۲.۱ نمایش کامپیوتری اعداد، با دقت متناهی

قبل از پیاده‌سازی عددی یک الگوریتم برای تحلیل یک مدل نیاز است دانشی پایه‌ای از چگونگی ذخیره‌سازی و نمایش اعداد و انجام محاسبات در ماشین داشته باشیم. در زندگی روزمره از اعداد در مبنای ۱۰ استفاده می‌کنیم اما کامپیوترها معمولاً بر مبنای دو استوار بوده و از حساب دودویی استفاده می‌کنند.

اعداد حقیقی را می‌توان در مبنای عدد صحیحی هم‌چون $\beta \geq 2$ به صورت یک رشته‌ی نامتناهی مانند

$$x = (-1)^{\sigma}(b_n b_{n-1} \cdots b_0.b_{-1}b_{-2}\cdots)_{\beta} \qquad (۲.۱)$$

نمایش داد جایی‌که $b_n, b_{n-1}, \cdots$ اعدادی صحیح در بازه‌ی $[0, \beta-1]$ بوده و $\sigma \in \{0,1\}$ علامت عدد را مشخص می‌کند. عدد حقیقی متناظر برابر است با

$$x = (-1)^{\sigma}\sum_{i=-\infty}^{n} b_i\beta^i = (-1)^{\sigma}(b_n\beta^n + b_{n-1}\beta^{n-1} + \cdots + b_0 + b_{-1}\beta^{-1} + b_{-2}\beta^{-2} + \cdots).$$

در این نمایش برخی قراردادها درنظر گرفته شده‌اند. بعنوان نمونه اگر عددی به تعدادی نامتناهی صفر متوالی ختم شود، آن‌ها را حذف می‌کنند. مثلا بجای $(12.25000\cdots)_{10}$ از $(12.25)_{10}$ استفاده می‌شود. همچنین صفرهای قبل از قسمت صحیح عدد یعنی قبل از بخش $(-1)^{\sigma}(b_n b_{n-1}\cdots b_0)_{\beta}$ نیز حذف می‌شوند. پس بجای $(0012.25\cdots)_{10}$ از $(12.25\cdots)_{10}$ و بجای $(000.0025\cdots)_{10}$ از $(0.0025\cdots)_{10}$ استفاده می‌شود.

وقتی یک عدد حقیقی به صورت (۲.۱) بیان شود، محل قرار گرفتن ممیز بسیار مهم است. دو قالب معروف برای اعداد عبارتند از قالب ممیز ثابت و قالب ممیز شناور. به طور ساده و غیرقیق، در قالب ممیز ثابت تعداد محل‌هایی از حافظه که بعد از علامت ممیز برای ذخیره‌ی اعداد، اختصاص می‌یابد ثابت است اما در قالب ممیز شناور با وجود اینکه تعداد محل‌هایی از حافظه که به کل عدد اختصاص می یابد ثابت است اما محل علامت ممیز انعطاف‌پذیر بوده و اعداد با تعداد ارقام متفاوت در قسمت بعد از ممیز قابل ذخیره‌سازی هستند. بعنوان مثال فرض کنید یک قالب ممیز ثابت دهدهی

بتواند دو رقم اعشاری را ذخیره کند. پس در این قالب چنانچه کاربر هر یک از اعداد

$$12345.67$$

$$67123.45$$

$$1.23$$

و شبیه آن را وارد کند، امکان ذخیره و نمایش آنها بدون خطا وجود داشته اما عددی مانند 1.234 نمی‌تواند بدون خطا ذخیره شود. در مقابل فرض کنید در یک قالب ممیزشناور دهدهی، هفت رقم برای ذخیره‌ی تمام ارقام عدد اختصاص یافته باشد. در چنین قالبی چنانچه کاربر هر یک از اعداد

$$1.234567$$

$$123456.7$$

$$0.00001234567$$

$$123456700000$$

را وارد کند، امکان ذخیره و نمایش آنها بدون خطا ذخیره وجود دارد. جزییات بیشتر را بعدا خواهیم دید، در این‌جا فقط به ذکر این نکته بسنده می‌کنیم که در قالب ممیز شناور، محل قرار گرفتن ممیز می‌تواند با استفاده از تغییر قسمت توانِ عدد، شناور باشد. از یک زیست‌شناس که با میکروسکوپ کار می‌کند تا اخترشناسی که فواصل اجرام در کهکشان‌های دوردست برایش مهم است با محاسبات علمی درگیر هستند. به همین دلیل مهم است که از قالبی در نمایش اعداد استفاده شود که بتواند اعداد با اندازه‌های از بسیار کوچک تا بسیار بزرگ را ذخیره کند. در نتیجه استفاده‌ی عملی از قالب ممیز ثابت که توان پایینی برای ذخیره‌ی اعداد با اندازه‌های متفاوت دارد بسیار محدودتر است.

یک عدد حقیقی در قالب ممیز شناور به صورت

$$x = (-1)^{\sigma}\ m \times \beta^{e}$$

نمایش داده می‌شود جایی که $(-1)^{\sigma}$ علامت عدد است، m را **مانتیس** و e را **توان** عدد x می‌نامند. در این قالب ممیز شناور، می‌توان در صورت لزوم با تغییری مناسب در توان عدد، مانتیس را به صورت

$$m = (b_0.b_1b_2\cdots)_\beta$$

نوشت و همین نکته مزیت بزرگی برای نوشتن اعداد در قالب ممیز شناور در قیاس با نمایش (۲.۱) به وجود می‌آورد. به بیان صریح‌تر، وقتی عدد را به صورت ممیز شناور می‌نویسیم، سیستم داخلی ماشین از شرِّ تعقیبِ محل قرارگرفتن ممیز راحت می‌شود چرا که در این قالب، ممیز همیشه بعد از اولین رقم مانتیس قرار می‌گیرد. پس برای جمع‌بندی بحث تاکنون، مجموعه‌ی اعداد حقیقی در مبنای β را می توان در قالب ممیز شناور به صورت

$$F_\beta = \{(-1)^\sigma \ m \times \beta^e \mid m = (b_0.b_1b_2\cdots)_\beta\}$$

نشان داد جایی‌که β عدد صحیحی بزرگ‌تر یا مساوی دو، $0 \leq b_i \leq \beta - 1$ برای هر i، و e می‌تواند هر عدد صحیحی باشد. توجه کنید که نمایش اعداد با این شرایط یکتا نیست. مثلا عدد دهدهی 123.4 می‌تواند با هر یک از نمایش‌های

$$(1.234)_{10} \times 10^2 = (0.1234)_{10} \times 10^3 = (0.01234)_{10} \times 10^4$$

در این قالب ذخیره شود[۷] چراکه هر سه عدد بالا شرایط مجموعه‌ی F_{10} را داشته و در نتیجه عضو این مجموعه هستند. برای منحصربفرد شدن نمایش اعداد، اِعمال شرط دیگری نیز لازم است و آن این‌که رقم پیشروی b_0 ناصفر باشد. اعداد ممیز شناوری که در این شرط صدق می‌کنند را **نرمال** می‌نامند. نمایش عدد 123.4 در دستگاه اعداد ممیزشناور نرمال دهدهی به صورت یکتای $(1.234)_{10} \times 10^2$ می‌باشد. مجموعه‌ی اعداد حقیقی $\mathbb{R}$ یک مجموعه‌ی نامتناهی ناشماراست در حالی که حافظه‌ی محدود یک کامپیوتر تنها می‌تواند مقداری متناهی از اطلاعات را ذخیره کند. پس در عمل کامپیوترها تنها خواهند توانست زیرمجموعه‌ای متناهی از اعضای F_β را ذخیره کرده و به صورت دقیق نمایش دهند. بعنوان گام بعدی برای نمایش اعداد در ماشین، مجموعه‌ی

$$F_{\beta,p} = \{x \in F_\beta \mid m = (b_0.b_1b_2\cdots b_{p-1})_\beta\}$$

[۷] دو نمایش $(123.4)_{10} = (12.34)_{10} \times 10^1$ و نظایر آن در شرط قرارگرفتن ممیز بلافاصله بعد از اولین رقم مانتیس، صدق نکرده و مجاز نیستند.

را درنظر می‌گیریم که در آن p، **دقت**[۸] دستگاه ممیز شناور نامیده می‌شود. این مجموعه هرچند شمارا اما هم‌چنان نامتناهی است چرا که هیچ کرانی روی قسمت توان اعداد عضو آن اِعمال نشده است. مجموعه‌ای که اعضای آن با طبیعتِ محدود حافظه‌ی ماشین سازگار باشند را می‌توان با مشخص کردن کران‌های پایین و بالا برای توان اعداد عضو $F_{\beta,p}$ ساخت: بدین منظور فرض کنید L کران پایین و L کران بالای مجاز برای توان باشد. در این صورت مجموعه‌ی ممیزشناور قابل نمایش به صورت دقیق در ماشین را می‌توان به صورت زیر تعریف کرد:

$$F_{\beta,p}^{L,U} = \{x \in F_{\beta,p} \mid L \leq e \leq U\}.$$

با توجه به تعاریف بالا واضح است که داریم:

$$F_{\beta,p}^{L,U} \subset F_{\beta,p} \subset F_\beta$$

یعنی از مجموعه‌ی ناشمارای نامتناهی $\mathbb{R}$ از اعداد حقیقی در ریاضیات چند مرحله عقب‌نشینی کرده‌ایم تا به مجموعه‌ی شمارای متناهی $F_{\beta,p}^{L,U}$ که مجموعه‌ی اعداد ماشین نامیده می‌شود برسیم.

مثال ۲. *اعداد نامنفی نرمال موجود در دستگاه $F_{2,3}^{-1,2}$ را (به صورت اعدادی در مبنای ۱۰) مشخص کنید.*

چون $\beta = 2$ و $p = 3$ پس سه بیت برای ذخیره‌ی مانتیس اعداد داریم: $m = (b_0.b_1b_2)_2$ و چون بدنبال اعداد نرمال هستیم پس $b_0 \neq 0$، یعنی بیت b_0 فقط می‌تواند مقدار یک را اختیار کند. برای سادگی، توان‌ها را در مبنای ۱۰ نشان می‌دهیم. به ازای $e = L = -1$ اعداد زیر را در این دستگاه داریم:

$$(1.00)_2 \times 2^{-1} = (1 \times 2^0 + 0 \times 2^{-1} + 0 \times 2^{-2}) \times 2^{-1} = 1 \times \frac{1}{2} = 0.5,$$
$$(1.01)_2 \times 2^{-1} = (1 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2}) \times 2^{-1} = \frac{5}{8} = 0.625,$$
$$(1.10)_2 \times 2^{-1} = (1 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2}) \times 2^{-1} = \frac{6}{8} = 0.75,$$
$$(1.11)_2 \times 2^{-1} = (1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2}) \times 2^{-1} = \frac{7}{8} = 0.875.$$

[۸] precision

به ازای $e=0$، اعداد زیر را داریم:

$$\begin{aligned}
(1.00)_2\times 2^0 &= (1\times 2^0+0\times 2^{-1}+0\times 2^{-2})=\frac{8}{8}=1,\\
(1.01)_2\times 2^0 &= (1\times 2^0+0\times 2^{-1}+1\times 2^{-2})=\frac{10}{8}=1.25,\\
(1.10)_2\times 2^0 &= (1\times 2^0+1\times 2^{-1}+0\times 2^{-2})=\frac{12}{8}=1.5,\\
(1.11)_2\times 2^0 &= (1\times 2^0+1\times 2^{-1}+1\times 2^{-2})=\frac{14}{8}=1.75.
\end{aligned}$$

به طرز مشابه برای $e=+1$ داریم:

$$\begin{aligned}
(1.00)_2\times 2^1 &= 2,\\
(1.01)_2\times 2^1 &= \frac{20}{8}=2.5,\\
(1.10)_2\times 2^1 &= \frac{24}{8}=3,\\
(1.11)_2\times 2^1 &= \frac{28}{8}=3.5.
\end{aligned}$$

و نهایتا به ازای $e=U=+2$ اعداد زیر را داریم:

$$\begin{aligned}
(1.00)_2\times 2^2 &= \frac{32}{8}=4,\\
(1.01)_2\times 2^2 &= \frac{40}{8}=5,\\
(1.10)_2\times 2^2 &= \frac{48}{8}=6,\\
(1.11)_2\times 2^2 &= \frac{56}{8}=7.
\end{aligned}$$

بنابراین اعداد نامنفی موجود در مجموعه‌ی $F_{2,3}^{-1,2}$ عبارتند از:

$$\{0.5, 0.625, 0.75, 0.875, 1, 1.25, 1.5, 1.75, 2, 2.5, 3, 3.5, 4, 5, 6, 7\}.$$

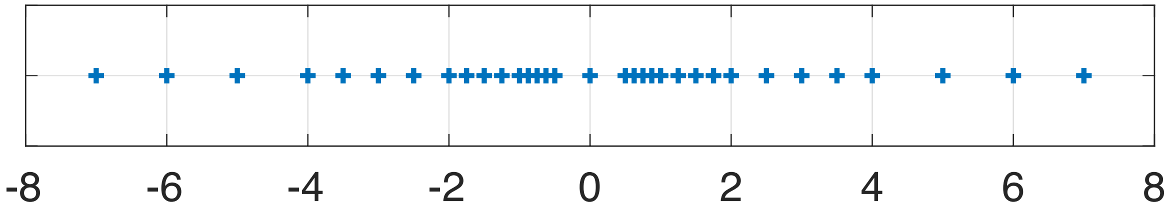
همان‌گونه که می‌بینیم کوچک‌ترین عدد نرمال مثبت در مجموعه‌ی $F_{2,3}^{-1,2}$ برابر است با

$$N_{\min}=(1.00)_2\times 2^{-1}=0.5$$

و بزرگ‌ترین عدد برابر است با

$$N_{\max} = (1.11)_2 \times 2^2 = 7.$$

تعداد کل اعداد (منفی و مثبت) عضو مجموعه‌ی شمارای $F_{2,3}^{-1,2}$ برابر است با ۳۲. این اعداد به انضمام صفر در شکل ۲.۱ نمایش داده شده‌اند.



شکل ۲.۱: اعداد ماشین مجموعه‌ی $F_{2,3}^{-1,2}$ در مثال ۲

نکته‌ی ۱. در مثال بالا فاصله‌ی بین اعداد متوالی متناظر با توان $e = -1$ برابر است با $2^{-3} = 0.125$. این فاصله برای اعداد نظیر با $e = 0$ برابر است با $2^{-2} = 0.25$. برای چهار عدد بعدی فاصله برابر است با $2^{-1} = 0.5$ و نهایتا فاصله‌ی اعداد ماشین نظیر با توان $e = U = +2$ برابر است با $2^0 = 1$. به طور کلی هرچند فاصله‌ی بین اعداد ممیزشناور عضو دستگاه یکنواخت نیست اما برای هر توان e فیکس، فاصله‌ی بین اعداد متعلق به $F_{\beta,p}^{L,U} \cap [\beta^e, \beta^{e+1}]$ یکنواخت است[۹].

نکته‌ی ۲. برای هر دو توان دلخواه m و n که در آن $L \le m, n \le U$، تعداد اعضای (کاردینالیتی) هر دو مجموعه‌ی $F_{\beta,p}^{L,U} \cap [\beta^m, \beta^{m+1}]$ و $F_{\beta,p}^{L,U} \cap [\beta^n, \beta^{n+1}]$ یکسان است.

نکته‌ی ۳. مجموعه‌ی اعداد ممیزشناور عضو $F_{\beta,p}^{L,U}$ نسبت به صفر متقارن است یعنی اگر $x \in F_{\beta,p}^{L,U}$ آن‌گاه $-x \in F_{\beta,p}^{L,U}$.

تمرین ۱. تعداد کل اعداد نرمال موجود در دستگاه ممیز شناور $F_{\beta,p}^{L,U}$ را بدست آورید.

دیدیم که هر عدد دودویی $(\beta = 2)$ ممیزشناور نرمال موجود در دستگاهی با دقت p را می‌توان به صورت

$$x = \pm(1.b_1 b_2 \cdots b_{p-2} b_{p-1})_2 \times 2^e$$

[۹] در حالت خاصی که $e = U$ باشد هرچند کران بالای بازه یعنی β^{U+1} عضو دستگاه $F_{\beta,p}^{L,U}$ نیست اما باز هم نکته‌ی ۱ به صورت مطرح‌شده صحیح است یعنی فاصله‌ی بین اعداد متعلق به $F_{\beta,p}^{L,U} \cap [\beta^U, \beta^{U+1}]$ یکسان است چراکه β^{U+1} اصولا عضو این اشتراک نیست. با استدلالی مشابه نکته‌ی ۲ برای حالت خاص $m = n = U$ نیز برقرار است.

نشان داد. واضح است که $1 = (1.00\cdots00)_2 \times 2^0$. پس کوچک‌ترین عدد نرمال بزرگ‌تر از یک برابر است با

$$(1.00\cdots01)_2 \times 2^0 = 1\times 2^0 + 0\times 2^{-1} + \cdots + 0\times 2^{-(p-2)} + 1\times 2^{-(p-1)} = 1 + 2^{-(p-1)}.$$

فاصله‌ی بین عدد یک و کوچک‌ترین عدد بزرگ‌تر از یک نقش بسیار مهمی در تحلیل خطای گردکردن در آنالیز عددی بازی می‌کند. این عدد را که **اپسیلون ماشین**[۱۰] نامیده می‌شود (در مبنای دو) برابر است با:

$$\varepsilon_M = 2^{-(p-1)}.$$

فاصله‌ی بین سایر اعداد متوالی عضو دستگاه ممیزشناور $F_{\beta,p}^{L,U}$ نیز ارتباط مستقیمی با اپسیلون ماشین دارد. این فاصله ارتباط نزدیکی با مفهوم مهم دیگری در تحلیل خطای گردکردن، به نام **واحد در آخرین مکان**[۱۱] دارد که با $ulp(x)$ نشان داده می‌شود و بیانگر وزن آخرین رقم مانتیس عدد نرمال x می‌باشد. می‌خواهیم فاصله‌ی بین هر دو عدد نرمال متوالی را در $F_{\beta,p}^{L,U}$ بیابیم. به طور کلی فاصله‌ی تمام اعداد ممیز شناور متوالی متعلق به $F_{\beta,p}^{L,U}\cap[\beta^e,\beta^{e+1}]$ یکسان است (در نکته‌ی ۱ هم این موضوع را دیدیم) [۱۲]. پس کافی است فاصله‌ی بین دو عدد ابتدایی عضو $[\beta^e,\beta^{e+1}]$ را بیابیم. فرض کنید $\tilde{x}$ اولین عدد عضو این بازه باشد یعنی $\tilde{x} = \beta^e = (1.00\cdots00)_\beta \times \beta^e$. پس اولین عدد بزرگ‌تر از $\tilde{x}$ برابر است با

$$(1.00\cdots01)_\beta \times \beta^e = (1 + 0 + \cdots 0 + 1\times\beta^{-(p-1)}) \times \beta^e = \tilde{x} + \beta^{-(p-1)} \times \beta^e$$

پس داریم:

$$ulp(x) = \beta^{e-p+1} = \varepsilon_M \beta^e.$$

دقت کنید که اگر $x > 0$ باشد، آن‌گاه $ulp(x)$ برابر است با فاصله‌ی بین x و عدد ممیز شناور بلافاصله بزرگ‌تر از آن و اگر $x < 0$ باشد، آن‌گاه $ulp(x)$ برابر است با فاصله‌ی بین x و عدد ممیز شناور

[۱۰] machine epsilon

[۱۱] unit in the last place

[۱۲] این فاصله در واقع برای تمام اعداد متعلق به فاصله‌ی بسته‌ی $[\beta^e,\beta^{e+1}]$ یکسان است مگر در حالت خاصی که $e = U$، که در این صورت، عدد نظیر عضو دستگاه نیست (در مثال قبل، به ازای $e = U = 2$، عدد 8 حاصل می‌شد که اصلا عضو دستگاه نبود). برای جلوگیری از ایجاد مشکل در این حالت خاص است که بحث را در مورد $F_{\beta,p}^{L,U}\cap[\beta^e,\beta^{e+1}]$ مطرح کرده‌ایم و نه فقط در مورد $[\beta^e,\beta^{e+1}]$.

بلافاصله کوچک‌تر از x. بعلاوه می‌توان دید که $\varepsilon_M = ulp(1)$. بار دیگر نکته‌ی ۱ را به یاد آورید. فواصلی که در آن‌جا ذکر شد همان $ulp(x)$ بودند. بعنوان نمونه دیدیم که $x = 5$ متناظر با توان $e = 2$ بود و به همین خاطر داریم:

$$ulp(5) = 2^{2-3+1} = 2^0 = 1.$$

به طرز مشابه داریم:

$$ulp(1.75) = 2^{0-3+1} = 2^{-2} = 0.25,$$

که فاصله‌ی $x = 1.75$ با کوچک‌ترین عددِ بزرگ‌تر از آن است.

ایده‌ی بیت پنهان

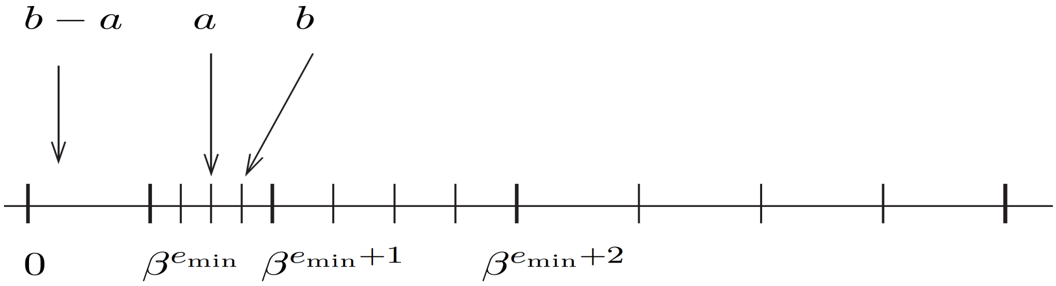
همانطور که گفتیم امروزه در کامپیوترها معمولا از مبنای دو برای نمایش اعداد استفاده می‌شود. پس تنها انتخابی که برای رقم پیشروی اعداد نرمال باقی می‌ماند $b_0 = 1$ است. پس اولین بیت تمام اعداد نرمال ممیز شناور یک می‌باشد. شانسی که این موضوع فراهم می‌سازد این است که این بیت که مقدارش همیشه یک است را به صورت ضمنی اعمال کرده و صریحا ذخیره نکنیم. در واقع اگر چنین نکنیم و بیت $b_0 = 1$ را صریحا ذخیره کنیم، فرصتِ صرفه‌جویی در یک محل حافظه (که محتویاتش را از قبل برای هر عدد نرمال می‌دانیم) را از دست داده‌ایم. البته که در کامپیوترها این صرفه جویی صورت پذیرفته و ذخیره‌سازی در حافظه از بیت بعدی شروع می‌شود. این بیت ضمنی که مقدارش همواره یک است و صریحا ذخیره نمی‌شود را **بیت پنهان**[۱۳] می‌نامند. ایده‌ی بیت پنهان تاثیر مستقیمی بر میزان دقت دستگاه اعداد ماشین خواهد داشت که آن‌را بعدا خواهیم دید.

اعداد زیرنرمال

بار دیگر مثال ۲ را به یاد آورید. دیدیم که کوچک‌ترین عدد نرمال مثبت در دستگاه $F_{2,3}^{-1,2}$ برابر با $N_{\min} = 0.5$ بود و فاصله‌ی آن با عدد بعدی برابر با 0.125 می باشد. با این حال فاصله‌ی بین $N_{\min}$ و صفر برابر با 0.5 است که در قیاس با 0.125 عدد بزرگی است. در واقع شکاف بزرگی در شکل ۲.۱ حول صفر وجود دارد و این نتیجه‌ی مستقیمی روی عدم برقراری برخی از مهمترین خواص ریاضی محاسبات با اعداد عضو دستگاه کلی $F_{\beta,p}^{L,U}$ دارد. به عنوان نمونه در وضعیت فعلی ممکن است a و b دو عدد متمایز عضو $F_{\beta,p}^{L,U}$ باشند اما با این وجود $b - a$ اصلا تعریف شده نباشد و یا داشته باشیم

[۱۳] hidden bit

$b-a=0$. در شکل ۳.۱ چنین موقعیتی را می‌بینیم جایی که $e_{\min}$ همان کوچک‌ترین توان مجاز یعنی L است، $\beta^{e_{\min}}=1$، $a=1.5$، $b=1.75$، و $\beta^{e_{\min}+1}=2$ مشابه مثال ۲ بوده و به وضوح



شکل ۳.۱: تفریق دو عدد عضو $F_{2,3}^{-1,2}$ وقتی فقط اعداد نرمال را داریم.

می‌بینیم که حاصل $b-a$ در شکافِ خالی از عدد ماشینی که حول صفر وجود دارد قرار گرفته. یک راه برای بهبود اوضاع معرفی اعدادی خاص به نام **اعداد زیرنرمال**[۱۴] است:

تعریف ۱. *یک عدد ممیزشناور غیرصفر در $F_{\beta,p}^{L,U}$، زیرنرمال نامیده می شود اگر بیت پیشروی آن صفر بوده و توان آن توان کمینه‌ی مجاز باشد یعنی دو شرط $b_0=0$ و $e=L$ با هم برقرار باشند.*

مثال ۳. *اعداد زیرنرمال نامنفیِ موجود در دستگاه $F_{2,3}^{-1,2}$ را (به صورت اعدادی در مبنای ۱۰) مشخص کنید.*

از آن‌جا که صفر طبق تعریف عددی زیرنرمال نیست مجموعه‌ی اعداد زیرنرمال عبارتند از

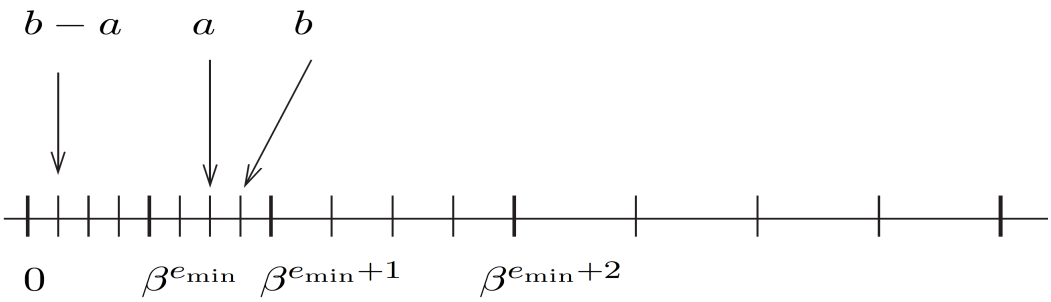
$$(0.01)_2\times 2^{-1}=(0\times 2^0+0\times 2^{-1}+1\times 2^{-2})\times 2^{-1}=0.125,$$
$$(0.10)_2\times 2^{-1}=(0\times 2^0+1\times 2^{-1}+0\times 2^{-2})\times 2^{-1}=0.25,$$
$$(0.11)_2\times 2^{-1}=(0\times 2^0+1\times 2^{-1}+1\times 2^{-2})\times 2^{-1}=0.375.$$

می بینیم که این اعداد زیرنرمال دقیقا در شکاف بین صفر و $N_{\min}=0.5$ قرار میگیرند. به طرز مشابه همین اعداد با علامت منفی، در شکاف بین بزرگ‌ترین عدد نرمال منفی یعنی -0.5 و عدد صفر قرار می‌گیرند. نکته‌ی مهم دیگر این‌که به طور کلی فاصله‌ی هر دو عدد زیرنرمال متوالی با فاصله‌ی بین کوچک‌ترین دو عدد نرمال یکسان می باشد. در این مثال خاص این فاصله به صورت اتفاقی با ε_M نیز یکی شد اما به طور کلی چنین نیست! □

بار دیگر به این پرسش که معرفی اعداد زیرنرمال چگونه می‌تواند مشکل صفرشدن حاصل تفریق برخی اعداد متمایز ماشین را حل کند باز می‌گردیم. شکل

[۱۴] subnorml (denormalized) numbers

را ببینید.



شکل ۴.۱: تفریق دو عدد عضو $F_{2,3}^{-1,2}$ وقتی اعداد زیرنرمال نیز اضافه شده‌اند.

به طور کلی وقتی اعداد زیرنرمال نیز موجود باشند غیرممکن است که حاصل تفریق دو عدد ماشین متمایز صفر شود. این البته فقط در مورد اعداد ماشین صدق می کند: چنانچه دو عدد حقیقی متمایز a و b اعداد ماشین نباشند همچنان این امکان وجود دارد که تفریق آنها صفر شود! در این مورد نیز بعدا مجددا بحث خواهیم کرد.

۱.۲.۱ خطای مطلق و نسبی

فرض کنید $\tilde{x}$ تقریبی برای عدد حقیقی x باشد. دو روش از مفیدترین ابزارهای اندازه گیری میزان درستیِ $\tilde{x}$ عبارتند از خطای مطلق

$$e(\tilde{x}) := |x-\tilde{x}|,$$

و خطای نسبی

$$\delta(\tilde{x}) := \frac{|x-\tilde{x}|}{|x|},$$

که وقتی $x=0$ تعریف نشده است. وقتی در یک محاسبه، کمیت‌ها بسیار کوچک یا بسیار بزرگ باشند و یا وقتی هم‌زمان با کمیت‌های کوچک و بزرگ سر و کار داشته باشیم، خطای نسبی، بیش از خطای مطلق اهمیت پیدا می‌کند. خطای نسبی بخاطر تقسیم موجود در تعریفش بدون واحد است. فرض کنید کمیت x با αx جایگزین شده و تقریب $\tilde{x}$ برای کمیت x نیز با $\alpha\tilde{x}$ جایگزین شود یعنی کیفیت تقریبِ $\alpha\tilde{x}$ برای کمیتِ αx با کیفیت تقریبِ $\tilde{x}$ برای کمیتِ x یکسان باشد. در این حالت خطای نسبی بدون تغییر باقی می‌ماند اما خطای مطلقِ تقریب جدید α برابر می‌شود. برای توضیح بیشتر دو موقعیت زیر را در نظر بگیرید. ابتدا فرض کنید $\tilde{x}=5.4599999\times10^{9}$ تقریبی برای مقدار

درست $x = 5.46 \times 10^9$ بوده باشد. داریم

$$e(\tilde{x}) = 10^2$$
$$\delta(\tilde{x}) = |\frac{e(\tilde{x})}{x}| = 1.8 \times 10^{-8}.$$

حال فرض کنید $\tilde{x} = 5.3 \times 10^{-11}$ تقریبی باشد که در محاسبه‌ای مشخص برای کمیت $x = 8 \times 10^{-15}$ به دست آمده. داریم

$$e(\tilde{x}) = 5.29 \times 10^{-11}$$
$$\delta(\tilde{x}) = \frac{5.29 \times 10^{-11}}{8 \times 10^{-15}} = 6.6 \times 10^{+3},$$

واقعیت این است که وقتی با محاسباتی در مقیاس 8×10^{-15} سر و کار داریم، خطایی (مطلق) به اندازه‌ی 5.29×10^{-11} نیز بزرگ است در حالی که وقتی با محاسباتی در مقیاس 10^9 سر و کار داریم، خطایی (مطلق) به اندازه‌ی 100 می‌تواند ناچیز باشد. عدد 8×10^{-15} در سناریوی دوم، فاصله‌ی بین پروتونها در هسته اتم است و 5.3×10^{-11}، فاصله‌ی بین پروتونها و الکترون‌ها در اتم هیدروژن است (هر دو فاصله بر حسب متر). بدیهی است که خطای (مطلقی) در حد 5.29×10^{-11} در چنین سطحی می‌تواند تفاوت‌هایی اساسی ایجاد کند. در سناریوی اول اما x فاصله‌ی مریخ از زمین بر حسب متر است. فرض کنید هدف این بوده که سفینه‌ای را در فاصله‌ی x متری از زمین بر سطح مریخ بنشانیم اما محلی که سفینه عملا بر آن فرود آمده در فاصله‌ی $\tilde{x}$ متری از زمین در سطح مریخ قرار دارد. با توجه به مساحت سطح مریخ که حدود 1.4×10^{14} مترمربع است، خطای (مطلق) به اندازه‌ی 100 متر چندان نگران‌کننده نمی‌باشد. به طور خلاصه خطای مطلق می‌تواند گمراه‌کننده باشد و معمولا خطای نسبی است که امکان داوری درست را به ما می‌دهد.

این بخش را با یادآوری مفهوم ارقام بامعنای یک عدد به پایان می‌بریم.

تعریف ۲. *ارقام بامعنای یک عدد عبارتند از اولین رقم ناصفر و تمام ارقام بعد از آن.*

بعنوان مثال عدد 0.0491 دارای سه رقم بامعناست و عدد 1.7320 دارای پنج رقم بامعنا می‌باشد. در آزمایشگاه فیزیک از این مفهوم برای نشان‌دادن تفاوت در میزان دقت یک وسیله‌ی اندازه‌گیری استفاده می‌کردیم. بعنوان نمونه وقتی می‌گفتیم در یک اندازه‌گیری طول 7.40 متر حاصل شده بطور ضمنی بیان می‌کردیم که وسیله‌ی استفاده‌شده برای سنجش این طول دقتی در حد صدم متر (یعنی سانتی متر) داشته ولی اگر نتیجه را به صورت 7.400 متر بیان می‌کردیم این منظور منتقل می شود که دقت

ابزار اندازه‌گیری در حد میلی‌متر داشته است.

۲.۲.۱ سبک‌های گردکردن: نگاشت اعداد حقیقی به اعداد ماشین

در محاسبات علمی اغلب با اعداد حقیقی سرو کار داریم. اما هرچند هر عدد ممیزشناور عضو دستگاه $F_{\beta,p}^{L,U}$، یک عدد حقیقی است، عکس این موضوع برقرار نیست. پس در عمل نیاز به نگاشتی هم‌چون

$$fl : \mathbb{R} \to F_{\beta,p}^{L,U}$$

از مجموعه‌ی ناشمارای اعداد حقیقی به مجموعه‌ی شمارای اعداد ماشین داریم. مرسوم است که برای مفید بودن چنین نگاشتی (که گردکردن نامیده می‌شود) دو شرط زیر اِعمال می‌شود. اولا اگر $x \in F_{\beta,p}^{L,U}$ آنگاه $fl(x) = x$ و ثانیا این‌که این نگاشت باید صعودی باشد یعنی اگر $x, y \in \mathbb{R}$ و $x \leq y$ آن‌گاه $fl(x) \leq fl(y)$. نتیجه‌ی مهم این دو شرط این است که درون بازه‌ی تولیدشده توسط x و $fl(x)$ یعنی در بازه‌های $[x, fl(x)]$ یا $[fl(x), x]$ هیچ عضو دیگری از $F_{\beta,p}^{L,U}$ وجود نخواهد داشت.

چهار نمونه‌ی معروف از نگاشت‌های گردکردن عبارتند از

- گردکردن به سمت پایین یا $-\infty$
- گردکردن به سمت بالا یا $+\infty$
- گردکردن به سمت صفر یا قطع‌کردن
- گردکردن به نزدیک‌ترین