ایدهی بیت پنهان

همانطور که گفتیم امروزه در کامپیوترها معمولا از مبنای دو برای نمایش اعداد استفاده می شود. پس تنها انتخابی که برای رقم پیشروی تمام اعداد نرمال ممیز شناور دودویی باقی می ماند $b_0=0$ یک است. شانسی که این موضوع فراهم می سازد این است که این بیت که مقدارش همیشه یک است را به صورت ضمنی اعمال کرده و صریحا ذخیره نکنیم. چرا که اگر چنین نکرده و بیت $b_0=0$ را صریحا ذخیره کنیم، فرصت صرفه جویی در یک محل حافظه (که محتویاتش از قبل برای تمام اعداد نرمال یکسان است) را از دست داده ایم. البته که در کامپیوترها این صرفه جویی صورت پذیرفته و ذخیره سازی در حافظه از بیت بعدی شروع می شود. این بیت ضمنی که مقدارش همواره یک است و صریحا ذخیره نمی شود را بیت پنهان می نامند. ایده ی بیت پنهان (همانطور که بعدا خواهیم دید) تاثیر مستقیمی بر میزان دقت دستگاه اعداد ماشین خواهد داشت.

اعداد زيرنرمال

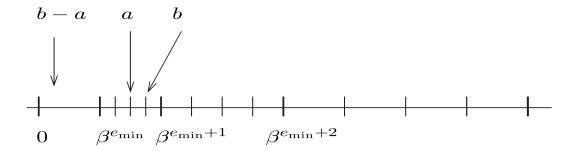
 $N_{\min}=N_{2,3}^{-1,2}$ برابر با جدد دوری برابر با ورید. دیدیم که کوچکترین عدد نرمال مثبت در دستگاه $P_{2,3}^{-1,2}$ برابر با عدد بعدی برابر با و 0.125 میباشد. با این حال فاصلهی بین N_{\min} و صفر برابر با 0.5 بود و فاصلهی آن با عدد بعدی برابر با و 0.125 میباشد. با این حال فاصلهی بین N_{\min} و صفر وجود 0.5 است که در قیاس با 0.5 عدد بزرگی است. در واقع شکاف بزرگی که در شکل ۲۰۱ حول صفر وجود دستگاه دارد، تاثیر مستقیمی روی عدم برقراری برخی از مهمترین خواص ریاضی محاسبات با اعداد عضو دستگاه این وجود $P_{\beta,p}^{L,U}$ باشند اما با $P_{\beta,p}^{L,U}$ دارد. به عنوان نمونه در وضعیت فعلی ممکن است $P_{\beta,p}^{L,U}$ باشند اما با این وجود $P_{\beta,p}^{L,U}$ باشند اما با در و یا آنگونه که بعد از بحث سبکهای گردکردن خواهیم دید، داشته باشیم $P_{\beta,p}^{L,U}$ به این معنا که حاصل $P_{\beta,p}^{L,U}$ به صفر گرد شود.) در شکل $P_{\beta,p}^{L,U}$ این موقعیت را در دستگاه اعداد $P_{\beta,p}^{L,U}$ مشاهده میکنیم جایی که $P_{\beta,p}^{L,U}$ مشابه مثال ۲ بوده و به وضوح میبینیم که حاصل $P_{\beta,p}^{L,U}$ مشابه مثال ۲ بوده و به وضوح میبینیم که حاصل $P_{\beta,p}^{L,U}$ معرفی اعدادی خاص به نام اعداد زیرنرمال ۲ 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 کمینه ی مجاز باشد یعنی دو شرط e=L و $b_0=0$ با هم برقرار باشند.

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

hidden bit $^{\prime}$

subnorml (denormalized) numbers (



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

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

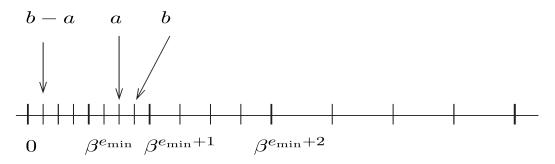
$$(0.01)_2 \times 2^{-1} = (0 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2}) \times 2^{-1} = \frac{1}{8} = 0.125,$$

$$(0.10)_2 \times 2^{-1} = (0 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2}) \times 2^{-1} = \frac{2}{8} = 0.25,$$

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

میبینیم که این اعداد زیرنرمال دقیقا در شکاف بین صفر و 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. نکته مهم دیگر این که به طور کلی فاصله می هر دو عدد زیرنرمال متوالی با فاصله ی 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 تعریفنشدهبودن یا صفرشدن حاصل تفریق برخی اعداد متمایز ماشین را حل کند باز میگردیم. همانطور که شکل ۴۰۱ نشان میدهد، وقتی اعداد زیرنرمال نیز موجود باشند، هیچگاه لزومی ندارد که حاصل تفریق دو عدد ماشین متمایز صفر شود (به صفر گرد شود).



شكل ۴.۱: تفريق دو عدد عضو $F_{2,3}^{-1,2}$ وقتى اعداد زيرنرمال نيز اضافه شدهاند.

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.

۱۰۲۰۱ خطای مطلق و نسبی

 \tilde{x} فرض کنید \tilde{x} تقریبی برای عدد حقیقی x باشد. دو روش از مفیدترین ابزارهای اندازهگیری میزان درستی 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 با همیت پیدا میکند. خطای نسبی \tilde{x} برای کمیت x نیز با x جایگزین شود یعنی کیفیت تقریب x برای کمیت x برای کمیت با کیفیت تقریب x برای کمیت x یکسان باشد. در این حالت خطای نسبی، بدون تغییر باقی می ماند اما خطای مطلق تقریب جدید، x برابر می شود. برای توضیح بیشتر دو موقعیت زیر را در نظر بگیرید. ابتدا فرض کنید x=0 بوده باشد. داریم مقدار درست x=0 بوده باشد. داریم

$$e(\tilde{x}) = 10^2,$$

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

 $x=8 imes 10^{-15}$ حال فرض کنید $ilde{x}=5.3 imes 10^{-15}$ تقریبی باشد که در یک محاسبه ی خاص برای کمیت $ilde{x}=5.3 imes 10^{-11}$ 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 $10^{-1} \times 0.00 \times 0.00$ نیز بزرگ است در حالی که وقتی با محاسباتی در مقیاس 0.00 سر و کار داریم، خطایی (مطلق) به اندازه ی 0.00 می تواند ناچیز باشد. عدد $0.00 \times 0.00 \times 0.00$ ناچیز باشد و $0.00 \times 0.00 \times 0.00$ ناچیز باشد. عدد و فاصله بر حسب متر). بدیهی است که خطای (مطلقی) در حد $0.00 \times 0.00 \times 0.00 \times 0.00$ ناچیز تفاوتهایی اساسی ایجاد کند. در سناریوی اول اما $0.00 \times 0.00 \times 0.00 \times 0.00 \times 0.00 \times 0.00$ ناچیز بنشانیم اما محلی که سفینه عملا بر آن این بوده که سفینه ای را در فاصله ی متری از زمین بر سطح مریخ بنشانیم اما محلی که سفینه عملا بر آن فرود آمده در فاصله ی $0.00 \times 0.00 \times 0.00$

این بخش را با یادآوریِ 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 اندازهگیری طول، نتیجهی 7.40 متر حاصل شده بطور ضمنی بیان میکردیم که وسیلهی استفاده شده برای این اندازهگیری، دقتی در حد صدم متر (یعنی سانتیمتر) داشته. به بیان دیگر روی ابزار اندازهگیری، دو عدد 7.40 نیز وجود داشته ولی مقداری که خوانده ایم 7.40 بوده است. از سوی دیگر اگر نتیجه به صورت 7.40 متر (که هرچند به لحاظ ریاضی با 7.40 یکسان است، اما یک رقم بامعنای بیشتر از آن دارد) بیان میشد، منظور این بود که ابزار استفاده شده، دقت اندازهگیری بیشتری (در میلیمتر) داشته است.

۲۰۲۰۱ سبکهای گردکردن: نگاشت اعداد حقیقی به اعداد ماشین

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

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:

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

و نتیجه ی مهم این دو شرط این است که درون بازه ی تولیدشده توسط x و fl(x) یعنی در بازههای fl(x) مهم این دو شرط این است که درون بازه ی $F_{\beta,p}^{L,U}$ وجود نخواهد داشت. [fl(x),x] یا [x,fl(x)]

چهار نمونهی معروف از نگاشتهای گردکردن عبارتند از

- $-\infty$ گردکردن به سمت پایین یا $-\infty$
 - $+\infty$ 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 رقم گرد کنیم. در این سبک، به سادگی ارقامی از مانتیس که بعد از b_{p-1} قرار دارند را حذف کرده و ارقام تا قبل از آن را نگه می داریم:

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

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 5 مستقیما از سبک قطع
کردن استفاده میکنیم: • اگر 5 می

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

- 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 معمولا از آنچه به نام گردکردن به نزدیک ترین زوج معروف است، استفاده می شود به این معنا که
 - جنانچه رقم b_{p-1} زوج باشد قرار می دهیم –

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

- 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

$$fl(1.23) = 1.2,$$
 $fl(1.25) = 1.2,$ $fl(1.28) = 1.3$

$$fl(1.34) = 1.3,$$
 $fl(1.35) = 1.4,$ $fl(1.36) = 1.4.$

حال به صورت مختصر سبک گردکردن به نزدیک ترین (زوج) را برای اعداد دودویی مرور می کنیم. فرض کنید عدد دودویی x با مانتیس x مانتیس x مانتیس x مانتیس x مانتیس x مانتیس x داده شده و بخواهیم آنرا تا x بیت گرد کنیم. بعلاوه فرض کنید x و عدد ماشین با مانتیس x رقمی هستند که حول x می باشند. شکل x در می توان تصور کرد. می خواهیم بدانیم از بین x و x کدام را بعنوان x برگزینیم. قاعده ی کلی باز هم این است که به اولین رقمی (بیتی) که قرار است حذف شود یعنی بیت x نگاه می کنیم.

• اگر b_p صفر بود آنگاه در واقع در وضعیتی هستیم که x از وسط بازهی $[x_-, x_+]$ کوچکتر است و به

tie-breaking\

همین خاطر به سمت پایین گرد میکنیم یعنی کافی است به سادگی b_p و همهی بیتهای بعد از آنرا حذف کنیم تا مانتیس به صورت $(b_0.b_1\cdots b_{p-1})_{\beta}$ حاصل شود.

- اگر بیت b_p یک بود و حداقل یکی دیگر از بیتهای بعد از آن نیز یک بود آنگاه میتوان نشان داد که x از وسط بازه ی $[x_-, x_+]$ بزرگتر است و به همین خاطر به سمت بالاگرد میکنیم یعنی یک واحد به بیت x اضافه میکنیم تا x به دست آید.
- اگر بیت b_p یک بود و همچنین تمام بیتهای بعد از آن صفر بود آنگاه در واقع گره رخ داده یعنی x_+ یک بود و همچنین تمام بیتهای بعد از آن صفر بود آنگاه در واقع گره رخ داده یعنی در وضعیتی هستیم که x دقیقا وسط بازه ی x_+ این است را انتخاب میکنیم. عددی که بیت x_+ آن زوج (یعنی صفر) است را انتخاب میکنیم.

مثال ۵. میخواهیم هر یک از سه عدد دودویی $\{1.11001, 1.11101, 1.11100\}$ را با سبک گردکردن به نزدیکترین (زوج) به p=3 رقم بامعنی گرد کنیم. از آنجا که چهارمین بیت عدد p=3 رقم بامعنی گرد کنیم. از آنجا که چهارمین بیت عدد $fl(1.11001)=(1.11)_2$

در مورد عدد 1.11101 از آنجا که چهارمین بیت یک بوده و بیت ناصفر دیگری هم بعد از آن وجود دارد پس به بالا گرد میکنیم یعنی داریم:

$$fl(1.11101) = (1.11)_2 + (0.01)_2 = (10.00)_2 = (1.00)_2 \times 2^1.$$

fl(1.11100)=1.11100 همچنین در مورد عدد 1.11100 گره داریم و با گردکردن به نزدیکترین زوج خواهیم داشت: 1.11100

توضیح بیشتر اینکه در این مثال میخواهیم به p=3 بیت گرد کنیم و با توجه به سه عدد داده شده داریم

$$[x_-, x_+] = [(1.11)_2, (1.00)_2 \times 2^1] = [(1.75)_{10}, (2)_{10}]$$

و نقطهی وسط بازه نیز برابر است با

$$\mu := (1.875)_{10} = (1.111)_2 = (1.11100 \cdots 0)_2.$$

 $[x_-,\mu]$ یعنی در بازه ی $[x_-,\mu]$ در نیمه ی ابتدایی بازه ی $[x_-,x_+]$ یعنی در بازه ی ایس واضح است که عدد اول یعنی x_- گرد شود. از سوی دیگر عدد 1.11101 بزرگتر از μ است و در نتیجه به بالا یعنی به x_+ گرد می شود.

۳.۲.۱ میزان خطای گردکردن

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

قضیه ۱۰۲۰۱ فرض کنید x یک عدد حقیقی در محدوده ی نرمال دستگاه ممیز شناور $F_{\beta,p}$ باشد. در این صورت میزان خطای نسبی گردکردن به یکی از دو صورت زیر کران دار می شود:

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

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

• اگر از سبک گردکردن به نزدیکترین استفاده شود آنگاه

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

اثبات. تنها حالتی که $2=\beta$ است را بیان می کنیم (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 می گیریم که عدد حقیقی x یک عدد ماشین نباشد و بدون کاستن از کلیت، فرض کنید x که در محدوده ی نرمال (یعنی بین کوچک ترین و بزرگ ترین 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 ماشین 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 می همچنین عدد ماشینِ بزرگتر از x_+

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

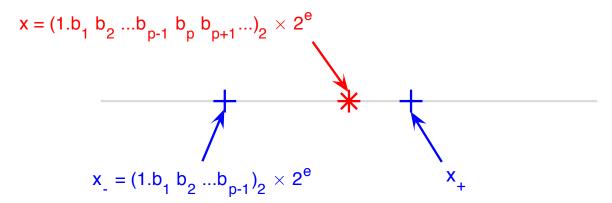
نمایش بالا برای x_+ با اضافه کردن ulp(x) به x_- که بزرگترین عدد ماشین کوچکتر از x_+ میباشد حاصل

شده است۱.

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

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

شکل ۵.۱ را ببینید.



شکل ۵۰۱: عدد حقیقی x و دو عدد ماشین مجاور آن در اثبات قضیهی ۱۰۲۰۱

 $x \geq (1.00\cdots 0)_2 \times 2^e = 2^e$ از سوی دیگر چون x ممکن است هر عدد نرمالی باشد پس داریم: $x \geq (1.00\cdots 0)_2 \times 2^e = 2^e$ یعنی

$$\frac{1}{|x|} \le \frac{1}{2^e}.\tag{4.1}$$

به کمک دو رابطهی (۳۰۱) و (۴۰۱) داریم:

$$\frac{|x - fl(x)|}{|x|} < \frac{2^{-(p-1)} \times 2^e}{2^e} = 2^{-(p-1)} = \varepsilon_M, \tag{(2.1)}$$

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_- یا بیکی که به نزدیکتر است بعنوان fl(x) تعیین خواهد بود و به همین دلیل فاصله x با x_- کوچکتر یا مساوی x_- نزدیکتر است بعنوان x_-

ا چگونگی تغییر مانتیس اعداد ماشین متوالی در مثال ۲ را به یاد آورید.

نصف فاصله ی بین x_{-} با x_{+} خواهد بود یعنی

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

کافی است این رابطه را بجای (۳۰۱) جایگزین کنیم تا حکم در مورد سبک گردکردن به نزدیکترین برقرار شود.