به نام پروردگار یگانه و یکتا

دانشگاه صنعتی اصفهان-دانشکده علوم ریاضی

محاسبات عددی

پدیدآورنده : رضا مختاری

ویرایش ششم : زمستان ۱۳۹۹ و بهار ۱۴۰۰

فهرست مطالب

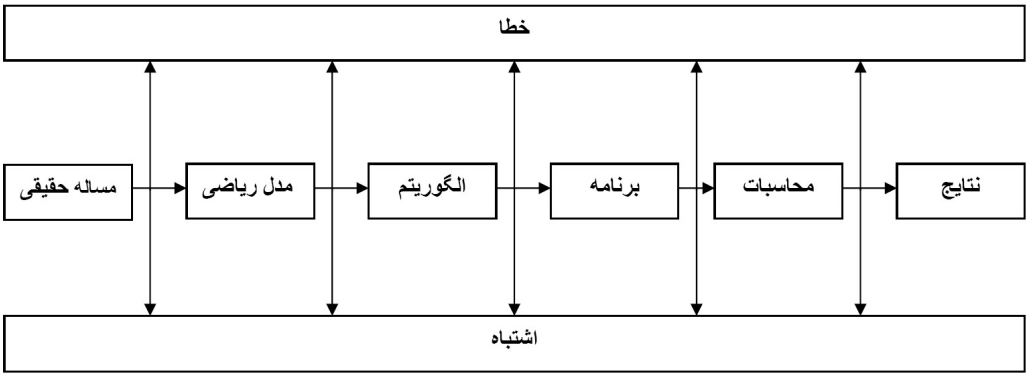
فصل ۱

خطاها

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

۱.۱ منابع تولید خطا

بیشتر مواقع در عمل با یک مسئله حقیقی (فیزیکی) مواجه هستیم و بنابر دلایلی، جواب تقریبی (عددی) آن را جستجو می‌کنیم. مراحل یافتن جواب تقریبی یک مسئله حقیقی در روندنمای آمده در شکل ۱.۱ خلاصه می‌شود. این روندنما مکان‌های احتمالی بروز خطا و همچنین اشتباهات را نیز نشان می‌دهد.



شکل ۱.۱: فرایند تولید جواب عددی (تقریبی)

تذکر ۱.۱ *اختلافی که بین جواب دقیق و تقریبی وجود دارد ممکن است از اشتباهات و خطاها ناشی شده باشد. اشتباه را می‌توان برطرف کرد ولی خطا بیشتر اوقات اجتناب‌ناپذیر است. به عنوان مثال قراردادن ۲۳۲۲ به جای ۲۲۳۲ یک اشتباه است و استفاده از ۳/۱۴ به جای عدد π که بسط ده‌دهی نامختوم دارد، موجب بروز خطا می‌شود.*

مراحل مختلف روندنمای آمده در شکل ۱.۱ را در مثال بعد دنبال می‌کنیم.

مثال ۱.۱ (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. به کمک برخی از قوانین و اصول فیزیک و ریاضیات، صرف نظر از مقاومت هوا و اصطکاک در لولا، مدل این مسئله به صورت*

$$ml\frac{d^۲\theta}{dt^۲} = -mg\sin\theta,$$

یا $\frac{d^۲\theta}{dt^۲} = -\frac{g}{l}\sin\theta$ به دست می‌آید که یک معادله دیفرانسیل عادی غیرخطی است. با فرض کوچک بودن θ یعنی

$$\theta = ۶^\circ \simeq ۰/۱۰۴۷^{\mathrm{rad}}, \quad \sin\theta \simeq ۰/۱۰۴۵,$$
$$\theta = ۱۵^\circ \simeq ۰/۲۶۲^{\mathrm{rad}}, \quad \sin\theta \simeq ۰/۲۵۹,$$

می‌توان فرض کرد $\sin\theta \simeq \theta^{\mathrm{rad}}$ و معادله دیفرانسیل غیرخطی را به صورت زیر نوشت

$$\frac{d^۲\theta}{dt^۲} + \omega^۲\theta = ۰, \qquad \omega^۲ = \frac{g}{l}.$$

این معادله دیفرانسیل خطی جوابی متناوب به صورت $\theta(t) = c_۱\sin\omega t + c_۲\cos\omega t$ دارد و بنابراین دوره تناوب آونگ ساده عبارت است از $T = \frac{۲\pi}{\omega} = ۲\pi\sqrt{\frac{l}{g}}$. △

با توجه به این مثال، می‌توان خطاها را از نقطه نظر منبع تولید به صورت زیر تقسیم‌بندی کرد.

انواع خطا

۱. ذاتی

- مدل (ناشی از صرف‌نظرها، چشم‌پوشی‌ها و ساده‌سازی‌ها مانند فرض $\sin\theta \simeq \theta^{rad}$)
- داده‌های مدل (ناشی از آزمایشات و اندازه‌گیری‌ها مثل g, l)

۲. محاسباتی

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

تذکر ۲.۱ *خطاهای ذاتی به محاسبات عددی مربوط نمی‌شود اما برای پرهیز از خطاهای محاسباتی و کنترل آن‌ها باید راه چاره‌ای پیدا کرد.*

تذکر ۳.۱ *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*[1] *در رایانه آن*

۲.۱ نمایش اعداد

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

قضیه ۱.۱ *هر عدد حقیقی مثبت x نمایشی به صورت*

$$\begin{aligned} x &= a_m\beta^m + a_{m-1}\beta^{m-1} + \cdots + a_1\beta^1 + a_0\beta^0 + a_{-1}\beta^{-1} + a_{-2}\beta^{-2} + \cdots \\ &= (a_m a_{m-1} \ldots a_1 a_0 / a_{-1} a_{-2} \ldots)_\beta, \end{aligned} \tag{۱.۱}$$

دارد که در آن $m \in \mathbb{Z}$، $a_i \in \{0, 1, 2, \ldots, \beta - 1\}$ و $a_m \neq 0$.

تذکر ۴.۱ *رابطه (۱.۱) به نمایش (بسط) عدد x در مبنای β معروف است و اگر $\beta = 10$ اختیار شود به آن، نمایش (بسط) ده‌دهی (اعشاری) گویند (متداول در زندگی روزمره) و در حالتی که $\beta = 2$ در نظر گرفته شود به آن، نمایش دودویی (باینری) گفته می‌شود (مبنای کار رایانه).*

قضیه ۲.۱ *نمایش یک عدد گویا (در هر مبنایی) یا مختوم است یا نامختوم متناوب.*

نتیجه ۱.۲.۱ *بسط یک عدد گنگ، نامختوم نامتناوب است.*

مثال ۲.۱ *به موارد زیر در مبناهای متفاوت توجه کنید*

$$\frac{2}{3} = 0/666\cdots = 0/\bar{6} = (0/2)_3, \quad \frac{3}{8} = 0/375 = (0/3)_8, \quad \sqrt{2} = 1/4142\cdots,$$
$$0/1 = (0/0\overline{0011})_2, \quad \frac{1}{4} = 0/25 = (0/01)_2, \quad \pi = 3/141592\cdots.$$

△

[1] Underflow

تذکر ۵.۱ *اگر چه فرض $a_m \neq ۰$ برای یکتایی نمایش (۱.۱) در نظر گرفته شده است، برای منحصر به فرد بودن نمایش (۱.۱) به فرض‌های دیگری نیز نیاز است. به مثال زیر توجه کنید*

$$۳/۴۷۹۹۹۹\cdots = ۳/۴۷\bar{۹} = ۳\times ۱۰^{۰} + ۴\times ۱۰^{-۱} + ۷\times ۱۰^{-۲} + ۹\times ۱۰^{-۳} + ۹\times ۱۰^{-۴} + \cdots$$
$$= ۳/۴۷ + \frac{۹\times ۱۰^{-۳}}{۱-۱۰^{-۱}} = ۳/۴۷ + ۰/۰۱ = ۳/۴۸.$$

یعنی برای اعداد $۳/۴۷\bar{۹}$ و $۳/۴۸$ یک نمایش وجود دارد. اگر فرض کنیم عدد صحیح j چنان وجود داشته باشد که $a_j = a_{j-۱} = \cdots = ۰$ (به عبارتی فرض کنیم بسط مختوم باشد) این مشکل برطرف می‌شود.

تذکر ۶.۱ *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 (امروزه مبنای ۲ و در قدیم مبناهای دیگری مانند ۱۶) کار می‌کنند. بنابراین مسئله تغییر مبنا مطرح می‌شود که ممکن است خطایی به دنبال داشته باشد که در اینجا از بررسی آن صرف نظر می‌کنیم.*

۳.۱ 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

$$x = \pm(a_n a_{n-۱}\ldots a_۱ a_۰/a_{-۱}a_{-۲}\ldots a_{-m})_\beta,$$

که در آن m و n 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 گرفت و تا حدودی به طور سلیقه‌ای با آن برخورد شد تا زمانی که استانداردی توسط $IEEE$[۴] وضع شد.

۱.۳.۱ 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 نمایش هر عدد در مبنای ۲، ابتدا یک ساختار به طول ۶۴ بیت در نظر گرفته می‌شود. اولین بیت به بیت علامت

[۲] Fixed point
[۳] Floating point
[۴] IEEE standard 754-1985
[۵] Double precision

معروف است و با s نمایش داده شود و بلافاصله بعد از آن ۱۱ بیت برای مشخصه[۶] در نظر گرفته می‌شود و با c نمایش داده می‌شود و ۵۲ بیت باقی‌مانده به نام مانتیس منظور می‌شود که آن را با f نشان می‌دهند. اگرچه برای مانتیس ۵۲ بیت در نظر گرفته شده است ولی در واقع ساختار $1+f$ موجب می‌شود که ۵۳ بیت داشته باشیم که آن یک بیت اضافه به بیت پنهان معروف است و برای یکتایی نمایش لازم است. برای هر عدد نمایشی به صورت $(-1)^s \times 2^e \times (1+f)$ در نظر گرفته می‌شود که در آن $e = c - 1023$ توان (نما) است. محدودیت $0 \le c \le (11\cdots1)_2 = 2^{11} - 1 = 2047$ موجب می‌شود که

$$-1023 \le e \le 1024.$$

از طرف دیگر محدودیت $0 \le f \le (0.\overbrace{11\cdots1}^{52 times})_2 < 2$ باعث می‌گردد که $1 \le (1+f)_2 \le (1.\overbrace{11\cdots1}^{52 times})_2 < 2$. بنابراین اگر کوچک‌ترین و بزرگ‌ترین عدد مثبت قابل نمایش را به ترتیب با mN (realmin در محیط MATLAB) و MN (realmax در محیط MATLAB) نشان دهیم آنگاه

$$mN = 2^{-1022} \times (1.0\cdots0)_2 \simeq 2.22507 \times 10^{-308},$$
$$MN = 2^{1024} \times (1.0\cdots0)_2 \simeq 1.79769 \times 10^{308}.$$

تذکر ۷.۱ *باید توجه داشت که توان -1023 برای نمایش صفر و توان 1024 با مانتیس مثبت برای نمایش ∞ (inf در محیط MATLAB) و صور مبهم (NaN در محیط MATLAB) مورد استفاده قرار می‌گیرد.*

تعریف ۱.۱ *در نمایش اعداد ماشینی، کوچک‌ترین عدد مثبت ماشینی که اگر به ۱ اضافه شود عددی بزرگ‌تر از ۱ به دست می‌آید به اپسیلون ماشین[۷] معروف است و با* eps *نمایش داده می‌شود.*

چون در نمایش ۶۴-بیتی بعد از ۱ عدد $(1.\underbrace{0\cdots0}_{51 times})_2$ قرار می‌گیرد، پس

$$\text{eps} = (1.\overbrace{0\cdots0}^{51 times}1)_2 - 1 = (0.\overbrace{0\cdots0}^{51 times}1)_2 = 2^{-52} \simeq 2.220446 \times 10^{-16}.$$

حال بزرگترین عدد صحیح مثبت M را تعیین می‌کنیم که هر عدد صحیح x با شرط $0 \le x \le M$ در این نمایش به طور دقیق قابل نمایش باشد. به وضوح، تمام اعداد صحیح نامنفی که بزرگتر از $2^{52} \times (1.11\cdots1)_2 = 2^{53} - 1$ نباشند به طور دقیق قابل نمایش هستند و به علاوه 2^{53} نیز به صورت $2^{53} \times (1.\underbrace{0\cdots0}_{52 times})_2$ قابل نمایش است. اما تعداد ارقام در مانتیس جهت نمایش $2^{53}+1$ کافی نیست (۵۳ رقم در مانتیس لازم است). بنابراین $M = 2^{53} \simeq 9.007199 \times 10^{15}$ و در نتیجه تمام اعداد صحیح ۱۵ رقمی و بسیاری از اعداد ۱۶ رقمی در این نمایش به طور دقیق قابل نمایش هستند (گفته می‌شود دقت در نمایش ۶۴-بیتی ۱۵ الی ۱۶ رقم است). البته بسیاری از توان‌های ۲ نیز به طور دقیق قابل نمایش هستند.

---

[۶] Characteristic

[۷] Machine epsilon

تذکر ۸.۱ *مجموعه اعداد با ممیز شناور را با $\mathbb{FP}$ 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 فاصله‌ی $(-mN, mN)$ تولید شود پیام پاریز[۸] و اگر عدد تولید شده از MN بزرگ‌تر یا از $-MN$ کوچک‌تر شود پیام سرریز[۹] صادر می‌شود. از توان‌های ۱۰۲۳− و ۱۰۲۴ برای نمایش پیام‌های پاریز و سرریز نیز استفاده می‌شود.*

تذکر ۹.۱ *بسیاری از نرم‌افزارها مانند* MATLAB *از محدودیت‌های سخت‌افزار پیروی می‌کنند و بعضی از نرم‌افزارها مانند* Mathematica*، محدودیت‌های سخت‌افزار را از طریق برنامه‌های نرم‌افزاری برطرف می‌کنند و به اصطلاح دقت را بالا می‌برند که ممکن است با کاهش سرعت انجام محاسبات همراه باشد.*

۲.۳.۱ اعداد ماشینی

به منظور سادگی نوشتار، برای هر عدد ماشینی[۱۰] (یا عناصر مجموعه اعداد $\mathbb{FP}$)، نمایشی به صورت

$$\pm \underbrace{{}^{0}/_{d_1 \cdots d_k}}_{\text{مانتیس}} \times 10^{n} \longleftarrow \text{نما (توان)}$$

در نظر گرفته می‌شود که در آن n به زیرمجموعه‌ای از $\mathbb{Z}$ مانند $[L, U]$ تعلق دارد و به ازای $i = 1, 2, \ldots, k$ داریم $d_i \in \{0, 1, \ldots, 9\}$ و برای یکتایی نمایش فرض می‌شود $d_1 \neq 0$. به این نمایش، نمایش ممیز شناور ده‌دهی نرمال‌شده گفته می‌شود و چنین اعدادی را، اعداد ماشینی ده‌دهی k-رقمی نیز می‌نامند. ایده چنین نمایشی، از نمایش علمی نرمال‌شده اعداد ناشی شده است. در نمایش علمی نرمال‌شده، هر عدد حقیقی مخالف صفر x را می‌توان به صورت

$$x = \pm \underbrace{{}^{0}/_{d_1 d_2 \cdots d_k d_{k+1} d_{k+2} \cdots}}_{\text{مانتیس}} \times 10^{n} \longleftarrow \text{نما (توان)}$$

نمایش داد که در آن $n \in \mathbb{Z}$ و به ازای $i = 1, 2, \ldots, k$ داریم $d_i \in \{0, 1, \ldots, 9\}$ و $d_1 \neq 0$. واضح است که برای نمایش x به صورت ممیز شناور ده‌دهی نرمال‌شده در یک ماشین (k-رقمی)، باید k رقم از مانتیس آن را حفظ کرده و بقیه را کنار گذاشت، که برای این کار روش‌های زیر موجود است

۱. روش قطع کردن (برش)[۱۱]

۲. روش گرد کردن معمولی[۱۲]

۳. روش گرد کردن به زوج[۱۳]

---

[۸] Underflow
[۹] Overflow
[۱۰] Machine number
[۱۱] Chopping
[۱۲] Rounding
[۱۳] Rounding to even

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 اِعمال شده، سپس اگر $d_{k+1} \geq ۵$ یک واحد به d_k اضافه می‌گردد. اما در روش گرد کردن به زوج، ابتدا روش قطع کردن اِعمال شده و در هر یک از حالت‌های زیر یک واحد به d_k اضافه می‌گردد

- $d_{k+1} > ۵$
- $d_{k+1} = ۵$ و رقم مخالف صفری در سمت راست d_{k+1} مشاهده شود
- $d_{k+1} = ۵$ و رقم مخالف صفری در سمت راست d_{k+1} مشاهده نشود و d_k فرد باشد.

قضیه ۳.۱ *اگر $x = ۰/d_۱d_۲d_۳\cdots d_kd_{k+1}\cdots \times ۱۰^n$ یک عدد حقیقی ناصفر در نمایش علمی نرمال‌شده باشد، $\widetilde{fl}(x) = ۰/d_۱d_۲d_۳\cdots d_k \times ۱۰^n$ عدد ماشینی k-رقمی متناظر با x است که از روش قطع کردن به دست می‌آید و*

$$\left|x - \widetilde{fl}(x)\right| \leq ۱ \times ۱۰^{n-k}.$$

هم‌چنین $fl(x) = ۰/\delta_۱\delta_۲\cdots\delta_k \times ۱۰^m$ ($m = n$ یا $m = n+۱$) عدد ماشینی k رقمی متناظر با x است که از روش گرد کردن معمولی به دست می‌آید و خواهیم داشت

$$|x - fl(x)| \leq ۵ \times ۱۰^{n-k-۱} = ۰/۵ \times ۱۰^{n-k}.$$

پرسش ۱.۱ *تفاوت ۴۷^{km} با ۴۷۰۰۰^{m} یا تفاوت ۳/۷ با ۳/۷۰ یا ۳/۷۰۰ در چیست؟*

تعریف ۲.۱ *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 است (تمام ارقام مانتیس در نمایش علمی نرمال‌شده).*

مثال ۳.۱ *عدد ۰/۰۰۰۷۰۴۵۰۰۰ حداقل ۴ رقم بامعنا و حداکثر ۷ رقم بامعنا دارد.* △

قرارداد ۱.۱ *rD یعنی r رقم اعشار (تنظیم ماشین حساب روی Mode fix r) و rS یعنی r رقم بامعنا (تنظیم ماشین حساب روی Mode sci r)* [۱۴].

مثال ۴.۱ *در جدول ۱.۱، چند عدد و مقدارهای تقریبی متناظر با آن‌ها با روش‌های قطع کردن ($\widehat{x}$) و گرد کردن معمولی ($\widetilde{x}$) با دقت $۳S$ و $۳D$ داده شده است.* △

تذکر ۱۰.۱ *از این به بعد، گرد کردن معمولی را به کار می‌بریم.*

[۱۴] D حرف اول کلمه Decimal و S حرف اول کلمه Significant

x	۲۸/۶۴۲۴	۰/۰۰۵۷۶۷۱	۴/۹۸۵۰	−۲۱۷۵/۳۴۵۱۲
$\widetilde{x}(۳S)$	۲۸/۶	۰/۰۰۵۷۷	۴/۹۹	−۲۱۸۰
$\widetilde{x}(۳D)$	۲۸/۶۴۲	۰/۰۰۶	۴/۹۸۵	−۲۱۷۵/۳۴۵
$\widehat{x}(۳S)$	۲۸/۶	۰/۰۰۵۷۶	۴/۹۸	−۲۱۷۰
$\widehat{x}(۳D)$	۲۸/۶۴۲	۰/۰۰۵	۴/۹۸۵	−۲۱۷۵/۳۴۵

جدول ۱.۱: مثال‌هایی از گرد کردن و قطع کردن معمولی

پرسش ۲.۱ *در نظر بگیرید $\widehat{x}$ تقریبی از x باشد. آیا هر چه تعداد ارقام بامعنای $\widehat{x}$ بیشتر باشد، می‌توان گفت $\widehat{x}$ تقریب بهتری (به معنای خطای کمتر) است؟*

تعریف ۳.۱ *فرض کنید $\widehat{x} = a_m \times ۱۰^m + \cdots + a_۱ \times ۱۰^۱ + a_۰ \times ۱۰^۰ + a_{-۱} \times ۱۰^{-۱} + \cdots + a_{-k} \times ۱۰^{-k}$ که در آن $a_m \neq ۰$، تقریبی از عدد حقیقی مثبت x باشد. منظور از تعداد ارقام با معنای درست $\widehat{x}$ نسبت به x بزرگ‌ترین nای $(n \in \mathbb{N}_۰)$ است که در نابرابری‌های زیر صدق کند.*

$$n \leq m + k + ۱, \qquad |x - \widehat{x}| \leq ۵ \times ۱۰^{m-n}$$

تذکر ۱۱.۱ *m در تعریف اخیر، نشان دهنده با ارزش‌ترین مکان عدد است.*

مثال ۵.۱ *از اعداد $\widehat{x} = ۹۹/۹۶$ و $\widetilde{x} = ۱۰۰/۷$ کدام یک تقریب بهتری برای $x = ۱۰۰$ است؟*

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

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

△

مثال ۶.۱ *تعداد ارقام بامعنای درست $\widehat{x} = ۱۰۰/۳۱$ را نسبت به $x = ۱۰۰/۳۱۰۴$ مشخص کنید.*

$$\widehat{x} = ۱۰۰/۳۱ \;\rightarrow\; m = ۲,\;\; |x - \widehat{x}| = ۰/۰۰۰۴ \leq ۵ \times ۱۰^{۲-n} \;\rightarrow\; n = ۶$$

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

تذکر ۱۲.۱ *در نظر بگیرید $\widehat{x}$ تقریبی از x باشد. هر چه تعداد ارقام بامعنای درست $\widehat{x}$ نسبت به x بیشتر باشد، $\widehat{x}$ تقریب بهتری خواهد بود.*

۴.۱ انواع خطا

تعریف ۴.۱ *اگر $\widehat{x}$ تقریبی از x باشد $\Delta x = |x - \widehat{x}|$ را خطای مطلق $\widehat{x}$ نسبت به x نامند. Δx منحصر به فرد است و در عمل بیشتر مواقع قابل تعیین نیست و به جای آن از هر عدد b_x استفاده می‌شود که کمتر از Δx نباشد. b_x منحصر*

به فرد نیست و به آن کران خطای مطلق گویند. بنابراین $\Delta x \leq b_x$ *و در نتیجه* $\widehat{x} - b_x \leq x \leq \widehat{x} + b_x$. *بعضی مواقع از نمایش* $x = \widehat{x} \pm b_x$ *استفاده می‌شود.*

مثال ۷.۱ *اگر عدد* $\widehat{x} = ۱/۷۳۲$ *را به عنوان تقریبی از* $x = \sqrt{۳}$ *در نظر بگیریم، آنگاه*

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

از طرفی می‌دانیم $۱/۷۳۲۱ < \sqrt{۳} < ۱/۷۳۲۰$ *بنابراین* $۰/۰۰۰۱ > \sqrt{۳} - ۱/۷۳۲ > ۰$ *پس* $b_x = ۰/۰۰۰۱$ *که معیاری برای نزدیکی* $۱/۷۳۲$ *به* $\sqrt{۳}$ *است.* △

پرسش ۳.۱ *آیا خطای مطلق معیار مناسبی برای مقایسه خطاها است؟*
پاسخ. *خیر. به عنوان مثال خطای مطلق یک صندوق‌دار بانک، تایپیست و دروازه‌بان را در نظر بگیرید.*

تعریف ۵.۱ *اگر* $\widehat{x} \neq ۰$ *تقریبی از* x *باشد* $\delta x = \frac{|x-\widehat{x}|}{|x|} = \frac{\Delta x}{|x|}$ *خطای نسبی* $\widehat{x}$ *نسبت به* x *نامیده می‌شود و مشابه خطای مطلق منحصر به فرد است و بیشتر مواقع در عمل قابل تعیین نیست و از کران خطای نسبی استفاده می‌شود.* $۱۰۰ \times \delta x$ *به درصد خطا معروف است.*

قضیه ۴.۱ *اگر* $\widehat{x}$ *تقریبی از* x *باشد و* b_x *یک کران خطای مطلق برای این تقریب باشد آنگاه*

$$\delta x \leq \frac{b_x}{|\widehat{x}| - b_x}.$$

به علاوه اگر b_x *نسبت به* $|\widehat{x}|$ *خیلی کوچک باشد*

$$\delta x \leq \frac{b_x}{|\widehat{x}|}.$$

مثال ۸.۱ *اگر عدد* $\widehat{x} = ۱/۷۳۲$ *را به عنوان تقریبی از* $x = \sqrt{۳}$ *در نظر بگیریم، آنگاه*

$$\delta x = \frac{\left|\sqrt{۳} - ۱/۷۳۲\right|}{\sqrt{۳}} = \frac{۱/۷۳۲۰۵۰۸۰۷۵\cdots - ۱/۷۳۲}{۱/۷۳۲۰۵۰۸۰۷۵\cdots} = \frac{۰/۰۰۰۰۵۰۸۰۷۵\cdots}{۱/۷۳۲۰۵۰۸۰۷۵\cdots}.$$

پس

$$\delta x = ۰/۰۰۰۰۲۹۳۳۳۷\cdots < ۰/۰۰۰۰۳.$$

اما با توجه به قضیه ۴.۱ و $b_x = ۰/۰۰۰۱$ *می‌توان نوشت*

$$\delta x \leq \frac{۰/۰۰۰۱}{۱/۷۳۲ - ۰/۰۰۰۱} = \frac{۰/۰۰۰۱}{۱/۷۳۱۹} = ۰/۰۰۰۰۵۷۷۴۰۰۵\cdots < ۰/۰۰۰۰۶$$

$$\delta x \leq \frac{۰/۰۰۰۱}{۱/۷۳۲} = ۰/۰۰۰۰۵۷۷۳۶۷۲\cdots < ۰/۰۰۰۰۶.$$

△

قضیه ۵.۱ *اگر $\widehat{x}$ تقریبی از x با n رقم بامعنای درست باشد و $\widehat{y} = ۱۰^t \times \widehat{x}$ و $y = ۱۰^t \times x$ که در آن t عددی صحیح است آنگاه $\widehat{y}$ نیز تقریبی از y با n رقم بامعنای درست است و خطای نسبی $\widehat{x}$ و $\widehat{y}$ برابر است.*

قضیه ۶.۱ *اگر $\widehat{x}$ گردشده x تا n رقم بامعنا باشد آنگاه $\widehat{x}$ دارای n رقم بامعنای درست است.*

قضیه ۷.۱ ارتباط دقت (خطای نسبی) با تعداد ارقام بامعنای درست
اگر $\widehat{x}$ دارای n رقم بامعنای درست باشد، آنگاه $\delta x < ۵ \times ۱۰^{-n}$ به شرط آن که ارقام بامعنای درست $\widehat{x}$ از یک رقم ۱ و $n-۱$ رقم صفر در جلوی آن تشکیل نشده باشد. برعکس اگر $\delta x \leq ۵ \times ۱۰^{-n-۱} = ۰/۵ \times ۱۰^{-n}$ آنگاه $\widehat{x}$ دارای دست کم n رقم بامعنای درست است.

مثال ۹.۱ *تقریبی از $\sqrt{۳}$ ارایه دهید که خطای نسبی آن از $۱۰^{-۴}$ کمتر باشد. بنابر قضیه ۷.۱ اگر $\widehat{x}$ تقریبی از $\sqrt{۳}$ باشد که ۵ رقم بامعنای درست داشته باشد آنگاه $\delta x < ۵ \times ۱۰^{-۵} < ۱۰^{-۴}$. از این‌رو، با توجه به قضیه ۶.۱ کافی است $\widehat{x}$ گرد شده $\sqrt{۳}$ تا ۵ رقم بامعنا باشد، یعنی $\widehat{x} = ۱/۷۳۲۱$.* △

۵.۱ خطای محاسبات (فرمول)

فرض کنید $z = f(x_۱, \ldots, x_n)$ تابعی باشد که می‌خواهیم آن را در نقطه $(x_۱, \ldots, x_n)$ ارزیابی کنیم. بدون کاستن از کلیت فرض کنید $\Delta x_i = x_i - \widehat{x}_i$ که در آن $\widehat{x}_i$ مقدار تقریبی x_i است. پس می‌توان نوشت

$$z = f(x_۱, \ldots, x_n) = f(\widehat{x}_۱ + \Delta x_۱, \ldots, \widehat{x}_n + \Delta x_n).$$

بنابر بسط تیلور توابع n متغیره داریم

$$z = f(\widehat{x}_۱, \ldots, \widehat{x}_n) + \left(\Delta x_۱ \frac{\partial f}{\partial x_۱} + \cdots + \Delta x_n \frac{\partial f}{\partial x_n}\right)(\widehat{x}_۱, \ldots, \widehat{x}_n) + R,$$

که در آن R جمله خطا بوده و شامل حاصل‌ضرب‌ها و توان‌های Δx_iها است و چون Δx_iها کوچک هستند از R چشم‌پوشی کرده، خواهیم داشت

$$f(x_۱, \ldots, x_n) - f(\widehat{x}_۱, \ldots, \widehat{x}_n) \simeq \left(\Delta x_۱ \frac{\partial f}{\partial x_۱} + \cdots + \Delta x_n \frac{\partial f}{\partial x_n}\right)(\widehat{x}_۱, \ldots, \widehat{x}_n),$$

و اگر $\widehat{z} = f(\widehat{x}_۱, \ldots, \widehat{x}_n)$ آنگاه

$$\Delta z = |z - \widehat{z}| \simeq \left|\left(\Delta x_۱ \frac{\partial f}{\partial x_۱} + \cdots + \Delta x_n \frac{\partial f}{\partial x_n}\right)(\widehat{x}_۱, \ldots, \widehat{x}_n)\right|,$$

و بلافاصله داریم

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

مثال ۱۰.۱ (**مستقیم**) *یک استوانه به شعاع قاعده $\frac{۴}{۳}$ و ارتفاع $\sqrt{۲}$ را در نظر بگیرید. اگر شعاع و ارتفاع استوانه و عدد π را با دقت $۴D$ وارد محاسبات کنیم، حجم این استوانه با چه خطایی به دست می‌آید؟*
می‌دانیم حجم یک استوانه از قاعده $\pi r^۲ h$ تعیین می‌شود که در آن r شعاع قاعده و h ارتفاع استوانه است. اگر تعریف کنیم $z = V(p, r, h) = pr^۲ h$، آنگاه باید V به ازای $p = \pi$، $r = \frac{۴}{۳}$ و $h = \sqrt{۲}$ ارزیابی شود. تمام محاسبات را با دقت $۴D$ دنبال می‌کنیم. بنابراین

$$\begin{aligned} p = \pi \quad &\rightarrow \quad \widehat{p} = ۳/۱۴۱۶ \\ r = \tfrac{۴}{۳} \quad &\rightarrow \quad \widehat{r} = ۱/۳۳۳۳ \\ h = \sqrt{۲} \quad &\rightarrow \quad \widehat{h} = ۱/۴۱۴۲ \end{aligned}$$

و داریم $\Delta p, \Delta r, \Delta h \leq ۰/۵ \times ۱۰^{-۴}$. بنابراین

$$\widehat{z} = V(\widehat{p}, \widehat{r}, \widehat{h}) = ۳/۴۱۱۶ \times ۱/۳۳۳۳^۲ \times ۱/۴۱۴۲ = ۷/۸۹۸۰.$$

از طرفی

$$\Delta z \simeq \Delta p \widehat{r}^۲ \widehat{h} + ۲\Delta r \widehat{p}\widehat{r}\widehat{h} + \Delta h \widehat{p}\widehat{r}^۲ \leq (\widehat{r}^۲ \widehat{h} + ۲\widehat{p}\widehat{r}\widehat{h} + \widehat{p}\widehat{r}^۲) \times ۰/۵ \times ۱۰^{-۴} < ۱۰^{-۳}.$$

در نتیجه حجم استوانه برابر است با $۷/۸۹۸ \pm ۰/۰۰۱$. به کمک یک ماشین حساب ۱۰ رقمی به دست می‌آوریم $z = ۷/۸۹۸۴۵۸۵۵۵$ و بنابراین خطای واقعی عبارت است از

$$\Delta z = |z - \widehat{z}| = ۴/۵۸۵۵۵۴ \times ۱۰^{-۴} < ۰/۰۰۱.$$

$\triangle$

مثال ۱۱.۱ (**معکوس**) *اعداد $x = \sqrt{۵}$ و $y = \frac{\pi}{۱۲}$ را با چه دقتی در نظر بگیریم تا مقدار $۶x^۲(\ln x + \sin ۲y)$ با دقت $۲D$ حساب شود؟ اگر فرض کنیم*

$$z = f(x, y) = ۶x^۲(\ln x + \sin ۲y),$$

آنگاه

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

و در نتیجه با فرض $x = ۲/۲$ *و* $y = ۰/۳$ *(*x *و* y *را با دقت* $۱S$ *یا* $۲S$ *در نظر می‌گیریم) داریم*

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

و برای برقراری نابرابری $\Delta z \leq ۰/۵ \times ۱۰^{-۲}$ *باید داشته باشیم*

$$\left|۴۸/۹\Delta x + ۴۷/۹\Delta y\right| \leq ۰/۵ \times ۱۰^{-۲}.$$

یک جواب نامعادله اخیر عبارت است از

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

بنابراین برای رسیدن به نتیجه مطلوب، x *و* y *را می‌توان با دقت* $۴D$ *در نظر گرفت.* △

مثال ۱۲.۱ *اگر در محاسبه* $z = ab^۲c^۳$*، خطای نسبی تقریب‌های* a*،* b *و* c *حداکثر* $۰/۰۱$ *باشد، بیشترین خطای نسبی قابل انتظار برای* z *چقدر است؟ با فرض* $z = f(a,b,c) = ab^۲c^۳$ *داریم*

$$\delta z \simeq \frac{\left|\Delta a\frac{\partial f}{\partial a} + \Delta b\frac{\partial f}{\partial b} + \Delta c\frac{\partial f}{\partial c}\right|}{|ab^۲c^۳|} = \frac{|b^۲c^۳\Delta a + ۲abc^۳\Delta b + ۳ab^۲c^۲\Delta c|}{|ab^۲c^۳|} \leq \delta a + ۲\delta b + ۳\delta c = ۰/۰۶.$$

△

۱.۵.۱ خطای اعمال ریاضی

هنگام کار با اعداد ماشینی (ممیز شناور) دقت شود که بعضی از اصول میدان $\mathbb{R}$ نظیر شرکت پذیری، منحصر به فرد بودن عضو خنثی و غیره برقرار نیست. به طور کلی، عبارت‌هایی که از نظر ریاضی معادل هستند ممکن است از نظر محاسباتی معادل نباشند. علت اصلی بروز این مشکلات، خطای گرد کردن است.

تعریف ۶.۱ *فرض کنید* A *و* B *دو عدد حقیقی،* a *و* b *تقریب‌هایی از آن‌ها و* $\otimes$ *بیان‌گر یک عمل دوتایی باشد. متناظر با* $A \otimes B$ *در ماشین عمل* $a \otimes^* b$ *انجام می‌شود و داریم*

$$|A \otimes B - a \otimes^* b| = |(A \otimes B - a \otimes b) + (a \otimes b - a \otimes^* b)| \leq \underbrace{|A \otimes B - a \otimes b|}_{\text{خطای منتشر شده}} + \underbrace{|a \otimes b - a \otimes^* b|}_{\text{خطای تولید شده}}$$

قضیه ۸.۱ *اگر* $\widehat{x}$ *و* $\widehat{y}$ *تقریب‌هایی از* x *و* y *بوده و همه این اعداد مثبت باشند، آنگاه*

۱. $\Delta(x \pm y) \leq \Delta x + \Delta y$، $\delta(x \pm y) \leq \frac{x}{|x \pm y|}\delta x + \frac{y}{|x \pm y|}\delta y$، $\delta(x + y) \leq \max\{\delta x, \delta y\}$،

۲. $\Delta(xy) \leq x\Delta y + y\Delta x$، $\delta(xy) \leq \delta x + \delta y$،

۳. $\Delta\left(\frac{x}{y}\right) \leq \frac{y\Delta x + x\Delta y}{y^۲}$، $\delta\left(\frac{x}{y}\right) \leq \delta x + \delta y$.

مثال ۱۳.۱ *اگر $\widehat{x}$ و $\widehat{y}$ هر یک n رقم بامعنای درست داشته باشند، حداقل تعداد ارقام بامعنای درست $\widehat{x}+\widehat{y}$ و $\widehat{x}\widehat{y}$ را تعیین کنید. بنابر قضیه ۷.۱، $\delta x < ۵ \times ۱۰^{-n}$ و $\delta y < ۵ \times ۱۰^{-n}$. هم‌چنین داریم*

$$\delta(x+y) \leq \max\{\delta x, \delta y\} < ۵ \times ۱۰^{-n} = ۵ \times ۱۰^{-(n-۱)-۱}.$$

پس $\widehat{x}+\widehat{y}$ حداقل $n-۱$ رقم بامعنای درست دارد. به علاوه می‌توان نوشت

$$\delta(\widehat{x}\widehat{y}) \leq \delta x + \delta y < ۵ \times ۱۰^{-n} + ۵ \times ۱۰^{-n} = ۱۰ \times ۱۰^{-n} = ۱۰^{-(n-۲)-۱} < ۵ \times ۱۰^{-(n-۲)-۱}.$$

بنابراین $\widehat{x}\widehat{y}$ دست کم $n-۲$ رقم بامعنای درست دارد. △

تذکر ۱۳.۱ *با توجه به خطای نسبی $\delta(x-y) \simeq \frac{\Delta(x-y)}{|x-y|}$ واضح است که اگر $\widehat{x}$ و $\widehat{y}$ دو عدد هم‌علامت نزدیک به هم باشند $|x-y|$ کوچک و در نتیجه $\delta(x-y)$ بزرگ خواهد شد و در نتیجه تعداد ارقام بامعنای $\widehat{x}-\widehat{y}$ کم خواهد بود. بنابراین در عمل بهتر است از تفاضل دو عدد هم‌علامت نزدیک به هم جلوگیری شود (تفاضل دو عدد هم‌علامت نزدیک به هم موجب از بین رفتن ارقام بامعنا[۱۵] می‌شود مانند $۱/۴۲ - ۱/۴۱ = ۰/۰۱$). اگر تفاضل اجتناب‌ناپذیر است باید عمل با دقت دو برابر (یا بیشتر) انجام شود.*

مثال ۱۴.۱ *جهت اجتناب از تفاضل در محاسبات می‌توان از اتحادها کمک گرفت. به عنوان مثال*

$$e^{a-b} = \frac{e^a}{e^b}, \qquad ۱ - \cos x = ۲\sin^۲\frac{x}{۲}, \qquad \ln a - \ln b = \ln\frac{a}{b}.$$

△

مثال ۱۵.۱ *با فرض $g(x) = x\left(\sqrt[۲]{۱+\frac{۱}{x}} - ۱\right)$، مقدار $g(۱۰^۹)$ را با دقت یک ماشین حساب ۱۰ رقمی به دست آورید. تمام محاسبات را با دقت $۹D$ انجام می‌دهیم. پس*

$$۱ + \frac{۱}{x} = ۱/۰۰۰۰۰۰۰۰۱, \qquad \sqrt[۲]{۱+\frac{۱}{x}} = ۱/۰۰۰۰۰۰۰۰۰.$$

بنابراین با دقت $۹D$، ماشین حساب نتیجه زیر را به دست می‌دهد

$$g(۱۰^۹) = ۱۰^۹(۱/۰۰۰۰۰۰۰۰۰ - ۱) = ۰/۰۰۰۰۰۰۰۰۰.$$

[۱۵] Cancellation

به این ترتیب با یک ماشین حساب ۱۰ رقمی $g(۱۰^۹) = ۰/۰۰۰۰۰۰۰۰۰$*، که نتیجه‌ای نادرست است. در حالی که اگر از یک رایانه استفاده شود، نتیجه درست* $g(۱۰^۹) = ۰/۳۳۳۳۳۳۳۳۳$ *به دست می‌آید. دلیل این خطای فاحش، تفاضل دو عدد نزدیک به هم در محاسبات است که منجر به از بین رفتن ارقام بامعنا می‌شود. برای به دست آوردن تقریب بهتر، روش محاسبه را به صورت زیر تغییر می‌دهیم*

$$g(x) = x\left(\sqrt[۳]{۱+\frac{۱}{x}} - ۱\right) \times \frac{\sqrt[۳]{(۱+\frac{۱}{x})^۲} + \sqrt[۳]{۱+\frac{۱}{x}} + ۱}{\sqrt[۳]{(۱+\frac{۱}{x})^۲} + \sqrt[۳]{۱+\frac{۱}{x}} + ۱} = \frac{۱}{\sqrt[۳]{(۱+\frac{۱}{x})^۲} + \sqrt[۳]{۱+\frac{۱}{x}} + ۱}.$$

در این صورت با توجه به

$$\left(۱+\frac{۱}{x}\right)^۲ = ۱/۰۰۰۰۰۰۰۰۲, \quad \sqrt[۳]{\left(۱+\frac{۱}{x}\right)^۲} = ۱/۰۰۰۰۰۰۰۰۰,$$

با دقت $۹D$ *نتیجه زیر به دست می‌آید*

$$g(۱۰^۹) = \frac{۱}{۱/۰۰۰۰۰۰۰۰۰ + ۱/۰۰۰۰۰۰۰۰۰ + ۱} = ۰/۳۳۳۳۳۳۳۳۳.$$

△

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 قایل شویم.

تذکر ۱۴.۱ *با توجه به قضیه ۸.۱، در محاسبات باید از ضرب اعداد بزرگ در اعداد تقریبی (تقسیم اعداد تقریبی به اعداد کوچک) پرهیز کرد.*

مثال ۱۶.۱ *در محاسبه* $x = ۱۰۰۰۰\pi$ *داریم*

$$\begin{aligned} \pi = ۳/۱۴ \quad &\to \quad \widehat{x} = ۳۱۴۰۰, \\ \pi = ۳/۱۴۲ \quad &\to \quad \widehat{x} = ۳۱۴۲۰, \\ \pi = ۳/۱۴۱۶ \quad &\to \quad \widehat{x} = ۳۱۴۱۶. \end{aligned}$$

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

تذکر ۱۵.۱ *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 ۱۷.۱ *به جای عبارت* $ax^۳ + bx^۲ + cx + d$ *از عبارت* $((ax+b)x+c)x+d$ *استفاده شود.* △

۲.۵.۱ تقریب توابع یک متغیره

قضیه ۹.۱ (تیلور با باقیمانده لاگرانژ) *فرض کنید $f \in C^n[a,b]$ (یعنی تابع f و مشتقات تا مرتبه n آن روی بازه $[a,b]$ پیوسته هستند) و $f^{(n+1)}$ بر (a,b) موجود باشد و $x_0 \in [a,b]$. در این صورت، به ازای هر $x \in [a,b]$، نقطه‌ای مانند $y(x)$ بین x_0 و x وجود دارد که*

$$f(x) = p_n(x) + R_n(x, x_0)$$

که در آن

$$p_n(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!}(x-x_0)^i$$

و

$$R_n(x, x_0) = \frac{f^{(n+1)}(y(x))}{(n+1)!}(x-x_0)^{n+1}.$$

در این‌جا p_n چندجمله‌ای تیلور مرتبه nام f حول x_0 و $R_n(x,x_0)$ جمله باقی‌مانده (یا خطای برش[۱۶]) متناظر با p_n نامیده می‌شود. اگر $n \to \infty$ آنگاه p_n به یک سری بی‌پایان تبدیل می‌شود که به آن سری تیلور f حول نقطه x_0 گویند. در این حالت، شرط بی‌نهایت بار مشتق‌پذیر بودن f در x_0 لازم است.

تذکر ۱۶.۱ *در قضیه قبل اگر $x_0 = 0$ آنگاه واژه تیلور به مکلورن تبدیل می‌شود.*

تذکر ۱۷.۱ *$R_n(x,x_0)$ مقدار خطا در استفاده از p_n به جای f را نشان می‌دهد. در عمل با یافتن کرانی برای جمله باقی‌مانده، در واقع برای خطای تقریب f با p_n کرانی پیدا می‌کنیم.*

تذکر ۱۸.۱ *ویژگی مهم چندجمله‌ای تیلور مرتبه nام آن است که p_n و مشتقات تا مرتبه‌ی nام آن با f و مشتقات تا مرتبه nام آن در نقطه x_0 برابر هستند.*

تذکر ۱۹.۱ *شکل دیگر (کاربردی) قضیه تیلور به صورت زیر است*

$$f(x_0+h) = f(x_0) + f'(x_0)h + \cdots + \frac{f^{(n)}(x_0)}{n!}h^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1} = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!}h^i + \frac{f^{(n+1)}(y)}{(n+1)!}h^{n+1}$$

و یا

$$f(x) = f(x_0+h) = f(x_0)+f'(x_0)h+\cdots+\frac{f^{(n)}(x_0)}{n!}h^n+\frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1} = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!}h^i+\frac{f^{(n+1)}(y)}{(n+1)!}h^{n+1},$$

که در آن $h = x - x_0$.

---

[۱۶] Truncation error

مثال ۱۸.۱ *فرض کنید* $f(x) = ۲ + ۴x - x^۳$. *آنگاه به وضوح برای* $n \geq ۳$ *داریم*

$$p_n(x) = ۲ + ۴x - x^۳, \qquad R_n(x, ۰) = ۰,$$

و برای $n = ۲$ *خواهیم داشت*

$$p_۲(x) = ۲ + ۴x, \qquad R_۲(x, ۰) = -x^۳,$$

و برای $n = ۱$ *می‌توان نوشت*

$$p_۱(x) = ۲ + ۴x, \qquad R_۱(x, ۰) = -۳x^۲ y(x),$$

که در آن $y(x)$ *نقطه‌ای بین* ۰ *و* x *است.* △

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

مثال ۱۹.۱ *مطلوب است محاسبه مقدار* $e^{\frac{\pi}{۱۰}}$ *با دقت* $۲D$.
روش اول- به کمک قضیه تیلور داریم

$$e^x = ۱ + x + \frac{x^۲}{۲!} + \frac{x^۳}{۳!} + \cdots + \frac{x^n}{n!} + \frac{x^{n+۱}}{(n+۱)!} e^{y(x)},$$

که در آن $۰ < y(x) < x$. *در نتیجه*

$$e^{\frac{\pi}{۱۰}} = ۱ + \frac{\pi}{۱۰} + \frac{(\frac{\pi}{۱۰})^۲}{۲!} + \cdots + \frac{(\frac{\pi}{۱۰})^n}{n!} + \frac{(\frac{\pi}{۱۰})^{n+۱}}{(n+۱)!} e^y,$$

که در آن $۰ < y < \frac{\pi}{۱۰} < ۱$. *چون* e^x *تابعی صعودی است پس* $۱ = e^۰ < e^y < e^۱ < ۳$. *از طرفی* $\frac{\pi}{۱۰} < ۰/۳۱۵$ *(دقت* $۳D$ *منظور می‌شود) و بنابراین*

$$\text{خطا} \leq \left| \frac{(\frac{\pi}{۱۰})^{n+۱}}{(n+۱)!} e^\xi \right| < \frac{۳ \times (۰/۳۱۵)^{n+۱}}{(n+۱)!}.$$

حال باید داشته باشیم $\frac{۳\times(۰/۳۱۵)^{n+۱}}{(n+۱)!} < ۰/۵ \times ۱۰^{-۲}$ *که نتیجه می‌دهد* $n \geq ۳$. *پس*

$$e^{\frac{\pi}{۱۰}} \simeq ۱ + ۰/۳۱۵ + \frac{(۰/۳۱۵)^۲}{۲!} + \frac{(۰/۳۱۵)^۳}{۳!} = ۱/۳۷۰,$$

و با دقت $۲D$ *داریم* $e^{\frac{\pi}{۱۰}} \simeq ۱/۳۷$ *که با جواب ماشین حساب یعنی* $۱/۳۶۹۱۰۷۷۷۱$ *کمتر از* $۰/۰۰۱$ *اختلاف دارد.*
روش دوم- به کمک سری تیلور داریم

$$e^x = ۱ + x + \frac{x^۲}{۲!} + \frac{x^۳}{۳!} + \cdots + \frac{x^n}{n!} + \cdots.$$

در نتیجه

$$e^{\frac{\pi}{۱۰}} = ۱ + \frac{\pi}{۱۰} + \frac{(\frac{\pi}{۱۰})^۲}{۲!} + \cdots + \frac{(\frac{\pi}{۱۰})^n}{n!} + \cdots.$$

با توجه به $\frac{\pi}{۱۰} < ۰/۳۱۵$ *(دقت* $۳D$ *منظور می‌شود) از* $\left|\frac{(۰/۳۱۵)^n}{n!}\right| < ۰/۵ \times ۱۰^{-۲}$ *نتیجه می‌شود* $n \geq ۴$*. پس*

$$e^{\frac{\pi}{۱۰}} \simeq ۱ + ۰/۳۱۵ + \frac{(۰/۳۱۵)^۲}{۲!} + \frac{(۰/۳۱۵)^۳}{۳!} + \frac{(۰/۳۱۵)^۴}{۴!}$$
$$= ۱/۳۱۵ + ۰/۰۵۰ + ۰/۰۰۵ + ۰/۰۰۰ = ۱/۳۷۰.$$

△

تذکر ۲۰.۱ *بعضی مواقع ممکن است تعداد جملاتی که از روش دوم به دست می‌آید کافی نباشد و بهتر است با جملات بیشتر هم مقایسه کرد.*

مثال ۲۰.۱ *(همگرایی سریع) می‌خواهیم تابع* $\cos x$ *را به ازای مقادیر* $|x| < \frac{\pi}{۴}$ *با دقت* $۵D$ *ارزیابی کنیم. با توجه به سری تیلور*

$$\cos x = \sum_{i=۰}^{\infty} (-۱)^i \frac{x^{۲i}}{(۲i)!},$$

و این که

$$\left|\frac{x^{۲n}}{(۲n)!}\right| < \frac{(\frac{\pi}{۴})^{۲n}}{(۲n)!} < \frac{۱/۶^{۲n}}{(۲n)!},$$

باید داشته باشیم

$$\frac{۱/۶^{۲n}}{(۲n)!} < ۰/۵ \times ۱۰^{-۵},$$

که نتیجه می‌دهد $n \geq ۶$*.* △

مثال ۲۱.۱ *(همگرایی کند) با توجه به*

$$\frac{۱}{۱+t} = ۱ - t + t^۲ - t^۳ + - \cdots,$$

داریم

$$\int_۰^x \frac{dt}{۱+t} = \int_۰^x (۱ - t + t^۲ - t^۳ + - \cdots) dt,$$

و یا

$$\ln(۱+x) = x - \frac{x^۲}{۲} + \frac{x^۳}{۳} - \frac{x^۴}{۴} + - \cdots.$$

برای ارزیابی $\ln(۱+x)$ *با دقت* $۵D$*، باید داشته باشیم*

$$\left|\frac{x^n}{n}\right| < ۰/۵ \times ۱۰^{-۵},$$

که برای $x = ۰/۹۹$ *نتیجه می‌دهد* $n \geq ۵۸۲$. △

مثال ۲۲.۱ *تابع* $Si(x) = \int_۰^x \frac{\sin t}{t} dt$ *برای* $x \geq ۰$ *را در نظر بگیرید. دست کم چند جمله از بسط مک‌لورن تابع* $f(t) = \sin t$ *لازم است تا* $Si(۱)$ *با دقت* $۶D$ *مشخص شود؟*

با اِعمال قضیه تیلور برای تابع f *داریم*

$$Si(۱) = \int_۰^۱ \frac{۱}{t} \left(f(۰) + \frac{f'(۰)}{۱!}t + \frac{f''(۰)}{۲!}t^۲ + \cdots + \frac{f^{(k)}(۰)}{k!}t^k + \frac{f^{(k+۱)}(y(t))}{(k+۱)!}t^{k+۱} \right) dt,$$

و یا

$$Si(۱) = \int_۰^۱ \frac{۱}{t} \left(t - \frac{t^۳}{۳!} + \frac{t^۵}{۵!} + \cdots + (-۱)^n \frac{t^{۲n+۱}}{(۲n+۱)!} + (-۱)^{n+۱} \frac{t^{۲n+۲}}{(۲n+۲)!} \sin(y(t)) \right) dt.$$

در نتیجه

$$Si(۱) = \int_۰^۱ \left(۱ - \frac{t^۲}{۳!} + \frac{t^۴}{۵!} + \cdots + (-۱)^n \frac{t^{۲n}}{(۲n+۱)!} \right) dt + \frac{(-۱)^{n+۱}}{(۲n+۲)!} \int_۰^۱ t^{۲n+۱} \sin(y(t)) dt.$$

بنابراین

$$Si(۱) = ۱ - \frac{۱}{۳ \times ۳!} + \frac{۱}{۵ \times ۵!} + \cdots + (-۱)^n \frac{۱}{(۲n+۱) \times (۲n+۱)!} + E,$$

که در آن

$$E = \frac{(-۱)^{n+۱}}{(۲n+۲)!} \int_۰^۱ t^{۲n+۱} \sin(y(t)) dt.$$

حال می‌توان نوشت

$$|E| = \frac{۱}{(۲n+۲)!} \left| \int_۰^۱ t^{۲n+۱} \sin(y(t)) dt \right| \leq \frac{۱}{(۲n+۲)!} \int_۰^۱ t^{۲n+۱} \left| \sin(y(t)) \right| dt.$$

پس

$$|E| \leq \frac{۱}{(۲n+۲)!} \int_۰^۱ t^{۲n+۱} dt = \frac{۱}{(۲n+۲) \times (۲n+۲)!}.$$

از $\frac{۱}{(۲n+۲) \times (۲n+۲)!} < ۰/۵ \times ۱۰^{-۶}$ *نتیجه می‌شود* $n \geq ۴$. *بنابراین*

$$Si(۱) \simeq ۱ - \frac{۱}{۳ \times ۳!} + \frac{۱}{۵ \times ۵!} - \frac{۱}{۷ \times ۷!} + \frac{۱}{۹ \times ۹!},$$

و یا

$$Si(۱) \simeq ۱ - ۰/۰۵۵۵۵۵۶ + ۰/۰۰۱۶۶۶۷ - ۰/۰۰۰۰۲۸۳ + ۰/۰۰۰۰۰۰۳ = ۰/۹۴۶۰۸۳۱.$$

پس با دقت $۶D$ *داریم* $Si(۱) \simeq ۰/۹۴۶۰۸۳$. △

تمرین

۱. 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؟

۲. مطلوب است تعیین تقریبی از عدد π با دقت $۳D$ به کمک بسط مکلورن تابع $f(x) = \tan^{-۱} x$.

۳. حداقل چند جمله از بسط مک‌لورن تابع $f(x) = e^{-\frac{x}{۲}}$ نیاز است تا $f(\ln(۲))$ با دقت $۳D$ به دست آید؟

۴. حداکثر خطای نسبی محاسبه حجم یک مخروط از فرمول $V = \frac{\pi}{۳}r^۲h$ چقدر است، هرگاه تقریبی از عدد π با خطای نسبی حداکثر $۰/۰۰۳$ داشته باشیم و کمیت‌های r و h را بتوان با حداکثر خطای نسبی $۰/۰۰۲$ اندازه‌گیری کرد.

۵. برای محاسبه عبارت $e^{۱-\ln(۲/۷۳)}$ در یک ماشین‌حساب با سه رقم بامعنا کدام گزینه مناسب‌تر است؟

الف) $e^{۱-\ln(۲/۷۳)}$ ب) $\frac{e}{۲/۷۳}$ ج) $e \times e^{-\ln(۲/۷۳)}$ د) $\frac{e}{e^{\ln(۲/۷۳)}}$

۶. برای همه گزینه‌ها $a_۰$ تقریبی از عدد π با دقت $۳S$ است. در کدام مورد خطای مطلق در محاسبه $a_{۱۰}$ کمتر است؟

الف) $a_n = ۱ + ۲a_{n-۱}$ ب) $a_n = ۱ + na_{n-۱}$ ج) $a_n = ۱ + a_{n-۱}$ د) $a_n = ۱ + \frac{۱}{۲}a_{n-۱}$

۷. برای محاسبه عبارت $\frac{۱-\cos(x)}{x}$ در کامپیوتر، وقتی x عدد مثبت کوچکی است، کدام گزینه مناسب‌تر است؟

الف) $\frac{۱-\cos(x)}{x}$ ب) $\frac{۲\sin^۲(\frac{x}{۲})}{x}$ ج) $\frac{\sin^۲(x)}{x(۱+\cos(x)}$ د) $\frac{۱}{x} - \frac{\cos(x)}{x}$

۸. برای محاسبه $\sqrt{x^۲ + \frac{۱}{x^۲}} - \sqrt{x^۲ - \frac{۱}{x^۲}}$ برای xهای خیلی بزرگ در ماشین کدام گزینه مناسب‌تر است؟

الف) $\frac{۱}{x^۲}$ ب) $\frac{۲}{x\left(\sqrt{x^۴+۱}+\sqrt{x^۴-۱}\right)}$ ج) $\sqrt{x^۲ + \frac{۱}{x^۲}} - \sqrt{x^۲ - \frac{۱}{x^۲}}$ د) $\frac{۱}{x}\left(\sqrt{x^۴ + ۱} - \sqrt{x^۴ - ۱}\right)$

۹. در یک ماشین با دقت $۴S$ و گرد کردن، کوچکترین عدد طبیعی x که در تساوی $۱ + x = x$ صدق می‌کند، کدام گزینه است؟ الف) $۰/۵۰۰۰ \times ۱۰^۴$ ب) $۰/۱۰۰۰ \times ۱۰^۵$ ج) $۰/۵۰۰۰ \times ۱۰^۵$ د) $۰/۱۰۰۰ \times ۱۰^۶$

۱۰. برای زوایای کمتر از ۶ درجه، $\sin(x)$ را با مقدار x بر حسب رادیان تقریب می‌زنند. کران بالای خطای مطلق این کار چقدر است؟

۱۱. حداقل چند جمله از بسط مک‌لورن تابع $f(x) = \sin(\frac{x}{۳})$ لازم است تا $\sin(۰/۱)$ با دقت $۴D$ مشخص شود؟

۱۲. 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 $۱۰^{-۲}$ مشخص کرد؟

الف) $۰/۵٪$ و $۰/۲۵٪$ ب) $۰/۶٪$ و $۰/۳٪$ ج) $۰/۵٪$ و $۰/۳۵٪$ د) $۰/۵۵٪$ و $۰/۲۵٪$

۱۳. می‌خواهیم کمیت‌های U و V را از فرمول‌های $U = x^۲ \div y$ و $V = xy^۳$ 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 $۱۰^{-۲}$ مشخص کرد؟

الف) ٪۰/۲ و ٪۰/۳۵ ب) ٪۰/۲۵ و ٪۰/۳ ج) ٪۰/۳ و ٪۰/۲۵ د) ٪۰/۳۵ و ٪۰/۲

۱۴. 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.

۱۵. حداکثر خطای نسبی محاسبه حجم یک کره از فرمول $V = \frac{۴}{۳}\pi r^۳$ چقدر است، هرگاه تقریبی از اعداد π و $\frac{۱}{۳}$ با خطای نسبی حداکثر ۰/۰۰۲ داشته باشیم و شعاع را بتوان با حداکثر خطای نسبی ۰/۰۰۲ اندازه‌گیری کرد؟

الف) ۰/۰۰۱ ب) ۰/۰۱۰ ج) ۰/۰۲۰ د) ۰/۰۰۲

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

الف) در محاسبه تقریبی $\frac{\sqrt{۲}}{۱۰۰۰۰۰}$ قطعا خطای زیادی داریم.

ب) در محاسبه تقریبی $۱۰۰۰۰۰\sqrt{۲}$ قطعا 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.

۱۷. در محاسبه عبارت $\frac{e^{x^۲} - \cos x}{x^۲}$ 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.

۱۸. در یک دستگاه ممیز شناور که در آن اعداد به صورت $\pm ۰/d_۱d_۲d_۳d_۴ \times ۱۰^n$ با $d_۱ \neq ۰$ و $-۷ \leq n \leq ۸$ و $۰ \leq d_i \leq ۹$ نمایش داده می‌شوند فاصله بین عدد ۱۰۰۰۰ و اولین عدد قابل نمایش بزرگتر از ۱۰۰۰۰ کدام است؟

الف) ۱۰ ب) ۰/۱ ج) ۰/۰۱ د) ۰/۰۰۵

۱۹. اگر مقدار $\ln(۱ + x)$ را با استفاده از دو جمله اول بسط مکلورن آن تقریب بزنیم، مقدار تقریبی و حداکثر خطای محاسبه $\ln(۱/۱)$ به ترتیب کدام است؟

الف) (۰/۰۰۱, ۰/۰۵) ب) (۰/۱, ۰/۰۵) ج) (۰/۰۵, ۰/۰۰۵) د) (۰/۱, ۰/۰۰۵)

۲۰. در محاسبه دوره تناوب آونگ ساده $T = ۲\pi\sqrt{\frac{l}{g}}$، 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) π, l, g اثر یکسانی دارند.

۲۱. در یک دستگاه ممیز شناور که در آن اعداد به صورت $\pm ۰/d_۱d_۲d_۳d_۴ \times ۱۰^n$ با $d_۱ \neq ۰$ و $۰ \leq d_i \leq ۹$ و $-۷ \leq n \leq ۸$ با گرد کردن نمایش داده می‌شوند، بزرگترین عدد x که در معادله $۵۰۰ + x = ۵۰۰$ صدق می‌کند، کدام است؟

الف) ۰/۰۴۹۹۹ ب) ۰/۰۹۹۹۹ ج) ۰/۰۵۰۰۰ د) ۰/۱۰۰۰

۲۲. 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.

۲۳. گزینه مناسب برای محاسبه $T = \frac{۱ - \cos x}{\sin x}$ به ازای مقادیر x نزدیک صفر کدام است؟
الف) ۰ ب) $\frac{x}{۲}$ ج) $\frac{\pi}{۲}$ د) $\frac{۱}{\sin x} - \cot x$

۲۴. در محاسبه تقریبی $\ln(۱/۰۰۱)$ با سه جمله از بسط تیلور، حداکثر خطا چقدر است؟
الف) $\frac{۱۰^{-۶}}{۳}$ ب) $\frac{۱۰^{-۹}}{۳}$ ج) $\frac{۱۰^{-۴}}{۳}$ د) $\frac{۱۰^{-۶}}{۲}$

۲۵. در محاسبه حجم ($V = \frac{۴}{۳}\pi r^۳$) 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 $\frac{۴}{۳}$ بیشتر است. د) هر سه اثر خطای نسبی یکسانی دارند.

۲۶. گزینه مناسب برای محاسبه $T = \frac{\cos(۲x) - ۱ + ۲x^۲}{x^۴}$ به ازای مقادیر x نزدیک صفر کدام است؟
الف) $\frac{۳}{۲}$ ب) $\frac{۲}{۳}$ ج) $\frac{۲x}{۳}$ د) $\frac{۳x}{۲}$

۲۷. برای محاسبه $f(۳)$ با خطایی کمتر از ۰/۰۰۰۷ چند جمله از بسط تیلور تابع f حول نقطه $x_۰ = ۲/۵$ لازم است (می‌دانیم $|f^{(n)}(x)| \leq \frac{۱}{n}$)؟
الف) دو جمله ب) سه جمله ج) پنج جمله د) چهار جمله

۲۸. در یک دستگاه ممیز شناور نرمال‌شده که هر عدد حقیقی به صورت $\pm ۰/d_۱d_۲d_۳d_۴ \times ۱۰^e$ با $d_۱ \neq ۰$ و $۰ \leq d_i \leq ۹$ برای $i = ۱,۲,۳,۴$ و $-۶ \leq e \leq ۷$، نمایش داده می‌شود، اعداد قابل نمایش قبل و بعد از ۱۰۰۰۰ کدام هستند؟
الف) ۹۹۹۹, ۱۰۰۰۱ ب) ۹۹۹۰, ۱۰۰۱۰ ج) ۹۹۹۹, ۱۰۰۱۰ د) ۹۹۹۰, ۱۰۰۰۱

۲۹. در دستگاه ممیز شناور نرمال‌شده سوال قبل، تعداد اعداد قابل نمایش در بازه $[۱, ۱۰]$ کدام است؟
الف) ۹۰۰۱ ب) ۹۰۰۰ ج) ۸۹۹۹ د) ۹۰۰۲