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 تمام اعداد نرمال ممیز شناورِ دودویی باقی می‌ماند $b_0 = 1$ یک است. 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. چرا اگر چنین نکرده و بیت $b_0 = 1$ را صریحا ذخیره کنیم، فرصتِ صرفه‌جویی در یک محل حافظه (که محتویاتش از قبل برای تمام اعداد نرمال یکسان است) را از دست داده‌ایم. البته که در کامپیوترها این صرفه‌جویی 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 (همانطور که بعدا خواهیم دید) تاثیر مستقیمی بر میزان دقت دستگاه اعداد ماشین خواهد داشت.

اعداد زیرنرمال

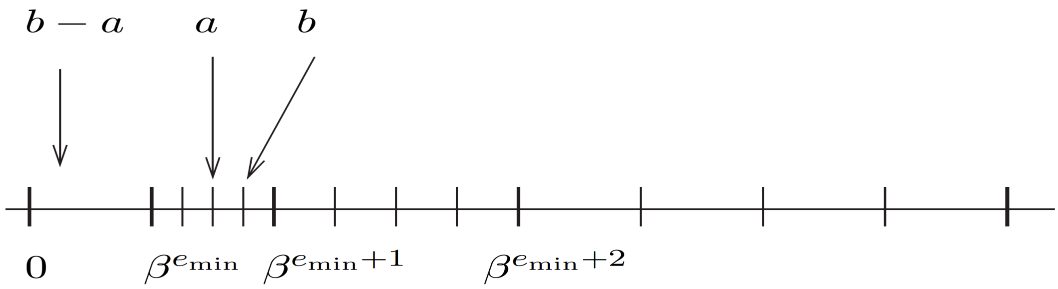
بار دیگر مثال ۲ را به یاد آورید. دیدیم که کوچک‌ترین عدد نرمال مثبت در دستگاه $F_{2,3}^{-1,2}$ برابر با $N_{\min} = 0.5$ بود و فاصله‌ی آن با عدد بعدی برابر با 0.125 می‌باشد. با این حال فاصله‌ی بین $N_{\min}$ و صفر برابر با 0.5 است که در قیاس با 0.125 عدد بزرگی است. در واقع شکاف بزرگی که در شکل ۱.۲ حول صفر وجود دارد، تاثیر مستقیمی روی عدم برقراری برخی از مهم‌ترین خواص ریاضی محاسبات با اعداد عضو دستگاه $F_{\beta,p}^{L,U}$ 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 آن‌گونه که بعد از بحث سبک‌های گردکردن خواهیم دید، داشته باشیم $b - a = 0$ به این معنا که حاصلِ $b - a$ به صفر گرد شود.) در شکل ۱.۳ این موقعیت را در دستگاه اعداد $F_{2,3}^{-1,2}$ مشاهده می‌کنیم جایی که $e_{\min}$ کوچک‌ترین توان مجاز یعنی L است، $\beta^{e_{\min}} = 0.5$، $a = 0.75$، $b = 0.875$ و $\beta^{e_{\min}+1} = 1$ مشابه مثال ۲ بوده و به وضوح 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 می‌شود 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.*

---

[۱۳] hidden bit

[۱۴] subnorml (denormalized) numbers



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

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

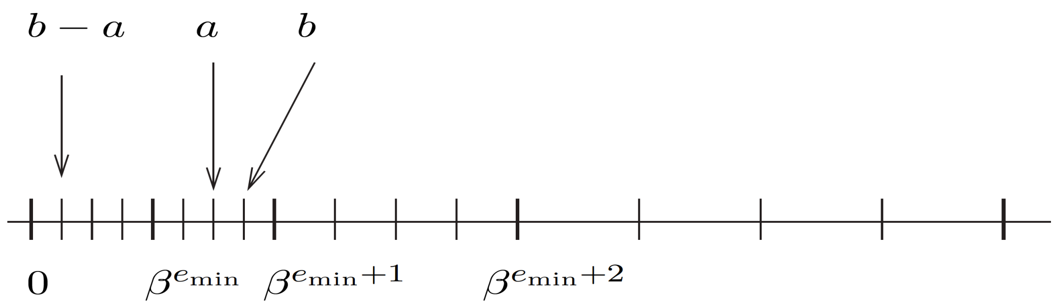
از آن‌جا که صفر طبق تعریف عددی زیرنرمال نیست مجموعه‌ی اعداد زیرنرمال عبارتند از

$$(0.01)_2 \times 2^{-1} = (0 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2}) \times 2^{-1} = \frac{1}{8} = 0.125,$$
$$(0.10)_2 \times 2^{-1} = (0 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2}) \times 2^{-1} = \frac{2}{8} = 0.25,$$
$$(0.11)_2 \times 2^{-1} = (0 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2}) \times 2^{-1} = \frac{3}{8} = 0.375.$$

می‌بینیم 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! $\square$

بار دیگر به این پرسش که معرفی اعداد زیرنرمال چگونه می‌تواند مشکل تعریف‌نشده‌بودن یا صفرشدنِ حاصل تفریق برخی اعداد متمایز ماشین را حل کند باز می‌گردیم. همان‌طور که شکل ۴.۱ نشان می‌دهد، وقتی اعداد زیرنرمال نیز موجود باشند، هیچگاه لزومی ندارد که حاصل تفریق دو عدد ماشین متمایز صفر شود (به صفر گرد شود).

این البته فقط در مورد اعداد ماشین صدق می‌کند: 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.



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

۱.۲.۱ خطای مطلق و نسبی

فرض کنید $\tilde{x}$ تقریبی برای عدد حقیقی x باشد. دو روش از مفیدترین ابزارهای اندازه‌گیری 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 $\tilde{x} = 5.4599999 \times 10^9$ تقریبی برای مقدار درست $x = 5.46 \times 10^9$ بوده باشد. داریم

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

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

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

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 پنج رقم بامعنا دارد. در آزمایشگاه فیزیک از این مفهوم برای نشان‌دادن تفاوت در میزان دقت یک وسیله‌ی اندازه‌گیری استفاده می‌کردیم. بعنوان نمونه وقتی می‌گفتیم در یک اندازه‌گیریِ طول، نتیجه‌ی 7.40 متر حاصل شده بطور ضمنی بیان می‌کردیم که وسیله‌ی استفاده‌شده برای این اندازه‌گیری، دقتی در حد صدم متر (یعنی سانتی‌متر) داشته ولی اگر نتیجه به صورت 7.400 متر بیان می‌شد، منظور این بود که ابزار استفاده‌شده، توان اندازه‌گیری در حد میلی‌متر را داشته است.

۲.۲.۱ سبک‌های گردکردن: نگاشت اعداد حقیقی به اعداد ماشین

در محاسبات علمی اغلب با اعداد حقیقی سرو کار داریم. هرچند هر عدد ممیزشناور عضو دستگاه $F_{\beta,p}^{L,U}$، یک عدد حقیقی است، عکس آن برقرار نیست. پس در عمل برای محاسبات با اعداد حقیقی در ماشین، ابتدا 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:

- اگر $x \in F_{\beta,p}^{L,U}$ آنگاه $fl(x) = x$.
- و ثانیا این‌که نگاشتِ گردکردن باید صعودی باشد یعنی اگر $x, y \in \mathbb{R}$ و $x \leq y$ آن‌گاه باید $fl(x) \leq fl(y)$.

و 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 مجموعه‌ی $F_{\beta,p}^{L,U}$ 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

در اینجا برای سادگی فقط روی دو سبک آخر تمرکز می‌کنیم. فرض کنید بخواهیم عدد حقیقی

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

که مانتیسش دارای بینهایت رقم است را با سبک قطع‌کردن به $p-1$ رقم گرد کنیم. در این سبک، به سادگی ارقامی از مانتیس که بعد از مکان $p-1$ هستند را حذف کرده و ارقام تا قبل از آن را نگه می‌داریم:

$$fl(x) = (-1)^{\sigma}(b_0.b_1b_2\cdots b_{p-1})_{\beta} \times \beta^{e}$$

برای توضیح سبک گردکردن به نزدیک‌ترین ابتدا حالتی که $\beta = 10$ است و عدد حقیقی

$$x = (-1)^\sigma (b_0.b_1b_2\cdots b_{p-1}b_pb_{p+1}\cdots)_{10} \times 10^e$$

را در نظر بگیرید. یکی از سه حالت زیر رخ می‌دهد:

- اگر $b_p < 5$ بود، مستقیما از سبک قطع‌کردن استفاده می‌کنیم:

$$fl(x) = (-1)^\sigma (b_0.b_1b_2\cdots b_{p-1})_{10} \times 10^e$$

- اگر $b_p > 5$ بود، ابتدا یک واحد به b_{p-1} اضافه کرده و سپس از قطع‌کردن تا p رقم استفاده می‌کنیم.

- در حالت خاصی که $b_p = 5$ باشد، کار گره می‌خورد و برای حل مشکل [۱۵]، معمولا از آنچه به نام گردکردن به نزدیک‌ترین زوج معروف است، استفاده می‌شود به این معنا که

  – چنانچه رقم b_{p-1} زوج باشد قرار می‌دهیم

$$fl(x) = (-1)^\sigma (b_0.b_1b_2\cdots b_{p-1})_{10} \times 10^e$$

  – اما چنانچه رقم b_{p-1} فرد باشد، یک واحد به آن اضافه کرده و سپس قطع می‌کنیم.

نتیجه اینکه در حالت خاصی که $b_p = 5$ گردکردن به گونه‌ای صورت می‌پذیرد که $fl(x)$ عددی زوج شود.

مثال ۴. *می‌خواهیم هر یک از شش عدد دهدهی* $\{1.23, 1.25, 1.28, 1.34, 1.35, 1.36\}$ *را با سبک گردکردن به نزدیکترین (زوج) به* $p = 2$ *رقم بامعنی گرد کنیم. با توجه به توضیحات قبل می‌توان دید که*

$$fl(1.23) = 1.2, \qquad fl(1.25) = 1.2, \qquad fl(1.28) = 1.3$$
$$fl(1.34) = 1.3, \qquad fl(1.35) = 1.4, \qquad fl(1.36) = 1.4.$$

[۱۵] tie-breaking

۳.۲.۱ میزان خطای گردکردن

قضیه‌ی مهم زیر میزان خطای ناشی از گردکردن هر عدد حقیقی را مشخص می‌کند.

قضیه ۱.۲.۱. *فرض کنید x یک عدد حقیقی در محدوده‌ی نرمال دستگاه ممیز شناور $F_{\beta,p}$ باشد. در این صورت میزان خطای نسبی گردکردن به یکی از دو صورت زیر کران‌دار می‌شود:*

- *اگر از یکی از دو سبک گردکردن به پایین یا بالا استفاده شود آنگاه*

$$\frac{|x-fl(x)|}{|x|} < \varepsilon_M.$$

- *اگر از یکی از سبک گردکردن به نزدیک‌ترین استفاده شود آنگاه*

$$\frac{|x-fl(x)|}{|x|} \leq \frac{\varepsilon_M}{2}.$$

اثبات. تنها حالتی که $\beta = 2$ است را بیان می‌کنیم (تعمیم به مبنای غیردو نیز سرراست است). در حالتی که عدد حقیقی x خود یک عدد ماشین باشد داریم $fl(x) = x$ و در نتیجه خطای گردکردن صفر بوده و قضیه به وضوح برقرار است. پس در ادامه حالتی را در نظر می‌گیریم که عدد حقیقی x یک عدد ماشین نباشد و بدون کاستن از کلیت، فرض کنید x که در محدوده‌ی نرمال (یعنی بین کوچک‌ترین و بزرگ‌ترین عدد نرمال)است، مثبت باشد. پس داریم

$$x = (1.b_1b_2\cdots b_{p-1}b_pb_{p+1}\cdots)_2 \times 2^e,$$

که می‌تواند مانتیسی با بینهایت رقم داشته باشد. نزدیک‌ترین عدد ماشینِ کوچک‌تر از x که آن‌را با x_- نشان می‌دهیم عبارت است از

$$x_- = (1.b_1b_2\cdots b_{p-1})_2 \times 2^e$$

هم‌چنین نزدیک‌ترین عدد ماشینِ بزرگ‌تر از x که آن‌را با x_+ نشان می‌دهیم عبارت است از

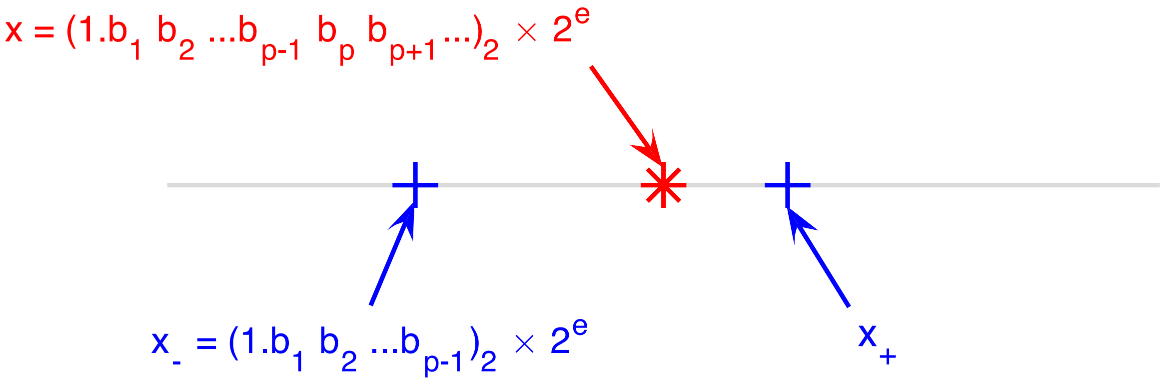
$$x_+ = \Big((1.b_1b_2\cdots b_{p-1})_2 + (0.00\cdots 01)_2\Big) \times 2^e.$$

نمایش بالا برای x_+ با اضافه‌کردن $ulp(x)$ به x_- که بزرگ‌ترین عدد ماشین کوچک‌تر از x_+ می‌باشد حاصل شده است.[۱۶]

چنانچه از یکی از سبک‌های گردکردن به پایین یا بالا استفاده شود آن‌گاه $fl(x)$ یکی از دو عدد ماشین x_- یا x_+ خواهد بود و در نتیجه فاصله‌ی x با $fl(x)$ از فاصله‌ی بین x_- با x_+ اکیدا کوچک‌تر خواهد بود یعنی

$$|x - fl(x)| < |x_+ - x_-| = 1 \times 2^{-(p-1)} \times 2^e = ulp(x). \qquad (۳.۱)$$

شکل ۵.۱ را ببینید.



شکل ۵.۱: عدد حقیقی x و دو عدد ماشین مجاور آن در اثبات قضیه‌ی ۱.۲.۱

از سوی دیگر چون x عددی نرمال است پس داریم: $x \geq (1.00\cdots0)_2 \times 2^e = 2^e$ یعنی

$$\frac{1}{|x|} \leq \frac{1}{2^e}. \qquad (۴.۱)$$

به کمک دو رابطه‌ی (۳.۱) و (۴.۱) داریم:

$$\frac{|x - fl(x)|}{|x|} < \frac{2^{-(p-1)} \times 2^e}{2^e} = 2^{-(p-1)} = \varepsilon_M, \qquad (۵.۱)$$

که حکم را در حالتی که از یکی از سبک‌های گردکردن به پایین یا بالا استفاده شود ثابت می‌کند.

در حالتی که از سبک گردکردن به نزدیک‌ترین استفاده شود از بین دو عدد ماشین x_- یا x_+ یکی که به x نزدیک‌تر است بعنوان $fl(x)$ تعیین خواهد بود و به همین دلیل فاصله‌ی x با $fl(x)$ کوچک‌تر

[۱۶] چگونگی تغییر مانتیس اعداد ماشین متوالی در مثال ۲ را به یاد آورید.

یا مساوی نصف فاصله‌ی بین x_- با x_+ خواهد بود یعنی

$$|x - fl(x)| \leq \frac{|x_+ - x_-|}{2}$$

کافی است این رابطه را بجای (۳.۱) جایگزین کنیم تا حکم در مورد سبک گردکردن به نزدیک‌ترین برقرار شود.

□