به نام خداوند بخشنده‌ی مهربان

آنالیز عددی

حسین‌زارع

۱

خطاها و دستگاه‌های نمایش اعداد

۱-۱ منابع خطا

در طی روند حل عددی یک مسئله، عوامل گوناگونی دست به دست هم می‌دهند تا جواب حاصل با جواب دقیق برابر نباشد. به‌طورکلی خطاهای موجود در جواب تقریبی یک مسئله از چهار منبع ناشی می‌شوند[۱]. این منابع عبارتند از:

- **مدلسازی مسئله**

اولین مرحله از حل هر مسئله‌ی واقعی، مدلسازی آن مسئله است. از آن‌جا که دخالت دادن تمام عوامل مؤثر در مسئله، احتمالاً به یک مدل پیچیده و گاه غیر مفید منجر می‌شود، برخی از عوامل کم‌اهمیت‌تر نادیده گرفته می‌شوند. برای مثال در مدلسازی مسئله‌ی «تعیین زمان تناوب حرکت یک آونگ ساده» در مکانیک، از جرم نخ متصل به آونگ، اصطکاک بین نخ و نقطه‌ی آویز و نیز مقاومت هوا صرف‌نظر می‌شود. بدیهی است که این عوامل، تأثیراتی هر چند کوچک در جواب مسئله خواهند گذاشت. بنابراین با در نظر نگرفتن آن‌ها خطایی در جواب مسئله به‌وجود می‌آید. این خطا را خطای مدل می‌نامیم.

- **داده‌های ورودی**

پارامترهای معلوم یک مدل اغلب کمیت‌هایی هستند که از طریق اندازه‌گیری به‌دست می‌آیند و چون وسایل اندازه‌گیری صد در صد دقیق نیستند، اندازه‌گیری‌ها با خطا همراه هستند. البته هرچه وسایل اندازه‌گیری دقیق‌تر باشند، این خطا کمتر می‌شود.

---

[۱] برخی کتاب‌ها منابع خطا را به سه دسته و برخی دیگر به پنج دسته تقسیم می‌کنند. در این دسته‌بندی که منسوب به جان فون نویمن* و هرمان گلدشتین* است [کتاب هوسهولدر]، منابع خطا به چهار دسته تقسیم شده است.

- **روند کردن اعداد**

انجام محاسبات تقریبی بر روی اعدادی که دارای نامتناهی رقم هستند، جز با انتخاب تعدادی متناهی از ارقام آن‌ها ممکن نیست. علاوه بر این، همواره در وسایل محاسباتی به‌دلیل محدودیت حافظه، تعداد محدودی از ارقام یک عدد قابل ذخیره‌سازی و نمایش است. از آن‌جا که بسیاری از اعداد حقیقی مانند $۱/۳$ ، $\sqrt{۲}$ ، e ، π و ... دارای نامتناهی رقم در بسط اعشاری یا دودویی خود هستند، می‌توان گفت که بیشتر اعداد به‌صورت تقریبی و در نتیجه توأم با خطا در نظر گرفته می‌شوند. به خطای ناشی شده در این موارد خطای نمایش اعداد می‌گوییم.

از طرف دیگر، گاهی در حین محاسبات به دلیل انجام یک عمل حسابی بر دو عدد حقیقی، حاصل عمل دارای نامتناهی رقم است که با توجه به آن‌چه که در بند قبل گفته شد، تنها تعدادی متناهی از ارقام آن قابل انتخاب است. خطای حاصل در این حالت را خطای تولید شده توسط عمل حسابی می‌نامیم. در یک ماشین محاسباتی، حتی نیازی نیست که حاصل یک عمل دارای نامتناهی رقم باشد، زیرا حتی با وجود متناهی بودن تعداد ارقام حاصل از یک عمل، اگر این تعداد رقم بیش از حافظه‌ی تخصیص داده شده برای نگهداری هر عدد باشد، نتیجه‌ی عمل با خطا در ماشین نمایش داده می‌شود.

در حالت کلی به خطای نمایش اعداد و خطای تولید شده توسط اعمال حسابی، خطاهای روند کردن می‌گوییم. خطاهای روند کردن آثار خود را در حین فرآیند محاسبات برجای می‌گذارند. زیرا اگر مقادیری که در یک محاسبه به کار می‌روند خود با خطایی همراه باشند، انجام عملیات حسابی بر آن‌ها موجب ایجاد خطاهای جدید می‌شود که خطای حاصل در این حالت را خطای منتشر شده توسط عمل حسابی می‌نامیم. انتشار خطا توسط توابع نیز صورت می‌گیرد که در ادامه‌ی فصل به‌طور کامل در مورد آن بحث خواهد شد.

- **روش عددی**

چنانچه حل تحلیلی یک مسئله‌ی ریاضی امکان‌پذیر یا از لحاظ محاسباتی مقرون به‌صرفه نباشد، از یک روش عددی برای تعیین جواب تقریبی مسئله استفاده می‌شود. برای این منظور ممکن است روش‌های عددی گوناگونی وجود داشته باشند. دقت تقریب حاصل از هر روش عددی به ویژگی‌های آن روش (از جمله پایداری و سرعت همگرایی) و مرحله‌ی توقف آن بستگی دارد. معمولاً در مباحث عددی، خطای روش نتیجه‌ی تقریب یک مسئله‌ی اولیه با یک مسئله‌ی ساده‌تر توسط گسسته‌سازی یا مختوم نمودن یک فرآیند نامتناهی است.

✎ **تذکر:** خطاها اغلب جزء اجتناب‌ناپذیری از حل عددی یک مسئله هستند و نباید آن‌ها را با اشتباهاتی که ممکن است در روند حل هر مسئله‌ای رخ دهند، یکی گرفت. از این رو همواره میان دو مفهوم خطا ($error$) و اشتباه ($mistake$) تمایز قائل می‌شویم و اشتباهات را جزء منابع خطا در نظر نمی‌گیریم.

خطاهای مدلسازی و داده‌های ورودی در آنالیز عددی مورد بررسی قرار نمی‌گیرند و افرادی که در علوم مختلف به تعیین مدل مسئله و اندازه‌گیری داده‌ها می‌پردازند، مسئول آن‌ها هستند. اما خطاهای ناشی از روند کردن اعداد و روش عددی به آنالیز عددی مربوط می‌شوند.

۱-۲ انواع خطا

بخش عمده‌ای از علم آنالیز عددی به یافتن تقریب مناسبی از جواب یک مسئله مربوط می‌شود. در این‌جا عبارت *«تقریب مناسبی»* باید به نحوی مطلوب تعریف شود. نخست تعریفی از تقریب یک عدد ارائه می‌دهیم.

تعریف ۱-۱:

فرض کنیم A و a دو عدد حقیقی باشند. a را یک تقریب A می‌گوییم هرگاه داشته باشیم $A - a \neq ۰$.

اگر $a < A$ آنگاه a را تقریب نقصانی A و چنانچه $a > A$ در آن‌صورت a را تقریب اضافی A می‌نامیم.

✍ **مثال ۱-۱:** در مورد عدد π داریم:

$$۳/۱۴ < \pi < ۳/۱۵$$

بنابراین عدد ۳/۱۴ یک تقریب نقصانی و عدد ۳/۱۵ یک تقریب اضافی عدد π می‌باشند.

از تعریف فوق معلوم می‌شود که تقریب یک عدد حقیقی منحصر به فرد نیست. بنابراین مناسب بودن یک تقریب به چه معناست؟ بدیهی است که اگر ناچار باشیم به یافتن جواب تقریبی یک مسئله رضایت دهیم، خواهان تقریبی هستیم که تا حدود زیادی با جواب دقیق مسئله مطابقت داشته باشد. بنابراین بایستی معیار یا معیارهایی برای مقایسه‌ی دقت تقریب‌های یک عدد داشته باشیم. انواع خطا در حقیقت همان معیارهای سنجش دقت یک تقریب هستند که در ادامه معرفی می‌شوند.

تعریف ۱-۲: اگر a تقریبی از عدد حقیقی A باشد، آنگاه خطای مطلق a را با $e(a)$ نشان داده و به‌صورت زیر تعریف می‌کنیم.

$$\boxed{e(a) = |A - a|.}$$

برای مثال اگر $A = ۱۰/۱۲$ ، $a = ۱۰/۰۵$ و $a' = ۱۰/۱۵$ ، آنگاه خطاهای مطلق a و a' به‌عنوان تقریب‌هایی از A برابرند با:

$$e(a) = \left|۱۰/۱۲ - ۱۰/۰۵\right| = ۰/۰۷,$$
$$e(a') = \left|۱۰/۱۲ - ۱۰/۱۵\right| = ۰/۰۳.$$

واضح است که هر چه $e(a)$ کوچک‌تر باشد، a تقریب دقیق‌تری برای A خواهد بود.

با اینکه خطای مطلق، معیار ساده‌ای برای تعیین دقت یک تقریب است، اما میزان دقت دو یا چند تقریب مختلف را نمی‌توان به کمک آن مقایسه کرد؛ زیرا این خطا فقط اختلاف بین مقدار واقعی و مقدار تقریبی را مشخص می‌کند و به اینکه این مقادیر چقدر بزرگ یا چقدر کوچک باشد، توجهی ندارد. برای مثال فرض کنید که $A = ۱۰۰۰$ و $B = ۱۰^{۱۰}$. اگر $a = ۱۵۰۰$ و $b = ۱۰^{۱۰} + ۵۰۰$ را به‌ترتیب به‌عنوان تقریب‌هایی از A و B در نظر بگیریم، در این‌صورت خطای مطلق هر دو تقریب برابر با ۵۰۰ است. اما آیا دقت a و b نیز یکسان است؟ واضح است که نه و دقت b به مراتب بیش‌تر از دقت a است. به‌عبارت دیگر عدد ۵۰۰ در مقایسه با عدد $۱۰^{۱۰}$ مقداری ناچیز می‌نماید

اما در مقایسه با عدد ۱۰۰۰، رقم قابل ملاحظه‌ای است. با این توضیحات به معیار مناسب‌تری برای تعیین دقت یک تقریب نیاز داریم. این معیار که نشان‌دهنده‌ی خطا در واحد کمیت است، خطای نسبی نامیده می‌شود.

تعریف ۱-۳: اگر a تقریبی از عدد حقیقی $A \neq ۰$ باشد، آنگاه خطای نسبی a را با $\delta(a)$ نشان داده و به‌صورت زیر تعریف می‌کنیم:

$$\boxed{\delta(a) = \frac{|A-a|}{|A|} = \frac{e(a)}{|A|}.}$$

برای مثال خطاهای نسبی $a = ۱۵۰۰$ به‌عنوان تقریبی از $A = ۱۰۰۰$ و $b = ۱۰^{۱۰} + ۵۰۰$ به‌عنوان تقریبی از $B = ۱۰^{۱۰}$ برابرند با:

$$\delta(a) = \frac{۵۰۰}{۱۰۰۰} = \frac{۱}{۲},$$
$$\delta(b) = \frac{۵۰۰}{۱۰^{۱۰}} = \frac{۱}{۲} \times ۱۰^{-۷}.$$

واضح است که هر چه $\delta(a)$ کوچک‌تر باشد، a تقریب دقیق‌تری برای A خواهد بود.

اگرچه کوچک بودن خطای نسبی یک تقریب نشان‌دهنده‌ی دقت بالای آن تقریب است، اما از بزرگ بودن اندازه‌ی آن نمی‌توان نادقیق بودن یک تقریب را نتیجه گرفت. به‌عبارت دیگر، گاهی ممکن است که یک تقریب a از عدد ناصفر A از دقت بسیار خوبی برخوردار باشد اما خطای نسبی آن بسیار زیاد باشد و آن هنگامی است که $|A|$ خود کوچک باشد. برای مثال فرض کنید $A = ۱۰^{-۱۱}$ و $a = ۱۰^{-۱۰}$. عدد a تا حد زیادی به A نزدیک و تقریب مناسبی از آن است. اکنون خطاهای مطلق و نسبی a را به‌عنوان تقریبی از A محاسبه می‌کنیم:

$$e(a) = |A-a| = |۱۰^{-۱۱} - ۱۰^{-۱۰}| = ۹ \times ۱۰^{-۱۱},$$
$$\delta(a) = \frac{e(a)}{|A|} = \frac{۹ \times ۱۰^{-۱۱}}{۱۰^{-۱۱}} = ۹.$$

در این‌جا خطای نسبی a در مقایسه با مقادیر A و a عدد بسیار بزرگی است و به نظر می‌رسد که خطای مطلق، دقت a را بهتر مشخص می‌کند. اکنون معیاری را معرفی می‌کنیم که می‌تواند هم برای مقادیر خیلی بزرگ و هم برای مقادیر خیلی کوچک A مورد استفاده قرار گیرد.

تعریف ۱-۴: اگر a تقریبی از عدد حقیقی A باشد، آنگاه خطای ترکیبی[۱] a را با $\mu(a)$ نشان داده و به‌صورت زیر تعریف می‌کنیم:

$$\boxed{\mu(a) = \frac{|A-a|}{۱+|A|} = \frac{e(a)}{۱+|A|}.}$$

با توجه به تعریف فوق می‌بینیم که اگر $|A| \ll ۱$، یعنی $|A|$ خیلی کوچک‌تر از یک باشد، آنگاه $\mu(a) \approx e(a)$ (یعنی خطای ترکیبی تقریباً برابر با خطای مطلق است) و چنانچه $|A| \gg ۱$، یعنی $|A|$ خیلی بزرگ‌تر از یک باشد،

[۱] به این خطا، خطای نسبی- مطلق نیز می‌گویند.

آن‌گاه $\mu(a) \approx \delta(a)$. هم‌چنین واضح است که $0 < \mu(a) \leq \min\{e(a), \delta(a)\}$.

تاکنون سه معیار برای سنجش دقت یک تقریب معرفی کرده‌ایم. یک معیار دیگر تعداد ارقام بامعنای درست است که چون از مفهوم گرد کردن اعداد ناشی می‌شود، در بخش‌های بعد معرفی خواهد شد. در معیارهایی که تاکنون بیان داشته‌ایم فرض بر این بوده است که مقادیر دقیق و تقریبی (به ترتیب A و a) برای ما معلوم هستند. اما در بسیاری از موارد A که جواب دقیق مسئله است، مجهول می‌باشد و تنها تقریب‌هایی از آن را در دست داریم (مانند ریشه‌ی یک معادله). از این‌رو عملاً تعیین هیچ‌یک از خطاهای گفته شده برای سنجش دقت این تقریب‌ها امکان‌پذیر نیست. به همین دلیل در روش‌های عددی اغلب کران بالایی برای خطای مطلق یا نسبی یک تقریب به‌دست می‌آوریم. بدیهی است که اگر کران بالای خطایی بسیار بزرگتر از واقعیت موجود باشد، نمی‌توان آن را برای برآورد معقولی از خطا مورد استفاده قرار داد و باید سعی کنیم تا کران بالای کوچک‌تری برای خطا به‌دست آوریم.

تعریف ۱-۵: اگر a تقریبی از عدد حقیقی A باشد، آن‌گاه هر عدد ناکمتر از $e(a)$ را یک خطای مطلق حدی a می‌نامیم و آن را با e_a نشان می‌دهیم. بنابراین $e(a) \leq e_a$ و e_a منحصر به‌فرد نیست.

توجه داریم که اگر e_a یک خطای مطلق حدی a به‌عنوان تقریبی از A باشد، آن‌گاه:

$$e(a) \leq e_a \Rightarrow |A - a| \leq e_a \Rightarrow a - e_a \leq A \leq a + e_a$$

یعنی A در بازه‌ی $[a - e_a, a + e_a]$ قرار دارد. طول این بازه $2e_a$ است؛ بنابراین هرچه e_a کوچک‌تر باشد، حدود دقیق‌تری برای A حاصل می‌شود.

✎ **تذکر:** بنابر قرارداد نامساوی $a - e_a \leq A \leq a + e_a$ را به اختصار به‌صورت زیر می‌نویسیم:

$$A = a \pm e_a.$$

✍ **مثال ۱-۲:** با فرض آن‌که $۳/۱۴۱ < \pi < ۳/۱۴۲$ ، یک کران بالا برای خطای مطلق $۳/۱۴$ به‌عنوان تقریبی از π به‌دست آورید.

حل: داریم

$$۳/۱۴۱ < \pi < ۳/۱۴۲ \Rightarrow ۳/۱۴۱ - ۳/۱۴ < \pi - ۳/۱۴ < ۳/۱۴۲ - ۳/۱۴$$
$$\Rightarrow ۰/۰۰۱ < \pi - ۳/۱۴ < ۰/۰۰۲ \Rightarrow \left|\pi - ۳/۱۴\right| < ۰/۰۰۲$$

بنابراین $e(۳/۱۴) < ۰/۰۰۲$.

معمولاً در اکثر روش‌های آنالیز عددی حدود جواب، یعنی کران بالا و پایینی برای جواب، قابل محاسبه است و از آن‌جا می‌توان مانند مثال فوق کران بالایی برای $e(a)$ به‌دست آورد.

تعریف ۱-۶: اگر a تقریبی از عدد حقیقی A باشد، آن‌گاه هر عدد ناکمتر از $\delta(a)$ را یک خطای نسبی حدی a می‌نامیم و آن را با δ_a نشان می‌دهیم. بنابراین $\delta(a) \leq \delta_a$ و δ_a منحصر به‌فرد نیست.

قضیه‌ی زیر نشان می‌دهد که چگونه می‌توان با دانستن خطای مطلق حدی یک تقریب، کران بالایی برای خطای نسبی آن به‌دست آورد.

قضیه ۱-۱: اگر a تقریبی از A و e_a یک خطای مطلق حدی a باشد، آنگاه:

$$\delta(a) \leq \frac{e_a}{|a| - e_a}.$$

اثبات: از خواص قدرمطلق می‌دانیم که $|a| - |A| \leq |A - a|$ و با توجه به فرض قضیه داریم:

$$|A - a| \leq e_a$$

بنابراین، $|a| - |A| \leq e_a$ و در نتیجه:

$$\begin{aligned} |a| - e_a \leq |A| &\Rightarrow \frac{۱}{|A|} \leq \frac{۱}{|a| - e_a} \\ &\Rightarrow \frac{e(a)}{|A|} \leq \frac{e(a)}{|a| - e_a} \left(\leq \frac{e_a}{|a| - e_a} \right) \\ &\Rightarrow \delta(a) \leq \frac{e_a}{|a| - e_a}. \end{aligned}$$

✍ **مثال ۱-۳:** با فرض آنکه $۳٫۱ < \sqrt{۱۰} < ۳٫۱۷$ ، یک کران بالا برای خطای نسبی ۳٫۱۵ به‌عنوان تقریبی از $\sqrt{۱۰}$ به‌دست آورید.

حل: داریم

$$\begin{aligned} &۳٫۱ < \sqrt{۱۰} < ۳٫۱۷ \Rightarrow ۳٫۱ - ۳٫۱۵ < \sqrt{۱۰} - ۳٫۱۵ < ۳٫۱۷ - ۳٫۱۵ \\ &\Rightarrow -۰٫۰۵ < \sqrt{۱۰} - ۳٫۱۵ < ۰٫۰۲ \Rightarrow \left|\sqrt{۱۰} - ۳٫۱۵\right| < ۰٫۰۵ \end{aligned}$$

بنابراین $e(۳٫۱۵) < ۰٫۰۵$. اکنون با توجه به قضیه‌ی قبل داریم:

$$\delta(۳٫۱۵) \leq \frac{۰٫۰۵}{۳٫۱۵ - ۰٫۰۵} = \frac{۰٫۰۵}{۳٫۱} = \frac{۱}{۶۲}.$$

✎ **تذکر:** اگر e_a در مقایسه با $|a|$ کوچک باشد، آنگاه $|a| - e_a \approx |a|$ و بنابراین می‌توان نوشت:

$$\delta(a) \lesssim \frac{e_a}{|a|}$$

(نماد $\lesssim$ به معنی «تقریباً کوچک‌تر از» می‌باشد.)

معمولاً اگر e_a در مقایسه با $|a|$ ناچیز باشد $\delta(a)$ تقریباً مساوی $\frac{e_a}{|a|}$ گرفته می‌شود. یعنی:

$$\delta(a) \approx \frac{e_a}{|a|}$$

برای نمونه، در مثال قبل می‌توانیم بنویسیم:

$$\delta(۳٫۱۵) \approx \frac{۰٫۰۵}{۳٫۱۵} = \frac{۵}{۳۱۵} = \frac{۱}{۶۳}.$$

۱-۳ دستگاه‌های نمایش اعداد

استفاده از دستگاه اعشاری(دهدهی)، متداول‌ترین شیوه‌ی نمایش اعداد در زندگی روزمره‌ی ماست. هنگامی که مقدار دقیق یک کمیت یا تقریبی از آن را می‌نویسیم از ده رقم ۰ ، ۱، ۲، ۳، ۴، ۵، ۶، ۷، ۸ و ۹ به همراه یک ممیز استفاده می‌کنیم. این در حالی است که در کامپیوترها معمولاً از دستگاه دودویی $(binary)$ با تنها دو رقم ۰ و ۱ برای نمایش اعداد استفاده می‌شود. از این رو شیوه‌ی نمایش اعداد در دستگاه‌های مختلف و چگونگی تبدیل آن‌ها از دستگاهی به دستگاه دیگر را مورد مطالعه قرار می‌دهیم.

تعریف ۱-۷: فرض کنید $\beta > ۱$ عددی طبیعی و A یک عدد حقیقی مثبت باشد. در این صورت منظور از نمایش A در مبنای β نمایش آن به صورت

$$A = a_m \times \beta^m + a_{m-۱} \times \beta^{m-۱} + \cdots$$

می‌باشد که در آن

$$a_m \neq ۰;\ m, a_i \in \mathbb{Z};\ ۰ \leq a_i \leq \beta - ۱;\ i = m, m-۱, \ldots$$

در بسط فوق چنانچه $\beta = ۱۰$ بسط اعشاری (دهدهی) A و چنانچه $\beta = ۲$ بسط دودویی A حاصل می‌شود. معمولاً نمایش فوق را به صورت $A = (a_m a_{m-۱} a_{m-۲} \ldots)_\beta$ می‌نویسند.

قضیه ۱-۲:(*وجود و یکتایی نمایش اعداد طبیعی در مبنای* $\beta > ۱$) اگر A عددی صحیح و مثبت باشد، آن‌گاه دارای نمایش منحصر به فرد زیر در مبنای $\beta > ۱$ است.

$$A = a_m \beta^m + a_{m-۱} \beta^{m-۱} + \cdots + a_۱ \beta + a_۰ \quad (*)$$

***اثبات:** ابتدا ثابت می‌کنیم A دارای چنین نمایشی است و سپس نشان می‌دهیم این نمایش منحصر به فرد است. اگر $A < \beta$ ، آن‌گاه $(*)$ به ازای $a_۰ = A$ و $m = ۰$ برقرار است. یعنی هر عدد A کوچکتر از β نمایش خودش در مبنای β است. برای $A \geq \beta$ فرض می‌کنیم تمام اعداد صحیح مثبت کوچکتر از A دارای نمایشی در مبنای β باشند (فرض استقرای قوی). بنابر قضیه‌ی تقسیم داریم $A = \beta q + r$ که در آن $۰ < q < A$ و $۰ \leq r \leq \beta - ۱$. طبق فرض استقرا، q نمایشی در مبنای β دارد. با جایگذاری نمایش q در رابطه‌ی بالا، نمایشی در مبنای β برای A حاصل می‌شود.

اکنون نشان می‌دهیم که نمایش A در مبنای β منحصر به فرد است. اگر $A < \beta$ آن‌گاه $A = a_۰$ و واضح است که A به جز این، نمایش دیگری ندارد. حال فرض کنید $A \geq \beta$ و نیز هر عدد کوچکتر از A دارای نمایش یکتایی در مبنای β باشد (فرض استقرا). هرگاه $(*)$ برقرار باشد داریم:

$$A = (a_m \beta^{m-۱} + a_{m-۱} \beta^{m-۲} + \cdots + a_۱) \times \beta + a_۰$$

و چون $۰ \leq r \leq \beta - ۱$ پس $a_۰$ باقیمانده و

$$(a_m \beta^{m-۱} + a_{m-۱} \beta^{m-۲} + \cdots + a_۱)$$

خارج قسمت تقسیم A بر β هستند. علاوه بر این، خارج قسمت بزرگتر از صفر است زیرا $A \geq \beta$. حال اگر A

دارای نمایش دیگری مانند $A = c_t\beta^t + a_{t-1}\beta^{t-1} + \cdots + c_1\beta + c_0$ در مبنای β باشد، آن‌گاه همانند آنچه که در مورد $(*)$ بحث شد، c_0 باقیمانده و

$$(c_t\beta^{t-1} + c_{t-1}\beta^{t-2} + \cdots + c_1)$$

خارج قسمت تقسیم A بر β هستند. اما در قضیه‌ی تقسیم، باقیمانده و خارج قسمت یکتا هستند. پس $a_0 = c_0$ و $a_m\beta^{m-1} + \cdots + a_1 = c_t\beta^{t-1} + \cdots + c_1$. چون طرفین تساوی اخیر، هر دو نمایش‌های یک عدد طبیعی کوچکتر از A در مبنای β هستند، بنابر فرض استقرا، این دو نمایش یکسان هستند. چون $a_0 = c_0$ پس $m = t$. بنابراین برای هر $0 \leq i \leq m$ داریم $a_i = c_i$. در نتیجه نمایش A در مبنای β منحصر به فرد است.

با توجه به مطالب فوق می‌توان روند تکراری زیر را برای نمایش عدد طبیعی A در مبنای β ارائه نمود:

۱ـ قرار می‌دهیم $i = 0$ و $q = [\frac{A}{\beta}]$. (q خارج‌قسمت تقسیم A بر β است.)

۲ـ تا زمانی که $q \neq 0$ ، اعمال زیر را به ترتیب انجام می‌دهیم:

$$\begin{aligned} a_i &= A - \beta q \\ A &\leftarrow q \\ q &\leftarrow [A/\beta] \\ i &\leftarrow i + 1 \end{aligned}$$

بدین ترتیب برای A نمایشی مانند $A = (a_m a_{m-1} \ldots a_1 a_0)_\beta$ به دست می‌آید. روش فوق برای نمایش عدد طبیعی A در مبنای β، روش تقسیم‌های متوالی نام دارد.

✍ **مثال ۱-۴:** نمایش دودویی عدد ۲۲ را به دست آورید.

حل:

i	A	$q = [\frac{A}{2}]$	$A - 2q$
۰	۲۲	۱۱	$0 = a_0$
۱	۱۱	۵	$1 = a_1$
۲	۵	۲	$1 = a_2$
۳	۲	۱	$0 = a_3$
۴	۱	۰	$1 = a_4$

باقیمانده خارج قسمت

$\longrightarrow$ به خارج قسمت ۰ رسیدیم.

بنابراین: $(10110)_2 = 22$.

قضیه ۱-۳: هر عدد $0 < r < 1$ دارای نمایش منحصر به فرد زیر در مبنای $\beta > 1$ است،

$$r = r_1\beta^{-1} + r_2\beta^{-2} + \cdots + r_m\beta^{-m} + \cdots \quad (**),$$

که در آن:

$$0 \leq r_i \leq \beta - 1;\ i = 1,2,\ldots;\ r_i \neq \beta - 1$$ بی‌نهایت بار.

***اثبات:** برای آن‌که ثابت کنیم r دارای چنین نمایشی است کافی است r_iها را معرفی کنیم. برای معرفی r_1 ابتدا طرفین (**) را در β ضرب می‌کنیم. خواهیم داشت:

$$\beta r = r_1 + r_2\beta^{-1} + r_3\beta^{-2} + \cdots + r_m\beta^{-(m-1)} + \cdots$$

فرض کنیم $h = r_2\beta^{-1} + r_3\beta^{-2} + \cdots + r_m\beta^{-(m-1)} + \cdots$.

از آن‌جا که $0 \le r_i \le \beta - 1$ و بی‌نهایت بار $r_i \neq \beta - 1$ ، داریم:

$$0 \le h < (\beta-1)(\beta^{-1} + \beta^{-2} + \cdots) = (\beta-1)\left(\frac{1}{\beta-1}\right) = 1$$

یعنی h عددی بزرگتر مساوی صفر و اکیداً کوچکتر از ۱ است. چون $\beta r = r_1 + h$ و $0 \le h < 1$ در نتیجه $r_1 = [\beta r]$. برای معرفی r_2 می‌نویسیم:

$$\beta r - r_1 = r_2\beta^{-1} + r_3\beta^{-2} + \cdots + r_m\beta^{-(m-1)} + \cdots$$

و سپس مشابه روشی که در مرحله‌ی قبل برای یافتن r_1 به‌کار گرفتیم، در این مرحله r_2 را بدست می‌آوریم:

$$r_2 = [\beta r - r_1]$$

اکنون اگر مجدداً $\beta r - r_1$ را r بنامیم، یعنی $r \leftarrow \beta r - r_1$ خواهیم داشت $r_2 = [\beta r]$ و می‌توانیم سایر r_i ها را به روشی مشابه به دست آوریم. واضح است که اگر در مرحله‌ای داشته باشیم $r = 0$، آن‌گاه نمایش r در مبنای β مختوم (دارای تعداد متناهی رقم) است. چنانچه $r \neq 0$ آن‌گاه نمایش r در مبنای β نامختوم (بی‌پایان) است و می‌توانیم روش فوق را تا هرجا که مایل باشیم تکرار کنیم.

با توجه به مطالب فوق، روند تکراری زیر را برای یافتن r_i ها داریم.

۱ـ قرار می‌دهیم $r_1 = [\beta r]$.

۲ـ برای $i \ge 2$ تا رسیدن به نتیجه‌ی مطلوب، اعمال زیر را انجام می‌دهیم:

$$r \leftarrow \beta r - r_{i-1}$$
$$r_i = [\beta r]$$

۳ـ r_i ها به ازای $i = 1, 2, \ldots$ نمایش عدد $0 < r < 1$ در مبنای β هستند.

تذکر: شرط بی‌نهایت بار $r_i \neq \beta - 1$ برای یکتایی بسط لازم است. زیرا مثلاً در مبنای $\beta = 10$ با بی‌نهایت بار $r_i = 9$ با توجه به فرمول حد مجموع یک سری هندسی داریم:

$$0/999\ldots = 0/9 + 0/09 + 0/009 + \cdots = \frac{0/9}{1 - 0/1} = 1$$

و اگر شرط مذکور را برداریم برای عدد ۱، دو بسط اعشاری خواهیم داشت. همچنین:

$$0/2999\ldots = 0/3 \quad , \quad 0/24999\ldots = 0/25$$

و به‌طور کلی، هر عدد اعشاری مختوم (با تعداد متناهی رقم) دارای دو بسط اعشاری خواهد بود.

✍ **مثال ۱-۵:** نمایش دودویی عدد ۰/۴ را به دست آورید.

حل:

i	r	$۲r$	$r_i = [۲r]$	$۲r - r_i$
۱	۰/۴	۰/۸	۰	۰/۸
۲	۰/۸	۱/۶	۱	۰/۶
۳	۰/۶	۱/۲	۱	۰/۲
۴	۰/۲	۰/۴	۰	۰/۴
۵	۰/۴			

مجدداً به $r = ۰/۴$ رسیدیم. پس نمایش دودویی ۰/۴ متناوب است. اعداد ستون r_i، به‌ترتیب از بالا به پایین بیانگر نمایش دودویی عدد ۰/۴ هستند، یعنی $۰/۴ = (۰/\overline{۰۱۱۰})_۲$.

نتیجه: فرض کنید A یک عدد حقیقی مثبت باشد. می‌توان A را به صورت مجموع جزءصحیح و جزءکسری آن نوشت. یعنی $A = [A] + (A)$ که در آن (A) نشان دهنده‌ی جزءکسری A است و برای آن داریم:

$$۰ \leq (A) < ۱$$

اکنون با توجه به دو قضیه‌ی قبل $[A]$ و (A) هر یک دارای نمایش منحصر به فردی در مبنای β هستند. بنابراین با مجموع این دو نمایش، نمایش منحصر به فردی برای A حاصل می‌شود. لذا هر عدد حقیقی مثبت دارای یک نمایش منحصر به فرد در مبنای $\beta > ۱$ است.

✍ **مثال ۱-۶:** نمایش دودویی عدد ۲۲/۴ را به دست آورید.

حل: برای نمایش دودویی عدد ۲۲/۴، با توجه به نتیجه‌ی قبل ابتدا باید آن را به صورت مجموع جزءصحیح و جزءکسری‌اش بنویسیم و سپس نمایش دو جزء آن را با هم جمع کنیم. با توجه به دو مثال قبل داریم:

$$۲۲/۴ = (۱۰۱۱۰/\overline{۰۱۱۰})_۲$$

✎ **تذکر:** در مبنای ۱۶ از حروف بزرگ الفبای لاتین به‌صورت زیر برای نمایش ارقام پس از ۹ استفاده می‌شود:

$$\begin{aligned} &\mathrm{A} = ۱۰, \quad \mathrm{B} = ۱۱, \quad \mathrm{C} = ۱۲, \\ &\mathrm{D} = ۱۳, \quad \mathrm{E} = ۱۴, \quad \mathrm{F} = ۱۵. \end{aligned}$$

هم‌چنین در مبنای ۲۶، از حروف انگلیسی کوچک $a - z$ به‌جای ارقام ۰ تا ۲۵ استفاده می‌شود.

قضیه ۱-۴: اگر نمایش عدد A در مبنای β مختوم یا نامختوم و متناوب باشد، A یک عدد گویاست.

اثبات: ابتدا فرض می‌کنیم نمایش A مختوم باشد. مثلاً $A = (a_m a_{m-۱} \dots a_۱ a_۰ / r_۱ r_۲ \dots r_n)_\beta$. بنابراین:

$$A = a_m \beta^m + a_{m-۱} \beta^{m-۱} + \cdots + a_۱ \beta + a_۰ + r_۱ \beta^{-۱} + r_۲ \beta^{-۲} + \cdots + r_n \beta^{-n}$$

طرفین تساوی را در β^n ضرب می‌کنیم.

$$\beta^n A = \beta^n(a_m\beta^m + a_{m-1}\beta^{m-1} + \cdots + a_1\beta + a_0) + (r_1\beta^{n-1} + r_2\beta^{n-2} + \cdots + r_n)$$

در نتیجه:

$$A = \frac{\beta^n(a_m\beta^m + a_{m-1}\beta^{m-1} + \cdots + a_1\beta + a_0) + (r_1\beta^{n-1} + r_2\beta^{n-2} + \cdots + r_n)}{\beta^n}$$

اکنون از اینکه صورت و مخرج کسر عددهای صحیح هستند نتیجه می‌شود A عددی گویاست.

حال فرض می‌کنیم که نمایش A نامختوم و متناوب باشد. مثلاً:

$$A = (a_m a_{m-1} \ldots a_1 a_0 / r_1 r_2 \ldots r_n \overline{c_1 c_2 \ldots c_k})_\beta$$

(توجه کنید که اگر m یا n صفر باشند، آنگاه $a_m \ldots a_0$ یا $r_1 \ldots r_n$ وجود نخواهد داشت.) از تعریف A داریم:

$$A = (a_m a_{m-1} \ldots a_1 a_0 / r_1 r_2 \ldots r_n)_\beta + (0/\underbrace{00\ldots0}_{n \text{ مرتبه}}\overline{c_1 c_2 \ldots c_k})_\beta$$

با توجه به قسمت قبل، نخستین عبارت سمت راست تساوی یک عدد گویاست. در مورد عبارت دوم داریم:

$$\begin{aligned}
(0/\underbrace{00\ldots0}_{n \text{ مرتبه}}\overline{c_1 c_2 \ldots c_k})_\beta &= \beta^{-n}(0/\overline{c_1 c_2 \ldots c_k})_\beta \\
&= \beta^{-n}(c_1 c_2 \ldots c_k)_\beta \ (\beta^{-k} + \beta^{-2k} + \cdots) \\
&= \beta^{-n}(c_1 c_2 \ldots c_k)_\beta \left(\frac{\beta^{-k}}{1-\beta^{-k}}\right) \\
&= \beta^{-n}(c_1 c_2 \ldots c_k)_\beta \left(\frac{1}{\beta^k - 1}\right) \\
&= \frac{(c_1 c_2 \ldots c_k)_\beta}{\beta^n}\left(\frac{1}{\beta^k - 1}\right)
\end{aligned}$$

که این هم یک عدد گویاست. بنابراین A مجموع دو عدد گویا و در نتیجه، خود عددی گویاست.

نتیجه: اگر A یک عدد گنگ باشد، نمایش A در مبنای β نامختوم و نامتناوب است.

تبدیل مبنای اعداد:

نحوه‌ی تبدیل یک عدد از مبنای ۱۰ به مبنای β را قبلاً در متن درس شرح داده‌ایم. برای تبدیل یک عدد از مبنای β به مبنای ۱۰ چهار حالت در نظر می‌گیریم:

۱. برای تبدیل عدد $A = (a_m a_{m-1} \ldots a_1 a_0)_\beta$ به مبنای ۱۰ از رابطه‌ی زیر استفاده می‌کنیم:

$$A = a_m\beta^m + a_{m-1}\beta^{m-1} + \cdots + a_1\beta + a_0.$$

۲. برای تبدیل عدد $A = (0/r_1 r_2 \ldots r_n)_\beta$ به مبنای ۱۰ از رابطه‌ی زیر استفاده می‌کنیم:

$$A = r_1\beta^{-1} + r_2\beta^{-2} + \cdots + r_n\beta^{-n}.$$

۳. برای تبدیل عدد $A = (0/r_1 r_2 \ldots r_n \overline{c_1 c_2 \ldots c_k})_\beta$ به مبنای ۱۰ از رابطه‌ی زیر استفاده می‌کنیم:

$$A=\frac{(r_1r_2\dots r_nc_1c_2\dots c_k)_\beta-(r_1r_2\dots r_n)_\beta}{(\underbrace{zz\dots z}_{k\text{ مرتبه}}\underbrace{\circ\circ\dots\circ}_{n\text{ مرتبه}})_\beta}$$

که در آن منظور از z رقم $\beta-1$ است. (این رابطه در ادامه ثابت می‌شود.)

۴. برای تبدیل عدد $A=(a_ma_{m-1}\dots a_1a_\circ/r_1r_2\dots r_n\overline{c_1c_2\dots c_k})_\beta$ به مبنای ۱۰ ابتدا بخش‌های صحیح و کسری A را جداگانه به مبنای ۱۰ می‌بریم و سپس آن‌ها را باهم جمع می‌کنیم.

نکته: به کمک نکات قبل می‌توان کسر متعارفی مولد یک عدد اعشاری را نیز تعیین نمود.

اثبات رابطه‌ی قسمت ۳:

$$A=(\circ/r_1r_2\dots r_n\overline{c_1c_2\dots c_k})_\beta=(\circ/r_1r_2\dots r_n)_\beta+(\circ/\underbrace{\circ\circ\dots\circ}_{n\text{ مرتبه}}\overline{c_1c_2\dots c_k})_\beta$$

$$\Longrightarrow A=\frac{(r_1r_2\dots r_n)_\beta}{\beta^n}+\frac{(c_1c_2\dots c_k)_\beta}{\beta^n}\left(\frac{1}{\beta^k-1}\right)=\frac{(\beta^k-1)(r_1r_2\dots r_n)_\beta+(c_1c_2\dots c_k)_\beta}{\beta^n(\beta^k-1)}$$

$$=\frac{(\beta^k)(r_1r_2\dots r_n)_\beta+(c_1c_2\dots c_k)_\beta-(r_1r_2\dots r_n)_\beta}{\beta^n(\beta^k-1)}=\frac{(r_1r_2\dots r_nc_1c_2\dots c_k)_\beta-(r_1r_2\dots r_n)_\beta}{\beta^n(\beta^k-1)}$$

از طرفی:

$$\begin{aligned}\beta^n(\beta^k-1)&=\beta^n(\beta-1)(\beta^{k-1}+\beta^{k-2}\dots+\beta+1)\\&=(\beta-1)\beta^{n+k-1}+(\beta-1)\beta^{n+k-2}+\dots+(\beta-1)\beta^n\\&=(\underbrace{(\beta-1)\dots(\beta-1)}_{k\text{ مرتبه}}\underbrace{\circ\circ\dots\circ}_{n\text{ مرتبه}})_\beta\end{aligned}$$

در نتیجه با قرار دادن $z:=\beta-1$ خواهیم داشت:

$$A=\frac{(r_1r_2\dots r_nc_1c_2\dots c_k)_\beta-(r_1r_2\dots r_n)_\beta}{(\underbrace{zz\dots z}_{k\text{ مرتبه}}\underbrace{\circ\circ\dots\circ}_{n\text{ مرتبه}})_\beta}$$

✍ **مثال ۱-۷:** کسر متعارفی معادل با عدد $۱/۲۳\overline{۴}$ را تعیین کنید.

حل:

$$۱/۲۳\overline{۴}=۱+\frac{۲۳۴-۲۳}{۹۰۰}=۱+\frac{۲۱۱}{۹۰۰}=\frac{۱۱۱۱}{۹۰۰}.$$

روش کلی برای تبدیل مبنای اعداد، استفاده از مبنای ۱۰ به عنوان مبنای واسطه است. به عبارت دیگر برای تبدیل عدد X از مبنای β_1 به مبنای β_2 ابتدا عدد X را از مبنای β_1 به مبنای ۱۰ و سپس از از مبنای ۱۰ به مبنای β_2 می‌بریم. البته در حالتی که β_1 یا β_2 توانی از دیگری باشد یا هر دو توانی از یک عدد دیگر باشند، راهکارهای سریع‌تری وجود دارد که به آن‌ها نیز اشاره خواهیم کرد.

✍ **مثال ۱-۸:** عدد $A=(۱۱۰۰۱/۱۰\overline{۱۱})_۲$ را به مبنای ۳ ببرید.

حل: ابتدا A را به مبنای ۱۰ می‌بریم. داریم:

$$(۱۱۰۰۱)_۲ = ۱ \times ۲^۴ + ۱ \times ۲^۳ + ۰ \times ۲^۲ + ۰ \times ۲^۱ + ۱ \times ۲^۰ = ۱۶ + ۸ + ۱ = ۲۵,$$
$$(۰/۱۰\overline{۱۱})_۲ = \frac{(۱۰۱۱)_۲ - (۱۰)_۲}{(۱۱۰۰)_۲} = \frac{۹}{۱۲} = ۰/۷۵.$$

بنابراین $A = ۲۵/۷۵$. قرار می‌دهیم $A_۱ = [A] = ۲۵$، $r = (A) = ۰/۷۵$ و ابتدا $A_۱$ را به مبنای ۳ می‌بریم:

		باقیمانده	خارج قسمت
i	$A_۱$	$q = \left[\frac{A_۱}{۳}\right]$	$A_۱ - ۳q$
۰	۲۵	۸	$۱ = a_۰$
۱	۸	۲	$۲ = a_۱$
۲	۲	۰	$۲ = a_۲$

بنابراین: $(۲۲۱)_۳ = ۲۵$. حال r را به مبنای ۳ می‌بریم:

i	r	$۳r$	$r_i = [۳r]$	$۳r - r_i$
۱	۰/۷۵	۲/۲۵	۲	۰/۲۵
۲	۰/۲۵	۰/۷۵	۰	۰/۷۵
۳	۰/۷۵			

مجدداً به $r = ۰/۷۵$ رسیدیم. پس نمایش عدد ۰/۷۵ در مبنای سه متناوب است. اعداد ستون r_i، به‌ترتیب از بالا به پایین بیانگر نمایش عدد ۰/۷۵ در مبنای ۳ هستند، یعنی $(۰/\overline{۲۰})_۳ = ۰/۷۵$. اکنون با توجه به دو قسمت قبل داریم:

$$A = (۱۱۰۰۱/۱۰\overline{۱۱})_۲ = ۲۵/۷۵ = (۲۲۱/\overline{۲۰})_۳$$

✍ **مثال ۱-۹:** نمایش در مبنای هشت عدد $A = (۶\mathrm{F}۹)_{۱۶}$ را تعیین کنید.

حل: داریم $\beta_۱ = ۱۶ = ۲^۴$ و $\beta_۲ = ۸ = ۲^۳$. حال ابتدا معادل دودویی هر یک از اعداد ۶، F و ۹ را با در نظر گرفتن چهار مکان می‌نویسیم. با این کار معادل دودویی عدد A مشخص می‌شود:

$$A = (۶\mathrm{F}۹)_{۱۶} = (\underbrace{۰۱۱۰}_{۶}\ \underbrace{۱۱۱۱}_{\mathrm{F}}\ \underbrace{۱۰۰۱}_{۹})_۲$$

اکنون در نمایش دودویی A از سمت راست سه رقم سه رقم جدا می‌کنیم و معادل آن را در مبنای ۸ قرار می‌دهیم:

$$A = (۶\mathrm{F}۹)_{۱۶} = (\underbrace{۰۱۱۰}_{۶}\ \underbrace{۱۱۱۱}_{\mathrm{F}}\ \underbrace{۱۰۰۱}_{۹})_۲ = (\underbrace{۰۱۱}_{۳}\ \underbrace{۰۱۱}_{۳}\ \underbrace{۱۱۱}_{۷}\ \underbrace{۰۰۱}_{۱})_۲ = (۳۳۷۱)_۸ .$$

✍ **مثال ۱-۱۰:** نمایش در مبنای چهار عدد $A = (۵۶۴/۷)_۸$ را تعیین کنید.

حل: داریم $\beta_1 = 8 = 2^3$ و $\beta_2 = 4 = 2^2$. پس ابتدا معادل دودویی هر یک از اعداد ۵، ۶، ۴ و ۷ را با در نظر گرفتن سه مکان می‌نویسیم. با این کار معادل دودویی عدد A مشخص می‌شود:

$$A = (564/7)_8 = (\underbrace{101}_{5}\ \underbrace{110}_{6}\ \underbrace{100}_{4}/\underbrace{111}_{7})_2$$

حال در قسمت صحیح عدد، از سمت راست و در قسمت کسری از سمت چپ، دو رقم دو رقم جدا می‌کنیم و معادل آن را در مبنای ۴ قرار می‌دهیم. توجه کنید که در قسمت صحیح (کسری)، در سمت چپ (راست) یک رقم اضافه می‌آید که با قرار دادن یک صفر در سمت چپ (راست) آن تعداد مکان‌های آن را به دو می‌رسانیم.

$$(11310/32)_4 = (\underbrace{10}_{2}\ \underbrace{11}_{3}\ \underbrace{00}_{0}/\underbrace{01}_{1}\ \underbrace{01}_{1}\ \underbrace{11}_{3}\ \underbrace{01}_{1}\ \underbrace{00}_{0}/\underbrace{11}_{3}\ \underbrace{11}_{3}) = (\underbrace{101}_{5}\ \underbrace{110}_{6}\ \underbrace{100}_{4}/\underbrace{111}_{7})_2 = (564/7)_8$$

۱-۴ نمایش ماشینی اعداد

در کامپیوترها مقدار حافظه‌ای که برای ذخیره‌ی هر عدد اختصاص داده می‌شود (طول کلمه) محدود است. به همین دلیل همواره تعداد محدودی از اعداد حقیقی قابل ذخیره در کامپیوتر هستند. به‌طور کلی برای ذخیره‌ی اعداد در کامپیوترها از یکی از دو سیستم نمایش ممیز ثابت و نمایش ممیز شناور استفاده می‌شود. البته سیستم نمایش ممیز ثابت بیشتر در کامپیوترهای اولیه مورد استفاده قرار می‌گرفت و امروزه در کامپیوترها از سیستم نمایش ممیز شناور استفاده می‌شود.

سیستم نمایش ممیز ثابت:

در این سیستم نمایش، اگر طول کلمه‌ی کامپیوتر ℓ باشد، n مکان آن برای نگهداری ارقام قبل از ممیز، m مکان آن برای نگهداری ارقام بعد از ممیز و ۱ مکان برای علامت (منفی یا مثبت) در نظر گرفته می‌شود که

$$\ell = m + n + 1$$

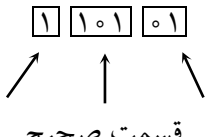
به عبارت دیگر مجموعه‌ی تمام اعداد در این سیستم نمایش به‌صورت زیر است:

$$\boxed{\mathbb{F}_0 = \left\{\pm(a_{n-1}a_{n-2}\dots a_1 a_0/b_1 b_2 \dots b_m)_\beta \mid a_i, b_j \in \{0,1,\dots,\beta-1\}\right\}}$$

توجه کنید که همواره، مکان علامت با یکی از ارقام ۰ و ۱ پر می‌شود که رقم ۰ برای علامت مثبت و رقم ۱ برای علامت منفی به‌کار می‌رود. مثلاً اگر در یک سیستم نمایش ممیز ثابت داشته باشیم:

$$\beta = 2, \quad m = 2, \quad n = 3,$$

آن‌گاه عدد $-5/25 = -(101/01)_2$ در این سیستم به‌صورت زیر نمایش داده می‌شود:



توجه کنید که در همین سیستم، مثلاً عدد ۱۰ قابل نمایش نیست زیرا $(۱۰۱۰)_۲ = ۱۰$ که نیاز به چهار مکان در بخش صحیح دارد. بنابراین سؤالی که در این‌جا طبیعی است پرسیده شود این است که اساساً چند عدد در این سیستم قابل نمایش هستند و بزرگترین و کوچکترین آن‌ها کدامند؟ برای پاسخ به این سؤال ابتدا قضیه‌ی زیر را بیان می‌کنیم.

قضیه ۱-۵: در یک سیستم نمایش ممیز ثابت با مبنای β که در آن، n مکان برای نگهداری ارقام قبل از ممیز، m مکان برای نگهداری ارقام بعد از ممیز و ۱ مکان برای علامت (منفی یا مثبت) در نظر گرفته می‌شود:

الف) تعداد اعداد حقیقی قابل نمایش عبارت است از $۲\beta^{m+n} - ۱$؛

ب) بزرگترین عدد مثبت قابل نمایش برابر است با $\beta^n - \beta^{-m}$؛

ج) کوچکترین عدد مثبت قابل نمایش برابر است با β^{-m}.

اثبات: الف) تعداد حالت‌های پر کردن مکان علامت برابر است با ۲ و هر یک از دیگر مکان‌ها را نیز می‌توان با یکی از β رقم ۰ تا $\beta - ۱$ پر کرد که چون تعداد این مکان‌ها $m+n$ است، طبق اصل ضرب این کار به β^{m+n} حالت امکان‌پذیر است. از طرفی $+۰$ و -۰ هر دو نمایش عدد صفر هستند. در نتیجه تعداد تمام اعداد قابل نمایش در این سیستم عبارت است از $۲\beta^{m+n} - ۱$.

ب) فرض می‌کنیم x_{max} بزرگترین عدد مثبت قابل نمایش در این سیستم باشد و قرار می‌دهیم $z := \beta - ۱$. لذا:

$$x_{max} = +(\underbrace{zz\dots z}_{n \text{ مرتبه}} / \underbrace{zz\dots z}_{m \text{ مرتبه}})_\beta = \sum_{i=-m}^{n-۱} z\beta^i = \sum_{i=-m}^{n-۱} (\beta-۱)\beta^i = \sum_{i=-m}^{n-۱} (\beta^{i+۱} - \beta^i) = \beta^n - \beta^{-m}.$$

ج) فرض می‌کنیم x_{min} کوچکترین عدد مثبت قابل نمایش در این سیستم باشد. بنابراین:

$$x_{min} = +(\underbrace{۰۰\dots۰}_{n \text{ مرتبه}} / \underbrace{۰۰\dots۰}_{m-۱ \text{ مرتبه}} ۱)_\beta = \beta^{-m}.$$

نکته: توزیع اعداد در سیستم نمایش ممیز ثابت یکنواخت است. به این معنی که تمام اعداد با فواصل یکسان در بازه‌ی $[-x_{max}, x_{max}]$ قرار دارند. فاصله‌ی هر دو عدد متوالی برابر β^{-m} است.

حال سؤالی را که قبلاً مطرح کردیم در مثال زیر پاسخ می‌دهیم.

✍ **مثال ۱-۱۱:** در سیستم نمایش ممیز ثابت با

$$\beta = ۲, \quad m = ۲, \quad n = ۳,$$

تعداد اعداد حقیقی قابل نمایش و نیز بزرگترین و کوچکترین اعداد مثبت قابل نمایش را مشخص کنید.

حل: تعداد اعداد حقیقی قابل نمایش در این سیستم برابر است با $۲ \times ۲^{۲+۳} - ۱ = ۶۳$. همچنین بزرگترین و کوچکترین اعداد مثبت قابل نمایش در این سیستم عبارتند از:

$$x_{max} = ۲^۳ - ۲^{-۲} = ۸ - ۰/۲۵ = ۷/۷۵, \qquad x_{min} = ۲^{-۲} = ۰/۲۵.$$

ضعف اصلی سیستم نمایش ممیز ثابت در این است که بازه‌ی $[-x_{max}, x_{max}] \approx [-\beta^n, \beta^n]$ محدوده‌ی چندان

بزرگی از اعداد حقیقی نیست مگر آنکه n بسیار بزرگ انتخاب شود که حافظه‌ی محدود کامپیوتر این اجازه را نمی‌دهد. این ضعف از آن‌جا ناشی می‌شود که توزیع اعداد در این سیستم یکنواخت است. به‌عبارت دیگر حساسیت این سیستم روی اعداد کوچک و بزرگ یکی است. برای رفع این مشکل از سیستم ممیز شناور استفاده می‌شود که در ادامه به معرفی آن می‌پردازیم.

سیستم نمایش ممیز شناور:

تعریف ۱-۸: نمایش ممیز شناور یک عدد حقیقی دلخواه مانند A در حالت کلی (مبنای β) به‌صورت زیر است:

$$A = \pm a \times \beta^e,$$

در نمایش فوق، a عددی بین صفر و یک در مبنای β است که به آن مانتیس می‌گویند و e عددی صحیح در مبنای β است که به آن نما گفته می‌شود[۱].

با توجه به تعریف فوق، نمایش ممیز شناور هر عدد حقیقی از چهار پارامتر علامت، مانتیس، مبنا و نما تشکیل می‌شود. هم‌چنین، یک عدد حقیقی ممکن است نمایش‌های ممیز شناور متفاوتی داشته باشد. مثلاً هر یک از اعداد $(۰/۰۰۱)_\beta \times \beta^۳$، $(۰/۰۱)_\beta \times \beta^۲$ و $(۰/۱)_\beta \times \beta^۱$ نمایش ممیز شناور عدد ۱ در مبنای β هستند. در حقیقت در نمایش ممیز شناور می‌توان ممیز را در طول ارقام مانتیس جابه‌جا نمود مشروط بر اینکه نما به‌طور مناسب تغییر کند. اگر ممیز یک رقم به سمت راست حرکت کند از نما یک واحد کم می‌شود. چنانچه شرط کنیم که اولین رقم پس از ممیز در مانتیس همواره مخالف صفر باشد (مگر در حالتی که مانتیس صفر باشد)، آنگاه هر عدد حقیقی، نمایش ممیز شناور یکتایی خواهد داشت که به آن نمایش ممیز شناور نرمال گفته می‌شود. بنابراین در نمایش ممیز شناور نرمال، حداقل مقدار مانتیس برابر با $(۰/۱)_\beta = \beta^{-۱}$ است.

درکامپیوترها به دلیل محدودیت حافظه، تعداد ارقام مانتیس و نما محدود انتخاب می‌شوند. بنابراین در یک سیستم ممیز شناور نرمال، اگر فرض کنیم که طول کلمه‌ی کامپیوتر ℓ باشد، آنگاه n مکان برای نگهداری ارقام نما، m مکان برای نگهداری ارقام مانتیس، ۱ مکان برای علامت مانتیس و ۱ مکان برای علامت نما در نظر گرفته می‌شود که $\ell = m + n + ۲$. به‌عبارت دیگر، مجموعه‌ی تمام اعداد قابل نمایش در این سیستم به‌صورت زیر است:

$$\boxed{\mathbb{F} = \left\{ \pm(۰/b_۱ b_۲ \dots b_m)_\beta \times \beta^{\pm(e_۱ e_۲ \dots e_n)_\beta} \mid b_i, e_j \in \{۰, ۱, \dots, \beta - ۱\}, b_۱ \neq ۰ \right\} \cup \{۰\}.}$$

در این سیستم نمایش نیز مکان‌های علامت با یکی از دو رقم ۰ و ۱ پر می‌شوند که رقم ۰ برای علامت مثبت و رقم ۱ برای علامت منفی به‌کار می‌رود. مثلاً اگر در یک سیستم نمایش ممیز شناور نرمال داشته باشیم:

$$\beta = ۲, \qquad m = ۳, \qquad n = ۲,$$

آنگاه عدد $-(۰/۱۱۰)_۲ \times ۲^{-(۰۱)_۲} = -۰/۳۷۵$ در این سیستم به‌صورت زیر نمایش می‌شود:

۱	۱۱۰	۱	۰۱

در نمایش فوق اولین رقم از چپ علامت مانتیس، سه رقم بعدی ارقام مانتیس (آن‌هایی که در سمت راست ممیز

[۱] چنانچه $\beta > a \geq ۱$ نمایش حاصل را نمایش علمی A می‌گویند.

در نمایش دودویی قرار دارند)، رقم بعدی علامت نما و دو رقم آخر ارقام نما را مشخص می‌کنند.

قضیه ۱-۶: در یک سیستم نمایش ممیز شناور نرمال با مبنای β که در آن، n مکان برای نگهداری ارقام نما، m مکان برای نگهداری ارقام مانتیس، و ۲ مکان برای علامت‌های مانتیس و نما (منفی یا مثبت) در نظر گرفته می‌شود:

الف) تعداد اعداد حقیقی قابل نمایش عبارت است از $۲(\beta-۱)\beta^{m-۱}(۲\beta^n-۱)+۱$؛

ب) بزرگترین عدد مثبت قابل نمایش برابر است با $\beta^{\beta^n-۱}(۱-\beta^{-m})$؛

ج) کوچکترین عدد مثبت قابل نمایش برابر است با $\beta^{-\beta^n}$.

اثبات: الف) ابتدا تعداد مانتیس‌های ممکن را می‌شماریم. علامت مانتیس می‌تواند مثبت یا منفی باشد (۲حالت). از آن‌جا که هر عدد ناصفر قابل نمایش در این سیستم، نرمال شده است، اولین رقم مانتیس نمی‌تواند صفر باشد. پس اولین رقم مانتیس یکی از ارقام ۱ تا $\beta-۱$ است ($\beta-۱$ حالت). سایر ارقام مانتیس می‌توانند هر یک از رقم‌های ۰ تا $\beta-۱$ را اختیار کنند (در مجموع $\beta^{m-۱}$ حالت). بنابراین طبق اصل ضرب، تعداد مانتیس‌های ممکن برابر با $۲(\beta-۱)\beta^{m-۱}$ است. حال تعداد نماهای ممکن را می‌شماریم. علامت نما می‌تواند مثبت یا منفی باشد (۲حالت) و هر یک از ارقام نما می‌توانند یکی از β رقم ۰ تا $\beta-۱$ را اختیار کنند (در مجموع β^nحالت). از طرفی در نما $+۰$ و -۰ هر دو نمایش عدد صفر هستند. در نتیجه تعداد نماهای ممکن برابر با $۲\beta^n-۱$ است. لذا تعداد تمام اعداد ناصفر قابل نمایش در این سیستم عبارت است از $۲(\beta-۱)\beta^{m-۱}(۲\beta^n-۱)$. اکنون با افزودن عدد صفر به مجموعه‌ی اعداد فوق، نتیجه‌ی مطلوب حاصل می‌شود.

ب) فرض می‌کنیم x_{max} بزرگترین عدد مثبت قابل نمایش در این سیستم باشد و قرار می‌دهیم $z:=\beta-۱$. لذا:

$$x_{max}=+(۰/\underbrace{zz\dots z}_{m\text{ مرتبه}})_\beta\times\beta^{+(\underbrace{zz\dots z}_{n\text{ مرتبه}})_\beta}=\left(\sum_{i=-m}^{-۱}z\beta^i\right)\beta^{(\sum_{i=۰}^{n-۱}z\beta^i)}=(۱-\beta^{-m})\beta^{\beta^n-۱}.$$

ج) فرض می‌کنیم x_{min} کوچکترین عدد مثبت قابل نمایش در این سیستم باشد. بنابراین با توجه به این‌که اولین رقم مانتیس نمی‌تواند صفر باشد داریم:

$$x_{min}=+(۰/۱\underbrace{۰۰\dots۰}_{m-۱\text{ مرتبه}})_\beta\times\beta^{-(\underbrace{zz\dots z}_{n\text{ مرتبه}})_\beta}=\beta^{-۱}\times\beta^{-(\beta^n-۱)}=\beta^{-\beta^n}.$$

نکته: توزیع اعداد در سیستم نمایش ممیز شناور نرمال یکنواخت نیست. در حقیقت اگر $x\in\mathbb{F}$ یک عدد مثبت و دارای نمایش $(۰/b_۱b_۲\dots b_m)_\beta\times\beta^e$ باشد، عدد بزرگ‌تر بلافاصله بعد از آن در این سیستم نمایش که با x_+ نشان داده می‌شود، دارای نمایش $(۰/b_۱b_۲\dots b_m+۰/\underbrace{۰۰\dots۰}_{m-۱\text{ مرتبه}}۱)_\beta\times\beta^e$ است. بنابراین فاصله‌ی بین دو عدد قابل نمایش متوالی x و x_+ برابر است با $(۰/\underbrace{۰۰\dots۰}_{m-۱\text{ مرتبه}}۱)_\beta\times\beta^e=\beta^{e-m}$. این فاصله به‌وضوح به نما وابسته است، یعنی با افزایش نما، فاصله‌ی بین اعداد بیشتر خواهد شد.

✍**مثال ۱-۱۲:** در سیستم نمایش ممیز شناور نرمال با

$$\beta = ۲, \quad m = ۳, \quad n = ۱,$$

الف) تعداد اعداد حقیقی قابل نمایش و نیز بزرگترین و کوچکترین اعداد مثبت قابل نمایش را مشخص کنید.

ب) اعداد مثبت قابل نمایش در این سیستم را لیست کرده و بر روی محور اعداد مشخص کنید.

حل: الف) تعداد اعداد حقیقی قابل نمایش در این سیستم برابر است با:

$$۲ \times (۲-۱) \times ۲^{۳-۱} \times (۲ \times ۲^{۱} - ۱) + ۱ = ۲۵,$$

همچنین بزرگترین و کوچکترین اعداد مثبت قابل نمایش در این سیستم عبارتند از:

$$x_{max} = (۱ - ۲^{-۳}) \times ۲ = ۲ - ۲^{-۲} = \frac{۷}{۴}, \qquad x_{min} = ۲^{-۲} = \frac{۱}{۴}.$$

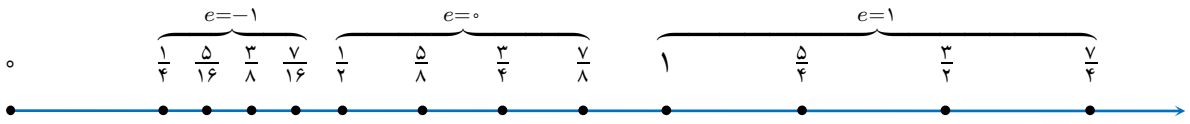
ب) هر عدد مثبت قابل نمایش در این سیستم به شکل $x = (۰/b_۱b_۲b_۳)_۲ \times ۲^{\pm e}$ است که در آن:

$$e, b_i \in \{۰,۱\}, b_۱ \neq ۰.$$

بنابراین اعداد مثبت قابل نمایش در این سیستم عبارتند از:

$۰/۱۰۰ \times ۲^{۱} = ۱$	$۰/۱۰۰ \times ۲^{۰} = \frac{۱}{۲}$	$۰/۱۰۰ \times ۲^{-۱} = \frac{۱}{۴}$
$۰/۱۰۱ \times ۲^{۱} = \frac{۵}{۴}$	$۰/۱۰۱ \times ۲^{۰} = \frac{۵}{۸}$	$۰/۱۰۱ \times ۲^{-۱} = \frac{۵}{۱۶}$
$۰/۱۱۰ \times ۲^{۱} = \frac{۳}{۲}$	$۰/۱۱۰ \times ۲^{۰} = \frac{۳}{۴}$	$۰/۱۱۰ \times ۲^{-۱} = \frac{۳}{۸}$
$۰/۱۱۱ \times ۲^{۱} = \frac{۷}{۴}$	$۰/۱۱۱ \times ۲^{۰} = \frac{۷}{۸}$	$۰/۱۱۱ \times ۲^{-۱} = \frac{۷}{۱۶}$

نمایش این اعداد بر روی محور، به‌صورت زیر است:



همان‌طور که در شکل فوق ملاحظه می‌شود، فاصله‌ی بین اعداد با نمای بزرگ‌تر، بیشتر از فاصله‌ی بین اعداد با نمای کوچک‌تر است و اعداد با نمای یکسان، به‌طور متساوی‌الفاصله از یک‌دیگر قرار دارند.

گاهی به‌جای آن‌که در نمایش ممیز شناور نرمال تعداد ارقام نما مشخص شود، محدوده‌ای مانند $L \leq e \leq U$ برای نما در نظر گرفته می‌شود که در آن $L < ۰$ و $U > ۰$ دو عدد صحیح هستند. یعنی در این حالت مجموعه‌ی تمام اعداد ممیز شناور نرمال به‌صورت زیر است:

$$\boxed{\mathbb{F} = \left\{ \pm(۰/b_۱b_۲ \dots b_m)_\beta \times \beta^e \mid b_i \in \{۰,۱,\dots,\beta-۱\}, b_۱ \neq ۰, L \leq e \leq U \right\} \cup \{۰\}.}$$

این سیستم نمایش را اغلب به همراه پارامترهای مشخص‌کننده‌ی آن و به‌صورت $\mathbb{F}(\beta, m, L, U)$ نشان می‌دهند.

ویژگی‌های سیستم نمایش ممیز شناور نرمال در این حالت به‌صورت زیر بیان می‌شوند:

الف) تعداد اعداد حقیقی قابل نمایش عبارت است از $۲(\beta-۱)\beta^{m-۱}(U-L+۱)+۱$؛

ب) بزرگترین عدد مثبت قابل نمایش برابر است با $(۱-\beta^{-m})\beta^{U}$؛

ج) کوچکترین عدد مثبت قابل نمایش برابر است با $\beta^{L-۱}$.

د) اگر $x\in\mathbb{F}$ یک عدد مثبت و دارای نمایش $(۰/b_۱b_۲\ldots b_m)_\beta\times\beta^e$ باشد، آنگاه فاصله‌ی بین x و x_+ برابر است با β^{e-m}. بنابراین توزیع اعداد در این سیستم نمایش یکنواخت نیست. کمترین فاصله‌ی ممکن بین دو عدد قابل نمایش به ازای $e=L$ و بیشترین فاصله‌ی ممکن به‌ازای $e=U$ رخ می‌دهد.

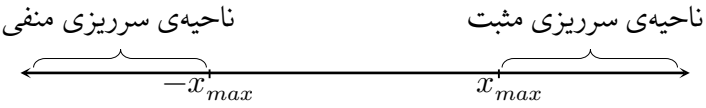
اعداد ماشینی و غیرماشینی:

تعریف ۱-۹: در یک ماشین محاسباتی که از یک سیستم نمایش اعداد استفاده می‌کند، تنها تعدادی متناهی از اعداد حقیقی را می‌توان نمایش داد. به اعداد قابل نمایش در ماشین، اعداد ماشینی و به اعداد غیر قابل نمایش در ماشین، اعداد غیرماشینی می‌گویند. مجموعه‌ی اعداد ماشینی را با A و مجموعه‌ی اعداد غیرماشینی را با A^c نشان می‌دهند.

سرریزی (over flow):

تعریف ۱-۱۰: اگر در یک سیستم نمایش برای عدد حقیقی x داشته باشیم $|x|>x_{max}$، آنگاه نمایش x در این سیستم امکان‌پذیر نیست. به چنین حالتی اصطلاحاً پدیده‌ی سرریزی گفته می‌شود.

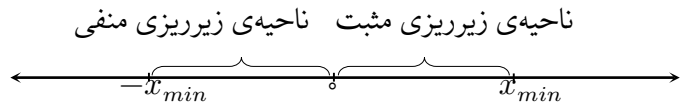
با رخ دادن پدیده‌ی سرریزی در یک ماشین، انجام عملیات توسط ماشین با دادن یک پیغام خطا متوقف می‌شود. پدیده‌ی سرریزی در سیستم نمایش ممیز شناور به علت محدود بودن نما از بالا رخ می‌دهد. ناحیه‌ی سرریزی را می‌توان به شکل زیر روی محور اعداد حقیقی مشخص کرد:



زیرریزی (under flow):

تعریف ۱-۱۱: اگر در یک سیستم نمایش برای عدد حقیقی x داشته باشیم $۰<|x|<x_{min}$، آنگاه نمایش x در این سیستم امکان‌پذیر نیست. به چنین حالتی اصطلاحاً پدیده‌ی زیرریزی گفته می‌شود.

با رخ دادن پدیده‌ی زیرریزی در یک ماشین، مقدار x برابر با صفر در نظر گرفته می‌شود. پدیده‌ی زیرریزی در سیستم نمایش ممیز شناور به علت محدود بودن نما از پایین رخ می‌دهد. ناحیه‌ی زیرریزی را می‌توان به شکل زیر روی محور اعداد حقیقی مشخص کرد. به ناحیه‌ی زیرریزی اصطلاحاً «حفره در صفر» نیز می‌گویند.



استاندارد IEEE برای نمایش اعداد ممیز شناور:

به‌لحاظ تاریخی، کامپیوترهای گوناگون انتخاب‌های متفاوتی در تعیین مبنا، کران‌های نما و ارقام مانتیس داشته‌اند. این تفاوت‌ها منجر به انتقال‌ناپذیری نرم‌افزارها روی ماشین‌های متفاوت می‌گردید تا این که در سال ۱۹۸۵با تلاش‌های گروهی از ریاضیدانان، دانشمندان علوم کامپیوتر و شرکت‌های تولید سخت‌افزار به سرپرستی ویلیام کاهان* از دانشگاه کالیفرنیا، استانداردی برای نمایش اعداد ممیز شناور تحت عنوان IEEE۷۵۴ به سازندگان سخت‌افزارها عرضه شد. هم‌اکنون در بیشتر کامپیوترها از این استاندارد استفاده می‌شود. استاندارد IEEE، چند قالب کلی با دقت‌های مختلف از جمله دقت معمولی، دقت مضاعف و دقت‌های معمولی و مضاعف توسعه یافته برای نمایش اعداد ارائه می‌کند. در این‌جا به منظور آشنایی بیشتر با شیوه‌ی نمایش اعداد در این استاندارد، نحوه‌ی نمایش در دقت معمولی و مضاعف را شرح می‌دهیم.

استاندارد IEEE برای دقت معمولی، سیستم ممیز شناور نرمال $\mathbb{F}(۲,۲۴,-۱۲۶,۱۲۷)$ و برای دقت مضاعف، سیستم ممیز شناور نرمال $\mathbb{F}(۲,۵۳,-۱۰۲۲,۱۰۲۳)$ را پیشنهاد می‌کند. نحوه‌ی ذخیره‌سازی اعداد در این استاندارد کمی با شیوه‌ی عادی متفاوت است که این تفاوت به منظور استفاده هرچه بهینه‌تر از حافظه‌ی ماشین است. مطابق این استاندارد، در دقت معمولی از ۳۲ بیت و در دقت مضاعف از ۶۴ بیت برای نمایش یک عدد استفاده می‌شود. هر نمایش از سه بخش تشکیل می‌شود که عبارتند از علامت (s)، نمای تعدیل یافته (c) و قسمت کسری مانتیس نرمال شده (f). این سه بخش با استفاده از روابط زیر مشخص می‌شوند و به‌صورت $\boxed{s|c|f}$ در کنار هم قرار می‌گیرند:

دقت معمولی: $$x = \pm(۱.f)_۲ \times ۲^{e} = (-۱)^{s}(۱.f)_۲ \times ۲^{c-۱۲۷}$$

دقت مضاعف: $$x = \pm(۱.f)_۲ \times ۲^{e} = (-۱)^{s}(۱.f)_۲ \times ۲^{c-۱۰۲۳}$$

همان‌طور که ملاحظه می‌کنید شکل کلی قالب‌های ذکر شده در دقت‌های معمولی و مضاعف، کمی شبیه به نمایش ممیز شناور نرمال است. فقط باید توجه داشت که در استاندارد IEEE مانتیس x به صورت $(۱.f)_۲$ نرمال شده است و تنها قسمتی از مانتیس که با f نشان داده شده است نمایش داده می‌شود. در واقع، چون اولین بیت مانتیس نرمال شده همواره برابر با ۱ است نیازی به ذخیره‌سازی آن نیست. در عوض، این بیت برای نمایش نما مورد استفاده قرار می‌گیرد. همچنین برای آنکه در ذخیره‌سازی نمای واقعی (e) نیازی به بیت علامت نباشد، این نما به صورت اریبی تعدیل و آنگاه ذخیره می‌شود. به عبارت دیگر یک مقدار ثابت به نام *اُریب* به نما (e) اضافه می‌شود تا یک عدد بدون علامت (c) در حافظه ذخیره شود. به‌سادگی می‌توان دید که برای دقت ساده اریب برابر با ۱۲۷ و برای دقت مضاعف برابر با ۱۰۲۳ است.

✍**مثال ۱-۱۳:** عدد $x = -۴۵/۷۵$ را در نظر بگیرید. می‌خواهیم این عدد را در استاندارد IEEE با دقت معمولی نمایش دهیم. برای این منظور، ابتدا نمایش دودویی آن را می‌یابیم. داریم $x = -(۱۰۱۱۰۱/۱۱)$. حال این عدد باید به فرم $(۱.f)_۲ \times ۲^{e}$ تبدیل شود. به‌سادگی می‌توان دید که $x = -(۱/۰۱۱۰۱۱۱) \times ۲^{۵}$.

اکنون از این تساوی باید مقادیر s، f و c را استخراج کنیم. با توجه به قالب کلی دقت ساده داریم:

$$s = ۱, \quad f = ۰۱۱۰۱۱۱, \quad e = c - ۱۲۷ = ۵$$

در نتیجه $c = ۱۳۲ = (۱۰۰۰۰۱۰۰)_۲$ و بنابراین: $x = ۱|۱۰۰۰۰۱۰۰|۰۱۱۰۱۱۱۰۰۰۰۰۰۰۰۰۰۰۰۰۰$.

✍**مثال ۱-۱۴:** عدد زیر در استاندارد IEEE با دقت معمولی نمایش داده شده است.

$$y = ۰|۱۰۰۰۰۰۰۱|۱۰۱۰۱۰۰۰۰۰۰۰۰۰۰۰۰۰$$

می‌خواهیم نمایش اعشاری آن را بیابیم. با توجه به نمایش فوق داریم:

$$s = ۰, \quad c = (۱۰۰۰۰۰۰۱)_۲ = ۱۲۹, \quad f = ۱۰۱۰۱$$

بنابراین y عددی مثبت است و $e = c - ۱۲۷ = ۱۲۹ - ۱۲۷ = ۲$. در نتیجه:

$$y = +(۱.f)_۲ \times ۲^e = (۱.۱۰۱۰۱)_۲ \times ۲^۲ = (۱۱۰.۱۰۱)_۲ = ۶٫۶۲۵$$

۱-۵ خطاهای نمایش اعداد

در این بخش، نحوه‌ی روند کردن اعداد را به دو شکل، یکی روند کردن در محاسبات تقریبی دستی و دیگری روند کردن ماشینی مطرح می‌کنیم و خطاهای روند کردن را در هر حالت مورد بررسی قرار می‌دهیم. همان‌طور که می‌دانید بیشتر اعداد حقیقی دارای نمایشی بی‌پایان هستند. به همین دلیل در محاسبات تقریبی نمی‌توان همه‌ی اعداد حقیقی را آن‌طور که واقعاً هستند وارد محاسبات نمود. در این موارد گاهی لازم است که تقریبی از یک عدد را با در نظر گرفتن تعدادی متناهی از ارقام اعشاری آن انتخاب کنیم. این عمل که آن را روند کردن می‌نامیم، به دو روش صورت می‌گیرد:

الف) قطع کردن

در این روش، بسط اعشاری عدد را با توجه به تعداد ارقامی که می‌توانیم یا می‌خواهیم نگهداری کنیم از رقم معینی قطع می‌کنیم. این رقم را اولین رقم ناخواسته می‌نامیم. بنابراین در روش قطع کردن، اولین رقم ناخواسته و ارقام پس از آن قطع می‌شوند.

✍**مثال ۱-۱۵:** قطع شده‌ی عدد ... $e = ۲٫۷۱۸۲$ تا دو رقم اعشار (یا تا ۲D)[۱] عبارت است از: ۲٫۷۱.

✍**مثال ۱-۱۶:** قطع شده‌ی عدد ... $\frac{۵}{۳} = ۱٫۶۶۶۶$ تا سه رقم اعشار (یا تا ۳D) عبارت است از: ۱٫۶۶۶.

ب) گرد کردن

در این روش، با توجه به مقدار اولین رقم ناخواسته به شرح زیر عمل می‌کنیم:

۱. اگر اولین رقم ناخواسته کوچک‌تر از ۵ باشد، آن رقم و ارقام پس از آن را قطع می‌کنیم.

۲. اگر اولین رقم ناخواسته بزرگ‌تر از ۵ باشد، به رقم قبلی آن یک واحد اضافه می‌کنیم و رقم ناخواسته و ارقام پس

[۱] D حرف اول واژه‌ی Decimal به معنای «اعشاری» است. از این رو نمایش یک عدد تا n رقم اعشار با نماد $(n\mathrm{D})$ بیان می‌شود.

از آن را قطع می‌کنیم.

۳. اگر اولین رقم ناخواسته مساوی با ۵ باشد و بعد از آن رقم مخالف صفری وجود داشته باشد، به رقم قبل از ۵ یک واحد اضافه می‌کنیم و رقم ۵ و ارقام پس از آن را قطع می‌کنیم.

۴. اگر اولین رقم ناخواسته مساوی با ۵ باشد و بعد از آن رقم مخالف صفری وجود نداشته باشد، آنگاه در صورتی که رقم قبل از ۵ فرد باشد یک واحد به آن اضافه می‌کنیم و در صورتی که رقم قبل از ۵ زوج باشد آن را تغییر نمی‌دهیم و رقم ۵ و ارقام پس از آن را قطع می‌کنیم.[۱]

✍ **مثال ۱-۱۷:** گرد شده‌ی عدد ... $e = ۲٫۷۱۸۲$ تا یک رقم اعشار عبارت است از: $e = ۲٫۷$ (1D).

✍ **مثال ۱-۱۸:** گرد شده‌ی عدد ... $e = ۲٫۷۱۸۲$ تا دو رقم اعشار عبارت است از: $e = ۲٫۷۲$ (2D).

✍ **مثال ۱-۱۹:** گرد شده‌ی عدد $w = ۰٫۷۵۰۲$ تا یک رقم اعشار عبارت است از: $w = ۰٫۸$ (1D).

✍ **مثال ۱-۲۰:** گرد شده‌ی عدد $x = ۱٫۴۳۵$ تا دو رقم اعشار عبارت است از: $x = ۱٫۴۴$ (2D).

✍ **مثال ۱-۲۱:** گرد شده‌ی عدد $y = ۰٫۸۴۵$ تا دو رقم اعشار عبارت است از: $y = ۰٫۸۴$ (2D).

✍ **مثال ۱-۲۲:** گرد شده‌ی عدد $z = ۳٫۹۹۵$ تا دو رقم اعشار عبارت است از: $z = ۴٫۰۰$ (2D).

مقایسه‌ی خطای قطع کردن با خطای گرد کردن

فرض کنید a گرد شده‌ی عدد ... $\frac{۲}{۳} = ۰٫۶۶۶$ و b قطع شده‌ی $\frac{۲}{۳}$ تا دو رقم اعشار باشند. داریم:

$$a = ۰٫۶۷ \Longrightarrow \left|\frac{۲}{۳} - a\right| = \frac{۱}{۳۰۰},$$
$$b = ۰٫۶۶ \Longrightarrow \left|\frac{۲}{۳} - b\right| = \frac{۲}{۳۰۰}.$$

یعنی فاصله‌ی b تا $\frac{۲}{۳}$ دو برابر فاصله‌ی a تا $\frac{۲}{۳}$ است. به عبارت دیگر، خطای قطع کردن تا دو برابر خطای گرد کردن است. به همین دلیل در عمل بیشتر از گرد کردن استفاده می‌شود (هرچند که قطع کردن ساده‌تر است).

قضیه ۱-۷: فرض کنید a و b به‌ترتیب گرد شده و قطع شده‌ی عدد A تا n رقم اعشار باشند. در این‌صورت:

$$e(a) \leq ۵ \times ۱۰^{-(n+۱)} = \frac{۱}{۲} \times ۱۰^{-n},$$
$$e(b) \leq ۱۰^{-n}.$$

اثبات: فرض کنید که $r = ۰/d_۱d_۲ \dots d_n d_{n+۱} \dots$ نمایش قسمت کسری A باشد و a را گرد شده‌ی A تا n رقم اعشار در نظر بگیرید. اگر $d_{n+۱}$ کوچک‌تر از ۵ باشد، آنگاه:

$$e(a) = |A - a| = ۰/\underbrace{۰۰ \dots ۰}_{n \text{ مرتبه}} d_{n+۱} \dots \leq ۵ \times ۱۰^{-(n+۱)}$$

---

[۱] وضع این قاعده که منسوب به گاوس* است، به این دلیل است که در یک مسئله با محاسبات زیاد، احتمال وقوع اعداد اعشاری با رقم زوج قبل از ۵ و احتمال وقوع اعداد اعشاری با رقم فرد قبل از ۵ با هم برابرند و از این رو میانگین خطای ناشی از گرد کردن متناظر با آن‌ها در محاسبات صفر است.

اگر d_{n+1} بزرگ‌تر از ۵ باشد به رقم d_n یک واحد اضافه می‌شود. بنابراین:

$$e(a) = |A - a| = ۱ \times ۱۰^{-n} - ۰/\underbrace{۰۰\ldots۰}_{n\text{ مرتبه}} d_{n+۱} \ldots \leq ۱۰^{-n} - ۵ \times ۱۰^{-(n+۱)} = ۵ \times ۱۰^{-(n+۱)}.$$

اگر $d_{n+۱} = ۵$، آن‌گاه با توجه به زوج یا فرد بودن d_n یکی از دو حالت قبل رخ می‌دهد و باز هم حکم برقرار است (در این حالت علامت‌های $\leq$ به $=$ تبدیل می‌شوند).

برای اثبات قسمت دوم قضیه، مجدداً فرض کنید که $r = ۰/d_۱ d_۲ \ldots d_n d_{n+۱} \ldots$ نمایش قسمت کسری A باشد و این بار b را قطع شده‌ی A تا n رقم اعشار در نظر بگیرید. در این‌صورت:

$$e(b) = |A - b| = ۰/\underbrace{۰۰\ldots۰}_{n\text{ مرتبه}} d_{n+۱} \ldots \leq ۱۰ \times ۱۰^{-(n+۱)} = ۱۰^{-n}.$$

نمایش علمی اعداد

گاهی اندازه‌ی یک کمیت ممکن است با مقادیر خیلی بزرگ یا خیلی کوچک بیان شود. مثلاً فاصله‌ی زمین تا خورشید در حدود ۱۵۰,۰۰۰,۰۰۰,۰۰۰ متر (یکصد و پنجاه میلیارد متر) است، یا مثلاً برای نوشتن جرم یک الکترون برحسب گرم باید بعد از ممیز ۲۷ صفر قرار داد و پس از آن عدد ۹۱۰۹ را نوشت. بدیهی است که نوشتن چنین عددهایی با صفرهای زیاد، علاوه بر آن که خواندن و نوشتن را مشکل می‌کنند، احتمال اشتباه را هم زیاد می‌کنند. به همین دلیل از نمایشی موسوم به نمایش علمی برای اعداد استفاده می‌شود. فرض کنید A عددی مخالف صفر باشد. واضح است که A را همواره می‌توان به‌صورت $A = \pm a \times ۱۰^e$ نوشت که در آن e عددی صحیح است و

$$۱ \leq a < ۱۰.$$

نمایش عدد A به‌صورت فوق را نمایش علمی A می‌نامند. در این نمایش a مانتیس و b نمای عدد A نامیده می‌شود. بنابراین در نمایش علمی، حداقل مقدار مانتیس برابر با ۱ است[۱].

✍ **مثال ۱-۲۳:** نمایش علمی فاصله‌ی زمین تا خورشید، عبارت است از: $d = ۱/۵ \times ۱۰^{۱۱}(\mathrm{m})$.

✍ **مثال ۱-۲۴:** نمایش علمی جرم الکترون عبارت است از: $m = ۹/۱۰۹ \times ۱۰^{-۲۸}(\mathrm{gr})$.

✍ **مثال ۱-۲۵:** نمایش علمی عدد ۰/۰۰۰۰۳۴۱ عبارت است از: $۳/۴۱ \times ۱۰^{-۵}$.

✍ **مثال ۱-۲۶:** نمایش علمی عدد ۲۰۵۰ عبارت است از: $۲/۰۵ \times ۱۰^{۳}$.

✍ **مثال ۱-۲۷:** نمایش علمی عدد ۱۰۰۰۰۰۰۰ عبارت است از: ۱×۱۰^{۷}.

---

[۱] اگر به‌جای عدد ۱۰ مبنای β در نظر گرفته شود، آن‌گاه نمایش علمی A در مبنای β حاصل می‌شود. در این صورت e عددی صحیح در مبنای β است و محدوده‌ی مانتیس عبارت است از $۱ \leq a < \beta$.

ارقام بامعنا

تعریف ۱-۱۲: ارقام بامعنای یک عدد شامل ممیز عبارتند از اولین رقم غیر صفر سمت چپ و ارقام صفر و غیرصفر پس از آن. ارقام بامعنای یک عدد صحیح بسته به این‌که در نمایش علمی آن، مانتیس شامل چند رقم بامعنا باشد، برابر با تعداد ارقام بامعنای مانتیس تعریف می‌شود.

✍**مثال ۱-۲۸:** اعداد زیر همگی شامل ۵ رقم بامعنا هستند:

$$۱٫۵۰۳۵, \quad ۱٫۵۰۳۰, \quad ۰٫۰۱۵۰۳۵, \quad ۰٫۰۰۱۵۰۳۰, \quad ۱۵۰٫۳۰, \quad ۱۵۰۳٫۰.$$

با توجه به تعریف فوق، اگر یک عدد صحیح مانند A شامل هیچ صفری در سمت راست خود نباشد، آنگاه تعداد ارقام بامعنای آن دقیقاً با تعداد ارقام بامعنای مانتیس خود در نمایش علمی‌اش برابر است. اما چنانچه حداقل یک صفر، ارقام سمت راست A را تشکیل دهند، آنگاه این صفرها ممکن است بامعنا یا بی‌معنا باشند.

✍**مثال ۱-۲۹:** عدد $A = ۱۵۰۳۰۰۰$ را در نظر بگیرید. این عدد **دست‌کم** چهار رقم بامعنا دارد. اما بدون هیچ اطلاع دیگری نمی‌توان گفت تعداد ارقام بامعنا بیشتر هستند یا خیر.

اگر نمایش علمی A به‌صورت $۱٫۵۰۳۰ \times ۱۰^{۶}$ در نظر گرفته شود، آنگاه تعداد ارقام بامعنای A برابر با ۵ است و در این‌صورت دو صفر سمت راست A بی‌معنا هستند و اگر نمایش علمی A به‌صورت $۱٫۵۰۳۰۰۰۰ \times ۱۰^{۶}$ در نظر گرفته شود، آنگاه تعداد ارقام بامعنای A برابر با هفت است و تمام رقم‌های A بامعنا هستند.

توجه داشته باشید که همواره در قطع کردن یا گرد کردن یک عدد داده شده تا n رقم بامعنا $(n\mathrm{S})$[۱] عدد اصلی با عددی شامل n رقم بامعنا جایگزین می‌شود.

✍**مثال ۱-۳۰:** گردشده‌ی عدد $u = ۲۵٫۸$ تا دو رقم بامعنا عبارت است از: $u = ۲۶$ (۲S).

✍**مثال ۱-۳۱:** گردشده‌ی عدد $v = ۰٫۷۵$ تا یک رقم بامعنا عبارت است از: $v = ۰٫۸$ (۱S).

✍**مثال ۱-۳۲:** گردشده‌ی عدد $w = ۴٫۰۳۴$ تا سه رقم بامعنا عبارت است از: $w = ۴٫۰۳$ (۳S).

✍**مثال ۱-۳۳:** گردشده‌ی عدد $x = ۰٫۳۰۵$ تا دو رقم بامعنا عبارت است از: $x = ۰٫۳۰$ (۲S).

✍**مثال ۱-۳۴:** گردشده‌ی عدد $y = ۲۰۸$ را تا دو رقم بامعنا به‌دست آورید.

حل: این یک مثال آموزشی جالب است، زیرا نمی‌توانیم ۲۱ را به‌عنوان جواب بنویسیم. با این‌که ۲۱ دارای دو رقم بامعناست ولی تقریب مطلوب y نیست. برای این منظور ابتدا نمایش علمی y را می‌نویسیم:

$$y = ۲٫۰۸ \times ۱۰^{۲}.$$

حال مانتیس را تا دو رقم بامعنا گرد می‌کنیم. جواب $۲٫۱ \times ۱۰^{۲}$ است.

✍**مثال ۱-۳۵:** گردشده‌ی عدد $z = ۲۱۶۹۹۸۰$ را تا ۵ رقم بامعنا به‌دست آورید.

---

[۱] S حرف اول واژه‌ی Significant به معنای «بامعنا» است. از این رو نمایش یک عدد تا n رقم بامعنا با نماد $(n\mathrm{S})$ بیان می‌شود.

حل: ابتدا نمایش علمی z را می‌نویسیم:

$$z = ۲/۱۶۹۹۸ \times ۱۰^{۶}.$$

حال مانتیس را تا پنج رقم بامعنا گرد می‌کنیم. بنابراین جواب برابر است با $۲/۱۷۰۰۰ \times ۱۰^{۶}$.

ارقام بامعنای درست یک تقریب

همان‌گونه که قبلاً گفته شد یکی از راه‌های سنجش دقت یک تقریب، تعداد ارقام بامعنای درست آن است. با توجه به این معیار، هرچه تعداد ارقام بامعنای درست یک تقریب بیشتر باشد، آن تقریب دقیق‌تر است. در این‌جا باید توجه داشت که دقت یک تقریب هیچ ارتباط معنی‌داری با تعداد ارقام بامعنای آن ندارد. مثلاً عدد $x = ۳/۷۱۸۲$ تقریبی از عدد e است که ۵ رقم بامعنا دارد، اما تقریب خوبی از e نیست و مثلاً عدد ۳ (تنها با یک رقم بامعنا) تقریب بهتری است. بنابراین برای آن‌که بهتر بتوانیم از روی ارقام یک تقریب به دقت آن پی ببریم، مفهوم تعداد ارقام بامعنای درست یک تقریب را تعریف می‌کنیم. این مفهوم را می‌توان به‌طور نادقیق این‌گونه تعریف کرد که تقریب a از A دارای n رقم بامعنای درست است اگر a و A هر دو به یک عدد سوم با n رقم بامعنا گرد شوند. این تعریف معمولاً کارساز و از نظر شهودی درست است، اما مثال زیر را در نظر بگیرید:

$$A = ۰/۹۹۴۹, \qquad a = ۰/۹۹۵۱,$$

با توجه به تعریف بالا a نمی‌تواند دو رقم با معنای درست داشته باشد. زیرا اگر هر دوی a و A را تا دو رقم با معنا گرد کنیم، A به عدد ۰/۹۹ گرد می‌شود در حالی که a به عدد ۱/۰ گرد می‌شود. اما در همین مثال، a دارای یک رقم با معنای درست و نیز سه رقم بامعنای درست است! تعریف دقیق زیر، این ابهام را برطرف می‌کند.

تعریف ۱-۱۳: اگر $a \neq ۰$ تقریبی از A با نمایش علمی $a = a_m/a_{m-۱}a_{m-۲} \ldots \times ۱۰^{m}$ و d تعداد ارقام بامعنای a باشد، آنگاه بزرگ‌ترین عدد صحیح نامنفی n که $n \leq d$ و

$$|A - a| \leq ۵ \times ۱۰^{m-n}$$

تعداد ارقام بامعنای درست a نامیده می‌شود.

✍**مثال ۱-۳۶:** اگر $A = ۰/۹۹۴۹$ آنگاه تعداد ارقام بامعنای درست $a = ۰/۹۹۵۱$ به‌عنوان تقریبی از A چقدر است؟

حل: ابتدا نمایش علمی a را برای تعیین m می‌نویسیم. با توجه به تعریف، m برابر با توان ۱۰ در نمایش علمی a است:

$$a = ۹/۹۵۱ \times ۱۰^{-۱} \Longrightarrow m = -۱$$

حال باید بزرگترین nی را بیابیم که $n \leq ۴$ و $|A - a| = ۲ \times ۱۰^{-۴} \leq ۵ \times ۱۰^{-۱-n}$. داریم:

$$۲ \times ۱۰^{-۴} \leq ۵ \times ۱۰^{-۱-n} \Longrightarrow ۲ \leq ۵ \times ۱۰^{۴-۱-n} \Longrightarrow n_{max} = ۳,$$

لذا تعداد ارقام بامعنای درست a برابر است با ۳.

✍**مثال ۱-۳۷:** ثابت اویلر، ... ۰٫۵۷۷۲۱۵۶۶۴ = γ را در نظر بگیرید[۱]. فرض کنید که به شما اعداد

$$a = ۰٫۵۷۷۲۱۵۴۹۴, \qquad a' = ۰٫۵۷۷۲۱۵۸۳۴$$

برای تقریب γ پیشنهاد می‌شود. تقریب مناسب‌تر را انتخاب کنید.

حل: واضح است که در این‌جا نمی‌توان خطای مطلق یا نسبی یا ترکیبی را حساب کرد؛ زیرا مقدار دقیق γ مجهول است. هم‌چنین توجه دارید که عددهای داده شده تا ۶ رقم بامعنا با γ یکسان هستند و نیز اختلاف رقم‌های هفتم به بعد آن‌ها یعنی ۴۹۴ و ۸۳۴ با ارقام متناظرشان در γ یعنی ۶۶۴ برابرند. تا این‌جا a و a' وضعیت یکسانی دارند. حال بیایید γ، a و a' را تا شش رقم بامعنا، یعنی شماره‌ی آخرین رقم بامعنای قطعی در دو تقریب داده شده، گرد کنیم:

$$\gamma = ۰٫۵۷۷۲۱۶ \ (۶\mathrm{S}),$$
$$a = ۰٫۵۷۷۲۱۵ \ (۶\mathrm{S}),$$
$$a' = ۰٫۵۷۷۲۱۶ \ (۶\mathrm{S}).$$

ملاحظه می‌شود که گرد شده‌ی a' و γ تا شش رقم بامعنا با هم یکسان شدند ولی برای a چنین اتفاقی نیفتاد. این یعنی تعداد ارقام بامعنای درست a' بیشتر از تعداد ارقام بامعنای درست a است. در نتیجه بهتر است a' را انتخاب کنید. این انتخاب از نظر ریاضی هم کاملاً درست است، زیرا سرانجام پس از ارقام ۶۶۴ در نمایش γ رقم مخالف صفری یافت می‌شود (چون اگر هرگز رقم مخالف صفری یافت نشود آن‌گاه نتیجه می‌شود که γ گویاست!). وجود همان رقم مخالف صفر، نزدیکی بیشتر a' به γ را ثابت می‌کند.

✍**مثال ۱-۳۸:** فرض کنید $A = ۸٫۰۰۰$، $a = ۷٫۹۹۷$ و $a' = ۸٫۰۸$. تعداد ارقام بامعنای درست a و a' را تعیین کنید.

حل: همان‌طور که مشاهده می‌شود که a' دو رقم بامعنای مساوی با ارقام A دارد (با حفظ ارزش هر رقم) اما هیچ‌یک از رقم‌های a مساوی با ارقام A نیست. آیا می‌توان گفت که ارقام درست a' بیشتر از ارقام درست a است؟ خواهیم دید که خیر. در این‌جا $e(a) = ۰٫۰۰۳$ و $e(a') = ۰٫۰۸$ و در واقع a باید تعداد ارقام درست بیشتری داشته باشد! اگر a را تا سه رقم بامعنا گرد کنیم، عدد A حاصل می‌شود. از این رو، a سه رقم بامعنای درست دارد. اگر a' را تا رقم یکان گرد کنیم، A حاصل می‌شود (توجه کنید که حتی گرد شده‌ی a' تا یک رقم اعشار، به عدد ۸٫۱ منجر می‌شود که مساوی با A نیست). یعنی a' تنها یک رقم با معنای درست دارد. اکنون تعداد ارقام بامعنای درست a و a' را با استفاده از تعریف نیز محاسبه می‌کنیم. چون $a = ۷٫۹۹۷ \times ۱۰^{۰}$ و $a' = ۸٫۰۸ \times ۱۰^{۰}$ بنابراین مقدار m برای هر دو صفر است. برای a داریم $|A - a| = e(a) = ۰٫۰۰۳$، بنابراین باید بزرگ‌ترین nی را بیابیم که در نامساوی زیر صدق کند:

---

[۱] این عدد در واقع برابر است با $\lim_{n\to\infty}(۱ + \frac{۱}{۲} + \cdots + \frac{۱}{n} - \ln n)$ و تاکنون گنگ یا گویا بودن آن معلوم نشده است.

$$۰/۰۰۳ \leq ۵ \times ۱۰^{-n}$$

بدیهی است که بزرگ‌ترین n برابر است با ۳. یعنی، a دارای ۳ رقم بامعنای درست است. برای a' نیز داریم:

$$e(a') = ۰/۰۸ = ۸ \times ۱۰^{-۲} < ۵ \times ۱۰^{۰-۱}$$

یعنی a' تنها یک رقم بامعنای درست دارد.

مثال ۱-۳۹: اگر $A = ۱۰۰$، $a = ۹۹/۹۸$ و $b = ۱۰۰/۶$. تعداد ارقام بامعنای درست a و b را تعیین کنید.

حل: در مورد a داریم:

$$a = ۹/۹۹۸ \times ۱۰^{۱} \Longrightarrow m = ۱,$$
$$e(a) = |A - a| = \left|۱۰۰ - ۹۹/۹۸\right| = ۰/۰۲.$$

بنابراین باید بزرگ‌ترین nی را بیابیم که $۰/۰۲ < ۵ \times ۱۰^{۱-n}$.

بدیهی است که $n = ۳$ جواب است. یعنی، a دارای ۳ رقم بامعنای درست است. در مورد b داریم:

$$b = ۱/۰۰۶ \times ۱۰^{۲} \Longrightarrow m = ۲,$$
$$e(b) = |A - b| = \left|۱۰۰ - ۱۰۰/۶\right| = ۰/۶.$$

بنابراین باید بزرگ‌ترین nی را بیابیم که $۰/۶ < ۵ \times ۱۰^{۲-n}$. در این‌جا $n = ۲$ جواب است. یعنی، b تنها ۲ رقم بامعنای درست دارد.

قضیه ۱-۸: اگر a تقریبی از A با n رقم بامعنای درست باشد و $b = ۱۰^k \times a$ و $B = ۱۰^k \times A$، که در آن k عددی صحیح است، آنگاه b تقریبی از B با n رقم بامعنای درست است و خطای نسبی a و b یکسان هستند.

اثبات: فرض کنید $a = a_m/a_{m-۱}a_{m-۲} \ldots \times ۱۰^m$ نمایش علمی a باشد. در این‌صورت

$$b = ۱۰^k \times a = a_m/a_{m-۱}a_{m-۲} \ldots \times ۱۰^{m+k},$$

یعنی نمای مربوط به b در نمایش علمی برابر با $m + k$ است. چون a دارای n رقم بامعنای درست است پس n بزرگ‌ترین عدد صحیحی است که در رابطه‌ی $|A - a| \leq ۵ \times ۱۰^{m-n}$ صدق می‌کند. بنابراین:

$$|B - b| = ۱۰^k|A - a| \leq ۵ \times ۱۰^{(m+k)-n},$$

و n بزرگ‌ترین عدد صحیحی است که در این نامساوی صدق می‌کند. از این‌رو b نیز دارای n رقم بامعنای درست است. قسمت دوم حکم به‌سادگی از تعریف خطای نسبی b نتیجه می‌شود.

قضیه ۱-۹: اگر a گرد شده‌ی عدد مثبت A تا n رقم بامعنا باشد، آنگاه a دارای n رقم بامعنای درست است.

اثبات: با توجه به قضیه‌ی قبل می‌توانیم فرض کنیم که:

$$a = ۰/c_۱c_۲c_۳ \ldots c_n, \qquad c_۱ \neq ۰$$

(اگر a چنین نباشد، با انتخاب عدد صحیح و مناسب k می‌توان $۱۰^k \times a$ را به این شکل در آورد). بنابراین در

مورد این a داریم $m = -۱$.

از طرف دیگر می‌دانیم که اگر a گرد شده‌ی A تا n رقم اعشار باشد، آنگاه $|A - a| \leq ۵ \times ۱۰^{-(n+۱)}$. اکنون با استفاده از این موضوع داریم:

$$|A - a| \leq ۵ \times ۱۰^{-(n+۱)} = ۵ \times ۱۰^{-n-۱},$$

که نتیجه می‌دهد a دارای n رقم بامعنای درست است. (توجه کنید که ممکن است داشته باشیم

$$|A - a| \leq ۵ \times ۱۰^{-۱-n'}$$

که در آن $n' > n$. اما چون a بیش از n رقم بامعنا ندارد، کمترین n و n' همان n می‌شود.)

قضیه ۱-۱۰: اگر $a > ۰$ تقریبی از A و دارای n رقم بامعنای درست باشد، آنگاه خطای نسبی a از ۵×۱۰^{-n} کمتر است، به شرط آنکه رقم‌های درست a شامل یک رقم یک و $n - ۱$ رقم صفر در سمت راست آن نباشد.

اثبات: با توجه به قضایای قبل می‌توان فرض کرد که:

$$a = b_n \dots b_۲ b_۱ / c_۱ c_۲ c_۳ \dots = b_n / b_{n-۱} \dots b_۲ b_۱ c_۱ c_۲ c_۳ \dots \times ۱۰^{n-۱}$$

که در آن $b_n \dots b_۲ b_۱$ عدد حاصل از n رقم بامعنای درست است. بنابراین نمای مربوط به این a برابر با $n - ۱$ است و با توجه به تعریف تعداد ارقام بامعنای درست هر تقریب داریم:

$$|A - a| \leq ۵ \times ۱۰^{(n-۱)-n} = ۰/۵. \qquad (۱)$$

با توجه به این‌که $|a| - |A| \leq |A - a|$، از نامساوی فوق نتیجه می‌شود:

$$|A| \geq |a| - ۰/۵ = a - ۰/۵,$$

اما بنابر فرض قضیه $b_n \dots b_۲ b_۱ \neq ۱۰^{n-۱}$، پس $a > b_n \dots b_۲ b_۱$. لذا $a \geq ۱۰^{n-۱} + ۱$ و در نتیجه:

$$|A| \geq a - ۰/۵ \geq ۱۰^{n-۱} + ۱ - ۰/۵ = ۱۰^{n-۱} + ۰/۵ > ۱۰^{n-۱}. \qquad (۲)$$

$$(۱), (۲) \Longrightarrow \delta(a) = \frac{|A - a|}{|A|} < \frac{۰/۵}{۱۰^{n-۱}} = ۵ \times ۱۰^{-n}.$$

در عمل به شرط انتهای این قضیه توجه نمی‌شود، زیرا به‌ندرت ممکن است تقریبی دارای این ویژگی باشد.

نتیجه: احکام دو قضیه‌ی قبل را می‌توان به صورت زیر خلاصه کرد:

$a > ۰$ گردشده‌ی A با n رقم بامعنا $\Longrightarrow \delta(a) < ۵ \times ۱۰^{-n}$ دارای n رقم بامعنای درست a.

✍ **مثال ۱-۴۰:** تقریبی از عدد π ارائه دهید که خطای نسبی آن کمتر از $۱۰^{-۳}$ باشد.

حل: با توجه به نامساوی $۱۰^{-۳} < ۵ \times ۱۰^{-۴}$، تقریب a از π را چنان ارائه می‌دهیم که دارای چهار رقم بامعنای درست باشد. چرا که در آن‌صورت، بنابر نتیجه‌ی قبل $\delta(a) < ۵ \times ۱۰^{-۴}$ و بنابراین $\delta(a) < ۱۰^{-۳}$. برای این منظور کافی است a را گردشده‌ی π تا چهار رقم بامعنا اختیار کنیم، که از آن‌جا $a = ۳/۱۴۲$ به دست می‌آید.

✍ **مثال ۱-۴۱:** تقریبی از عدد $\sqrt{۱۰}$ ارائه دهید که خطای نسبی آن کمتر از $۱۰^{-۴}$ باشد.

حل:

$$۵ \times ۱۰^{-n} < ۱۰^{-۴} \Longrightarrow ۵ < ۱۰^{n-۴} \Longrightarrow n_{min} = ۵$$

بنابراین کافی است تقریبی از $\sqrt{۱۰}$ با پنج رقم بامعنای درست ارائه دهیم. از این‌رو گردشده‌ی $\sqrt{۱۰}$ را تا پنج رقم بامعنا می‌نویسیم:

$$\sqrt{۱۰} = ۳/۱۶۲۲۷۷۶۶۰۱۶ \ldots = ۳/۱۶۲۳ \ (۵\mathrm{S}).$$

قضیه ۱-۱۱: اگر $a > ۰$ تقریبی از A باشد به‌طوری که $\delta(a) \leq ۰/۵ \times ۱۰^{-n}$، آن‌گاه a حداقل n رقم بامعنای درست دارد.

اثبات: فرض کنید $a = a_m/a_{m-۱}a_{m-۲} \ldots \times ۱۰^m$ نمایش علمی a باشد. با توجه به این‌که $a_i \leq ۹$ داریم:

$$a < ۹/۹۹ \ldots \times ۱۰^m = ۱۰ \times ۱۰^m = ۱۰^{m+۱},$$

از طرفی می‌توان نوشت $\delta(a) \approx \frac{e(a)}{a}$ که این به همراه فرض قضیه نتیجه می‌دهد:

$$e(a) \approx a \times \delta(a) \leq ۱۰^{m+۱} \times ۰/۵ \times ۱۰^{-n} = ۵ \times ۱۰^{m-n}.$$

چون ممکن است عددی بزرگ‌تر از n نیز در نامساوی بالا صدق کند، نتیجه می‌گیریم که a حداقل n رقم بامعنای درست دارد.

در برخی از روش‌های عددی (مانند حل دستگاه‌های معادلات خطی) می‌توان خطای نسبی یک تقریب و در برخی دیگر از روش‌ها (مانند حل معادلات غیرخطی به روش نیوتن) می‌توان تعداد ارقام بامعنای درست یک تقریب را به‌دست آورد. در این موارد، کاربرد دو قضیه‌ی اخیر بدین صورت است که با داشتن یکی از دو اطلاع خطای نسبی یا تعداد ارقام بامعنای درست یک تقریب، می‌توان دیگری را نیز برآورد نمود.

روند کردن ماشینی:

فرآیند محاسبات و اعلام نتیجه‌ی محاسبه توسط ماشین، به کمک عددهای ماشینی صورت می‌گیرد. اما محدودیت در تعداد ارقام مانتیس سبب می‌شود که مجموعه‌ی اعداد ماشینی تنها زیرمجموعه‌ای متناهی از اعداد گویای مختوم متعلق به بازه‌ی $\Omega = [-x_{max}, x_{max}]$ باشند. بنابراین در ماشین‌های محاسباتی امکان نگهداری و نمایش بسیاری از عددهای حقیقی وجود ندارد. از این رو، عددهای غیر ماشینی با اعداد ماشینی تقریب زده می‌شوند و فرآیند محاسبات بر روی تقریب آن‌ها صورت می‌گیرد.

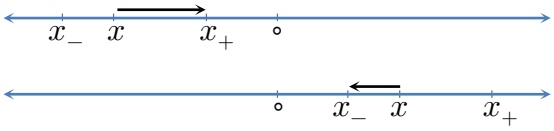
فرض کنید که حالت‌های سرریز یا زیرریز رخ ندهند و $x \in \Omega - A$ که A مجموعه‌ی اعداد ماشینی است. در این‌صورت می‌توان با تعریف نگاشت نمایش‌ساز $\mathrm{fl}: \Omega \longrightarrow A$ عدد x را با یک عدد ماشینی تقریب زد و به اصطلاح x را قابل نمایش نمود.[۱] عمل نگاشت fl را **روند کردن** می‌نامیم.

[۱] توجه کنید که دامنه‌ی نگاشت fl برابر با Ω است، یعنی نگاشت fl برای اعداد ماشینی نیز تعریف می‌شود. در حقیقت $x \in A \Longleftrightarrow \mathrm{fl}(x) = x$.

ضابطه‌ی نگاشت fl به شکل‌های مختلفی تعریف می‌شود، از جمله روند کردن به سمت صفر (بریدن یا قطع کردن)، روند کردن به نزدیک‌ترین عدد ماشینی (گرد کردن)، روند کردن به بالا (یا به سمت $+\infty$) و روند کردن به پایین (یا به سمت $-\infty$) که در ادامه نحوه‌ی تعریف هر یک از آن‌ها را بیان می‌کنیم. فرض کنید $x \in \Omega - A$، و x_+ و x_- به‌ترتیب نشان‌دهنده‌ی نزدیک‌ترین اعداد ماشینی بزرگ‌تر و کوچک‌تر از x باشند. در این‌صورت:

۱ــ در روند کردن به سمت صفر (قطع کردن)، x با اولین عددی که بر سر راه حرکت آن به سمت صفر قرار دارد تقریب زده می‌شود. به عبارت دیگر، اگر fl_c بیانگر این نگاشت باشد داریم:

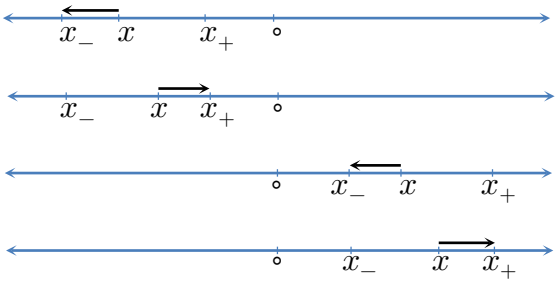
$$\mathrm{fl}_c(x) = \begin{cases} x_+; & x < 0, \\ x_-; & x > 0. \end{cases}$$



۲ــ در گرد کردن، x با نزدیک‌ترین عدد ماشینی مجاور خود تقریب زده می‌شود. اگر این نگاشت را با fl_r نشان دهیم، داریم:

$$\mathrm{fl}_r(x) = \begin{cases} x_-; & x < x_*, \\ x_+; & x > x_*, \\ x_+ \text{ or } x_-; & x = x_*. \end{cases}$$

که در آن $x_* = (x_- + x_+)/2$. همان‌طور که ملاحظه می‌کنید در این روش چنانچه $x = x_*$، یعنی x دقیقاً بین x_+ و x_- واقع شده باشد، می‌توان به دلخواه گرد کردن را به بالا یا به پایین انجام داد. اما معمولاً در این حالت از قرارداد «گرد کردن به زوج» استفاده می‌شود. به این ترتیب که هر یک از اعداد x_+ و x_- که آخرین رقم سمت راست مانتیس آن زوج باشد، به عنوان تقریب x انتخاب می‌شود. می‌توان دید که با این قرارداد، نیمی از اعداد غیر ماشینی به بالا و نیم دیگر به پایین گرد می‌شوند. به همین دلیل، این قرارداد نتیجه‌ای مانند قرارداد گاوس در گردکردن عادی دارد. توجه کنید که معمولاً $\mathrm{fl}_r(x)$ را می‌توان مانند گردکردن معمولی، با توجه به ارقام ناخواسته در مانتیس (اگر مانتیس m رقمی باشد، رقم $m+1$ام به بعد)، به‌سادگی تعیین کرد و محاسبه‌ی صریح x_* لازم نیست.



۳ــ در روند کردن به بالا (یا به سمت $+\infty$)، x با اولین عددی که بر سر راه حرکت آن به سمت $+\infty$ قرار دارد تقریب زده می‌شود (یعنی در این حالت، در شکل اخیر جهت تمام پیکان‌ها به سمت راست است). ضابطه‌ی این نگاشت به صورت $\mathrm{fl}_u(x) = x_+$ است.

۴ــ در روند کردن به پایین (یا به سمت $-\infty$)، x با اولین عددی که بر سر راه حرکت آن به سمت $-\infty$ قرار

دارد تقریب زده می‌شود (یعنی در این حالت، در شکل اخیر جهت تمام پیکان‌ها به سمت چپ است). ضابطه‌ی این نگاشت به صورت $\mathrm{fl}_d(x) = x_-$ است.

✍**مثال ۱-۴۲:** سیستم نمایش ممیز شناور نرمال $\mathbb{F}(۲,۳,-۱,۲)$ را در نظر بگیرید. در این صورت:

الف) اگر $w = ۰/۱۱۰۱۰۱ \times ۲^{۱}$، آن‌گاه:

$$\mathrm{fl}_c(w) = ۰/۱۱۰ \times ۲^{۱}, \qquad \mathrm{fl}_r(w) = ۰/۱۱۱ \times ۲^{۱},$$
$$\mathrm{fl}_u(w) = ۰/۱۱۱ \times ۲^{۱}, \qquad \mathrm{fl}_d(w) = ۰/۱۱۰ \times ۲^{۱}.$$

ب) اگر $x = ۰/۱۰۱۰۱ \times ۲^{-۱}$، آن‌گاه:

$$\mathrm{fl}_c(x) = ۰/۱۰۱ \times ۲^{-۱}, \qquad \mathrm{fl}_r(x) = ۰/۱۰۱ \times ۲^{-۱},$$
$$\mathrm{fl}_u(x) = ۰/۱۱۰ \times ۲^{-۱}, \qquad \mathrm{fl}_d(x) = ۰/۱۰۱ \times ۲^{-۱}.$$

ج) اگر $y = -۰/۱۱۰۱ \times ۲^{۱}$، آن‌گاه:

$$\mathrm{fl}_c(y) = -۰/۱۱۰ \times ۲^{۱}, \qquad \mathrm{fl}_r(y) = -۰/۱۱۰ \times ۲^{۱},$$
$$\mathrm{fl}_u(y) = -۰/۱۱۰ \times ۲^{۱}, \qquad \mathrm{fl}_d(y) = -۰/۱۱۱ \times ۲^{۱}.$$

توجه کنید که در این‌جا، $y_+ = -۰/۱۱۰ \times ۲^{۱}$، $y_- = -۰/۱۱۱ \times ۲^{۱}$. بنابراین:

$$\begin{aligned} y_* &= -\left[۰/۱۱۰ \times ۲^{۱} + \frac{۱}{۲}(۰/۰۰۱ \times ۲^{۱})\right] \\ &= -\left[۰/۱۱۰ \times ۲^{۱} + ۰/۰۰۱ \times ۲^{۰}\right] \\ &= -\left[۰/۱۱۰ \times ۲^{۱} + ۰/۰۰۰۱ \times ۲^{۱}\right] \\ &= -۰/۱۱۰۱ \times ۲^{۱}. \end{aligned}$$

لذا $y = y_*$ و در نتیجه برای گرد کردن از قرارداد «گرد کردن به زوج» استفاده می‌کنیم.

د) اگر $z = ۰/۱۱۱۱۰۱۱ \times ۲^{۱}$، آن‌گاه:

$$\mathrm{fl}_c(z) = ۰/۱۱۱ \times ۲^{۱}, \qquad \mathrm{fl}_r(z) = ۰/۱۰۰ \times ۲^{۲},$$
$$\mathrm{fl}_u(z) = ۰/۱۰۰ \times ۲^{۲}, \qquad \mathrm{fl}_d(z) = ۰/۱۱۱ \times ۲^{۱}.$$

با توجه به مثال فوق و شکل‌های قبل به‌سادگی می‌توان دید که روش گرد کردن برای تقریب اعداد غیرماشینی، نسبت به سایر روش‌ها از دقت بالاتری برخوردار است (از این رو در ماشین‌های محاسباتی، به‌صورت پیش‌فرض از روش گرد کردن استفاده می‌شود). در ادامه، ضمن بررسی خطاهای مطلق و نسبی هر یک از روش‌ها، این مطلب را ثابت می‌کنیم. سیستم ممیز شناور نرمال زیر را برای نمایش اعداد در نظر بگیرید:

$$\boxed{\mathbb{F} = \left\{\pm(۰/b_۱ b_۲ \dots b_m)_\beta \times \beta^e \mid b_i \in \{۰,۱,\dots,\beta-۱\}, b_۱ \neq ۰, L \le e \le U\right\} \cup \{۰\}}$$

اگر x عددی حقیقی و غیر ماشینی با نمایش ممیز شناور نرمال $x = \pm(۰/b_۱ b_۲ \dots b_m b_{m+۱} \dots)_\beta \times \beta^e$ باشد،

آن‌گاه خطای مطلق بریدن آن برابر با قدرمطلق اختلاف x و $\mathrm{fl}_c(x)$ و خطای مطلق گرد کردن آن برابر با قدرمطلق اختلاف x و $\mathrm{fl}_r(x)$ است. بنابراین:

$$|x-\mathrm{fl}_c(x)| = (۰/\underbrace{۰۰\ldots۰}_{m \text{ مرتبه}} b_{m+۱}\ldots)_\beta \times \beta^e \le (۰/\underbrace{۰۰\ldots۰}_{m-۱ \text{ مرتبه}} ۱)_\beta \times \beta^e = \beta^{e-m},$$

$$|x-\mathrm{fl}_r(x)| \le \frac{۱}{۲}(x_+ - x_-) = \frac{۱}{۲}\beta^{e-m}.$$

اکنون با توجه به این‌که $|x| \ge (۰/۱\underbrace{۰۰\ldots۰}_{m-۱ \text{ مرتبه}})_\beta \times \beta^e = \beta^{e-۱}$، کران‌های بالای زیر را برای خطای نسبی $\mathrm{fl}_c(x)$ و $\mathrm{fl}_r(x)$ خواهیم داشت:

$$\frac{|x-\mathrm{fl}_c(x)|}{|x|} \le \frac{\beta^{e-m}}{\beta^{e-۱}} = \beta^{۱-m},$$
$$\frac{|x-\mathrm{fl}_r(x)|}{|x|} \le \frac{۱}{۲}\frac{\beta^{e-m}}{\beta^{e-۱}} = \frac{۱}{۲}\beta^{۱-m}.$$

همان‌طور که ملاحظه می‌کنید، کران بالای خطای نسبی به اندازه‌ی عدد (نما) بستگی ندارد و تنها به ارقام مانتیس وابسته است. بنابراین در سیستم ممیز شناور، اگرچه فاصله‌ی دو عدد متوالی بزرگ، زیاد است اما تأثیری در خطای نسبی ندارد. این یک مزیت برای اعداد ممیز شناور به شمار می‌رود.

در مورد خطاهای مطلق و نسبی روند کردن به بالا و پایین توجه کنید که اگر x عددی منفی (مثبت) باشد، آن‌گاه روند شده‌ی x به بالا (پایین) و بریده شده‌ی آن هر دو یکسان هستند. بنابراین می‌توان از تقارن اعداد ممیز شناور نسبت به مبدأ مختصات نتیجه گرفت که کران‌هایی که برای خطای مطلق و نسبی در حالت بریدن به دست آمدند، برای گرد کردن به سمت بالا و پایین نیز برقرارند. یعنی:

$$|x-\mathrm{fl}_u(x)| \le \beta^{e-m}, \qquad \frac{|x-\mathrm{fl}_u(x)|}{|x|} \le \beta^{۱-m},$$
$$|x-\mathrm{fl}_d(x)| \le \beta^{e-m}, \qquad \frac{|x-\mathrm{fl}_d(x)|}{|x|} \le \beta^{۱-m}.$$

اپسیلون ماشین و روند واحد

تعریف ۱-۱۴: به فاصله‌ی بین عدد یک و عدد ممیز شناور بلافاصله پس از آن در سیستم نمایش $\mathbb{F}(\beta, m, L, U)$ اپسیلون ماشین می‌گوییم و آن را با نماد ϵ_M یا eps نشان می‌دهیم. با توجه به این‌که عدد یک در این سیستم به‌صورت $(۰/۱\underbrace{۰۰\ldots۰۰}_{m-۱ \text{ مرتبه}})_\beta \times \beta^۱ = ۱$ نمایش داده می‌شود، این فاصله عبارت است از:

$$\epsilon_M = ۱_+ - ۱ = (۰/۱\underbrace{۰۰\ldots۰}_{m-۲ \text{ مرتبه}}۱)_\beta \times \beta^۱ - (۰/۱\underbrace{۰۰\ldots۰۰}_{m-۱ \text{ مرتبه}})_\beta \times \beta^۱ = \beta^{۱-m}.$$

همان‌طور که ملاحظه می‌کنید اپسیلون ماشین تنها به مبنا و تعداد ارقام مانتیس بستگی دارد.

اکنون فرض کنید x عددی مخالف صفر و متعلق به Ω باشد و نگاشت نمایش‌ساز fl را در نظر بگیرید. قرار می‌دهیم:

$$\delta = \frac{\mathrm{fl}(x) - x}{x},$$

در این صورت $\mathrm{fl}(x) = x(1+\delta)$ و $|\delta| \leq u$ که در آن:

$$u = \begin{cases} \beta^{1-m} = \epsilon_M; & \text{اگر } \mathrm{fl}(x) = \mathrm{fl}_c(x) \\ \frac{1}{2}\beta^{1-m} = \frac{\epsilon_M}{2}; & \text{اگر } \mathrm{fl}(x) = \mathrm{fl}_r(x) \end{cases}$$

عدد u در این‌جا، روند واحد نامیده می‌شود و علاوه بر مبنا و تعداد ارقام مانتیس به نوع نگاشت نمایش‌ساز بستگی دارد. بنابراین، اگر u روند واحد وابسته به نگاشت نمایش‌ساز fl باشد، به ازای هر $0 \neq x \in \Omega$ می‌توان نوشت:

$$\boxed{\mathrm{fl}(x) = x(1+\delta); \qquad |\delta| \leq u}$$

تعریف $\mathrm{fl}(x)$ بدین صورت و استفاده از آن، نخستین بار توسط ویلکینسون* انجام شده‌است. این رابطه هم‌چنین در تحلیل خطای تولید شده توسط اعمال حسابی در کامپیوتر بسیار اهمیت دارد. نکته‌ی مهمی که از این رابطه استنتاج می‌شود این است که هرچه ϵ_M کوچک‌تر باشد، u و در نتیجه خطای نسبی روند کردن کمتر است. یعنی کوچک بودن اندازه‌ی اپسیلون ماشین به‌گونه‌ای بیانگر دقت آن ماشین است و چون اپسیلون ماشین تنها به دو عامل مبنا و تعداد ارقام مانتیس بستگی دارد، با افزایش تعداد ارقام مانتیس می‌توان دقت ماشین را بالا برد. به همین دلیل است که وقتی دقت بالاتری در محاسبات مد نظر است، از **دقت مضاعف** استفاده می‌شود. در دقت مضاعف، برای نمایش یک عدد ممیز شناور از دو کلمه‌ی کامپیوتر استفاده می‌شود و در این حالت مانتیس حداقل دو برابر بیت نسبت به **دقت معمولی** در اختیار دارد. برخی از کامپیوترها قادرند اعداد را حتی تا ۴ برابر دقت معمولی ذخیره کنند.

نکته‌ی دیگر این‌که با توجه به تعریف $\mathrm{fl}(x)$ و با قرار دادن $x = 1$ داریم:

$$1 = \mathrm{fl}(1) = 1 + \delta; \qquad |\delta| \leq u,$$

بنابراین:

$$\mathrm{fl}(1+\delta) = \mathrm{fl}\big(\mathrm{fl}(1)\big) = \mathrm{fl}(1) = 1; \qquad |\delta| \leq u,$$

یعنی در یک سیستم نمایش ممیز شناور نرمال به‌همراه نگاشت نمایش‌ساز fl، روند واحد بزرگ‌ترین مقداری است که اگر با عدد ۱ جمع شود، هم‌چنان عدد ۱ به عنوان حاصل‌جمع نمایش داده می‌شود. به‌عبارت دیگر:

$$\boxed{u = \sup\{\delta | \mathrm{fl}(1+\delta) = 1\}}$$

نتیجه: اگر $|\delta| \leq u$، آن‌گاه نمایش $1+\delta$ برابر با ۱ است. هم‌چنین با توجه به این‌که

$$x + y = x(1 + \frac{y}{x})$$

اگر داشته باشیم $|\frac{y}{x}| \leq u$ آن‌گاه نمایش $x+y$ با نمایش x یکسان خواهد بود.

محاسبه‌ی تقریبی روند واحد با استفاده از برنامه‌ی زیر امکان‌پذیر است. این برنامه کوچک‌ترین عدد ممیز شناور x را محاسبه می‌کند که به ازای آن $\text{fl}(1+x) > 1$. هم‌چنین از این برنامه برای تقریب eps ماشین می‌توان استفاده کرد. توجه کنید که eps ماشین و روند واحد دو مفهوم مجزا هستند. با این وجود، اختلاف این مقادیر در بسیاری از کامپیوترها به‌قدری کوچک است که معمولاً تمایزی بین آن‌ها گذارده نمی‌شود.

```
x := 1;
while 1 + x > 1
        x := x/2;
end
```

۱-۶ تولید و انتشار خطا

همان‌طور که پیش‌تر گفته شد در ماشین‌های محاسباتی عددهای غیر ماشینی با اعداد ماشینی تقریب زده می‌شوند و فرآیند محاسبات بر روی تقریب آن‌ها صورت می‌گیرد. در این بخش می‌خواهیم این مطلب را به‌طور دقیق‌تر بررسی کنیم و ببینیم که خطاهای موجود در مراحل یک فرآیند محاسباتی چگونه به‌وجود می‌آیند و آیا می‌توان برای کاهش آن‌ها یا حداقل جلوگیری از افزایش آن‌ها در مراحل بعدی محاسبات کاری انجام داد یا خیر؟ برای این منظور ابتدا به‌طور مختصر نحوه‌ی انجام محاسبات بر روی اعداد ممیز شناور را تشریح می‌کنیم و سپس به بررسی خطای موجود در حاصل یک محاسبه‌ی عددی می‌پردازیم.

حساب ممیز شناور

انجام محاسبات بر روی اعداد ممیز شناور را حساب ممیز شناور می‌نامند. ماشینی را در نظر بگیرید که از سیستم ممیز شناور نرمال $\mathbb{F}(\beta, m, L, U)$ استفاده می‌کند. هم‌چنین فرض کنید $*$ بیانگر یکی از چهار عمل حسابی اصلی باشد، یعنی $* \in \{+, -, \times, \div\}$. حال اگر X و Y دو عدد حقیقی باشند و بخواهیم $X * Y$ را به‌کمک این ماشین محاسبه کنیم، آن‌گاه مراحل زیر توسط ماشین انجام می‌شود:

۱ـ ممکن است X یا Y ماشینی نباشند، بنابراین در نخستین مرحله آن‌ها با $x = \text{fl}(X)$ و $y = \text{fl}(Y)$ جایگزین می‌شوند. لذا محاسبه‌ی $X * Y$ به $x * y$ تغییر می‌یابد.

۲ـ $x * y$ با استفاده از حساب ممیز شناور و با دقت مضاعف محاسبه می‌شود (در عمل، بسیاری از کامپیوترها اعمال حسابی را در ثبات‌های خاص که تعداد بیت‌های بیشتری نسبت به اعداد معمولی ماشین دارند انجام می‌دهند. این بیت‌ها که اصطلاحاً بیت پشتیبان نامیده می‌شوند به اعداد اجازه می‌دهند که موقتاً دقت بیشتری داشته باشند و تعداد آن‌ها در حالت کلی از ماشینی به ماشین دیگر متفاوت است). در این‌جا نیز ممکن است که نتیجه‌ی عمل عددی ماشینی نباشد. لذا پس از انجام عمل حسابی، نتیجه‌ی عمل با استفاده از نگاشت نمایش ساز ماشین به یک عدد ماشینی تبدیل می‌شود. در نتیجه $x * y$ با $x \circledast y = \text{fl}(x * y)$ جای‌گزین می‌شود.