## ۲.۳.۱ برخی از خواص نامتعارف حساب ممیز شناور

۱. جمع مميزشناور، لزوما شركتپذير نيست يعنى

 $\exists a, b, c \in F_{\beta, p}^{L, U} \ s.t. \ (a \oplus b) \oplus c \neq a \oplus (b \oplus c).$ 

بعنوان مثال دستگاه بازیچه ی  $F_{2,3}^{-1,2}$  را با سبک گردکردن به نزدیکترین (زوج) در نظر بگیرید. داریم:

$$0.5 \oplus (2.5 \oplus 0.75) = 0.5 \oplus fl(3.25) = 0.5 \oplus 3 = fl(3.5) = 3.5,$$

$$(0.5 \oplus 2.5) \oplus 0.75 = fl(3) \oplus 0.75 = 3 \oplus 0.75 = fl(3.75) = 4,$$

در حالی که هیچیک از این دو نیز جواب درست ریاضی نیستند!

٢. ضرب مميزشناور، لزوما شركتيذير نيست يعنى

 $\exists a, b, c \in F_{\beta, p}^{L, U} \ s.t. \ (a \otimes b) \otimes c \neq a \otimes (b \otimes c).$ 

۳. ضرب ممیزشناور، لزوما بر جمع ممیزشناور پخشپذیر نیست یعنی

 $\exists a,b,c \in F_{\beta,p}^{L,U} \ s.t. \quad a \otimes (b \oplus c) \neq (a \otimes b) \oplus (a \otimes c).$ 

- ۴. ترتیب انجام عملیات ممیزشناور، گاهی اوقات بر درستی نتیجه تاثیرگذار است.
  - ۵. خاصیت حذفی عمل جمع (و همینطور عمل ضرب) لزوما برقرار نیست یعنی

 $\exists a,b,c \in F_{\beta,p}^{L,U} \ s.t. \quad a \oplus b = a \oplus c \ \& \ b \neq c.$ 

 $\exists a,b,c \in F_{\beta,p}^{L,U} \ s.t. \quad a \otimes b = a \otimes c \ \& \ b \neq c.$ 

۶. حاصل ضرب یک عدد ممیز شناور در معکوسش لزوما مساوی یک نیست.

همانگونه که قبلا گفتیم اگر بخواهیم خواص عملیات حسابی ممیزشناور را کنکاش کنیم نیاز چندانی

نیست که جزئیات غیرضروروی چگونگی انجام عملیاتی که واقعا در مبنای دو در ماشین صورت میپذیرد را بدانیم به مورت زیر استفاده کنیم: برای را بدانیم حاصل x\*y در ماشینی با دقت p چقدر خواهد شد، میتوانیم x\*y را با دقت بینهایت رقم انجام دهیم و تنها پاسخ نهایی را یک بار به عددی متعلق به  $F_{\beta,p}^{L,U}$  گرد کنیم:

تمرین ۲. سه عدد

$$a = +2.3371258 \times 10^{-5}$$
$$b = +3.3678429 \times 10^{+1}$$

 $c = -3.3677811 \times 10^{+1}$ 

را در دستگاه  $F_{10,8}^{-5,+5}$  در نظر گرفته و مقدار  $a \oplus b \oplus c$  و  $a \oplus b \oplus c$  را به دست آورده و مقایسه کنید. فرض کنید قواعد استاندارد آی-تریپل-ای در دستگاه  $F_{10,8}^{-5,+5}$  برقرار است.

## ۳.۳.۱ برخی فجایع و رخدادهای ناشی از استفادهی نامناسب از حساب ممیزشناور

با اینکه خطاهای گردکردن معمولا کوچک هستند، وقتی در الگوریتمهای طولانی و پیچیده چنین خطاهایی تکرار و انباشته میگردند، میتوانند آثار فاجعهباری داشته باشند. در اینجا برخی از اتفاقاتی که در جهان واقعی بخاطر خطاهای محاسبات کامپیوتری رخ دادهاند را مرور میکنیم:

۱. انفجار موشک آریان۵ در ۴ ژوئن ۱۹۹۶ در گینهی فرانسه. این انفجار در اثر خطای سرریز در کامپیوتر تنظیمکنندهی مسیر حرکت موشک رخ داد. این موشک که توسط آژانس فضایی اروپا و با هزینه ۷ میلیون دلار ساخته شده بود، حدود ۴۰ ثانیه پس از پرتاب در ارتفاع ۳۷۰۰ متری منفجر شد. دو هفته بعد از این رخداد گروهی از بازرسان گزارش خود را از دلایل انفجار موشک ارائه کردند. در گزارش بیان شده که یک عدد ممیزشناور ۶۴ بیتی مربوط به شتاب افقی موشک نسبت به سکوی پرتاب باید پس از یک تبدیل در محل یک عدد صحیح ۱۶ بیتی ذخیره میشد. این عدد، بزرگتر از ۲۷۶۷ که بزرگترین عدد قابل ذخیره در ۱۶ بیت است، بوده و در نتیجه با خطای سرریز رخ داده است و انحراف موشک از مسیر و انفجار آن در نتیجهی این ایراد نرمافزاری بوده است.

ا مثلا این که اگر در مراحل میانی محاسبه ای با دو عدد نرمال، عددی زیرنرمال ظاهر شد ماشین چه خواهد کرد؟ و یا اینکه استفاده از بیت نگهبان یا بیت گردکردن یا بیت چسبناک (که در تضمین خاصیت بیشترین کیفیت چهار عمل اصلی است) دقیقا به چه صورت می باشد؟

۲۰ شکست ماموریت موشک آمریکایی در جریان جنگ خلیج فارس در ۲۵ فوریه ۱۹۹۱۰ موشک آمریکایی پاتریوت که در واقع یک موشک ضدموشک است، قرار بود پس از پرتاب از ظهران عربستان، موشک اسکادی را که توسط ارتش عراق پرتاب شده بود ردگیری کند اما بواسطهی اشتباه در محاسبهی زمان نتوانست موشک اسکاد را مورد اصابت قرار دهد و در نتیجه 28 سرباز آمریکایی کشته شدند. در واقع زمان محاسبه شده توسط ساعت داخلی سیستم در واحد 01 برابر یک ثانیه اندازه گیری شده و نهایتا در عدد  $\frac{1}{10}$  ضرب می شده تا زمان بر حسب ثانیه بدست آید. هرچند بسط دهدهی عدد  $\frac{1}{10}$  فقط یک رقم بامعنا دارد، بسط دودویی آن نامتناهی است:

$$(0.1)_{10} = 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9} + 2^{-12} + 2^{-13} + \cdots$$
$$= (1.10011001100 \cdots)_2 \times 2^{-4} = (1.100\overline{1100})_2 \times 2^{-4}.$$

در سیستم موشک پاتریوت، عدد  $\frac{1}{10}$  بعد از قطعشدن در یک ثبات ۲۴ بیتی ذخیره میشد. این خطای گردکردن البته کوچک بوده است. اما از آنجا که باتری موشک پاتریوت به مدت ۱۰۰ ساعت در وضعیت آمادهباش بود، زمان پرتاب موشک پاتریوت باید در 10برابر تعداد ثانیههای موجود در ۱۰۰ ساعت (یعنی 3600000 = 0.000 0.000 0.000 ضرب میشده و ضرب در این عدد بزرگ باعث میشود که خطای کوچک گردکردن، بزرگ شده و نهایتا موشک پاتریوت با تاخیری 0.34 ثانیهای پرتاب شود. از سوی دیگر موشک اسکاد در هر ثانیه تقریبا 1676 متر را میپیماید و در نتیجه در مدت زمان 0.000 ثانیه، بیش از 0.000 متررا طی میکند و این فاصله خارج از برد پوشش داده شده توسط یک موشک پاتریوت است. درنتیجه موشک اسکاد نهایتا به هدف اصابت میکند.

۳. تغییر احزاب تشکیل دهنده ی پارلمان آلمان در سال ۱۹۹۲ در اثر خطای گردکردن. در سیستم پیچیده ی انتخابات آلمان اگر حزبی کمتر از پنج درصد آراء رابدست آورد، نمی تواند وارد پارلمان شده و کلیه آراء آن حزب حذف شده و کرسیهای پارلمانی مربوطه بین سایر احزاب بصورت خاصی پخش می شود . پس از اعلام نتایج انتخابات ۵ آوریل ۱۹۹۲، اعلام می شود که حزب سبزها توانسته پنج درصد آراء را بدست آورد اما بعد از نیمه شب یکی از اعضای کمیته ی انتخابات متوجه شد که حزب سبزها در واقع توانسته بوده %4.97 آراء را بدست آورد، اما برنامه ای که محاسبات را انجام می داده، تنها یک رقم بعداز ممیز اعشار را پس از گرد کردن به سمت بالا نهایتا چاپ می کرده و به همین دلیل عدد %4.97 را به پنج درصد تبدیل کرده بود. پس از مشخص شدن این اشتباه حزب سبزها نتوانست وارد پارلمان شود و حزب GPD توانست اکثریت پارلمان را به خود اختصاص دهد.

## ۴.۱ عدد وضعیت مساله

میخواهیم بدانیم یک مسئله ی ریاضی چقدر به خطاهای اندک کردکردن حساس است. برخی از پرسشهای مرتبط و مهمی که مطرح میشوند عبارتند از:

- تا چه حد می توان روی میزان درستی جواب عددی یک مسئله ی ریاضی حساب کرد؟
- آیا اصولا حل مسالهای که به خطاهای گردکردن حساسیت زیادی دارد با روشهای عددی (که مرتبا درگیر خطاهای اندک گردکردن میباشند) ایده ی خوبی است؟

عدد وضعیت ایک مساله، نشان میدهد که عدم اطمینان در دادههای مساله تا چه حد میتواند جواب مساله را تغییر دهد، یعنی میزان حساسیت خروجی مساله را به تغییرات اندک در ورودی آن توصیف میکند. پس مفهوم عدد وضعیت برای یک مساله متناظر است با مفهوم پیوستگی برای یک تابع. (به زبانی نه کاملا دقیق) عدد وضعیت با تعیین نسبت

مشخص مىشود.

بعنوان مثال، مسئله ی تعیین مقدار تابع  $\mathbb{R} \to \mathbb{R}$  در نقطه ی x را در نظر گرفته و فرض کنید x با ایجاد اختلالی اندک در ورودی x حاصل شده باشد (مثلا به صورت  $x = x + \Delta x$ ). واضح است که:

$$\frac{f(x) - f(\tilde{x})}{f(x)} = \frac{f(x) - f(\tilde{x})}{x - \tilde{x}} \times \frac{x}{f(x)} \times \frac{x - \tilde{x}}{x}$$

که در آن ارتباط بین اختلال ورودی و خروجی مشخص است:

$$\frac{f(x) - f(\tilde{x})}{f(x)} = \underbrace{\frac{f(x) - f(\tilde{x})}{x - \tilde{x}}}_{\text{times of the size of the size}} \times \frac{x}{f(x)} \times \underbrace{\frac{x - \tilde{x}}{x}}_{\text{times of the size}} \times \frac{x - \tilde{x}}{f(x)}$$

$$= \underbrace{x - \tilde{x}}_{\text{times of the size}} \times \frac{x - \tilde{x}}{f(x)} \times \underbrace{\frac{x - \tilde{x}}{x}}_{\text{times of the size}} \times \frac{x - \tilde{x}}{f(x)} \times \underbrace{\frac{x - \tilde{x}}{x}}_{\text{times of the size}} \times \frac{x - \tilde{x}}{f(x)} \times \underbrace{\frac{x - \tilde{x}}{x}}_{\text{times of the size}} \times \underbrace{\frac{x - \tilde{x}}{x}}_{\text{times of times of the size}} \times \underbrace{\frac{x - \tilde{x}}{x}}_{\text{times of times of time$$

condition number

نماد مرسوم برای عدد وضعیت مسئله ی ارزیابی تابع f در نقطه ی x است. با توجه به فرمول  $\kappa_f(x)$  قبل داریم:

$$\frac{|f(x) - f(\tilde{x})|}{|f(x)|} \approx \kappa_f(x) \times \frac{|x - \tilde{x}|}{|x|} \tag{9.1}$$

که در آن

$$\kappa_f(x) = \frac{|x| |f'(x)|}{|f(x)|}$$

به طور کلی عدد وضعیت نسبی یک مساله را میتوان از اندازهی

مشتق (خروجی بر حسب ورودی) 
$$\times \frac{eرودی}{خروجی}$$
 مشتق (۱۰.۱)

تخمین زد. مسالهای که عدد وضعیتش کوچک باشد را مسئلهی خوش وضع مینامیم. به همین ترتیب یک مساله را بدوضع نامیم اگر عدد وضعیتش بزرگ باشد. اینکه چه عدد وضعیتی بزرگ محسوب میشود بسته به موقعیت، متفاوت است. در این مورد در ادامه بحث خواهیم کرد. بطور کلی اینکه ((چه عدد وضعیتی نگرانکننده است؟)) مرتبط خواهد بود با اینکه چه میزان درستی از جواب مساله انتظار داریم و همچنین اینکه دقت دستگاه ممیزشناوری که از آن استفاده میکنیم چقدر است. توجه کنید که عدد وضعیت مساله موضوعی ریاضی است در بطن خود مساله و به خودی خود هیچ ارتباطی با الگوریتمی که بعدا برای حل آن مساله برخواهیم گزید ندارد.

از تعریف کلیای که برای عدد وضعیت کردیم و یا همچنین از رابطهی (۹.۱) میتوان فهمید که عدد وضعیت، در واقع به طور تقریبی، میزان درشتشدنِ خطاهای گردکردن ورودی x را در خروجی (x) اندازهگیری میکند. چنانچه از طرفین رابطهی (۹.۱) لگاریتم در مبنای ۱۰ بگیریم، خواهیم داشت:

$$-\log_{10}\left(\frac{|f(x) - f(\tilde{x})|}{|f(x)|}\right) \approx -\log_{10}\left(\frac{|x - \tilde{x}|}{|x|}\right) - \log_{10}\left(\kappa_f(x)\right) \tag{11.1}$$

در این رابطه نکات زیر حائز اهمیت هستند:

f(x) و  $f( ilde{x})$  و المحال و الم

- $\cdot x$  همچنین  $\log_{10}\left(rac{|x- ilde{x}|}{|x|}
  ight)$  عبارت است از تعداد تقریبی ارقام دهدهی یکسان x و
- از سوی دیگر طبق قضیه ی ۱۰۲۰۱ میدانیم که خطای نسبی حاصل از گردکردن هر عددی حقیقی به یک عدد ماشین با ایسیلون ماشین کراندار می شود.
  - به یاد آورید که  $-\log_{10}(\varepsilon_M)$  در قالب دوگانهی آی-تریپل-ای تقریبا برابر است با ۱۶.

پس به عنوان یک قاعده ی سردستی، چنانچه عدد وضعیت یک مساله برابر با  $10^k$  باشد آنگاه تعداد ارقام درستی که میتوان (با دقت دوگانه) از خروجی مساله انتظار داشت تقریبا برابر خواهد بود با k-16. اگر عدد وضعیت مساله ای  $10^8$  باشد در بهترین حالت میتوان حدود k رقم درست از خروجی مساله انتظار داشت و یا اگر عدد وضعیت مساله ای  $10^{16}$  باشد اصولا تلاش برای حل عددی آن مساله بیهوده است.

مثال ۱۰۴۰۱، در مورد میزان حساسیت مسئله ی ارزیابی تابع  $f(x) = \exp(-x)$  به خطاهای اندک دن) بحث کنید. (گردکردن) بحث کنید

داريم

$$f'(x) = -\exp(-x) \to \kappa_f(x) = |\frac{x \exp(-x)}{\exp(-x)}| = |x|.$$

پس برای  $x \approx 10^n$  وقتی  $x \approx 10^n$  وقتی  $x \approx 10^n$  این مساله برای  $x \approx 10^n$  عددی بزرگ باشد بدوضع میباشد.

مثال ۲.۴.۱. در مورد میزان حساسیت مسئله ی ارزیابی تابع  $f(x) = \log(x)$  به اختلالی کوچک در نزدیکی  $x \approx 1$ 

داريم

$$f'(x) = \frac{1}{x} \to \kappa_f(x) = \left| \frac{xf'(x)}{f(x)} \right| = \left| \frac{x\frac{1}{x}}{\log(x)} \right| = \left| \frac{1}{\log(x)} \right|.$$

پس برای x pprox 1 چون  $\log(x)$  عددی کوچک خواهد بود، مساله بدوضع است.

نکته ی ۱.۴۰۱. معمولا هرگاه از عدد وضعیت صحبت می شود منظور عدد وضعیت نسبی است. با این حال عدد وضعیت مطلق که معمولا با نماد  $\hat{\kappa}_f(x)$  نشان داده می شود نیز گاها مفید است (مثلا آنچنان که در مسئله ی ریشه یابی در فصل بعد خواهیم دید).

$$\underbrace{f(x)-f(\tilde{x})}_{\text{cept}}=\underbrace{\frac{f(x)-f(\tilde{x})}{x-\tilde{x}}}_{\text{cept}} \times \underbrace{x-\tilde{x}}_{\text{cept}}$$
 اختلال مطلق ورودی تقریب مشتق تابع  $f$  در نقطه ی

پس عدد وضعیت مطلق مسئله ی ارزیابی تابع f در نقطه ی x برابر است با

$$\hat{\kappa}_f(x) = |f'(x)|.$$

## ۱۰۴۰۱ پدیدهی حذف

منظور از حذف $^{\prime}$ ، خنثی شدن ارقام بامعنا در حساب با دقت متناهی است وقتی حاصل تفریق دو عدد همعلامت ِ نزدیک به هم (اعداد با ارقام پیشروی یکسان) محاسبه می شود. همین اتفاق در جمع با دقت متناهی دو عدد با علامتهای متضاد که از نظر اندازه به هم نزدیک هستند نیز رخ می دهد. فرض کنید k رقم پیشرو از ارقام بامعنای دو عدد ماشین x و y یکسان باشند و بخواهیم حاصل تفریق آنها را در حساب ممیز شناور بیابیم. داریم:

$$(d_0.d_1d_2\cdots d_{k-1}\ d_k\ d_{k+1}\cdots d_{p-1})_{\beta}\times\beta^e$$

$$\ominus (d_0.d_1d_2\cdots d_{k-1}\ c_k\ c_{k+1}\cdots c_{p-1})_{\beta}\times\beta^e$$

$$= (0.00\cdots 0\ f_k\ f_{k+1}\cdots f_{p-1})_{\beta}\times\beta^e$$

که وقتی نرمال میشود برابر خواهد بود با

$$(f_k.f_{k+1}\cdots f_{p-1}??\cdots?)_{\beta}\times\beta^{e-k}$$

پس k رقم داریم که بجایشان هرچه بگذاریم نادرست خواهند بود. پس دو عدد x و y دارای p رقم بامعنا بودند ولی حاصل تفریق آنها تنها دارای p-k رقم بامعنای درست است و k رقم پایانی آن صفرهایی خواهند بود که به صورت مصنوعی جلوی مانتیس قرار گرفته اند.

در زنجیرهای از محاسبات علمی با حساب ممیز شناور قویا ممکن است x و y خود تقریبهایی از y عدادی دیگر بوده باشند. در چنین موقعیتی قاعدتا امکان بیشتری وجود دارد که ارقام پیشروی x و y درست بوده باشند و ارقام پایانی آنها از اعداد اصلی متفاوت و در نتیجه غلط بوده باشند چرا که ارقام پایانی یک عدد در محاسبات ممیزشناور بیشترین تاثیر را از خطاهای گردکردن می پذیرند. آنچه نگرانی از پیش تشدید می کند این است که y رقم بامعنایی که واقعا حاصل اجرای یک

cancellation\

محاسبه بودند، دقیقا از همان ارقامی از x و y به دست آمدهاند که احتمال نادرست بودنشان بیشتر بوده است. پس حتی به درستی آن رقمهای غیرمصنوعی نیز چندان نمی توان خوشبین بود! در واقع گاهی بجای حذف (خنثی شدن) از نام پدیده ی ((حذف فاجعه بار ()) استفاده می شود.

بعنوان یک مثال عددی به یاد آورید که در قسمتی از حل تمرین ۲ قرار بود دو عدد

$$b = +3.3678429 \times 10^{+1}$$

$$c = -3.3677811 \times 10^{+1}$$

را در دستگاه  $F_{10,8}^{-5,+5}$  با هم جمع کنیم. داریم:

$$3.3678429 \times 10^{+1}$$

 $\ominus 3.3677811 \times 10^{+1}$ 

 $= 0.0000618 \times 10^{+1}$ 

و با نرمال کردن آن، پاسخ c = 0 دارای هشت رقم قبل نرمال کردن آن، پاسخ c = 0 دارای هشت رقم قبل نرمال کردن آن، پاسخ c = 0 دارای سه رقم بامعنا هستند ولی حاصل جمع، قبل از نرمال سازی تنها دارای سه رقم بامعنا ست چرا که پنج رقم پیشروی c = 0 با هم خنثی شدهاند. پس دو عددی که هشت رقم بامعنا دارند را با هم جمع کرده و نهایتا بعد از نرمال سازی پاسخی به دست آورده ایم که تنها سه رقمش درست است.

قبل از پایان بحث پدیده ی حذف، خوب است ارتباط احتمالی بین خطرناک بودن پدیده ی حذف با عدد وضعیت مسئله ی ریاضی ِ تفریق دو عدد را بررسی کنیم. مسئله ی محاسبه ی مقدار تابع f(x) = x - c که در آن x - c عددی ثابت است را درنظر بگیرید. داریم:

$$\kappa_f(x) = \frac{|x|(1)}{|x-c|} = |\frac{x}{x-c}|$$

اگر x و c خیلی به هم نزدیک باشند، مخرج کسر بالا خیلی به صفر نزدیک شده و در نتیجه عدد وضعیت، خیلی بزرگ خواهد شد. در غیر این صورت مسئله ی تفریق دو عدد، یک مسئله ی بدوضع نخواهد بود. پس ناخوشایند بودن پدیده ی حذف شگفت آور نیست: این متناظر با مساله ای بدوضع است.

catastrophic cancellation

 $a\oplus (b\oplus c)$  و هم در محاسبه و  $(a\oplus b)\oplus c$  تمرین ۳. بار دیگر تمرین ۲ را مرور کنید. هم در محاسبه و  $a\oplus (b\oplus c)$  و هم در محاسبه و  $a\oplus (b\oplus c)$  و هم در کدام قسمت از فرمول  $a\oplus (b\oplus c)$  پدیده و خذف رخ می دهد. و توضیحات قبل مشخص کردیم که در کدام قسمت از فرمول و پدیده و باید و نام درخ می دهد.

- در کدام قسمت از فرمول دیگر یعنی  $(a\oplus b)\oplus c$  یعنی دخف رخ می دهد؟ . ۱
- ۲. کدامیک از این دو پدیده ی حذف ذکرشده مخربتر است؟ (مثلا با تعیین عدد وضعیت هر یک از دو رخداد حذف)

این امکان وجود دارد که پدیده ی حذف تاثیر مخرب زیادی نداشته باشد. ممکن است بتوان به سادگی آنرا از بین برد و یا ممکن است به راحتی قابل برطرفکردن نباشد. اما باتوجه به فراوانی عمل تفریق در مسائل و الگوریتمهای محاسبات علمی، بهتر است در زمان طراحی الگوریتمهای عددی مواظب رخداد آن بوده و حتی الامکان از آن پیشگیری کنیم. بعدا در بحث پایداری عددی الگوریتمها، پدیده ی حذف را در فرمول دبیرستانی دلتا برای یافتن ریشه ی چندجملهایهای درجه دو شناسایی و راهکار جلوگیری از آن را خواهیم دید.