apsolutna greška i apsolutna uvjetovanost

Apsolutna, odnosno, relativna uvjetovanost problema mjeri koliko je problem osjetljiv na odgovarajuće promjene polaznih podataka.

- Apsolutna greška: $\|\Delta x\|$, $\|\Delta y\|$, (svaka norma u svom prostoru), gdje je

$$\Delta x = \hat{x} - x, \quad \Delta y = \hat{y} - y.$$

- Apsolutna uvjetovanost:

$$\kappa_{\text{abs}}(x) := \frac{\|\Delta y\|}{\|\Delta x\|}.$$

Za male greške $\|\Delta x\|$, veza s derivacijom $f'(x)$ je očita!

Relativna greška i relativna uvjetovanost

U praksi se češće koristi **relativna** mjera za grešku (na primjer, zbog aritmetike računala).

● Relativna greška:

$$\delta_x := \frac{\|\Delta x\|}{\|x\|}, \quad \delta_y := \frac{\|\Delta y\|}{\|y\|}.$$

● Relativna uvjetovanost:

$$\kappa_{\text{rel}}(x) := \frac{\|\delta_y\|}{\|\delta_x\|}.$$

Problem je **dobro uvjetovan** u relativnom smislu ako je

● κ_{rel} što je moguće **manji**, (barem) za $\delta_x \rightarrow 0$.

Landauov simbol — red veličine

Za zapis “reda veličine” vrijednosti neke funkcije u okolini neke točke koristimo tzv. Landauov simbol.

Definicija. Neka su $g, h : \mathbb{R}^m \rightarrow \mathbb{R}^n$ funkcije, $\|\cdot\|_{\mathbb{R}^m}$ i $\|\cdot\|_{\mathbb{R}^n}$ norme i neka je $x_0 \in \mathbb{R}^m$.

Ako postoje konstante $\delta > 0$ i $C > 0$, takve da za sve x vrijedi

$$\|x - x_0\|_{\mathbb{R}^m} \leq \delta \quad \Longrightarrow \quad \|g(x)\|_{\mathbb{R}^n} \leq C \|h(x)\|_{\mathbb{R}^n},$$

onda kažemo da je

“funkcija g reda veličine \mathcal{O} od h , kad x teži prema x_0 ”

i to pišemo ovako

$$g(x) = \mathcal{O}(h(x)) \quad (x \rightarrow x_0).$$

Landauov simbol (nastavak)

Napomena. Umjesto znaka “=”, **korektno** bi bilo pisati \in , tj.

$$g(x) \in \mathcal{O}(h(x)) \quad (x \rightarrow x_0).$$

Također, često se piše “veliko” O , umjesto “pisanog” \mathcal{O} .

Primjer. Za $m = n = 1$ je

$$\sin x = \mathcal{O}(x) \quad (x \rightarrow a), \quad \text{za sve } a \in \mathbb{R},$$

$$x^2 + 3x = \mathcal{O}(x) \quad (x \rightarrow 0),$$

$$x^2 - x - 6 = \mathcal{O}(x - 3), \quad (x \rightarrow 3).$$

Uvjetovanost i Taylorov teorem

Istražimo **uvjetovanost** problema računanja vrijednosti funkcije $f : \mathbb{R} \rightarrow \mathbb{R}$ u nekoj točki x .

Promatramo ponašanje funkcije f za **male** perturbacije Δx u okolini točke x . Neka je Δy pripadna perturbacija funkcijske vrijednosti $y = f(x)$, tj.

$$f(x + \Delta x) = y + \Delta y.$$

Neka je f još **dva puta neprekidno derivabilna** oko x .

Korištenjem **Taylorovog** polinoma stupnja 1 dobivamo da je

$$\begin{aligned} \Delta y &= f(x + \Delta x) - f(x) \\ &= f'(x) \Delta x + \frac{f''(x + \vartheta \Delta x)}{2!} (\Delta x)^2, \quad \vartheta \in (0, 1). \end{aligned}$$

Uvjetovanost i Taylorov teorem (nastavak)

Za male perturbacije Δx , **apsolutni** oblik ove relacije je

$$\Delta y = f'(x) \Delta x + O((\Delta x)^2),$$

odakle slijedi da je $f'(x)$ ili $|f'(x)|$ **apsolutna** uvjetovanost funkcije f (za male Δx).

Ako je $x \neq 0$ i $y \neq 0$, dijeljenjem s x i y izlazi **relativna** forma

$$\frac{\Delta y}{y} = \frac{x f'(x)}{f(x)} \cdot \frac{\Delta x}{x} + O\left(\left(\frac{\Delta x}{x}\right)^2\right)$$

pa **relativnu** uvjetovanost funkcije f možemo definirati kao

$$\kappa_{\text{rel}}(x) = (\text{cond } f)(x) := \left| \frac{x f'(x)}{f(x)} \right|.$$

Uvjetovanost – primjer

Napomena. U **relativnoj** formi koristimo da je

$$\frac{(\Delta x)^2}{x} = x \left(\frac{\Delta x}{x} \right)^2 = O \left(\left(\frac{\Delta x}{x} \right)^2 \right)$$

za sve dovoljno **male** relativne perturbacije $\Delta x/x$.

Primjer. **Relativna** uvjetovanost funkcije

$$f(x) = \ln x$$

jednaka je

$$(\text{cond } f)(x) = \left| \frac{x f'(x)}{f(x)} \right| = \left| \frac{1}{\ln x} \right|,$$

što je **veliko** za $x \approx 1$.

Pitanje: **Apsolutna** uvjetovanost?

Uvjetovanost višedimenzionalnog problema

Što u više dimenzija?

Kad je $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$, problem postaje složeniji. Uz oznake

$$x = (x_1, x_2, \dots, x_m)^T \in \mathbb{R}^m, \quad y = (y_1, y_2, \dots, y_n)^T \in \mathbb{R}^n,$$

preslikavanje f možemo komponentno zapisati kao

$$y_k = f_k(x_1, x_2, \dots, x_m), \quad k = 1, 2, \dots, n.$$

Ponovno, pretpostavljamo da svaka funkcija f_k ima

- neprekidne parcijalne derivacije po svim komponentnim varijablama x_ℓ u točki x , do barem drugog reda.

Najdetaljniju analizu dobivamo gledajući promjene

- svake komponentne funkcije f_k po svakoj varijabli x_ℓ .

Finija analiza — svaki izlaz po svakom ulazu

Promjena koju uzrokuje mala relativna perturbacija varijable x_ℓ u funkciji f_k ista je kao za funkciju jedne varijable.

Relativna uvjetovanost tog problema je

$$\gamma_{k\ell}(x) := (\text{cond}_{k\ell} f)(x) := \left| \frac{x_\ell}{f_k(x)} \cdot \frac{\partial f_k}{\partial x_\ell} \right|.$$

Ako to napravimo za sve varijable x_ℓ i za svaku funkciju f_k , dobivamo matricu brojeva uvjetovanosti

$$\Gamma(x) = [\gamma_{k\ell}(x)] \in \mathbb{R}_+^{n \times m}.$$

Da bismo iz matrice $\Gamma(x)$ dobili jedan broj, koristimo neku normu i definiramo

$$(\text{cond } f)(x) := \|\Gamma(x)\|.$$

Grublja analiza — po normi

Grublju analizu — s **manje parametara**, dobivamo po ugledu na jednodimenzionalnu, promatranjem

- **apsolutnih i relativnih perturbacija vektora** u smislu **norme**.

Relativnu perturbaciju vektora $x \in \mathbb{R}^m$ “**po normi**” definiramo kao

$$\frac{\|\Delta x\|}{\|x\|}, \quad \Delta x = (\Delta x_1, \Delta x_2, \dots, \Delta x_m)^T,$$

pri čemu je $\|\cdot\|$ bilo koja vektorska norma, a komponente Δx_ℓ vektora perturbacije Δx su **male** u odnosu na komponente x_ℓ vektora x . Tada je i $\|\Delta x\|/\|x\|$ **malo** (obrat **ne vrijedi**).

Isto napravimo i za vektor $y \in \mathbb{R}^n$, tj. gledamo $\|\Delta y\|/\|y\|$.

Taylorov razvoj komponentnih funkcija

Sada možemo pokušati povezati relativnu perturbaciju od y s relativnom perturbacijom od x .

Za male perturbacije Δx , iz početka Taylorovog razvoja funkcije f_k dobivamo

$$\Delta y_k = f_k(x + \Delta x) - f_k(x) \approx \sum_{\ell=1}^m \frac{\partial f_k}{\partial x_\ell} \Delta x_\ell.$$

Ovu relaciju možemo zapisati u vektorsko-matričnom obliku

$$\Delta y \approx \frac{\partial f}{\partial x} \cdot \Delta x,$$

gdje je $\frac{\partial f}{\partial x} = J_f(x)$ Jacobijeva matrica funkcije f u točki x .

Jacobijeva matrica preslikavanja

Preciznije, ta **Jacobijeva** matrica jednaka je:

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

Apsolutne perturbacije po normi

Iz približne jednakosti za **male** perturbacije

$$\Delta y \approx \frac{\partial f}{\partial x} \cdot \Delta x,$$

uzimanjem bilo koje **operatorske** ili **konzistentne** norme izlazi

$$\|\Delta y\| \approx \left\| \frac{\partial f}{\partial x} \cdot \Delta x \right\| \lesssim \left\| \frac{\partial f}{\partial x} \right\| \cdot \|\Delta x\|.$$

Za **operatorske** norme, prethodna nejednakost **oštra**, tj. **postoji** perturbacija Δx za koju se ona **dostiže**.

Oдавde vidimo da **normu Jacobijeve** matrice možemo uzeti kao **apsolutnu** uvjetovanost “**po normi**”.

Uočite: Za $m = n = 1$ dobivamo **isto** kao i ranije!

Relativne perturbacije po normi

Kao i ranije, ako je $x \neq 0$ i $y \neq 0$, dijeljenjem s $\|x\|$ i $\|y\|$ dobivamo da za **relativne perturbacije po normi** vrijedi

$$\frac{\|\Delta y\|}{\|y\|} \lesssim \frac{\|x\|}{\|f(x)\|} \cdot \left\| \frac{\partial f}{\partial x} \right\| \cdot \frac{\|\Delta x\|}{\|x\|}.$$

To opravdava definiciju **relativne** uvjetovanosti “**po normi**” u obliku

$$(\text{cond } f)(x) := \frac{\|x\|}{\|f(x)\|} \cdot \left\| \frac{\partial f}{\partial x} \right\|.$$

Ova uvjetovanost je **mного grublja** nego $\|\Gamma(x)\|$, jer **norma** pokušava “**uništiti**” detalje o komponentama vektora.

Ako su komponente **bitno različitih** redova veličina, samo **najveće** po apsolutnoj vrijednosti igraju **neku ulogu**.

Primjer

Primjer. Ispitajmo **relativnu** uvjetovanost problema računanja vrijednosti funkcije $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, gdje je

$$y_1 = f_1(x_1, x_2) = \frac{1}{x_1} + \frac{1}{x_2}, \quad y_2 = f_2(x_1, x_2) = \frac{1}{x_1} - \frac{1}{x_2}.$$

U **finijoj** analizi, **matrica** uvjetovanosti $\Gamma(x)$ ima elemente

$$\begin{aligned} \gamma_{11} &= \left| \frac{x_2}{x_1 + x_2} \right|, & \gamma_{12} &= \left| \frac{x_1}{x_1 + x_2} \right|, \\ \gamma_{21} &= \left| \frac{x_2}{x_2 - x_1} \right|, & \gamma_{22} &= \left| \frac{x_1}{x_2 - x_1} \right|, \end{aligned}$$

što ukazuje na **lošu uvjetovanost** (engl. ill-conditioning) za $x_1 \approx \pm x_2$, uz uvjet da $|x_1|$ (a onda i $|x_2|$) **nisu mali**.

Primjer (nastavak)

Za “globalni” broj uvjetovanosti $\|\Gamma(x)\|_F$ dobivamo

$$\|\Gamma(x)\|_F = \sqrt{2} \frac{x_1^2 + x_2^2}{|x_1^2 - x_2^2|},$$

što ponovno pokazuje istu **lošu uvjetovanost** za $x_1 \approx \pm x_2$.

Ako uzmemo **relativnu** uvjetovanost po ∞ -normi, onda je

$$(\text{cond } f)(x) = \frac{\max\{|x_1|, |x_2|\} \cdot (x_1^2 + x_2^2)}{\max\{|x_1 + x_2|, |x_2 - x_1|\} \cdot |x_1 x_2|}.$$

Uvrstimo li $x_1 \approx \pm x_2$, dobivamo da je $(\text{cond } f)(x) \approx 1$, što vodi na **pogrešan** zaključak da je problem **dobro uvjetovan** i neosjetljiv na perturbacije za $x_1 \approx \pm x_2$.

Primjer uvjetovanosti problema

Problem — računanje integrala (Gautschi)

Ispitajmo **uvjetovanost** problema računanja integrala

$$I_n = \int_0^1 \frac{t^n}{t+5} dt$$

za zadani prirodni broj $n \in \mathbb{N}$.

U **ovom** obliku, problem je napisan kao preslikavanje iz \mathbb{N} u \mathbb{R} i **ne** “paše” ranijem pojmu **problema**.

- Domena **nije** \mathbb{R} , nego \mathbb{N} (diskretan skup), pa nema smisla govoriti o neprekidnosti, derivabilnosti i sl.

Zato prvo **transformiramo** problem.

Rekurzija za integral

Nađimo **vezu** između I_k i I_{k-1} , s tim da I_0 **znamo** izračunati

$$I_0 = \int_0^1 \frac{1}{t+5} dt = \ln(t+5) \Big|_0^1 = \ln \frac{6}{5}.$$

Za početak, očito vrijedi da je

$$\frac{t}{t+5} = 1 - \frac{5}{t+5},$$

Množenjem obje strane s t^{k-1} dobivamo

$$\frac{t^k}{t+5} = t^{k-1} - 5 \frac{t^{k-1}}{t+5}.$$

Rekurzija za integral (nastavak)

Na kraju, **integracijom** na segmentu $[0, 1]$ izlazi

$$I_k = \int_0^1 t^{k-1} dt - 5I_{k-1} = \frac{1}{k} - 5I_{k-1}, \quad k = 1, 2, \dots, n.$$

Dakle, I_k je **rješenje** (linearne, nehomogene) **diferencijske** **jednadžbe prvog reda**

$$y_k = -5y_{k-1} + \frac{1}{k}, \quad k = 1, 2, \dots,$$

uz **početni** uvjet $y_0 = I_0$.

Ovo gore je **dvočlana rekurzivna relacija** za y_k (pogledajte priču o rekurzivnim relacijama u **Diskretnoj matematici**).

Rekurzija unaprijed — zapis funkcijama

Varijacija početnog uvjeta definira niz funkcija f_k , $y_k = f_k(y_0)$.

Zanima nas relativna uvjetovanost funkcije f_n u točki $y_0 = I_0$, u ovisnosti o $n \in \mathbb{N}$. Razlog:

- I_0 nije egzaktno prikaziv u računalu,
- umjesto I_0 , spremi se aproksimacija \hat{I}_0 ,
- rezultat — neka aproksimacija $\hat{I}_n = f_n(\hat{I}_0)$.

Indukcijom ili supstitucijom unatrag lako se dokaže da vrijedi

$$y_n = f_n(y_0) = (-5)^n y_0 + p_n, \quad p_n = \sum_{k=1}^n \frac{(-5)^{n-k}}{k},$$

gdje je p_n ovisi samo o nehomogenim članovima rekurzije, ali ne i o početnom uvjetu y_0 .

Rekurzija unaprijed — relativna uvjetovanost

Relativna uvjetovanost funkcije f_n u točki y_0 je

$$(\text{cond } f_n)(y_0) = \left| \frac{y_0 f'_n(y_0)}{y_n} \right| = \left| \frac{y_0 (-5)^n}{y_n} \right|.$$

Iz definicije integrala slijedi: I_n monotonno padaju po n , čak

$$\lim_{n \rightarrow \infty} I_n = 0.$$

Zbrajanjima dobivamo sve manje i manje brojeve! U $y_0 = I_0$ je

$$(\text{cond } f_n)(I_0) = \frac{I_0 \cdot 5^n}{I_n} > \frac{I_0 \cdot 5^n}{I_0} = 5^n.$$

Zaključak: f_n je vrlo loše uvjetovana u $y_0 = I_0$, i to tim gore kad n raste.

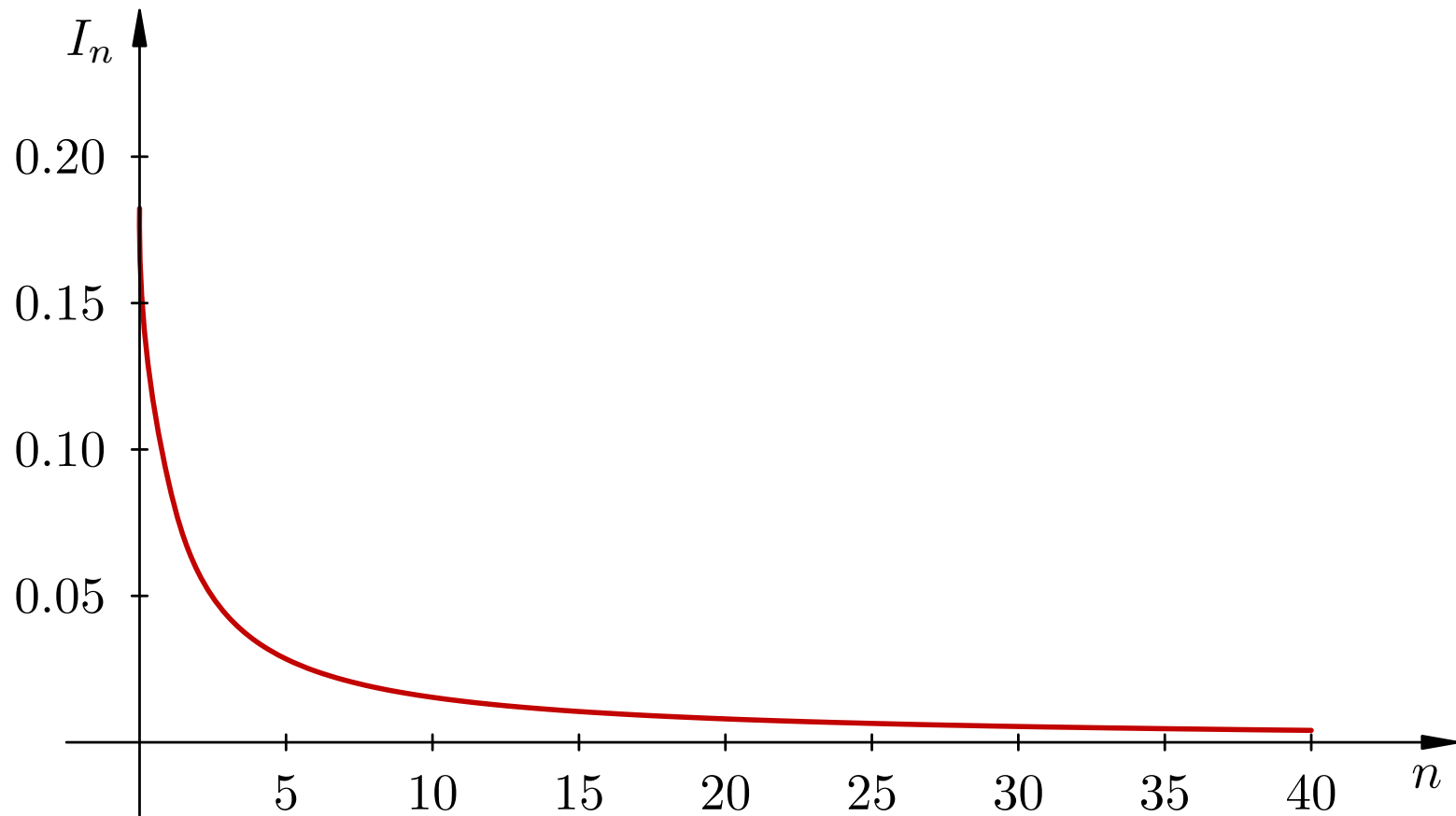
Rekurzija unaprijed — rezultati

Pitanje: Kako se loša uvjetovanost vidi, kad stvarno računamo $f_n(I_0)$?

Slikice!

Pokaži program i rezultate!

Točne vrijednosti integrala I_n



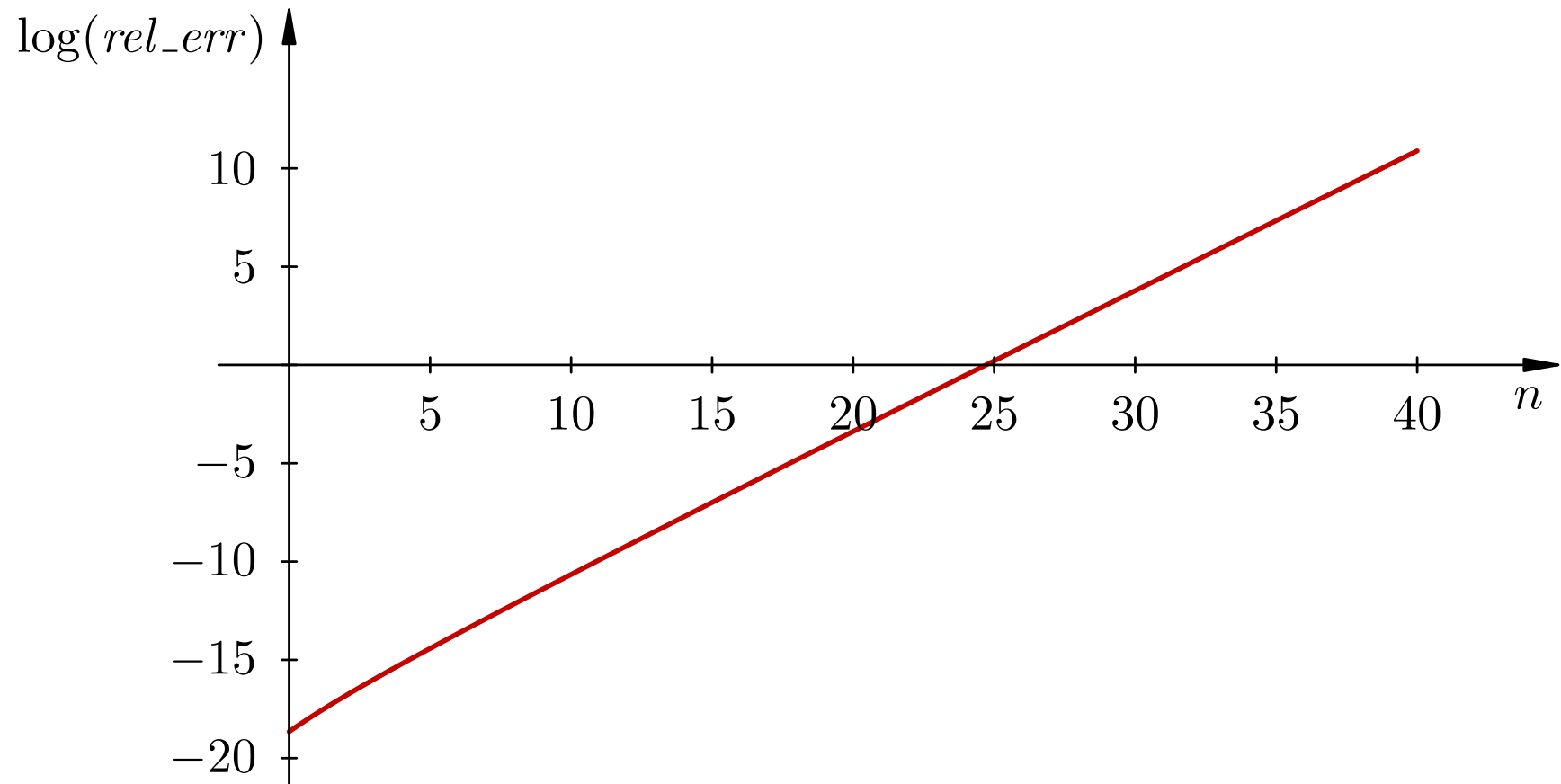
egzaktne/točne vrijednosti integrala I_n

Rekurzija unaprijed — numerički rezultati

Izračunate vrijednosti u tipu `extended` ($u \approx 5.42 \cdot 10^{-20}$).
 Crvene znamenke su pogrešne!

n	\hat{I}_n	I_n	rel. greška
0	1.82321556793954626E-1	1.82321556793954626E-1	2.2E-19
1	8.83922160302268688E-2	8.83922160302268689E-2	-2.0E-18
2	5.80389198488656562E-2	5.80389198488656553E-2	1.5E-17
..
22	7.38035060732479776E-3	7.29738306511145509E-3	1.1E-02
23	6.57650783294122857E-3	6.99134554400794192E-3	-5.9E-02
24	8.78412750196052383E-3	6.70993894662695705E-3	3.1E-01
25	-3.92063750980261915E-3	6.45030526686521474E-3	-1.6E+00
..
39	-6.32992112791892692E+7	4.18374034921478077E-3	-1.5E+10
40	3.16496056420946346E+8	4.08129825392609613E-3	7.8E+10

Rekurzija unaprijed za I_n



(\log_{10}) relativne greške izračunate vrijednosti
integrala I_n rekurzijom unaprijed

Rekurzija unaprijed — komentar rezultata

Izračunata vrijednost \hat{I}_0 ima

- vrlo **malu** relativnu grešku — samo **nekoliko** u .

Međutim, ta mala greška “**eksplodira**” vrlo **brzo**,

- jer se **pojačava** s faktorom **5** u **svakoj** iteraciji.

Isto vrijedi i za **sve** greške zaokruživanja iza toga, samo je **ukupni** faktor pojačanja malo manji (kasnije su nastale).

Stvarni problem i **bitna** razlika od primjera “**sin 24 π** ”:

- Ovdje **nema velikih omjera** brojeva u algoritmu.

Brojevi I_n relativno **sporo** padaju — omjer I_0/I_{40} je ispod **50**.
Po tome, očekivali bismo gubitak točnosti od oko **2** decimale,

- a stvarno imamo **užasno** i još “**nevidljivo**” kraćenje.

Rekurzija unatrag — zapis funkcijama

Može li se loša uvjetovanost izbjeći?

● Može — okretanjem rekurzije, unaprijed \mapsto unatrag!

Treba uzeti neki $\nu > n$ i “silazno” računati

$$y_{k-1} = \frac{1}{5} \left(\frac{1}{k} - y_k \right), \quad k = \nu, \nu - 1, \dots, n + 1.$$

Ovo, u principu, smijemo koristiti i za $n = 0$, tj. računati y_0 .

Problem: Kako izračunati početnu vrijednost y_ν ?

Nova rekurzija definira niz funkcija $g_{n,\nu}$, koje vežu y_n i y_ν , uz $\nu > n$, tj.

$$y_n = g_{n,\nu}(y_\nu).$$

Rekurzija unatrag — relativna uvjetovanost

Relativna uvjetovanost za $g_{n,\nu}$ je

$$(\text{cond } g_{n,\nu})(y_\nu) = \left| \frac{y_\nu (-1/5)^{\nu-n}}{y_n} \right|, \quad \nu > n.$$

Za $y_\nu = I_\nu$ dobivamo da je $y_n = I_n$, a iz monotonosti I_n slijedi

$$(\text{cond } g_{n,\nu})(I_\nu) = \frac{I_\nu}{I_n} \cdot \left(\frac{1}{5}\right)^{\nu-n} < \left(\frac{1}{5}\right)^{\nu-n}, \quad \nu > n,$$

što je ispod 1, tj. relativne greške se prigušuju.

- Prigušenje grešaka ide s faktorom $1/5$ po svakoj iteraciji!
- To vrijedi i za greške zaokruživanja napravljene u ranijim iteracijama (u aritmetici računala).

Rekurzija unatrag — početna vrijednost

Ako je \hat{I}_ν neka aproksimacija za I_ν , onda za **relativne perturbacije** vrijedi

$$\left| \frac{\hat{I}_n - I_n}{I_n} \right| = (\text{cond } g_{n,\nu})(I_\nu) \cdot \left| \frac{\hat{I}_\nu - I_\nu}{I_\nu} \right| < \left(\frac{1}{5} \right)^{\nu-n} \cdot \left| \frac{\hat{I}_\nu - I_\nu}{I_\nu} \right|.$$

Zbog **linearnosti** funkcije $g_{n,\nu}$, ova relacija vrijedi za **bilo kakve** perturbacije, a ne samo za **male**.

- **Početna** vrijednost \hat{I}_ν uopće **ne mora biti blizu** prave I_ν .
- Možemo uzeti $\hat{I}_\nu = 0$, čime smo napravili **relativnu** grešku od **100%** (tj. **1**) u **početnoj** vrijednosti ...

Rekurzija unatrag — točnost i start ν

- ... a još uvijek dobivamo \hat{I}_n s **relativnom** greškom

$$\left| \frac{\hat{I}_n - I_n}{I_n} \right| < \left(\frac{1}{5} \right)^{\nu-n}, \quad \nu > n.$$

- Povoljnim izborom ν , ocjenu na desnoj strani možemo napraviti **po volji malom** — ispod **tražene točnosti** ε .
- Dovoljno je uzeti $\hat{I}_\nu = 0$ i

$$\nu \geq n + \frac{\log(1/\varepsilon)}{\log 5},$$

a zatim računamo vrijednosti

$$\hat{I}_{k-1} = \frac{1}{5} \left(\frac{1}{k} - \hat{I}_k \right), \quad k = \nu, \nu - 1, \dots, n + 1.$$

Rekurzija unatrag — rezultati

Pitanje: Kako se **dobra** uvjetovanost **vidi**, kad stvarno računamo $g_{n,\nu}(I_\nu)$?

Pokaži program i rezultate za $\varepsilon = 10^{-19}$!

● Početna vrijednost je $\hat{I}_\nu = 0$.

● Za ovaj ε dobijemo

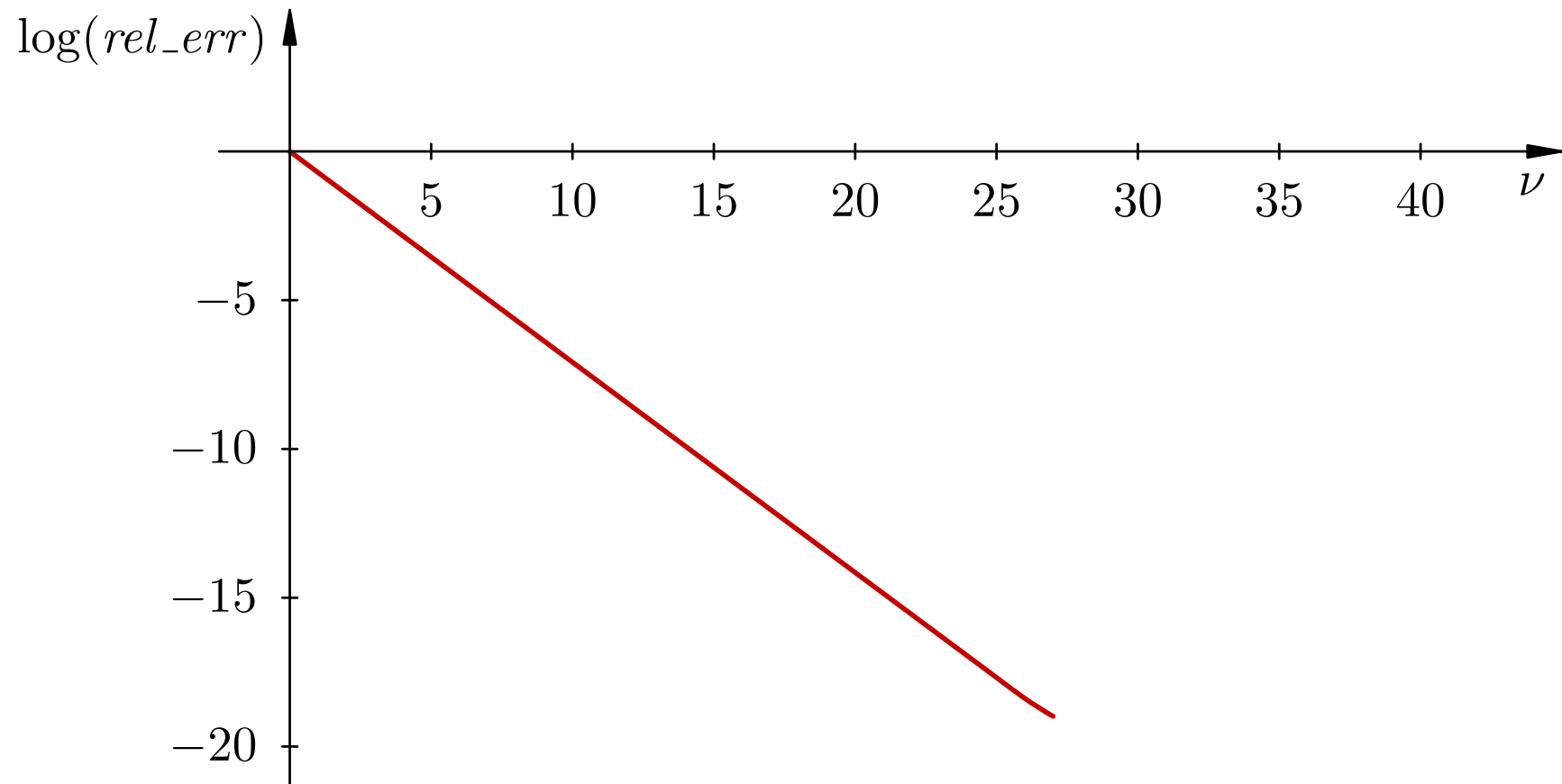
$$\nu \geq n + \frac{\log(1/\varepsilon)}{\log 5} \approx n + 28.$$

Dakle, “**silazno**” računamo **28** vrijednosti.

Usput, to su vrijednosti za I_n iz ranije tablice i **sve** prikazane znamenke su **točne** (ima ih **18**). Dakle, “ništa se **ne vidi**”.

Greška je samo **nekoliko** u , jer je $\varepsilon \approx 2u$ i imamo **3** operacije.

Rekurzija unazad za I_{40} — ovisno o startu ν



(\log_{10}) relativne greške izračunate vrijednosti
integrala I_{40} obratnom rekurzijom za $I_{40+\nu} = 0$

Rekurzivno računanje — završne napomene

Rekurzije prvog i (posebno) drugog reda se vrlo često koriste u praksi

- ne samo za računanje vrijednosti **integrala**,
- već za računanje raznih **specijalnih** funkcija (poput **Besselovih**) i **ortogonalnih** polinoma (v. kasnije).

Zato **oprez** ...

- treba znati nešto o **stabilnosti** rekurzije, **prije računanja!**

“Trik” **okretanja** rekurzije poznat je kao **Millerov algoritam**. Prvi puta je iskorišten baš za računanje **Besselovih** funkcija.

Primjer širenja grešaka zaokruživanja

Primjer širenja grešaka

Primjer. Vrijednost

$$f_n(x) = (x - n)^{10}, \quad n = 0, \dots, 10,$$

računamo u aritmetici računala (**extended**) u **okolini** točke n .

● Točka n je **deseterostruka** nultočka polinoma f_n .

Egzaktno gledano, graf funkcije $(x - n)^{10}$ = translaterani graf funkcije x^{10} za n jedinica **udesno**.

Funkcijske vrijednosti $f_n(x)$ možemo izračunati na **više** načina.

- Oni su **matematički ekvivalentni**,
- ali **nisu numerički jednaki** i ne daju iste rezultate.

Primjer širenja grešaka (nastavak)

Vrijednosti $f_n(x)$ računat ćemo na dva načina:

- binarnim potenciranjem vrijednosti $(x - n)$ na desetu potenciju — savršeno stabilno (ovo oduzimanje je dobro),
- razvojem po potencijama od x , korištenjem binomne formule

$$(x - n)^{10} = \sum_{k=0}^{10} \binom{10}{k} (-n)^{10-k} \cdot x^k,$$

s tim da polinom na desnoj strani računamo Hornerovom shemom.

Problem: U okolini točke n , $(x - n)^{10}$ je vrlo mali broj.

Članovi u sumi na desnoj strani su alternirajući po predznaku i rastu s porastom $n \implies$ mora doći do sve većeg kraćenja.

Primjer širenja grešaka (nastavak)

Kako izgledaju **izračunati** grafovi funkcije f_n ?

- 🔴 **zeleno** — graf dobiven **binarnim potenciranjem**,
- 🔴 **crveno** — graf dobiven **Hornerom** iz **binomne formule**.

Za svaki n crtamo **dvije** slike **izračunatih** grafova:

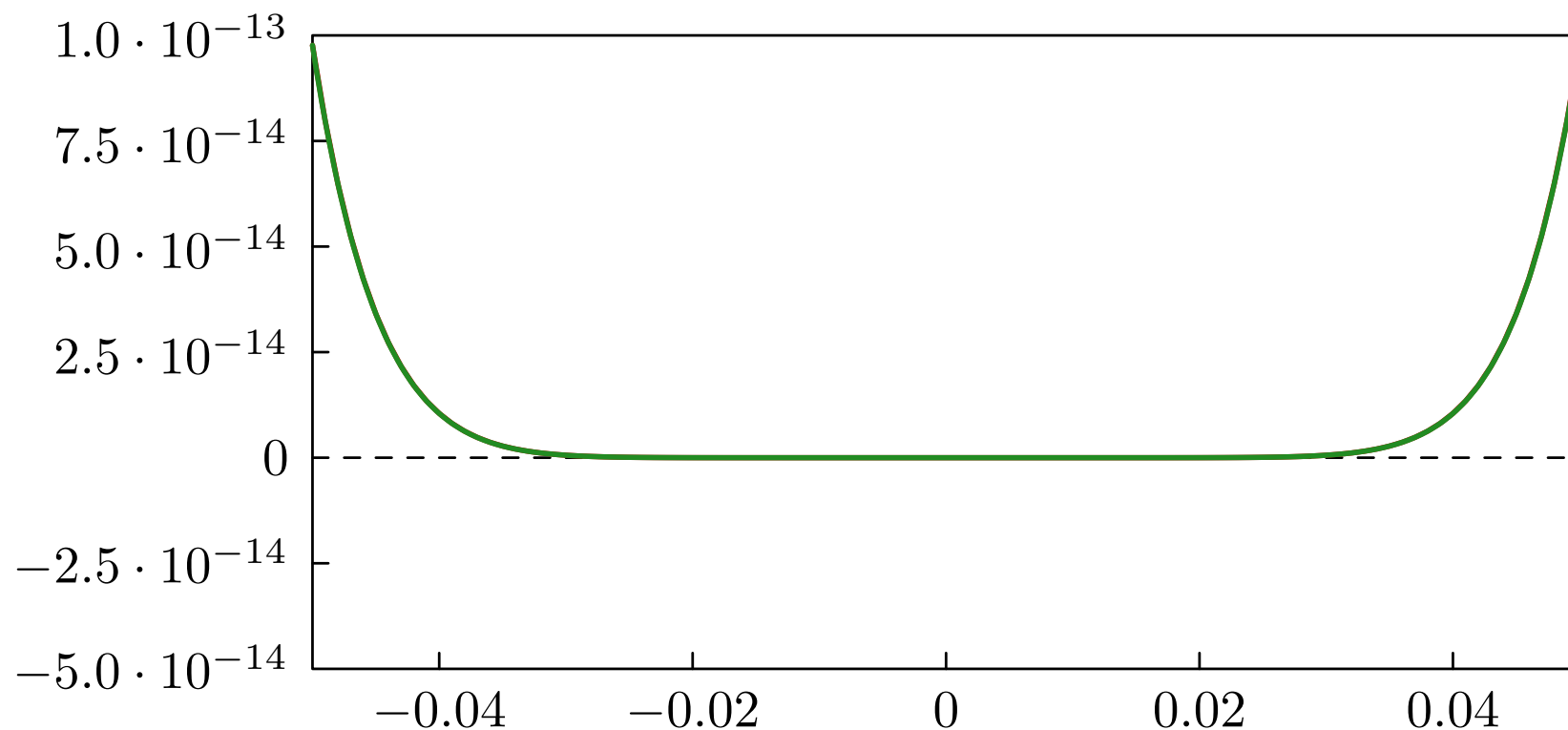
- 🔴 na intervalu $[n - 0.05, n + 0.05]$,
- 🔴 na intervalu $[n - r_n, n + r_n]$, gdje je r_n odabran tako da ovaj interval sadrži **numeričke nultočke** od f_n .

Obratite pažnju na **skale** po x i y !

- 🔴 Kako n raste, interval za **nultočke** se **širi** — r_n raste,
- 🔴 a funkcijske vrijednosti u tom intervalu postaju sve **udaljenije** od **nule** — **veće** po apsolutnoj vrijednosti.

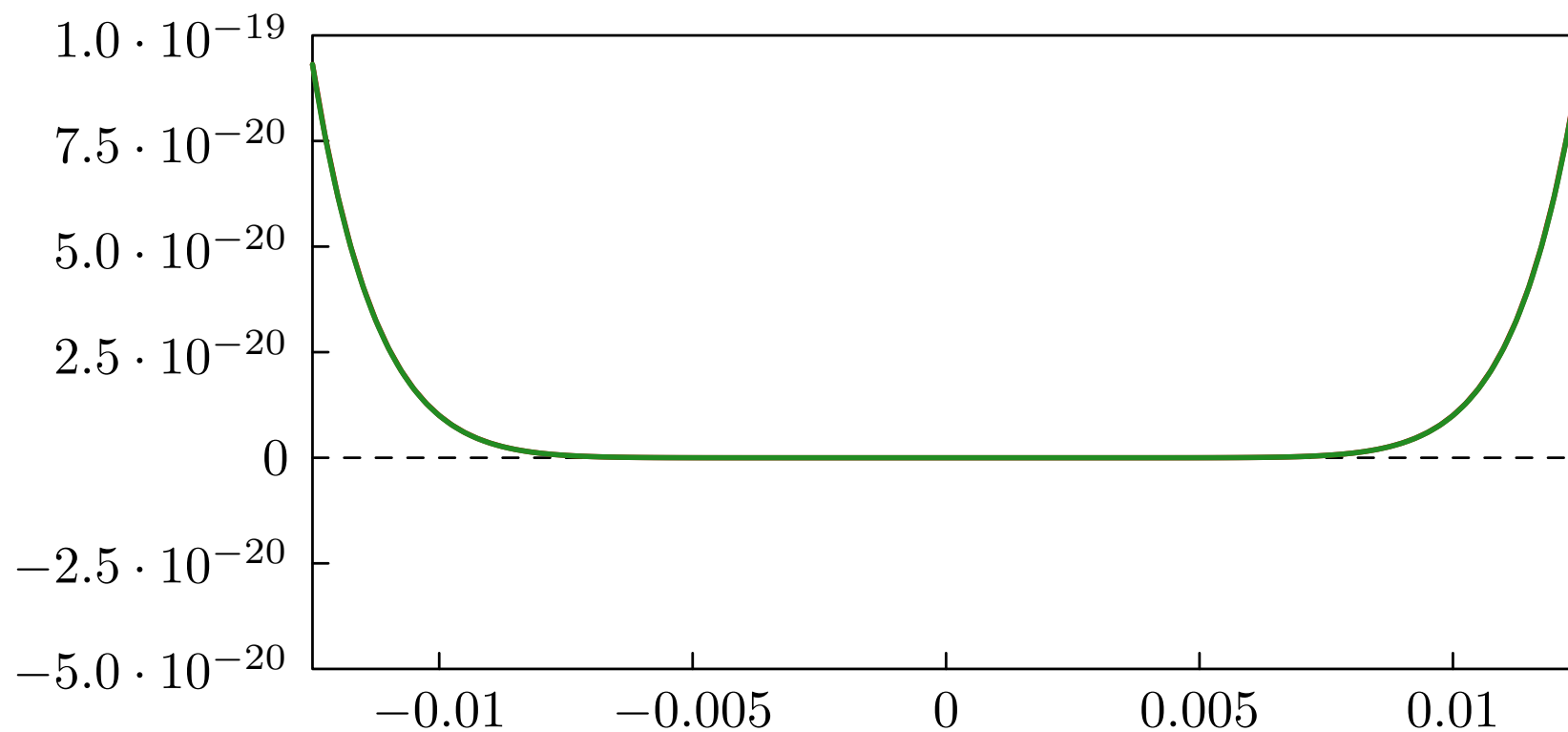
Primjer širenja grešaka — $n = 0, r = 0.05$

$$(x - 0)^{10} = x^{10}$$



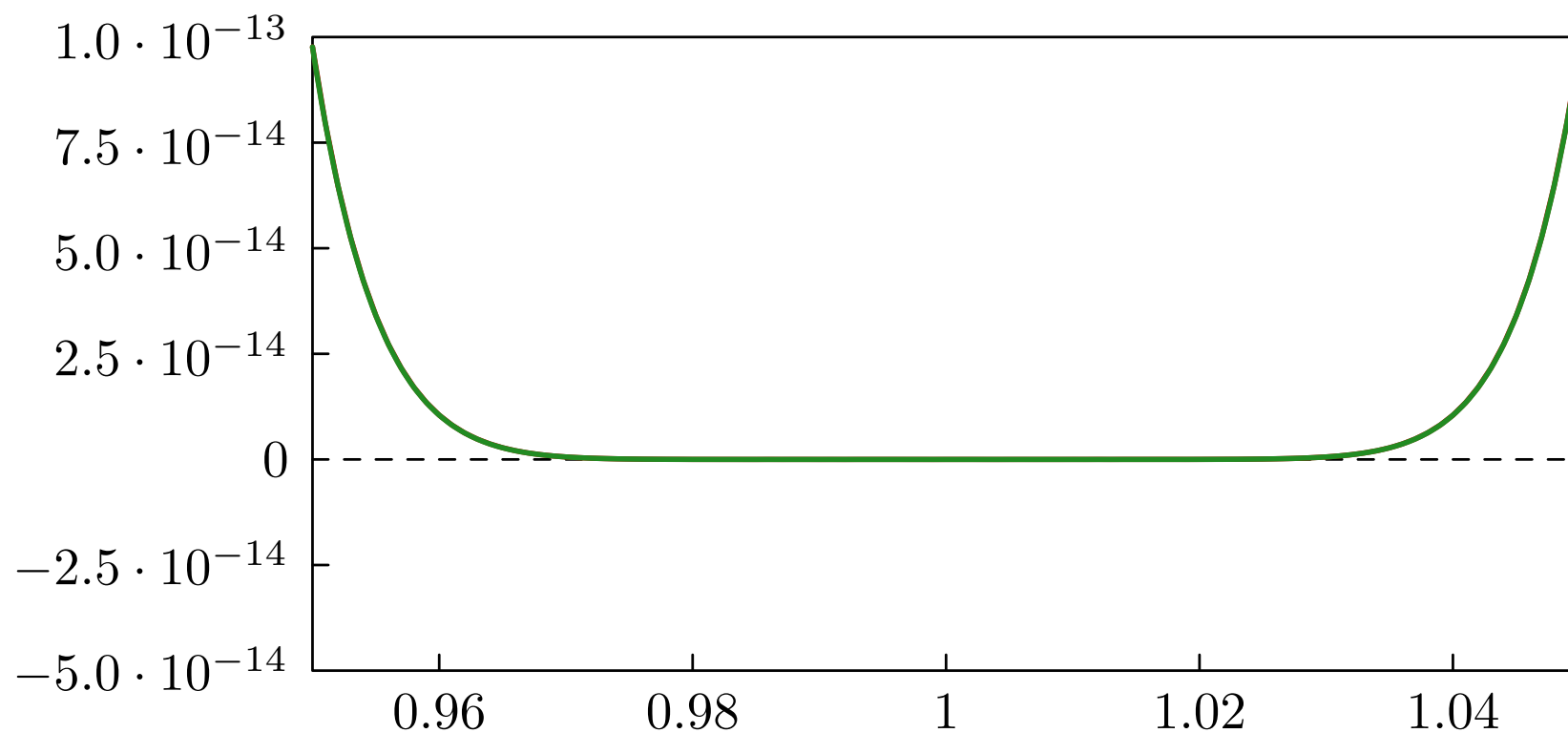
Primjer širenja grešaka — $n = 0$, $r_0 = 0.0125$

$$(x - 0)^{10} = x^{10}$$



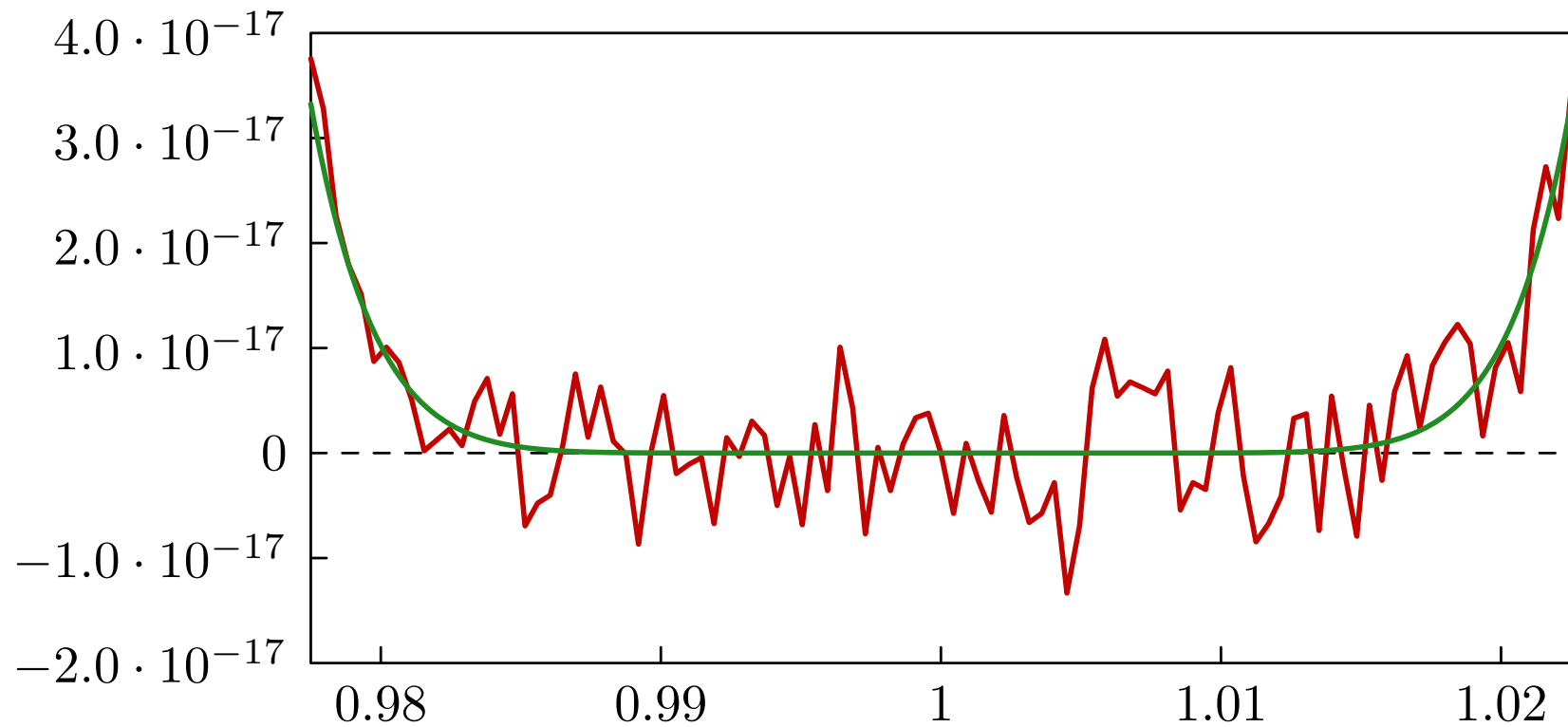
Primjer širenja grešaka — $n = 1, r = 0.05$

$$\begin{aligned}(x - 1)^{10} = & x^{10} - 10x^9 + 45x^8 - 120x^7 + 210x^6 \\ & - 252x^5 + 210x^4 - 120x^3 \\ & + 45x^2 - 10x + 1\end{aligned}$$



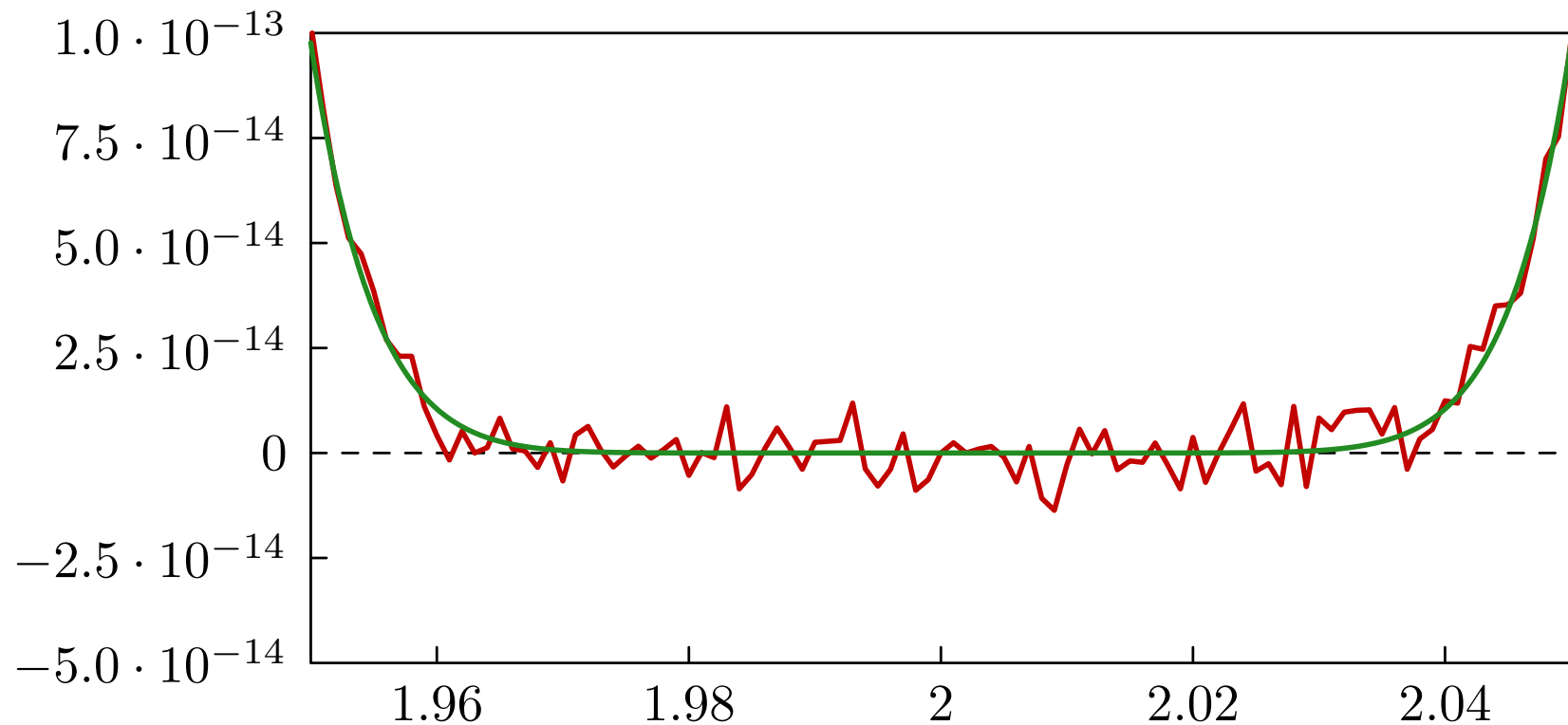
Primjer širenja grešaka — $n = 1, r_1 = 0.0225$

$$\begin{aligned}(x - 1)^{10} &= x^{10} - 10x^9 + 45x^8 - 120x^7 + 210x^6 \\ &\quad - 252x^5 + 210x^4 - 120x^3 \\ &\quad + 45x^2 - 10x + 1\end{aligned}$$



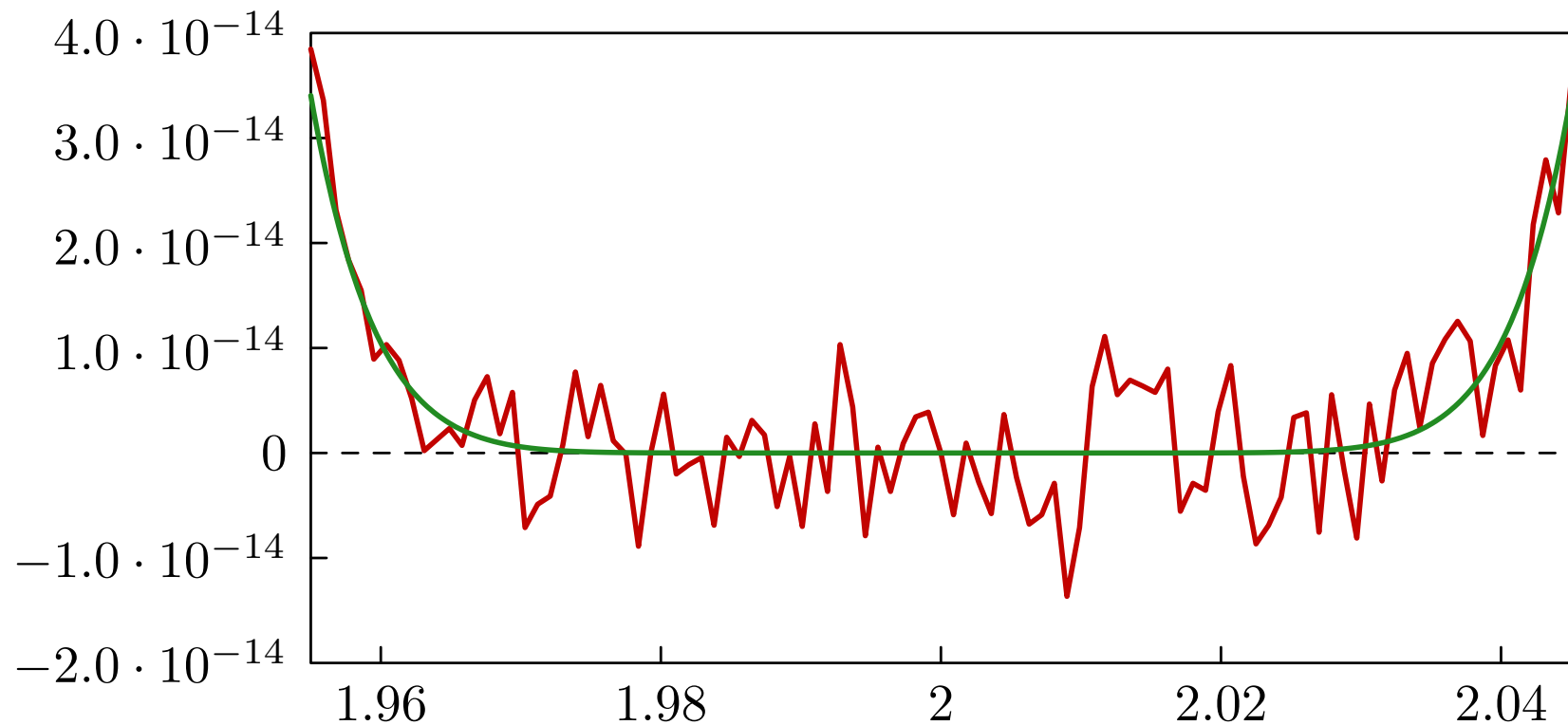
Primjer širenja grešaka — $n = 2, r = 0.05$

$$(x - 2)^{10} = x^{10} - 20x^9 + 180x^8 - 960x^7 + 3360x^6 \\ - 8064x^5 + 13440x^4 - 15360x^3 \\ + 11520x^2 - 5120x + 1024$$



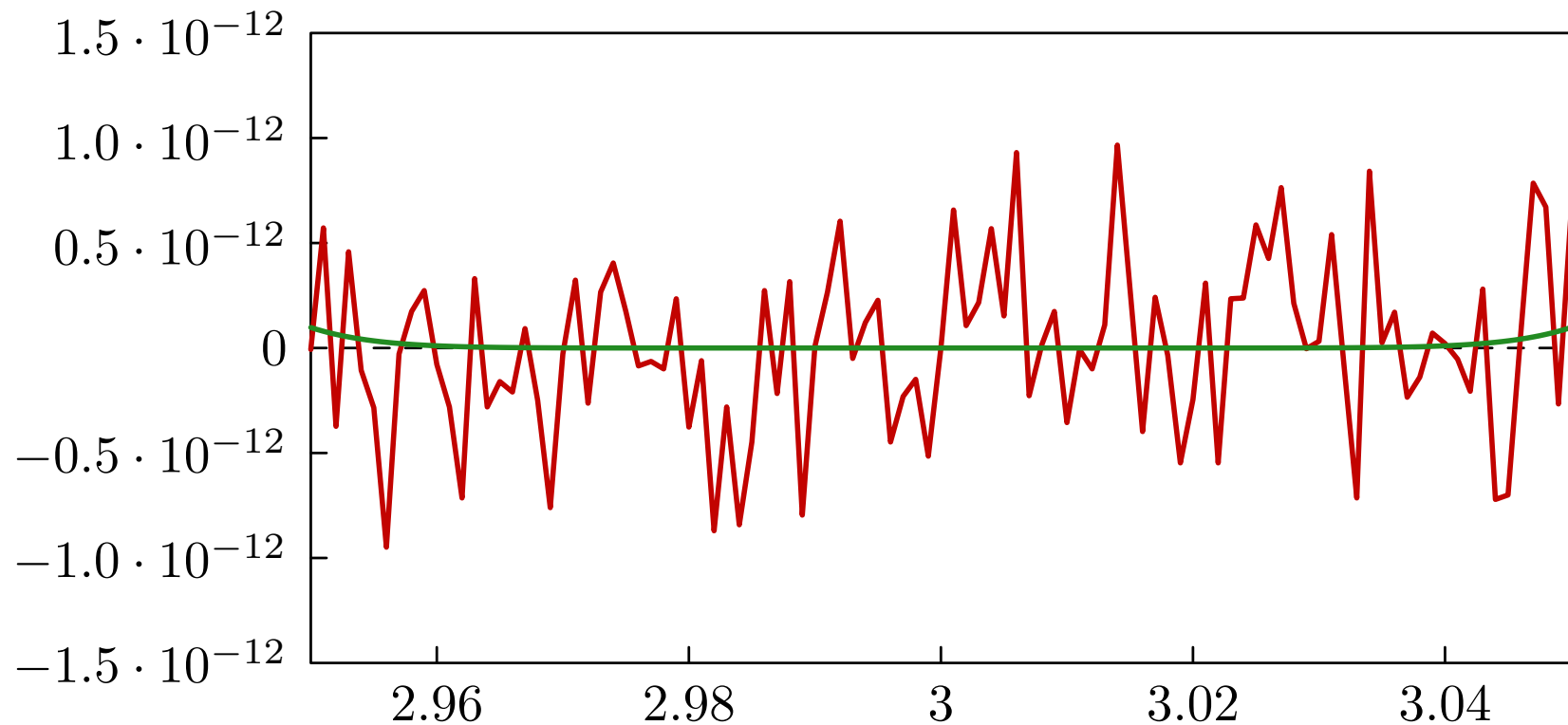
Primjer širenja grešaka — $n = 2, r_2 = 0.045$

$$(x - 2)^{10} = x^{10} - 20x^9 + 180x^8 - 960x^7 + 3360x^6 \\ - 8064x^5 + 13440x^4 - 15360x^3 \\ + 11520x^2 - 5120x + 1024$$



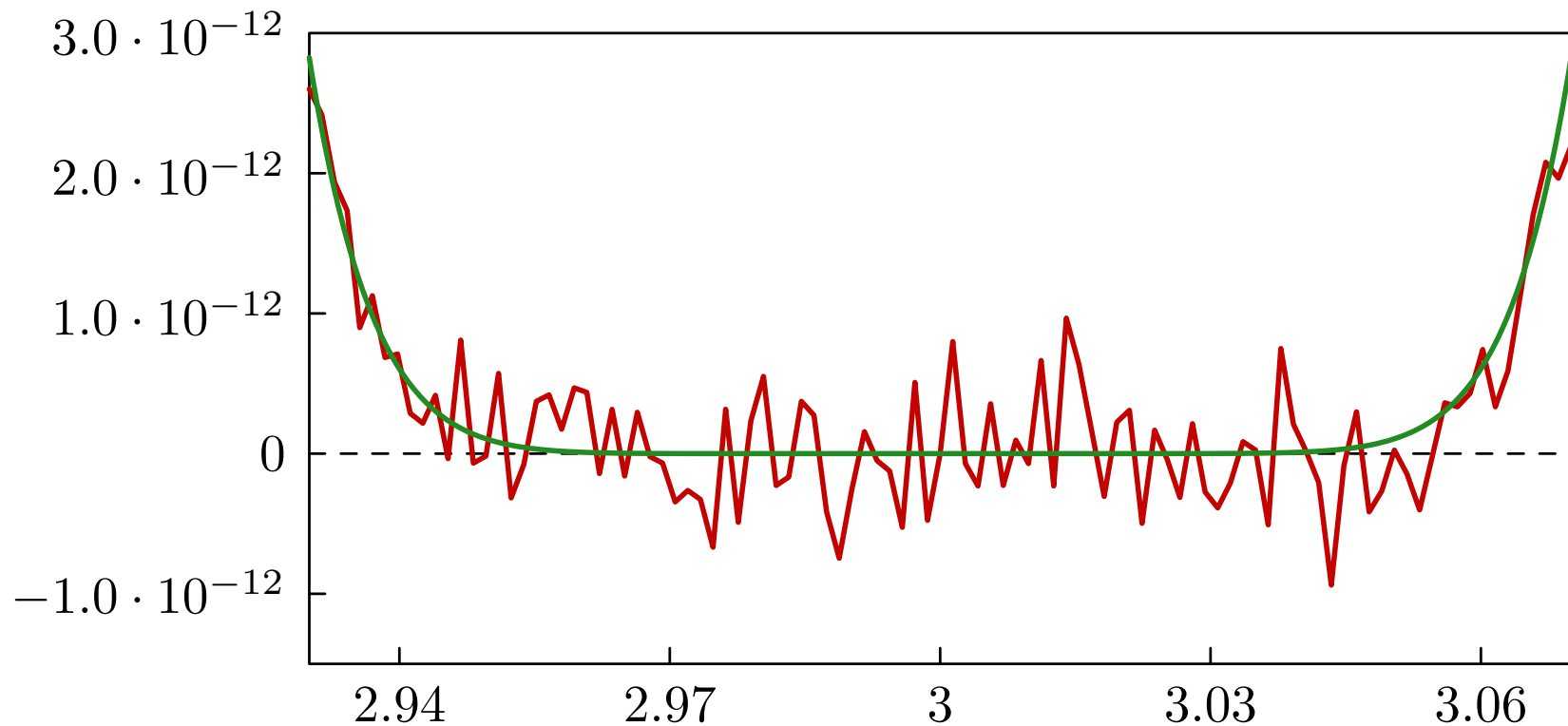
Primjer širenja grešaka — $n = 3, r = 0.05$

$$\begin{aligned}(x - 3)^{10} = & x^{10} - 30x^9 + 405x^8 - 3240x^7 + 17010x^6 \\ & - 61236x^5 + 153090x^4 - 262440x^3 \\ & + 295245x^2 - 196830x + 59049\end{aligned}$$



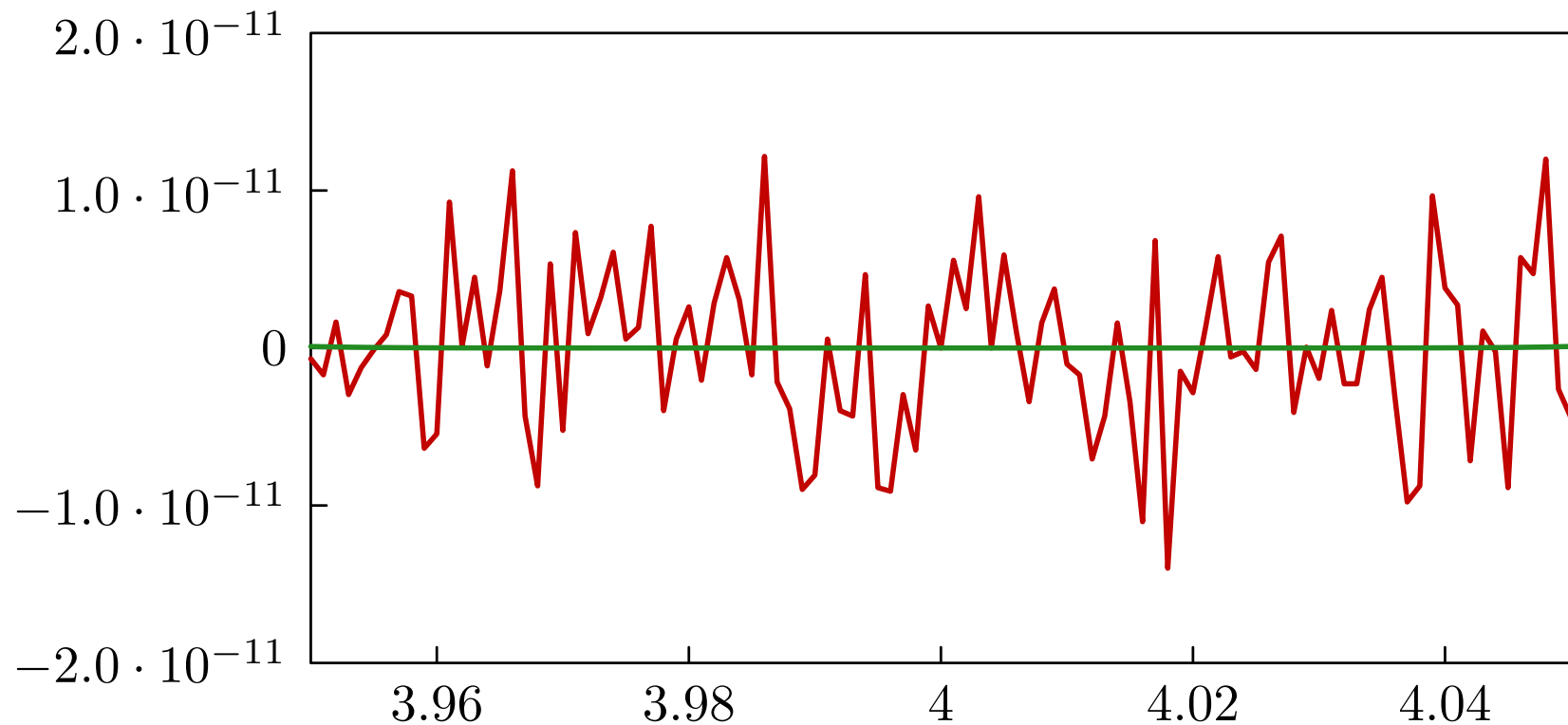
Primjer širenja grešaka — $n = 3, r_3 = 0.07$

$$\begin{aligned}(x - 3)^{10} = & x^{10} - 30x^9 + 405x^8 - 3240x^7 + 17010x^6 \\ & - 61236x^5 + 153090x^4 - 262440x^3 \\ & + 295245x^2 - 196830x + 59049\end{aligned}$$



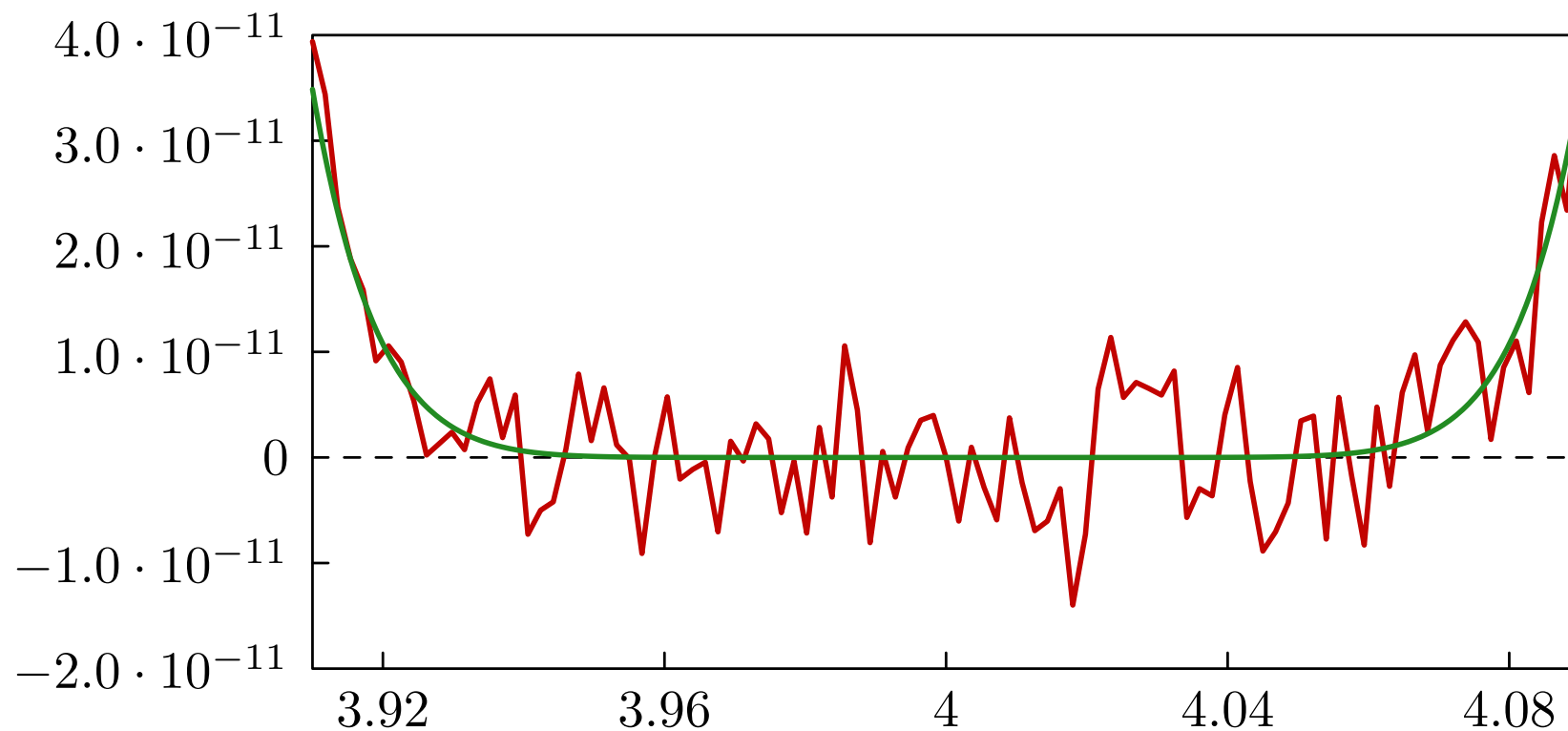
Primjer širenja grešaka — $n = 4, r = 0.05$

$$\begin{aligned}(x - 4)^{10} = & x^{10} - 40x^9 + 720x^8 - 7680x^7 + 53760x^6 \\ & - 258048x^5 + 860160x^4 - 1966080x^3 \\ & + 2949120x^2 - 2621440x + 1048576\end{aligned}$$



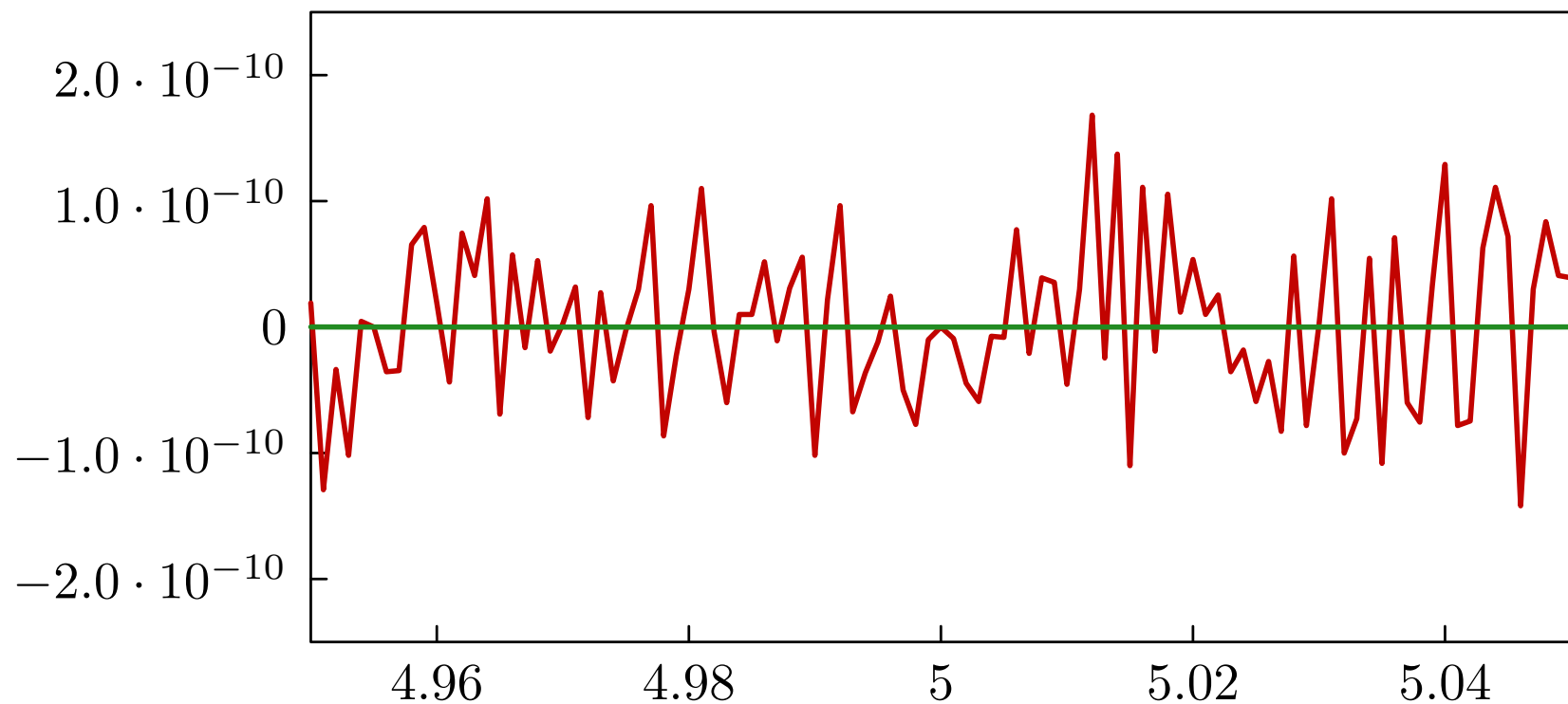
Primjer širenja grešaka — $n = 4, r_4 = 0.09$

$$\begin{aligned}(x - 4)^{10} = & x^{10} - 40x^9 + 720x^8 - 7680x^7 + 53760x^6 \\ & - 258048x^5 + 860160x^4 - 1966080x^3 \\ & + 2949120x^2 - 2621440x + 1048576\end{aligned}$$



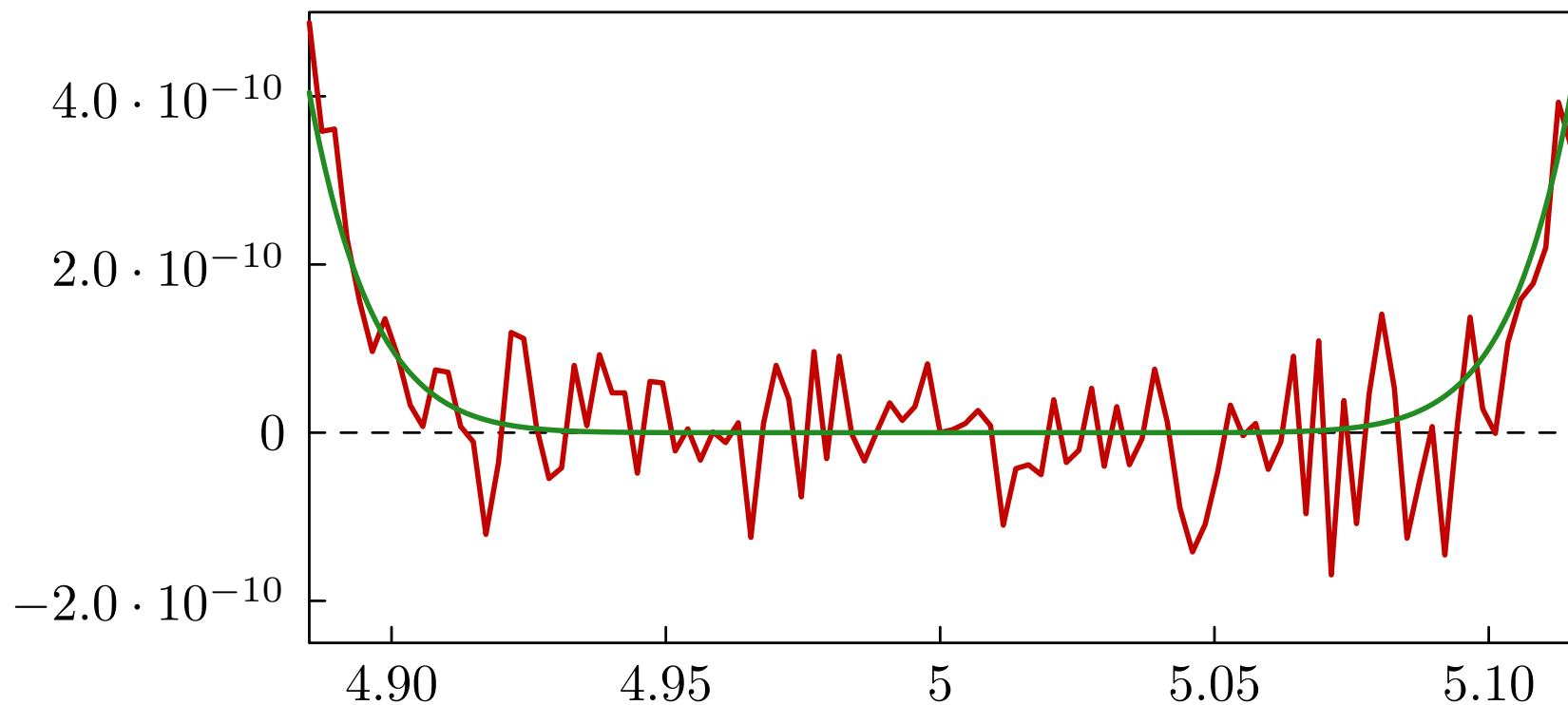
Primjer širenja grešaka — $n = 5, r = 0.05$

$$\begin{aligned}(x - 5)^{10} = & x^{10} - 50x^9 + 1125x^8 - 15000x^7 + 131250x^6 \\ & - 787500x^5 + 3281250x^4 - 9375000x^3 \\ & + 17578125x^2 - 19531250x + 9765625\end{aligned}$$



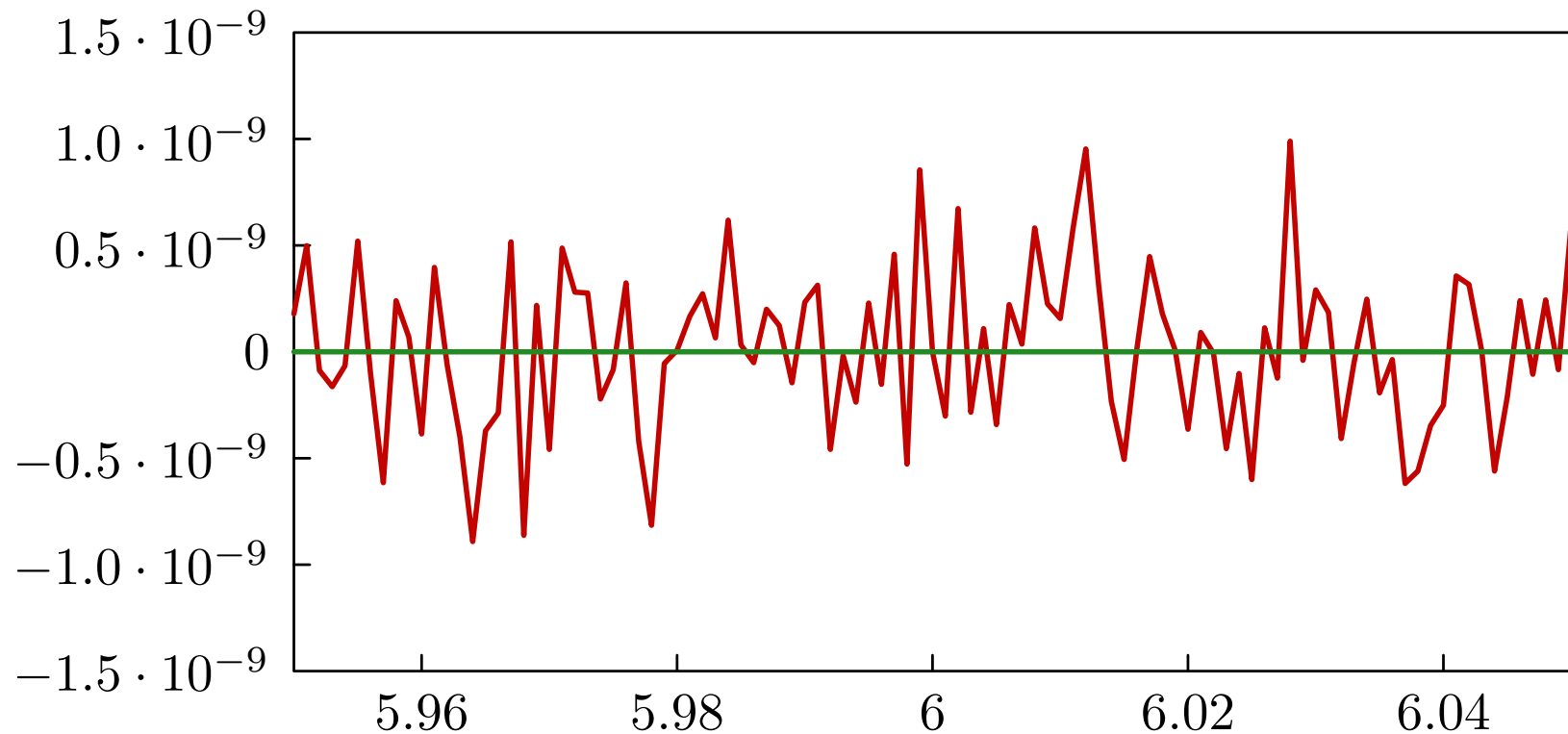
Primjer širenja grešaka — $n = 5, r_5 = 0.115$

$$\begin{aligned}(x - 5)^{10} = & x^{10} - 50x^9 + 1125x^8 - 15000x^7 + 131250x^6 \\ & - 787500x^5 + 3281250x^4 - 9375000x^3 \\ & + 17578125x^2 - 19531250x + 9765625\end{aligned}$$



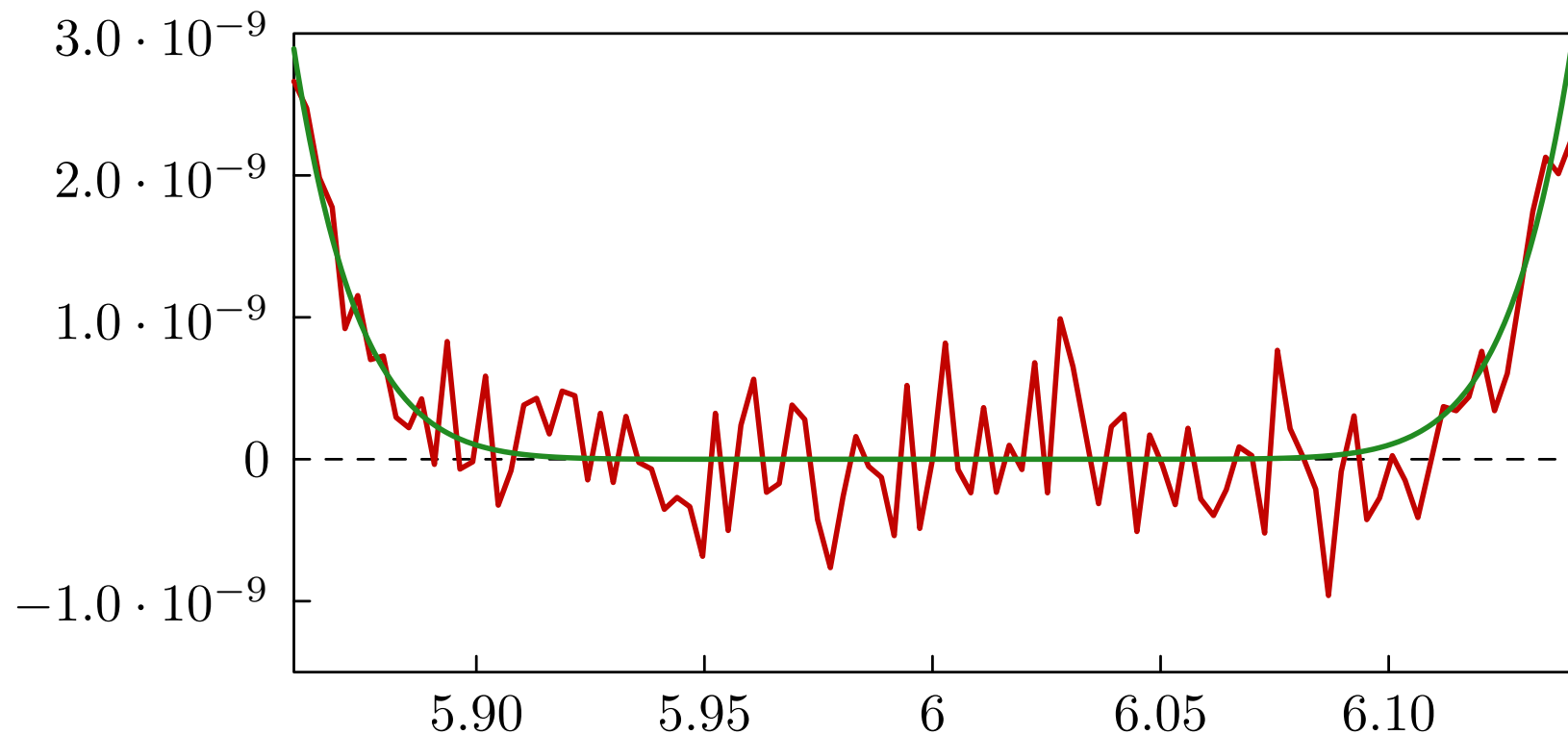
Primjer širenja grešaka — $n = 6, r = 0.05$

$$\begin{aligned}(x - 6)^{10} = & x^{10} - 60x^9 + 1620x^8 - 25920x^7 + 272160x^6 \\ & - 1959552x^5 + 9797760x^4 - 33592320x^3 \\ & + 75582720x^2 - 100776960x^1 + 60466176\end{aligned}$$



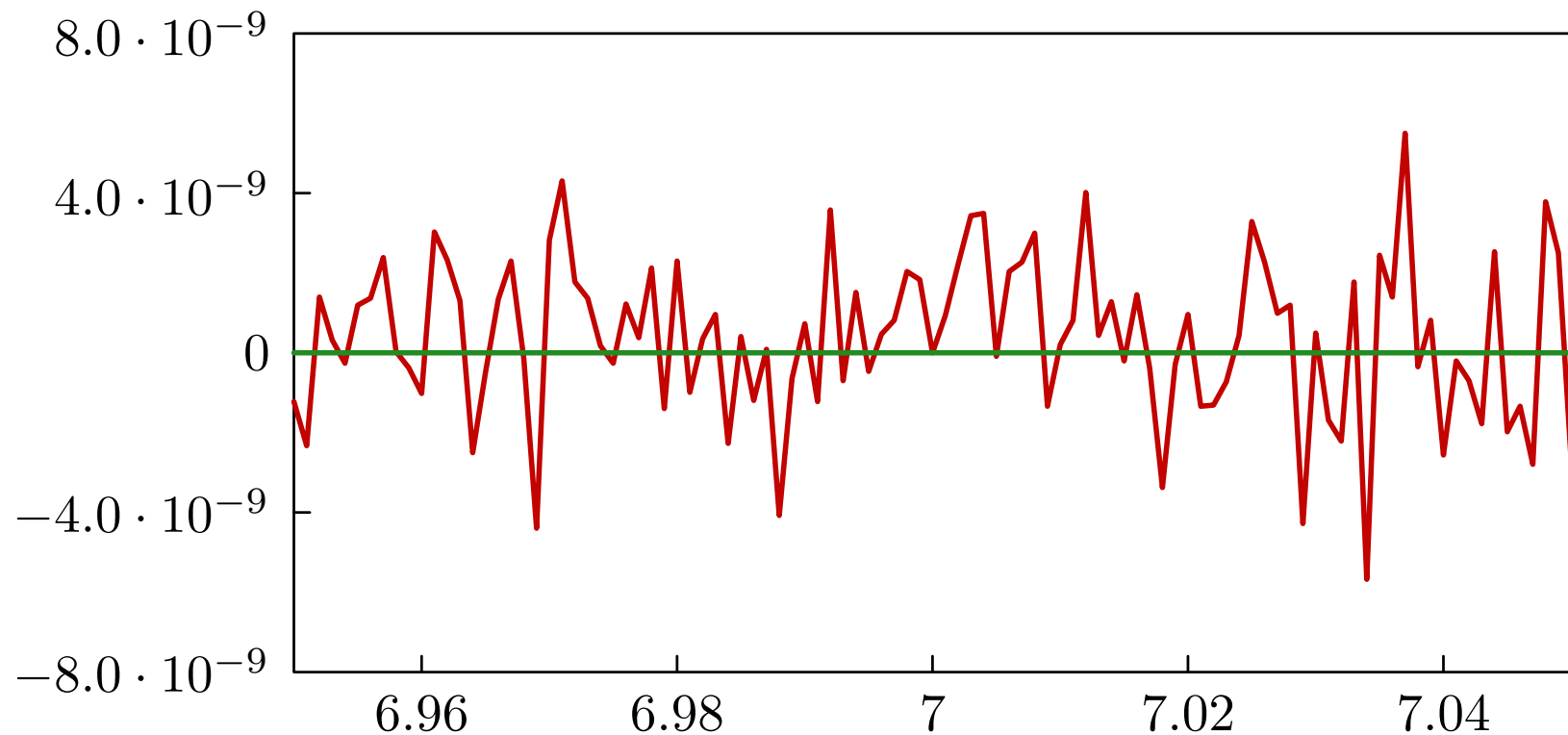
Primjer širenja grešaka — $n = 6, r_6 = 0.14$

$$(x - 6)^{10} = x^{10} - 60x^9 + 1620x^8 - 25920x^7 + 272160x^6 \\ - 1959552x^5 + 9797760x^4 - 33592320x^3 \\ + 75582720x^2 - 100776960x^1 + 60466176$$



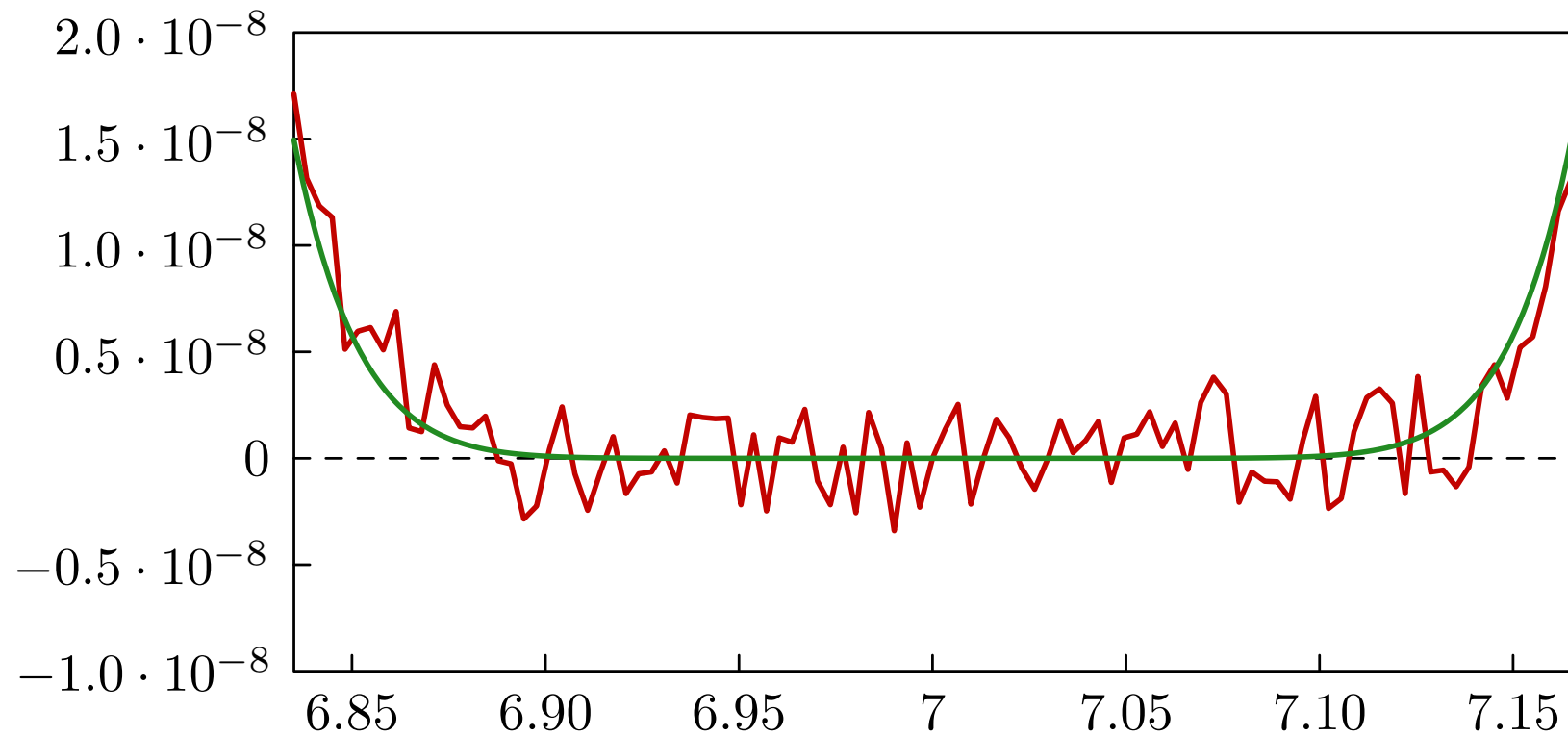
Primjer širenja grešaka — $n = 7, r = 0.05$

$$\begin{aligned}(x - 7)^{10} = & x^{10} - 70x^9 + 2205x^8 - 41160x^7 + 504210x^6 \\ & - 4235364x^5 + 24706290x^4 - 98825160x^3 \\ & + 259416045x^2 - 403536070x^1 + 282475249\end{aligned}$$



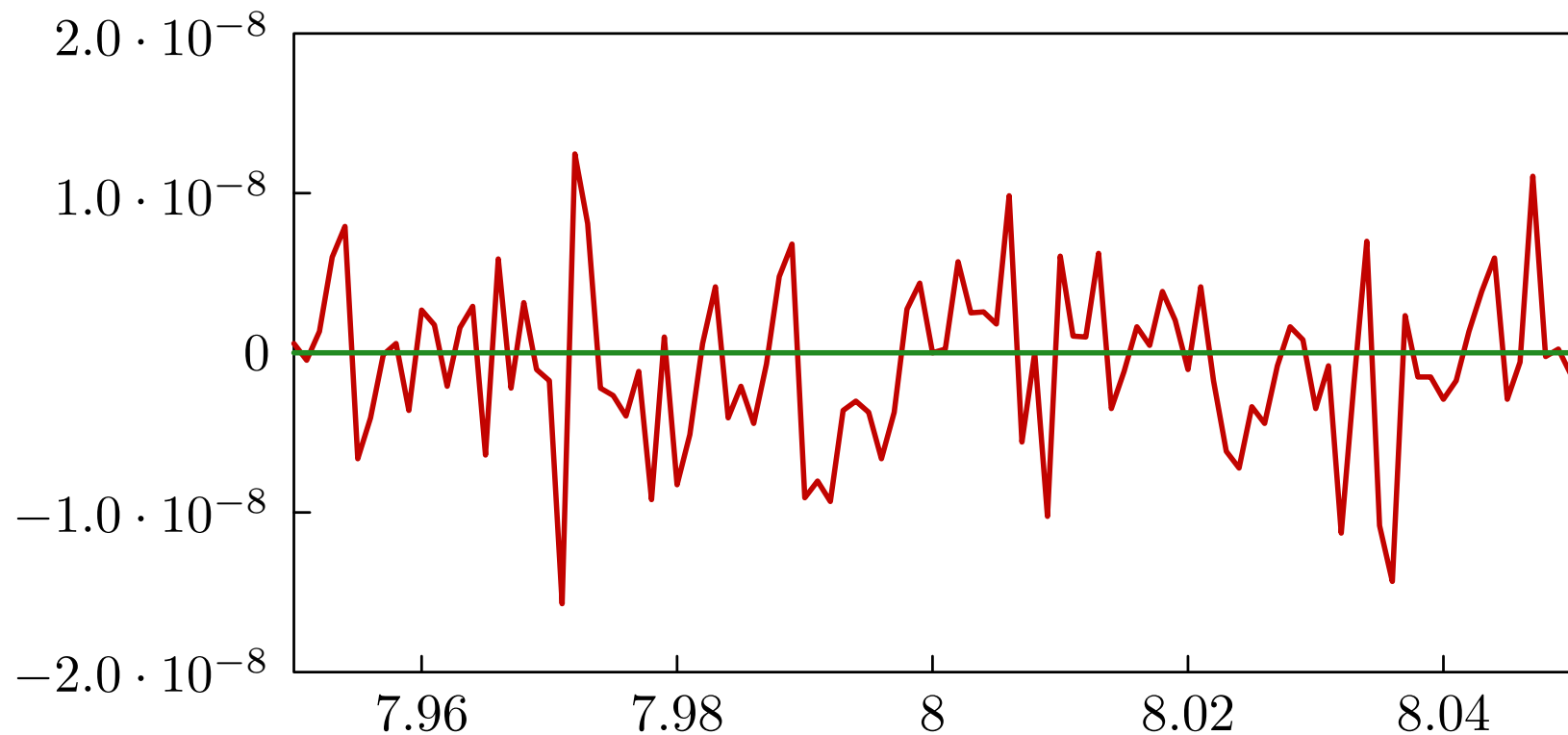
Primjer širenja grešaka — $n = 7, r_7 = 0.165$

$$\begin{aligned}(x - 7)^{10} = & x^{10} - 70x^9 + 2205x^8 - 41160x^7 + 504210x^6 \\ & - 4235364x^5 + 24706290x^4 - 98825160x^3 \\ & + 259416045x^2 - 403536070x^1 + 282475249\end{aligned}$$



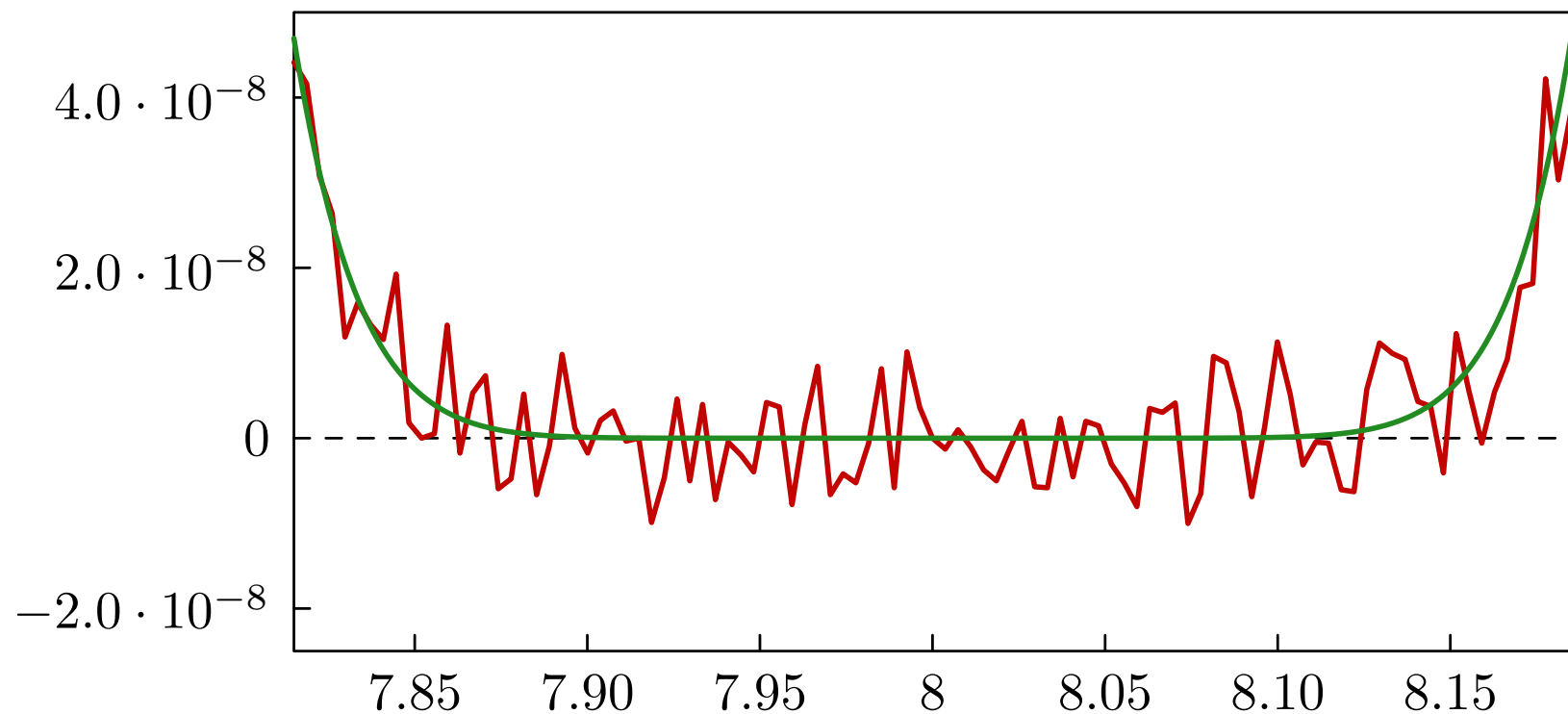
Primjer širenja grešaka — $n = 8, r = 0.05$

$$\begin{aligned}(x - 8)^{10} = & x^{10} - 80x^9 + 2880x^8 - 61440x^7 + 860160x^6 \\ & - 8257536x^5 + 55050240x^4 - 251658240x^3 \\ & + 754974720x^2 - 1342177280x^1 + 1073741824\end{aligned}$$



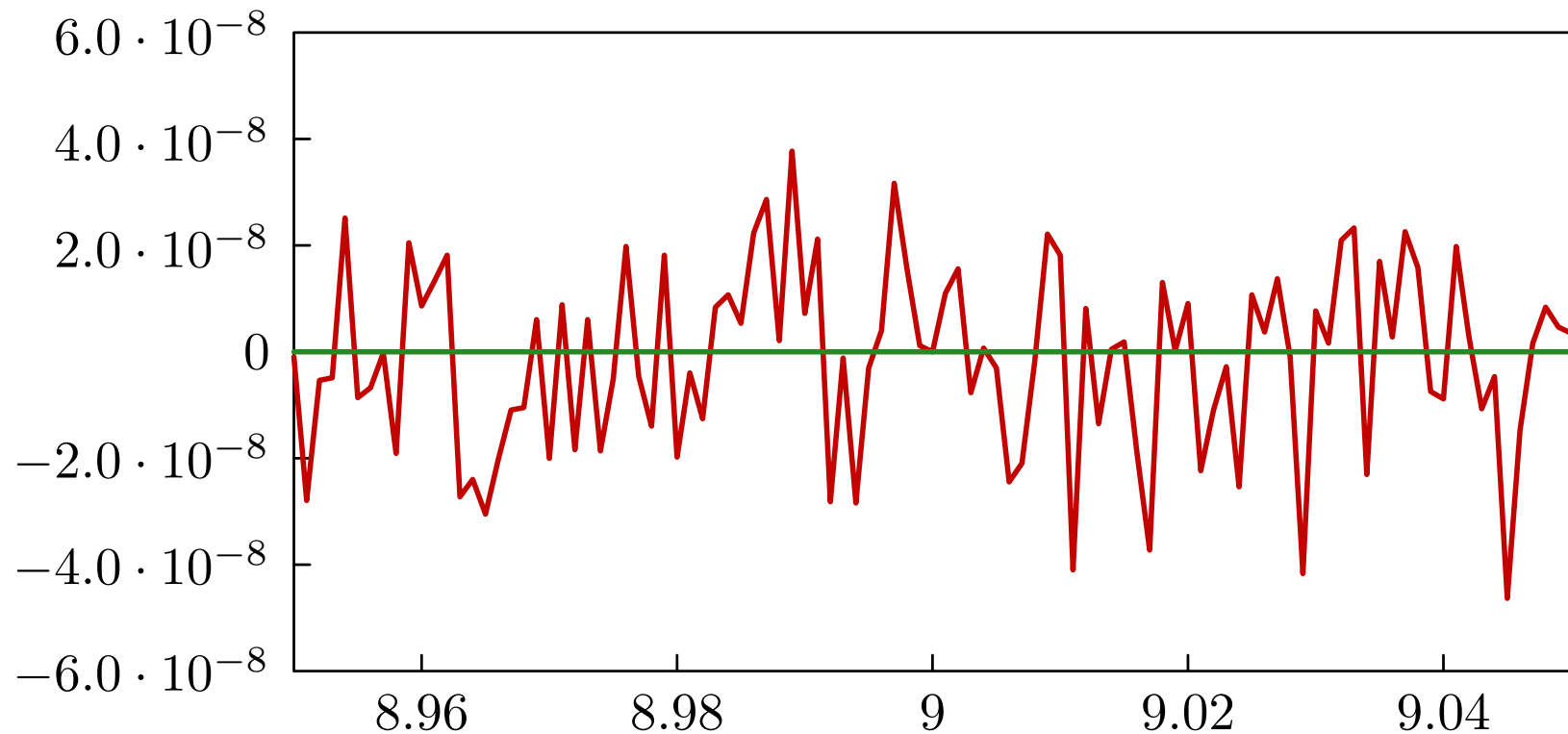
Primjer širenja grešaka — $n = 8, r_8 = 0.185$

$$\begin{aligned}(x - 8)^{10} = & x^{10} - 80x^9 + 2880x^8 - 61440x^7 + 860160x^6 \\ & - 8257536x^5 + 55050240x^4 - 251658240x^3 \\ & + 754974720x^2 - 1342177280x^1 + 1073741824\end{aligned}$$



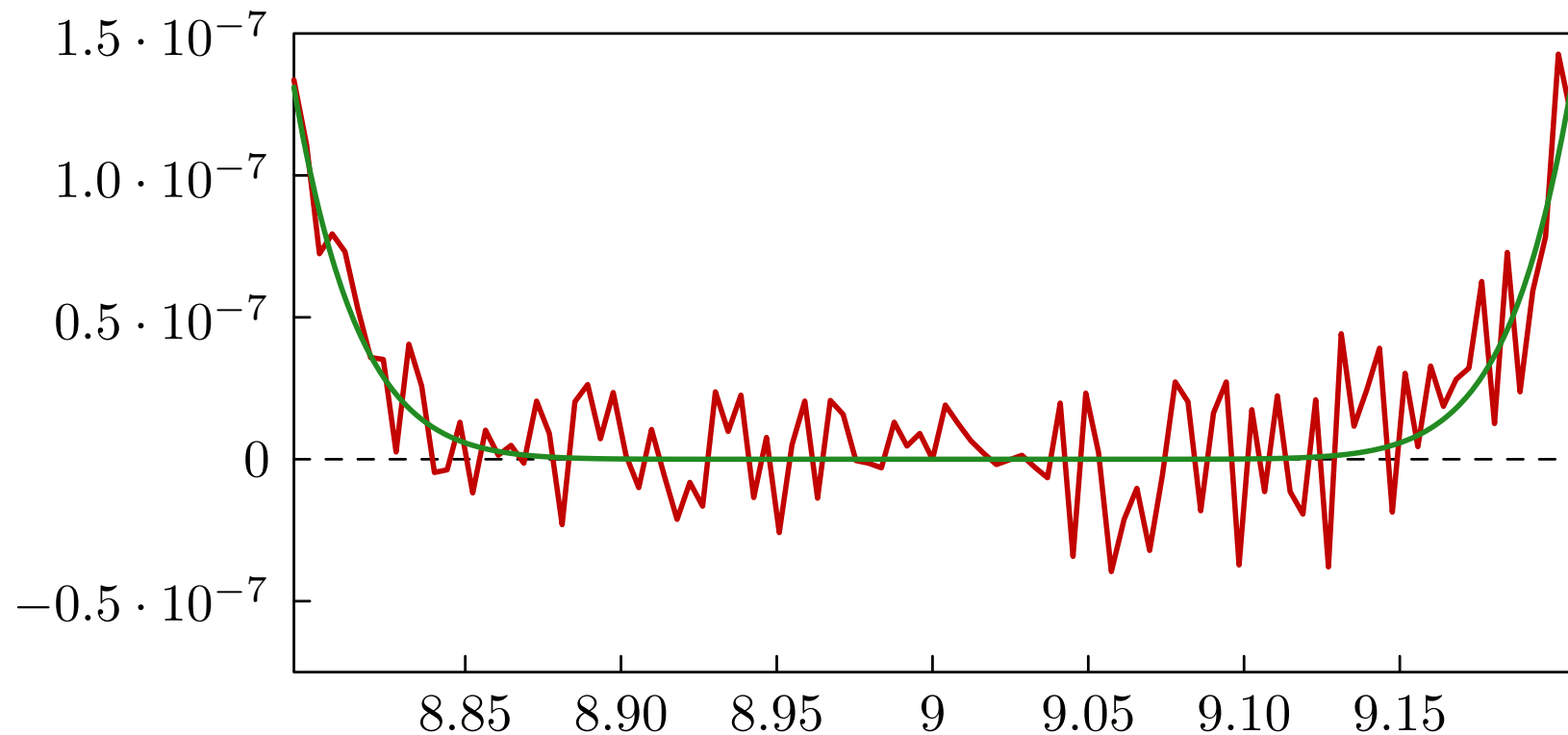
Primjer širenja grešaka — $n = 9, r = 0.05$

$$\begin{aligned}(x - 9)^{10} = & x^{10} - 90x^9 + 3645x^8 - 87480x^7 + 1377810x^6 \\ & - 14880348x^5 + 111602610x^4 - 573956280x^3 \\ & + 1937102445x^2 - 3874204890x^1 + 3486784401\end{aligned}$$



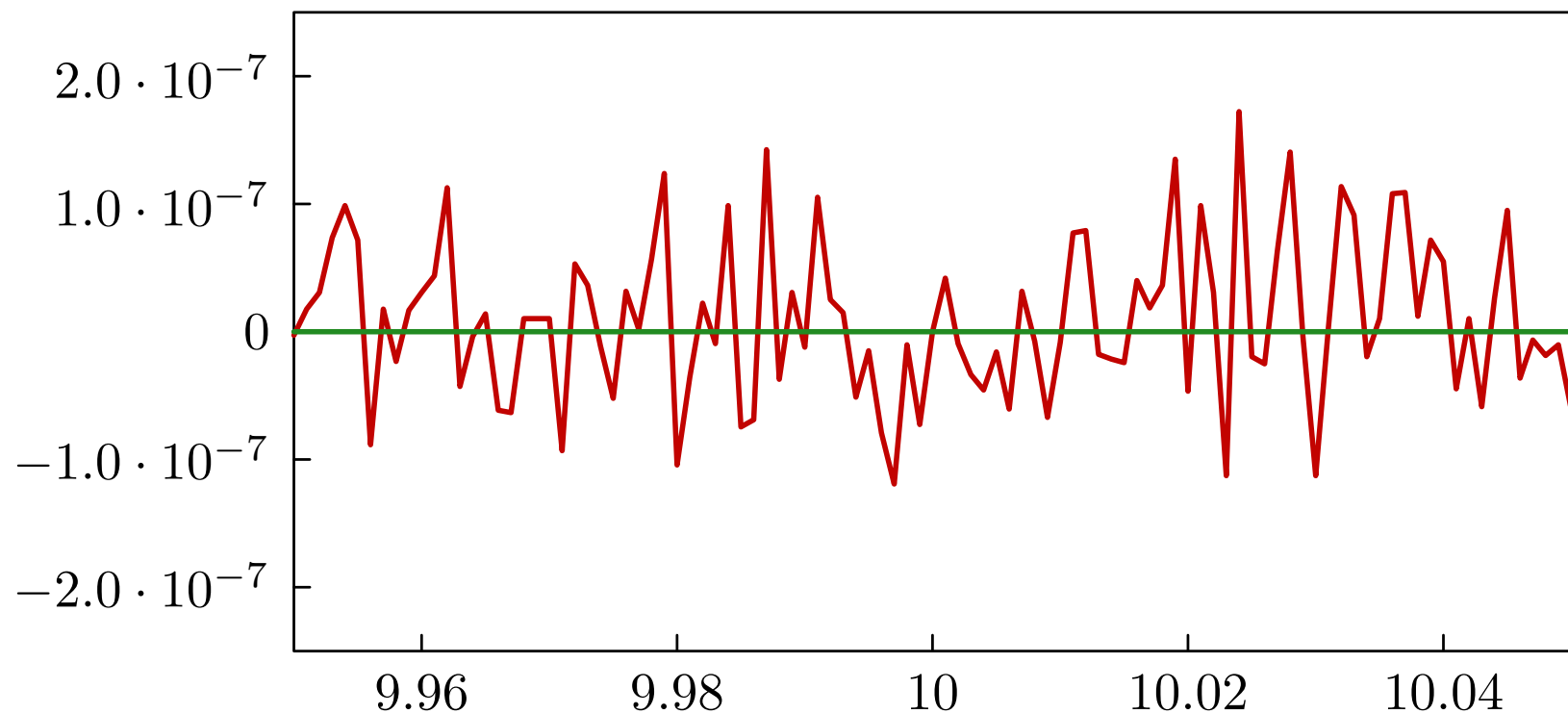
Primjer širenja grešaka — $n = 9$, $r_9 = 0.205$

$$\begin{aligned}(x - 9)^{10} = & x^{10} - 90x^9 + 3645x^8 - 87480x^7 + 1377810x^6 \\ & - 14880348x^5 + 111602610x^4 - 573956280x^3 \\ & + 1937102445x^2 - 3874204890x^1 + 3486784401\end{aligned}$$



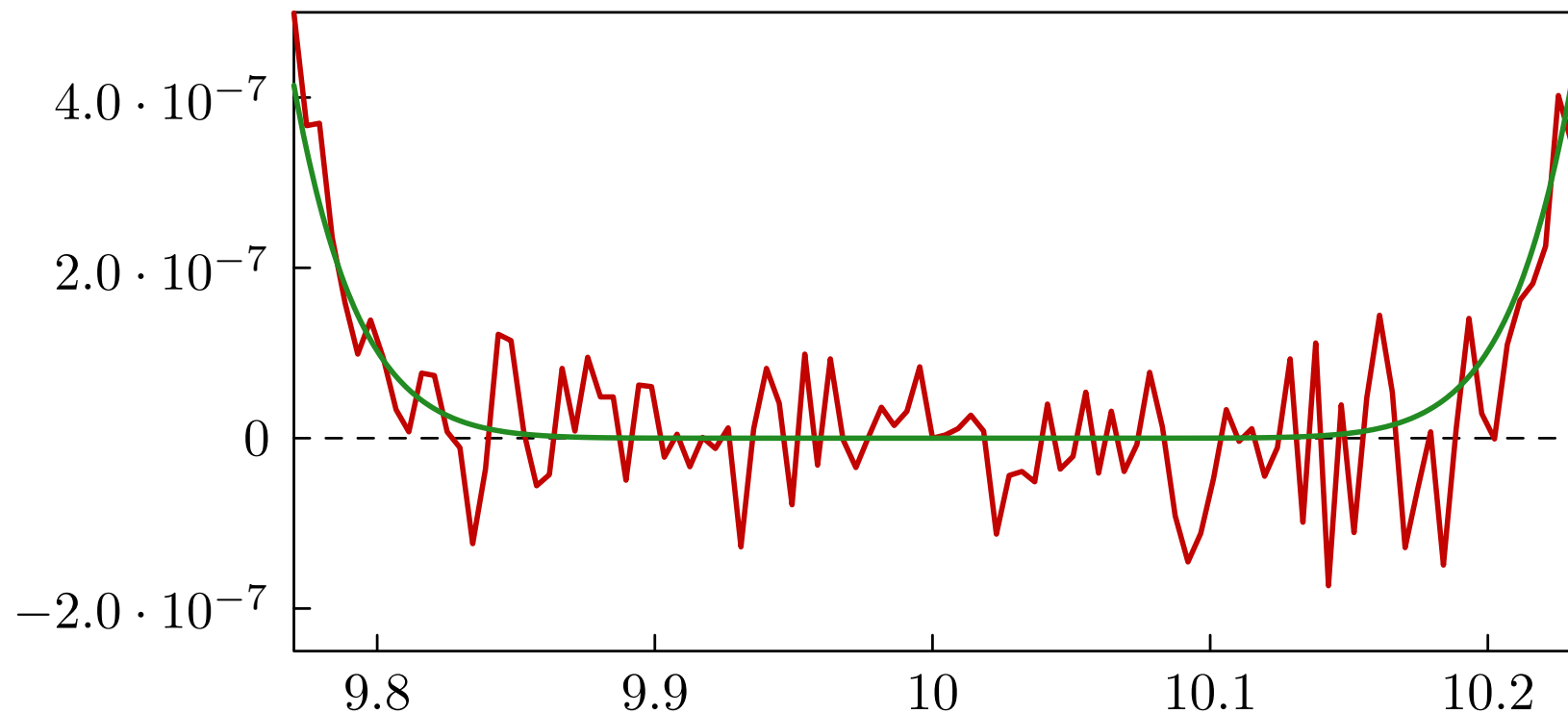
Primjer širenja grešaka — $n = 10, r = 0.05$

$$\begin{aligned}(x - 10)^{10} = & x^{10} - 100x^9 + 4500x^8 - 120000x^7 + 2100000x^6 \\ & - 25200000x^5 + 210000000x^4 - 1200000000x^3 \\ & + 4500000000x^2 - 10000000000x^1 + 10000000000\end{aligned}$$



Primjer širenja grešaka — $n = 10, r_{10} = 0.23$

$$\begin{aligned}(x - 10)^{10} = & x^{10} - 100x^9 + 4500x^8 - 120000x^7 + 2100000x^6 \\ & - 25200000x^5 + 210000000x^4 - 1200000000x^3 \\ & + 4500000000x^2 - 10000000000x^1 + 10000000000\end{aligned}$$



Primjeri izbjegavanja kraćenja

Primjer: Kvadratna jednadžba

Kvadratna jednadžba

Uzmimo da treba riješiti (realnu) kvadratnu jednadžbu

$$ax^2 + bx + c = 0,$$

gdje su a , b i c zadani, i vrijedi $a \neq 0$.

Matematički gledano, problem je lagan: imamo 2 rješenja

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Numerički gledano, problem je mnogo izazovniji:

- ni uspješno računanje po ovoj formuli,
- ni točnost izračunatih korijena,

ne možemo uzeti “zdravo za gotovo”.

Kvadratna jednadžba — problem

Primjer. Rješavamo kvadratnu jednadžbu $x^2 - 56x + 1 = 0$.

U dekadskoj aritmetici s $p = 5$ značajnih znamenki dobijemo

$$x_1 = \frac{56 - \sqrt{3132}}{2} = \frac{56 - 55.964}{2} = 0.018000,$$

$$x_2 = \frac{56 + \sqrt{3132}}{2} = \frac{56 + 55.964}{2} = 55.982.$$

Točna rješenja su

$$x_1 = 0.0178628\dots \quad \text{i} \quad x_2 = 55.982137\dots$$

Apsolutno **manji** od ova dva korijena — x_1 , ima **samo dvije** točne znamenke (**kraćenje**). Drugi je “savršeno” **točan**.

Kvadratna jednadžba — popravak

Prvo izračunamo **većeg** po apsolutnoj vrijednosti, po formuli

$$x_2 = \frac{-(b + \text{sign}(b)\sqrt{b^2 - 4ac})}{2a},$$

a **manjeg** po apsolutnoj vrijednosti, izračunamo iz

$$x_1 \cdot x_2 = \frac{c}{a}$$

(Vieta), tj. formula za x_1 je

$$x_1 = \frac{c}{x_2 a}.$$

Opasnog **kraćenja** za x_1 više **nema!**

Kvadratna jednadžba (nastavak)

Ovo je bila samo **jedna**, od (barem) **tri** “opasne” točke za računanje. Preostale **dvije** su:

- “kvadriranje” pod korijenom — mogućnost za **overflow**.
Rješenje — “skaliranjem”.
- **oduzimanje** u diskriminanti (**kraćenje**) — **nema** jednostavnog rješenja. Naime, “krivac” **nije** aritmetika.
 - To je samo odraz tzv. **nestabilnosti** problema, jer tad imamo **dva bliska korijena** koji su **vrlo osjetljivi** na male **perturbacije** koeficijenata jednadžbe.
 - Na primjer, pomak c = pomak grafa “**gore–dolje**”. **Mali** pomak rezultira **velikom** promjenom korijena!

Neki primjeri izbjegavanja kraćenja

Primjer. Treba izračunati

$$y = \sqrt{x + \delta} - \sqrt{x},$$

gdje su x i δ zadani ulazni podaci, s tim da je $x > 0$,

• a $|\delta|$ vrlo mali broj.

U ovoj formuli, očito, dolazi do velike greške zbog kraćenja — zaokruživanje korijena prije oduzimanja.

Ako formulu “deracionaliziramo” u oblik

$$y = \frac{\delta}{\sqrt{x + \delta} + \sqrt{x}},$$

problema više nema!

Neki primjeri izbjegavanja kraćenja

Primjer. Treba izračunati

$$y = \cos(x + \delta) - \cos x,$$

gdje su x i δ zadani ulazni podaci, s tim da je $|\cos x|$ razumno velik,

• a $|\delta|$ vrlo mali broj.

Opet, dolazi do velike greške zbog kraćenja.

Ako formulu napišmo u “produktnom” obliku

$$y = -2 \sin \frac{\delta}{2} \sin \left(x + \frac{\delta}{2} \right),$$

problema više nema!

Primjer za nultočke polinoma

Svojstvene vrijednosti i nultočke polinoma

U linearnoj algebri, svojstvene vrijednosti zadane matrice A se računaju “na ruke” kao

• nultočke karakterističnog polinoma te matrice

$$k_A(\lambda) = \det(\lambda I - A) = 0.$$

Prvo, računanjem determinante, nađemo “standardni” oblik karakterističnog polinoma, preko koeficijenata

$$k_A(\lambda) = \lambda^n + c_{n-1}\lambda^{n-1} + \dots + c_1\lambda + c_0,$$

a onda tražimo nultočke $\lambda_1, \dots, \lambda_n$ tog polinoma.

Oprez: Nultočke polinoma mogu biti vrlo osjetljive na male perturbacije u koeficijentima polinoma.

Primjer — Wilkinsonov polinom

Primjer. Uzmimo tzv. **Wilkinsonov** polinom stupnja $n = 20$,

$$P_{20}(\lambda) = (\lambda - 1) \cdot (\lambda - 2) \cdots (\lambda - 19) \cdot (\lambda - 20).$$

Iz ovog “multiplikativnog” oblika odmah čitamo da su **nultočke** tog polinoma, redom, prirodni brojevi

$$\lambda_i = i, \quad i = 1, \dots, 20.$$

Ovaj oblik polinoma — kao **produkt linearnih faktora**, je

- idealno **stabilan** na male perturbacije “polaznih” podataka,
- jer su ti podaci upravo **nultočke** polinoma!

Wilkinsonov polinom — razvijen po potencijama

Kad polinom P_{20} “razvijemo” po potencijama od λ , tj. zapišemo preko **koeficijenata** c_j , dobivamo

$$P_{20}(\lambda) = \lambda^{20} + c_{19}\lambda^{19} + \cdots + c_1\lambda + c_0,$$

s koeficijentima:

$$c_{19} = -(1 + 2 + \cdots + 19 + 20) = -210,$$

$$\vdots$$

$$c_0 = (-1) \cdot (-2) \cdots (-19) \cdot (-20) = 20!$$

Baš to je oblik kojeg bismo, na primjer, izračunali iz pripadne matrice. Poanta:

🔴 Ovdje se **nultočke** baš i “**ne vide**” odmah ...

Treba ih **izračunati!**

Egzaktni koeficijenti Wilkinsonovog polinoma

Točne vrijednosti koeficijenata c_j su

$c_0 =$	2432 90200 81766 40000	$c_{10} =$	1 30753 50105 40395
$c_1 =$	-8752 94803 67616 00000	$c_{11} =$	-13558 51828 99530
$c_2 =$	13803 75975 36407 04000	$c_{12} =$	1131 02769 95381
$c_3 =$	-12870 93124 51509 88800	$c_{13} =$	-75 61111 84500
$c_4 =$	8037 81182 26450 51776	$c_{14} =$	4 01717 71630
$c_5 =$	-3599 97951 79476 07200	$c_{15} =$	-16722 80820
$c_6 =$	1206 64780 37803 73360	$c_{16} =$	533 27946
$c_7 =$	-311 33364 31613 90640	$c_{17} =$	-12 56850
$c_8 =$	63 03081 20992 94896	$c_{18} =$	20615
$c_9 =$	-10 14229 98655 11450	$c_{19} =$	-210

Koeficijenti su “jedva” prikazivi u tipu **extended**, a sigurno nisu egzaktno prikazivi u manjim tipovima, poput **double**.

Mala perturbacija koeficijenta c_{19}

U polinomu P_{20} napravimo

• jednu jedinu perturbaciju veličine 2^{-23} u koeficijentu c_{19} , tako da dobijemo polinom

$$\tilde{P}_{20}(\lambda) = P_{20}(\lambda) - 2^{-23}\lambda^{19}.$$

Pripadna relativna perturbacija koeficijenta c_{19} je

• reda veličine 2^{-30} , odnosno, 10^{-9} .

Reklo bi se — zaista mala perturbacija!

Kako izgledaju nultočke tog perturbiranog polinoma \tilde{P}_{20} , tj.

• jesu li se i nultočke “malo” promijenile?

Nažalost, ne!

Nestabilnost nultočka Wilkinsonovog polinoma

Egzaktne nultočke polinoma \tilde{P}_{20} , na 9 decimala, su

1.00000 0000	6.00000 6944	10.09526 6145 \pm 0.64350 0904 i
2.00000 0000	6.99969 7234	11.79363 3881 \pm 1.65232 9728 i
3.00000 0000	8.00726 7603	13.99235 8137 \pm 2.51883 0070 i
4.00000 0000	8.91725 0249	16.73073 7466 \pm 2.81262 4894 i
4.99999 9928	20.84690 8101	19.50243 9400 \pm 1.94033 0347 i

Od 20 realnih nultočka polinoma P_{20} , dobili smo

- samo 10 realnih — prvih 9 i zadnja,
- i 5 parova kompleksnih, s vrlo “nezanemarivim” imaginarnim dijelovima.

Ni govora o “maloj” perturbaciji!

Svojstvene vrijednosti matrica — pouka

Zato se, u praksi, **svojstvene vrijednosti** matrice A

- **nikad** (ili gotovo nikad) **ne** računaju kao
- **nultočke** karakterističnog polinoma k_A .

Za taj problem postoji gomila **raznih** numeričkih metoda, ovisno o tipu matrice i raznim drugim stvarima.