به نام پروردگار یگانه و یکتا

دانشکده علوم ریاضی

آنالیز عددی پیشرفته

گردآورنده : رضا مختاری

ویرایش ششم : پاییز ۱۴۰۰

فهرست مطالب

فصل ۱

مفاهیم مقدماتی

دی رفت و باز نیاید

فردا را اعتماد نشاید

حال را غنیمـت دان

که دیـــــر نپایـــــد

خواجه عبدا... انصاری

در این فصل پس از ارایه مراجع و تاریخچه‌ای از آنالیز عددی، مفاهیم، تعاریف و قضایایی که در ادامه کار به آن‌ها نیاز داریم را مرور می‌کنیم.

۱.۱ مراجع آنالیز عددی

سه برخورد متفاوت با سرفصل آنالیز عددی عبارتند از

- کاربردی و الگوریتمی (روش‌گرایانه)

۱. Introductory Computer Methods & NA, Pennington, 1970 ،

۲. Introduction to NA: An Algorithmic Approach, Deboor, 1972 ،

۳. Numerical Recipes: The Art of Scientific Computing, Press & Teukolsky & Vetterling & Flannery, 2007،

۴. NA for Engineers and Scentists, Miller, 2014 .

- نظری (ایده‌گرایانه)

۱. Theoretical NA, Wendroff, 1967 ،

۲. Theoretical NA, Linz, 1979 ،

۳. A theoretical Introduction to NA, Ryaben'kii & Tsynkov, 2007 ،

۴. Theoretical NA: A Functional Analysis Framework, Atkinson & Han, 2009 .

- تلفیقی از دو دیدگاه قبلی

۱. Elements of NA, Henrichi, 1964 ،

۲. A First Course in NA, Ralston & Rabinowits, 1978 ،

۳. An Introduction to NA, Atkinson, 1989 ،

۴. NA: Mathematics of Scientific Computing, Kincaid & Cheney, 1991 ،

۵. Numerical Mathematics, Quarteroni & Sacco & Saleri, 2007 .

۶. Introduction to NA, Stoer & Bulirsch, 1993 & 2002 & 2008 ،

۷. An Introduction to Numerical Methods and Analysis, Epperson, 2013 ،

۸. NA, Burdan & Faires, 2016 ،

۹. A Consice Introduction to NA, Faul, 2016 .

۲.۱ تاریخچه آنالیز(محاسبات) عددی

آن چه در ادامه می‌آید و از آن به عنوان تاریخچه آنالیز(محاسبات) عددی یاد شده، وقایعی است که بی‌شک با آنالیز عددی در ارتباط بوده و در گسترش آن بی‌تاثیر نبوده‌اند. خواننده علاقمند می‌تواند تاریخچه کامل را در کتاب‌های تاریخ ریاضیات دنبال کند.

۱. تهیه نتایج به صورت اعداد (بشر اولیه!؟)

۲. کارهای بابلی‌ها و مصری‌ها در رابطه با نجوم و مهندسی راه و ساختمان

- لوح بابلی حدود ۲۰۰۰ سال قبل از میلاد (مجذور اعداد صحیح از ۱ تا ۶۰)
- ثبت کسوف و خسوف از حدود ۷۵۰ سال قبل از میلاد
- توجه مصریان به کسر و ابداع روش نابجایی

۳. کارهای ارشمیدس در حدود ۲۲۰ سال قبل از میلاد ($۳\frac{۱۰}{۷۱} < \pi < ۳\frac{۱}{۷}$)

۴. ابداع روش تکراری محاسبه $\sqrt{a}$ توسط هِرون در حدود ۱۰۰ سال قبل از میلاد

۵. کارهای فیثاغورسیان در رابطه با مجموع عددی سری‌ها

۶. روش دیوفانتوس برای حل معادله درجه دوم در حدود سال ۲۵۰ میلادی

۷. نمادگذاری عددی عربی (هندی)

۸. ساخت جدول‌های مثلثاتی قبل از قرن ۱۰

۹. پیدایش جبر در قرن ۱۶

۱۰. قرن ۱۷

- انقلاب صنعتی (کارهای نیوتن، اویلر، لاگرانژ، گاوس و بسل)
- ساخت جدول لگاریتم توسط نپر
- اختراع خط‌کش محاسبه توسط اوترد
- اختراع ماشین‌حساب توسط پاسکال و لایب‌نیتز
- محاسبات با سری‌های نامتناهی

۱۱. پایه‌ریزی تفاضلات متناهی توسط ژاکوب استرلینگ و بروک تیلور در اوایل قرن ۱۸

۱۲. پیش‌بینی وجود و موضع سیاره نپتون توسط آدامز و لوریه در سال ۱۸۴۵ میلادی

۱۳. پیشرفت و تولید وسیع ماشین‌حساب در اواخر قرن ۱۹

۱۴. چاپ کتاب آنالیز ریاضی عددی[۱] توسط اسکاربورو [۲۱] در سال ۱۹۳۰ میلادی

۱۵. جنگ جهانی دوم و پیدایش رایانه در سال ۱۹۴۰ میلادی

۱۶. معرفی واژه Numerical Analysis توسط UCLA در سال ۱۹۴۷ میلادی

[۱] Numerical mathematical analysis

۳.۱ مفاهیم اساسی

در این بخش پس از بررسی چند تعریف از آنالیز عددی، مفاهیمی اساسی و مرتبط با آنالیز عددی را مطرح می‌کنیم.

تعریف ۱.۱ *تعاریف متنوع آنالیز عددی عبارتند از*

۱. *آنالیز عددی نظریه روش‌های ساخت‌یافته (سازنده)[۲] در آنالیز ریاضی، آنالیز تابعی، جبر، جبرخطی و سایر شاخه‌های ریاضی است[۱۲].*

۲. *آنالیز عددی آنالیز الگوریتم‌های پیوسته است [۲۳].*

۳. *آنالیز عددی مطالعه کمی جواب تقریبی مسایل ریاضی با توجه به خطاها و کران‌های آن‌ها است [۲۵].*

۴. *آنالیز عددی به استنتاج ریاضی، توصیف و تحلیل روش‌هایی می‌پردازد که جواب‌های عددی مسایل ریاضی را تولید می‌کنند [۱].*

۵. *آنالیز عددی مطالعه روش‌های تقریبی و دقت آن‌ها است [۶].*

۶. *آنالیز عددی شامل مطالعه، توسعه و تجزیه و تحلیل الگوریتم‌ها برای به دست آوردن جواب‌های عددی مسایل مختلف ریاضی است. آنالیز عددی ریاضیات محاسبات علمی است [۱۴].*

۷. *آنالیز عددی مطالعه جواب تقریبی مسایل ریاضی با در نظر گرفتن حد و اندازه خطاهای ممکن است [۳].*

۸. *آنالیز عددی مطالعه الگوریتم‌ها برای مسایل ریاضیات پیوسته (متمایز از ریاضیات گسسته) است [۲۴].*

آتکینسون[۳] آنالیز عددی‌دان معاصر، برای شناخت یک موضوع دیدگاه‌های متفاوت زیر را مطرح می‌کند.

۱. (مورد قدیمی) نظری (دیدگاه دانشمندان)

۲. (مورد قدیمی) تجربی (دیدگاه مهندسان)

۳. (مورد جدید) محاسباتی (دیدگاه متخصصان آنالیز عددی)

به عبارتی وی علوم محاسباتی را به علوم نظری و علوم تجربی اضافه کرده و معتقد است این علوم به صورت زیر با هم در ارتباط هستند.

$$\text{علوم نظری} \rightleftarrows \text{علوم محاسباتی} \rightleftarrows \text{علوم تجربی} \rightleftarrows \text{علوم نظری}$$

الگوریتم واژه‌ای است که در آنالیز (محاسبات) عددی بسیار مورد استفاده قرار می‌گیرد و در ادامه به تعریف آن می‌پردازیم.

تعریف ۲.۱ *الگوریتم در آنالیز عددی،*

- *به عنوان توصیفی کامل و بدون ابهام از روش ساختن جواب یک مسئله ریاضی تعریف می‌شود [۱].*
- *فرآیند بدون ابهامی است که دنباله‌ای متناهی از گام‌هایی که با ترتیب مشخصی اجرا می‌شوند را تشریح می‌کند [۲].*

در آنالیز عددی فقط با الگوریتم‌های عددی سر و کار داریم. الگوریتم‌های عددی، الگوریتم‌هایی هستند که خروجی آن‌ها جواب عددی یا تقریبی یک مسئله است. الگوریتم عددی همگرا، به آن دسته از الگوریتم‌های عددی گفته می‌شود که در هر تکرار، عناصر دنباله‌ای را تولید می‌کنند که آن دنباله به جواب واقعی مسئله همگرا باشد. باید توجه داشت که ماهیت این دنباله‌ها وابسته به جواب مسئله است و ممکن است عدد، تابع، بردار، ماتریس و

[۲] Constructive methods : روشی که علاوه بر وجود، راه رسیدن به جواب را نیز مشخص می‌کند.
[۳] K.E. Atkinson

غیره باشد. از شبه‌کد[۴] برای توصیف الگوریتم استفاده می‌شود. شبه‌کد را می‌توان زبانی بین زبان‌های سطح بالای رایانه‌ای و زبان محاوره تلقی کرد. هنگام مواجه شدن با یک مسئله حقیقی، سه سوال بحث‌برانگیزی که برای یک متخصص آنالیز عددی مطرح می‌شود، عبارتند از:

۱. آیا جواب مسئله وجود دارد؟

۲. آیا جواب یکتا است؟

۳. حساسیت جواب در برابر اختلالات جزئی (کوچک) داده‌های ورودی در چه حد است؟

تعریف ۳.۱ *فرض کنید d و $S(d)$ به ترتیب بیان‌گر داده‌های ورودی و جواب متناظر با داده‌ها، برای مسئله‌ای دلخواه باشند[۵]. هم‌چنین فرض کنید $d+\delta d$ داده‌های ورودی اختلال‌یافته (تغییر داده‌شده) و $S(d+\delta d)$ جواب نظیر آن باشد. معیار اندازه‌گیری تغییرات اعمال‌شده در داده‌ها و اختلاف جواب‌های متناظر را به ترتیب با $\|\delta d\|$ و $\|S(d+\delta d)-S(d)\|$ که اعداد نامنفی هستند، نشان می‌دهیم. به ازای داده‌های مفروض d، مسئله خوش‌طرح (قیافه)[۶] نامیده می‌شود اگر دو شرط زیر برقرار باشد*

الف) به ازای تمام داده‌های متعلق به یک همسایگی حول d جواب یکتا موجود باشد، یعنی عددی مانند $\epsilon > 0$ چنان موجود باشد که به ازای هر δd که $\|\delta d\| < \epsilon$ مقدار $S(d+\delta d)$ موجود و یکتا باشد.

ب) جواب یعنی $S(d)$ به طور پیوسته به داده‌ها یعنی d بستگی داشته باشد یعنی اگر $\|\delta d\| \to 0$ آن‌گاه داشته باشیم $\|S(d+\delta d)-S(d)\| \to 0$.

تذکر ۱.۱ *در صورتی که به ازای مجموعه‌ای از داده‌ها، بیش از یک جواب موجود باشد، نمی‌توانیم تشخیص دهیم روش عددی ما کدام جواب را تولید می‌کند. ممکن است اطلاعات حقیقی (تجربی) پیرامون مسئله، مشخص کند که جواب مسئله، موجود و یکتا است. اگر متخصص در این مرحله به جواب نرسید، مدل ریاضی مسئله را مورد بررسی قرار داده و برای تعیین وجود و یکتایی جواب در مباحث نظری در جستجوی یافتن قضایایی است که تکلیف وجود و یکتایی جواب را روشن می‌کنند. به طور معمول با افزودن شرایطی، می‌توان جواب‌های اضافی را کنار گذاشت. به عنوان مثال، در ریشه‌یابی یک معادله جبری درجه دو یا بیشتر، می‌توان هدف را تعیین ریشه با بزرگ‌ترین اندازه قرار داد.*

تذکر ۲.۱ *اگر شرط ب برقرار نباشد، استفاده از یک الگوریتم عددی کار دشواری است، زیرا در این صورت داده‌هایی بسیار نزدیک به d وجود دارند که جواب‌های متناظر آنها متمایز از $S(d)$ هستند.*

تعریف ۴.۱ *مسئله‌ای که به ازای مجموعه‌ای از داده‌ها خوش‌طرح است، خوش‌حالت (وضع)[۷] نامیده می‌شود اگر هر تغییر جزئی در داده‌های ورودی منجر به تغییر کوچکی در جواب شود و در صورتی که تغییر جواب فاحش باشد، مسئله را بدحالت (وضع)[۸] (مریض حال) نامند.*

تذکر ۳.۱ *مفاهیم خوش‌طرحی (بدطرحی) و خوش‌حالت (بدحالت) بودن نسبی هستند و نه تنها به مسئله بلکه به داده‌ها نیز بستگی دارند.*

تعریف ۵.۱ *جواب یک مسئله خوش‌طرح اغلب بستگی لیپشیتس به داده‌ها دارد یعنی اعداد $L > 0$ و $\epsilon > 0$ موجود هستند به طوری که به ازای هر δd که $\|\delta d\| < \epsilon$ خواهیم داشت*

$$\|S(d+\delta d)-S(d)\| \le L\,\|\delta d\|\,.$$

---

[۴] Pseudocode

[۵] ماهیت d و $S(d)$ بر حسب نوع مسئله، ممکن است عدد، تابع، بردار، ماتریس و غیره باشد.

[۶] Well-posed

[۷] Well-conditioned

[۸] Ill-conditioned

L به عدد حالت (وضعیت)[۹] معروف است و اگر L خیلی بزرگ نباشد، مسئله خوش‌حالت است. یک مثال جالب در برنامه ill-conditioned.nb *بررسی شده است.*

تعریف ۶.۱ *اگر هنگام به کار بردن یک الگوریتم در حل یک مسئله خوش‌حالت، اختلال کوچکی در داده‌های ورودی، منجر به تولید تغییرات بزرگی در نتایج نهایی گردد، آن الگوریتم ناپایدار[۱۰] و در غیر این صورت پایدار[۱۱] نامیده می‌شود. اگر الگوریتمی به ازای انتخاب مشخصی از داده‌ها پایدار گردد، آن را به طور مشروط پایدار[۱۲] نامند. برنامه* stable.nb *را ببینید.*

۴.۱ پیش‌نیازهای ریاضی

در این بخش به معرفی اصطلاحات و بیان قضایای مورد نیاز می‌پردازیم.

تعریف ۷.۱ *با این فرض که $n \in \mathbb{N}_\circ$ و $X \subseteq \mathbb{R}$، چند فضای پرکاربرد از توابع به صورت زیر تعریف می‌شوند.*

$$C(X) = \{f : \Omega \to \mathbb{R} | \text{پیوسته باشد } \Omega \text{ در } f\}$$

$$C[a,b] = \{f : [a,b] \to \mathbb{R} | \text{پیوسته باشد } [a,b] \text{ در } f\}$$

$$C^n(X) = \{f : \Omega \to \mathbb{R} | \text{تا مشتق مرتبه } n \text{ ام } f \text{ در } \Omega \text{ پیوسته باشد}\}$$

$$C^n[a,b] = \{f : [a,b] \to \mathbb{R} | \text{پیوسته باشد } [a,b] \text{ در } f \text{ ام } n \text{ تا مشتق مرتبه}\}$$

$$C^\infty(\Omega) = \{f : X \to \mathbb{R} | \text{مشتق } f \text{ از هر مرتبه‌ای در } \Omega \text{ پیوسته باشد}\}$$

$$C^\infty[a,b] = \{f : [a,b] \to \mathbb{R} | \text{مشتق } f \text{ از هر مرتبه‌ای در } [a,b] \text{ پیوسته باشد}\}$$

در $\Omega \subseteq \mathbb{R}^d$ که در آن $d > ۱$، تعاریف متناظری تعمیم داده می‌شود.

تذکر ۴.۱ *توابع چندجمله‌ای، گویا، مثلثاتی، نمایی و لگاریتمی در رده $C^\infty(\Omega)$ قرار دارند، که در آن Ω دامنه تعریف این توابع است.*

قضیه ۱.۱ *فرض کنید $f \in C[a,b]$. f در $[a,b]$، ماکزیمم و مینیمم مطلق خود را اختیار می‌کند.*

قضیه ۲.۱ (مقدار میانی) *فرض کنید $f \in C[a,b]$ و $f(a) \neq f(b)$. f هر مقدار بین $f(a)$ و $f(b)$ را در نقطه‌ای بین a و b به خود می‌گیرد.*

قضیه ۳.۱ *فرض کنید $f \in C[a,b]$ و f بر (a,b) مشتق‌پذیر باشد. اگر $f(a) = f(b) = k$ آن‌گاه عدد x در (a,b) موجود است که $f'(x) = ۰$. این قضیه برای $k = ۰$، به قضیه رُل[۱۳] معروف است.*

قضیه ۴.۱ (تعمیم قضیه رُل) *فرض کنید $f \in C[a,b]$ و تا مشتق مرتبه n ام f بر (a,b) موجود باشد. اگر $n+۱$ نقطه $x_۰, \ldots, x_n$، متمایز در $[a,b]$ باشند و $f(x_۰) = f(x_۱) = \cdots = f(x_n)$، آن‌گاه عدد x در (a,b) موجود است به طوری که $f^{(n)}(x) = ۰$.*

---

[۹] Condition-number
[۱۰] Unstable
[۱۱] Stable
[۱۲] Conditionally stable
[۱۳] Rolle

قضیه ۵.۱ (مقدار میانگین) *فرض کنید $f \in C[a,b]$ و f بر (a,b) مشتق‌پذیر باشد. نقطه‌ای مانند c در (a,b) موجود است به طوری که*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

قضیه ۶.۱ (کوشی[۱۴]) *فرض کنید $f, g \in C[a,b]$ و f و g بر (a,b) مشتق‌پذیر باشند و g' در (a,b) ناصفر باشد. نقطه‌ای مانند c در (a,b) موجود است به طوری که*

$$\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

قضیه ۷.۱ (مقدار میانگین برای انتگرال) *فرض کنید $f \in C[a,b]$. نقطه‌ای مانند c در $[a,b]$ موجود است به طوری که*

$$\int_a^b f(x)dx = (b-a)f(c).$$

قضیه ۸.۱ (مقدار میانگین تعمیم‌یافته برای انتگرال) *فرض کنید $f \in C[a,b]$ و g تابعی انتگرال‌پذیر باشد که در $[a,b]$ تغییر علامت ندهد. نقطه‌ای مانند c در $[a,b]$ موجود است به طوری که*

$$\int_a^b f(x)g(x)dx = f(c)\int_a^b g(x)dx.$$

قضیه ۹.۱ (اساسی حساب دیفرانسیل و انتگرال) *فرض کنید $f \in C[a,b]$ و $F(x) = \int_a^x f(t)dt$. F در (a,b) مشتق‌پذیر بوده و $F'(x) = f(x)$.*

قضیه ۱۰.۱ (تیلور با باقیمانده لاگرانژ) *فرض کنید $f \in C^n[a,b]$ و $f^{(n+1)}$ بر (a,b) موجود باشد و $x_0 \in [a,b]$. در این صورت، به ازای هر $x \in [a,b]$، نقطه‌ای مانند $\xi(x)$ بین x_0 و x وجود دارد که*

$$f(x) = P_n(x) + R_n(x, x_0)$$

که در آن

$$P_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!}(x - x_0)^i$$

و

$$R_n(x, x_0) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)^{n+1}.$$

در این‌جا $P_n(x)$ چندجمله‌ای تیلور مرتبه nام f حول x_0 و $R_n(x, x_0)$ جمله باقی‌مانده (یا خطای برش[۱۵]) متناظر با $P_n(x)$ نامیده می‌شود. اگر $n \to \infty$ آن‌گاه $P_n(x)$ به یک سری بی‌پایان تبدیل می‌شود که به آن سری تیلور f حول x_0 گویند. در این حالت، شرط بی‌نهایت بار مشتق‌پذیر بودن f در x_0 لازم است.

---

[۱۴] Cauchy

[۱۵] Truncation error

تذکر ۵.۱ *در قضیه قبل اگر $x_\circ = \circ$ آن‌گاه واژه تیلور به مکلورن تبدیل می‌شود.*

تذکر ۶.۱ *$R_n(x, x_\circ)$ مقدار خطا در استفاده از P_n به جای f را نشان می‌دهد. در عمل با یافتن کرانی برای جمله باقی‌مانده، در واقع برای خطای تقریب f با P_n کرانی پیدا می‌کنیم.*

تذکر ۷.۱ *ویژگی مهم چندجمله‌ای تیلور مرتبه nاُم آن است که P_n و مشتقات تا مرتبه nاُم آن با f و مشتقات تا مرتبه nاُم آن در نقطه $x_\circ$ برابر هستند.*

تذکر ۸.۱ *شکل دیگر (کاربردی) قضیه تیلور به صورت زیر است*

$$f(x+h) = f(x) + f'(x)h + \cdots + \frac{f^{(n)}(x)}{n!}h^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1} = \sum_{i=\circ}^{n} \frac{f^{(i)}(x)}{i!}h^i + \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1}.$$

و یا

$$f(x_\circ) = f(x+h) = f(x) + f'(x)h + \cdots + \frac{f^{(n)}(x)}{n!}h^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1} = \sum_{i=\circ}^{n} \frac{f^{(i)}(x)}{i!}h^i + \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1},$$

که در آن $h = x_\circ - x$.

مثال ۱.۱ *فرض کنید $f(x) = ۲ + ۴x - x^۳$. آن‌گاه به وضوح برای $n \geq ۳$ داریم*

$$P_n(x) = ۲ + ۴x - x^۳, \qquad R_n(x, \circ) = \circ,$$

و برای $n = ۲$ خواهیم داشت

$$P_۲(x) = ۲ + ۴x, \qquad R_۲(x, \circ) = -x^۳,$$

و برای $n = ۱$ می‌توان نوشت

$$P_۱(x) = ۲ + ۴x, \qquad R_۱(x, \circ) = -۳x^۲\xi(x),$$

که در آن $\xi(x)$ نقطه‌ای بین $\circ$ و x است. △

قضیه ۱۱.۱ (تیلور با باقیمانده انتگرالی) *اگر $f \in C^{n+۱}[a, b]$ آن‌گاه به ازای هر $x, x_\circ \in [a, b]$ داریم*

$$f(x) = \sum_{i=\circ}^{n} \frac{f^{(i)}(x_\circ)}{i!}(x - x_\circ)^i + R_n(x, x_\circ)$$

که در آن

$$R_n(x, x_\circ) = \frac{۱}{n!}\int_{x_\circ}^{x} f^{(n+۱)}(t)(x-t)^n dt.$$

برهان. به کمک انتگرال‌گیری جزء به جزء به ازای هر n داریم

$$R_n(x) = -\frac{1}{n!} f^n(x_\circ)(x - x_\circ)^n + R_{n-1}(x).$$

با استفاده مکرر از این رابطه خواهیم داشت

$$R_n(x) = -\sum_{i=1}^{n} \frac{f^{(i)}(x_\circ)}{i!}(x - x_\circ)^i + R_\circ(x).$$

حکم با توجه به رابطه زیر برقرار است

$$R_\circ(x) = \int_{x_\circ}^{x} f'(t)dt = f(x) - f(x_\circ).$$

□

قضیه ۱۲.۱ (تیلور توابع دومتغیره) *اگر $f \in C^{n+1}(\Omega)$ که در آن $\Omega = [a,b] \times [c,d] \subseteq \mathbb{R}^2$، آن‌گاه به ازای هر نقطه $(x+h, y+k)$ در Ω داریم*

$$f(x+h, y+k) = \sum_{i=0}^{n} \frac{1}{i!} \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} \right)^i f(x,y) + R_n(h,k),$$

که در آن

$$R_n(h,k) = \frac{1}{(n+1)!} \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} \right)^{n+1} f(x+\theta h, y+\theta k),$$

و θ عددی بین ۰ و ۱ است. هم‌چنین داریم

$$\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} \right)^0 f(x,y) = f(x,y),$$
$$\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} \right)^1 f(x,y) = \left(h\frac{\partial f}{\partial x} + k\frac{\partial f}{\partial y} \right)(x,y),$$
$$\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} \right)^2 f(x,y) = \left(h^2\frac{\partial^2 f}{\partial x^2} + 2hk\frac{\partial^2 f}{\partial x \partial y} + k^2\frac{\partial^2 f}{\partial y^2} \right)(x,y), \ldots.$$

تذکر ۹.۱ *این قضیه برای توابع چندمتغیره به سادگی تعمیم‌پذیر است.*

هنگامی که قصد داریم دو دنباله را با هم مقایسه کنیم از تعریف زیر استفاده می‌کنیم.

تعریف ۸.۱ *اگر $\{\alpha_n\}$ و $\{\beta_n\}$ دو دنباله همگرا به x باشند، آن‌گاه بنابر تعریف داریم*

$$\forall \epsilon > 0,\ \exists N_1 \ :\ \forall n\,(n \geq N_1 \to |\alpha_n - x| < \epsilon),$$
$$\forall \epsilon > 0,\ \exists N_2 \ :\ \forall n\,(n \geq N_2 \to |\beta_n - x| < \epsilon).$$

اگر به ازای هر $\epsilon > 0$ داشته باشیم $N_1 < N_2$، آن‌گاه $\{\alpha_n\}$ سریعتر از $\{\beta_n\}$ به x همگرا است.

متاسفانه چون یافتن N به سادگی امکان‌پذیر نیست، این تعریف بیشتر جنبه نظری دارد و در عمل نمی‌توان از آن سود برد. خوشبختانه در عمل تعریف زیر راه‌گشا است.

تعریف ۹.۱ *فرض کنید $\{x_n\}$ و $\{\alpha_n\}$ دو دنباله متفاوت باشند. می‌نویسیم*

$$x_n = O(\alpha_n),$$

و می‌خوانیم x_n اُی بزرگ α_n است هرگاه ثابت‌های C و N چنان موجود باشند که داشته باشیم

$$|x_n| \leq C\,|\alpha_n|\,, \qquad \forall n \geq N.$$

اگر برای هر n داشته باشیم $\alpha_n \neq ۰$ آن‌گاه می‌توان گفت نسبت $|x_n/\alpha_n|$ کران‌دار (با کران C) باقی می‌ماند همان‌طور که $n \to \infty$. هر چند که در این تعریف $\{\alpha_n\}$ یک دنباله دلخواه است ولی در عمل، بیشتر مواقع $\alpha_n = \frac{۱}{n^p}$ اختیار می‌شود که در آن $p > ۰$ و در صدد یافتن بزرگ‌ترین مقدار p هستیم به طوری که

$$x_n = O\left(\frac{۱}{n^p}\right).$$

گاهی مواقع می‌نویسیم

$$x_n = o(\alpha_n),$$

و می‌خوانیم x_n اُی کوچک α_n است هرگاه داشته باشیم $\lim_{n\to\infty}(x_n/\alpha_n) = ۰$.

تذکر ۱۰.۱ *بیشتر مواقع دو دنباله بیان‌شده در تعریف قبل همگرا به صفر هستند و اگر در این صورت داشته باشیم $x_n = O(\alpha_n)$ آن‌گاه می‌توان گفت همگرایی x_n به صفر متناسب با همگرایی α_n به صفر است و یا همگرایی x_n به صفر مانند همگرایی α_n به صفر است در حالی که اگر داشته باشیم $x_n = o(\alpha_n)$ آن‌گاه می‌توان گفت همگرایی x_n به صفر سریعتر از همگرایی α_n به صفر است.*

مثال ۲.۱ *برای $n \in \mathbb{N}_۰$ فرض کنید*

$$\alpha_n = \frac{n+۱}{n^۲} \quad \text{و} \quad \hat{\alpha}_n = \frac{n+۳}{n^۳}.$$

اگرچه $\lim_{n\to\infty}\alpha_n = ۰ = \lim_{n\to\infty}\hat{\alpha}_n$ ولی $\{\hat{\alpha}_n\}$ سریعتر از $\{\alpha_n\}$ به صفر همگرا است. به جدول زیر نگاه کنید.

n	۱	۲	۳	۴	۵	۱۰۰
α_n	۲/۰۰۰۰۰	۰/۷۵۰۰۰	۰/۴۴۴۴۴	۰/۳۱۲۵۰	۰/۲۴۰۰۰	۰/۰۱۰۱
$\hat{\alpha}_n$	۴/۰۰۰۰۰	۰/۶۲۵۰۰	۰/۲۲۲۲۲	۰/۱۰۹۳۸	۰/۰۶۴۰۰	۰/۰۰۰۱۰۳

اگر قرار دهیم $\beta_n = ۱/n$ و $\hat{\beta}_n = ۱/n^۲$ آن‌گاه

$$|\alpha_n - ۰| = \frac{n+۱}{n^۲} \leq \frac{n+n}{n^۲} = ۲\frac{۱}{n} = ۲\beta_n,$$

و

$$|\hat{\alpha}_n - ۰| = \frac{n+۳}{n^۳} \leq \frac{n+۳n}{n^۳} = ۴\frac{۱}{n^۲} = ۴\hat{\beta}_n,$$

بنابراین

$$\alpha_n = ۰ + O\left(\frac{۱}{n}\right) \quad \text{و} \quad \hat{\alpha}_n = ۰ + O\left(\frac{۱}{n^۲}\right).$$

هم‌چنین می‌توان نوشت $\hat{\alpha}_n = o(\alpha_n)$. △

مثال ۳.۱ *بسط تیلور* $\ln(x)$ *حول* $x_\circ = ۱$ *به صورت زیر است*

$$\ln(x) = \ln(۱ + (x-۱)) = (x-۱) - \frac{۱}{۲}(x-۱)^۲ + \frac{۱}{۳}(x-۱)^۳ - \frac{۱}{۴}(x-۱)^۴ + \cdots.$$

در نتیجه می‌توان نوشت

$$\ln ۲ - \sum_{k=۱}^{n-۱} (-۱)^{k-۱} \frac{۱}{k} = O\left(\frac{۱}{n}\right),$$

که نمونه‌ای از همگرایی خیلی کند است. از طرف دیگر به کمک بسط مکلورن e^x *داریم*

$$e - \sum_{k=\circ}^{n-۱} \frac{۱}{k!} = O\left(\frac{۱}{n!}\right),$$

که مثالی از یک همگرایی خیلی تند است. △

تعریف ۱۰.۱ *فرض کنید* f *و* g *دو تابع دلخواه باشند. می‌نویسیم*

$$f(x) = O(g(x)), \qquad (x \to \infty),$$

هرگاه ثابت‌های C *و* r *چنان موجود باشند که داشته باشیم*

$$|f(x)| \le C\,|g(x)|, \qquad \forall x \ge r.$$

به طور کلی می‌نویسیم

$$f(x) = O(g(x)), \qquad (x \to x_\circ),$$

هرگاه ثابت C *و یک همسایگی حول* $x_\circ$ *چنان موجود باشند که به ازای هر* x *در آن همسایگی داشته باشیم*

$$|f(x)| \le C\,|g(x)|.$$

هم‌چنین

$$f(x) = o(g(x)), \qquad (x \to x_\circ),$$

به این معنی است که $\lim_{x\to x_\circ} f(x)/g(x) = \circ$.

مثال ۴.۱ *چون برای* $x \ge ۱$ *داریم* $\sqrt{x^۲+۱} \le \sqrt{۲}x$ *پس می‌توان نوشت*

$$\sqrt{x^۲+۱} = O(x), \qquad (x \to \infty).$$

△

مثال ۵.۱ *با اعمال قضیه مکلورن برای تابع* $\cos(x)$ *خواهیم داشت*

$$\cos h = ۱ - \frac{۱}{۲}h^۲ + \frac{۱}{۲۴}h^۴ \cos \xi(h),$$

که در آن $\xi(h)$ *عددی بین* ۰ *و* h *است. در نتیجه*

$$\cos h + \frac{۱}{۲}h^۲ = ۱ + \frac{۱}{۲۴}h^۴ \cos \xi(h),$$

و یا می‌توان نوشت

$$\cos h + \frac{۱}{۲}h^۲ = ۱ + O(h^۴),$$

زیرا $\frac{۱}{۲۴}h^۴ \leq \frac{۱}{۲۴}h^۴$، $\left|(\cos h + \frac{۱}{۲}h^۲) - ۱\right| = \left|\frac{۱}{۲۴}\cos \xi(h)\right| h^۴$. *در نتیجه زمانی که* $h \to ۰$، $\cos h + \frac{۱}{۲}h^۲$ *سریعتر به* ۱ *میل می‌کند تا* $h^۴$ *به صفر.* △

تمرین ۱۰.۱ *درستی روابط زیر را بررسی کنید*

$$\frac{۱}{n \ln n} = o\left(\frac{۱}{n}\right),$$

$$\sin h = h - \frac{h^۳}{۶} + O(h^۵), \qquad (h \to ۰),$$

$$e^{-۱/h} = o(h^۲), \qquad (h \to \infty).$$

تعریف ۱۱.۱ *فرض کنید* $\{x_n\}$ *دنباله‌ای همگرا به* x *باشد و برای هر* n *داشته باشیم* $x_n \neq x$. *این دنباله به طور خطی به* x *همگرا است هرگاه ثابت* $C \in (۰, ۱)$ *چنان وجود داشته باشد که*

$$\lim_{n\to\infty} \frac{|x_{n+۱} - x|}{|x_n - x|} = C.$$

در این حالت C *به نرخ همگرایی*[۱۶] *معروف است. برای* $C = ۰$ *همگرایی زیرخطی*[۱۷] *(*$C = ۱$*) (زیرخطی*[۱۸]*) نامیده می‌شود. اگر اعداد مثبت و ثابت* C *و* $p > ۱$ *وجود داشته باشند به طوری که*

$$\lim_{n\to\infty} \frac{|x_{n+۱} - x|}{|x_n - x|^p} = C$$

آن‌گاه مرتبه همگرایی[۱۹] *این دنباله* p *است.* C *به ثابت خطای مجانبی*[۲۰] *معروف است. اگر* $p = ۲$ *(*$p = ۳$*) همگرایی مربعی (مکعبی) نامیده می‌شود (*$C < ۱$ *لزومی ندارد).*

---

[۱۶] Rate of convergence
[۱۷] Superlinear
[۱۸] Sublinear
[۱۹] Order of convergence
[۲۰] Asymptotic error constant

مثال ۶.۱ *فرض کنید $\{x_n\}$ دنباله‌ای با همگرایی خطی به صفر با*

$$\lim_{n\to\infty} \frac{|x_{n+1}|}{|x_n|} = ۰/۵,$$

و $\{\hat{x}_n\}$ دنباله‌ای با همگرایی مربعی به صفر با همان ثابت خطای مجانبی باشد یعنی

$$\lim_{n\to\infty} \frac{|\hat{x}_{n+1}|}{|\hat{x}_n|^۲} = ۰/۵.$$

برای سادگی فرض می‌کنیم برای $n \geq N$ داریم

$$\frac{|x_{n+1}|}{|x_n|} \simeq ۰/۵, \qquad \frac{|\hat{x}_{n+1}|}{|\hat{x}_n|^۲} \simeq ۰/۵.$$

بنابراین می‌توان نوشت

$$|x_n - ۰| = |x_n| \simeq ۰/۵\,|x_{n-۱}| \simeq (۰/۵)^۲\,|x_{n-۲}| \simeq \cdots \simeq (۰/۵)^n\,|x_۰|\,,$$

و

$$\begin{aligned} |\hat{x}_n - ۰| = |\hat{x}_n| &\simeq ۰/۵\,|\hat{x}_{n-۱}|^۲ \simeq (۰/۵)(۰/۵\,|\hat{x}_{n-۲}|^۲)^۲ = (۰/۵)^۳\,|\hat{x}_{n-۲}|^۴ \\ &\simeq (۰/۵)^۳(۰/۵\,|\hat{x}_{n-۲}|^۲)^۴ = (۰/۵)^۷\,|\hat{x}_{n-۳}|^۸ \\ &\simeq \cdots \simeq (۰/۵)^{۲^n-۱}\,|\hat{x}_۰|^{۲^n}\,. \end{aligned}$$

ادامه این مثال را در برنامه $oc.nb$ دنبال کنید. △

مثال ۷.۱ *دنباله $a_n = \frac{۱}{۲^n}$ با نرخ همگرایی $\frac{۱}{۲}$، همگرایی خطی به صفر دارد. همگرایی دنباله $b_n = \frac{۱}{۲^{۲^n}}$ به صفر زیرخطی است. در اصل همگرایی این دنباله مربعی است. دنباله $c_n = \frac{۱}{n+۱}$ همگرایی زیرخطی به صفر دارد.* △

تمرین ۲.۱ *در مورد مرتبه و نرخ همگرایی دنباله $a_n = \frac{۱}{۴^{\left[\frac{n}{۲}\right]}}$ چه می‌توان گفت؟*

تمرین ۳.۱ *نشان دهید در زیر، مرتبه همگرایی دقیقتر تعریف شده است.*
فرض کنید $\{x_n\}$ دنباله‌ای همگرا به x باشد و برای هر n داشته باشیم $x_n \neq x$. اگر اعداد مثبت و ثابت C و $p > ۱$ وجود داشته باشند به طوری که برای n به اندازه کافی بزرگ داشته باشیم

$$|x_{n+۱} - x| \leq C\,|x_n - x|^p\,,$$

آن‌گاه گوییم مرتبه همگرایی این دنباله p است.

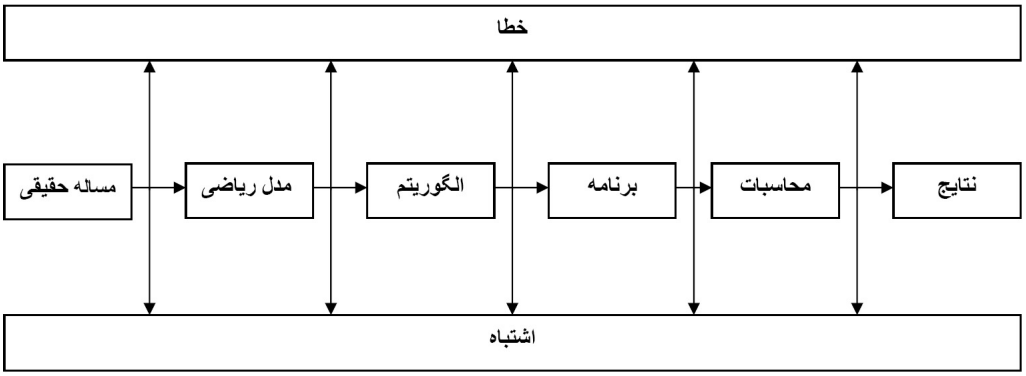
فصل ۲

خطاها

زمانی که به دست آوردن جواب واقعی یک مسئله غیرممکن است و یا مقرون به صرفه نیست، سعی می‌کنیم به کمک روش‌های عددی، یک جواب تقریبی برای مسئله پیدا کنیم. این فرایند منجر به تولید خطا[۱] می‌شود. در این فصل قصد داریم منابع تولید خطا و انواع خطا را شناسایی کرده و تا حدی از انتشار خطا جلوگیری کنیم.

۱.۲ منابع تولید خطا

بیشتر مواقع در عمل با یک مسئله حقیقی (فیزیکی) مواجه هستیم و بنا به دلایلی، جواب عددی (تقریبی) آن را جستجو می‌کنیم. مراحل یافتن جواب عددی چنین مسئله‌ای در روندنمای آمده در شکل ۱.۲ خلاصه می‌شود. این روندنما مکان‌های احتمالی بروز خطا و اشتباه[۲] را نیز نشان می‌دهد. مراحل مختلف این روندنما را در مثال بعد دنبال می‌کنیم.



شکل ۱.۲: فرایند تولید جواب عددی (تقریبی)

مثال ۱.۲ (مسئله حقیقی) *می‌خواهیم دوره تناوب حرکت نوسانی و متناوب یک آونگ ساده به جرم m و طول l را به دست آوریم. فرض کنید $\theta(t)$ جابجایی زاویه‌ای آونگ از خط قائم در زمان t باشد. صرف نظر از مقاومت هوا و اصطکاک در لولا، مدل این مسئله به صورت*

---
[۱] Error
[۲] Bug

یا $ml\frac{d^۲\theta}{dt^۲} = -mg\sin\theta$

$$\frac{d^۲\theta}{dt^۲} = -\frac{g}{l}\sin\theta, \tag{۱.۲}$$

در می‌آید که یک معادله دیفرانسیل غیرخطی است. با فرض کوچک بودن θ یعنی

$$\theta = ۶^\circ \simeq ۰/۱۰۴۷^{rad}, \quad \sin\theta \simeq ۰/۱۰۴۵, \qquad \theta = ۱۵^\circ \simeq ۰/۲۶۲^{rad}, \quad \sin\theta \simeq ۰/۲۵۹,$$

می‌توان فرض کرد $\sin\theta \simeq \theta^{rad}$ و معادله دیفرانسیل را به صورت $\frac{d^۲\theta}{dt^۲} + \omega^۲\theta = ۰$ نوشت که در آن $\omega^۲ = \frac{g}{l}$. این معادله دیفرانسیل خطی جوابی به صورت

$$\theta(t) = A\sin\omega t + B\cos\omega t,$$

دارد که از آن نتیجه می‌گیریم $T_L = \frac{۲\pi}{\omega} = ۲\pi\sqrt{\frac{l}{g}}$. از طرف دیگر با حل معادله دیفرانسیل غیرخطی (۱.۲) خواهیم داشت

$$T_N = ۲\pi\sqrt{\frac{l}{g}}\left(۱ + \frac{۱^۲}{۲^۲}\sin^۲\frac{\theta_m}{۲} + \frac{۱^۲\times۳^۲}{۲^۲\times۴^۲}\sin^۴\frac{\theta_m}{۲} + \cdots\right),$$

که در آن θ_m جابجایی زاویه‌ای ماکزیمم آونگ است. جالب است توجه داشته باشیم که T_L اولین جمله سری T_N است. در ادامه الگوریتم ۱.۲ را برای این مدل طراحی می‌کنیم.

الگوریتم ۱.۲ الگوریتم یافتن دوره تناوب حرکت یک آونگ ساده.

- **ورودی.** l، θ_m و مرز خطا یعنی ϵ
- **خروجی.** مقدار T_L، T_N و اختلاف آن‌ها

۱) قرار دهید $g = ۹/۸$، $p = ۳/۱۴$ و $T_L = ۲p\sqrt{\frac{l}{g}}$

۲) قرار دهید $k = ۱$، $M = \left(\frac{۲k-۱}{۲k}\right)^۲$، $T_۱ = T_L$، $C = T_L \times M \times \sin^{۲k}\frac{\theta_m}{۲}$ و $T_۲ = T_۱ + C$

۳) تا زمانی که $|C| > \epsilon$ گام ۴ را تکرار کنید

۴) قرار دهید $k = k + ۱$، $M = M \times \left(\frac{۲k-۱}{۲k}\right)^۲$، $T_۱ = T_۲$، $C = T_L \times M \times \sin^{۲k}\frac{\theta_m}{۲}$ و $T_۲ = T_۱ + C$

۵) قرار دهید $T_N = T_۲$ و T_L، T_N و $T_N - T_L$ را چاپ کرده، متوقف شوید

حال برای الگوریتم ۱.۲ برنامه ۱.۲ را با استفاده از نرم‌افزار Mathematica می‌نویسیم. پس از اجرای برنامه ۱.۲ در محیط نرم‌افزار Mathematica محاسبات شروع شده و به ازای $\epsilon = ۰/۰۰۰۱$ نتایج گزارش شده در جدول ۱.۲ تولید می‌شود. باید توجه داشت که برای $\theta_m < ۱۵^\circ$ اختلاف T_l و T_N از $۰/۰۰۵$ بیشتر نیست. △

```
l = Input["طول آونگ را بر حسب متر وارد کنید"];
Thetam = Input["جابجایی زاویه‌ای ماکزیمم را بر حسب درجه وارد کنید"];
Epsilon = Input["مقدار مجاز خطا را وارد کنید"];
st = Sin[Thetam * Pi/۳۶۰]; g = ۹٫۸; TL = ۲ * Pi * Sqrt[l/g]; k = ۱;
M = ((۲k-۱)/(۲k))^۲; CC = TL * M * st^(۲k); T۱ = TL; T۲ = T۱ + CC;
While[CC > Epsilon, {
  k++; M* = ((۲k-۱)/(۲k))^۲; T۱ = T۲; CC = TL * M * st^(۲k); T۲ = T۱ + CC; }];
TN = T۲;
Print["T_L = ", TL, "   T_N = ", TN, "   T_N - T_L = ", TN - TL];
```

برنامه ۱٫۲: برنامه یافتن دوره تناوب حرکت یک آونگ ساده.

با توجه به روندنمای ارایه شده در شکل ۱.۲ و مثال ۱.۲، خطاها از نظر منابع تولید به صورت زیر تقسیم‌بندی می‌شوند

$l(cm)$	θ_m	T_L	T_N	$T_N - T_L$	$l(cm)$	θ_m	T_L	T_N	$T_N - T_L$
۱۰	۳	۰٫۶۳۴۷۰	۰٫۶۳۴۸۱	۰٫۰۰۰۱۱	۱۰	۱۲	۰٫۶۳۴۷۰	۰٫۶۳۶۴۴	۰٫۰۰۱۷۴
۲۰	۳	۰٫۸۹۷۶۰	۰٫۸۹۷۷۵	۰٫۰۰۰۱۵	۲۰	۱۲	۰٫۸۹۷۶۰	۰٫۹۰۰۰۷	۰٫۰۰۲۴۷
۳۰	۳	۱٫۰۹۹۳۳	۱٫۰۹۹۵۲	۰٫۰۰۰۱۹	۳۰	۱۲	۱٫۰۹۹۳۳	۱٫۱۰۲۳۵	۰٫۰۰۳۰۲
۱۰	۶	۰٫۶۳۴۷۰	۰٫۶۳۵۱۳	۰٫۰۰۰۴۳	۱۰	۱۵	۰٫۶۳۴۷۰	۰٫۶۳۷۴۳	۰٫۰۰۲۷۳
۲۰	۶	۰٫۸۹۷۶۰	۰٫۸۹۸۲۱	۰٫۰۰۰۶۱	۲۰	۱۵	۰٫۸۹۷۶۰	۰٫۹۰۱۴۶	۰٫۰۰۳۸۶
۳۰	۶	۱٫۰۹۹۳۳	۱٫۱۰۰۰۸	۰٫۰۰۰۷۵	۳۰	۱۵	۱٫۰۹۹۳۳	۱٫۱۰۴۰۶	۰٫۰۰۴۷۳
۱۰	۹	۰٫۶۳۴۵۰	۰٫۶۳۵۷۰	۰٫۰۰۱۲۰	۱۰	۱۸	۰٫۶۳۴۷۰	۰٫۶۳۸۶۳	۰٫۰۰۳۹۳
۲۰	۹	۰٫۸۹۷۶۰	۰٫۸۹۹۰۰	۰٫۰۰۱۴۰	۲۰	۱۸	۰٫۸۹۷۶۰	۰٫۹۰۳۱۷	۰٫۰۰۵۵۷
۳۰	۹	۱٫۰۹۹۳۳	۱٫۱۰۱۰۳	۰٫۰۰۱۷۰	۳۰	۱۸	۱٫۰۹۹۳۳	۱٫۱۰۶۱۵	۰٫۰۰۶۸۲

جدول ۱.۲: دوره تناوب آونگ ساده با طول‌های مختلف

۱. ذاتی

- مدل (ناشی از صرف‌نظرها، چشم‌پوشی‌ها و ساده‌سازی‌ها مانند فرض $\sin\theta \simeq \theta^{rad}$)
- داده‌های مدل (ناشی از آزمایشات و اندازه‌گیری‌ها مثل g, l)

۲. محاسباتی

- نمایش اعداد (مانند $\pi = ۳٫۱۴$)
- اعمال ریاضی (به عنوان مثال $\frac{l}{g}$)
- روش‌های (الگوریتم‌های) عددی (محاسباتی) (مثل خطای روش محاسبه $\sqrt{\frac{l}{g}}$)

تذکر ۱.۲ *خطاهای ذاتی به متخصص آنالیز عددی مربوط نمی‌شود اما برای پرهیز از خطاهای محاسباتی و کنترل آن‌ها باید راه چاره‌ای پیدا کرد.*

تذکر ۲.۲ *اختلاف جواب واقعی و تقریبی از اشتباهات و خطاها ناشی می‌شود. اشتباه را می‌توان برطرف نمود ولی خطا ممکن است اجتناب‌ناپذیر باشد. به عنوان مثال قراردادن ۲۳۲۳ به جای ۳۲۳۲ یک اشتباه است حال آن که عدد π بسط ده‌دهی نامختوم دارد و در وسایل محاسباتی باید یک بسط ده‌دهی مختوم برای آن در نظر گرفت که این موجب بروز خطا می‌شود.*

تذکر ۳.۲ *برای اطلاعات بیشتر در خصوص اثرات مخرب خطاها و اشتباهات به آدرس‌های زیر مراجعه کنید.*

https://www.iro.umontreal.ca/ mignotte/IFT2425/Disasters.html

https://www.computerworld.com/article/3412197/top-software-failures-in-recent-history.html

در اینجا فقط به دو مورد زیر اشاره می‌شود.

- *عدم موفقیت موشک پاتریوت در جنگ خلیج فارس سال ۱۹۹۱ (۲۸ کشته و ۱۰۰ زخمی) به دلیل وقوع خطای گرد کردن در محاسبات مسیر*
- *شکست ماموریت موشک آریان ۵ فرانسه در سال ۱۹۹۶ (۵۰۰ میلیون دلار خسارت مادی) به دلیل وقوع پاریز[۳] در رایانه آن*

۲.۲ نمایش اعداد

در این بخش به بررسی نمایش اعداد حقیقی می‌پردازیم. اثبات برخی از قضایا را می‌توان در سایر مراجع آنالیز عددی یافت.

تعریف ۱.۲ *هر عدد حقیقی مثبت x نمایشی معروف به نمایش مکانی[۴] به صورت*

$$\begin{aligned} x &= a_m\beta^m + a_{m-1}\beta^{m-1} + \cdots + a_1\beta^1 + a_0\beta^0 + a_{-1}\beta^{-1} + a_{-2}\beta^{-2} + \cdots \\ &= (a_m a_{m-1} \ldots a_1 a_0 / a_{-1} a_{-2} \ldots)_\beta, \end{aligned} \tag{۲.۲}$$

دارد که در آن $\beta > 1$ یک عدد طبیعی معروف به مبنا، $m \in \mathbb{Z}$ و $a_i \in \{0, 1, 2, \ldots, \beta - 1\}$. برای یکتایی نمایش لازم است که $a_m \neq 0$ و عدد صحیح j چنان وجود داشته باشد که $a_{j-1} = a_j = \cdots = 0$، یا به ازای هر N به اندازه کافی بزرگ وجود داشته باشد $N \leq j$ به طوری که $a_{-j} \neq \beta - 1$.

تذکر ۴.۲ *رابطه (۲.۲) به نمایش (بسط) عدد x در مبنای β معروف است و اگر $\beta = 10$ اختیار شود به آن، نمایش (بسط) دهدهی (اعشاری) گویند (متداول در زندگی روزمره) و در حالتی که $\beta = 2$ در نظر گرفته شود به آن، نمایش دودویی (باینری) گفته می‌شود (مبنای کار رایانه).*

تذکر ۵.۲ *شرط دوم یکتایی نمایش، لازم است چه در غیر این صورت به عنوان مثال می‌توان نوشت*

$$\begin{aligned} 3/479999\ldots = 3/47\bar{9} &= 3 \times 10^0 + 4 \times 10^{-1} + 7 \times 10^{-2} + 9 \times 10^{-3} + 9 \times 10^{-4} + \cdots \\ &= 3/47 + \frac{9 \times 10^{-2}}{1 - 10^{-1}} = 3/47 + 0/01 = 3/48. \end{aligned}$$

قضیه ۱.۲ *بسط یک عدد گویا در هر مبنایی یا مختوم است یا نامختوم متناوب.*

---

[۳] Underflow

[۴] Positional representation

نتیجه ۱.۱.۲ *بسط عدد گنگ، نامختوم نامتناوب است.*

مثال ۲.۲

$$\frac{۱}{۳} = ۰/\overline{۳} = (۰/۱)_۳,\ \frac{۳}{۸} = ۰/۳۷۵,\ \frac{۱}{۱۰} = ۰/۱ = (۰/۰\overline{۰۰۱۱})_۲,\ \sqrt{۲} = ۱/۴۱۴۲\cdots,\ \pi = ۳/۱۴۱۵۹۲\cdots.$$

△

قضیه ۲.۲ *به ازای هر $\epsilon > ۰$ و هر عدد حقیقی x عدد حقیقی y چنان وجود دارد که $|x-y| < \epsilon$.*

هنگام کار با رایانه (ماشین حساب) اعداد را در مبنای ۱۰ وارد کرده و انتظار داریم نتایج (خروجی) نیز در همین مبنا نمایش داده شود، ولی این وسایل با مبنای دیگری (امروزه مبنای ۲ و در قدیم مبنای ۱۶ نیز استفاده می‌شده) کار می‌کنند. بنابراین مسئله تغییر مبنا مطرح می‌شود.

پرسش ۱.۲ *مبنای ۲ در مقایسه با مبنای ۱۶ چه معایب و مزایایی دارد؟*

تغییر مبنا از مبنای ۲ به مبنای ۱۰ به کمک بسط به سادگی امکان‌پذیر است.

مثال ۳.۲

$$\begin{aligned}(۱۰۱۱۰۰۱/۱۱۱۰۰۱)_۲ &= ۲^۶ + ۲^۴ + ۲^۳ + ۲^۰ + ۲^{-۱} + ۲^{-۲} + ۲^{-۳} + ۲^{-۶} \\ &= (۶۴ + ۱۶ + ۸ + ۱ + ۰/۵ + ۰/۲۵ + ۰/۱۲۵ + ۰/۰۱۵۶۲۵) \\ &= ۸۹/۸۹۰۶۲۵.\end{aligned}$$

△

برای تغییر مبنا از ۱۰ به ۲، اگر عدد داده شده در مبنای ۱۰، صحیح مثبت باشد، با تقسیم‌های پشت سر هم بر ۲ عدد متناظر در مبنای ۲ به دست می‌آید.

مثال ۴.۲

$$۲۳ = ۱۱ \times ۲ + ۱, \qquad ۱۱ = ۵ \times ۲ + ۱, \qquad ۵ = ۲ \times ۲ + ۱, \qquad ۲ = ۱ \times ۲ + ۰.$$

پس $(۱۰۱۱۱)_۲ = ۲۳$. △

اگر عدد داده شده x در مبنای ۱۰ در فاصله $(۰, ۱)$ باشد و عدد متناظر در مبنای ۲ به صورت $(۰/b_۱b_۲b_۳\cdots)_۲$ فرض شود آن‌گاه

$$x = b_۱ \times ۲^{-۱} + b_۲ \times ۲^{-۲} + b_۳ \times ۲^{-۳} + \cdots,$$

و در نتیجه

$$۲x = b_۱ + b_۲ \times ۲^{-۱} + b_۳ \times ۲^{-۲} + \cdots.$$

با قرار دادن

$$y = b_۲ \times ۲^{-۱} + b_۳ \times ۲^{-۲} + \cdots,$$

خواهیم داشت

$$\circ \leq y < ۱ \times ۲^{-۱} + ۱ \times ۲^{-۲} + \cdots = \frac{۲^{-۱}}{۱ - ۲^{-۱}} = ۱.$$

پس $[y] = \circ$. از طرف دیگر داریم $۲x = b_۱ + y$ و بلافاصله از $[۲x] = [b_۱ + y]$ نتیجه می‌شود $b_۱ = [۲x]$ زیرا $b_۱$ عددی صحیح است ($b_۱$ یا صفر است یا یک). با قرار دادن $x = y$ و تکرار این روند می‌توان $b_۲$، $b_۳$ و $\cdots$ را به دست آورد. شرط توقف این روند تکراری زمانی است که x صفر شود یا تناوب به وجود آید و یا ممکن است بسط نامختوم نامتناوب باشد و پس از چند تکرار مجبور شویم روند را متوقف کنیم.

مثال ۵.۲ *این روند را برای حالت‌هایی که نمایش، مختوم یا نامختوم متناوب است دنبال می‌کنیم.*

با توجه به جدول ۲.۲ می‌توان نوشت $(۰/۱۰۰۱۰۱)_۲ = ۰/۵۷۸۱۲۵$ *و* $(۰/۰۰۰۱۱\overline{۰})_۲ = ۰/۱$. △

i	x	$۲x$	b_i
۱	۰/۱	۰/۲	۰
۲	۰/۲	۰/۴	۰
۳	۰/۴	۰/۸	۰
۴	۰/۸	۱/۶	۱
۵	۰/۶	۱/۲	۱
۶	۰/۲	۰/۴	۰
	۰/۴		

i	x	$۲x$	b_i
۱	۰/۵۷۸۱۲۵	۱/۱۵۶۲۵	۱
۲	۰/۱۵۶۲۵	۰/۳۱۲۵	۰
۳	۰/۳۱۲۵	۰/۶۲۵	۰
۴	۰/۶۲۵	۱/۲۵	۱
۵	۰/۲۵	۰/۵	۰
۶	۰/۵	۱/۰	۱
	۰		

جدول ۲.۲: نمایش دودویی اعداد کوچک‌تر از واحد

تمرین ۱.۲ *برنامه‌ای بنویسید که عدد حقیقی x و عدد طبیعی n را از ورودی گرفته و نمایش دودویی عدد x را تولید کند و اگر نمایش نامختوم نامتناوب است حداکثر تا n رقم نمایش، چاپ شود.*

تعریف ۲.۲ *برای نمایش اعداد در ماشین، ابتدا نمایشی معروف به نمایش ممیز ثابت[۵] در نظر گرفته شد که در آن نظیر برخی از اعداد حقیقی مثبت x می‌توان یک عدد ماشینی به صورت زیر در نظر گرفت*

$$\overline{x} = (a_{n-۱} \cdots a_۱ a_۰ / a_{-۱} a_{-۲} \cdots a_{-t})_\beta,$$

که در آن $\beta > ۱$ یک عدد طبیعی معروف به مبنا، $a_i \in \{۰, ۱, ۲, \ldots, \beta - ۱\}$ و t و n اعداد طبیعی مشخص و ثابتی هستند. در اصل در این نمایش مکان ممیز مشخص و ثابت است و فرض می‌شود a_iها همگی صفر نیستند.

مجموعه تمام اعداد ماشینی با نمایش ممیز ثابت با $\mathbb{F}_\circ = \mathbb{F}_\circ(\beta, n, t)$ نمایش داده می‌شود و برخی از نکات مهم در این نمایش عبارتند از

- $\mathrm{Card}(\mathbb{F}_\circ) = ۲ \times \beta^n \times \beta^t - ۱$،
- با این فرض که $\gamma = \beta - ۱$، بزرگترین و کوچکترین عدد عبارتند از

$$x_{\max} = (\gamma \cdots \gamma / \gamma \cdots \gamma)_\beta = \gamma \sum_{i=-t}^{n-۱} \beta^i = \beta^n - \beta^{-t} \simeq \beta^n, \qquad x_{\min} = (\circ \cdots \circ / \circ \cdots ۱)_\beta = \beta^{-t},$$

[۵] Fixed point representation

- اعداد در مجموعه $[-x_{\max}, -x_{\min}] \cup [x_{\min}, x_{\max}]$ هم‌فاصله و با فاصله β^{-t} قرار دارند،
- برای نمایش اعداد بسیار بزرگ (کوچک) در این نمایش با مشکل مواجه می‌شویم و در نتیجه این نمایش برای محاسبات علمی مناسب نیست ولی برای بسیاری از کاربردها مانند حسابداری، این نمایش سودمند است و هنوز هم ماشین‌هایی بر این اساس ساخته می‌شوند.

تعریف ۳.۲ *هر عدد حقیقی مثبت x نمایشی معروف به نمایش علمی[۶] به صورت*

$$x = (۰/a_۱ \cdots a_t a_{t+۱} \cdots)_\beta \times \beta^e,$$

دارد که در آن $\beta > ۱$ یک عدد طبیعی معروف به مبنا و $a_i \in \{۰, ۱, ۲, \ldots, \beta - ۱\}$. $e \in \mathbb{Z}$ به توان (نما) معروف است و $m = ۰/a_۱ \cdots a_t a_{t+۱} \cdots$ مانتیس (جزء کسری) نام دارد و $\frac{۱}{\beta} \le m < ۱$ و برای یکتایی نمایش لازم است که $a_۱ \neq ۰$ (در این حالت گفته می‌شود نمایش نرمال‌شده است).

نمایش مکانی نقطه شروعی برای تعریف نمایش ممیز ثابت بود حال آنکه برای تعریف نمایش پرکاربرد ممیز شناور[۷] از نمایش علمی ایده گرفته شد. در این نمایش اعداد بسیار بزرگ (کوچک) به سادگی قابل نمایش هستند و یک روش متداول برای نمایش اعداد در رایانه است که از بدو پیدایش مورد توجه سازندگان سخت‌افزار رایانه قرار گرفته است و چنین به نظر می‌رسد که تا حدودی به طور سلیقه‌ای با آن برخورد شده است.

تعریف ۴.۲ *نظیر برخی از اعداد حقیقی مثبت x نمایشی معروف به نمایش ممیز شناور به صورت*

$$\widehat{x} = (۰/a_۱ \cdots a_t)_\beta \times \beta^e = m \times \beta^{e-t},$$

تعریف می‌شود که در آن $\beta > ۱$ یک عدد طبیعی معروف به مبنا، e یعنی توان به زیرمجموعه‌ای از $\mathbb{Z}$ تعلق دارد، $t \in \mathbb{N}$ تعداد ارقام بامعنای مانتیس یعنی $۰/a_۱ \cdots a_t$ را نشان می‌دهد، $a_i \in \{۰, ۱, ۲, \ldots, \beta - ۱\}$ و برای یکتایی نمایش لازم است که $a_۱ \neq ۰$ (نمایش نرمال‌شده). به وضوح برای $m = (a_۱ \cdots a_t)_\beta$ داریم $\beta^{t-۱} \le m \le \beta^t - ۱$ و اعداد صحیح L و U چنان وجود دارند که $L \le e \le U$. همچنین $a_۱$ به باارزش‌ترین رقم و a_t به کم‌ارزش‌ترین رقم معروف است.

مجموعه تمام اعداد ماشینی با نمایش ممیز شناور همراه با صفر با $\mathbb{F} = \mathbb{F}(\beta, t, L, U)$ نمایش داده می‌شود و برخی از نکات مهم در این نمایش عبارتند از

- $\mathrm{Card}(\mathbb{F}) = ۲ \times (\beta - ۱) \times \beta^{\mathrm{t}-۱}(\mathrm{U} - \mathrm{L} + ۱) + ۱$،
- با این فرض که $\gamma = \beta - ۱$، بزرگترین و کوچکترین عدد عبارتند از

$$x_{\max} = (۰/\gamma \cdots \gamma)_\beta \times \beta^U = \gamma \beta^U \sum_{i=۱}^{t} \beta^{-i} = \beta^U (۱ - \beta^{-t}) \simeq \beta^U, \quad x_{\min} = (۰/۱ ۰ \cdots ۰)_\beta \times \beta^L = \beta^{L-۱},$$

- اعداد در مجموعه $\mathbb{F}$ هم‌فاصله نیستند و فاصله آنها به توان بستگی دارد. هرچه e بزرگتر (کوچکتر) باشد فاصله اعداد بیشتر (کمتر) است.

باید توجه داشت که از یک طرف صفر به مجموعه اعداد ممیز شناور اضافه شده است و از طرف دیگر در یک همسایگی حول صفر هیچ عددی قابل نمایش نیست. به منظور برطرف کردن این مشکلات، تعریف زیر مطرح می‌شود.

---

[۶] Scientific representation

[۷] Floating point

تعریف ۵.۲ *اگر در تعریف اعداد با نمایش ممیز شناور $a_1 = 0$ فقط برای $e = L$ مجاز باشد، مجموعه بزرگتری از اعداد ماشینی معروف به اعداد با نمایش غیرنرمال‌شده ممیز شناور[۸] به دست می‌آید که با $\mathbb{F}_D = \mathbb{F}_D(\beta, t, L, U)$ نمایش داده می‌شود.*

باید توجه داشت که صفر در این نمایش وجود دارد، یکتایی نمایش حفظ شده است و به علاوه $\mathbb{F} \subset \mathbb{F}_D$. برخی از نکات مهم در این نمایش عبارتند از

- $\mathrm{Card}(\mathbb{F}_\mathrm{D}) = \mathrm{Card}(\mathbb{F}) + 2(\beta^{\mathrm{t-1}} - 1)$،
- $x_{\min} = \beta^{L-t}$،
- می‌توان اعدادی با مانتیس بین ۱ (متناظر با حالت $0\cdots01$) و $1 - \beta^{t-1}$ (متناظر با حالت $0\gamma\cdots\gamma$) که به بازه $(-\beta^{L-1}, \beta^{L-1})$ تعلق دارند نیز نمایش داد.

مثال ۶.۲ *مجموعه اعداد $\mathbb{F}_D(2, 3, -1, 2)$ را مشخص کنید.*

در مجموعه $\mathbb{F}(2, 3, -1, 2)$، $16 = (2 + 1 + 1) \times 2^{2-1} \times (2 - 1)$ عدد مثبت وجود دارد که عبارتند از

$$x_{\max} = (0/111)_2 \times 2^2 = \tfrac{7}{2}, \quad (0/110)_2 \times 2^2 = 3, \quad (0/101)_2 \times 2^2 = \tfrac{5}{2}, \quad (0/100)_2 \times 2^2 = 2,$$
$$\tfrac{7}{4}, \quad \tfrac{3}{2}, \quad \tfrac{5}{4}, \quad 1, \quad \tfrac{7}{8}, \quad \tfrac{3}{4}, \quad \tfrac{5}{8}, \quad \tfrac{1}{2}, \quad \tfrac{7}{16}, \quad \tfrac{3}{8}, \quad \tfrac{5}{16}, \quad x_{\min} = (0/100)_2 \times 2^{-1} = \tfrac{1}{4}.$$

اعداد غیرنرمال‌شده نامنفی عبارتند از

$$(0/011)_2 \times 2^{-1} = \frac{3}{16}, \quad (0/010)_2 \times 2^{-1} = \frac{1}{8}, \quad (0/001)_2 \times 2^{-1} = \frac{1}{16}, \quad (0/000)_2 \times 2^{-1} = 0.$$

تعریف ۶.۲ *فاصله عدد ۱ و عدد قابل نمایش بعد از آن به اپسیلون ماشین معروف است و با ϵ_M نمایش داده می‌شود و به عنوان مثال در $\mathbb{F}(\beta, t, L, U)$ مقدار آن به صورت زیر به دست می‌آید* △

$$(0/10\cdots01)_\beta \times \beta^1 - (0/10\cdots0)_\beta \times \beta^1 = (0/0\cdots01)_\beta \times \beta^1 = \beta^{1-t}.$$

تنوع انتخاب کمیت‌های β، t، L و U موجب شد که ماشین‌های محاسباتی (رایانه‌های) گوناگونی تولید شوند که در برخی موارد نتایج اجرای یک برنامه بر روی ماشین‌های مختلف، متفاوت بود. به همین دلیل IEEE[۹] در سال ۱۹۸۵ استاندارد ۷۵۴ را وضع و سازندگان این وسایل را به پیروی از آن ترغیب کرد. این استاندارد در سال ۱۹۸۹ توسط IEC[۱۰] تائید شد و به صورت استاندارد IEC559 مطرح گردید. در این استاندارد دو قالب[۱۱] موسوم به دقت ساده[۱۲] و دقت دو برابر[۱۳] برای اعداد با نمایش ممیز شناور پیشنهاد شده است.

۱.۲.۲ نمایش ۶۴-بیتی ممیز شناور

این نمایش پیش از این به دقت دو برابر (مضاعف) معروف بوده و متناظر با نوع double در زبان C است. در این نمایش برای هر عدد در مبنای ۲، ابتدا یک ساختار به طول ۶۴ بیت در نظر گرفته می‌شود. اولین بیت به بیت علامت معروف است و با s نمایش داده می‌شود و بلافاصله بعد از

[۸] Floating point de-normalized
[۹] Institute of Electrical and Electronics Engineers
[۱۰] International Electronical Commission
[۱۱] Format
[۱۲] Single precision
[۱۳] Double precision

آن ۱۱ بیت برای مشخصه[۱۴] در نظر گرفته می‌شود و با c نمایش داده می‌شود و ۵۲ بیت باقی‌مانده برای مانتیس کنار گذاشته می‌شود و آن را با f نشان می‌دهند.

سپس برای هر عدد نمایشی به صورت $(-۱)^s \times ۲^e \times (۱ + f)$ در نظر گرفته می‌شود و توان به کمک رابطه $e = c - ۱۰۲۳$ به دست می‌آید. با توجه به محدودیت $۰ \leq c \leq (۱۱۰۰۰۰۱)_۲ = ۲۰۴۷ = ۲^{۱۱} - ۱$ داریم $-۱۰۲۳ \leq e \leq ۱۰۲۴$. از طرف دیگر محدودیت $(۰/۱۱۰۰۰۰۱)_۲ \leq f \leq ۰$ باعث می‌گردد که $۲ > (۱/۱۱۰۰۰۰۱)_۲ \geq (۱ + f) \geq ۱$. به نظر می‌رسد برای مانتیس ۵۲ بیت در اختیار داریم ولی در واقع ساختار $۱ + f$ موجب می‌شود که ۵۳ بیت داشته باشیم که آن یک بیت اضافه به بیت پنهان معروف است و برای یکتایی نمایش لازم است. اگر کوچک‌ترین و بزرگ‌ترین عدد مثبت قابل نمایش را به ترتیب با mN و MN نشان دهیم آن‌گاه

$$mN = ۲^{-۱۰۲۲} \times (۱/۰۰۰۰۰)_۲ \simeq ۲/۲۲۵۰۷ \times ۱۰^{-۳۰۸},$$

$$MN = ۲^{۱۰۲۴} \times (۱/۰۰۰۰۰)_۲ \simeq ۱/۷۹۷۶۹ \times ۱۰^{۳۰۸}.$$

تذکر ۶.۲ *در این نمایش $\epsilon_M = ۲^{-۵۲} \simeq ۲/۲۲۰۴۴۶ \times ۱۰^{-۱۶}$ (همان eps در نرم‌افزار MATLAB) و از توان -۱۰۲۳ برای نمایش صفر و پیام پاریز[۱۵] (زمانی که اعدادی با اندازه خیلی کوچک تولید می‌شوند) و از توان ۱۰۲۴ با مانتیس مثبت برای نمایش ∞ (inf در MATLAB)، صور مبهم (NaN در MATLAB) و پیام سرریز[۱۶] (زمانی که اعدادی با اندازه خیلی بزرگ تولید می‌شوند) استفاده می‌شود.*

در ادامه به دنبال بزرگ‌ترین عدد صحیح مثبت M هستیم که هر عدد صحیح x با شرط $۰ \leq x \leq M$ در این نمایش به طور دقیق قابل‌نمایش باشد. به وضوح، تمام اعداد صحیح نامنفی که بزرگتر از $(۱/۱۱۰۰۰۰۱)_۲ \times ۲^{۵۲} = ۲^{۵۳} - ۱$ نباشند به طور دقیق قابل‌نمایش هستند و به علاوه $۲^{۵۳}$ نیز به صورت $(۱/۰۰۰۰۰)_۲ \times ۲^{۵۳}$ قابل‌نمایش است. اما تعداد ارقام در مانتیس جهت نمایش $۱ + ۲^{۵۳}$ کافی نیست (۵۳ رقم در مانتیس لازم است). بنابراین $M = ۲^{۵۳} \simeq ۹/۰۰۷۱۹۹ \times ۱۰^{۱۵}$ و در نتیجه تمام اعداد صحیح ۱۵ رقمی و بسیاری از اعداد ۱۶ رقمی در این نمایش به طور دقیق قابل‌نمایش هستند (گفته می‌شود دقت در نمایش ۶۴-بیتی ۱۵ الی ۱۶ رقم است). البته باید توجه داشت که بسیاری از توان‌های ۲ نیز به طور دقیق قابل‌نمایش هستند.

تذکر ۷.۲ محور *اعداد با ممیز شناور بر خلاف محور اعداد حقیقی، متناهی و گسسته است. بعضی از نرم‌افزارها مانند* Mathematica، *محدودیت‌های سخت‌افزار را به کمک برنامه‌های نرم‌افزاری برطرف می‌کنند و به اصطلاح دقت را بالا می‌برند که ممکن است کمی سرعت انجام محاسبات پایین بیاید. برنامه* uoflow.nb *را ببینید.*

تمرین ۲.۲ *آیا $\mathbb{F}(۲, ۵۳, -۱۰۲۳, ۱۰۲۴)$ بیانگر مجموعه اعداد ممیز شناور ۶۴-بیتی است؟*

تمرین ۳.۲ *در نمایش ۳۲-بیتی (دقت ساده متناظر با نوع* float *در زبان* C*) هشت بیت برای مشخصه منظور می‌شود. تمام کمیت‌های معرفی‌شده در این بخش را، در این نمایش تعیین کنید. آیا $\mathbb{F}(۲, ۲۴, -۱۲۷, ۱۲۸)$ بیانگر مجموعه اعداد ممیز شناور ۳۲-بیتی است؟*

برای جزئیات بیشتر به مرجع [۱۷] و برای جزئیات کاملتر به استاندارد بیان‌شده مراجعه کنید.

---

[۱۴] Characteristic

[۱۵] Underflow

[۱۶] Overflow

۲.۲.۲ اعداد ماشینی

اعداد $\mathbb{F} = \mathbb{F}(\beta, t, L, U)$ به اعداد ماشینی t-رقمی نیز معروف هستند. به وضوح $\mathbb{F} \subset \mathbb{R}$ و اگر $x \in \mathbb{R}$ و $x \notin \mathbb{F}$، باید نگاشتی مانند $r : \mathbb{R} \to \mathbb{F}$ ساخت به گونه‌ای که

$$|x - r(x)| \leq |x - y|, \qquad \forall y \in \mathbb{F}.$$

اگر نمایش علمی x به صورت $x = (۰/a_۱ \cdots a_t a_{t+۱} \cdots)_\beta \times \beta^e$ باشد، عدد ماشینی t-رقمی نظیر x با نگه داشتن t رقم از مانتیس و کنار گذاشتن بقیه ارقام به دست می‌آید (البته با این فرض که برای نمایش e مشکلی نداشته باشیم) و برای این کار روش‌های زیر موجود است

۱. روش قطع کردن (برش)[۱۷]

۲. روش گرد کردن معمولی[۱۸]

۳. روش گرد کردن به زوج[۱۹]

در روش قطع کردن، t رقم از مانتیس حفظ شده و بقیه ارقام کنار گذاشته می‌شود در حالی که در روش گرد کردن معمولی، ابتدا روش قطع کردن اعمال شده، سپس اگر $a_{t+۱} \geq \frac{\beta}{۲}$ یک واحد به a_t اضافه می‌گردد. اما در روش گرد کردن به زوج، ابتدا روش قطع کردن اعمال شده و در هر یک از حالت‌های زیر یک واحد به a_t اضافه می‌گردد

- $a_{t+۱} > \frac{\beta}{۲}$،
- $a_{t+۱} = \frac{\beta}{۲}$ و رقم مخالف صفری در سمت راست $a_{t+۱}$ مشاهده شود،
- $a_{t+۱} = \frac{\beta}{۲}$ و رقم مخالف صفری در سمت راست $a_{t+۱}$ مشاهده نشود و a_t فرد باشد.

امروزه روش‌های قطع کردن (به دلیل خطای زیاد) و گرد کردن به زوج (به دلیل عدم توجیحات قابل قبول) منسوخ شده‌اند (اگرچه زمانی رایانه‌هایی بر مبنای آنها ساخته شد). بنابراین می‌توان نگاشت r را به صورت زیر در نظر گرفت

$$r(x) = \begin{cases} (۰/a_۱ \cdots a_t)_\beta \times \beta^e, & a_{t+۱} < \frac{\beta}{۲}, \\ (۰/a_۱ \cdots a_t + ۰/۰ \cdots ۰۱)_\beta \times \beta^e, & a_{t+۱} \geq \frac{\beta}{۲}. \end{cases}$$

پرسش ۲.۲ *تفاوت* ۴۷^{km} *با* ۴۷۰۰۰^m *یا تفاوت* ۳/۷ *با* ۳/۷۰ *یا* ۳/۷۰۰ *در چیست؟*

تعریف ۷.۲ *منظور از ارقام بامعنای یک عدد مخالف صفر، ارقام مخالف صفر، صفرهای بین دو رقم مخالف صفر و صفرهایی است که در سمت راست عدد به منظور نشان دادن نوعی دقت قرار داده می‌شوند (تمام ارقام مانتیس).*

مثال ۷.۲ *عدد* ۰/۰۰۰۷۰۴۵۰۰ *حداقل ۴ رقم بامعنا و حداکثر ۶ رقم بامعنا دارد.* △

---

[۱۷] Chopping
[۱۸] Rounding
[۱۹] Rounding to even

تعریف ۸.۲ *اگر $\widehat{x}$ تقریبی از x باشد، $e_a(x) = |x - \widehat{x}|$ را خطای مطلق $\widehat{x}$ نسبت به x نامند. e_a منحصر به فرد است و در عمل بیشتر مواقع قابل تعیین نیست و به جای آن از هر عدد b_a استفاده می‌شود که کمتر از $e_a(x)$ نباشد. b_a یکتا نیست و به کران خطای مطلق معروف است. بنابراین $e_a(x) \leq b_a$ و در نتیجه $\widehat{x} - b_a \leq x \leq \widehat{x} + b_a$. بعضی مواقع از نمایش $x = \widehat{x} \pm b_a$ استفاده می‌شود.*

مثال ۸.۲ *اگر عدد $\widehat{x} = ۱/۷۳۲$ به عنوان تقریبی از $x = \sqrt{۳}$ باشد، آن‌گاه*

$$e_a(x) = \left|\sqrt{۳} - ۱/۷۳۲\right| = ۱/۷۳۲۰۵۰۸۰۷۵\cdots - ۱/۷۳۲ = ۰/۰۰۰۰۵۰۸۰۷۵\cdots.$$

از طرفی می‌دانیم $۱/۷۳۲۱ < \sqrt{۳} < ۱/۷۳۲۰$، بنابراین $۰/۰۰۰۱ > \sqrt{۳} - ۱/۷۳۲ > ۰$. در نتیجه $b_a = ۰/۰۰۰۱$ معیاری برای نزدیکی $۱/۷۳۲$ به $\sqrt{۳}$ است. △

پرسش ۳.۲ *آیا خطای مطلق معیار مناسبی برای مقایسه خطاها است؟*

پاسخ. *خیر. به عنوان مثال خطای مطلق یک صندوق‌دار بانک، تایپیست و دروازه‌بان را در نظر بگیرید.*

تعریف ۹.۲ *اگر $\widehat{x}$ تقریبی از $x \neq ۰$ باشد، $e_r(x) = \frac{|x-\widehat{x}|}{|x|} = \frac{e_a(x)}{|x|}$ خطای نسبی $\widehat{x}$ نسبت به x نامیده می‌شود و مشابه خطای مطلق منحصر به فرد است و بیشتر مواقع در عمل قابل تعیین نیست و از کران خطای نسبی استفاده می‌شود. $۱۰۰ \times e_r(x)$ به درصد خطا معروف است.*

قضیه ۳.۲ *اگر $\widehat{x}$ تقریبی از x باشد و b_a یک کران خطای مطلق برای این تقریب باشد، آن‌گاه*

$$e_r(x) \leq \frac{b_a}{|\widehat{x}| - b_a}.$$

به علاوه اگر b_a نسبت به $|\widehat{x}|$ خیلی کوچک باشد داریم

$$e_r(x) \leq \frac{b_a}{|\widehat{x}|}.$$

برهان. به کمک تعریف b_a و با استفاده از خواص نابرابری‌ها نشان دهید. □

مثال ۹.۲ *عدد $\widehat{x} = ۱/۷۳۲$ را به عنوان تقریبی از $x = \sqrt{۳}$ در نظر گرفته، داریم*

$$e_r(x) = \frac{\left|\sqrt{۳} - ۱/۷۳۲\right|}{\sqrt{۳}} = \frac{۱/۷۳۲۰۵۰۸۰۷۵\cdots - ۱/۷۳۲}{۱/۷۳۲۰۵۰۸۰۷۵\cdots} = \frac{۰/۰۰۰۰۵۰۸۰۷۵\cdots}{۱/۷۳۲۰۵۰۸۰۷۵\cdots}.$$

پس

$$e_r(x) = ۰/۰۰۰۰۲۹۳۳۳۷\cdots < ۰/۰۰۰۰۳.$$

اما با توجه به قضیه ۳.۲ و $b_a = ۰/۰۰۰۱$ می‌توان نوشت

$$e_r(x) \leq \frac{۰/۰۰۰۱}{۱/۷۳۲ - ۰/۰۰۰۱} = \frac{۰/۰۰۰۱}{۱/۷۳۱۹} = ۰/۰۰۰۰۵۷۷۴۰۰۵\cdots < ۰/۰۰۰۰۶$$

و یا

$$e_r(x) \leq \frac{۰/۰۰۰۱}{۱/۷۳۲} = ۰/۰۰۰۰۵۷۷۳۶۷۲\cdots < ۰/۰۰۰۰۶.$$

△

تعریف ۱۰.۲ *اگر $\widehat{x} = (۰/a_۱ \cdots a_t)_\beta \times \beta^e$ تقریبی برای x باشد، بزرگترین عدد طبیعی n (حداکثر t) که در نابرابری*

$$e_a(x) \leq ۰/۵ \times \beta^{e-n},$$

صدق کند، تعداد ارقام بامعنای درست $\widehat{x}$ نسبت به x نامیده می‌شود.

پرسش ۴.۲ *در نظر بگیرید $\widehat{x}$ تقریبی از x باشد. آیا هر چه تعداد ارقام بامعنای $\widehat{x}$ بیشتر باشد، می‌توان گفت $\widehat{x}$ تقریب بهتری است؟*
پاسخ. *خیر. با توجه به تعریف، هرچه تعداد ارقام بامعنای درست $\widehat{x}$ نسبت به x بیشتر باشد، خطای $\widehat{x}$ کمتر است.*

مثال ۱۰.۲ *از اعداد $\widehat{x} = ۹۹/۹۶$ و $\widetilde{x} = ۱۰۰/۷$ کدام یک تقریب بهتری برای $x = ۱۰۰$ است؟*

$$\widehat{x} = ۹۹/۹۶ \quad \rightarrow \quad e = ۲, \quad |x - \widehat{x}| = ۰/۰۴ \leq ۰/۵ \times ۱۰^{۲-n} \quad \rightarrow \quad n = ۳,$$
$$\widetilde{x} = ۱۰۰/۷ \quad \rightarrow \quad e = ۳, \quad |x - \widetilde{x}| = ۰/۷ \leq ۰/۵ \times ۱۰^{۳-n} \quad \rightarrow \quad n = ۲.$$

بنابراین $\widehat{x}$ تقریب بهتری برای x است. △

مثال ۱۱.۲ *تعداد ارقام بامعنای درست $\widehat{x} = ۱۰۰/۳۱$ را نسبت به $x = ۱۰۰/۳۱۰۴$ مشخص کنید.*

$$\widehat{x} = ۱۰۰/۳۱ \quad \rightarrow \quad e = ۳, \quad |x - \widehat{x}| = ۰/۰۰۰۴ \leq ۰/۵ \times ۱۰^{۳-n} \quad \rightarrow \quad n = ۶,$$

و چون $\widehat{x}$ فقط پنج رقم بامعنا دارد پس $n = ۵$. △

قضیه ۴.۲ *اگر $\widehat{x}$ گردشده x تا n رقم بامعنا باشد، آن‌گاه $\widehat{x}$ دارای n رقم بامعنای درست است.*

قضیه ۵.۲ *اگر کمیت تقریبی و دقیق را در(به) توانی از مبنا ضرب (تقسیم) کنیم، خطای نسبی تغییر نمی‌کند، ولی خطای مطلق در(به) آن توان ضرب (تقسیم) می‌شود. همچنین تعداد ارقام بامعنای درست نیز تغییر نمی‌کند.*

ارتباط خطای نسبی(دقت) با تعداد ارقام بامعنای درست، در قضیه بعد بیان شده است.

قضیه ۶.۲ *اگر کمیت تقریبی $\widehat{x}$ دارای n رقم بامعنای درست نسبت به کمیت دقیق x باشد، آن‌گاه $e_r(x) < ۰/۵ \times \beta^{۱-n}$ به شرط آن که ارقام بامعنای درست $\widehat{x}$ از یک رقم ۱ و $n-۱$ رقم صفر در جلوی آن تشکیل نشده باشد. برعکس اگر $e_r(x) \leq ۰/۵ \times \beta^{-n}$، آن‌گاه $\widehat{x}$ نسبت به x، دارای دست کم n رقم بامعنای درست است.*

مثال ۱۲.۲ *تقریبی از $\sqrt{۳}$ ارایه دهید که خطای نسبی آن از $۱۰^{-۴}$ کمتر باشد.*
بنابر قضیه ۲.۶ اگر $\widehat{x}$ تقریبی از $x = \sqrt{۳}$ باشد که ۵ رقم بامعنای درست داشته باشد آن‌گاه $e_r(x) < ۵ \times ۱۰^{-۵} < ۱۰^{-۴}$. از این‌رو، با توجه به قضیه ۲.۴، کافی است $\widehat{x}$ گردشده $\sqrt{۳}$ تا ۵ رقم بامعنا باشد، یعنی $\widehat{x} = ۱/۷۳۲۱$. △

تمرین ۴.۲ *نشان دهید*

$$e_a(r(x)) \leq ۰/۵ \times \beta^{e-t}, \qquad e_r(r(x)) \leq ۰/۵ \times \beta^{۱-t} = ۰/۵ \times \epsilon_M. \tag{۳.۲}$$

اگر $r(x)$ از روش قطع کردن به دست آمده باشد، این کران‌های خطا چگونه تغییر می‌کنند؟

تعریف ۱۱.۲ *کمیت $۰/۵ \times \epsilon_M$ به رُند واحد[۲۰] یا واحد گردکردن[۲۱] یا دقت ماشین[۲۲] معروف است. به نظر می‌رسد دلیل این نام‌گذاری آن است که اگر $۰/۵ \times \epsilon_M \leq \epsilon$ آن‌گاه $r(۱ + \epsilon) = ۱$.*

تذکر ۸.۲ *گاهی مواقع اتفاق می‌افتد که $r(x) \notin \mathbb{F}$. برای نمونه به موارد زیر در $\mathbb{F} = \mathbb{F}(۱۰, ۳, -۹, ۹)$ توجه کنید.*

$$r(۰/۱۲۳۴ \times ۱۰^{۱۲}) = ۰/۱۲۳ \times ۱۰^{۱۲} \notin \mathbb{F},$$
$$r(۰/۹۹۹۸ \times ۱۰^{۹}) = ۰/۱۰۰ \times ۱۰^{۱۰} \notin \mathbb{F},$$
$$r(۰/۰۱۲۳ \times ۱۰^{-۹}) = ۰/۱۲۳ \times ۱۰^{-۱۰} \notin \mathbb{F},$$
$$r(۰/۱۲۳۴ \times ۱۰^{-۱۱}) = ۰/۱۲۳ \times ۱۰^{-۱۱} \notin \mathbb{F}.$$

در دو حالت اول سرریز و در دو حالت دوم پاریز اتفاق افتاده است. دو مورد آخر به صورت زیر قابل برطرف کردن هستند

$$r(۰/۰۱۲۳ \times ۱۰^{-۹}) = ۰/۰۱۲ \times ۱۰^{-۹} \in \mathbb{F},$$
$$r(۰/۱۲۳۴ \times ۱۰^{-۱۱}) = ۰ \in \mathbb{F}.$$

ولی متاسفانه دیگر روابط (۳.۲) برقرار نیستند.

رایانه‌ها با پدیده سرریز(پاریز) به عنوان نوعی بی‌نظمی در محاسبات برخورد می‌کنند. وقوع این پدیده‌ها ممکن است به توقف محاسبات منجر شود. با مقیاس کردن مناسب داده‌های ورودی و بررسی نتایج میانی محاسبات، تا حدودی می‌توان از وقوع این پدیده‌ها جلوگیری کرد. به عنوان نمونه برای محاسبه $\|x\| = \sqrt{x_۱^۲ + x_۲^۲}$ عبارت $\|x\| = |x_۱| \sqrt{۱ + (x_۲/x_۱)^۲}$ یا $\|x\| = |x_۲| \sqrt{۱ + (x_۱/x_۲)^۲}$ مناسب‌تر است. چون در برخورد با این پدیده‌ها برای هر روش عددی راه‌کارهای حفاظتی مخصوصی نیاز است و این پدیده‌ها به ندرت اتفاق می‌افتند، فرض می‌کنیم هیچ محدودیتی روی e (توان در نمایش ممیز شناور) نداشته باشیم و از تعریف زیر استفاده می‌کنیم.

تعریف ۱۲.۲ *برای $x \in \mathbb{R}$ به گونه‌ای که $x_{\min} \leq |x| \leq x_{\max}$، تعریف می‌کنیم*

$$r(x) = x(۱ + \epsilon), \qquad |\epsilon| \leq ۰/۵ \times \epsilon_M.$$

بنابراین $x\epsilon$ خطای مطلق تقریب x با $r(x)$ و ϵ خطای نسبی این تقریب است.

---
[۲۰] Unit round
[۲۱] Roundoff unit
[۲۲] Machine precision

تذکر ۹.۲ *به نظر می‌رسد ایده این تعریف از خطای نسبی آمده در (۳.۲) ناشی شده است. زیرا داریم*

$$\frac{|x - r(x)|}{|x|} \leq ۰/۵ \times \epsilon_M,$$

و یا

$$-۰/۵ \times \epsilon_M \leq \frac{x - r(x)}{x} \leq ۰/۵ \times \epsilon_M,$$

و از آنجا

$$(-۰/۵ \times \epsilon_M)x \leq x - r(x) \leq (۰/۵ \times \epsilon_M)x,$$

در نتیجه

$$x(۱ - ۰/۵ \times \epsilon_M) \leq r(x) \leq x(۱ + ۰/۵ \times \epsilon_M).$$

تعریف ۱۳.۲ *اعمال جمع، منها، ضرب و تقسیم در یک ماشین محاسباتی به صورت اعمال ممیز شناور به کار می‌روند که برای $x, y \in \mathbb{F}$ به صورت زیر تعریف می‌شوند*

$$\begin{cases} r(x+y) = (x+y)(۱+\epsilon_۱), \\ r(x-y) = (x-y)(۱+\epsilon_۲), \\ r(x*y) = (x*y)(۱+\epsilon_۳), \\ r(x/y) = (x/y)(۱+\epsilon_۴), \end{cases} \qquad |\epsilon_i| \leq ۰/۵ \times \epsilon_M, \quad i = ۱,۲,۳,۴.$$

تذکر ۱۰.۲ *اگر E یک عبارت شامل تعدادی اعمال ریاضی باشد، منظور از $r(E)$ آن است که تمام اعمال در آن عبارت باید به صورت اعمال ممیز شناور در نظر گرفته شوند. به عنوان نمونه داریم*

$$r(x * (y + z)) = (x * ((y + z)(۱ + \epsilon_۱)))(۱ + \epsilon_۲), \qquad \forall x, y, z \in \mathbb{F},$$

که در آن $|\epsilon_۱|, |\epsilon_۲| \leq ۰/۵ \times \epsilon_M$. همچنین منظور از $r(\sqrt{x})$، $r(\sin x)$ و سایر اعمال مشابه، اعمال r بر روی نتیجه توابع تقریبی نظیر است که در ماشین استفاده می‌شود.

۳.۲ انتشار خطا

بسیاری از اصول موضوعه میدان اعداد حقیقی مانند یکتایی عضو خنثی، شرکت‌پذیری و غیره، در مجموعه اعداد ماشینی برقرار نیستند. به عبارتی روابطی که از نظر ریاضی معادل هستند از نظر محاسباتی نتایج متفاوتی تولید می‌کنند. بنابراین از نظر عددی مهم است که بین طرح‌های ارزیابی مختلف حتی اگر از دیدگاه نظری معادل باشند تفاوت قائل باشیم. به مثال زیر از مرجع [۲۲] توجه کنید.

مثال ۱۳.۲ *فرض کنید می‌خواهیم در یک ماشین با $t = ۸$ و $\beta = ۱۰$، مجموع سه عدد زیر را به دست آوریم.*

$$a = ۰/۲۳۳۷۱۲۵۸ \times ۱۰^{-۴}, \qquad b = ۰/۳۳۶۷۸۴۲۹ \times ۱۰^{۲}, \qquad c = -۰/۳۳۶۷۷۸۱۱ \times ۱۰^{۲}.$$

با یک مقایسه ساده با جواب دقیق یعنی $a+b+c = ۰/۶۴۱۳۷۱۲۵۸ \times ۱۰^{-۳}$*، متوجه می‌شویم که مقدار*

$$r(a+(b+c)) = r(۰/۲۳۳۷۱۲۵۸ \times ۱۰^{-۴} + ۰/۶۱۸۰۰۰۰۰ \times ۱۰^{-۳}) = ۰/۶۴۱۳۷۱۲۶ \times ۱۰^{-۳},$$

دقیق‌تر از مقدار

$$r((a+b)+c) = r(۰/۳۳۶۷۸۴۵۲ \times ۱۰^{۲} - ۰/۳۳۶۷۷۸۱۱ \times ۱۰^{۲}) = ۰/۶۴۱۰۰۰۰۰ \times ۱۰^{-۳},$$

است. در واقع خطای ناشی از رابطه اخیر به علت پدیده از بین رفتن ارقام بامعنا[۲۳] *است. این پدیده در تفاضل دو عدد هم علامت نزدیک به هم اتفاق می‌افتد.* △

تعریف ۱۴.۲ *اعمال جمع، منها، ضرب و تقسیم و همچنین اعمال* $\sqrt{\cdot}$، $\sin$ *و غیره به اعمال مقدماتی*[۲۴] *معروف هستند.*

حال قصد داریم مفهوم الگوریتم را قدری دقیق‌تر بیان کنیم و به همین منظور فرض کنید در یک مسئله نتایج $y_۱, \ldots, y_m$ از داده‌های ورودی $x_۱, \ldots, x_n$ به دست می‌آیند. با تعریف

$$x = [x_۱ \cdots x_n]^T, \qquad y = [y_۱ \cdots y_m]^T,$$

حل مسئله یعنی تعیین $y = \phi(x)$ که در آن $\phi : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$ یک تابع برداری چندمتغیره است، یعنی

$$y_j = \phi_j(x_۱, \ldots, x_n), \qquad j = ۱, \ldots, m.$$

فرض کنید ϕ_j و مشتقات جزئی مرتبه اول آن برای $j = ۱, \ldots, m$، بر Ω پیوسته باشند، $\widehat{x}$ تقریبی از x باشد،

$$e_a(x) = x - \widehat{x}, \qquad e_a(x_i) = x_i - \widehat{x}_i, \qquad i = ۱, \ldots, n,$$

بیانگر خطای مطلق و

$$e_r(x_i) = \frac{x_i - \widehat{x}_i}{x_i}, \qquad x_i \neq ۰, \qquad i = ۱, \ldots, n,$$

نشان‌دهنده خطای نسبی باشد. به کمک بسط تیلور توابع چندمتغیره داریم

$$e_a(y_j) = y_j - \widehat{y}_j = \phi_j(x) - \phi_j(\widehat{x}) \doteq \sum_{i=۱}^{n} (x_i - \widehat{x}_i) \frac{\partial \phi_j(\widehat{x})}{\partial x_i} = \sum_{i=۱}^{n} e_a(x_i) \frac{\partial \phi_j(\widehat{x})}{\partial x_i}, \quad j = ۱, \ldots, m,$$

که می‌توان آن را به شکل فشرده $e_a(y) \doteq J e_a(x)$ نوشت که در آن

$$J = \left[\frac{\partial \phi_j(\widehat{x})}{\partial x_i} \right]_{m \times n},$$

---

[۲۳] Cancellation

[۲۴] Elementary operations

ماتریس ژاکوبین تابع ϕ و منظور از $\doteq$ تقریب مرتبه اول است. کمیت $\frac{\partial \phi_j(\widehat{x})}{\partial x_i}$ نشان‌دهنده میزان حساسیت y_j نسبت به اختلال $e_a(x_i)$ از x_i است. حال اگر $x_i \neq ۰$ و $y_j \neq ۰$ آن‌گاه می‌توان نوشت

$$e_r(y_j) \doteq \sum_{i=۱}^{n} \frac{x_i}{\phi_j(\widehat{x})} \frac{\partial \phi_j(\widehat{x})}{\partial x_i} e_r(x_i), \qquad j = ۱, \ldots, m,$$

کمیت $\frac{x_i}{\phi_j(\widehat{x})} \frac{\partial \phi_j(\widehat{x})}{\partial x_i}$ نشان‌دهنده آن است که خطای نسبی $e_r(x_i)$ به چه اندازه روی خطای نسبی $e_r(y_j)$ اثرگذار است. این ضرایب بزرگ‌نمایی نسبت به تغییرات $x_i \to \lambda x_i$ و $y_j \to \mu y_j$ ثابت باقی می‌مانند و به اعداد حالت(وضعیت)[۲۵] معروف هستند. اگر تمام اعداد حالت کوچک باشند، مسئله $y = \phi(x)$ خوش‌حالت و در غیر این صورت بدحالت است. این آنالیز خطا به آنالیز خطای دیفرانسیلی[۲۶] یا آنالیز خطای پیشرو[۲۷] معروف است و در ارتباط با آن باید به نکات زیر توجه داشت

- در مقابل این آنالیز خطا، یک آنالیز خطای قدیمی موسوم به آنالیز خطای پسرو[۲۸] در برخی مراجع یافت می‌شود که نخستین بار ویلکینسون[۲۹] در مباحث جبرخطی مطرح کرد. در این آنالیز خطا سعی بر آن است که $e_a(x)$ به گونه‌ای یافت شود که $y + e_a(y) = \phi(x + e_a(x))$،
- این طور نیست که یک مسئله یا خوش‌حالت باشد یا بدحالت، بلکه بسته به داده‌های مسئله، ممکن است مسئله خوش‌حالت یا بدحالت باشد،
- برای یک مسئله بدحالت، تغییرات کوچک روی داده‌ها به خطای بزرگی روی جواب منجر می‌شود،
- این تعریف خوش‌حالت بودن خیلی کلی است و دو ایراد دارد

  ۱. باید $x_i \neq ۰$ و $y_j \neq ۰$،

  ۲. باید mn عدد بررسی شوند که برای برخی مسایل مانند مسئله حل دستگاه خطی ممکن است بررسی آنها زمان‌گیر باشد،

  به همین دلیل بهتر است با توجه به مسئله، حالت آن بررسی شود. به عنوان مثال در مسئله حل دستگاه خطی $Ax = b$ کمیت $\kappa = \|A\| \, \|A^{-۱}\|$ عدد وضعیت مسئله است و با توجه به بزرگ یا کوچک بودن آن، مسئله بدحالت یا خوش‌حالت خواهد بود.

به کمک آنالیز خطای پیشرو می‌توان به دو پرسش زیر پاسخ داد.

۱. (مستقیم) اگر خطای مطلق(نسبی) x_iها معلوم باشد، خطای مطلق(نسبی) y_jها چقدر خواهد بود؟

۲. (وارون) برای آنکه خطای مطلق(نسبی) y_jها از ϵ کمتر باشد، حداکثر خطای مطلق(نسبی) x_iها چقدر می‌تواند باشد؟

مثال ۱۴.۲ *یک استوانه به شعاع قاعده $\frac{۴}{۳}$ و ارتفاع $\sqrt{۲}$ را در نظر بگیرید. اگر شعاع و ارتفاع استوانه و عدد π را با دقت $۵S$[۳۰] وارد محاسبات کنیم، حجم این استوانه با چه خطایی به دست می‌آید؟*

می‌دانیم حجم یک استوانه به کمک قاعده $\pi r^۲ h$ تعیین می‌شود که در آن r شعاع قاعده و h ارتفاع استوانه است. اگر تعریف کنیم

$$z = V(p, r, h) = p r^۲ h$$

---

[۲۵] Condition numbers
[۲۶] Differential error analysis
[۲۷] Forward error analysis
[۲۸] Backward error analysis
[۲۹] Wilkinson
[۳۰] منظور ۵ رقم بامعنا است. S حرف نخست واژه significat است.

آن‌گاه باید V به ازای $p=\pi$، $r=\frac{۴}{۳}$ و $h=\sqrt{۲}$ ارزیابی شود. چون دقت $۵S$ مدنظر است

$$\widehat{p}=۳/۱۴۱۶, \qquad \widehat{r}=۱/۳۳۳۳, \qquad \widehat{h}=۱/۴۱۴۲,$$

به طوری که $e_a(p), e_a(r), e_a(h) \leq ۰/۵ \times ۱۰^{-۴}$. بنابراین

$$\widehat{z}=V(\widehat{p},\widehat{r},\widehat{h})=۳/۴۱۱۶ \times ۱/۳۳۳۳^{۲} \times ۱/۴۱۴۲ = ۷/۸۹۸۰.$$

از طرفی

$$e_a(z) \doteq e_a(p)\widehat{r}^{۲}\widehat{h} + ۲e_a(r)\widehat{p}\widehat{r}\widehat{h} + e_a(h)\widehat{p}\widehat{r}^{۲} \leq (\widehat{r}^{۲}\widehat{h} + ۲\widehat{p}\widehat{r}\widehat{h} + \widehat{p}\widehat{r}^{۲}) \times ۰/۵ \times ۱۰^{-۴} \lesssim ۰/۰۰۱۰.$$

در نتیجه حجم استوانه با خطایی کمتر از $۰/۰۰۱۰$، برابر است با $۷/۸۹۸۰$. به کمک یک ماشین حساب ۱۰ رقمی نتیجه $z=۷/۸۹۸۴۵۸۵۵۵$ به دست می‌آید و بنابراین خطای واقعی عبارت است از

$$e_a(z)=|z-\widehat{z}|=۴/۵۸۵۵۵۴ \times ۱۰^{-۴} < ۰/۰۰۱۰.$$

$\triangle$

تمرین ۵.۲ یک استوانه به شعاع قاعده $\frac{۴}{۳}$ و ارتفاع $\sqrt{۲}$ در نظر بگیرید. اگر بخواهیم حجم این استوانه را با دقت $۴S$ به دست آوریم، شعاع قاعده و ارتفاع استوانه و حتی عدد π را با چه دقتی وارد محاسبات کنیم؟

مثال ۱۵.۲ اعداد $x=\sqrt{۵}$ و $y=\frac{\pi}{۱۱}$ را با چه دقتی در نظر بگیریم تا مقدار $۶x^{۲}(\ln x+\sin ۲y)$ با خطایی حداکثر به اندازه $۰/۵\times۱۰^{-۲}$ حساب شود.

اگر فرض کنیم

$$z=\phi(x,y)=۶x^{۲}(\ln x+\sin ۲y),$$

آن‌گاه

$$\frac{\partial\phi}{\partial x}=۱۲x(\ln x+\sin ۲y)+۶x, \qquad \frac{\partial\phi}{\partial y}=۱۲x^{۲}\cos ۲y,$$

و در نتیجه با فرض $x=۲/۲$ و $y=۰/۳$ (x و y را با دقت $۱S$ یا $۲S$ در نظر می‌گیریم) داریم

$$e_a(z) \doteq \left| e_a(x)\frac{\partial f(۲/۲,۰/۳)}{\partial x} + e_a(y)\frac{\partial f(۲/۲,۰/۳)}{\partial y} \right| \simeq \left| ۴۸/۹e_a(x) + ۴۷/۹e_a(y) \right|,$$

و برای برقراری نابرابری $e_a(z) \leq ۰/۵ \times ۱۰^{-۲}$ باید داشته باشیم

$$\left| ۴۸/۹e_a(x) + ۴۷/۹e_a(y) \right| \leq ۰/۵ \times ۱۰^{-۲}.$$

یک جواب نامعادله اخیر عبارت است از

$$e_a(x) \leq ۰/۵ \times ۱۰^{-۴}, \qquad e_a(y) \leq ۰/۵ \times ۱۰^{-۴}.$$

△

مثال ۱۶.۲ *برای محاسبه* $y = \phi(a,b,c) = a^۲b^۳c^۴$ *داریم* $e_r(y) \simeq ۲e_r(a) + ۳e_r(b) + ۴e_r(c)$. △

به دو مثال بعد از مرجع [۲۲] توجه کنید.

مثال ۱۷.۲ *مسئله* $y = \phi(a,b,c) = a + b + c$ *با توجه به*

$$e_r(y) \doteq \frac{a}{a+b+c}e_r(a) + \frac{b}{a+b+c}e_r(b) + \frac{c}{a+b+c}e_r(c),$$

خوش‌حالت است اگر جمعوندهای a, b, c *در مقایسه با* $a+b+c$ *کوچک باشند. به عنوان نمونه اگر* $a = ۱۰^n$، $b = -a$ *و* $c = ۱۰^{-n}$، *آن‌گاه هرچه* n *بزرگتر باشد، مسئله بدحالت‌تر خواهد بود.* △

مثال ۱۸.۲ *به منظور یافتن ریشه با اندازه کوچکتر معادله* $y^۲ + ۲py - q = ۰$، *مسئله زیر را در نظر می‌گیریم*

$$y = \phi(p,q) = -p + \sqrt{p^۲ + q}.$$

چون $\frac{\partial \phi}{\partial p} = -۱ + \frac{p}{\sqrt{p^۲+q}} = \frac{-y}{\sqrt{p^۲+q}}$ *و* $\frac{\partial \phi}{\partial q} = \frac{۱}{۲\sqrt{p^۲+q}}$ *داریم*

$$e_r(y) \doteq \frac{-p}{\sqrt{p^۲+q}}e_r(p) + \frac{q}{۲y\sqrt{p^۲+q}}e_r(q) = \frac{-p}{\sqrt{p^۲+q}}e_r(p) + \frac{p+\sqrt{p^۲+q}}{۲\sqrt{p^۲+q}}e_r(q).$$

برای $q \geq ۰$، *اندازه ضرایب بزرگ‌نمایی (اعداد حالت) حداکثر یک است و در نتیجه مسئله خوش‌حالت است ولی برای* $q < ۰$، *هر چه* q *به* $-p^۲$ *نزدیکتر باشد، مسئله بدحالت‌تر خواهد بود.* △

ماتریس هیلبرت(در مسایل تقریب ظاهر می‌شود)، ماتریس واندرموند(در درونیابی چندجمله‌ای به دست می‌آید) و چندجمله‌ای خائن ویلکینسون(در ریشه‌یابی به آن برخورد می‌کنیم) نمونه‌هایی از معروفترین مسایل هستند که از خود بدحالتی آشکاری نشان می‌دهند.

۱.۳.۲ خطای اعمال ریاضی

به کمک آنالیز خطای پیشرو به سادگی می‌توان تخمین خطای چهار عمل اصلی را به صورت زیر به دست آورد

$$\begin{aligned}
&y = \phi(x_۱, x_۲) = x_۱ \pm x_۲, \quad e_a(y) \doteq e_a(x_۱) \pm e_a(x_۲), && e_r(y) \doteq \frac{x_۱}{x_۱ \pm x_۲}e_r(x_۱) \pm \frac{x_۲}{x_۱ \pm x_۲}e_r(x_۲),\\
&y = \phi(x_۱, x_۲) = x_۱ \times x_۲, \quad e_a(y) \doteq x_۲e_a(x_۱) + x_۱e_a(x_۲), && e_r(y) \doteq e_r(x_۱) + e_r(x_۲),\\
&y = \phi(x_۱, x_۲) = x_۱/x_۲, \quad e_a(y) \doteq \frac{x_۲e_a(x_۱) - x_۱e_a(x_۲)}{x_۲^۲}, && e_r(y) \doteq e_r(x_۱) - e_r(x_۲).
\end{aligned}$$

مثال ۱۹.۲ *اگر اعداد مثبت a و b هر یک n رقم بامعنای درست نسبت به اعداد A و B داشته باشند، حداقل تعداد ارقام بامعنای درست $a+b$ و ab را تعیین کنید.*

بنابر قضیه ۶.۲، $e_r(a) \le ۵ \times ۱۰^{-n}$ و $e_r(b) \le ۵ \times ۱۰^{-n}$. با توجه به کران خطای نسبی عمل جمع داریم

$$e_r(a+b) \le \max\{e_r(a), e_r(b)\} \le ۵ \times ۱۰^{-n} = ۰/۵ \times ۱۰^{-(n-۱)}.$$

پس $a+b$ حداقل $n-۱$ رقم بامعنای درست دارد. به علاوه می‌توان نوشت

$$e_r(ab) \le e_r(a) + e_r(b) = ۵ \times ۱۰^{-n} + ۵ \times ۱۰^{-n} = ۱۰ \times ۱۰^{-n} = ۱۰^{-(n-۲)-۱} < ۰/۵ \times ۱۰^{-(n-۲)}.$$

بنابراین ab دست کم $n-۲$ رقم بامعنای درست دارد. △

تذکر ۱۱.۲ *با توجه به کران خطای نسبی عمل منها، باید از تفاضل دو عدد هم‌علامت نزدیک به هم جلوگیری کرد زیرا موجب می‌شود ارقام بامعنای نتیجه از بین برود. به وضوح ضرایب بزرگ‌نمایی (اعداد حالت) بزرگ هستند و تولید چنین خطایی ممکن است زیانبار نباشد ولی انتشار آن می‌تواند خطرناک (به این مفهوم که رشد خطای زیادی به همراه دارد) باشد. باید توجه داشت که با یک بررسی ساده مشخص می‌شود که تفاضل دو عدد هم‌علامت که نزدیک به هم نباشند خطرناک نیست. بعضی مواقع می‌توان جهت اجتناب از عمل تفاضل در محاسبات، از اتحادها کمک گرفت. به عنوان مثال*

$$e^{x-y} = \frac{e^x}{e^y}, \qquad ۱ - \cos x = ۲\sin^۲\frac{x}{۲}, \qquad \ln x - \ln y = \ln\frac{x}{y}.$$

اگر عمل منها اجتناب‌پذیر نباشد باید سعی کرد عمل منها با دقت دو برابر (یا بیشتر) انجام شود.

مثال ۲۰.۲ *در صورتی که $g(x) = x\left(\sqrt[۳]{۱+\frac{۱}{x}} - ۱\right)$، مقدار $g(۱۰^۹)$ را با دقت $۱۰S$ به دست آورید.*

به سادگی داریم

$$۱ + \frac{۱}{x} = ۱/۰۰۰۰۰۰۰۰۱, \quad \sqrt[۳]{۱+\frac{۱}{x}} = ۱/۰۰۰۰۰۰۰۰۰.$$

بنابراین ماشین حساب[۳۱] نتیجه زیر را به دست می‌دهد

$$g(۱۰^۹) = ۱۰^۹(۱/۰۰۰۰۰۰۰۰۰ - ۱) = ۰/۰۰۰۰۰۰۰۰۰.$$

به این ترتیب با یک ماشین حساب ۱۰ رقمی $g(۱۰^۹) = ۰/۰۰۰۰۰۰۰۰۰۰$، که نتیجه‌ای نادرست است. در حالی که اگر از یک رایانه استفاده شود، نتیجه درست $g(۱۰^۹) = ۰/۳۳۳۳۳۳۳۳۳۳$ به دست می‌آید. دلیل این خطای فاحش، تفاضل دو عدد نزدیک به هم در محاسبات است که به از بین رفتن ارقام بامعنا منجر می‌شود. برای به دست آوردن تقریب بهتر، روش محاسبه را به صورت زیر تغییر می‌دهیم

$$g(x) = x\left(\sqrt[۳]{۱+\frac{۱}{x}} - ۱\right) \times \frac{\sqrt[۳]{(۱+\frac{۱}{x})^۲} + \sqrt[۳]{۱+\frac{۱}{x}} + ۱}{\sqrt[۳]{(۱+\frac{۱}{x})^۲} + \sqrt[۳]{۱+\frac{۱}{x}} + ۱} = \frac{۱}{\sqrt[۳]{(۱+\frac{۱}{x})^۲} + \sqrt[۳]{۱+\frac{۱}{x}} + ۱}.$$

[۳۱] ماشین حساب‌های جدید مقدار متفاوتی به دست می‌دهند.

در این صورت با توجه به

$$\left(۱+\frac{۱}{x}\right)^{۲}=۱/۰۰۰۰۰۰۰۰۲,\quad \sqrt[۲]{\left(۱+\frac{۱}{x}\right)^{۲}}=۱/۰۰۰۰۰۰۰۰۰,$$

نتیجه زیر به دست می‌آید

$$g(۱۰^{۹})=\frac{۱}{۱/۰۰۰۰۰۰۰۰۰+۱/۰۰۰۰۰۰۰۰۰+۱}=۰/۳۳۳۳۳۳۳۳۳.$$

△

این مثال نه تنها نشان می‌دهد که ممکن است یک ماشین نتایج نادرستی تولید کند بلکه به خوبی نشان می‌دهد که به دلیل خطای گرد کردن، ممکن است محاسبه دو عبارت مختلف که از نظر ریاضی هم‌ارز هستند به نتایج متفاوتی منجر شود. از این رو باید از نظر عددی بین الگوریتم‌هایی که از نظر ریاضی هم‌ارز هستند تفاوت قائل شویم.

تذکر ۱۲.۲ *با توجه به خطای نسبی اعمال ضرب و تقسیم، باید در محاسبات از ضرب (تقسیم) اعداد بزرگ (کوچک) در (به) اعداد تقریبی پرهیز کرد.*

مثال ۲۱.۲ *در محاسبه ۱۰۰۰۰π با توجه به*

$$\begin{aligned}\pi=۳/۱۴ \quad &\rightarrow \quad ۱۰۰۰۰\times\pi=۳۱۴۰۰,\\ \pi=۳/۱۴۲ \quad &\rightarrow \quad ۱۰۰۰۰\times\pi=۳۱۴۲۰,\\ \pi=۳/۱۴۱۶ \quad &\rightarrow \quad ۱۰۰۰۰\times\pi=۳۱۴۱۶,\end{aligned}$$

درمی‌یابیم که در تقریب اول خطایی به اندازه ۱۶ واحد، در تقریب دوم خطایی نزدیک به ۴ واحد و در تقریب سوم خطایی کمتر از ۰/۱ مرتکب شده‌ایم. △

تذکر ۱۳.۲ *چون هر عمل محاسباتی خطایی به همراه دارد، یک قاعده کلی دیگر آن است که از حجم محاسبات تا آنجا که ممکن است کاسته شود.*

مثال ۲۲.۲ *به جای عبارت $a_۰+a_۱x+\cdots+a_nx^n$ از عبارت $(\cdots(a_nx+a_{n-۱})x+\cdots+a_۱)x+a_۰$ استفاده شود که تعداد ضرب کمتری دارد. این راه‌کار به قاعده هورنر[۳۲] معروف است.* △

۲.۳.۲ تقریب توابع یک متغیره

به کمک آنالیز خطای پیشرو می‌توان کران خطای برخی از اعمال را به سادگی به دست آورد. به عنوان نمونه در محاسبه جذر یک عدد داریم

$$y=\phi(x)=\sqrt{x},\qquad e_a(y)\doteq\frac{۱}{۲\sqrt{x}}e_a(x),\qquad e_r(y)\doteq\frac{۱}{۲}e_r(x).$$

[۳۲] Horner's rule

بنابر قضیه تیلور، می‌توان به جای کار کردن با یک تابع پیچیده از چندجمله‌ای تیلور نظیر آن استفاده کرد. مثال‌هایی که در ادامه خواهند آمد چگونگی این تقریب را نشان می‌دهند.

مثال ۲۳.۲ *مطلوب است محاسبه مقدار $e^{\frac{\pi}{۱۰}}$ با دقت $۳S$.*

به کمک قضیه تیلور داریم

$$e^x = ۱ + x + \frac{x^۲}{۲!} + \frac{x^۳}{۳!} + \cdots + \frac{x^n}{n!} + \frac{x^{n+۱}}{(n+۱)!}e^{\xi},$$

که در آن $۰ < \xi < x$. در نتیجه

$$e^{\frac{\pi}{۱۰}} = ۱ + \frac{\pi}{۱۰} + \frac{(\frac{\pi}{۱۰})^۲}{۲!} + \cdots + \frac{(\frac{\pi}{۱۰})^n}{n!} + \frac{(\frac{\pi}{۱۰})^{n+۱}}{(n+۱)!}e^{\xi},$$

که در آن $۰ < \xi < \frac{\pi}{۱۰} < ۱$. چون e^x تابعی صعودی است پس $۱ = e^۰ < e^{\xi} < e^۱ < ۳$. از طرفی $\frac{\pi}{۱۰} < ۰/۳۱۵$ و بنابراین

$$\text{خطا} \leq \left|\frac{(\frac{\pi}{۱۰})^{n+۱}}{(n+۱)!}e^{\xi}\right| < \frac{۳ \times (۰/۳۱۵)^{n+۱}}{(n+۱)!}.$$

حال باید داشته باشیم $\frac{۳\times(۰/۳۱۵)^{n+۱}}{(n+۱)!} < ۰/۵ \times ۱۰^{-۲}$ که نتیجه می‌دهد $n \geq ۳$. پس

$$e^{\frac{\pi}{۱۰}} \simeq ۱ + ۰/۳۱۵ + \frac{(۰/۳۱۵)^۲}{۲!} + \frac{(۰/۳۱۵)^۳}{۳!} = ۱/۳۷.$$

△

مثال ۲۴.۲ *(همگرایی سریع) می‌خواهیم تابع $\cos x$ را به ازای مقادیر $|x| < \frac{\pi}{۴}$ با دقت $۵D$ ارزیابی کنیم. با توجه به سری تیلور*

$$\cos x = \sum_{i=۰}^{\infty}(-۱)^i \frac{x^{۲i}}{(۲i)!},$$

و این که

$$\left|\frac{x^{۲n}}{(۲n)!}\right| < \frac{(\frac{\pi}{۴})^{۲n}}{(۲n)!} < \frac{۱/۶^{۲n}}{(۲n)!},$$

باید داشته باشیم

$$\frac{۱/۶^{۲n}}{(۲n)!} < ۰/۵ \times ۱۰^{-۵},$$

که نتیجه می‌دهد $n \geq ۶$. △

مثال ۲۵.۲ *(همگرایی کند) با توجه به*

$$\frac{۱}{۱+t} = ۱ - t + t^۲ - t^۳ + - \cdots,$$

داریم

$$\int_۰^x \frac{dt}{۱+t} = \int_۰^x (۱ - t + t^۲ - t^۳ + - \cdots)dt,$$

و یا

$$\ln(۱+x) = x - \frac{x^۲}{۲} + \frac{x^۳}{۳} - \frac{x^۴}{۴} + - \cdots.$$

برای ارزیابی $\ln(1+x)$ *با دقت* $5D$*، باید داشته باشیم*

$$\left|\frac{x^n}{n}\right| < 0/5 \times 10^{-5},$$

که برای $x = 0/99$ *نتیجه می‌دهد* $n \geq 582$*.* △

مثال ۲۶.۲ *تابع* $Si(x) = \int_0^x \frac{\sin t}{t} dt$ *را که در آن* $x \geq 0$*، در نظر بگیرید. دست کم چند جمله از بسط مک‌لورن تابع* $f(t) = \sin t$ *لازم است تا* $Si(1)$ *با دقت* $6D$ *مشخص شود؟*

با اِعمال قضیه تیلور برای تابع f *داریم*

$$Si(1) = \int_0^1 \frac{1}{t}\left(f(0) + \frac{f'(0)}{1!}t + \frac{f''(0)}{2!}t^2 + \cdots + \frac{f^{(k)}(0)}{k!}t^k + \frac{f^{(k+1)}(\xi(t))}{(k+1)!}t^{k+1}\right) dt,$$

و یا

$$Si(1) = \int_0^1 \frac{1}{t}\left(t - \frac{t^3}{3!} + \frac{t^5}{5!} + \cdots + (-1)^n \frac{t^{2n+1}}{(2n+1)!} + (-1)^{n+1}\frac{t^{2n+2}}{(2n+2)!}\sin(\xi(t))\right) dt.$$

در نتیجه

$$Si(1) = \int_0^1 \left(1 - \frac{t^2}{3!} + \frac{t^4}{5!} + \cdots + (-1)^n \frac{t^{2n}}{(2n+1)!}\right) dt + \frac{(-1)^{n+1}}{(2n+2)!}\int_0^1 t^{2n+1}\sin(\xi(t))dt.$$

بنابراین

$$Si(1) = 1 - \frac{1}{3 \times 3!} + \frac{1}{5 \times 5!} + \cdots + (-1)^n \frac{1}{(2n+1)\times(2n+1)!} + E,$$

که در آن

$$E = \frac{(-1)^{n+1}}{(2n+2)!}\int_0^1 t^{2n+1}\sin(\xi(t))dt.$$

حال می‌توان نوشت

$$|E| = \frac{1}{(2n+2)!}\left|\int_0^1 t^{2n+1}\sin(\xi(t))dt\right| \leq \frac{1}{(2n+2)!}\int_0^1 t^{2n+1}\left|\sin(\xi(t))\right| dt.$$

پس

$$|E| \leq \frac{1}{(2n+2)!}\int_0^1 t^{2n+1}dt = \frac{1}{(2n+2)\times(2n+2)!}.$$

از $\frac{1}{(2n+2)\times(2n+2)!} < 0/5 \times 10^{-6}$ *نتیجه می‌شود* $n \geq 4$*. بنابراین*

$$Si(1) \simeq 1 - \frac{1}{3\times 3!} + \frac{1}{5\times 5!} - \frac{1}{7\times 7!} + \frac{1}{9\times 9!},$$

و یا

$$Si(1) \simeq 1 - 0/0555556 + 0/0016667 - 0/0000283 + 0/0000003 = 0/9460831.$$

△

تمرین ۶.۲ *مطلوب است تعیین تقریبی از عدد π با دقت $3S$ به کمک بسط مکلورن تابع $f(x) = \tan^{-1} x$.*

۴.۲ بررسی پایداری یک الگوریتم

این بخش را با تعریف دیگری از الگوریتم آغاز می‌کنیم.

تعریف ۱۵.۲ *یک دنباله متناهی از اعمال مقدماتی که تعیین می‌کنند چگونه جواب یک مسئله به کمک داده‌های ورودی حساب می‌شود، الگوریتم (عددی) نامیده می‌شود.*

در ادامه قصد داریم یک الگوریتم را دقیق‌تر بررسی کنیم. در هر مرحله از محاسبات یک مجموعه عملوندی[۳۳] از اعداد وجود دارد که یا همان داده‌های ورودی (x_i) یا نتایج میانی تولیدشده هستند. یک عمل مقدماتی بر روی یک یا دو عنصر از یک مجموعه عملوندی اثر کرده و یک عدد جدید تولید می‌کند که یا یک نتیجه نهایی (y_j) یا یک نتیجه میانی است. عناصری از مجموعه عملوندی که در ادامه محاسبات به آنها نیازی نیست، از مجموعه عملوندی کنار گذاشته می‌شوند. اولین مجموعه عملوندی فقط شامل داده‌های ورودی یعنی $x_1, \ldots, x_n$ و آخرین مجموعه عملوندی فقط شامل داده‌های خروجی یعنی $y_1, \ldots, y_m$ است. اگرچه در یک مجموعه عملوندی ترتیب (و حتی تکرار) مهم نیست، یک مجموعه عملوندی را با یک بردار نمایش می‌دهیم

$$x^{(0)} = x, \qquad x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_{n_i}^{(i)} \end{bmatrix} \in \mathbb{R}^{n_i}.$$

هر عمل مقدماتی متناظر با یک نگاشت بر روی یک مجموعه عملوندی است و یک عمل مقدماتی را با یک نگاشت مقدماتی به صورت زیر نظیر می‌کنیم

$$\begin{cases} \phi^{(i)} : D_i \subseteq \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}, \\ \phi^{(i)}(x^{(i)}) = x^{(i+1)}. \end{cases}$$

صرف نظر از جایگشت‌های بی‌اهمیت $x^{(i)}$ و $x^{(i+1)}$، $\phi^{(i)}$ به طور یکتا تعریف می‌شود. اعمال مقدماتی متنوع در یک الگوریتم موجب می‌شود که ϕ به یک دنباله از نگاشت‌های مقدماتی تجزیه شود

$$\begin{cases} \phi = \phi^{(r)} \circ \phi^{(r-1)} \circ \cdots \circ \phi^{(0)}, \\ \phi^{(i)} : D_i \subseteq \mathbb{R}^{n_i} \to D_{i+1} \subseteq \mathbb{R}^{n_{i+1}}, \end{cases}$$

که در آن $n_0 = n$ و $n_{r+1} = m$.

مثال ۲۷.۲ *برای مسئله $y = \phi(a, b, c) = a + b + c$ متناظر با الگوریتم $(a + b) + c$ نگاشت‌های*

$$\phi^{(0)}(a, b, c) = \begin{bmatrix} a + b \\ c \end{bmatrix}, \quad \phi^{(1)}(u, v) = u + v,$$

[۳۳] Operand set

و متناظر با الگوریتم $a+(b+c)$ نگاشت‌های

$$\phi^{(0)}(a,b,c)=\begin{bmatrix} a \\ b+c \end{bmatrix}, \quad \phi^{(1)}(u,v)=u+v,$$

در نظر گرفته می‌شوند. △

مثال ۲۸.۲ *برای مسئله $y=\phi(a,b)=a^2-b^2$ متناظر با الگوریتم a^2-b^2 نگاشت‌های*

$$\phi^{(0)}(a,b)=\begin{bmatrix} a^2 \\ b \end{bmatrix}, \quad \phi^{(1)}(u,b)=\begin{bmatrix} u \\ b^2 \end{bmatrix}, \quad \phi^{(2)}(v,w)=v-w,$$

و متناظر با الگوریتم $(a+b)(a-b)$ نگاشت‌های

$$\phi^{(0)}(a,b)=\begin{bmatrix} a \\ b \\ a+b \end{bmatrix}, \quad \phi^{(1)}(a,b,u)=\begin{bmatrix} u \\ a-b \end{bmatrix}, \quad \phi^{(2)}(v,w)=vw,$$

در نظر گرفته می‌شوند که می‌توان آنها را به صورت زیر خلاصه‌تر کرد

$$\phi^{(0)}(a,b)=\begin{bmatrix} a^2 \\ b^2 \end{bmatrix}, \quad \phi^{(1)}(u,v)=u-v, \qquad \phi^{(0)}(a,b)=\begin{bmatrix} a+b \\ a-b \end{bmatrix}, \quad \phi^{(1)}(u,v)=uv.$$

باید توجه داشت که $\phi^{(0)}$هایی که در روابط اخیر تعریف شده‌اند، مقدماتی نیستند ولی چون متناظر با اعمال مقدماتی مستقل از هم هستند، به کار بردن آنها به منظور خلاصه‌تر شدن الگوریتم‌ها مانعی ندارد. △

با این امید که بتوان معیاری برای بررسی کیفیت یک الگوریتم به دست آورد، ابتدا باید دید چرا الگوریتم‌های مختلف در حل یک مسئله، نتایج متفاوتی تولید می‌کنند. به بیان ساده شاید بتوان گفت که انتشار خطا برای الگوریتم‌های مختلف، نقش متفاوتی بازی می‌کند. به مثال زیر از مرجع [۲۲] توجه کنید.

مثال ۲۹.۲ *در ادامه مثال ۱۳.۲ برای $|\epsilon_1|,|\epsilon_2| \le 0/5\times\epsilon_M$ داریم*

$$\widehat{y}=r((a+b)+c)=((a+b)(1+\epsilon_1)+c)(1+\epsilon_2)=(a+b+c)(1+\frac{a+b}{a+b+c}\epsilon_1(1+\epsilon_2)+\epsilon_2),$$

و اگر $e_r(y)=\frac{\widehat{y}-y}{y}$، به وضوح

$$e_r(y)\doteq\frac{a+b}{a+b+c}\epsilon_1+\epsilon_2.$$

برای یک الگوریتم ضرایب $\frac{a+b}{a+b+c}$ و ۱ و برای الگوریتم دیگر ضرایب ۱ و $\frac{b+c}{a+b+c}$، موجب رشد خطاهای گرد کردن میانی می‌شوند. برای اعداد در نظر گرفته‌شده داریم

$$\frac{a+b}{a+b+c}\simeq 0/5\times 10^5, \qquad \frac{b+c}{a+b+c}\simeq 0/97,$$

که نشان می‌دهد چرا الگوریتم $a+(b+c)$ به نتیجه دقیق‌تری منجر شده است. △

در ادامه قصد داریم انتشار خطاهای گرد کردن میانی را در یک الگوریتم بررسی کنیم. همان طور که بیان شد، یک الگوریتم در حل مسئله $y=\phi(x)$ متناظر با دنباله‌ای از نگاشت‌های مقدماتی است، یعنی

$$x=x^{(0)}\to x^{(1)}=\phi^{(0)}(x^{(0)})\to\cdots\to y=x^{(r+1)}=\phi^{(r)}(x^{(r)}).$$

با این فرض که $\phi^{(i)}$ روی D_i به طور پیوسته مشتق‌پذیر باشد، تعریف می‌کنیم

$$\psi^{(i)}:D_i\to\mathbb{R}^m,\qquad \psi^{(i)}=\phi^{(r)}\circ\phi^{(r-1)}\circ\cdots\circ\phi^{(i)}.$$

$\psi^{(i)}$ به نگاشت باقی‌مانده مرحله iام معروف است و $\psi^{(0)}=\phi$. اگر $J\phi^{(i)}$ و $J\psi^{(i)}$ به ترتیب بیانگر ژاکوبین $\phi^{(i)}$ و $\psi^{(i)}$ باشند، به کمک قاعده زنجیره‌ای داریم

$$J\phi(x)=J\phi^{(r)}(x^{(r)})\cdot J\phi^{(r-1)}(x^{(r-1)})\cdots J\phi^{(0)}(x^{(0)}),$$
$$J\psi^{(i)}(x)=J\phi^{(r)}(x^{(r)})\cdot J\phi^{(r-1)}(x^{(r-1)})\cdots J\phi^{(i)}(x^{(i)}).$$

انتظار داریم در مرحله iام، $x^{(i+1)}$ به دست آید ولی با در نظر گرفتن خطای داده‌های ورودی و خطای گرد کردن میانی، $\widehat{x}^{(i+1)}=r(\phi^{(i)}(\widehat{x}^{(i)}))$ تولید می‌شود. بنابراین

$$e_a(x^{(i+1)}):=r(\phi^{(i)}(\widehat{x}^{(i)}))-x^{(i+1)}=(r(\phi^{(i)}(\widehat{x}^{(i)}))-\phi^{(i)}(\widehat{x}^{(i)}))+(\phi^{(i)}(\widehat{x}^{(i)})-\phi^{(i)}(x^{(i)})).\qquad (۴.۲)$$

بنابراین باید دو نوع خطا انتظار داشته باشیم که به خطای (گرد کردن) تولیدشده[۳۴](متناظر با عبارت اول در سمت راست (۴.۲)) و خطای منتشرشده[۳۵](متناظر با عبارت دوم در سمت راست (۴.۲)) معروف هستند.

مشابه آنالیز خطای پیشرو، به سادگی برای خطای منتشرشده داریم

$$\phi^{(i)}(\widehat{x}^{(i)})-\phi^{(i)}(x^{(i)})\doteq J\phi^{(i)}(\widehat{x}^{(i)})e_a(x^{(i)}),$$

و برای محاسبه خطای تولیدشده، ابتدا می‌دانیم برای $j=1,\dots,n_{i+1}$ اگر $\phi_j^{(i)}:D_i\to\mathbb{R}$ یک نگاشت مقدماتی باشد، آنگاه خواهیم داشت

$$r(\phi_j^{(i)}(u))=(1+\epsilon_j)\phi_j^{(i)}(u),\qquad |\epsilon_j|\le 0/5\times\epsilon_M.$$

حال برای

$$\phi^{(i)}(u)=[\phi_1^{(i)}(u)\ \cdots\ \phi_{n_{i+1}}^{(i)}(u)]^T,$$

که یا فقط یک نگاشت مقدماتی یا تعدادی نگاشت مقدماتی مستقل از هم را شامل می‌شود، می‌توان نوشت

$$r(\phi^{(i)}(u))=(I+E_{i+1})\phi^{(i)}(u),$$

---

[۳۴] Produced error
[۳۵] Propageted error

که در آن I یک ماتریس همانی $n_{i+1} \times n_{i+1}$ و $E_{i+1} = \mathrm{diag}[\epsilon_1, \ldots, \epsilon_{n_{i+1}}]$ یک ماتریس قطری $n_{i+1} \times n_{i+1}$ است. بنابراین

$$r(\phi^{(i)}(\widehat{x}^{(i)})) - \phi^{(i)}(\widehat{x}^{(i)}) = E_{i+1}\phi^{(i)}(\widehat{x}^{(i)}),$$

و چون $E_{i+1}\phi^{(i)}(\widehat{x}^{(i)}) \doteq E_{i+1}\phi^{(i)}(x^{(i)})$ (زیرا جملات خطای تقریب $\phi^{(i)}(x^{(i)})$ با $\phi^{(i)}(\widehat{x}^{(i)})$ در عناصر قطری E_{i+1} ضرب شده و خطاهایی با مرتبه بالاتر تولید می‌کنند که قابل چشم‌پوشی هستند)، می‌توان نوشت

$$r(\phi^{(i)}(\widehat{x}^{(i)})) - \phi^{(i)}(\widehat{x}^{(i)}) \doteq E_{i+1}\phi^{(i)}(x^{(i)}) = E_{i+1}x^{(i+1)} := e_{i+1}.$$

در این رابطه e_{i+1} نقش خطای مطلق و E_{i+1} نقش خطای نسبی را بازی می‌کند. در نتیجه خطای مرحله iام عبارت است از

$$e_a(x^{(i+1)}) \doteq e_{i+1} + J\phi^{(i)}(\widehat{x}^{(i)})e_a(x^{(i)}) = E_{i+1}x^{(i+1)} + J\phi^{(i)}(\widehat{x}^{(i)})e_a(x^{(i)}), \qquad i \geq 0.$$

بنابراین با توجه به $e_a(x^{(0)}) = e_a(x)$ داریم

$$\begin{aligned}
e_a(x^{(1)}) &\doteq J\phi^{(0)}(\widehat{x}^{(0)})e_a(x^{(0)}) + e_1,\\
e_a(x^{(2)}) &\doteq J\phi^{(1)}(\widehat{x}^{(1)})(J\phi^{(0)}(\widehat{x}^{(0)})e_a(x^{(0)}) + e_1) + e_2,\\
&\vdots,\\
e_a(x^{(r+1)}) &\doteq J\phi^{(r)}(\widehat{x}^{(r)}) \cdots J\phi^{(0)}(\widehat{x}^{(0)})e_a(x^{(0)}) + J\phi^{(r)}(\widehat{x}^{(r)}) \cdots J\phi^{(1)}(\widehat{x}^{(1)})e_1 + \cdots + J\phi^{(r)}(\widehat{x}^{(r)})e_r + e_{r+1}.
\end{aligned}$$

حال به کمک نگاشت‌های باقی‌مانده، می‌توان اثر خطای داده‌های ورودی $e_a(x)$ و خطای گرد کردن میانی e_i، روی نتیجه نهایی $y = x^{(r+1)} = \phi(x)$ را در رابطه مهم زیر خلاصه کرد

$$\begin{aligned}
e_a(y) &\doteq J\phi(\widehat{x})e_a(x) + J\psi^{(1)}(\widehat{x}^{(1)})e_1 + \cdots + J\psi^{(r)}(\widehat{x}^{(r)})e_r + e_{r+1}\\
&= J\phi(\widehat{x})e_a(x) + J\psi^{(1)}(\widehat{x}^{(1)})E_1x^{(1)} + \cdots + J\psi^{(r)}(\widehat{x}^{(r)})E_rx^{(r)} + E_{r+1}y.
\end{aligned}$$

بنابراین بزرگی اندازه ماتریس ژاکوبین $J\psi^{(i)}(\widehat{x}^{(i)})$ نشان‌دهنده اثر خطای گرد کردن میانی e_i روی جواب است.

مثال ۳۰.۲ *در مثال ۲۸.۲ دو الگوریتم برای حل مسئله $y = \phi(a,b) = a^2 - b^2$ ارائه شده است. برای الگوریتم $a^2 - b^2$ داریم*

$$x = x^{(0)} = \begin{bmatrix} a \\ b \end{bmatrix}, \qquad x^{(1)} = \begin{bmatrix} a^2 \\ b \end{bmatrix}, \qquad x^{(2)} = \begin{bmatrix} a^2 \\ b^2 \end{bmatrix}, \qquad x^{(3)} = y = a^2 - b^2.$$

به سادگی خواهیم داشت

$$\psi^{(1)}(u,v) = u - v^2, \qquad \psi^{(2)}(u,v) = u - v,$$

و از آنجا

$$J\phi(x) = [2a \ -2b]^T, \qquad J\psi^{(1)}(x^{(1)}) = [1 \ -2b]^T, \qquad J\psi^{(2)}(x^{(2)}) = [1 \ -1]^T.$$

بنابراین

$$E_۱ = \begin{bmatrix} \epsilon_۱ & ۰ \\ ۰ & ۰ \end{bmatrix}, \qquad e_۱ = E_۱ x^{(۱)} = \begin{bmatrix} \epsilon_۱ a^۲ \\ ۰ \end{bmatrix},$$

$$E_۲ = \begin{bmatrix} ۰ & ۰ \\ ۰ & \epsilon_۲ \end{bmatrix}, \qquad e_۲ = E_۲ x^{(۲)} = \begin{bmatrix} ۰ \\ \epsilon_۲ b^۲ \end{bmatrix},$$

و $e_۳ = \epsilon_۳(a^۲ - b^۲)$ *به طوری که* $|\epsilon_۱|, |\epsilon_۲|, |\epsilon_۳| \leq ۰/۵ \times \epsilon_M$. *در نتیجه چون* $e_a(x) = [e_a(a)\ e_a(b)]^T$ *خواهیم داشت*

$$e_a(y) \doteq ۲a e_a(a) - ۲b e_a(b) + a^۲\epsilon_۱ - b^۲\epsilon_۲ + (a^۲ - b^۲)\epsilon_۳.$$

برای الگوریتم $(a-b)(a+b)$ *داریم*

$$x = x^{(۰)} = \begin{bmatrix} a \\ b \end{bmatrix}, \qquad x^{(۱)} = \begin{bmatrix} a+b \\ a-b \end{bmatrix}, \qquad x^{(۲)} = y = a^۲ - b^۲.$$

به سادگی خواهیم داشت

$$\psi^{(۱)}(u,v) = uv, \qquad J\psi^{(۱)}(x^{(۱)}) = [a-b\ a+b]^T, \qquad J\phi(x) = [۲a\ -۲b]^T.$$

بنابراین

$$E_۱ = \begin{bmatrix} \epsilon_۱ & ۰ \\ ۰ & \epsilon_۲ \end{bmatrix}, \qquad e_۱ = E_۱ x^{(۱)} = \begin{bmatrix} \epsilon_۱(a+b) \\ \epsilon_۲(a-b) \end{bmatrix},$$

و $e_۲ = \epsilon_۳(a^۲ - b^۲)$ *به طوری که* $|\epsilon_۱|, |\epsilon_۲|, |\epsilon_۳| \leq ۰/۵ \times \epsilon_M$. *در نتیجه خواهیم داشت*

$$e_a(y) \doteq ۲a e_a(a) - ۲b e_a(b) + (a^۲ - b^۲)(\epsilon_۱ + \epsilon_۲ + \epsilon_۳).$$

△

تعریف ۱۶.۲ *برای هر الگوریتم کمیت* $J\psi^{(۱)}(\widehat{x}^{(۱)})e_۱ + \cdots + J\psi^{(r)}(\widehat{x}^{(r)})e_r + e_{r+۱}$ *به اثر کلی خطای گرد کردن*[۳۶] *آن الگوریتم معروف است. از دو الگوریتم در حل یک مسئله، آن الگوریتمی که اثر کلی خطای گرد کردن کوچک‌تری داشته باشد از نظر عددی قابل‌اعتمادتر*[۳۷] *است، البته به شرط آن که داده‌های ورودی هر دو الگوریتم یکسان باشند.*

مثال ۳۱.۲ *در ادامه مثال ۳۰.۲، اثر کلی خطای گرد کردن برای الگوریتم اول حداکثر* $(a^۲ + b^۲ + |a^۲ - b^۲|)\frac{\epsilon_M}{۲}$ *و برای الگوریتم دوم حداکثر* $۳\,|a^۲ - b^۲|\,\frac{\epsilon_M}{۲}$ *است. بنابراین اگر*

$$۳\,|a^۲ - b^۲|\,\frac{\epsilon_M}{۲} < (a^۲ + b^۲ + |a^۲ - b^۲|)\frac{\epsilon_M}{۲},$$

[۳۶] Total effect of rounding
[۳۷] Numerically more trustworthy

که هم‌ارز است با

$$۲\left|a^۲-b^۲\right|<a^۲+b^۲,$$

و یا

$$-a^۲-b^۲<۲(a^۲-b^۲)<a^۲+b^۲,$$

که نتیجه می‌دهد $\frac{۱}{۳}<\left(\frac{a}{b}\right)^۲<۳$ *. در این صورت الگوریتم دوم از الگوریتم اول و در غیر این صورت الگوریتم اول از الگوریتم دوم قابل‌اعتمادتر است. برای* $a=۰/۳۲۳۷$ *و* $b=۰/۳۱۳۴$ *در یک ماشین با* $\beta=۱۰$ *و* $t=۴$ *داریم*

$$r(a^۲-b^۲)=۰/۶۵۸۰\times ۱۰^{-۲},\qquad r((a+b)(a-b))=۰/۶۵۶۲\times ۱۰^{-۲},$$

و چون $\left(\frac{a}{b}\right)^۲\simeq ۱/۰۷$ *پس الگوریتم دوم از نظر عددی قابل‌اعتمادتر است. این مطلب با توجه به مقدار دقیق‌تر* $y=۰/۶۵۶۲۱۳\times ۱۰^{-۲}$ *تائید می‌شود.* △

صرف نظر از اینکه چه الگوریتمی در حل مسئله استفاده می‌شود، کران خطای زیر را داریم

$$|E_{r+۱}y|\leq|y|\,\frac{\epsilon_M}{۲},\qquad |y|=[|y_۱|\ \cdots\ |y_m|].$$

از طرفی چون ممکن است داده‌های ورودی اعداد ماشینی(قابل‌نمایش به طور دقیق) نباشند، کران خطای زیر را می‌توان در نظر گرفت

$$|e_a(x)|\leq|x|\,\frac{\epsilon_M}{۲}.$$

بنابراین بدون توجه به اینکه چه الگوریتمی برای حل مسئله در نظر گرفته می‌شود، باید همواره خطای

$$e_a^\circ(y):=(|J\phi(x)|\,|x|+|y|)\frac{\epsilon_M}{۲},$$

معروف به خطای ذاتی[۳۸] را از هر الگوریتمی انتظار داشته باشیم.

تعریف ۱۷.۲ *خطای گرد کردن* e_i *بی‌ضرر*[۳۹] *است، هرگاه سهم آن در محاسبه* $e_a(y)$*، حداکثر به اندازه خطای ذاتی باشد، یعنی*

$$\left|J\psi^{(i)}(\widehat{x}^{(i)})e_i\right|\simeq e_a^\circ(y),$$

و اگر e_i*ها همه بی‌ضرر باشند، الگوریتم از نظر عددی پایدار(خوش‌رفتار)*[۴۰] *نامیده می‌شود.*

مثال ۳۲.۲ *در ادامه مثال ۳۱.۲، با توجه به خطای ذاتی، یعنی*

$$e_a^\circ(y)=([۲\,|a|\ \ ۲\,|b|]\begin{bmatrix}|a|\\|b|\end{bmatrix}+\left|a^۲-b^۲\right|)\frac{\epsilon_M}{۲}=(۲a^۲+۲b^۲+\left|a^۲-b^۲\right|)\frac{\epsilon_M}{۲},$$

[۳۸] Inherent error
[۳۹] Harmless
[۴۰] Numerically stable(well-behaved)

نتیجه می‌گیریم هر دو الگوریتم از نظر عددی پایدار هستند. △

تذکر ۱۴.۲ *جهت آشنایی با سایر مفاهیم مرتبط مانند گراف نظیر یک الگوریتم، الگوریتم خوش‌خیم*[۴۱] *و غیره به مقاله [۴] مراجعه کنید.*

مثال‌های زیر از مرجع [۲۲] را به دقت بررسی کنید.

مثال ۳۳.۲ *می‌خواهیم ریشه با اندازه کوچکتر معادله* $y^۲ + ۲py - q = ۰$ *را بیابیم که در آن* $p, q > ۰$ *و* $p \gg q$*. یعنی قصد داریم مسئله خوش‌حالت* $y = \phi(p, q) = -p + \sqrt{p^۲ + q}$ *را حل کنیم. ابتدا الگوریتم زیر به ذهن می‌رسد*

$$x = x^{(۰)} = \begin{bmatrix} p \\ q \end{bmatrix},\ x^{(۱)} = \begin{bmatrix} p \\ q \\ p^۲ \end{bmatrix},\ x^{(۲)} = \begin{bmatrix} p \\ p^۲ + q \end{bmatrix},\ x^{(۳)} = \begin{bmatrix} p \\ \sqrt{p^۲ + q} \end{bmatrix},\ x^{(۴)} = y = -p + \sqrt{p^۲ + q},$$

و از آنجا داریم

$$\psi^{(۱)}(u, v, w) = -u + \sqrt{w + v}, \qquad \psi^{(۲)}(u, v) = -u + \sqrt{v}, \qquad \psi^{(۳)}(u, v) = v - u,$$

و بنابراین

$$J\psi^{(۱)}(x^{(۱)}) = [-۱\ \frac{۱}{۲\sqrt{p^۲ + q}}\ \frac{۱}{۲\sqrt{p^۲ + q}}]^T, \quad J\psi^{(۲)}(x^{(۲)}) = [-۱\ \frac{۱}{۲\sqrt{p^۲ + q}}]^T, \quad J\psi^{(۳)}(x^{(۳)}) = [-۱\ ۱]^T,$$

پس

$$e_۱ = \begin{bmatrix} ۰ \\ ۰ \\ \epsilon_۱ p^۲ \end{bmatrix}, \quad e_۲ = \begin{bmatrix} ۰ \\ \epsilon_۲(p^۲ + q) \end{bmatrix}, \quad e_۳ = \begin{bmatrix} ۰ \\ \epsilon_۳\sqrt{p^۲ + q} \end{bmatrix}, \quad e_۴ = \epsilon_۴ y.$$

در نتیجه

$$e_a(y) \doteq \frac{-y}{\sqrt{p^۲ + q}} e_a(p) + \frac{۱}{۲\sqrt{p^۲ + q}} e_a(q) + \frac{p^۲}{۲\sqrt{p^۲ + q}}\epsilon_۱ + \frac{\sqrt{p^۲ + q}}{۲}\epsilon_۲ + \sqrt{p^۲ + q}\epsilon_۳ + e_۴.$$

از طرفی چون

$$e_a^{\circ}(y) = (\frac{py}{\sqrt{p^۲ + q}} + \frac{q}{۲\sqrt{p^۲ + q}} + y)\frac{\epsilon_M}{۲},$$

پس داریم

$$\frac{e_a^{\circ}(y)}{y} = (\frac{p}{\sqrt{p^۲ + q}} + \frac{p + \sqrt{p^۲ + q}}{۲\sqrt{p^۲ + q}} + ۱)\frac{\epsilon_M}{۲},$$

و بنابراین

$$\frac{\epsilon_M}{۲} < \frac{e_a^{\circ}(y)}{y} < ۳\frac{\epsilon_M}{۲}.$$

متناظر با عمل $\sqrt{\ }$ *خطای* $\frac{\sqrt{p^۲+q}}{-p+\sqrt{p^۲+q}}\epsilon_۳ = k\epsilon_۳$ *تولید می‌شود که در آن* $k = \frac{p\sqrt{p^۲+q}+p^۲+q}{q} \simeq \frac{۲p^۲}{q}$ *و چون* $p \gg q$ *به وضوح* $k \gg ۳$ *و این حاکی از مضر بودن خطای* $e_۳$ *و در نتیجه ناپایداری عددی الگوریتم است. حال با داده ورودی* $x = x^{(۰)} = [p\ q]^T$ *الگوریتم زیر را در*

[۴۱] Benign

نظر می‌گیریم

$$x^{(1)} = \begin{bmatrix} p \\ q \\ p^2 \end{bmatrix},\ x^{(2)} = \begin{bmatrix} p \\ q \\ p^2+q \end{bmatrix},\ x^{(3)} = \begin{bmatrix} p \\ q \\ \sqrt{p^2+q} \end{bmatrix},\ x^{(4)} = \begin{bmatrix} q \\ p+\sqrt{p^2+q} \end{bmatrix},\ y = \frac{q}{p+\sqrt{p^2+q}},$$

و از آنجا داریم

$$\psi^{(3)}(u,v,w) = \frac{v}{u+w}, \qquad J\psi^{(3)}(x^{(3)}) = [\frac{-q}{(p+\sqrt{p^2+q})^2}\ \frac{1}{p+\sqrt{p^2+q}}\ \frac{-q}{(p+\sqrt{p^2+q})^2}]^T,$$

و چون $e_3 = [0\ 0\ \epsilon_3\sqrt{p^2+q}]^T$، متناظر با عمل $\sqrt{\cdot}$ خطای $k\epsilon_3 = \frac{-q\sqrt{p^2+q}}{-p+\sqrt{p^2+q}}\epsilon_3$ تولید می‌شود که در آن $k = -\frac{\sqrt{p^2+q}}{p+\sqrt{p^2+q}}$. به وضوح $|k| < 1$ حاکی از بی‌ضرر بودن خطای گرد کردن نظیر است و با بررسی اثر سایر e_iها می‌توان نتیجه گرفت این الگوریتم از نظر عددی پایدار است. برنامه sqr.nb را ببینید. △

مثال ۳۴.۲ برای $k \in \mathbb{N}$ و $x \in \mathbb{R}$ داده‌شده، مقدار $\cos(kx)$ را از رابطه زیر به دست می‌آوریم

$$\cos(m+1)x = 2\cos x\cos(mx) - \cos(m-1)x, \qquad m = 1,\dots,k-1.$$

اگر در محاسبه $\cos x$ خطایی داشته باشیم، چقدر در محاسبه $\cos(kx)$ اثر دارد؟

برای مسئله $y = \phi(x) = \cos(kx)$، با تعریف $x^{(0)} = x$ و $\phi^{(0)}(x) = \cos x := c$ به وضوح داریم

$$\psi^{(1)}(c) = \cos(kx) = \cos(k\cos^{-1}c), \qquad J\psi^{(1)}(c)e_1 = \frac{k\sin(kx)}{\sin x}\cos x\,\epsilon_1 = k\cot x\sin(kx)\epsilon_1, \qquad |\epsilon_1| \le \frac{\epsilon_M}{2}.$$

با توجه به $e_a^*(y) = (k\,|x\sin(kx)| + |\cos(kx)|)\frac{\epsilon_M}{2}$، واضح است که برای $|x|$ کوچک و $|kx| \simeq 1$، خطای e_1 مضر و در نتیجه الگوریتم ناپایدار است. △

مثال ۳۵.۲ برای $k \in \mathbb{N}$ و $x \in \mathbb{R}$ داده‌شده، مقدار $\cos(kx)$ و $\sin(kx)$ را از روابط بازگشتی زیر به دست می‌آوریم

$$\begin{cases} \cos(mx) = \cos x\cos(m-1)x - \sin x\sin(m-1)x, \\ \sin(mx) = \sin x\cos(m-1)x + \cos x\sin(m-1)x, \end{cases} \qquad m = 1,\dots,k. \tag{۵.۲}$$

اگر در محاسبه $\cos x$ و $\sin x$ خطایی داشته باشیم، چقدر در محاسبه $\cos(kx)$ و $\sin(kx)$ اثر دارد؟

یک ماتریس دوران به صورت زیر در نظر بگیرید

$$U = \begin{bmatrix} c & -s \\ s & c \end{bmatrix},$$

که در آن $c = c_1 = \cos x$ و $s = s_1 = \sin x$. U یک ماتریس متعامد[۴۲] است، یعنی $U^{-1} = U^T$ و به کمک آن می‌توان روابط بازگشتی

[۴۲] Orthogonal matrix

(۵.۲) را به صورت زیر خلاصه کرد

$$\begin{bmatrix} c_m \\ s_m \end{bmatrix} = U \begin{bmatrix} c_{m-1} \\ s_{m-1} \end{bmatrix}, \qquad m = 1, \dots, k,$$

که در آن $c_m = \cos(mx)$ و $s_m = \sin(mx)$. بنابراین

$$\begin{bmatrix} c_k \\ s_k \end{bmatrix} = U^k \begin{bmatrix} c_0 \\ s_0 \end{bmatrix} = U^k \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

از طرفی بنابر

$$\frac{\partial U}{\partial c} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I, \qquad \frac{\partial U}{\partial s} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} := A.$$

داریم

$$\frac{\partial U^k}{\partial s} = \frac{\partial}{\partial s}(UU^{k-1}) = \frac{\partial U}{\partial s}U^{k-1} + U\frac{\partial U^{k-1}}{\partial s} = AU^{k-1} + U\frac{\partial}{\partial s}(UU^{k-2}) = AU^{k-1} + U\left(\frac{\partial U}{\partial s}U^{k-2} + U\frac{\partial U^{k-2}}{\partial s}\right).$$

با تکرار این فرایند خواهیم داشت

$$\frac{\partial U^k}{\partial s} = AU^{k-1} + UAU^{k-2} + \cdots + U^{k-1}A = kAU^{k-1},$$

زیرا $AU = UA$ و چون $\frac{\partial U}{\partial c} = I$ داریم $\frac{\partial U^k}{\partial c} = kU^{k-1}$. بنابراین

$$\frac{\partial U^k}{\partial c} = k\begin{bmatrix} c_{k-1} & -s_{k-1} \\ s_{k-1} & c_{k-1} \end{bmatrix}, \qquad \frac{\partial U^k}{\partial s} = k\begin{bmatrix} -s_{k-1} & -c_{k-1} \\ c_{k-1} & -s_{k-1} \end{bmatrix}.$$

برای مسئله $y = \phi(x) = \begin{bmatrix} \cos(kx) \\ \sin(kx) \end{bmatrix}$، با تعریف $x^{(0)} = x$ و $\phi^{(0)}(x) = \begin{bmatrix} \cos x \\ \sin x \end{bmatrix}$ به وضوح داریم

$$\psi^{(1)}\begin{bmatrix} c \\ s \end{bmatrix} = U^k \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} c_k \\ s_k \end{bmatrix},$$

و از آنجا می‌توان نوشت

$$J\psi^{(1)}(x^{(1)})e_1 = \frac{\partial U^k}{\partial c}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\cos x\,\epsilon_1 + \frac{\partial U^k}{\partial s}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\sin x\,\epsilon_2 = \epsilon_1 k\cos x\begin{bmatrix} c_{k-1} \\ s_{k-1} \end{bmatrix} + \epsilon_2 k\sin x\begin{bmatrix} -s_{k-1} \\ c_{k-1} \end{bmatrix}.$$

با توجه به

$$e_a^{\circ}(y) = \begin{bmatrix} k\,|x\sin(kx)| + |\cos(kx)| \\ k\,|x\cos(kx)| + |\sin(kx)| \end{bmatrix} \frac{\epsilon_M}{۲},$$

واضح است که برای $|x|$ کوچک و $|kx| \simeq ۱$ داریم $|k\sin x| \simeq ۱$ و $|k\cos x| \simeq k$ و بنابراین خطای محاسبه $\sin x$ بی‌ضرر ولی خطای محاسبه $\cos x$ مضر و در نتیجه الگوریتم ناپایدار است. به عنوان تمرین بررسی کنید که محاسبه $\cos(kx)$ با این الگوریتم برای $|x|$ کوچک و $|kx| \simeq ۱$ نسبت به الگوریتم مثال قبل از نظر عددی قابل اعتمادتر است. △

مثال ۳۶.۲ برای $k \in \mathbb{N}$ و $x \in \mathbb{R}^+$ داده‌شده، به منظور محاسبه مقادیر $\cos(kx)$ و $\sin(kx)$ به کمک روابط بازگشتی

$$\begin{cases} \cos(m+۱)x = \cos x\cos(mx) - \sin x\sin(mx), \\ \sin(m+۱)x = \sin x\cos(mx) + \cos x\sin(mx), \end{cases} \qquad m = ۱, \ldots, k-۱,$$

و روابط

$$\begin{cases} dc_{m+۱} & := \cos(m+۱)x - \cos(mx), \\ & = ۲(\cos x - ۱)\cos(mx) - \sin x\sin(mx) - \cos x\cos(mx) + \cos(mx), \\ & = -۴\sin^۲(\frac{x}{۲})\cos(mx) + (\cos(mx) - \cos(m-۱)x), \end{cases}$$

و

$$\begin{cases} ds_{m+۱} & := \sin(m+۱)x - \sin(mx), \\ & = ۲(\cos x - ۱)\sin(mx) + \sin x\cos(mx) - \cos x\sin(mx) + \sin(mx), \\ & = -۴\sin^۲(\frac{x}{۲})\sin(mx) + (\sin(mx) - \sin(m-۱)x), \end{cases}$$

روابط بازگشتی زیر معرفی می‌شوند

$$\begin{cases} c_m := c_{m-۱} + dc_m, & dc_{m+۱} := t\,c_m + dc_m, \\ s_m := s_{m-۱} + ds_m, & ds_{m+۱} := t\,s_m + ds_m, \end{cases} \qquad m = ۱, \ldots, k,$$

که در آن

$$c_۰ = ۱,\ s_۰ = ۰,\ t = ۲dc_۱,\ dc_۱ = -۲\sin^۲(\frac{x}{۲}),\ ds_۱ = \sin x = ۲\sin(\frac{x}{۲})\cos(\frac{x}{۲}) = \sqrt{-dc_۱(۲+dc_۱)}.$$

برای مسئله $y = \phi(x) = \begin{bmatrix} \cos(kx) \\ \sin(kx) \end{bmatrix}$، با تعریف $x^{(۰)} = x$ و $s := \sin(\frac{x}{۲}) = \phi^{(۰)}(x)$، اگر در محاسبه $\sin(\frac{x}{۲})$ خطایی داشته باشیم، چقدر در محاسبه $\cos(kx)$ و $\sin(kx)$ اثر دارد؟

به وضوح داریم

$$\psi^{(۱)}(s) = \begin{bmatrix} c_k \\ s_k \end{bmatrix},$$

و از آنجا می‌توان نوشت

$$J\psi^{(1)}(s)e_1 = \begin{bmatrix} \frac{\partial c_k}{\partial s} \\ \frac{\partial s_k}{\partial s} \end{bmatrix} s\,\epsilon = \begin{bmatrix} \frac{\partial c_k}{\partial x}\frac{\partial x}{\partial s} \\ \frac{\partial s_k}{\partial x}\frac{\partial x}{\partial s} \end{bmatrix} s\,\epsilon = 2k\tan(\frac{x}{2}) \begin{bmatrix} -s_k \\ c_k \end{bmatrix} \epsilon.$$

واضح است که برای x کوچک و $kx \simeq 1$، خطای محاسبه s بی‌ضرر بوده و در نتیجه الگوریتم پایدار است. به عنوان تمرین بررسی کنید که محاسبه $\cos(kx)$ با این الگوریتم برای x کوچک و $kx \simeq 1$ نسبت به الگوریتم مثال قبل از نظر عددی قابل اعتمادتر است. △

برنامه cskx.nb را بررسی کنید. مثال 5 صفحه 25 مرجع [۲۲] را ببینید.

۵.۲ حساب بازه‌ای

به کمک آنالیز خطای پیشرو به سادگی می‌توان اثر خطاهای گرد کردن میانی یک الگوریتم را در حل یک مسئله نه چندان پیچیده مشخص کرد. تعیین اثر کلی خطای گرد کردن میانی الگوریتم‌هایی که از پیچیدگی بالایی برخوردار باشند به عبارتی الگوریتم‌هایی که شامل تعداد زیادی خطاهای گرد کردن میانی هستند به سادگی امکان‌پذیر نیست. به کمک حساب بازه‌ای[۴۳] می‌توان با در نظر گرفتن خطای داده‌های ورودی و خطاهای گرد کردن میانی، برای خطای مطلق یک الگوریتم یک کران بالا به دست آورد. حساب بازه‌ای بر اساس این واقعیت است که هر عدد حقیقی $a \in \mathbb{R}$ که یا یک داده ورودی است یا یک نتیجه میانی یا نهایی، مقداری نامعلوم است و تنها اطلاعی که از آن داریم یک بازه کوچک شامل آن عدد است. بنابراین در حساب بازه‌ای متناظر با عدد a بازه $\tilde{a} = [a', a'']$ در نظر گرفته می‌شود که در آن a' و a'' اعداد ماشینی هستند و $a \in \tilde{a}$. متناظر با هر عملی روی اعداد باید عملی روی بازه‌ها تعریف کرد. به منظور تعریف عمل دوتایی $\circ \in \{+, -, \times, /\}$ روی بازه‌های $\tilde{a}$ و $\tilde{b}$ قرار می‌دهیم $\tilde{c} := \tilde{a} \circ \tilde{b}$ به طوری که $\{a \circ b | a \in \tilde{a}\ \&\ b \in \tilde{b}\} \subset \tilde{c}$.

مثال ۳۷.۲ *برای عمل جمع داریم $[a', a''] + [b', b''] := [c', c'']$ که در آن*

$$c' = \max\{\gamma' \in \mathbb{F} | \gamma' \le a' + b'\}, \qquad c'' = \min\{\gamma'' \in \mathbb{F} | \gamma'' \ge a'' + b''\},$$

و برای عمل ضرب داریم $[a', a''] \times [b', b''] := [c', c'']$ که در آن برای $a', b' > 0$

$$c' = \max\{\gamma' \in \mathbb{F} | \gamma' \le a' \times b'\}, \qquad c'' = \min\{\gamma'' \in \mathbb{F} | \gamma'' \ge a'' \times b''\}.$$

سایر اعمال به طریق مشابه تعریف می‌شوند. △

تذکر ۱۵.۲ *بیشتر مواقع می‌توان از تعاریف زیر استفاده کرد*

$$\begin{cases} [a', a''] + [b', b''] = [a' + b', a'' + b''], \\ [a', a''] - [b', b''] = [a' - b'', a'' - b'], \\ [a', a''] \times [b', b''] = [\min\{a' \times b', a' \times b'', a'' \times b', a'' \times b''\}, \max\{a' \times b', a' \times b'', a'' \times b', a'' \times b''\}], \\ [a', a'']/[b', b''] = [a', a''] \times [\frac{1}{b''}, \frac{1}{b'}], \qquad \text{if} \qquad 0 \notin [b', b'']. \end{cases}$$

[۴۳] Interval arithmetic

زمانی که یک الگوریتم بر اساس حساب بازه‌ای دنبال شود در پایان یک بازه به دست می‌آید که نه تنها جواب مسئله به آن بازه تعلق دارد بلکه می‌توان به کمک آن بازه کرانی برای خطای مطلق جواب نیز تعیین کرد. اگرچه این کران خطا قابل اعتماد است ولی در بیشتر حالات ممکن است بدبینانه[۴۴] باشد. بنابراین کافی نیست که فقط اعمال حسابی با اعمال بازه‌ای نظیر جایگزین شوند بلکه باید ملاحظاتی نیز در نظر گرفت و راه‌کارهایی به کار گرفت.

مثال ۳۸.۲ *برای ارزیابی تابع* $y = \phi(x) = x^3 - 3x^2 + 3x$ *ابتدا بهتر است به کمک قاعده هورنر آن را به صورت* $y = \phi(x) = ((x-3)x+3)x$ *یا*

$$u = x - 3, \qquad v = u \times x, \qquad w = v + 3, \qquad y = w \times x,$$

بازنویسی کرد. بازه نظیر اعداد دقیق مانند ۳ یا به صورت $[3,3]$ *یا با انتخاب مناسبی از* $\epsilon > 0$ *به صورت* $[3-\epsilon, 3+\epsilon]$ *در نظر گرفته می‌شود. حال برای ارزیابی* ϕ *در نقطه* $x \in \tilde{x} = [0.9, 1.1]$ *داریم*

$$\begin{cases} \tilde{u} = \tilde{x} - [3,3] = [-2.1, -1.9], & \tilde{v} = \tilde{u} \times \tilde{x} = [-2.31, -1.71], \\ \tilde{w} = \tilde{v} + [3,3] = [0.69, 1.29], & \tilde{y} = \tilde{w} \times \tilde{x} = [0.6219, 1.4191]. \end{cases}$$

اگرچه بازه به دست آمده شامل جواب مسئله است ولی بزرگ بودن این بازه دلالت بر خطای زیادی دارد. از طرفی چون $\phi'(x) = 3(x-1)^2 \geq 0$ *پس* ϕ *یک تابع صعودی است و باید داشته باشیم*

$$\{\phi(x) | x \in \tilde{x}\} = [\phi(0.9), \phi(1.1)] = [0.999, 1.001],$$

یعنی $\phi : [0.9, 1.1] \to [0.999, 1.001]$*. اگر با اعداد ماشینی دو رقمی کار کنیم، به ازای* $x = 0.9$ *داریم*

$$u = -2.1, \qquad v = -1.9, \qquad w = 1.1, \qquad y = 0.99,$$

و به ازای $x = 1.1$ *خواهیم داشت*

$$u = -1.9, \qquad v = -2.1, \qquad w = 0.9, \qquad y = 0.99.$$

حال با بازنویسی تابع ϕ *به صورت* $y = \phi(x) = 1 + (x-1)^3$*، به ازای* $\tilde{x} = [0.9, 1.1]$ *داریم*

$$\begin{cases} \tilde{u} = \tilde{x} - [1,1] = [-0.1, 0.1], & \tilde{v} = \tilde{u} \times \tilde{u} = [-0.01, 0.01], \\ \tilde{w} = \tilde{v} \times \tilde{u} = [-0.001, 0.001], & \tilde{y} = \tilde{w} + [1,1] = [0.999, 1.001]. \end{cases}$$

اگر این بار نیز با اعداد ماشینی دو رقمی کار کنیم، به ازای $x = 0.9$ *داریم*

$$u = -0.1, \qquad v = 0.01, \qquad w = -0.001, \qquad y = 1.0,$$

[۴۴] Pessimistic

و به ازای $x = ۱/۱$ خواهیم داشت

$$u = ۰/۱, \qquad v = ۰/۰۱, \qquad w = ۰/۰۰۱, \qquad y = ۱/۰.$$

△

تذکر ۱۶.۲ *همان طور که در مثال قبل دیده شد، برای دنبال کردن یک الگوریتم در حساب بازه‌ای نه تنها باید اعمال حسابی را با اعمال بازه‌ای نظیر جایگزین کرد بلکه باید راه‌کارهای مناسبی نیز به کار گرفت. برای اطلاعات بیشتر در مورد حساب بازه‌ای، به مرجع [۱۸] مراجعه کنید.*

پروژه ۱.۲ *بازنویسی الگوریتم‌های عددی کارشناسی با حساب بازه‌ای.*

سمینار ۱.۲ *حساب بازه‌ای گذشته و حال با تکیه بر کاربردهای جدید.*

سمینار ۲.۲ *تخمین خطای گرد کردن به صورت آماری.*

۶.۲ تقسیم‌بندی مسایل در ریاضیات محاسباتی

بسیاری از مسایل مطرح‌شده در ریاضیات کاربردی و به ویژه آنالیز عددی به صورت زیر مدل می‌شوند

$$T : X \to Y, \qquad Tx = y, \tag{۶.۲}$$

که در آن X و Y دو فضای برداری نرمدار مناسب هستند که بر اساس ورودی و خروجی مسئله تعریف می‌شوند و T عملگری است که متناظر با مسئله تعریف می‌شود. مسایلی که در عمل با آنها برخورد می‌کنیم به سه دسته اساسی زیر تقسیم می‌شوند

۱. مسایل مستقیم[۴۵]: در این مسایل T و x معلوم هستند و y مجهول است. ارزیابی توابع یک یا چندمتغیره عددی یا برداری یا ماتریسی، محاسبه مشتق یا انتگرال یک تابع یک یا چندمتغیره از این نوع به شمار می‌روند،

۲. مسایل وارون[۴۶]: در این مسایل T و y معلوم بوده و x مجهول است. انواع معادلات، دستگاه معادلات خطی/غیرخطی، معادلات دیفرانسیل عادی یا جزئی، معادلات انتگرال، مسایل ویژه‌مقدار و مقدار تکین جزء مسایل وارون هستند،

۳. مسایل شناسایی[۴۷]: در این مسایل x و y معلوم و T مجهول است. مسئله درونیابی معروفترین مسئله از نوع شناسایی است. به بیان مهندسی سیستم، چون ورودی و خروجی معلوم هستند، این مسئله به مهندسی وارون[۴۸] یا جعبه سیاه[۴۹] نیز معروف است. برخی از مسایل کمینه‌سازی تابعی[۵۰]، برنامه‌ریزی ریاضی و آمار به طور غیرمستقیم در زمره مسایل شناسایی به حساب می‌آیند.

---

[۴۵] Direct problems
[۴۶] Inverse problems
[۴۷] Identification problems
[۴۸] Inverse engineering
[۴۹] Black box
[۵۰] Functional minimization

مسایل شناسایی از پیچیدگی بالایی برخوردارند در حالی که مسایل مستقیم ساده بوده و بسیاری از ابهامات و پرسش‌های مربوط به آنها تاکنون بررسی شده‌اند ولی با این حال سرعت پیشرفت پژوهش در این راستا کند است. بسیاری از مسایلی که امروزه با آنها مواجه می‌شویم از نوع مسایل وارون هستند زیرا این مسایل از اهمیت بالایی برخوردارند و یک جایگاه مرکزی در آنالیز عددی برای خود باز کرده‌اند. نوع خطی این مسایل هم از دیدگاه نظری هم از دیدگاه عددی به خوبی مطالعه و بررسی شده است، ولی هنوز مسایلی در این راستا جهت مطالعه و تحقیق وجود دارد. امروزه مسایل وارون غیرخطی راستای بیشتر پژوهش‌ها را به خود منعطف کرده است زیرا این مسایل پیچیده از جذابیت خاصی برخوردارند. باید توجه داشت که گاهی مواقع ممکن است حل یک مسئله از یک نوع به سادگی امکان‌پذیر نباشد و در این صورت سعی می‌کنیم آن مسئله را به یک مسئله از همان نوع یا حتی نوع دیگر تبدیل کرده و به حل آن مبادرت ورزیم.

مثال ۳۹.۲ *مسئله مقدار اولیه*

$$\begin{cases} \dfrac{dx(t)}{dt} = f(t)x(t), & t > ۰, \\ x(۰) = ۱, \end{cases}$$

که از نوع وارون است به کمک روش جداسازی متغیرها به مسئله مستقیم زیر تبدیل می‌شود

$$x(t) = e^{\int_۰^t f(s)ds}.$$

این در حالی است که مسئله جدید با مشتق‌گیری به همان مسئله مقدار اولیه تبدیل می‌شود. △

مثال ۴۰.۲ *مسئله مقدار اولیه*

$$\begin{cases} \dfrac{dx(t)}{dt} = f(t, x(t)), & t > ۰, \\ x(۰) = ۱, \end{cases}$$

که از نوع وارون است به کمک انتگرال‌گیری به مسئله وارون زیر تبدیل می‌شود که یک معادله انتگرال است.

$$x(t) = ۱ + \int_۰^t f(s, x(s))ds$$

مسئله جدید با مشتق‌گیری به همان مسئله مقدار اولیه تبدیل می‌شود. △

تعریف ۱۸.۲ *اگر X و Y دو فضای برداری باشند، فضای حاصل‌ضرب $X \times Y$ به صورت زیر تعریف می‌شود*

$$X \times Y := \{(x, y) | x \in X,\ y \in Y\}.$$

فضای $X \times Y$ همراه با جمع برداری

$$(x_۱, y_۱) + (x_۲, y_۲) = (x_۱ + x_۲, y_۱ + y_۲), \qquad \forall x_۱, x_۲ \in X, \quad \forall y_۱, y_۲ \in Y,$$

و ضرب اسکالر

$$\alpha(x, y) = (\alpha x, \alpha y), \qquad \forall \alpha \in \mathbb{R}, \qquad \forall x \in X, \quad \forall y \in Y,$$

یک فضای برداری است.

مثال ۴۱.۲ *مسئله مقدار اولیه*

$$\begin{cases} \dfrac{dx(t)}{dt} = f(t, x(t)), & ۰ < t < ۱, \\ x(۰) = \alpha, \end{cases}$$

را به صورت عملگری بنویسید.

با معرفی $X = C^۱[۰, ۱]$، $Y = C[۰, ۱] \times \mathbb{R}$ *و* $Tx(t) = \left(\dfrac{dx(t)}{dt} - f(t, x(t)), x(۰)\right)$، *به وضوح* $T : X \to Y$ *و شکل عملگری مسئله به صورت* $Tx(t) = (۰, \alpha)$ *خلاصه می‌شود.* △

تمرین ۷.۲ *شکل عملگری مسایل مقدار مرزی یا مقدار اولیه-مرزی*

$$\begin{cases} -\Delta u(x, y) = f(x, y), & (x, y) \in \Omega, \\ u(x, y) = g(x, y), & (x, y) \in \partial\Omega, \end{cases} \qquad \begin{cases} u_t(x, y, t) = \Delta u(x, y, t) + f(x, y, t), & (x, y, t) \in \Omega \times \mathbb{R}^+, \\ u(x, y, ۰) = u_۰(x, y), & (x, y) \in \Omega, \\ u(x, y, t) = g(x, y, t), & (x, y, t) \in \partial\Omega, \end{cases}$$

را بنویسید که در آن $\Omega \in \mathbb{R}^۲$ *یک مجموعه باز و همبند و* $\partial\Omega$ *مرز* Ω *است و* $\Delta = \frac{\partial^۲}{\partial x^۲} + \frac{\partial^۲}{\partial y^۲}$. *این مسایل را برای حالتی که* $\Omega \in \mathbb{R}^۱$ *یا* $\Omega \in \mathbb{R}^۳$ *نیز بازنویسی کرده و شکل عملگری آنها را بنویسید.*

تمرین ۸.۲ *مسئله ویژه‌مقدار را به صورت عملگری بنویسید.*

بیشتر مواقع فضاهای X و Y نامتناهی‌البعد بوده و به جای آنها از فضاها (زیرفضاهای) متناهی‌البعد X_n و Y_n استفاده می‌کنیم و در واقع مسئله عملگری (۶.۲) را به مسئله ساده‌تر

$$T_n : X_n \to Y_n, \qquad T_n x_n = y_n,$$

گسسته‌سازی می‌کنیم.

مثال ۴۲.۲ *مسئله انتگرال‌گیری معین* $y = Tx$ *که در آن* T *عملگری از* $X = C[a, b]$ *به توی* $Y = \mathbb{R}$ *است و* $Tx = \int_a^b x(t)dt$، *را در نظر بگیرید. این مسئله به کمک یک قاعده انتگرال‌گیری به صورت زیر گسسته‌سازی می‌شود*

$$X_n = \mathbb{R}^n, \qquad Y_n = \mathbb{R}, \qquad x_n = [x(t_۱) \ \cdots \ x(t_n)]^T, \qquad y_n = T_n x_n = \sum_{i=۱}^{n} a_i x(t_i).$$

△

تعریف ۱۹.۲ *اگر در حل یک مسئله وارون داشته باشیم* $\|x - x_n\| \le B(n)$، *تابع* $B(n)$ *به کران بالای خطا*[۵۱] *یا تخمین اولیه (پیشین)*[۵۲] *معروف است. بیشتر مواقع داریم* $B(n) = cn^{-p}$ *و در این حالت می‌نویسیم* $x = x_n + O(n^{-p})$ *و می‌گوییم روش خطایی از مرتبه* p *دارد.*

مثال ۴۳.۲ *در روش دوبخشی در یافتن ریشه یکتای معادله* $x(t) = ۰$ *متعلق به بازه* $[a, b]$ *داریم* $|t - t_n| \le \frac{b-a}{۲^n}$. *همچنین در روش انتگرال‌گیری ذوزنقه یعنی*

$$T(h) = \frac{h}{۲}\left(x(a) + ۲\sum_{i=۱}^{n-۱} x(t_i) + x(b)\right),$$

[۵۱] Upper error bound
[۵۲] A periori estimate

که در آن

$$h = (b-a)/n, \qquad t_i = a + ih, \quad i = ۰, \dots, n,$$

داریم

$$\left| \int_a^b x(t)dt - T(h) \right| \leq \frac{(b-a)^۳}{۱۲n^۲} \max_{t\in[a,b]} |x''(t)| .$$

بنابراین اگر $x \in C^۲[a,b]$ روش ذوزنقه خطایی از مرتبه دو دارد. △

تذکر ۱۷.۲ بعضی مواقع داریم $B(n) = B(n,x)$. این کران جنبه نظری داشته و در عمل کاربرد ندارد ولی اگر $B(n) = B(n, x_n)$ که در این حالت به کران (تخمین) ثانویه (پسین)[۵۳] معروف است و به کمک اطلاعاتی از مسئله اصلی یا اعمال فرضیاتی، به یک کران قابل محاسبه تبدیل می‌شود.

مثال ۴۴.۲ در روش نقطه ثابت برای یافتن تنها ریشه $x(t) = ۰$ در بازه $[a,b]$، ابتدا معادله را به صورت $t = z(t)$ بازنویسی کرده و تلاش می‌کنیم نقطه ثابت z یعنی همان ریشه x را بیابیم. قضیه نقطه ثابت را جهت یادآوری بیان می‌کنیم.

قضیه ۷.۲ آ) اگر $z : [a,b] \longrightarrow [a,b]$ و $z \in C[a,b]$ آن‌گاه z در $[a,b]$ دست کم یک نقطه ثابت دارد؛

ب) همچنین اگر $z'(x)$ بر (a,b) موجود باشد و عدد ثابت و مثبت L چنان وجود داشته باشد که

$$\forall\, t \in (a,b), \qquad |z'(t)| \leq L < ۱,$$

آن‌گاه نقطه ثابت z در $[a,b]$ یکتا است؛

پ) به علاوه دنباله $\{t_n\}_{n\geq ۱}$ تولید شده توسط فرایند $t_n = z(t_{n-۱})$ به ازای هر $t_۰$ در $[a,b]$ به نقطه ثابت z یعنی t همگرا است و

$$\begin{cases} |t - t_n| \leq L^n \, |t - t_۰| , \\ |t - t_n| \leq \dfrac{L^n}{۱-L} \, |t_۱ - t_۰| , \\ |t - t_n| \leq \dfrac{L}{۱-L} \, |t_n - t_{n-۱}| . \end{cases}$$

اولین نابرابری دلالت بر کران بالای خطا دارد و به کمک آن همگرایی ثابت می‌شود، نابرابری دوم یک تخمین اولیه و نابرابری سوم یک تخمین ثانویه است. △

تذکر ۱۸.۲ اگر کران بالای خطا با خطای واقعی فاصله زیادی داشته باشد، بدبینانه[۵۴] و در غیر این صورت واقع‌بینانه[۵۵] است.

تعریف ۲۰.۲ اگر داشته باشیم $x - x_n = e_n + g_n$ که در آن $\lim_{n\to\infty} \|g_n\| / \|e_n\| = ۰$، آنگاه e_n به تخمین خطای مجانبی[۵۶] معروف است.

---

[۵۳] A posteriori estimate
[۵۴] Pessimestic
[۵۵] Realistic
[۵۶] Asymptotic error estimate

مثال ۴۵.۲ *به کمک بسط تیلور داریم*

$$x''(t_\circ) = \frac{x(t_\circ - h) - ۲x(t_\circ) + x(t_\circ + h)}{h^۲} + C_۲h^۲x^{(۴)}(t_\circ) + C_۴h^۴x^{(۶)}(t_\circ) + \cdots.$$

با معرفی $e_h = C_۲h^۲x^{(۴)}(t_\circ)$ *و* $g_h = C_۴h^۴x^{(۶)}(t_\circ) + \cdots$ *واضح است که* $|g_h/e_h| \to ۰$ *هرگاه* $h \to ۰$ *و* x *به اندازه کافی هموار باشد. بنابراین* e_h *تخمین خطای مجانبی و از مرتبه دو است.* △