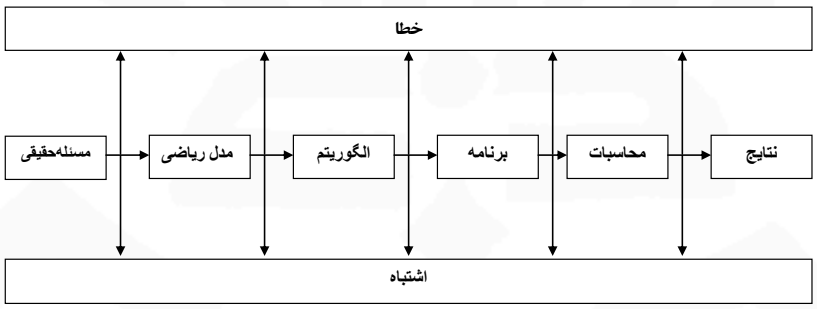
فصل ۲

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 ۲۳۲۳ به جای ۲۳۳۲ 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*

$$\frac{d^2\theta}{dt^2} = -\frac{g}{l}\sin\theta, \tag{۱.۲}$$

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

$$\theta = 6^\circ \simeq 0/1047^{rad} \ \rightarrow\ \sin\theta \simeq 0/1045, \quad \theta = 15^\circ \simeq 0/262^{rad} \ \rightarrow\ \sin\theta \simeq 0/259,$$

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

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

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 $k = 1$، $M = \left(\frac{2k-1}{2k}\right)^2$، $T_1 = T_L$، $C = T_L \times M \times \sin^{2k}\frac{\theta_m}{2}$ و $T_2 = T_1 + C$.

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

۴) قرار دهید $k = k + 1$، $M = M \times \left(\frac{2k-1}{2k}\right)^2$، $T_1 = T_2$، $C = T_L \times M \times \sin^{2k}\frac{\theta_m}{2}$ و $T_2 = T_1 + C$.

۵) قرار دهید $T_N = T_2$ و T_L، T_N 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 گزارش‌شده در جدول ۱.۲ را تولید کرد. باید توجه داشت که برای $\theta_m < 15^\circ$ اختلاف T_l و T_N از $0/005$ بیشتر نیست.* △

با توجه به روندنمای ارائه‌شده 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 صرف‌نظرها، 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 $\frac{l}{g}$)
- خطای روش‌های (الگوریتم‌های) عددی (محاسباتی) (مانند خطای روش محاسبه $\sqrt{\frac{l}{g}}$)

```
l = Input["طول آونگ را بر حسب متر وارد کنید"];
θm = Input["جابجایی زاویه‌ای ماکزیمم را بر حسب درجه وارد کنید"];
ϵ = Input["مقدار مجاز خطا را وارد کنید"];
st = Sin[θm * Pi/۱۸۰];
g = ۹/۸;  TL = ۲ * π * √(l/g);  k = ۱;  M = ((۲k−۱)/(۲k))^۲;
CC = TL * M * st^(۲k);  T۱ = TL;  T۲ = T۱ + CC;
While[CC > ϵ, {
  k++;
  M* = ((۲k−۱)/(۲k))^۲;
  T۱ = T۲;
  CC = TL * M * st^(۲k);
  T۲ = T۱ + CC;
}];
TN = T۲;
Print["T_L = ", TL, "  T_N = ", TN, "  T_N − T_L = ", 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

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*

$$\begin{aligned} x &= a_m\beta^m + a_{m-1}\beta^{m-1} + \cdots + a_1\beta^1 + a_0\beta^0 + a_{-1}\beta^{-1} + a_{-2}\beta^{-2} + \cdots \\ &= (a_m a_{m-1} \ldots a_1 a_0 / a_{-1} a_{-2} \ldots)_\beta, \end{aligned} \tag{۲.۲}$$

دارد که در آن $m \in \mathbb{Z}$ و $a_i \in \{0, 1, 2, \ldots, \beta - 1\}$. برای منحصر به فرد بودن این نمایش لازم است که $a_m \neq 0$ و عدد صحیح j چنان وجود داشته باشد که $a_{j-1} = a_j = 0 \cdots$، 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 سر و 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 ۲.۲ *شرط دوم منحصر به فرد بودن نمایش، لازم است چه در غیر این صورت به عنوان مثال می‌توان نوشت*

$$\begin{aligned} ۳/۴۷۹۹۹\cdots = ۳/۴۷\bar{۹} &= ۳ \times ۱۰^{0} + ۴ \times ۱۰^{-1} + ۷ \times ۱۰^{-2} + ۹ \times ۱۰^{-3} + ۹ \times ۱۰^{-4} + \cdots \\ &= ۳/۴۷ + \tfrac{۹\times ۱۰^{-3}}{۱-۱۰^{-1}} = ۳/۴۷ + ۰/۰۱ = ۳/۴۸. \end{aligned}$$

یعنی بسط ده‌دهی بیان‌شده را می‌توان برای دو عدد $۳/۴۷۹۹۹\cdots$ و $۳/۴۸$ 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.*

$$\frac{۲}{۳} = ۰/۶۶۶\cdots = ۰/\bar{۶} = (۰/۲)_۳, \quad \frac{۳}{۸} = ۰/۳۷۵ = (۰/۰۱۱)_۲, \quad \frac{۱}{۷} = ۰/\overline{۱۴۲۸۵۷} = (۰/۱)_۷,$$
$$\sqrt{۲} = ۱/۴۱۴۲۱۳\cdots, \qquad \pi = ۳/۱۴۱۵۹۲\cdots, \qquad \exp(۱) = e = ۲/۷۱۸۲۸۱\cdots.$$

عدد $۰/۱۹۱۹۱۹\cdots$ گویا است ولی اعداد $۰/۱۲۲۳۳۳۴۴۴۴\cdots$ و $۰/۱۰۲۰۰۳۰۰۰۴۰۰۰۰۰\cdots$ 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 می‌کنند. بنابراین مسئله تغییر مبنا مطرح می‌شود.

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 می‌بریم.*

$$(۱۰۱۱۰۰۱/۱۱۱۰۰۱)_۲ = ۲^۶ + ۲^۴ + ۲^۳ + ۲^۰ + ۲^{-۱} + ۲^{-۲} + ۲^{-۳} + ۲^{-۶}$$
$$= ۶۴ + ۱۶ + ۸ + ۱ + ۰/۵ + ۰/۲۵ + ۰/۱۲۵ + ۰/۰۱۵۶۲۵ = ۸۹/۸۹۰۶۲۵$$

بنابراین تغییر مبنا از مبنای ۲ به مبنای ۱۰ به کمک بسط به سادگی امکان‌پذیر است. △

مثال ۴.۲ *در این مثال یک عدد صحیح مثبت را از مبنای ۱۰ به مبنای ۲ می‌بریم.*

$$۲۳ = ۱۱ \times ۲ + ۱, \qquad ۱۱ = ۵ \times ۲ + ۱, \qquad ۵ = ۲ \times ۲ + ۱, \qquad ۲ = ۱ \times ۲ + ۰.$$

پس $(۱۰۱۱۱)_۲ = ۲۳$. *بنابراین از تقسیم‌های متوالی بر ۲ استفاده می‌کنیم.* △

اگر عدد داده‌شده x در مبنای ۱۰ در فاصله $(۰, ۱)$ باشد و عدد متناظر در مبنای ۲ به صورت $(۰/b_۱b_۲b_۳\cdots)_۲$ فرض شود آن‌گاه

$$x = b_۱ \times ۲^{-۱} + b_۲ \times ۲^{-۲} + b_۳ \times ۲^{-۳} + \cdots,$$

و در نتیجه $۲x = b_۱ + b_۲ \times ۲^{-۱} + b_۳ \times ۲^{-۲} + \cdots$ و با قرار دادن $y = b_۲ \times ۲^{-۱} + b_۳ \times ۲^{-۲} + \cdots$ خواهیم داشت

$$۰ \le y < ۱ \times ۲^{-۱} + ۱ \times ۲^{-۲} + \cdots = \frac{۲^{-۱}}{۱ - ۲^{-۱}} = ۱.$$

پس $[y] = ۰$. از طرف دیگر داریم $۲x = b_۱ + y$ و بلافاصله از $[۲x] = [b_۱ + y]$ نتیجه می‌شود $b_۱ = [۲x]$ زیرا $b_۱$ عددی صحیح است ($b_۱$ یا صفر است یا یک). با قرار دادن $x = y$ و تکرار این روند می‌توان $b_۲$، $b_۳$ و سایر رقم‌ها 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.

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*

$$۰/۵۷۸۱۲۵ = (۰/۱۰۰۱۰۱)_۲, \qquad ۰/۱ = (۰/۰\overline{۰۰۱۱})_۲.$$

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

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.*

i	x	$۲x$	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		

i	x	$۲x$	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		

جدول ۲.۲: نمایش دودویی اعداد کوچک‌تر از واحد

۳.۲ نمایش اعداد در رایانه

برای نمایش اعداد در ماشین، ابتدا نمایشی به نام ممیز ثابت[۱] در نظر گرفته شد که در آن هر عدد حقیقی x به صورت زیر نمایش داده می‌شود

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

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)[۲] 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 استانداردی است که IEEE[۳] وضع کرده است.

۱.۳.۲ نمایش ۶۴-بیتی ممیز شناور

این نمایش پیش از این به دقت دو برابر (مضاعف)[۴] 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[۵] در نظر گرفته می‌شود و با c نمایش داده می‌شود و ۵۲ بیت باقی‌مانده نیز برای مانتیس کنار گذاشته می‌شود و آن را با f نشان می‌دهند.

سپس برای هر عدد نمایشی به صورت $(-۱)^s \times ۲^e \times (۱ + ۰/f)_۲$ در نظر گرفته می‌شود که در آن $e = c - ۱۰۲۳$ توان (نما) است. محدودیت $۰ \le c \le (۱۰۰۰۱)_۲ = ۲^{۱۱} - ۱ = ۲۰۴۷$ موجب می‌شود که $-۱۰۲۳ \le e \le ۱۰۲۴$. از

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

طرف دیگر محدودیت $(۰/\overbrace{۱۰۰۰۱}^{۵۲times})_۲ \le f \le (۰/\overbrace{۱۰۰۰۱}^{۵۲times})_۲ < ۲$ باعث می‌گردد که $۱ \le (۱+۰/f)_۲ \le (۱/\overbrace{۱۰۰۰۱}^{۵۲times})_۲ < ۲$. بنابراین اگر کوچک‌ترین و بزرگ‌ترین عدد مثبت قابل نمایش را به ترتیب با $+mN$ و $+MN$ نشان دهیم آن‌گاه

$$+mN = ۲^{-۱۰۲۲} \times ۱/۰ \simeq ۲/۲۲۵۰۷۴ \times ۱۰^{-۳۰۸}, \qquad +MN = ۲^{۱۰۲۴} \times ۱/۰ \simeq ۱/۷۹۷۶۹۳ \times ۱۰^{۳۰۸}.$$

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

$$+۰ = (۱/\underbrace{۱۰۰۰۱}_{۵۲times})_۲ \times ۲^{-۱۰۲۳}, \qquad +\infty = (۱/\underbrace{۰۰۰۰۰}_{۵۱times}۱)_۲ \times ۲^{۱۰۲۴}.$$

تمرین ۲.۲ *نمایش $-mN$، $-MN$، -۰ 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 ماشین[۶] معروف است و با* eps *نمایش داده می‌شود. نصف* eps*، به رُند واحد[۷] یا واحد گرد کردن[۸] یا دقت ماشین[۹] معروف است.*

چون در نمایش ۶۴ـبیتی بعد از ۱ عدد $(۱ \overbrace{۰۰۰۰۰}^{۵۱times}۱/)_۲$ قرار می‌گیرد، پس

$$\text{eps} = (۱/\underbrace{۰۰۰۰۰}_{۵۱times}۱)_۲ - ۱ = (۰/\underbrace{۰۰۰۰۰}_{۵۱times}۱)_۲ = ۲^{-۵۲} \simeq ۲/۲۲۰۴۴۶ \times ۱۰^{-۱۶}.$$

در ادامه بزرگ‌ترین عدد صحیح مثبت M را تعیین می‌کنیم که هر عدد صحیح x با شرط $۰ < x \le M$ را بتوان به طور دقیق نمایش داد. با توجه به موارد بیان‌شده، تمام اعداد صحیح نامنفی که بزرگ‌تر از $۲^{۵۲} \times ۲(۱/۱۰۰۰۱)_۲ = ۲^{۵۳} - ۱$ نباشند به طور دقیق قابل‌نمایش هستند و به علاوه $۲^{۵۳}$ نیز به صورت $۲^{۵۳} \times ۲(۱/۰)_۲$ قابل‌نمایش است. اما تعداد ارقام در مانتیس جهت نمایش $۲^{۵۳}+۱$ کافی نیست (۵۳ رقم در مانتیس لازم است). بنابراین $M = ۲^{۵۳} \simeq ۹/۰۰۷۱۹۹ \times ۱۰^{۱۵}$ و در نتیجه تمام اعداد صحیح ۱۵ رقمی و بسیاری از اعداد ۱۶ رقمی در این نمایش به طور دقیق قابل‌نمایش هستند به بیان دیگر نمایش ۶۴ـبیتی ۱۵ الی ۱۶ رقم دقت دارد. بنابراین اعدادی که بیشتر از ۱۶ رقم داشته باشند به طور دقیق قابل‌نمایش نیستند.

تذکر ۳.۲ *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 $(-۰, +۰)$ تولید شود پیام پاریز[۱۰] و اگر عدد تولیدشده از $+MN$ بزرگ‌تر یا از $-MN$ کوچک‌تر باشد پیام سرریز[۱۱] صادر می‌شود. برنامه* uoflow.nb *را ببینید.*

تذکر ۴.۲ *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.*

---

[۶] Machine epsilon
[۷] Unit round
[۸] Roundoff unit
[۹] Machine precision
[۱۰] Underflow
[۱۱] Overflow

تمرین ۳.۲ *در نمایش ۳۲ـبیتی (دقت معمولی[۱۲] متناظر با نوع* float *در زبان* C*) هشت بیت برای مشخصه منظور می‌شود. تمام کمیت‌های معرفی‌شده 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 $\pm a \times ۱۰^b$ 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 $L \le b \le U$ که در آن L و U با توجه به محدودیت‌های سخت‌افزاری و نرم‌افزاری تعیین می‌شوند. $a = ۰/d_۱ \cdots d_k$ بیانگر مانتیس (جزء کسری) است و به ازای $i = ۱, \ldots, k$ داریم $d_i \in \{۰, ۱, \ldots, ۹\}$ و برای یکتایی نمایش فرض می‌شود $d_۱ \neq ۰$. بنابراین $۰/۱ \le a < ۱$. این نمایش، به نمایش ممیز شناور ده‌دهی نرمال‌شده نیز معروف است و چنین اعدادی را، اعداد ماشینی ده‌دهی kـرقمی 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 $x = \pm a \times ۱۰^b$ نمایش داد که در آن $b \in \mathbb{Z}$ و $a = ۰/d_۱ \cdots d_k d_{k+۱} \cdots$ که در آن به ازای $i = ۱, ۲, \ldots$ داریم $d_i \in \{۰, ۱, \cdots, ۹\}$ و $d_۱ \neq ۰$. واضح است که برای نمایش x به صورت ممیز شناور ده‌دهی نرمال‌شده 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)[۱۳]، روش گرد کردن معمولی[۱۴]، روش گرد کردن به زوج[۱۵].

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 $d_{k+۱} \ge ۵$ 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

- $d_{k+۱} > ۵$
- $d_{k+۱} = ۵$ و رقم مخالف صفری در سمت راست $d_{k+۱}$ مشاهده شود
- $d_{k+۱} = ۵$ و رقم مخالف صفری در سمت راست $d_{k+۱}$ مشاهده نشود و d_k فرد باشد.

قضیه ۳.۲ *اگر $x = ۰/d_۱ \cdots d_k d_{k+۱} \cdots \times ۱۰^n$ یک عدد حقیقی ناصفر در نمایش علمی نرمال‌شده باشد آن‌گاه $c(x) = ۰/d_۱ \cdots d_k \times ۱۰^n$ و $r(x) = ۰/\delta_۱ \cdots \delta_k \times ۱۰^m$ ($m = n$ یا $m = n+۱$) اعداد ماشینی kـرقمی متناظر با x هستند که به ترتیب از روش قطع کردن و گرد کردن معمولی به دست می‌آیند و خواهیم داشت*

$$|x - c(x)| \le ۱ \times ۱۰^{n-k}, \qquad |x - r(x)| \le ۵ \times ۱۰^{n-k-۱} = ۰/۵ \times ۱۰^{n-k}.$$

پرسش ۲.۲ *تفاوت ۴۷^{km} با ۴۷۰۰۰^m یا تفاوت ۳/۷ با ۳/۷۰ یا ۳/۷۰۰ در چیست؟*

---

[۱۲] Single precision
[۱۳] Chopping
[۱۴] Rounding
[۱۵] Rounding to even

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 ۳.۲: مثال‌هایی از گرد کردن و قطع کردن معمولی

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 مانتیس در نمایش علمی نرمال‌شده).*

مثال ۶.۲ *عدد ۰/۰۰۰۷۰۴۵۰۰۰ حداقل ۴ رقم بامعنا و حداکثر ۷ رقم بامعنا دارد.* △

قرارداد ۱.۲ *rD یعنی r رقم اعشار و rS یعنی r رقم بامعنا[۱۶].*

مثال ۷.۲ *در جدول ۳.۲، 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 ۵.۲ *از این به بعد، همانند بسیاری از وسایل محاسباتی از گرد کردن معمولی استفاده می‌کنیم.*

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

تعریف ۳.۲ *فرض کنید $\widehat{x} = d_m \times ۱۰^m + \cdots + d_۱ \times ۱۰^۱ + d_۰ \times ۱۰^۰ + d_{-۱} \times ۱۰^{-۱} + \cdots + d_{-k} \times ۱۰^{-k}$ که در آن $d_m \neq ۰$، تقریبی برای $x > ۰$ باشد (در اینجا m نشان‌دهنده با ارزش‌ترین مکان است). تعداد ارقام با معنای درست $\widehat{x}$ نسبت به x، بزرگ‌ترین عدد طبیعی n است که در نابرابری‌های زیر صدق کند.*

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

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

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

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

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

بنابراین $\widehat{x}$ تقریب بهتری برای x است زیرا تعداد ارقام بامعنای درست بیشتری دارد. △

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

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

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

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

۴.۲ انواع خطا

تعریف ۴.۲ *اگر $\widehat{x}$ تقریبی از x باشد $\Delta x = |x - \widehat{x}|$ خطای مطلق $\widehat{x}$ 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 $\Delta x \le b_x$ و در نتیجه $\widehat{x} - b_x \le x \le \widehat{x} + b_x$ و یا می‌نویسیم $x = \widehat{x} \pm b_x$.*

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

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

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 ۴.۲ *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 ۵.۲ *اگر $\widehat{x} \neq ۰$ تقریبی از $x \neq ۰$ باشد $\delta x = \dfrac{|x - \widehat{x}|}{|x|} = \dfrac{\Delta x}{|x|}$ خطای نسبی $\widehat{x}$ 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 $\widehat{x}$ تقریبی از x باشد و b_x یک کران خطای مطلق برای این تقریب باشد آن‌گاه*

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

به علاوه اگر b_x نسبت به $|\widehat{x}|$ خیلی کوچک باشد می‌توان نوشت $\delta x \le \dfrac{b_x}{|\widehat{x}|}$.

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

مثال ۱۱.۲ *برای $\widehat{x} = ۱/۷۳۲$ به عنوان تقریبی از $x = \sqrt{۳}$ می‌توان نوشت*

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

پس $\delta x = ۰/۰۰۰۰۲۹۳۳۳۷\cdots < ۰/۰۰۰۰۳$ و با توجه به قضیه ۴.۲ و $\Delta x = ۰/۰۰۰۱$ خواهیم داشت

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

و یا

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

△

قضیه ۵.۲ *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 بوده و خطای نسبی $\widehat{x}$ و $\widehat{y}$ برابر است.*

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

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

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

مثال ۱۲.۲ *تقریبی از $\sqrt{۳} = ۱/۷۳۲۰۵\cdots$ ارائه دهید که خطای نسبی آن از $۱۰^{-۴}$ کمتر باشد.*
بنابر قضیه ۷.۲ 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 ۶.۲ کافی است $\widehat{x}$ گردشده $\sqrt{۳}$ تا ۵ رقم بامعنا باشد، یعنی $\widehat{x} = ۱/۷۳۲۱$. △

۵.۲ خطای محاسبات

فرض کنید $z = f(x_۱, \ldots, x_n)$ تابعی باشد که می‌خواهیم آن را در نقطه $(x_۱, \ldots, x_n)$ ارزیابی کنیم و $x_i - \widehat{x}_i = \Delta x_i$ 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

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

تذکر ۶.۲ *اگر می‌خواهیم جوابی از یک مسئله را با دقت rD به دست آوریم، باید محاسبات میانی را با دقت $(r+۱)D$ 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 $\frac{۴}{۳}$ 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 $z = V(p,r,h) = pr^۲h$ آن‌گاه باید V به ازای $p=\pi$، $r=\frac{۴}{۳}$ و $h=\sqrt{۲}$ ارزیابی شود. محاسبات میانی را با دقت $۴D$ دنبال می‌کنیم (ماشین حساب را به صورت MODE FIX 4 تنظیم می‌کنیم). بنابراین

$$p=\pi \quad\to\quad \widehat{p}=۳/۱۴۱۶, \qquad r=\frac{۴}{۳} \quad\to\quad \widehat{r}=۱/۳۳۳۳, \qquad h=\sqrt{۲} \quad\to\quad \widehat{h}=۱/۴۱۴۲.$$

به طوری که $\Delta p, \Delta r, \Delta h \le ۰/۵ \times ۱۰^{-۴}$. بنابراین

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

از طرفی

$$\Delta z \simeq \Delta p\widehat{r}^۲\widehat{h} + ۲\Delta r\widehat{p}\widehat{r}\widehat{h} + \Delta h\widehat{p}\widehat{r}^۲ \le (\widehat{r}^۲\widehat{h} + ۲\widehat{p}\widehat{r}\widehat{h} + \widehat{p}\widehat{r}^۲) \times ۰/۵ \times ۱۰^{-۴} \simeq ۹/۹۷۳۱ \times ۱۰^{-۴}.$$

در نتیجه حجم استوانه با دقت $۳D$ برابر است با $۷/۸۹۸$ و حداکثر خطای مرتکب‌شده عبارت است از

$$\Delta z + ۰/۵ \times ۱۰^{-۴} \lesssim ۰/۰۰۱۰.$$

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

$$e(z) = |z - \widehat{z}| = ۴/۵۸۵۵۵۴ \times ۱۰^{-۴} < ۰/۰۰۱۰.$$

$\triangle$

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 $y=\frac{\pi}{۱۲}$ 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 $۰/۵ \times ۱۰^{-۲} \leq |۴۸/۹\Delta x + ۴۷/۹\Delta y|$. یک جواب این نامعادله 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، می‌توان 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 شرکت‌پذیری، یکتایی 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 ۶.۲ *فرض کنید x و y دو عدد حقیقی، $\widehat{x}$ و $\widehat{y}$ تقریب‌هایی از آن‌ها و $\otimes$ بیانگر یک عمل دوتایی باشد. متناظر با $x \otimes y$ در ماشین عمل $\widehat{x}\widehat{\otimes}\widehat{x}$ انجام می‌شود و می‌توان نوشت*

$$\left|x \otimes y - \widehat{x}\widehat{\otimes}\widehat{y}\right| = \left|(x \otimes y - \widehat{x} \otimes \widehat{y}) + (\widehat{x} \otimes \widehat{y} - \widehat{x}\widehat{\otimes}\widehat{y})\right| \leq |x \otimes y - \widehat{x} \otimes \widehat{y}| + \left|\widehat{x} \otimes \widehat{y} - \widehat{x}\widehat{\otimes}\widehat{y}\right|.$$

بنابراین خطا دو بخش دارد: خطای منتشرشده یا انباشته‌شده (جمله اول) و خطای تولیدشده یا گرد کردن (جمله دوم).

قضیه ۸.۲ *اگر $\widehat{x}$ و $\widehat{y}$ تقریب‌هایی از x و y بوده و همه این اعداد مثبت باشند، آن‌گاه*

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

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

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

برهان. به عنوان تمرین. □

مثال ۱۵.۲ *$\widehat{x}$ و $\widehat{y}$ هر یک n رقم بامعنای درست دارند. حداقل تعداد ارقام بامعنای درست $\widehat{y} + \widehat{x}$ و $\widehat{x}\widehat{y}$ را تعیین کنید.*
بنابر قضیه ۷.۲، $\delta x < ۵ \times ۱۰^{-n}$ و $\delta y < ۵ \times ۱۰^{-n}$. هم‌چنین بنابر قضیه ۸.۲ می‌توان نوشت

$$\delta(x + y) \leq \max\{\delta x, \delta y\} \leq ۵ \times ۱۰^{-n} = ۰/۵ \times ۱۰^{-(n-۱)}.$$

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

$$\delta(xy) \leq \delta x + \delta y < ۵ \times ۱۰^{-n} + ۵ \times ۱۰^{-n} = ۱۰ \times ۱۰^{-n} = ۱۰^{-(n-۲)-۱} < ۰/۵ \times ۱۰^{-(n-۲)}.$$

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

تذکر ۷.۲ *با توجه به خطای نسبی $\delta(x-y) \simeq \frac{\Delta(x-y)}{|x-y|}$ 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 بامعنای 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 $۱/۴۲ - ۱/۴۱ = ۰/۰۱$). 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*

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

△

مثال ۱۷.۲ *در صورتی که $g(x) = x\left(\sqrt[۳]{۱+\frac{۱}{x}} - ۱\right)$، مقدار $g(۱۰^۹)$ را با دقت $۸D$ به دست آورید.*
همان‌گونه که بیان شد، چون دقت $۸D$ مورد نظر است، محاسبات میانی را با دقت $۹D$ انجام می‌دهیم

$$۱ + \frac{۱}{x} = ۱/۰۰۰۰۰۰۰۰۱, \quad \sqrt[۳]{۱+\frac{۱}{x}} = ۱/۰۰۰۰۰۰۰۰۰.$$

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

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

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 به از بین رفتن ارقام بامعنا منجر می‌شود. برای به دست آوردن تقریب بهتر، روش محاسبه را به صورت زیر تغییر می‌دهیم

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

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

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

و بنابراین با دقت $8D$ داریم $g(۱۰^۹) = ۰/۳۳۳۳۳۳۳۳۳$. △

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 قائل شویم.

تذکر ۸.۲ *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 $x = ۱۰۰۰۰ \times \pi$ می‌توان نوشت*

$$\pi \simeq ۳/۱۴ \quad \to \quad x \simeq ۳۱۴۰۰, \quad \pi \simeq ۳/۱۴۲ \quad \to \quad x \simeq ۳۱۴۲۰, \quad \pi \simeq ۳/۱۴۱۶ \quad \to \quad x \simeq ۳۱۴۱۶.$$

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 ۹.۲ *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 $ax^۳ + bx^۲ + cx + d$ 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 چندجمله‌ای تیلور نظیر آن استفاده کرد. مثال‌هایی که در ادامه خواهند آمد چگونگی این تقریب را نشان می‌دهند.

مثال ۲۰.۲ *مطلوب است محاسبه مقدار $e^{\frac{\pi}{۱۰}}$ با دقت $2D$.*
روش اول- به کمک قضیه تیلور داریم

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

که در آن $۰ < \xi(x) < x$. در نتیجه

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

که در آن $۰ < \xi < \frac{\pi}{۱۰} < ۱$. چون e^x تابعی صعودی است پس $۱ = e^۰ < e^\xi < e^۱ < ۳$. از طرفی $\frac{\pi}{۱۰} < ۰/۳۱۵$ (دقت $3D$ 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*

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

△

مثال ۲۱.۲ *(همگرایی سریع) می‌خواهیم تابع* $\cos x$ *را به ازای مقادیر* $|x| < \frac{\pi}{۴}$ *با دقت* $۵D$ *ارزیابی کنیم. با توجه به سری تیلور*

$$\cos x = \sum_{i=۰}^{\infty} (-۱)^i \frac{x^{۲i}}{(۲i)!} = ۱ + \frac{x^۲}{۲!} - \frac{x^۴}{۴!} + \cdots,$$

و این که $\frac{۱/۶^{۲n}}{(۲n)!} < \frac{(\frac{\pi}{۴})^{۲n}}{(۲n)!} < \left|\frac{x^{۲n}}{(۲n)!}\right|$ *باید داشته باشیم* $۰/۵ \times ۱۰^{-۵} > \frac{۱/۶^{۲n}}{(۲n)!}$ *که نتیجه می‌دهد* $n \geq ۶$. △

مثال ۲۲.۲ *(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 ۲۳.۲ *تابع* $Si(x) = \int_۰^x \frac{\sin t}{t} dt,\ x \geq ۰$ *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*

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

و یا

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

در نتیجه

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

بنابراین

$$Si(۱) = ۱ - \frac{۱}{۳\times۳!} + \frac{۱}{۵\times۵!} + \cdots + (-۱)^n\frac{۱}{(۲n+۱)\times(۲n+۱)!} + E,$$

که در آن $E = \frac{(-۱)^{n+۱}}{(۲n+۲)!}\int_۰^۱ t^{۲n+۱}\sin(\xi(t))dt$. حال می‌توان نوشت

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

پس

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

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

$$Si(۱) \simeq ۱ - \frac{۱}{۳\times۳!} + \frac{۱}{۵\times۵!} - \frac{۱}{۷\times۷!} + \frac{۱}{۹\times۹!},$$

و یا

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

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

۶.۲ تمرین‌ها

۱. حاصل $(۰/\overline{۰z})_\beta$ که در آن $z = \beta - ۱$ را به دست آورید.

۲. فاصله بین اعداد در دستگاه ممیز ثابت $\pm(a_۴a_۳a_۲a_۱a_۰/b_۱b_۲b_۳b_۴)_۲$ را به دست آورید.

۳. آیا $\frac{۱}{\beta+۱}$ می‌تواند مانتیس یک عدد نرمال‌شده در مبنای β باشد؟ چرا؟

۴. بزرگ‌ترین عدد p را به دست آورید که نمایش آن در دستگاه نمایش اعداد نرمال‌شده در مبنای β با m رقم مانتیس و $l \le e \le u$ 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 $\pm(0/b_1 \cdots b_m)_\beta \times \beta^e$ که در آن $l \le e \le u$ را محاسبه کنید.

۶. اگر نمایش دو عدد x و $x+y$ در دستگاه نمایش اعداد نرمال در مبنای β با m رقم مانتیس و $l \le e \le u$ 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 مانتیس 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

$$(0/b_1 b_2 \cdots b_m \overline{c_1 c_2 \cdots c_k})_\beta = \frac{(b_1 b_2 \cdots b_m c_1 c_2 \cdots c_k)_\beta - (b_1 b_2 \cdots b_m)_\beta}{(\underbrace{zz\cdots z}_{k}\underbrace{0\cdots 0}_{m})_\beta}, \qquad z = \beta - 1$$

۹. مقدار محاسبه‌شده برای

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

و $p \ge 1$ در یک ماشین با روند عدد یک برابر 2^{1-t} به ازای چه مقادیری از $|h|$ برابر ۰ است؟

۱۰. درستی عبارات $\frac{|c(x)-x|}{|x|} < \beta^{1-k}$ و $\frac{|r(x)-x|}{|x|} < \frac{1}{2}\beta^{1-k}$ را بررسی کنید.

۱۱. برای محاسبه مقدار $1 - \cos(x)$ در نزدیکی ۰ از چه عبارتی استفاده می‌کنید؟ توضیح دهید.

۱) ۰ ۲) $\frac{x}{2} - \frac{x^2}{2}$ ۳) $x^3 - \frac{x^2}{2}$ ۴) $\frac{\sin^2(x)}{1+\cos(x)}$

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

۱۳. برای محاسبه تقریبی $\frac{1}{(\sqrt{2}+1)^6}$ کدام عبارت تقریب دقیق‌تری به دست می‌دهند؟

۱) $99 - 70\sqrt{2}$ ۲) $\frac{1}{99+70\sqrt{2}}$ ۳) $(\sqrt{2}-1)^6$

۱۴. برای تقریب تابع $\cos(x)$ به ازای $|x| \le 0/1$ با استفاده از بسط تیلور حول نقطه ۰ به منظور دست‌یابی به خطای برشی حدود 10^{-5} به چه تعداد از جملات سری تیلور نیاز داریم؟

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

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