

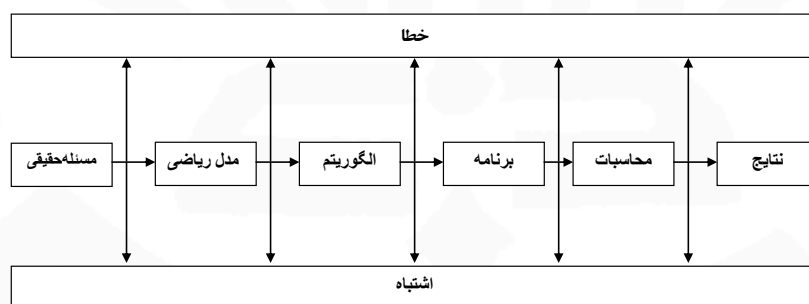
فصل ۲

خطاها

زمانی که به دست آوردن جواب واقعی یک مسئله غیرممکن است و یا مقرون به صرفه نیست، سعی می‌کنیم به کمک روش‌های عددی، یک جواب تقریبی برای مسئله پیدا کنیم. این فرایند منجر به تولید خطا می‌شود. در این فصل قصد داریم منابع تولید خطا و انواع خطا را شناسایی کرده و تا حدی از انتشار خطا جلوگیری کنیم.

۱.۲ منابع تولید خطا

بیشتر مواقع در عمل با یک مسئله حقیقی (فیزیکی) مواجه هستیم و بنا به دلایلی، جواب عددی یا تقریبی آن را جستجو می‌کنیم. مراحل یافتن جواب عددی چنین مسئله‌ای در روندنمای آمده در شکل ۱.۲ خلاصه می‌شود. این روندنما مکان‌های احتمالی بروز خطا و اشتباه را نیز نشان می‌دهد.



شکل ۱.۲: فرایند تولید جواب عددی (تقریبی)

تذکر ۱.۲ در مدل‌سازی مسائل باید از بروز خطاهای ذاتی تا حد ممکن جلوگیری کرد در حالی که پرهیز از خطاهای محاسباتی و کنترل آن‌ها، از وظایف متخصص آنالیز عددی است. تفاوت بین جواب واقعی و تقریبی از اشتباهات و خطاها ناشی می‌شود. اشتباه را می‌توان برطرف کرد ولی خطا ممکن است اجتناب‌ناپذیر باشد. به عنوان مثال قرار دادن 2323 به جای 2332 یک اشتباه است حال آن که عدد π بسط دهدهی نامختوم دارد و در وسایل محاسباتی باید یک بسط دهدهی مختوم برای آن در نظر گرفت که این موجب بروز خطا می‌شود.

مراحل مختلف این روندنما را در مثال بعد دنبال می‌کنیم.

مثال ۱.۲ (مسئله حقیقی) می‌خواهیم دوره تناوب حرکت نوسانی و متناوب یک آونگ ساده به جرم m و طول l را به دست آوریم. فرض کنید $\theta(t)$ جابجایی زاویه‌ای آونگ در زمان t باشد. صرف نظر از مقاومت هوا و اصطکاک در لولا، مدل این مسئله به صورت $ml \frac{d^2\theta}{dt^2} = -mg \sin \theta$ یا

$$\frac{d^2\theta}{dt^2} = -\frac{g}{l} \sin \theta, \quad (1.2)$$

در می‌آید که یک معادله دیفرانسیل غیرخطی است. با فرض کوچک بودن θ یعنی

$$\theta = 6^\circ \simeq 0.1047 \text{ rad} \rightarrow \sin \theta \simeq 0.1045, \quad \theta = 15^\circ \simeq 0.262 \text{ rad} \rightarrow \sin \theta \simeq 0.259,$$

می‌توان فرض کرد $\sin \theta \simeq \theta^{rad}$ و معادله دیفرانسیل را به صورت $\frac{d^2\theta}{dt^2} + \omega^2 \theta = 0$ نوشت که در آن $\omega^2 = \frac{g}{l}$. این معادله دیفرانسیل خطی جوابی به صورت $\theta(t) = A \sin \omega t + B \cos \omega t$ دارد که از آن نتیجه می‌گیریم $T_L = 2\pi \sqrt{\frac{l}{g}}$. از طرف دیگر با حل معادله دیفرانسیل غیرخطی (۱.۲) خواهیم داشت

$$T_N = 2\pi \sqrt{\frac{l}{g}} \left(1 + \frac{1^2}{2^2} \sin^2 \frac{\theta_m}{2} + \frac{1^2 \times 3^2}{2^2 \times 4^2} \sin^4 \frac{\theta_m}{2} + \dots \right),$$

که در آن θ_m جابجایی زاویه‌ای ماکزیمم آونگ است. جالب است توجه داشته باشیم که T_L اولین جمله سری T_N است. در ادامه یک الگوریتم برای این مدل ریاضی طراحی می‌کنیم.

الگوریتم ۱.۲ الگوریتم یافتن دوره تناوب حرکت یک آونگ ساده.

• ورودی: l , θ_m و مرز خطا یعنی ϵ

• خروجی: مقدار T_L , T_N و اختلاف آن‌ها

(۱) قرار دهید $g = 9.8$, $p = 3.14$ و $T_L = 2p \sqrt{\frac{l}{g}}$

(۲) قرار دهید $k = 1$, $M = \left(\frac{2k-1}{2k}\right)^2$, $T_k = T_L$, $C = T_L \times M \times \sin^{2k} \frac{\theta_m}{2}$ و $T_{k+1} = T_k + C$

(۳) تا زمانی که $|C| > \epsilon$ گام ۴ را تکرار کنید.

(۴) قرار دهید $k = k + 1$, $M = M \times \left(\frac{2k-1}{2k}\right)^2$, $T_k = T_{k+1}$ و $C = T_L \times M \times \sin^{2k} \frac{\theta_m}{2}$ و $T_{k+1} = T_k + C$

(۵) قرار دهید $T_N = T_{k+1}$ و T_L و $T_N - T_L$ را چاپ کنید.

حال برای الگوریتم ۱.۲ یک برنامه با استفاده از بسته نرم‌افزاری Mathematica می‌نویسیم. پس از اجرای برنامه ۱.۲ در محیط Mathematica محاسبات شروع شده و به ازای $\epsilon = 0.0001$ می‌توان نتایج گزارش شده در جدول ۱.۲ را تولید کرد. باید توجه داشت که برای $\theta_m < 15^\circ$ اختلاف T_N و T_L از 0.0005 بیشتر نیست. Δ

با توجه به روندنمای ارائه شده در شکل ۱.۲ و مثال ۱.۲، خطاها از نظر منابع تولید به صورت زیر تقسیم‌بندی می‌شوند.

۱. خطای ذاتی

• خطای مدل (ناشی از صرف نظرها، چشم‌پوشی‌ها و ساده‌سازی‌ها مانند فرض $\sin \theta \simeq \theta^{rad}$)

• خطای داده‌های مدل (ناشی از آزمایشات و اندازه‌گیری‌ها مانند g, l)

۲. خطای محاسباتی

- خطای نمایش اعداد (مانند $\pi = ۳/۱۴$)
- خطای اعمال ریاضی (به عنوان مثال $\frac{1}{g}$)
- خطای روش‌های (الگوریتم‌های) عددی (محاسباتی) (مانند خطای روش محاسبه $(\sqrt{\frac{l}{g}})$)

```

l = Input["طول آونگ را بر حسب متر وارد کنید"];
theta_m = Input["جایابی زاویه‌ای ماکزیمم را بر حسب درجه وارد کنید"];
epsilon = Input["مقدار مجاز خطا را وارد کنید"];
st = Sin[theta_m * Pi / ۳۶۰];
g = ۹/۸; TL = ۲ * pi * sqrt[l/g]; k = ۱; M = (frac{k-1}{k})^۲;
CC = TL * M * st^k; T1 = TL; T2 = T1 + CC;
While[CC > epsilon, {
    k++;
    M = (frac{k-1}{k})^۲;
    T1 = T2;
    CC = TL * M * st^k;
    T2 = T1 + CC;
}];
TN = T2;
Print["TL = ", TL, " TN = ", TN, " TN - TL = ", TN - TL];

```

برنامه ۱.۲: برنامه یافتن دوره تناوب حرکت یک آونگ ساده.

$l(cm)$	θ_m	T_L	T_N	$T_N - T_L$	$l(cm)$	θ_m	T_L	T_N	$T_N - T_L$
۱۰	۳	۰/۶۳۴۷۰	۰/۶۳۴۸۱	۰/۰۰۰۱۱	۱۰	۱۲	۰/۶۳۴۷۰	۰/۶۳۶۴۴	۰/۰۰۱۷۴
۲۰	۳	۰/۸۹۷۶۰	۰/۸۹۷۷۵	۰/۰۰۰۱۵	۲۰	۱۲	۰/۸۹۷۶۰	۰/۹۰۰۰۷	۰/۰۰۲۴۷
۳۰	۳	۱/۰۹۹۳۳	۱/۰۹۹۵۲	۰/۰۰۰۱۹	۳۰	۱۲	۱/۰۹۹۳۳	۱/۱۰۲۳۵	۰/۰۰۳۰۲
۱۰	۶	۰/۶۳۴۷۰	۰/۶۳۵۱۳	۰/۰۰۰۴۳	۱۰	۱۵	۰/۶۳۴۷۰	۰/۶۳۷۴۳	۰/۰۰۲۷۳
۲۰	۶	۰/۸۹۷۶۰	۰/۸۹۸۲۱	۰/۰۰۰۶۱	۲۰	۱۵	۰/۸۹۷۶۰	۰/۹۰۱۴۶	۰/۰۰۳۸۶
۳۰	۶	۱/۰۹۹۳۳	۱/۱۰۰۰۸	۰/۰۰۰۷۵	۳۰	۱۵	۱/۰۹۹۳۳	۱/۱۰۴۰۶	۰/۰۰۴۷۳
۱۰	۹	۰/۶۳۴۵۰	۰/۶۳۵۷۰	۰/۰۰۱۲۰	۱۰	۱۸	۰/۶۳۴۷۰	۰/۶۳۸۶۳	۰/۰۰۳۹۳
۲۰	۹	۰/۸۹۷۶۰	۰/۸۹۹۰۰	۰/۰۰۱۴۰	۲۰	۱۸	۰/۸۹۷۶۰	۰/۹۰۳۱۷	۰/۰۰۵۵۷
۳۰	۹	۱/۰۹۹۳۳	۱/۱۰۱۰۳	۰/۰۰۱۷۰	۳۰	۱۸	۱/۰۹۹۳۳	۱/۱۰۶۱۵	۰/۰۰۶۸۲

جدول ۱.۲: دوره تناوب آونگ ساده با طول‌های مختلف

تذکر ۲.۲ برای اطلاعات بیشتر در خصوص اثرات مخرب خطاها و اشتباهات به آدرس‌های زیر مراجعه کنید.

<https://www.iro.umontreal.ca/mignotte/IFT2425/Disasters.html>

<https://www.computerworld.com/article/3412197/top-software-failures-in-recent-history.html>

در اینجا فقط به دو مورد زیر اشاره می‌شود.

- عدم موفقیت موشک پاتریوت در جنگ خلیج فارس سال ۱۹۹۱ (۲۸ کشته و ۱۰۰ زخمی) به دلیل وقوع خطای گرد کردن در محاسبات مسیر
- شکست ماموریت موشک آریان ۵ فرانسه در سال ۱۹۹۶ (۵۰۰ میلیون دلار خسارت مادی) به دلیل وقوع پاریز^۱ در رایانه آن

۲.۲ نمایش اعداد

در این بخش به بررسی نمایش اعداد حقیقی می‌پردازیم. اثبات برخی از قضایا را می‌توان در سایر مراجع یافت.

قضیه ۱.۲ هر عدد حقیقی مثبت x نمایشی به صورت

$$\begin{aligned} x &= a_m \beta^m + a_{m-1} \beta^{m-1} + \dots + a_1 \beta^1 + a_0 \beta^0 + a_{-1} \beta^{-1} + a_{-2} \beta^{-2} + \dots \\ &= (a_m a_{m-1} \dots a_1 a_0 / a_{-1} a_{-2} \dots)_\beta, \end{aligned} \quad (2.2)$$

دارد که در آن $m \in \mathbb{Z}$ و $a_i \in \{0, 1, 2, \dots, \beta - 1\}$. برای منحصر به فرد بودن این نمایش لازم است که $a_m \neq 0$ و عدد صحیح j چنان وجود داشته باشد که $a_{j-1} = a_j = 0, \dots$ یا به ازای هر N به اندازه کافی بزرگ وجود داشته باشد $N \leq j$ به طوری که $a_{-j} \neq \beta - 1$.

رابطه (۲.۲) به نمایش (بسط) عدد x در مبنای β معروف است. اگر $\beta = 10$ اختیار شود به (۲.۲) نمایش (بسط) ده‌دهی (اعشاری) گویند که در زندگی روزمره با آن سرو کار داریم. در حالتی که $\beta = 2$ در نظر گرفته شود، (۲.۲) به نمایش دودویی (باینری) عدد x معروف بوده و این مبنا، اساس کار رایانه و وسایل دیجیتال است.

تذکر ۳.۲ شرط دوم منحصر به فرد بودن نمایش، لازم است چه در غیر این صورت به عنوان مثال می‌توان نوشت

$$\begin{aligned} 3,47999\dots &= 3,47\bar{9} = 3 \times 10^0 + 4 \times 10^{-1} + 7 \times 10^{-2} + 9 \times 10^{-3} + 9 \times 10^{-4} + \dots \\ &= 3,47 + \frac{9 \times 10^{-3}}{1 - 10^{-1}} = 3,47 + 0,1 = 3,57. \end{aligned}$$

یعنی بسط ده‌دهی بیان شده را می‌توان برای دو عدد $3,47999\dots$ و $3,57$ در نظر گرفت.

قضیه ۲.۲ اگر بسط ده‌دهی یک عدد مختوم یا نامختوم متناوب باشد، آن عدد گویا است.

نتیجه ۱.۲.۲ بسط دهدهی یک عدد گنگ، نامختوم نامتناوب است.

مثال ۲.۲ به نمایش اعداد زیر توجه کنید.

$$\frac{2}{3} = 0,666\dots = 0,6\bar{6} = (0,2)_3, \quad \frac{3}{8} = 0,375 = (0,11)_2, \quad \frac{1}{7} = 0,142857\dots = (0,1)_7,$$

$$\sqrt{2} = 1,414213\dots, \quad \pi = 3,141592\dots, \quad \exp(1) = e = 2,718281\dots$$

عدد $0,191919\dots$ گویا است ولی اعداد $0,1223334444\dots$ و $0,10200300040000\dots$ گنگ هستند.

△

هنگام کار با رایانه (ماشین حساب) اعداد را در مبنای ۱۰ وارد کرده و انتظار داریم نتایج (خروجی) نیز در همین مبنا نمایش داده شود ولی این وسایل با مبنای دیگری مانند ۲، ۸ و ۱۶ کار می کنند. بنابراین مسئله تغییر مبنا مطرح می شود.

پرسش ۱.۲ مبنای ۲ در مقایسه با مبنای ۱۶ چه معایب و مزایایی دارد؟

مثال ۳.۲ در این مثال یک عدد در مبنای ۲ را به مبنای ۱۰ می بریم.

$$(1011001/111001)_2 = 2^6 + 2^4 + 2^3 + 2^0 + 2^{-1} + 2^{-2} + 2^{-3} + 2^{-6}$$

$$= 64 + 16 + 8 + 1 + 0,5 + 0,25 + 0,125 + 0,015625 = 89,890625$$

△

بنابراین تغییر مبنا از مبنای ۲ به مبنای ۱۰ به کمک بسط به سادگی امکان پذیر است.

مثال ۴.۲ در این مثال یک عدد صحیح مثبت را از مبنای ۱۰ به مبنای ۲ می بریم.

$$23 = 11 \times 2 + 1, \quad 11 = 5 \times 2 + 1, \quad 5 = 2 \times 2 + 1, \quad 2 = 1 \times 2 + 0.$$

△

پس $23 = (10111)_2$. بنابراین از تقسیم های متوالی بر ۲ استفاده می کنیم.

اگر عدد داده شده x در مبنای ۱۰ در فاصله $(0, 1)$ باشد و عدد متناظر در مبنای ۲ به صورت $(0, b_1 b_2 b_3 \dots)_2$ فرض شود آنگاه

$$x = b_1 \times 2^{-1} + b_2 \times 2^{-2} + b_3 \times 2^{-3} + \dots,$$

و در نتیجه $2x = b_1 + b_2 \times 2^{-1} + b_3 \times 2^{-2} + \dots$ و با قرار دادن $y = b_2 \times 2^{-1} + b_3 \times 2^{-2} + \dots$ خواهیم داشت

$$0 \leq y < 1 \times 2^{-1} + 1 \times 2^{-2} + \dots = \frac{2^{-1}}{1 - 2^{-1}} = 1.$$

پس $[y] = 0$. از طرف دیگر داریم $2x = b_1 + y$ و بلافاصله از $[2x] = [b_1 + y]$ نتیجه می شود $b_1 = [2x]$ زیرا b_1 عددی صحیح است (b_1 یا صفر است یا یک). با قرار دادن $x = y$ و تکرار این روند می توان b_2, b_3 و سایر رقم ها را به دست آورد. شرط توقف این روند تکراری زمانی است که x صفر شود یا تناوب به وجود آید و یا ممکن است بسط نامختوم نامتناوب باشد و پس از چند تکرار مجبور شویم روند را متوقف کنیم.

i	x	$2x$	b_i
۱	۰٫۵۷۸۱۲۵	۱٫۱۵۶۲۵	۱
۲	۰٫۱۵۶۲۵	۰٫۳۱۲۵	۰
۳	۰٫۳۱۲۵	۰٫۶۲۵	۰
۴	۰٫۶۲۵	۱٫۲۵	۱
۵	۰٫۲۵	۰٫۵	۰
۶	۰٫۵	۱٫۰	۱
	۰		

i	x	$2x$	b_i
۱	۰٫۱	۰٫۲	۰
۲	۰٫۲	۰٫۴	۰
۳	۰٫۴	۰٫۸	۰
۴	۰٫۸	۱٫۶	۱
۵	۰٫۶	۱٫۲	۱
۶	۰٫۲	۰٫۴	۰
	۰٫۴		

جدول ۲.۲: نمایش دودویی اعداد کوچک‌تر از واحد

مثال ۵.۲ در این مثال این روند را برای حالت‌هایی دنبال می‌کنیم که عدد نمایشی مختوم یا نامختوم متناوب دارد. با توجه به جدول ۲.۲ می‌توان نوشت

$$۰٫۵۷۸۱۲۵ = (۰٫۱۰۰۱۰۱)۲, \quad ۰٫۱ = (۰٫۰۰۰۱۱)۲.$$

با توجه به نتایج این مثال، نمایش یک عدد گویا در هر مبنایی یا مختوم است یا نامختوم متناوب [۱]. Δ

تمرین ۱.۲ برنامه‌ای بنویسید که عدد حقیقی x و عدد طبیعی n را از ورودی گرفته و نمایش دودویی عدد x را تولید کند و اگر نمایش نامختوم نامتناوب است حداکثر تا n رقم نمایش، چاپ شود.

۳.۲ نمایش اعداد در رایانه

برای نمایش اعداد در ماشین، ابتدا نمایشی به نام ممیز ثابت^۲ در نظر گرفته شد که در آن هر عدد حقیقی x به صورت زیر نمایش داده می‌شود

$$x = \pm(a_{n-1} \dots a_1 a_0 / a_{-1} a_{-2} \dots a_{-m})_{\beta},$$

که در آن n و m اعداد مشخص و ثابتی هستند. در اصل در این نمایش مکان ممیز مشخص و ثابت است. برای نمایش اعداد بسیار بزرگ (کوچک) در این نمایش با مشکل مواجه می‌شویم و در نتیجه این نمایش برای محاسبات علمی مناسب نیست ولی برای بسیاری از کاربردها (مانند حسابداری) این نمایش سودمند است و هنوز هم ماشین‌هایی بر این اساس ساخته می‌شوند.

یک روش جدید و متداول برای نمایش اعداد در رایانه، نمایش ممیز (نقطه) شناور (سیار)^۳ است که از بدو پیدایش مورد توجه سازندگان سخت‌افزار رایانه قرار گرفته است و چنین به نظر می‌رسد که تا حدودی به طور سلیقه‌ای با آن برخورد شده است. آنچه در ادامه درباره این نمایش خواهد آمد بر اساس استاندارد IEEE^۴ وضع کرده است.

^۲ Fixed point

^۳ Floating point

^۴ IEEE standard 754-1985

۱.۳.۲ نمایش ۶۴-بیتی ممیز شناور

این نمایش پیش از این به دقت دو برابر (مضاعف)^۵ معروف بوده و متناظر با نوع double در زبان C است. در این نمایش برای ذخیره هر عدد در مبنای ۲، ابتدا یک ساختار به طول ۶۴ بیت در نظر گرفته می‌شود. اولین بیت به بیت علامت معروف است و با s نمایش داده می‌شود و بلافاصله بعد از آن ۱۱ بیت برای مشخصه^۶ در نظر گرفته می‌شود و با c نمایش داده می‌شود و ۵۲ بیت باقی‌مانده نیز برای ماتیس کنار گذاشته می‌شود و آن را با f نشان می‌دهند. سپس برای هر عدد نمایشی به صورت $(-1)^s \times 2^e \times (1 + \circ/f)_2$ در نظر گرفته می‌شود که در آن $e = c - 1023$ (نما) است. محدودیت $2^{11} - 1 = 2047 = e$ موجب می‌شود که $0 \leq c \leq (10001)_2 = 17$ و $1023 \leq e \leq 1024$ از طرف دیگر محدودیت $0 \leq f \leq (\circ/10001)_2$ باعث می‌گردد که $2 < (1/10001)_2 \leq (1 + \circ/f)_2 \leq 1$. بنابراین اگر کوچک‌ترین و بزرگ‌ترین عدد مثبت قابل نمایش را به ترتیب با $+mN$ و $+MN$ نشان دهیم آن‌گاه

$$+mN = 2^{-1022} \times 1/0 \simeq 2/225074 \times 10^{-308}, \quad +MN = 2^{1024} \times 1/0 \simeq 1/797693 \times 10^{308}.$$

همچنین برای نمایش صفر و بی‌نهایت می‌توان از قراردادهای زیر استفاده کرد

$$+0 = (1/10001)_2 \times 2^{-1023}, \quad +\infty = (1/00001)_2 \times 2^{1024}.$$

تمرین ۲.۲ نمایش $-mN$ ، $-MN$ ، -0 و $-\infty$ را بنویسید.

تعریف ۱.۲ در نمایش اعداد ماشینی، کوچک‌ترین عدد مثبت ماشینی که اگر به ۱ اضافه شود عددی بزرگ‌تر از ۱ به دست می‌آید به اپسیلون ماشینی^۷ معروف است و با eps نمایش داده می‌شود. نصف eps به رُند واحد^۸ یا واحد گرد کردن^۹ یا دقت ماشینی^{۱۰} معروف است.

چون در نمایش ۶۴-بیتی بعد از ۱ عدد $(1/00001)_2$ قرار می‌گیرد، پس

$$\text{eps} = (1/00001)_2 - 1 = (\circ/00001)_2 = 2^{-52} \simeq 2/220446 \times 10^{-16}.$$

در ادامه بزرگ‌ترین عدد صحیح مثبت M را تعیین می‌کنیم که هر عدد صحیح x با شرط $0 < x \leq M$ را بتوان به طور دقیق نمایش داد. با توجه به موارد بیان‌شده، تمام اعداد صحیح نامنفی که بزرگ‌تر از $2^{52} \times (1/10001)_2 = 2^{52} - 1$ نباشند به طور دقیق قابل نمایش هستند و به علاوه 2^{52} نیز به صورت $(1/0)_2 \times 2^{52}$ قابل نمایش است. اما تعداد ارقام در ماتیس جهت نمایش $2^{52} + 1$ کافی نیست (۵۳ رقم در ماتیس لازم است). بنابراین $M = 2^{52} \simeq 9/007199 \times 10^{15}$ و در نتیجه تمام اعداد صحیح ۱۵ رقمی و بسیاری از اعداد ۱۶ رقمی در این نمایش به طور دقیق قابل نمایش هستند

Double precision^۵
Characteristic^۶
Machine epsilon^۷
Unit round^۸
Roundoff unit^۹
Machine precision^{۱۰}

به بیان دیگر نمایش ۶۴-بیتی ۱۵ الی ۱۶ رقم دقت دارد. بنابراین اعدادی که بیشتر از ۱۶ رقم داشته باشند به طور دقیق قابل‌نمایش نیستند.

تذکر ۴.۲ مجموعه اعداد با ممیز شناور را با \mathbb{F} نمایش می‌دهیم. محور اعداد با ممیز شناور بر خلاف محور اعداد حقیقی، متناهی و گسسته است و بر خلاف آنچه به نظر می‌رسد، نقاط روی این محور هم‌فاصله نیستند. اگر در هنگام انجام محاسبات، عددی در فاصله $(+0, -0)$ تولید شود پیام پاریز^{۱۱} و اگر عدد تولید شده از MN بزرگ‌تر یا از $-MN$ کوچک‌تر باشد پیام سرریز^{۱۲} صادر می‌شود. برنامه uoflow.nb را ببینید.

تذکر ۵.۲ بعضی از نرم‌افزارها مانند Mathematica، محدودیت‌های سخت‌افزار را از طریق برنامه‌های نرم‌افزاری برطرف می‌کنند و به اصطلاح دقت را بالا می‌برند که ممکن است تا حدودی سرعت محاسبات کاهش یابد.

تمرین ۳.۲ در نمایش ۳۲-بیتی (دقت معمولی^{۱۳} متناظر با نوع float در زبان C) هشت بیت برای مشخصه منظور می‌شود. تمام کمیت‌های معرفی شده در این بخش را برای این نمایش تعیین کنید.

۲.۳.۲ اعداد ماشینی

به اعداد \mathbb{F} اعداد ماشینی نیز گفته می‌شود و برای سادگی، نمایشی به صورت $\pm a \times 10^b$ برای آن‌ها در نظر گرفته می‌شود. به عبارتی یک عدد ماشینی با زوج مرتب (a, b) نظیر می‌شود. در اینجا b توان (نما) است و به زیرمجموعه‌ای از \mathbb{Z} تعلق دارد. به بیان دقیق‌تر $L \leq b \leq U$ که در آن L و U با توجه به محدودیت‌های سخت‌افزاری و نرم‌افزاری تعیین می‌شوند. نمایش فرض می‌شود $a = 0.d_1 \dots d_k$ بیانگر مانتیس (جزء کسری) است و به ازای $i = 1, \dots, k$ داریم $d_i \in \{0, 1, \dots, 9\}$ و برای یکتایی نمایش فرض می‌شود $d_1 \neq 0$. بنابراین $0.1 \leq a < 1$. این نمایش، به نمایش ممیز شناور ده‌دهی نرمال شده نیز معروف است و چنین اعدادی را، اعداد ماشینی ده‌دهی k -رقمی می‌نامند. ایده چنین نمایشی، از نمایش علمی نرمال شده اعداد ناشی شده است. در نمایش علمی نرمال شده، هر عدد حقیقی مخالف صفر x را می‌توان به صورت $x = \pm a \times 10^b$ نمایش داد که در آن $b \in \mathbb{Z}$ و $a = 0.d_1 \dots d_k d_{k+1} \dots$ که در آن به ازای $i = 1, 2, \dots$ داریم $d_i \in \{0, 1, \dots, 9\}$ و $d_1 \neq 0$. واضح است که برای نمایش x به صورت ممیز شناور ده‌دهی نرمال شده باید k رقم از مانتیس آن را حفظ کرده و رقم‌های اضافی را به کمک یکی از روش‌های زیر کنار گذاشت

روش قطع کردن (برش)^{۱۴}، روش گرد کردن معمولی^{۱۵}، روش گرد کردن به زوج^{۱۶}.

در روش قطع کردن، k رقم از مانتیس حفظ و بقیه کنار گذاشته می‌شود در حالی که در روش گرد کردن معمولی، ابتدا روش قطع کردن اعمال شده، سپس اگر $d_{k+1} \geq 5$ یک واحد به d_k اضافه می‌گردد. اما در روش گرد کردن به زوج، ابتدا روش قطع کردن اعمال شده و در هر یک از حالت‌های زیر یک واحد به d_k اضافه می‌گردد

$$d_{k+1} > 5 \bullet$$

$$d_{k+1} = 5 \bullet \text{ و رقم مخالف صفری در سمت راست } d_{k+1} \text{ مشاهده شود}$$

^{۱۱} Underflow

^{۱۲} Overflow

^{۱۳} Single precision

^{۱۴} Chopping

^{۱۵} Rounding

^{۱۶} Rounding to even

x	۲۸,۶۴۲۴	۰,۰۰۵۷۶۷۱	۴,۹۸۵۰	-۲۱۷۵,۳۴۵۱۲
$\tilde{x}(3S)$	۲۸,۶	۰,۰۰۵۷۷	۴,۹۹	-۲۱۸۰
$\tilde{x}(3D)$	۲۸,۶۴۲	۰,۰۰۰۶	۴,۹۸۵	-۲۱۷۵,۳۴۵
$\hat{x}(3S)$	۲۸,۶	۰,۰۰۵۷۶	۴,۹۸	-۲۱۷۰
$\hat{x}(3D)$	۲۸,۶۴۲	۰,۰۰۰۵	۴,۹۸۵	-۲۱۷۵,۳۴۵

جدول ۳.۲: مثال‌هایی از گرد کردن و قطع کردن معمولی

• $d_{k+1} = 5$ و رقم مخالف صفری در سمت راست d_{k+1} مشاهده نشود و d_k فرد باشد.

قضیه ۳.۲ اگر $x = 0/d_1 \dots d_k d_{k+1} \dots \times 10^n$ یک عدد حقیقی ناصفر در نمایش علمی نرمال شده باشد آن‌گاه $c(x) = 0/d_1 \dots d_k \times 10^n$ و $r(x) = 0/d_1 \dots d_k \times 10^m$ (یا $m = n$ یا $m = n + 1$) اعداد ماشینی k -رقمی متناظر با x هستند که به ترتیب از روش قطع کردن و گرد کردن معمولی به دست می‌آیند و خواهیم داشت

$$|x - c(x)| \leq 1 \times 10^{n-k}, \quad |x - r(x)| \leq 5 \times 10^{n-k-1} = 0,5 \times 10^{n-k}.$$

پرسش ۲.۲ تفاوت 47^{km} با 47000^m یا تفاوت $3,7$ با $3,70$ یا $3,700$ در چیست؟

تعریف ۲.۲ منظور از ارقام بامعنای یک عدد مخالف صفر، ارقام مخالف صفر، صفرهای بین دو رقم مخالف صفر و صفرهایی است که در سمت راست عدد به منظور نشان دادن نوعی دقت قرار داده می‌شوند (تمام ارقام مانع در نمایش علمی نرمال شده).

مثال ۶.۲ عدد $0,0007045000$ حداقل ۴ رقم بامعنا و حداکثر ۷ رقم بامعنا دارد. \triangle

قرارداد ۱.۲ rD یعنی r رقم اعشار و rS یعنی r رقم بامعنا^{۱۷}.

مثال ۷.۲ در جدول ۳.۲، چند عدد و مقدارهای تقریبی متناظر با آن‌ها با روش‌های قطع کردن (\tilde{x}) و گرد کردن معمولی (\hat{x}) با دقت $3S$ و $3D$ داده شده است. \triangle

تذکر ۶.۲ از این به بعد، همانند بسیاری از وسایل محاسباتی از گرد کردن معمولی استفاده می‌کنیم.

پرسش ۳.۲ در نظر بگیرید \hat{x} تقریبی از x باشد. آیا هر چه تعداد ارقام بامعنای \hat{x} بیشتر باشد، \hat{x} تقریب بهتری است؟

تعریف ۳.۲ فرض کنید $\hat{x} = d_m \times 10^m + \dots + d_1 \times 10^1 + d_0 \times 10^0 + d_{-1} \times 10^{-1} + \dots + d_{-k} \times 10^{-k}$ که در آن $d_m \neq 0$ تقریبی برای $x > 0$ باشد (در اینجا m نشان‌دهنده با ارزش‌ترین مکان است). تعداد ارقام بامعنای درست \hat{x} نسبت به x ، بزرگ‌ترین عدد طبیعی n است که در نابرابری‌های زیر صدق کند.

$$|x - \hat{x}| \leq 5 \times 10^{m-n}, \quad n \leq m + k + 1$$

به وضوح هر چه تعداد ارقام بامعنای درست \hat{x} نسبت به x بیشتر باشد، \hat{x} تقریب بهتری خواهد بود.

^{۱۷} D حرف اول کلمه Decimal و S حرف اول کلمه Significant

مثال ۸.۲ از اعداد $\hat{x} = ۹۹/۹۶$ و $\tilde{x} = ۱۰۰/۷$ کدام یک تقریب بهتری برای $x = ۱۰۰$ است؟

$$\hat{x} = ۹۹/۹۶ \rightarrow m = ۱, |x - \hat{x}| = ۰/۰۴ \leq ۵ \times ۱۰^{۱-n} \rightarrow n = ۳$$

$$\tilde{x} = ۱۰۰/۷ \rightarrow m = ۲, |x - \tilde{x}| = ۰/۷ \leq ۵ \times ۱۰^{۲-n} \rightarrow n = ۲$$

بنابراین \hat{x} تقریب بهتری برای x است زیرا تعداد ارقام بامعنای درست بیشتری دارد. \triangle

مثال ۹.۲ تعداد ارقام بامعنای درست $\hat{x} = ۱۰۰/۳۱$ را نسبت به $x = ۱۰۰/۳۱۰۴$ مشخص کنید.

$$\hat{x} = ۱۰۰/۳۱ \rightarrow m = ۲, |x - \hat{x}| = ۰/۰۰۰۰۴ \leq ۵ \times ۱۰^{۲-n} \rightarrow n = ۶$$

و چون \hat{x} فقط پنج رقم بامعنا دارد پس $n = ۵$. \triangle

۴.۲ انواع خطا

تعریف ۴.۲ اگر \hat{x} تقریبی از x باشد $\Delta x = |x - \hat{x}|$ خطای مطلق \hat{x} نسبت به x نامیده می‌شود. Δx یکتا بوده و در عمل بیشتر مواقع قابل تعیین نیست و به جای آن از هر عدد b_x استفاده می‌شود که از Δx کمتر نباشد. b_x یکتا نیست و به آن کران خطای مطلق گویند. بنابراین $\Delta x \leq b_x$ و در نتیجه $\hat{x} - b_x \leq x \leq \hat{x} + b_x$ و یا می‌نویسیم $x = \hat{x} \pm b_x$.

مثال ۱۰.۲ برای $\hat{x} = ۱/۷۳۲$ به عنوان تقریبی از $x = \sqrt{۳}$ می‌توان نوشت

$$\Delta x = \left| \sqrt{۳} - ۱/۷۳۲ \right| = ۱/۷۳۲۰۵۰۸۰۷۵\dots - ۱/۷۳۲ = ۰/۰۰۰۰۵۰۸۰۷۵\dots$$

از طرفی می‌دانیم $۱/۷۳۲۱ < \sqrt{۳} < ۱/۷۳۲۰$ بنابراین $۰/۰۰۰۰۱ < \sqrt{۳} - ۱/۷۳۲ < ۰/۰۰۰۰۱$ پس $b_x = ۰/۰۰۰۰۱$ که معیاری برای نزدیکی $۱/۷۳۲$ به $\sqrt{۳}$ است. \triangle

پرسش ۴.۲ آیا خطای مطلق معیار مناسبی برای مقایسه خطاها است؟

پاسخ. خیر. به عنوان مثال خطای مطلق یک صندوق دار بانک، تایپیست و دروازه بان را در نظر بگیرید.

تعریف ۵.۲ اگر $\hat{x} \neq ۰$ تقریبی از $x \neq ۰$ باشد $\delta x = \frac{|x - \hat{x}|}{|x|} = \frac{\Delta x}{|x|}$ خطای نسبی \hat{x} نسبت به x نامیده می‌شود و یکتا بوده و بیشتر مواقع قابل تعیین نیست و از کران خطای نسبی استفاده می‌شود. $۱۰۰ \times \delta x$ به درصد خطا معروف است.

قضیه ۴.۲ اگر \hat{x} تقریبی از x باشد و b_x یک کران خطای مطلق برای این تقریب باشد آنگاه

$$\delta x \leq \frac{b_x}{|\hat{x}|}$$

به علاوه اگر b_x نسبت به $|\hat{x}|$ خیلی کوچک باشد می‌توان نوشت $\delta x \leq \frac{b_x}{|\hat{x}|}$.

برهان. بنابر تعریف b_x و با استفاده از خواص نابرابری‌ها واضح است. □

مثال ۱۱.۲ برای $\hat{x} = ۱,۷۳۲$ به عنوان تقریبی از $x = \sqrt{۳}$ می‌توان نوشت

$$\delta x = \frac{|\sqrt{۳} - ۱,۷۳۲|}{\sqrt{۳}} = \frac{۱,۷۳۲۰۵۰۸۰۷۵\dots - ۱,۷۳۲}{۱,۷۳۲۰۵۰۸۰۷۵\dots} = \frac{۰,۰۰۰۰۵۰۸۰۷۵\dots}{۱,۷۳۲۰۵۰۸۰۷۵\dots}$$

پس $\delta x = ۰,۰۰۰۰۰۲۹۳۳۳۷\dots < ۰,۰۰۰۰۰۳$ و با توجه به قضیه ۴.۲ و $\Delta x = ۰,۰۰۰۰۱$ خواهیم داشت

$$\delta x \leq \frac{۰,۰۰۰۰۱}{۱,۷۳۲ - ۰,۰۰۰۰۱} = \frac{۰,۰۰۰۰۱}{۱,۷۳۱۹} = ۰,۰۰۰۰۰۵۷۷۴۰۰۵\dots < ۰,۰۰۰۰۰۶$$

و یا

$$\delta x \leq \frac{۰,۰۰۰۰۱}{۱,۷۳۲} = ۰,۰۰۰۰۰۵۷۷۳۶۷۲\dots < ۰,۰۰۰۰۰۶.$$

△

قضیه ۵.۲ اگر \hat{x} تقریبی از x با n رقم بامعنای درست باشد و $\hat{y} = ۱۰^t \times \hat{x}$ و $y = ۱۰^t \times x$ که در آن t عددی صحیح است آن‌گاه \hat{y} نیز تقریبی از y با n رقم بامعنای درست بوده و خطای نسبی \hat{x} و \hat{y} برابر است.

قضیه ۶.۲ اگر \hat{x} گردشده x تا n رقم بامعنا باشد آن‌گاه \hat{x} دارای n رقم بامعنای درست است.

ارتباط خطای نسبی با تعداد ارقام بامعنای درست در قضیه بعدی مشخص می‌شود.

قضیه ۷.۲ اگر \hat{x} دارای n رقم بامعنای درست باشد، آن‌گاه $\delta x < ۵ \times ۱۰^{-n}$ به شرط آن که ارقام بامعنای درست \hat{x} یک رقم ۱ و $n-۱$ رقم صفر جلوی آن نباشد. اگر $\delta x \leq ۰,۵ \times ۱۰^{-n}$ آن‌گاه \hat{x} حداقل n رقم بامعنای درست دارد.

مثال ۱۲.۲ تقریبی از $\sqrt{۳} = ۱,۷۳۲۰۵\dots$ ارائه دهید که خطای نسبی آن از $۱۰^{-۴}$ کمتر باشد.

بنابر قضیه ۷.۲ اگر \hat{x} تقریبی از $\sqrt{۳}$ باشد که ۵ رقم بامعنای درست داشته باشد آن‌گاه $۱۰^{-۴} < ۵ \times ۱۰^{-۵} < \delta x$. از

این‌رو، با توجه به قضیه ۶.۲ کافی است \hat{x} گردشده $\sqrt{۳}$ تا ۵ رقم بامعنا باشد، یعنی $\hat{x} = ۱,۷۳۲۱$. △

۵.۲ خطای محاسبات

فرض کنید $z = f(x_1, \dots, x_n)$ تابعی باشد که می‌خواهیم آن را در نقطه (x_1, \dots, x_n) ارزیابی کنیم و $x_i - \hat{x}_i = \Delta x_i$ که در آن \hat{x}_i مقدار تقریبی x_i است. پس می‌توان نوشت

$$z = f(x_1, \dots, x_n) = f(\hat{x}_1 + \Delta x_1, \dots, \hat{x}_n + \Delta x_n).$$

بنابر بسط تیلور توابع n متغیره داریم

$$z = f(\hat{x}_1, \dots, \hat{x}_n) + \left(\Delta x_1 \frac{\partial f}{\partial x_1} + \dots + \Delta x_n \frac{\partial f}{\partial x_n} \right) (\hat{x}_1, \dots, \hat{x}_n) + R,$$

که در آن R جمله خطا بوده و شامل حاصل ضرب‌ها و توان‌های Δx_i ها است و چون Δx_i ها کوچک هستند از R چشم‌پوشی کرده، خواهیم داشت

$$f(x_1, \dots, x_n) - f(\hat{x}_1, \dots, \hat{x}_n) \simeq \left(\Delta x_1 \frac{\partial f}{\partial x_1} + \dots + \Delta x_n \frac{\partial f}{\partial x_n} \right) (\hat{x}_1, \dots, \hat{x}_n),$$

و اگر $\hat{z} = f(\hat{x}_1, \dots, \hat{x}_n)$ آن‌گاه

$$\Delta z = |z - \hat{z}| \simeq \left| \left(\Delta x_1 \frac{\partial f}{\partial x_1} + \dots + \Delta x_n \frac{\partial f}{\partial x_n} \right) (\hat{x}_1, \dots, \hat{x}_n) \right|,$$

و بلافاصله داریم

$$\delta z = \frac{\Delta z}{|z|} \simeq \left| \frac{\left(\Delta x_1 \frac{\partial f}{\partial x_1} + \dots + \Delta x_n \frac{\partial f}{\partial x_n} \right) (\hat{x}_1, \dots, \hat{x}_n)}{f(\hat{x}_1, \dots, \hat{x}_n)} \right|.$$

تذکر ۷.۲ اگر می‌خواهیم جوابی از یک مسئله را با دقت rD به دست آوریم، باید محاسبات میانی را با دقت $(r+1)D$ و حتی بیشتر انجام دهیم و نتیجه نهایی را با دقت rD گرد کنیم. اگر محاسبات میانی با دقت rD انجام شود، به دلیل اثرات خطای گرد کردن اعتباری به رقم (ارقام) سمت راست جواب نیست. به طور مشابه، هنگام کار با دقت rS باید ملاحظاتی را در نظر گرفت.

مثال ۱۳.۲ (مستقیم) یک استوانه به شعاع قاعده $\frac{4}{3}$ و ارتفاع $\sqrt{2}$ را در نظر بگیرید. اگر شعاع و ارتفاع استوانه و عدد π را با دقت $4D$ وارد محاسبات کنیم، حجم این استوانه با چه خطایی به دست می‌آید؟ می‌دانیم حجم یک استوانه از قاعده $\pi r^2 h$ تعیین می‌شود که در آن r شعاع قاعده و h ارتفاع استوانه است. اگر تعریف کنیم $z = V(p, r, h) = \pi r^2 h$ باید V به ازای $p = \pi$ ، $r = \frac{4}{3}$ و $h = \sqrt{2}$ ارزیابی شود. محاسبات میانی را با دقت $4D$ دنبال می‌کنیم (ماشین حساب را به صورت MODE FIX 4 تنظیم می‌کنیم). بنابراین

$$p = \pi \rightarrow \hat{p} = 3,1416, \quad r = \frac{4}{3} \rightarrow \hat{r} = 1,3333, \quad h = \sqrt{2} \rightarrow \hat{h} = 1,4142.$$

به طوری که $\Delta p, \Delta r, \Delta h \leq 0,5 \times 10^{-4}$. بنابراین

$$\hat{z} = V(\hat{p}, \hat{r}, \hat{h}) = 3,1416 \times 1,3333^2 \times 1,4142 = 7,8980.$$

از طرفی

$$\Delta z \simeq \Delta p \hat{r}^2 \hat{h} + 2 \Delta r \hat{p} \hat{r} \hat{h} + \Delta h \hat{p} \hat{r}^2 \leq (\hat{r}^2 \hat{h} + 2 \hat{p} \hat{r} \hat{h} + \hat{p} \hat{r}^2) \times 0,5 \times 10^{-4} \simeq 9,9731 \times 10^{-4}.$$

در نتیجه حجم استوانه با دقت $3D$ برابر است با $7,898$ و حداکثر خطای مرکب شده عبارت است از

$$\Delta z + 0,5 \times 10^{-4} \lesssim 0,0010.$$

به کمک یک ماشین حساب ۱۰ رقمی نتیجه $z = ۷/۸۹۸۴۵۸۵۵۵$ به دست می آید و بنابراین

$$e(z) = |z - \hat{z}| = ۴/۵۸۵۵۵۴ \times ۱۰^{-۴} < ۰/۰۰۰۱۰.$$

△

تمرین ۴.۲ (معکوس) یک استوانه به شعاع قاعده $\frac{۴}{۳}$ و ارتفاع $\sqrt{۲}$ در نظر بگیرید. اگر بخواهیم حجم این استوانه را با دقت ۳D به دست آوریم، شعاع قاعده و ارتفاع استوانه و حتی عدد π را با چه دقتی وارد محاسبات کنیم؟

مثال ۱۴.۲ (معکوس) اعداد $x = \sqrt{۵}$ و $y = \frac{\pi}{۱۱}$ را با چه دقتی در نظر بگیریم تا مقدار $۶x^۲(\ln x + \sin ۲y)$ با دقت ۲D حساب شود.

اگر فرض کنیم $z = f(x, y) = ۶x^۲(\ln x + \sin ۲y)$ آن گاه

$$\frac{\partial f}{\partial x} = ۱۲x(\ln x + \sin ۲y) + ۶x, \quad \frac{\partial f}{\partial y} = ۱۲x^۲ \cos ۲y,$$

و در نتیجه با فرض $x = ۲/۲$ و $y = ۰/۳$ و x و y را با دقت ۱S یا ۲S در نظر می گیریم داریم

$$\Delta z \simeq \left| \Delta x \frac{\partial f(۲/۲, ۰/۳)}{\partial x} + \Delta y \frac{\partial f(۲/۲, ۰/۳)}{\partial y} \right| \simeq |۴۸/۹ \Delta x + ۴۷/۹ \Delta y|,$$

و برای برقراری نابرابری $\Delta z \leq ۰/۵ \times ۱۰^{-۲}$ باید داشته باشیم $|۴۸/۹ \Delta x + ۴۷/۹ \Delta y| \leq ۰/۵ \times ۱۰^{-۲}$. یک جواب این نامعادله عبارت است از

$$\Delta x \leq ۰/۵ \times ۱۰^{-۴}, \quad \Delta y \leq ۰/۵ \times ۱۰^{-۴}.$$

△

بنابراین برای رسیدن به نتیجه مطلوب، می توان x و y را با دقت ۴D در نظر گرفت.

۱.۵.۲ خطای اعمال ریاضی

هنگام کار با اعداد ماشینی (ممیز شناور) باید توجه داشت که بعضی از اصول میدان \mathbb{R} نظیر شرکت پذیری، یکتایی عضو خنثی و غیره برقرار نیست. به طور کلی، عبارت هایی که از نظر ریاضی معادل هستند ممکن است از نظر محاسباتی معادل نباشند. علت اصلی بروز این مشکلات، خطای گرد کردن است.

تعریف ۶.۲ فرض کنید x و y دو عدد حقیقی، \hat{x} و \hat{y} تقریب هایی از آن ها و \otimes بیانگر یک عمل دوتایی باشد. متناظر با $x \otimes y$ در ماشین عمل $\hat{x} \otimes \hat{y}$ انجام می شود و می توان نوشت

$$|x \otimes y - \hat{x} \otimes \hat{y}| = |(x \otimes y - \hat{x} \otimes \hat{y}) + (\hat{x} \otimes \hat{y} - \hat{x} \otimes \hat{y})| \leq |x \otimes y - \hat{x} \otimes \hat{y}| + |\hat{x} \otimes \hat{y} - \hat{x} \otimes \hat{y}|.$$

بنابراین خطا دو بخش دارد: خطای منتشر شده یا انباشته شده (جمله اول) و خطای تولید شده یا گرد کردن (جمله دوم).

قضیه ۸.۲ اگر \hat{x} و \hat{y} تقریب هایی از x و y بوده و همه این اعداد مثبت باشند، آن گاه

$$1. \delta(x \pm y) \leq \frac{x}{|x \pm y|} \delta x + \frac{y}{|x \pm y|} \delta y \text{ و } \delta(x + y) \leq \max\{\delta x, \delta y\}, \Delta(x \pm y) \leq \Delta x + \Delta y$$

$$2. \delta(xy) \leq \delta x + \delta y, \Delta(xy) \leq y\Delta x + x\Delta y$$

$$3. \delta\left(\frac{x}{y}\right) \leq \delta x + \delta y, \Delta\left(\frac{x}{y}\right) \leq \frac{y\Delta x + x\Delta y}{y^2}$$

□ برهان. به عنوان تمرین.

مثال ۱۵.۲ \hat{x} و \hat{y} هر یک n رقم بامعناى درست دارند. حداقل تعداد ارقام بامعناى درست $\hat{x} + \hat{y}$ و $\hat{x}\hat{y}$ را تعیین کنید. بنابر قضیه ۷.۲، $\delta x < 5 \times 10^{-n}$ و $\delta y < 5 \times 10^{-n}$. هم‌چنین بنابر قضیه ۸.۲ می‌توان نوشت

$$\delta(x + y) \leq \max\{\delta x, \delta y\} \leq 5 \times 10^{-n} = 0.5 \times 10^{-(n-1)}$$

پس $\hat{x} + \hat{y}$ حداقل $n - 1$ رقم بامعناى درست دارد. به علاوه می‌توان نوشت

$$\delta(xy) \leq \delta x + \delta y < 5 \times 10^{-n} + 5 \times 10^{-n} = 10 \times 10^{-n} = 10^{-(n-2)-1} < 0.5 \times 10^{-(n-2)}$$

△ بنابراین $\hat{x}\hat{y}$ دست کم $n - 2$ رقم بامعناى درست دارد.

تذکر ۸.۲ با توجه به خطای نسبی $\frac{\Delta(x-y)}{|x-y|} \simeq \delta(x-y)$ واضح است که اگر x و y دو عدد هم‌علامت و به هم نزدیک باشند $|x - y|$ کوچک و در نتیجه $\delta(x - y)$ بزرگ خواهد شد و در نتیجه تعداد ارقام بامعناى درست $x - y$ کم خواهد بود. بنابراین در عمل بهتر است از تفاضل دو عدد هم‌علامت نزدیک به هم جلوگیری شود (به اصطلاح تفاضل دو عدد نزدیک به هم موجب از بین رفتن ارقام بامعنا می‌شود مانند $1.41 - 1.42 = 0.01$). اگر تفاضل اجتناب‌پذیر نیست باید دقت را بالا برد (دقت دو برابر یا بیشتر). البته به کمک قضیه ۸.۲، می‌توان پدیده از بین رفتن ارقام بامعنا را بهتر تفسیر کرد.

مثال ۱۶.۲ جهت اجتناب از تفاضل در محاسبات می‌توان از اتحادها کمک گرفت. به عنوان مثال

$$e^{x-y} = \frac{e^x}{e^y}, \quad 1 - \cos x = 2 \sin^2 \frac{x}{2}, \quad \ln x - \ln y = \ln \frac{x}{y}$$

△

مثال ۱۷.۲ در صورتی که $g(x) = x \left(\sqrt{1 + \frac{1}{x}} - 1 \right)$ مقدار $g(10^9)$ را با دقت $8D$ به دست آورید. همان‌گونه که بیان شد، چون دقت $8D$ مورد نظر است، محاسبات میانی را با دقت $9D$ انجام می‌دهیم

$$1 + \frac{1}{x} = 1.000000001, \quad \sqrt{1 + \frac{1}{x}} = 1.0000000005$$

بنابراین با دقت $9D$ ، ماشین حساب نتیجه زیر را به دست می‌دهد

$$g(10^9) = 10^9 (1.0000000005 - 1) = 0.0000000005$$

به این ترتیب با یک ماشین حساب ۱۰ رقمی $g(10^9) = 0.0000000000$ که نتیجه‌ای نادرست است. در حالی که اگر از یک رایانه استفاده شود، نتیجه درست $g(10^9) = 0.3333333333$ است. دلیل این خطای فاحش، تفاضل دو عدد نزدیک به هم در محاسبات است که به از بین رفتن ارقام بامعنا منجر می‌شود. برای به دست آوردن تقریب بهتر، روش محاسبه را به صورت زیر تغییر می‌دهیم

$$g(x) = x \left(\sqrt[3]{1 + \frac{1}{x}} - 1 \right) \times \frac{\sqrt[3]{(1 + \frac{1}{x})^2} + \sqrt[3]{1 + \frac{1}{x}} + 1}{\sqrt[3]{(1 + \frac{1}{x})^2} + \sqrt[3]{1 + \frac{1}{x}} + 1} = \frac{1}{\sqrt[3]{(1 + \frac{1}{x})^2} + \sqrt[3]{1 + \frac{1}{x}} + 1}$$

در این صورت با توجه به

$$\left(1 + \frac{1}{x}\right)^2 = 1.0000000002, \quad \sqrt[3]{\left(1 + \frac{1}{x}\right)^2} = 1.0000000000,$$

با دقت ۹D نتیجه زیر به دست می‌آید

$$g(10^9) = \frac{1}{1.0000000000 + 1.0000000000 + 1} = 0.3333333333,$$

△

و بنابراین با دقت ۸D داریم $g(10^9) = 0.3333333333$.

این مثال نه تنها نشان می‌دهد که ممکن است یک ماشین حساب هم نتایج نادرستی تولید کند بلکه به خوبی نشان می‌دهد که به دلیل خطای گرد کردن، ممکن است محاسبه با دو روش مختلف که از نظر ریاضی هم‌ارز هستند به نتایج متفاوتی منجر شوند. از این رو باید از نظر عددی بین الگوریتم‌هایی که از نظر ریاضی هم‌ارز هستند تفاوت قائل شویم. تذکر ۹.۲ با توجه به قضیه ۸.۲، در محاسبات باید از ضرب اعداد بزرگ در اعداد تقریبی (تقسیم اعداد تقریبی به اعداد کوچک) پرهیز کرد.

مثال ۱۸.۲ برای به دست آوردن تقریبی از $x = 10000 \times \pi$ می‌توان نوشت

$$\pi \approx 3/14 \rightarrow x \approx 31400, \quad \pi \approx 3/142 \rightarrow x \approx 31420, \quad \pi \approx 3/1416 \rightarrow x \approx 31416.$$

در تقریب اول خطایی به اندازه‌ی ۱۶ واحد، در تقریب دوم خطایی نزدیک به ۴ واحد و در تقریب سوم خطایی کمتر از ۰/۱ مرتکب شده‌ایم. △

تذکر ۱۰.۲ چون هر عمل محاسباتی خطایی به همراه دارد، یک قاعده کلی دیگر آن است که از حجم محاسبات تا آنجا که ممکن است کاسته شود.

△

مثال ۱۹.۲ به جای عبارت $ax^3 + bx^2 + cx + d$ از عبارت $((ax + b)x + c)x + d$ استفاده شود.

۲.۵.۲ تقریب توابع یک متغیره

بنابر قضیه تیلور، می‌توان به جای کار کردن با یک تابع پیچیده از چند جمله‌ای تیلور نظیر آن استفاده کرد. مثال‌هایی که در ادامه خواهند آمد چگونگی این تقریب را نشان می‌دهند.

مثال ۲۰.۲ مطلوب است محاسبه مقدار $e^{\frac{\pi}{10}}$ با دقت ۲D. روش اول- به کمک قضیه تیلور داریم

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!} e^{\xi(x)},$$

که در آن $0 < \xi(x) < x$ در نتیجه

$$e^{\frac{\pi}{10}} = 1 + \frac{\pi}{10} + \frac{(\frac{\pi}{10})^2}{2!} + \dots + \frac{(\frac{\pi}{10})^n}{n!} + \frac{(\frac{\pi}{10})^{n+1}}{(n+1)!} e^{\xi},$$

که در آن $0 < \xi < \frac{\pi}{10} < 1$. چون e^x تابعی صعودی است پس $1 < e^1 < e^{\xi} < e^0 = 1$. از طرفی $0.315 < \frac{\pi}{10}$ (دقت ۳D منظور می شود) و بنابراین

$$\text{خطا} \leq \left| \frac{(\frac{\pi}{10})^{n+1}}{(n+1)!} e^{\xi} \right| < \frac{3 \times (0.315)^{n+1}}{(n+1)!}.$$

حال باید داشته باشیم $\frac{3 \times (0.315)^{n+1}}{(n+1)!} < 0.5 \times 10^{-2}$ که نتیجه می دهد $n \geq 3$ پس

$$e^{\frac{\pi}{10}} \simeq 1 + 0.315 + \frac{(0.315)^2}{2!} + \frac{(0.315)^3}{3!} = 1.370,$$

و با دقت ۲D داریم $e^{\frac{\pi}{10}} \simeq 1.37$ که با جواب ماشین حساب یعنی 1.369107771 کمتر از 0.001 اختلاف دارد. روش دوم- به کمک سری تیلور داریم

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^n}{n!} + \dots.$$

در نتیجه

$$e^{\frac{\pi}{10}} = 1 + \frac{\pi}{10} + \frac{(\frac{\pi}{10})^2}{2!} + \dots + \frac{(\frac{\pi}{10})^n}{n!} + \dots.$$

با توجه به $0.315 < \frac{\pi}{10}$ (دقت ۳D منظور می شود) از $\frac{(0.315)^n}{n!} < 0.5 \times 10^{-2}$ نتیجه می شود $n \geq 4$ پس

$$e^{\frac{\pi}{10}} \simeq 1 + 0.315 + \frac{(0.315)^2}{2!} + \frac{(0.315)^3}{3!} + \frac{(0.315)^4}{4!} = 1.315 + 0.050 + 0.005 + 0.000 = 1.370.$$

△

مثال ۲۱.۲ (همگرایی سریع) می خواهیم تابع $\cos x$ را به ازای مقادیر $|x| < \frac{\pi}{4}$ با دقت ۵D ارزیابی کنیم. با توجه به سری تیلور

$$\cos x = \sum_{i=0}^{\infty} (-1)^i \frac{x^{2i}}{(2i)!} = 1 + \frac{x^2}{2!} - \frac{x^4}{4!} + \dots,$$

و این که $\frac{1}{(2n)!} < 0.5 \times 10^{-5}$ باید داشته باشیم که نتیجه می دهد $n \geq 6$. △

مثال ۲۲.۲ (همگرایی کند) با توجه به

$$\frac{1}{1+t} = 1 - t + t^2 - t^3 + \dots,$$

داریم

$$\int_0^x \frac{dt}{1+t} = \int_0^x (1 - t + t^2 - t^3 + \dots) dt,$$

و یا

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots.$$

برای ارزیابی $\ln(1+x)$ با دقت δD ، باید داشته باشیم $|\frac{x^n}{n}| < 0.5 \times 10^{-5}$ که برای $x = 0.99$ نتیجه می‌دهد $n \geq 582$. \triangle

مثال ۲۳.۲ تابع $Si(x) = \int_0^x \frac{\sin t}{t} dt$ ، $x \geq 0$ را در نظر بگیرید. دست کم چند جمله از بسط مک‌لورن تابع $f(t) = \sin t$ لازم است تا $Si(1)$ با دقت δD مشخص شود؟
با اعمال قضیه تیلور برای تابع f داریم

$$Si(1) = \int_0^1 \frac{1}{t} \left(f(0) + \frac{f'(0)}{1!}t + \frac{f''(0)}{2!}t^2 + \dots + \frac{f^{(k)}(0)}{k!}t^k + \frac{f^{(k+1)}(\xi(t))}{(k+1)!}t^{k+1} \right) dt,$$

و یا

$$Si(1) = \int_0^1 \frac{1}{t} \left(t - \frac{t^3}{3!} + \frac{t^5}{5!} + \dots + (-1)^n \frac{t^{2n+1}}{(2n+1)!} + (-1)^{n+1} \frac{t^{2n+2}}{(2n+2)!} \sin(\xi(t)) \right) dt.$$

در نتیجه

$$Si(1) = \int_0^1 \left(1 - \frac{t^2}{3!} + \frac{t^4}{5!} + \dots + (-1)^n \frac{t^{2n}}{(2n+1)!} \right) dt + \frac{(-1)^{n+1}}{(2n+2)!} \int_0^1 t^{2n+1} \sin(\xi(t)) dt.$$

بنابراین

$$Si(1) = 1 - \frac{1}{3 \times 3!} + \frac{1}{5 \times 5!} + \dots + (-1)^n \frac{1}{(2n+1) \times (2n+1)!} + E,$$

که در آن $E = \frac{(-1)^{n+1}}{(2n+2)!} \int_0^1 t^{2n+1} \sin(\xi(t)) dt$ حال می‌توان نوشت

$$|E| = \frac{1}{(2n+2)!} \left| \int_0^1 t^{2n+1} \sin(\xi(t)) dt \right| \leq \frac{1}{(2n+2)!} \int_0^1 t^{2n+1} |\sin(\xi(t))| dt.$$

پس

$$|E| \leq \frac{1}{(2n+2)!} \int_0^1 t^{2n+1} dt = \frac{1}{(2n+2) \times (2n+2)!}.$$

از $10^{-6} \times 0.5 < \frac{1}{(2n+2) \times (2n+2)!}$ نتیجه می‌شود $n \geq 4$ بنابراین

$$Si(1) \approx 1 - \frac{1}{3 \times 3!} + \frac{1}{5 \times 5!} - \frac{1}{7 \times 7!} + \frac{1}{9 \times 9!},$$

$$Si(1) \simeq 1 - 0.0555556 + 0.0016667 - 0.0000282 + 0.0000003 = 0.9460831.$$

△

پس با دقت $6D$ داریم $Si(1) \simeq 0.946083$.

۶.۲ تمرین‌ها

۱. حاصل $(\circ/\circ\bar{z})_\beta$ که در آن $z = \beta - 1$ را به دست آورید.
۲. فاصله بین اعداد در دستگاه ممیز ثابت $(a_1 a_2 a_3 a_4 \dots / b_1 b_2 b_3 b_4 \dots)_\beta$ را به دست آورید.
۳. آیا $\frac{1}{\beta+1}$ می‌تواند ماننسیس یک عدد نرمال شده در مبنای β باشد؟ چرا؟
۴. بزرگ‌ترین عدد p را به دست آورید که نمایش آن در دستگاه نمایش اعداد نرمال شده در مبنای β با m رقم ماننسیس و $l \leq e \leq u$ برابر خود آن عدد باشد.
۵. فرض کنید x_+ و x_- به ترتیب اعداد ماشینی بلافاصله بعد و قبل از عدد غیرماشینی x باشند. کمترین مقدار $x_+ - x_-$ در دستگاه اعداد نرمال $\pm(\circ/b_1 \dots b_m)_\beta \times \beta^e$ که در آن $l \leq e \leq u$ را محاسبه کنید.
۶. اگر نمایش دو عدد x و $x+y$ در دستگاه نمایش اعداد نرمال در مبنای β با m رقم ماننسیس و $l \leq e \leq u$ یکسان باشد، آن‌گاه چه رابطه‌ای بین x و y برقرار است.
۷. در یک دستگاه ممیز شناور نرمال برای اعداد حقیقی در مبنای ۲ با t رقم ماننسیس و روش گرد کردن را در نظر بگیرید. اگر x و y هر دو در این دستگاه قابل نمایش باشند، چه شرطی برقرار باشد تا نمایش $x + 2^{-t}y$ همان نمایش x باشد؟
۸. نشان دهید تساوی زیر برقرار است

$$(\circ/b_1 b_2 \dots b_m \overline{c_1 c_2 \dots c_k})_\beta = \frac{(b_1 b_2 \dots b_m c_1 c_2 \dots c_k)_\beta - (b_1 b_2 \dots b_m)_\beta}{\underbrace{(z z \dots z)_k}_{k} \underbrace{(z \dots z \circ \dots \circ)_m}_{m}}, \quad z = \beta - 1$$

۹. مقدار محاسبه شده برای

$$\frac{f(a + 2^{-p}h) - f(a - 2^{-p}h)}{h 2^{1-p}}$$

و $p \geq 1$ در یک ماشین با روند عدد یک برابر 2^{1-t} به ازای چه مقادیری از $|h|$ برابر \circ است؟

$$10. \text{ درستی عبارات } \frac{|r(x) - x|}{|x|} < \frac{1}{4} \beta^{1-k} \text{ و } \frac{|c(x) - x|}{|x|} < \beta^{1-k} \text{ را بررسی کنید.}$$

۱۱. برای محاسبه مقدار $1 - \cos(x)$ در نزدیکی \circ از چه عبارتی استفاده می‌کنید؟ توضیح دهید.

$$\frac{\sin^2(x)}{1 + \cos(x)} \quad (4) \quad x^3 - \frac{x^2}{3} \quad (3) \quad \frac{x}{2} - \frac{x^2}{2} \quad (2) \quad \circ \quad (1)$$

۱۲. برای محاسبه $\sum_{k=1}^{2n} (-1)^k x^k$ با یک وسیله محاسباتی، زمانی که $0 < x < 1$ و n خیلی بزرگ باشد، از چه عبارتی استفاده کنیم مناسب تر است؟

۱۳. برای محاسبه تقریبی $\frac{1}{(\sqrt{3}+1)^6}$ کدام عبارت تقریب دقیق‌تری به دست می‌دهند؟

(۱) $99 - 7\sqrt{2}$ (۲) $\frac{1}{99+7\sqrt{2}}$ (۳) $(\sqrt{3}-1)^6$

۱۴. برای تقریب تابع $\cos(x)$ به ازای $|x| \leq 0.1$ با استفاده از بسط تیلور حول نقطه 0 به منظور دستیابی به خطای برشی حدود 10^{-5} به چه تعداد از جملات سری تیلور نیاز داریم؟

۱۵. \hat{x} و \hat{y} به ترتیب تقریبی از x و y هستند. اگر این اعداد مثبت باشند، در چه صورت خطای مطلق \widehat{xy} به عنوان تقریبی از xy در حد خطای مطلق $\hat{x} + \hat{y}$ است؟

۱۶. یک استوانه به شعاع قاعده $\frac{4}{3}$ و ارتفاع $\sqrt{3}$ در نظر بگیرید. اگر بخواهیم حجم این استوانه را با دقت $3D$ به دست آوریم، شعاع قاعده و ارتفاع استوانه و حتی عدد π را با چه دقتی وارد محاسبات کنیم؟

۱۷. مطلوب است تعیین تقریبی از عدد π با دقت $3D$ به کمک بسط مک‌لورن تابع $f(x) = \tan^{-1} x$.

۱۸. حداقل چند جمله از بسط مک‌لورن تابع $f(x) = e^{-\frac{x}{2}}$ نیاز است تا $f(\ln(2))$ با دقت $3D$ به دست آید؟

۱۹. حداکثر خطای نسبی محاسبه حجم یک مخروط از فرمول $V = \frac{\pi}{3} r^2 h$ چقدر است، هرگاه تقریبی از عدد π با خطای نسبی حداکثر 0.003 داشته باشیم و کمیت‌های r و h را بتوان با حداکثر خطای نسبی 0.002 اندازه‌گیری کرد.

۲۰. برای محاسبه عبارت $e^{1-\ln(2/7^3)}$ در یک ماشین حساب با سه رقم بامعنا کدام گزینه مناسب‌تر است؟

الف) $e^{1-\ln(2/7^3)}$ (ب) $\frac{e}{7^3}$ (ج) $e \times e^{-\ln(2/7^3)}$ (د) $\frac{e}{\ln(2/7^3)}$

۲۱. برای همه گزینه‌ها a_0 تقریبی از عدد π با دقت $3S$ است. در کدام مورد خطای مطلق در محاسبه a_{10} کمتر است؟

الف) $a_n = 1 + 2a_{n-1}$ (ب) $a_n = 1 + na_{n-1}$ (ج) $a_n = 1 + a_{n-1}$ (د) $a_n = 1 + \frac{1}{n}a_{n-1}$

۲۲. برای محاسبه عبارت $\frac{1-\cos(x)}{x}$ در کامپیوتر، وقتی x عدد مثبت کوچکی است، کدام گزینه مناسب‌تر است؟

الف) $\frac{1-\cos(x)}{x}$ (ب) $\frac{2 \sin^2(\frac{x}{2})}{x}$ (ج) $\frac{\sin^2(x)}{x(1+\cos(x))}$ (د) $\frac{1}{x} - \frac{\cos(x)}{x}$

۲۳. برای محاسبه $\sqrt{x^2 + \frac{1}{x^2}} - \sqrt{x^2 - \frac{1}{x^2}}$ برای x های خیلی بزرگ در ماشین کدام گزینه مناسب‌تر است؟

الف) $\frac{1}{x^2}$ (ب) $\frac{2}{x(\sqrt{x^2+1} + \sqrt{x^2-1})}$ (ج) $\sqrt{x^2 + \frac{1}{x^2}} - \sqrt{x^2 - \frac{1}{x^2}}$ (د) $\frac{1}{x} (\sqrt{x^2+1} - \sqrt{x^2-1})$

۲۴. در یک ماشین با دقت $4S$ و گرد کردن، کوچکترین عدد طبیعی x که در تساوی $1+x=x$ صدق می‌کند، کدام گزینه است؟ الف) $10^4 \times 0.5000$ (ب) $10^5 \times 0.1000$ (ج) $10^5 \times 0.5000$ (د) $10^6 \times 0.1000$

۲۵. برای زوایای کمتر از 6 درجه، $\sin(x)$ را با مقدار x بر حسب رادیان تقریب می‌زنند. کران بالای خطای مطلق این کار چقدر است؟

۲۶. حداقل چند جمله از بسط مک‌لورن تابع $f(x) = \sin(\frac{x}{2})$ لازم است تا $\sin(0.1)$ با دقت $4D$ مشخص شود؟

۲۷. درصد خطای محاسبه عدد π و اندازه‌گیری شعاع دایره به ترتیب در هر یک از گزینه‌های زیر داده شده است. با انتخاب کدام گزینه می‌توان مساحت دایره را با حداکثر خطای نسبی 10^{-2} مشخص کرد؟

(الف) 0.5% و 0.25% (ب) 0.6% و 0.3% (ج) 0.5% و 0.35% (د) 0.55% و 0.25%

۲۸. می‌خواهیم کمیت‌های U و V را از فرمول‌های $U = x^2 \div y$ و $V = xy^2$ محاسبه کنیم. درصد خطای x و y به ترتیب در هر یک از گزینه‌های زیر داده شده است. با انتخاب کدام گزینه می‌توان U و V را با حداکثر خطای نسبی 10^{-2} مشخص کرد؟

(الف) 0.2% و 0.35% (ب) 0.25% و 0.3% (ج) 0.3% و 0.25% (د) 0.35% و 0.2%

۲۹. کدام یک از گزینه‌های زیر در حساب ممیز شناور درست است؟

(الف) عمل جمع دارای خاصیت شرکت پذیری است. (ب) عضو خنثای عمل جمع یکتا است.
(ج) عمل جمع خاصیت جابجایی دارد. (د) جمع دو عدد مثبت نزدیک به هم دارای خطای زیادی است.

۳۰. حداکثر خطای نسبی محاسبه حجم یک کره از فرمول $V = \frac{4}{3}\pi r^3$ چقدر است، هرگاه تقریبی از اعداد π و $\frac{1}{3}$ با خطای نسبی حداکثر 0.002 داشته باشیم و شعاع را بتوان با حداکثر خطای نسبی 0.002 اندازه‌گیری کرد؟

(الف) 0.001 (ب) 0.010 (ج) 0.020 (د) 0.002

۳۱. کدام یک از گزینه‌های زیر در حساب ممیز شناور درست نیست؟

(الف) در محاسبه تقریبی $\frac{\sqrt{2}}{100000}$ قطعاً خطای زیادی داریم.
(ب) در محاسبه تقریبی $100000\sqrt{2}$ قطعاً خطای زیادی داریم.
(ج) در تفاضل دو عدد منفی نزدیک به هم، انتشار خطا رخ می‌دهد.
(د) در تفاضل دو عدد مثبت نزدیک به هم، انتشار خطا رخ می‌دهد.

۳۲. در محاسبه عبارت $\frac{e^{x^2} - \cos x}{x^2}$ کدام گزینه صحیح است؟

(الف) به ازای تمام مقادیر x انتشار خطا رخ نمی‌دهد. (ب) برای مقادیر بزرگ $|x|$ انتشار خطا رخ می‌دهد.
(ج) برای مقادیر کوچک $|x|$ انتشار خطا رخ می‌دهد. (د) به ازای همه مقادیر $|x|$ انتشار خطا رخ می‌دهد.

۳۳. در یک دستگاه ممیز شناور که در آن اعداد به صورت $\pm 0.d_1d_2d_3d_4 \times 10^n$ با $d_1 \neq 0$ و $-7 \leq n \leq 8$ و $0 \leq d_i \leq 9$ نمایش داده می‌شوند فاصله بین عدد 10000 و اولین عدد قابل نمایش بزرگتر از 10000 کدام است؟

(الف) 10 (ب) 0.1 (ج) 0.01 (د) 0.005

۳۴. اگر مقدار $\ln(1+x)$ را با استفاده از دو جمله اول بسط مکولرن آن تقریب بزنیم، مقدار تقریبی و حداکثر خطای محاسبه $\ln(1/1)$ به ترتیب کدام است؟

(الف) $(0.001, 0.005)$ (ب) $(0.1, 0.05)$ (ج) $(0.05, 0.005)$ (د) $(0.1, 0.005)$

۳۵. در محاسبه دوره تناوب آونگ ساده $T = 2\pi\sqrt{\frac{l}{g}}$ ، درباره اثر خطای نسبی مقادیر π, l, g در خطای نسبی T چه می‌توان گفت؟

الف) اثر خطای نسبی π بیشتر است. (ب) اثر خطای نسبی l بیشتر است.
 ج) اثر خطای نسبی g بیشتر است. (د) اثر یکسانی دارند. π, l, g

۳۶. در یک دستگاه ممیز شناور که در آن اعداد به صورت $\pm 0/d_1 d_2 d_3 d_4 \times 10^n$ با $d_1 \neq 0$ و $0 \leq d_i \leq 9$ و $-7 \leq n \leq 8$ با گرد کردن نمایش داده می‌شوند، بزرگترین عدد x که در معادله $500 + x = 500$ صدق می‌کند، کدام است؟

الف) 0.04999 (ب) 0.09999 (ج) 0.05000 (د) 0.10000

۳۷. اگر مقدار دقیق و تقریبی در توانی از ده ضرب شوند، آنگاه

الف) خطای نسبی به توانی از ده تقسیم می‌شود. (ب) خطای مطلق به توانی از ده تقسیم می‌شود.
 ج) خطای نسبی تغییری نمی‌کند. (د) خطای مطلق تغییری نمی‌کند.

۳۸. گزینه مناسب برای محاسبه $T = \frac{1 - \cos x}{\sin x}$ به ازای مقادیر x نزدیک صفر کدام است؟

الف) 0 (ب) $\frac{x}{3}$ (ج) $\frac{\pi}{3}$ (د) $\frac{1}{\sin x} - \cot x$

۳۹. در محاسبه تقریبی $\ln(1/0.01)$ با سه جمله از بسط تیلور، حداکثر خطا چقدر است؟

الف) $\frac{10^{-6}}{3}$ (ب) $\frac{10^{-9}}{3}$ (ج) $\frac{10^{-4}}{3}$ (د) $\frac{10^{-6}}{2}$

۴۰. در محاسبه حجم $(V = \frac{4}{3}\pi r^3)$ یک کره به شعاع r ، درباره اثر خطای نسبی داده‌ها در خطای نسبی V چه می‌توان گفت؟

الف) اثر خطای نسبی r بیشتر است. (ب) اثر خطای نسبی π بیشتر است.
 ج) اثر خطای نسبی $\frac{4}{3}$ بیشتر است. (د) هر سه اثر خطای نسبی یکسانی دارند.

۴۱. گزینه مناسب برای محاسبه $T = \frac{\cos(2x) - 1 + 2x^2}{2x^4}$ به ازای مقادیر x نزدیک صفر کدام است؟

الف) $\frac{2}{3}$ (ب) $\frac{2x^4}{3}$ (ج) $\frac{2x}{3}$ (د) $\frac{3x}{2}$

۴۲. برای محاسبه $f(3)$ با خطایی کمتر از 0.0007 چند جمله از بسط تیلور تابع f حول نقطه $x_0 = 2/5$ لازم است (می‌دانیم $|f^{(n)}(x)| \leq \frac{1}{n}$)؟

الف) دو جمله (ب) سه جمله (ج) پنج جمله (د) چهار جمله

۴۳. در یک دستگاه ممیز شناور نرمال شده که هر عدد حقیقی به صورت $\pm 0/d_1 d_2 d_3 d_4 \times 10^e$ با $d_1 \neq 0$ و $0 \leq d_i \leq 9$ برای $i = 1, 2, 3, 4$ و $-6 \leq e \leq 7$ ، نمایش داده می‌شود، اعداد قابل نمایش قبل و بعد از 10000 کدام هستند؟

الف) $10001, 9999$ (ب) $10010, 9990$ (ج) $10010, 9999$ (د) $10001, 9990$

۴۴. در دستگاه ممیز شناور نرمال شده سوال قبل، تعداد اعداد قابل نمایش در بازه $[1, 10]$ کدام است؟

الف) 9001 (ب) 9000 (ج) 8999 (د) 9002