یادداشت های درس آنالیز عددی پیشرفته ©

جواد فرضی [1]
دانشگاه صنعتی سهند، بخش ریاضی

بخش اول : خطا و انتشار آن

فهرست مندرجات

[1] $farzi@sut.ac.ir$

۱ خطا

تعیین دقت نتایج عددی یکی از اهداف بسیار مهم در آنالیز عددی است. با مطالعه در این زمینه می‌توان چند نوع خطا که در محدود کردن دقت نتایج موثرند را بر شمرد:

۱) خطاهای داده های ورودی (errors in the input data)

۲) خطاهای گرد کردن (roundoff errors)

۳) خطاهای تقریب (approximation errors)

خطاهای ورودی از کنترل محاسبات خارج است. خطاهای گرد کردن مربوط به محاسبه با اعدادی است که نمایش آنها در ماشین به تعداد ارقام متناهی محدود شده است که گریزی از این واقعیت نیست. برای نوع سوم مساله P را در نظر می‌گیریم. بسیاری از روشها حتی با این فرض که محاسبات بدون خطای گرد کردن باشد به جواب دقیق این مساله منجر نمی‌شوند و به جای مساله P مساله ساده‌تر $\tilde{P}$ را حل می کنند. برای مثال مساله P، محاسبه سری نامتناهی

$$e = ۱ + \frac{۱}{۱!} + \frac{۱}{۲!} + \frac{۱}{۳!} + \dots \qquad (۱)$$

را می‌توان با مساله $\tilde{P}$، محاسبه سری تا یک تعداد متناهی از جملات سری جایگزین کرد. چنین خطای تقریبی خطای قطع (truncation error) نام دارد. البته خطای قطع در زمینه های دیگری مانند نمایش متناهی عدد نیز مطرح می‌شود. بسیاری از مسائل تقریب P با گسسته سازی (discretizing) مساله اصلی P بدست می آیند. انتگرالها با مجموع های متناهی و معادلات دیفرانسیل بامعادلات تفاضلی تقریب می‌شوند. در چنین حالتهایی خطای تقریب با عنوان خطای گسسته سازی مطرح می‌شود. در اینجا تاثیر خطاهای ورودی و خطای گرد کردن بر روی نتایج عددی را بررسی می‌کنیم. به طور طبیعی خطای تقریب را در کنار روش تقریب ارائه شده در فصول بعدی بررسی خواهیم کرد.

۱.۱ نمایش اعداد

نمایش عدد x در مبنای ۲

$$x = \pm(\alpha_n ۲^n + \alpha_{n-۱} ۲^{n-۱} + \dots + \alpha_۰ ۲^۰ + \alpha_{-۱} ۲^{-۱} + \alpha_{-۲} ۲^{-۲} + \dots)$$
$$\alpha_i = ۰ \quad or \quad ۱.$$

برای مثال نمایش ۱۸.۵ در مبنای دو به صورت زیر است:

$$۱۸.۵ = ۱ \times ۲^۴ + ۰ \times ۲^۳ + ۰ \times ۲^۲ + ۱ \times ۲^۱ + ۰ \times ۲^۰ + ۱ \times ۲^{-۱}$$

مثال ۴ $= ۳.۹۹۹\dots$ نشان می‌دهد که نمایش ده‌دهی یک عدد ممکن است منحصر به فرد نباشد. همین در مورد نمایش دودویی هم درست است. برای دوری از چنین ابهامهایی نمایش متناهی را در نظر خواهیم گرفت. همچنین مثالهایی وجود دارد که نمایش یک عدد در یک دستگاه متناهی است ولی نمایش همان عدد در دستگاه دیگر متناهی نیست.

اعداد ماشینی ماشین‌های دیجیتالی با توجه به محدودیت حافظه در ذخیره کردن و محاسبات داخلی خود مکانهای ثابت و متناهی را برای اعداد اختصاص می‌دهند که به آن کلمه گفته می‌شود. این عدد n در برخی ماشین‌ها دو یا سه برابر قرار داده می‌شود. از این طول n به شیوه‌های مختلف می‌توان برای نمایش اعداد استفاده کرد.

- نمایش ممیز ثابت

در این نمایش تعداد n_1 رقم پیش و تعداد n_2 رقم پس از ممیز اعشار(دودویی) وجود دارد که $n = n_1 + n_2$.
برای مثال با $n = 10$، $n_1 = 4$ و $n_2 = 6$ داریم

$$\begin{array}{rcc} & \overbrace{}^{n_1} & \overbrace{}^{n_2} \\ ۳۰.۴۲۱ \rightarrow & ۰۰۳۰ & ۴۲۱۰۰۰ \\ ۰.۰۴۳۷ \rightarrow & ۰۰۰۰ & ۰۴۳۷۰۰ \end{array}$$

- نمایش ممیز سیار

$$x = a \times ۱۰^b (x = a \times ۲^b), \qquad |a| < ۱ \tag{۲}$$

در این نمایش عدد صحیح b جایگاه ممیز را نسبت به مانتیس a نشان می‌دهد. برای مثال نمایش عدد ۳۰.۴۲۱ به صورت ممیز سیار عبارتست از ۰.۳۰۴۲۱×۱۰^۲ و عدد ۱۸.۵ را در مبنای دو با ممیز سیار به صورت ۰.۱۰۰۱۰۱×۲^{۱۰۱} می‌توان نمایش داد.
در هر ماشینی تعداد مکانهای ثابت و متناهی t و e، $n = t + e$، در حافظه به ترتیب برای نمایش مانتیس و توان وجود دارد. برای مثال با $t = ۴$ و $e = ۲$ نمایش ممیز سیار ۵۴۲۰ به صورت ۰.۵۴۲۰×۱۰^{۰۴} است.
در صورتی که رقم اول مانتیس ناصفر باشد نمایش ممیز سیار به صورت نرمال است. در این صورت در رابطه (۲) داریم $|a| \geq ۱۰^{-۱}$ $(|a| \geq ۲^{-۱})$. در این صورت ارقام با معنی یک عدد عبارتست از ارقام مانتیس بدون شمارش صفرهای پیشرو.
زیر مجموعه $A \subseteq R$ از اعداد حقیقی که قابل نمایش در یک ماشین مفروض باشند اعداد ماشینی نامیده می‌شوند. در یک ماشین ۳۲ و ۶۴ بیتی مکانهای حافظه برای ذخیره کردن یک عدد به صورت ممیز سیار(۲)، به ترتیب، به این صورت اختصاص می‌یابد:

	۳۲ بیتی	۶۴ بیتی
علامت عدد حقیقی x	۱	۱
علامت نما b	۱	۱
نما $\|b\|$	۷	۱۰
مانتیس $\|a\|$	۲۳	۵۲

۲.۱ حساب ممیز سیار و خطای گرد کردن

مجموعه اعداد A که قابل نمایش در ماشین باشند متناهی‌اند. این سئوال پیش می‌آید که چگونه می‌توان یک عدد $x \notin A$ که عدد ماشینی نیست را با یک عدد ماشینی مانند $g \in A$ نمایش داد. این سئوال نه تنها در خواندن اعداد از سوی رایانه، بلکه به هنگام محاسبات و نمایش نتایج میانی در رایانه نیز پیش می‌آید. به عبارتی با مثالهایی می‌توان نشان داد که مجموعه اعداد ماشینی نسبت به اعمال چهارگانه ریاضی بسته نیست. مثال زیر وقوع خطا به هنگام خواندن را نشان می‌دهد:

$$\frac{۱}{۱۰} = (۰.۰۰۰۱۱۰۰۱۱۰۰۱۱۰۰۱...)_۲ \qquad (۳)$$

در انتهای این بخش مثال‌هایی از حالت‌های دیگر ارائه می‌شود.

طبیعی است بخواهیم که تقریب عدد $x \notin A$ با یک عدد ماشینی $rd(x) \in A$ در رابطه زیر صدق کند

$$|x - rd(x)| \leq |x - g| \quad \forall g \in A \qquad (۴)$$

در بیشتر مواقع چنین تقریبی را می‌توان با گرد کردن بدست آورد.

مثال $(t = ۴)$

$$rd(۰.۱۴۲۸۵ \times ۱۰^{۰}) = ۰.۱۴۲۹ \times ۱۰^{۰},$$
$$rd(۳.۱۴۱۵۹ \times ۱۰^{۰}) = ۰.۳۱۴۲ \times ۱۰^{۱},$$
$$rd(۰.۱۴۲۸۴۲ \times ۱۰^{۲}) = ۰.۱۴۲۸ \times ۱۰^{۲}.$$

به طور کلی، برای بدست آوردن $rd(x)$ در یک رایانه t—رقمی به صورت زیر عمل می‌کنیم: در آغاز $x \notin A$ را به صورت نرمال شده $x = a \times ۱۰^{b}$ که $|a| \geq ۱۰^{-۱}$ نمایش می‌دهیم. فرض کنید نمایش اعشاری $|a|$ به صورت زیر باشد

$$|a| = ۰.\alpha_۱\alpha_۲ \ldots \alpha_i\alpha_{i+۱} \ldots, \quad ۰ \leq \alpha_i \leq ۹, \quad \alpha_۱ \neq ۰.$$

در این صورت با معرفی

$$a' := \begin{cases} ۰.\alpha_۱\alpha_۲ \ldots \alpha_t, & ۰ \leq \alpha_{t+۱} \leq ۴, \\ ۰.\alpha_۱\alpha_۲ \ldots \alpha_t + ۱۰^{-t}, & \alpha_{t+۱} \geq ۵. \end{cases}$$

تقریب مورد نظر به صورت زیر ارائه می‌شود

$$\tilde{rd}(x) := sign(x).a' \times ۱۰^{b}.$$

با توجه به $|a| \geq ۱۰^{-۱}$ خطای نسبی $\tilde{rd}(x)$ در نامساوی زیر صدق می‌کند:

$$\left|\frac{\tilde{rd}(x) - x}{x}\right| \leq \frac{۵ \times ۱۰^{-(t+۱)}}{|a|} \leq ۵ \times ۱۰^{-t}.$$

با فرض $eps := ۵ \times ۱۰^{-t}$ این عبارت را می‌توان به صورت زیر نوشت

$$\tilde{rd}(x) = x(۱ + \varepsilon), \quad |\varepsilon| \leq eps. \qquad (۵)$$

کمیت $eps := ۵ \times ۱۰^{-t}$ دقت ماشین نامیده می‌شود. در دستگاه دودویی به طور مشابه می‌توان $\tilde{rd}(x)$ را تعریف کرد: با شروع از $x = a \times ۲^b$ که $۲^{-۱} \leq |a| < ۱$ و نمایش دودویی $|a|$ داریم

$$|a| = ۰.\alpha_۱\alpha_۲ \ldots \alpha_t\alpha_{t+۱} \ldots, \quad \alpha_i = ۰ \; or \; ۱, \quad \alpha_۱ = ۱.$$

با تشکیل a'

$$a' := \begin{cases} ۰.\alpha_۱\alpha_۲ \ldots \alpha_t, & \alpha_{t+۱} = ۰, \\ ۰.\alpha_۱\alpha_۲ \ldots \alpha_t + ۲^{-t}, & \alpha_{t+۱} = ۱. \end{cases}$$

$$\tilde{rd}(x) := sign(x).a' \times ۲^b.$$

دوباره رابطه (۵) با $eps := ۲^{-t}$ برقرار است.
اگر $\tilde{rd}(x) \in A$ یک عدد ماشینی باشد، آنگاه $\tilde{rd}$ دارای خاصیت (۴) است که روند گرد کردن درست را نشان می‌دهد، در این حالت داریم

$$\tilde{rd}(x) := rd(x) \quad \forall x \in A.$$

با توجه به اینکه مکانهای موجود برای نمایش توان تعداد محدود e است، همواره اعدادی وجود دارند که $x \notin A$ و $\tilde{rd}(x) \notin A$.
مثال $(e = ۲، t = ۴)$

$$\begin{aligned} (a) \quad \tilde{rd}(۰.۳۱۷۹۴ \times ۱۰^{۱۱۰}) &= ۰.۳۱۷۹ \times ۱۰^{۱۱۰} \notin A. \\ (b) \quad \tilde{rd}(۰.۹۹۹۹۷ \times ۱۰^{۹۹}) &= ۰.۱۰۰۰ \times ۱۰^{۱۰۰} \notin A. \\ (c) \quad \tilde{rd}(۰.۰۱۲۳۴۵ \times ۱۰^{-۹۹}) &= ۰.۱۲۳۵ \times ۱۰^{-۱۰۰} \notin A. \\ (d) \quad \tilde{rd}(۰.۵۴۳۲۱ \times ۱۰^{-۱۱۰}) &= ۰.۵۴۳۲ \times ۱۰^{-۱۱۰} \notin A. \end{aligned}$$

در حالت (a) و (b) نما یک مقدار بسیار بزرگ است که در فضای اختصاص داده شده نمی گنجد: این دو مورد مثالهایی از سر ریز نما (exponent overflow) هستند. در حالت (b) سرریز نما بعد از گرد کردن اتفاق می افتد. حالتهایی (c) و (d) مثالهای از زیرریز نما(exponent overflow) هستند. در حالتهای (c) و (d) با تعریف زیر می‌توان از زیرریز نما جلوگیری کرد

$$\begin{aligned} (c) \quad rd(۰.۰۱۲۳۴۵ \times ۱۰^{-۹۹}) &= ۰.۰۱۲۳ \times ۱۰^{-۹۹} \in A, \\ (d) \quad rd(۰.۵۴۳۲۱ \times ۱۰^{-۱۱۰}) &= ۰ \in A. \end{aligned}$$

در این صورت rd در (۵) صدق نخواهد کرد. چون خطای نسبی $rd(x)$ ممکن است از eps تجاوز کند. باصرف نظر از سرریز و زیرریز، گرد کردن را به صورت زیر تعریف می‌کنیم

$$rd(x) := \tilde{rd}(x).$$

با در نظر گرفتن مقیاس های مناسب و کنترل داده‌های ورودی و بررسی‌های مناسب به هنگام محاسبات با این مشکلات مواجه نمی‌شویم. بنابراین با توجه به اینکه این موارد در عمل چندان اتفاق نمی افتد در ادامه بحث فرض می‌کنیم $e = \infty$. با چنین شرایطی، $rd := \tilde{rd}$ دستوری برای گرد کردن فراهم می‌کند که در شرایط زیر صدق می‌کند

$$rd : R \rightarrow A, \tag{۶}$$

$$\tilde{rd}(x) = x(۱ + \varepsilon), \quad |\varepsilon| \leq eps \quad \forall x \in R. \tag{۷}$$

مشاهده کردیم که در صورتی که نتیجه اعمال حسابی $x \pm y$، $x \times y$ و x/y لزوما اعداد ماشینی نیستند، حتی اگر اعداد x و y اعداد ماشینی باشند. بنابراین انتظار نداریم که اعمال حسابی را در یک ماشین به دقت انجام شود. در اینجا می‌توان اعمال جایگزین $+^*$، $-^*$، $\times^*$ و $/^*$ که اعمال ممیز سیار نام دارند را در نظر گرفت که اعمال حسابی را تا حد ممکن تقریب می‌کنند. بر اساس نگاشت گرد کردن rd می‌توان این اعمال را به صورت زیر تعریف کرد

$$\begin{aligned} x +^* y &:= rd(x + y), \\ x -^* y &:= rd(x - y), \qquad \forall x, y \in A \\ x \times^* y &:= rd(x * y), \\ x /^* y &:= rd(x/y), \end{aligned} \tag{۸}$$

لذا از (۶) نتیجه می‌شود

$$\begin{aligned} x +^* y &:= (x + y)(۱ + \varepsilon_۱), \\ x -^* y &:= (x - y)(۱ + \varepsilon_۲), \qquad |\varepsilon_i| \leq eps. \\ x \times^* y &:= (x * y)(۱ + \varepsilon_۳), \\ x /^* y &:= (x/y)(۱ + \varepsilon_۴), \end{aligned} \tag{۹}$$

منظور از $fl(E)$ محاسبه عبارت E با حساب ممیز سیار است. اگر E یک عبارت مقدماتی باشد، $fl(E) = rd(E)$. اعمال ممیز سیار در قوانین شناخته شده اعمال حسابی صدق نمی‌کنند. برای مثال،

$$x +^* y = x \quad if \ \ |y| < \frac{eps}{B}|x|, \quad x, y \in A,$$

که در آن B مبنای دستگاه اعداد است. دقت ماشین را می توان به صورت کوچکترین عدد مثبت ماشینی g که $۱ +^* g > ۱$ تعریف کرد:

$$eps = \min\{g \in A | ۱ +^* g > ۱, \ g > ۰\}.$$

اعمال ممیز سیار لزوما جابجایی و توزیع پذیر نیستند.
مثال ($t = ۸$) با فرض

$$
\begin{aligned}
a &:= ۰.۲۳۳۷۱۲۵۸ \times ۱۰^{-۴},\\
b &:= ۰.۳۳۶۷۸۴۲۹ \times ۱۰^{۲},\\
c &:= -۰.۳۳۶۷۷۸۱۱ \times ۱۰^{۲},
\end{aligned}
$$

داریم

$$
\begin{aligned}
a +^* (b +^* c) &= ۰.۲۳۳۷۱۲۵۸ \times ۱۰^{-۴} +^* ۰.۶۱۸۰۰۰۰۰ \times ۱۰^{-۳},\\
&= ۰.۶۴۱۳۷۱۲۶ \times ۱۰^{-۳},\\
(a +^* b) +^* c &= ۰.۳۳۶۷۸۴۵۲ \times ۱۰^{۲} -^* ۰.۳۳۶۷۷۸۱۱ \times ۱۰^{۲},\\
&= ۰.۶۴۱۰۰۰۰۰ \times ۱۰^{-۳}.
\end{aligned}
$$

مقدار دقیق عبارتست از

$$a + b + c = ۰.۶۴۱۳۷۱۲۵۸ \times ۱۰^{-۳}.$$

در تفریق دو عدد $x, y \in A$ با علامت یکسان باید مواظب خنثی سازی(cancellation) بود. این وقتی پیش می آید که x و y یک یا چند رقم پیشرو مشترک با یک نما داشته باشند، برای مثال

$$
\begin{aligned}
x &= ۰.۳۱۵۸۷۶ \times ۱۰^{۱},\\
y &= ۰.۳۱۴۲۸۹ \times ۱۰^{۱}.
\end{aligned}
$$

تفریق موجب از بین رفتن ارقام پیشرو مشترک می‌شود. مقدار دقیق $y - x$ یک عدد ماشینی است، لذا خطای گردکردن جدیدی پیش نمی آید: $y -^* x = y - x$. بنابراین، تفریق در حالت خنثی سازی یک عمل کاملا بی‌خطر است. با این حال، خنثی سازی به دلیل انتشار خطاهای قبلی بسیار خطرناک است.

اگر نمایش عدد ماشینی نرمال $\pm ۰.\alpha_۱\alpha_۲ \ldots \alpha_s \times \beta^e$ را در ماشین‌های مختلف در نظر بگیریم با فرض $m \le e \le M$ جدول زیر را خواهیم داشت

ماشین	β	s	m	M	eps
VAX	۲	۲۴	−۱۲۸	۱۲۷	۶.۰×۱۰^{-۰۸}
VAX	۲	۵۶	−۱۲۸	۱۲۷	۱.۴×۱۰^{-۱۷}
CRAY-1	۲	۴۸	−۱۶۳۸۴	۱۶۳۸۳	۳.۶×۱۰^{-۱۵}
IBM 3081	۱۶	۶	−۶۴	۶۳	۹.۵×۱۰^{-۰۷}
IBM 3081	۱۶	۱۴	−۶۴	۶۳	۲.۲×۱۰^{-۱۶}
IEEE					
single	۲	۲۴	−۱۲۵	۱۲۸	۶.۰×۱۰^{-۰۸}
double	۲	۵۳	−۱۰۲۱	۱۰۲۴	۱.۱×۱۰^{-۱۶}

۲ انتشار خطا

۱.۲ مقدمه

در عمل وقتی توابع مرکب از چندین تابع را با داده‌های اولیه خطادار محاسبه می‌کنیم در هر مرحله‌ای خطایی مرتکب می‌شویم تا جایی که در بعضی از موارد نتیجه حاصل به طور کامل غلط است. بنابراین لازم است قبل از محاسبه یک تابع از رفتار آن آگاه باشیم. اینجاست که مفهوم ناپایداری عددی ظاهر می‌شود؛ با یک بیان ساده و نه عبارت دقیق، گوییم یک روند عددی ناپایدار است در صورتی که خطاهای کوچک حاصل از یک مرحله این روند محاسباتی دقت محاسبات در مراحل بعدی را به صورت محسوس کاهش دهد.

مثال. دنباله عددی زیر را در نظر می‌گیریم

$$\begin{cases} x_۰ = ۱, & x_۱ = \frac{۱}{۳} \\ x_{n+۱} = \frac{۱۳}{۳}x_n - \frac{۴}{۳}x_{n-۱}, & n \geq ۱. \end{cases} \qquad (۱۰)$$

با حل این رابطه بازگشتی داریم

$$x_n = (\frac{۱}{۳})^n \qquad (۱۱)$$

اگر برای محاسبه این دنباله از (۱۰) استفاده کنیم در یک ماشین ۳۲ بیتی نتایج زیر بدست می‌آید که بعضی از نتایج بدست آمده به صورت کامل نادرست است.

$x_۰$	=	۱.۰۰۰۰۰۰۰۰					
$x_۱$	=	۰.۳۳۳۳۳۳۳	۷ رقم بامعنا	$x_۸$	=	۰.۰۰۰۳۷۵۷	۰ رقم بامعنا
$x_۲$	=	۰.۱۱۱۱۱۱۲	۶ رقم بامعنا	$x_۹$	=	۰.۰۰۰۹۴۳۷	
$x_۳$	=	۰.۰۳۷۰۳۷۳	۵ رقم بامعنا	$x_{۱۰}$	=	۰.۰۰۳۵۸۸۷	
$x_۴$	=	۰.۰۱۲۳۴۶۶	۴ رقم بامعنا	$x_{۱۱}$	=	۰.۰۱۴۲۹۲۷	
$x_۵$	=	۰.۰۰۴۱۱۸۷	۳ رقم بامعنا	$x_{۱۲}$	=	۰.۰۵۷۱۵۰۲	
$x_۶$	=	۰.۰۰۱۳۸۵۷	۲ رقم بامعنا	$x_{۱۳}$	=	۰.۲۲۸۵۹۳۹	
$x_۷$	=	۰.۰۰۰۵۱۳۱	۱ رقم بامعنا	$x_{۱۴}$	=	۰.۹۱۴۳۷۳۵	
				$x_{۱۵}$	=	۳.۶۵۷۴۹۳	نادرست! با خطای نسبی $۱۰^۸$

این نتایج را می‌توان با کد ساده زیر در MATLAB محاسبه کرد. بر اساس این برنامه $x_{۲۹} = -۱۴.۲۷۳$:

```
x(1) = 1;
x(2) = 1/3;
for n=2:30
    x(n+1) = 13/3*x(n)-4/3*x(n-1);
end
x'
```

این الگوریتم ناپایدار است. خطای موجود در x_n در محاسبه x_{n+1} با عامل $\frac{13}{3}$ افزایش می‌یابد. یعنی خطا در x_1 با عامل $(\frac{13}{3})^{14}$ به x_{15} انتشار می یابد. با توجه به اینکه خطای مطلق x_1 تقریبا برابر 10^{-8} است و در x_{15} خطا به عامل $(\frac{13}{3})^{14}$ که تقریبا برابر است با 10^9 ضرب می‌شود لذا، خطا در x_{15} تقریبا برابر است با ۱۰. بنابراین، نتایج بدست آمده از نرم افزارهای پیشرفته در صورتی قابل اعتماد است که از نظر تئوری رفتار خطا بررسی شده باشد.

مثال. مثال دیگر برای ناپایداری عددی عبارتست از محاسبه انتگرال

$$y_n = \int_0^1 x^n e^x dx, \quad n \geq 0 \tag{۱۲}$$

اگر از انتگرال گیری جز به جز برای y_{n+1} استفاده کنیم داریم:

$$y_{n+1} = e - (n+1)y_n \tag{۱۳}$$

با استفاده از این رابطه و $y_0 = e - 1$ داریم:

$$y_1 = e - y_0 = e - (e-1) = 1$$

با شروع از $y_1 = 1$ بر روی یک ماشین ۳۲ بیتی مقادیر $y_2, y_3, \ldots, y_{15}$ با رابطه (۱۳) محاسبه می‌کنیم. بعضی از این نتایج عبارتند از

$$y_2 = .7182817$$
$$y_{11} = 1.422453$$
$$y_{15} = 39711.43$$

این نتایج نمی‌تواند درست باشد. از (۱۲) واضح است که $0 < \ldots < y_2 < y_1$ و $\lim_{n\to\infty} y_n = 0$.

۲.۲ عدد حالت

این بحث را با یک سوال آغاز می‌کنیم. در محاسبه تابع $f(x)$ اگر x کمی پریشیده شود تاثیر آن بر $f(x)$ چه خواهد بود؟ اگر در این سوال هدف خطای مطلق باشد از قضیه مقدار میانگین داریم:

$$f(x+h) - f(x) = f'(\xi)h \approx hf'(x) \tag{۱۴}$$

بنابراین در صورتی که $f'(x)$ زیاد بزرگ نباشد تاثیر این پریشیدگی روی $f(x)$ کم است. با این وجود، خطای نسبی وضعیت را بهتر نمایان می‌کند. وقتی x به میزان h پریشیده می‌شود $f(x)$ به $f(x+h)$ پریشیده می‌شود و خطای نسبی برابر است با

$$\frac{f(x+h)-f(x)}{f(x)} \approx \frac{hf'(x)}{f(x)} = [\frac{xf'(x)}{f(x)}](\frac{h}{x}) \tag{۱۵}$$

عامل $\frac{xf'(x)}{f(x)}$ **عدد حالت** این مساله است.

مثال. عدد حالت برای محاسبه تابع $f(x) = \sin^{-1} x$ چیست؟

$$\frac{xf'(x)}{f(x)} = \frac{x}{\sqrt{1-x^2}\sin^{-1} x} \tag{۱۶}$$

اگر x نزدیک ۱ باشد، $\sin^{-1} x \approx \pi/2$ و عدد حالت به بی‌نهایت میل خواهد کرد. لذا، در نزدیکی $x = 1$ خطاهای نسبی کوچک به خطاهای نسبی بزرگ در $\sin^{-1} x$ منجر خواهد شد.

مثال. در این مثال مساله پیدا کردن ریشه یک تابع را در نظر می‌گیریم. این مساله در بخش‌های بعدی به صورت الگوریتمی بررسی خواهد شد. فرض کنید توابع f و g در همسایگی r متعلق به رده توابع C^2 باشند، که r ریشه f است. فرض می‌کنیم r یک ریشه ساده باشد یعنی؛ $f'(r) \neq 0$. اگر تابع f را به $F \equiv f + \varepsilon g$ پریشیده کنیم ریشه جدید در کجا خواهد بود؟ فرض کنیم ریشه جدید $r + h$ باشد؛ می‌خواهیم یک فرمول تقریبی برای h بدست آوریم. پریشیدگی h در معادله $F(r + h) = 0$، به طور معادل در رابطه زیر، صدق می‌کند:

$$f(r + h) + \varepsilon g(r + h) = 0$$

با توجه به اینکه f و g در C^2 هستند می‌توان بسط تیلور را برای $F(r + h) = 0$ نوشت:

$$[f(r) + hf'(r) + \frac{1}{2}h^2 f''(\xi)] + \varepsilon[g(r) + hg'(r) + \frac{1}{2}h^2 g''(\eta)] = 0$$

با صرف نظر از جملات شامل h^2 و توجه به این نکته که $f(r) = 0$، داریم

$$h \approx -\varepsilon \frac{g(r)}{f'(r) + \varepsilon g'(r)} \approx -\varepsilon \frac{g(r)}{f'(r)}$$

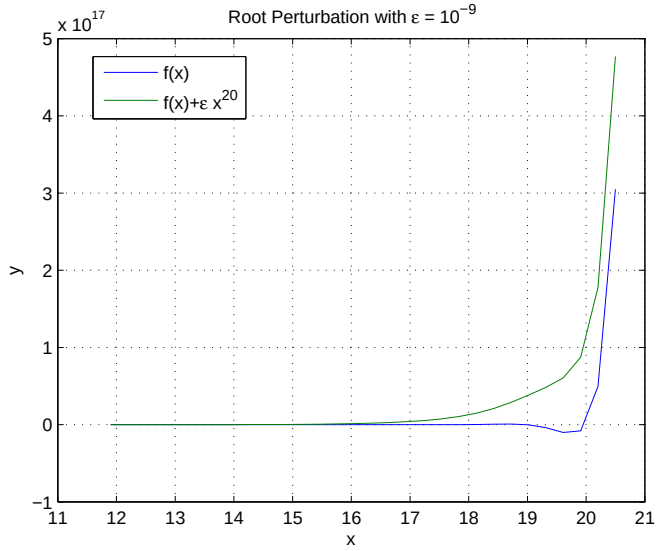
برای بررسی بهتر مثال زیر را در نظر می‌گیریم

$$\begin{aligned} f(x) &= \prod_{k=1}^{20} (x - k) = (x - 1)(x - 2) \dots (x - 20) \\ g(x) &= x^{20} \end{aligned}$$

به وضوح ریشه‌های f برابر است با $1, 2, \dots, 20$. اگر f را با $f + \varepsilon g$ جایگزین کنیم ریشه $r = 20$ چه تاثیری روی ریشه $r = 20$ خواهد داشت؟ جواب عبارتست از

$$h \approx -\varepsilon \frac{g(20)}{f'(20)} = -\varepsilon \frac{20^{20}}{19!} \approx -10^9 \varepsilon$$

بنابراین، یک تغییر ε در ضریب x^{20} در $f(x)$ یک پریشیدگی به اندازه $10^9 \varepsilon$ در ریشه ۲۰ ایجاد می‌کند. در نتیجه، ریشه‌های این چندجمله‌ای به پریشیدگی در ضرایب بسیار حساس هستند.



شکل فوق انحراف ریشه $x = ۲۰$ را با پریشیدگی $\varepsilon = ۱۰^{-۹}$ نشان می‌دهد. با این پریشیدگی تغییر قابل توجهی در مکان این ریشه بوجود می‌آید.

۳.۲ پایداری الگوریتم و خطای ذاتی

تابع مفروض $\phi : D \subseteq R^n \to R^m$، $y = \phi(x)$ را در نظر می‌گیریم و آن را به دنباله‌ای از توابع مقدماتی تجزیه می‌کنیم. برای راحتی نمادهایی را هم معرفی می‌کنیم

$$\phi^{(i)} : D_i \to D_{i+۱}, \quad i = ۰, ۱, \ldots, r, \quad D_۰ = D, D_i \subseteq R^{n_i}, D_{r+۱} \subseteq R^{n_{r+۱}} = R^m$$
$$\phi = \phi^{(r)} \circ \phi^{(r-۱)} \circ \ldots \circ \phi^{(۰)}$$
$$x^{(i)} = \begin{pmatrix} x_۱^{(i)} \\ \vdots \\ x_{n_i}^{(i)} \end{pmatrix}, \quad \phi^{(i)}(x^{(i)}) = x^{(i+۱)}$$
$$x = x^{(۰)} \to \phi^{(۰)}(x^{(۰)}) = x^{(۱)} \to \ldots \to \phi^{(r)}(x^{(r)}) = x^{(r+۱)} = y$$

چنین تجزیه‌ای منحصربفرد نیست. مثال زیر را در نظر می‌گیریم

مثال. برای محاسبه $\phi(a, b, c) = a + b + c$ از دو الگوریتم زیر می‌توان استفاده کرد

$$\phi^{(۰)}(a, b, c) = \begin{pmatrix} a+b \\ c \end{pmatrix} \in R^۲, \quad \phi^{(۱)}(u, v) = u + v \in R$$
$$\phi^{(۰)}(a, b, c) = \begin{pmatrix} a \\ b+c \end{pmatrix} \in R^۲, \quad \phi^{(۱)}(u, v) = u + v \in R.$$

همانطور که قبلا هم بیان شد این دو الگوریتم از نظر عددی یکسان نیستند. در این بخش با بررسی خطاهای گرد کردن در روند محاسبه یک تابع بر اساس توابع مقدماتی به یک معیار کلی برای ارزیابی الگوریتم‌های مختلف برای حل یک مساله می‌رسیم.

فرض کنید

$$\phi(x) = \begin{pmatrix} \phi_1(x_1, \ldots, x_n) \\ \vdots \\ \phi_m(x_1, \ldots, x_n) \end{pmatrix}, \quad \phi^{(i)}(u) = \begin{pmatrix} \phi_1^{(i)}(u) \\ \vdots \\ \phi_{n_{i+1}}^{(i)}(u) \end{pmatrix}$$

اگر $\tilde{x}$ تقریب x باشد برای محاسبه $y = \phi(x)$ در عمل به دلیل وجود خطاهای گردکردن $\tilde{y} = \phi(\tilde{x})$ محاسبه می‌شود. بنابراین می‌توان تاثیر خطاهای اولیه $\Delta x_i = \tilde{x_i} - x_i$ بر y را بدست آورد. به عبارت دیگر می‌توان تقریبی برای خطای $\Delta y_i = \tilde{y_i} - y_i$ تعیین کرد. با بسط تیلور و صرف نظرکردن از جملات مرتبه بالا داریم

$$\begin{aligned} \Delta y_i &= \tilde{y_i} - y_i = \phi_i(\tilde{x}) - \phi_i(x) = \sum_{j=1}^{n} (\tilde{x_i} - x_i) \frac{\partial \phi_i(x)}{\partial x_j} \\ &= \sum_{j=1}^{n} \frac{\partial \phi_i(x)}{\partial x_j} \Delta x_i, \quad i = 1, \ldots, m, \end{aligned} \tag{۱۷}$$

اگر به صورت ماتریسی بنویسیم داریم

$$\Delta y = \begin{bmatrix} \Delta y_1 \\ \vdots \\ \Delta y_m \end{bmatrix} \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial \phi_m}{\partial x_1} & \cdots & \frac{\partial \phi_m}{\partial x_n} \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \vdots \\ \Delta x_m \end{bmatrix} = D\phi(x) \Delta x. \tag{۱۸}$$

که در آن $D\phi(x)$ ماتریس ژاکوبی است. با تقسیم طرفین (۱۷) به $y_i = \phi_i(x)$ داریم

$$\varepsilon_{y_i} = \sum_{j=1}^{n} \frac{x_j}{\phi_i(x)} \frac{\partial \phi_i(x)}{\partial x_j} \varepsilon_{x_j}, \quad i = 1, \ldots, m. \tag{۱۹}$$

که ε_{x_i} و ε_{y_i} به ترتیب خطای نسبی x_i و y_i هستند. برای ادامه بحث به مفهوم نگاشت مانده نیاز خواهیم داشت که به صورت زیر تعریف می‌شود

$$\psi^{(i)} = \phi^{(r)} \circ \phi^{(r-1)} \circ \ldots \circ \phi^{(i)} : D_i \to R^m, \quad i = 0, 1, 2, \ldots, r. \tag{۲۰}$$

در این صورت $\psi^{(0)} = \phi$ و $D\phi^{(i)}$ و $D\psi^{(i)}$ ماتریسهای ژاکوبی نگاشت‌های $\phi^{(i)}$ و $\psi^{(i)}$ هستند. با توجه به

$$D(f \circ g)(x) = Df(g(x)).Dg(x),$$

داریم

$$D\phi(x) = D\phi^{(r)}(x^{(r)}).D\phi^{(r-1)}(x^{(r-1)}). \ldots .\phi^{(0)}(x), \tag{۲۱}$$

$$D\psi^{(i)}(x^{(i)}) = D\phi^{(r)}(x^{(r)}).D\phi^{(r-1)}(x^{(r-1)}). \ldots .\phi^{(i)}(x^{(i)}), \quad i = 0, 1, 2, \ldots, r. \tag{۲۲}$$

خطاهای ورودی و خطاهای گردکردن مقادیر $x^{(i)}$ را پریشیده می‌کنند و مقادیر تقریبی $\tilde{x}^{(i)}$ با $\tilde{x}^{(i+1)} = fl(\phi^{(i)}(\tilde{x}^{(i)}))$ محاسبه می‌شوند. با داشتن خطاهای $\Delta x^{(i)} = \tilde{x}^{(i)} - x^{(i)}$ داریم

$$\Delta x^{(i+1)} = [fl(\phi^{(i)}(\tilde{x}^{(i)}) - \phi^{(i)}(\tilde{x}^{(i)})] + [\phi^{(i)}(\tilde{x}^{(i)}) - \phi^{(i)}(x^{(i)})] \tag{۲۳}$$

با صرف نظرکردن از جملات مرتبه بالا، طبق (۱۸) داریم

$$\phi^{(i)}(\tilde{x}^{(i)}) - \phi^{(i)}(x^{(i)}) = D\phi^{(i)}(x^{(i)})\Delta x^{(i)}. \qquad (۲۴)$$

با توجه به اینکه نگاشت $\phi^{(i)}$ یک نگاشت مقدماتی است داریم

$$fl(\phi^{(i)}(u)) = rd(\phi^{(i)}(u)). \qquad (۲۵)$$

اگر (۲۵) را به صورت مولفه‌ای بنویسیم داریم

$$fl(\phi_j^{(i)}(u)) = rd(\phi_j^{(i)}(u)) = (۱ + \varepsilon_j)\phi_j^{(i)}(u), \qquad |\varepsilon_j| < eps, j = ۱, ۲, \ldots n_{i+۱} \qquad (۲۶)$$

لذا (۲۵) را می‌توان به صورت زیر نوشت

$$fl(\phi^{(i)}(u)) = (I + E_{i+۱})\phi^{(i)}(u) \qquad (۲۷)$$

که در آن I ماتریس همانی و $E_{i+۱}$ ماتریس خطای قطری است

$$E_{i+۱} = \begin{bmatrix} \varepsilon_۱ & & & ۰ \\ & \varepsilon_۲ & & \\ ۰ & & \ddots & \\ & & & \varepsilon_{n_{i+۱}} \end{bmatrix}, \qquad |\varepsilon_j| \le eps. \qquad (۲۸)$$

بنابراین برای عبارت اول در طرف راست رابطه (۲۳) داریم

$$fl(\phi^{(i)}(\tilde{x}^{(i)})) - \phi^{(i)}(\tilde{x}^{(i)}) = E_{i+۱}.\phi^{(i)}(\tilde{x}^{(i)}).$$

علاوه بر آن، اگر عبارت $\phi^{(i)}(\tilde{x}^{(i)})$ را حول $x^{(i)}$ بسط داده و به $E_{i+۱}$ ضرب کنیم جملات مرتبه بالای خطا پدیدار می‌شوند که با چشم‌پوشی از آنها نتیجه می‌شود $E_{i+۱}.\phi^{(i)}(\tilde{x}^{(i)}) \approx E_{i+۱}.\phi^{(i)}(x^{(i)})$. لذا، داریم

$$fl(\phi^{(i)}(\tilde{x}^{(i)})) - \phi^{(i)}(\tilde{x}^{(i)}) = E_{i+۱}.\phi^{(i)}(x^{(i)}) = E_{i+۱}.x^{(i+۱)} =: \alpha_{i+۱}. \qquad (۲۹)$$

$\alpha_{i+۱}$ را می‌توان به عنوان خطای مطلق گردکردن که در محاسبه $\phi^{(i)}$ تازه تولید شده، تعبیر کرد و همچنین اعضای قطری $E_{i+۱}$ خطاهای نسبی متناظر هستند. در نتیجه، بنابر (۲۳)، (۲۴) و (۲۹) عبارت $\Delta x^{(i+۱)}$ را می‌توان به صورت تقریب مرتبه اول زیر بیان کرد

$$\Delta x^{(i+۱)} = \alpha_{i+۱} + D\phi^{(i)}(x^{(i)}).\Delta x^{(i)} = E_{i+۱}.x^{(i+۱)} + D\phi^{(i)}(x^{(i)}).\Delta x^{(i)},$$
$$i \ge ۰, \qquad \Delta x^{(۰)} := \Delta x.$$

و لذا داریم

$$\begin{aligned} \Delta y = \Delta x^{(i+۱)} &= \alpha_{r+۱} + D\phi^{(r)}(x^{(r)}).\Delta x^{(r)} \\ &= \alpha_{r+۱} + D\phi^{(r)}(x^{(r)}).(\alpha_r + D\phi^{(r-۱)}(x^{(r-۱)}).\Delta x^{(r-۱)}) \\ &\vdots \\ &= \alpha_{r+۱} + D\phi^{(r)}(x^{(r)}).\alpha_r + D\phi^{(r)}(x^{(r)}).D\phi^{(r-۱)}(x^{(r-۱)}).\alpha_{r-۱} + \ldots \\ &\qquad + D\phi^{(r)}(x^{(r)}) \ldots D\phi^{(۱)}(x^{(۱)}).\alpha_۱ + D\phi^{(r)}(x^{(r)}) \ldots D\phi^{(۰)}(x^{(۰)}).\Delta x. \end{aligned}$$

بنابراین با استفاده از نگاشت مانده داریم

$$\begin{aligned}\Delta y &= \alpha_{r+1} + D\psi^{(r)}(x^{(r)}).\alpha_r + \ldots + D\psi^{(1)}(x^{(1)}).\alpha_1 + D\phi(x).\Delta x \\ &= E_{r+1}.y + D\psi^{(r)}(x^{(r)}).E_r x^{(r)} + \ldots + D\psi^{(1)}(x^{(1)}).E_1 x^{(1)} + D\phi(x).\Delta x. \qquad (۳۰)\end{aligned}$$

اگر الگوریتم‌های مختلفی برای محاسبه ϕ وجود داشته باشد، در عبارت خطا $D\phi$ تغییر نخواهد کرد. با این وجود، ماتریسهای ژاکوبی $D\psi^{(i)}$ که انتشار خطای گردکردن را ارزیابی می‌کنند تغییر خواهند کرد، و لذا تاثیر کلی خطای گردکردن به صورت زیر خواهد بود

$$\alpha_{r+1} + D\psi^{(r)}(x^{(r)}).\alpha_r + \ldots + D\psi^{(1)}(x^{(1)}).\alpha_1. \qquad (۳۱)$$

برای محاسبه $\phi(x)$، به ازای مجموعه‌ای از داده‌های x، یک الگوریتم را در مقایسه با الگوریتم دیگر به طور عددی قابل اعتماد(trustworthy) گوییم اگر تاثیر کلی خطاهای گرد کردن (۳۱) برای الگوریتم اول در مقایسه با الگوریتم دوم کمتر باشد.

در عبارت خطای (۳۰) جمله اول، بدون وابستگی به الگوریتم استفاده شده، دارای کران زیر است

$$|E_{r+1}y| \leq |y|eps,$$

لذا اندازه خطای $|y|eps$ برای هر الگوریتمی قابل انتظار است. علاوه بر آن، با استفاده از اعداد با مفسر t- رقمی داده‌های ورودی دارای خطای زیر هستند (جز در حالتی که داده‌های اولیه اعداد ماشینی باشند)

$$\Delta^{(0)}x \leq |x|eps,$$

در نتیجه برای هر الگوریتمی خطای زیر همواره پیش‌بینی می‌شود

$$\Delta^{(0)}y := [|D\phi(x)|.|x| + |y|]eps \qquad (۳۲)$$

$\Delta^{(0)}y$ خطای ذاتی y نامیده می‌شود. چون این خطا همواره وجود دارد، لذا نباید انتظار داشت که خطاهای گردکردن در محاسبات میانی از $\Delta^{(0)}y$ کمتر باشند. بنابراین خطاهای گردکردن α_i یا E_i در (۳۰) را بی‌ضرر گوییم اگر خطای کلی Δy حداکثر از مرتبه خطای ذاتی $\Delta^{(0)}y$ باشد:

$$|D\psi^{(i)}(x^{(i)}).\alpha_i| = |D\psi^{(i)}(x^{(i)}).E_i x^{(i)}| \approx \Delta^{(0)}y.$$

اگر خطاهای گردکردن یک الگوریتم بی‌ضرر باشند، در این صورت الگوریتم را خوش‌رفتار یا به طور عددی پایدار گوییم.

مثال. محاسبه $\phi(a,b) = a^2 - b^2 = (a+b)(a-b)$ از دو الگوریتم زیر را در نظرمی‌گیریم

الگوریتم۱: $\phi^{(0)}(a,b) = \begin{bmatrix} a^2 \\ b \end{bmatrix}$, $\phi^{(1)}(u,v) = \begin{bmatrix} u \\ v^2 \end{bmatrix}$, $\phi^{(2)}(u,v) = u - v$

الگوریتم۲: $\phi^{(0)}(a,b) = \begin{bmatrix} a \\ b \\ a+b \end{bmatrix}$, $\phi^{(1)}(a,b,u) = \begin{bmatrix} u \\ a-b \end{bmatrix}$, $\phi^{(2)}(u,v) = u.v$

برای محاسبه خطاهای کلی گردکردن و خطای ذاتی داریم

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

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

$$D\phi(x) = (2a, -2b),$$

$$D\psi^{(1)}(x^{(1)}) = (1, -2b), \qquad D\psi^{(2)}(x^{(2)}) = (1, -1)$$

با توجه به $fl(\phi^{(0)}(x^{(0)})) - \phi^{(0)}(x^{(0)}) = \begin{bmatrix} a \times^* a \\ b \end{bmatrix} - \begin{bmatrix} a^2 \\ b \end{bmatrix}$ داریم

$$\alpha_1 = \begin{bmatrix} \varepsilon_1 a^2 \\ 0 \end{bmatrix}, \qquad E_1 = \begin{bmatrix} \varepsilon_1 & 0 \\ 0 & 0 \end{bmatrix},$$

و به طور مشابه،

$$\alpha_2 = \begin{bmatrix} 0 \\ \varepsilon_2 b^2 \end{bmatrix}, \qquad E_2 = \begin{bmatrix} 0 & 0 \\ 0 & \varepsilon_2 \end{bmatrix},$$

$$\alpha_3 = \varepsilon_3(a^2 - b^2), \qquad |\varepsilon_i| \le eps, \qquad i = 1, 2, 3.$$

با $\Delta x = \begin{bmatrix} \Delta a \\ \Delta b \end{bmatrix}$ از (۳۰) داریم

$$\Delta y = 2a\Delta a - 2b\Delta b + a^2\varepsilon_1 - b^2\varepsilon_2 + (a^2 - b^2)\varepsilon_3. \tag{۳۳}$$

و تاثیر کلی خطاهای گردکردن عبارتست از

$$|a^2\varepsilon_1 - b^2\varepsilon_2 + (a^2 - b^2)\varepsilon_3| \le (a^2 + b^2 + |a^2 - b^2|)eps. \tag{۳۴}$$

به طور مشابه برای الگوریتم ۲ داریم

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

$$\psi^{(1)}(u,v) = u.v,$$

$$D\phi(x) = (2a, -2b), \qquad D\psi^{(1)}(x^{(1)}) = (a - b, a + b),$$

$$\alpha_1 = \begin{bmatrix} \varepsilon_1(a + b) \\ \varepsilon_2(a - b) \end{bmatrix}, \qquad \alpha_2 = \varepsilon_3(a^2 - b^2), \qquad E_1 = \begin{bmatrix} \varepsilon_1 & 0 \\ 0 & \varepsilon_2 \end{bmatrix}, \qquad |\varepsilon_i| \le eps,$$

دوباره از (۳۰) داریم

$$\Delta y = 2a\Delta a - 2b\Delta b + (a^2 - b^2)(\varepsilon_1 + \varepsilon_2 + \varepsilon_3). \tag{۳۵}$$

$$|(a^2 - b^2)(\varepsilon_1 + \varepsilon_2 + \varepsilon_3)| \le 3|a^2 - b^2|eps. \tag{۳۶}$$

با مقایسه (۳۴) و (۳۶) به راحتی می‌توان نشان داد که اگر $|\frac{a}{b}|^2 > 3$ یا $|\frac{a}{b}|^2 < \frac{1}{3}$ الگوریتم ۲ و در غیر این صورت الگوریتم ۱ قابل اعتمادتر است.

۴.۲ جمع سری و واگرایی عددی

بنا بر تعریف، سری $\sum_{n=۱}^{\infty} x_n$ به عدد s همگراست هرگاه، به ازای $\varepsilon > ۰$ دلخواه عدد صحیح $N > ۰$ وجود داشته باشد به طوری که

$$|s - \sum_{n=۱}^{N} x_n| < \varepsilon \qquad (۳۷)$$

بنابراین با جمع تعداد کافی از جملات این سری می‌توان به اندازه دلخواه به مقدار واقعی سری دست یافت. اما این موضوع با اضافه کردن جملات گرد شده درست نیست. الگوریتم جمع کردن این سری با ترتیب معمولی عبارتست از

$$s_{k+۱} = s_k + x_{k+۱}$$

که $s_k = \sum_{n=۱}^{k} x_n$ مجموع جزئی سری مورد نظر است.
دو منبع اصلی خطا در جمع سریها وجود دارد:

- در صورتی که همه جمله‌ها مثبت باشند، خطاها انباشته می‌شوند و در مجموع های جزئی خطاهای بزرگی را پدید می‌آورند.
- در سری‌های متناوب که دارای جملات با علامت متناوب هستند، بر اساس تحلیل ریاضی

$$s_{k+۱} = s_k + x_{k+۱} = s_{k-۱} + x_{k+۱} + x_k$$

(علامت x_k با $x_{k+۱}$ متفاوت است) نشان می‌دهد که در صورتی که $|x_{k+۱}| \approx |x_k|$ خطای خنثی سازی ارقام بامعنی اتفاق می‌افتد.

در اینجا دو مثال از واگرایی عددی ناشی از گردکردن ارائه می‌کنیم.
مثال۱ . سری زیر را در نظر می‌گیریم [۴]

$$\begin{aligned} &۰.۵ - ۰.۴ \\ &+ \underbrace{۰.۰۵ - ۰.۰۴۹ + ۰.۰۵ - ۰.۰۴۹ + \ldots + ۰.۰۵ - ۰.۰۴۹}_{۲۰ \text{ جمله}} \\ &+ \underbrace{۰.۰۰۵ - ۰.۰۰۴۹۹ + ۰.۰۰۵ - ۰.۰۰۴۹۹ + \ldots + ۰.۰۰۵ - ۰.۰۰۴۹۹}_{۲۰۰ \text{ جمله}} \\ &+ \ldots \end{aligned} \qquad (۳۸)$$

به وضوح جمع این سری عبارتست از

$$s = ۰.۱ + ۱۰ \times ۰.۰۰۱ + ۱۰۰ \times ۰.۰۰۰۰۱ + \ldots = \frac{۱}{۹} \qquad (۳۹)$$

اگر جملات سری تا n رقم اعشار گرد شوند، جملات سطر $(n+۱)$ام به بعد ناپدید می‌شوند. در سطرهای باقی مانده شرایط زیر حاکم است: جملات از سطر ۱ تا سطر $[\frac{n+۱}{۲}]$ تاثیری از گرد کردن نخواهند داشت. جملات از سطر $[\frac{n+۱}{۲}] + ۱$ تا n دوتا دوتا همدیگر را حذف می‌کنند(خطای خنثی سازی). در نهایت در سطر $n+۱$ مجموع برابر است

با $۱ = ۱۰^{n} \times ۱۰^{-n}$. بنابراین مجموع جملات با گرد کردن تا n رقم اعشار حداقل ۱ خواهد شد.

قضیه. فرض کنید $x_۰$، $x_۱$،...،x_n اعداد مثبت ماشینی در یک رایانه باشند که خطای گردکردن آن ε باشد. آنگاه خطای نسبی گردکردن در محاسبه $\sum_{k=۰}^{n} x_i$ به صورت معمولی، حداکثر برابر است با $(۱ + \varepsilon)^{n} - ۱$. این کمیت تقریبا برابر $n\varepsilon$ است.

مثال ۲. محاسبه $\sin(۴۰)$ را با سری زیر در نظر می‌گیریم

$$\sin x = \sum_{n=۰}^{\infty} \frac{(-۱)^{n}}{(۲n+۱)!} x^{۲n+۱}$$

با در نظر گرفتن ۸۰ جمله از این سری با دقت معمولی برابر است با $\sin(۴۰.۰) = ۵.۲۳۴۴۳۱۴E+۰۸$، اما مقدار دقیق آن برابر است با $\sin(۴۰.۰) = ۰.۷۴۵۱۱۳۱$.

۳ مسائل

۱) در محاسبه سری نامتناهی $\sum_{n=۱}^{\infty} x_n$، جواب با خطای مطلق کمتر از ε مورد نیاز است. آیا توقف جمع کردن تا جملاتی که اندازه آنها کمتر از ε است روش مطمئنی است؟ پاسخ را با سری $\sum_{n=۱}^{\infty} (۰.۹۹)^{n}$ نشان دهید.

۲) مساله قبلی را در حالتی که جملات سری به طور متناوب مثبت و منفی باشند و $|x_n|$ یکنوا به صفر همگرا شود بررسی کنید.(از قضیه همگرایی سریهای متناوب از ریاضیات عمومی استفاده کنید.)

۳) چرا از دست رفتن ارقام بامعنی در تقریب زیر مهم نیست

$$x - \sin x \approx (\frac{x^{۳}}{۶})(۱ - (\frac{x^{۲}}{۲۰})(۱ - \frac{x^{۲}}{۴۲})).$$

۴) ضرب نقطه‌ای دو بردار زیر را به چند طریق محاسبه کنید:

$$x = [۲.۷۱۸۲۸۱۸۲۸, -۳.۱۴۱۵۹۲۶۵۴, ۱.۴۱۴۲۱۳۵۶۲, ۰.۵۷۷۲۱۵۶۶۴۹, ۰.۳۰۱۰۲۹۹۹۵۷]$$

$$y = [۱۴۸۶.۲۴۹۷, ۸۷۸۳۶۶.۹۸۷۹, -۲۲.۳۷۴۹۲, ۴۷۷۳۷۱۴.۶۴۷, ۰.۰۰۰۱۸۵۰۴۹]$$

الف) $\sum_{i=۱}^{n} x_i y_i$

ب) $\sum_{i=n}^{۱} x_i y_i$

ج) با ترتیب بزرگترین — به — کوچکترین (اعداد مثبت را به ترتیب بزرگترین به کوچکترین جمع کنید و سپس اعداد منفی را از کوچکترین به بزرگترین جمع کنید و دو مجموع جزئی را با هم جمع کنید).

د) کوچکترین — به — بزرگترین (ترتیب عکس روش قبلی)

دقت معمولی و دقت مضاعف را برای جوابها بکار ببرید. نتایج را با جواب درست تا ۷ رقم اعشار ۱.۰۰۶۵۷۱×۱۰^{-۹} مقایسه کنید.

۵) در یک رایانه محاسبات طوری انجام می‌شود که خطای گرد کردن دارای کران زیر است

$$x_۱ \ op \ x_۲ \to (x_۱ \ op \ x_۲)(۱ + \varepsilon), \qquad |\varepsilon| \le ۲^{-۳۶}$$

که op نمایانگر $+$، $-$، $\times$ یا $/$ است. فرض کنید خطای نسبی محاسبه تابع $\exp(x)$ برابر $۱۰^{-۱۰}$ است.

الف) برای خطای نسبی محاسبه $\sinh x$ با $\frac{۱}{۲}(e^x - e^{-x})$، $|x| \le ۱$، کران بالایی، به صورت تابعی از x، ارائه کنید.

ب) برای مقادیر کوچک x، محاسبه $\sinh x$ به صورت $x + \frac{x^۳}{۶}$ در مقایسه با روش (الف) دارای خطای نسبی کمتری است. برای این خطا کران بالایی به صورت تابعی از x ارائه کنید.

ج) برای دو روش $x_۰$ را طوری محاسبه کنید که هر دو دارای خطای یکسان باشند. مقدار این کران را محاسبه کنید.

جوابها:

الف) $|R/\sinh x| \le v \coth x + u$ که $u = ۲^{-۳۶}$ و $v = ۱۰^{-۱۰}$

ب) $|R/\sinh x| \le (u(x + ۲x^۳/۳) + x^۵/۱۲۰)/\sinh x$

ج) $x_۰ = ۰.۰۲۶$ که نتیجه می‌دهد $|R/\sinh x| \lesssim ۰.۳۸ \times ۱۰^{-۸}$.

۶) محاسبه تابع $f(x) = e^x - ۱ - ۰.۹۹x$ را برای مجموعه‌ای بزرگ از مقادیر $x > ۰$ در نظرمی‌گیریم. به دلیل خنثی سازی برای مقادیر کوچک x باید از سری تیلور استفاده شود. برای $x > ۰$ و $n \ge ۲$ رابطه $g_۰(x) < f(x) < g_۱(x)$ برقرار است، که در آن

$$g_\theta(x) = ۰.۰۱x + \frac{x^۲}{۲!} + \ldots + \frac{x^{n-۱}}{(n-۱)!} + \frac{x^n}{n!}e^{\theta x}$$

برنامه‌ای بنویسید که به صورت زیر محدوده مقادیر $x \le x_m$ را تعیین کند که بسط سری باید مورد استفاده قرار گیرد(مقادیر محاسبه شده با $*$ مشخص شده‌اند):

(i) مقادیر $g_۰^*(x_i)$، $f^*(x_i)$ و $g_۱^*(x_i)$ را برای $x_i = ۱۰^{-۱۰+۰.۱i}$، $i = ۱, ۲, ۳, \ldots$ محاسبه کنید. تعداد جملات n_i در سری را طوری انتخاب کنید که

$$g_۱^*(x_i) - g_۰^*(x_i) \le \frac{۱}{۲}۱۰^{-۷}(g_۱^*(x_i) + g_۰^*(x_i)).$$

(ii) هرگاه رابطه $g_۰^*(x_i) < f^*(x_i) < g_۱^*(x_i)$ سه مرتبه متوالی، برای مثال برای $i = m-۱, m, m+۱$، صدق کرد، آنگاه مفادیر x_m و تعداد جملات n_m مورد استفاده را چاپ کرده و خاتمه پیدا کند.

۷) سری زیر را تا ۵ رقم اعشار درست محاسبه کنید

$$\sum_{n=۱}^{\infty}(-۱)^{n-۱}\frac{۱}{n^۲ - \frac{۱}{۳۶}}$$

جواب: ۰.۸۴۹۵۵

۸) مقدار $\sum_{n=۱}^{۱۰^۶} n^{-\frac{۱}{۲}}$ را با حداکثر خطای $۱۰^{-۵}$ محاسبه کنید. تخمین خطاها را ارائه کنید.
جواب: $\sum = ۱۹۹۸.۵۴۰۱۴۴ \pm ۱۰^{-۶}$

۹) مقدار $\prod_{n=۱}^{\infty} \frac{(۲n)^{\frac{۱}{۲n}}}{(۲n+۱)^{\frac{۱}{(۲n+۱)}}}$ را تا چهار رقم اعشار محاسبه کنید.
جواب: ۱.۱۷۳۳۵۷ ± ۱۰^{-۶}
(جواب دقیق عبارتست از $\exp((\gamma - \frac{۱}{۲}\log ۲)\log ۲)$ که γ ثابت اویلر است).

۱۰) می‌دانیم

$$\lim_{n\to\infty} \frac{\sin\frac{۱}{n}}{\frac{۱}{n}} = ۱,$$

یعنی برای هر $\varepsilon > ۰$ عدد طبیعی $n_۰$ وجود دارد به طوری که به ازای هر $n \geq n_۰$

$$|\frac{\sin\frac{۱}{n}}{\frac{۱}{n}}| < \varepsilon$$

کوچکترین مقدار ممکن $n_۰$ را برای $\varepsilon = ۱۰^{-۶}$ بدست آورید.
جواب: $n_۰ = ۴۰۹$.

۱۱) برای تقریب e^x در $۰ \leq x \leq ۱$ دو جمله از بسط تیلور e^x حول $x = a$ را در نظر می‌گیریم. a را چگونه باید انتخاب کرد تا بیشترین خطا کمینه باشد؟ a را تا دو رقم اعشار درست محاسبه کنید.
جواب: $a \approx ۰.۵۴$

۱۲) با رابطه بازگشتی زیر مقدار $a_{۱۰۰}$ را تا ۲ رقم بامعنی درست محاسبه کنید

$$\begin{cases} a_n - \frac{۳}{۲}a_{n-۱} - a_{n-۲} = ۰ \\ a_۰ = ۰ \\ a_۱ = \frac{۵}{۲} \end{cases} \qquad (۴۰)$$

جواب: $a_{۱۰۰} \approx ۱.۳ \times ۱۰^{۳۰}$.

۱۳) تا ۳ رقم اعشار کوچکترین مقدار مثبت x را پیدا کنید به طوری که

$$\frac{\tan x}{x} \geq ۳.$$

جواب: $x = ۱.۳۲۴$.

۴ بررسی مراجع

برای محاسبه

$$s = a_۱ + a_۲ + \ldots + a_n$$

با جمعوندهای نامنفی به شیوه‌های مختلف می‌توان عمل کرد. در مرجع [۲] نشان داده شده است که جمع این مقادیر با ترتیب صعودی دارای خطای کمتری است و کران خطای نسبی عبارتست از

$$\frac{|s - fl(s)|}{s} \leq (\log_2 n + 1)eps,$$

که eps دقت ماشین است. با این حال می‌توان از روشهای دقیق‌تری هم استفاده کرد که دارای محاسبات بیشتر است و نیاز به حافظه بیشتری دارد.

استفاده از فرمول اویلر—ماکلورن برای جمع سریها در مرجع [۲] بحث شده است.

مراجع

[1] D . A. Macdonald, *A Note On The Summation Of Slowly Convergent Alternating Series*, BIT 36 :4 (1996), 766-774 .

[2] O. Caprani, *Roundoff Errors In Floating-Point Summation*, BIT 15 (1975), 5-9.

[3] Stoer. *Introduction to Numerical Analysis*

[4] Riesel H. *A case of numerical divergence*, springer BIT,Volume 1 No. 2 June, 1961.

[5] Kincaid D., Cheney W. *Numerical Analysis*, (Brooks/Cole, 1991).

[6] R. L. Burden and J. D. Faires. *Numerical Analysis*, (Brooks/Cole, 1997).

[۷] هوسکینگ، *نخستین گامها در آنالیز عددی*، ترجمه دکتر بابلیان و دکتر میرنیا، نشردانشگاهی.