آشنایی با آنالیز عددی

کندال ای. اتکینسن

ترجمهٔ دکتر علی دانایی

آشنایی با

آنالیز عددی

کندال ای. اتکینسن

ترجمهٔ دکتر علی دانایی

مرکز نشر دانشگاهی

An Introduction to Numerical Analysis
Second Edition
Kendall E. Atkinson
John Wiley & Sons, 1989

آشنایی با آنالیز عددی
تألیف کندال ای. اتکینسن
ترجمهٔ دکتر علی دانایی
ویراستهٔ دکتر منوچهر وصال، دکتر محمدهادی شفیعیها
طراح جلد: شکوه پیله‌فروشها
نسخه‌پرداز: زهرا دلاوری
تهیهٔ نمایه: فاطمه پیوندی
حروفچین: الهه عموحسن
ناظر چاپ: علی صادقی
مرکز نشر دانشگاهی
چاپ اول ۱۳۸۷
تعداد ۱۰۰۰
۹۵۰۰ تومان
لیتوگرافی: سعید
چاپ: دایره سفید

فهرست‌نویسی پیش از انتشار کتابخانهٔ ملی جمهوری اسلامی ایران

سرشناسه: اتکینسون، کندل ای. Atkinson, Kendall E.
عنوان و نام پدیدآور: آشنایی با آنالیز عددی / کندال ای. اتکینسن؛ ترجمهٔ علی دانایی؛ ویراستهٔ منوچهر وصال، محمدهادی شفیعیها.
مشخصات نشر: تهران: مرکز نشر دانشگاهی، ۱۳۸۷.
مشخصات ظاهری: هشت، ۷۹۲ ص.
فروست: مرکز نشر دانشگاهی؛ ۱۲۹۳. ریاضی، آمار، و رایانه؛ ۱۵۵.
شابک: 978-964-01-1293-9
وضعیت فهرست‌نویسی: فیپا
یادداشت: عنوان اصلی: An introduction to numerical analysis, 2nd ed, c1989.
یادداشت: کتابنامه.
موضوع: آنالیز عددی.
شناسه افزوده: دانایی، علی، ۱۳۱۹- ، مترجم.
شناسه افزوده: مرکز نشر دانشگاهی.
رده‌بندی کنگره: ۱۳۸۷ ۵آ۲الف/QA۲۹۷
رده‌بندی دیویی: ۵۱۸
شماره کتابشناسی ملی: ۱۱۴۲۳۹۲

بسم‌الله‌الرحمن‌الرحیم

فهرست

پیشگفتار مترجم

کتابی که در اختیار شماست به‌چند دلیل برای ترجمه برگزیده شده است.

۱ـ کتابی است جامع در موضوعات مختلف آنالیز عددی که دانشجویان رشته‌های ریاضی، علوم کامپیوتر و مهندسی می‌توانند از آن استفاده کنند.

۲ـ کتابی است که برای پژوهشگران می‌تواند راهنمای بسیار مفیدی باشد.

۳ـ در موضوعی که وارد شده ابتدا به‌طور ساده آن را بیان کرده و دانشجو یا پژوهشگر را به‌آخرین دستاوردهای آن موضوع رهبری می‌کند و این هدف را با معرفی مقالات وکتابهای سودمند که در پایان هر فصل آورده شده به نحو بسیار خوبی به انجام می‌رساند.

اساتید بزرگوار جناب آقای دکتر محمدهادی شفیعیها و جناب آقای دکتر منوچهر وصال به تناوب ویراستاری ادبی و علمی این کتاب را در مرکز نشر دانشگاهی با دقت و حوصلهٔ بسیار انجام داده‌اند و ضعف کار مترجم را برطرف نموده‌اند که از این بابت از این اساتید بزرگوار صمیمانه سپاسگزاری می‌شود و برای آنان طول عمر با برکت و توفیقات بیشتر آرزومندم. ازکارکنان مرکز نشر دانشگاهی که با صبوری و حوصله در تایپ، غلطگیری و چاپ کتاب همکاری داشته‌اند تشکر می‌کنم.

از خوانندگانی که با راهنماییهای خود می‌توانند کاستیهای احتمالی را برطرف سازند پیشاپیش سپاسگزارم

علی دانایی
عضو هیأت علمی دانشگاه شیخ‌بهائی

پیشگفتار

ساختار اصلی ویراست دوم این کتاب، با چند بخش جدید که اضافه گردیده، اساساً مانند ویراست اول است. تمام بخشها بازنویسی شده‌اند، گاهی بسیار هم جدّی این کار انجام گرفته و در بعضی مواقع مطالبی حذف شده‌اند. به‌علاوه، برخی از موضوعات جدید معرفی شده‌اند. مثلاً درونیابی مثلثاتی و FFT (بخش ۹.۶)، مشتقگیری عددی (بخش ۷.۵)، روش خطوط (بخش ۹.۶)، مسائل مقادیر مرزی (بخش ۱۱.۶)، روش گرادیان مزدوج (بخش ۹.۸) و حل دستگاههای معادلات خطی به‌روش کمترین مربعات (بخش ۷.۹). همچنین یک پیوست در مورد نرم‌افزار ریاضی اضافه شده است که بعضی بسته‌های نرم‌افزار شناخته‌تری را تشریح می‌کند. مراجع برای هر فصل به‌روز تبدیل شده است تا آخرین کتابها و آثار خواندنی پژوهشی را منعکس کند.

با عنایت به نرم‌افزار ریاضی، قویاً توصیه می‌کنم دانشجویان در حد امکان، برنامه‌های تجارتی با کیفیت بالای آنالیز عددی در یکی از کتابخانه‌های موجود را مورد استفاده قرار دهند. آنهایی که بیشتر به‌کار گرفته شده‌اند در IMSL و NAG می‌باشند که در پیوست تحت عنوان نرم‌افزار ریاضی در مورد آنها بحث شده است. در این کتاب تعدادی از بسته‌های نرم‌افزاری آنالیز عددی که برای اهداف خاص است توضیح داده شده‌اند. این نرم‌افزارها، وقتی کدهای اصلی مورد نیاز باشند، ارزشمندند، گرچه معمولاً برنامه‌های موجود در کتابخانه‌های تجارتی، برای برنامه‌سازی از مسائل، که دانشجویان با آن مواجه می‌شوند کافی‌اند. به‌علاوه کتابخانه‌های تجارتی معمولاً برای استفاده ساده‌ترند. همچنین بسیاری از بسته‌هایی که به منظور خاص تهیه شده و در این کتاب توضیح داده شده‌اند در کتابخانه‌های تجارتی مهم جای گرفته‌اند.

۱

خطا: منشأ، انتشار و تحلیل آن

موضوع آنالیز عددی تهیهٔ روشهای محاسباتی برای مطالعه و حل مسائل ریاضی است. در این کتاب روشهای عددی متداولترین مسائل ریاضی را مطالعه می‌کنیم و خطاهای موجود در این روشها را بررسی می‌نماییم. چون تقریباً تمام محاسبات امروزه با رایانه‌های رقمی انجام می‌گیرند از ضرورتهای این کار در انجام روشهای عددی نیز بحث خواهیم کرد.

مطالعهٔ خطا مطلبی اساسی در آنالیز عددی است. بیشتر روشهای عددی جوابهایی به‌دست می‌دهند که فقط تقریبی از جواب درست مورد نظرند و درک، و در صورت امکان، توانایی برآورد یا کراندارکردن خطای حاصل حائز اهمیت است. در این فصل انواع مختلف خطاهایی که ممکن است در مسئله‌ای رخ دهند بررسی خواهند شد. نمایش اعداد در رایانه‌ها و درکنار آن خطا در حساب رایانه‌یی نیز بررسی می‌شود. قضایای کلی دربارهٔ انتشار خطاها در محاسبات، با بررسی مفصل خطا در روشهای مجموعیابی داده می‌شوند. بالاخره مفاهیم پایداری و مشروط کردن مسائل و روشهای عددی معرفی و تشریح خواهند شد. بخش اول شامل مقدمات ریاضی لازم برای این کار در فصول بعدی است.

۱.۱ مقدمات ریاضی

این بخش شامل مروری است بر قضایای حسابان که در این کتاب به‌کار برده می‌شوند. ابتدا بعضی

از قضایای مقدار میانگین را بیان می‌کنیم، سپس قضیهٔ تیلر برای توابع یک و دو متغیره را مورد بحث قرار می‌دهیم. این بخش را با نمادهایی که در فصول بعد کاربرد پیدا خواهند کرد پایان می‌دهیم.

قضیهٔ ۱.۱ (مقدار میانی) گیریم $f(x)$ بر بازهٔ متناهی $a \leq x \leq b$ پیوسته باشد، m و M را به‌صورت زیر تعریف می‌کنیم

$$m = \inf_{a \leq x \leq b} f(x), \qquad M = \operatorname{Sup}_{a \leq x \leq b} f(x)$$

پس برای هر عدد ζ در بازهٔ $[m, M]$ دست کم یک نقطهٔ ξ در $[a, b]$ وجود دارد به گونه‌ای که

$$f(\xi) = \zeta$$

به‌ویژه، نقاطی مانند $\underline{x}$ و $\overline{x}$ در $[a, b]$ هستند که برای آنها

$$m = f(\underline{x}), \qquad M = f(\overline{x})$$

قضیهٔ ۲.۱ (مقدار میانگین) گیریم $f(x)$ به ازای $a \leq x \leq b$ پیوسته و به ازای $a < x < b$ مشتقپذیر باشد. در این صورت دست کم یک نقطه چون ξ در (a, b) وجود دارد که برای آن

$$f(b) - f(a) = f'(\xi)(b - a)$$

قضیهٔ ۳.۱ (مقدار میانگین انتگرال) گیریم $w(x)$ نامنفی و بر بازهٔ $[a, b]$ انتگرالپذیر باشد، و $f(x)$ در $[a, b]$ پیوسته باشد. در این صورت برای مقداری مانند $\xi \in [a, b]$ داریم

$$\int_a^b w(x) f(x) dx = f(\xi) \int_a^b w(x) dx$$

این قضایا در بیشتر کتابهای درسی مقدماتی حساب دیفرانسیل و انتگرال مورد بحث قرار گرفته‌اند و بنابراین ما از اثبات آنها صرف‌نظر می‌کنیم. بعضی از نتایج این قضایا در مسائل آخر این فصل بررسی شده‌اند.

یکی از مهمترین ابزارهای آنالیز عددی قضیهٔ تیلر و سری مربوطه به آن است که در سرتاسر این کتاب به‌کار برده شده است. این قضیه یک روش نسبتاً ساده‌ای برای تقریب توابع $f(x)$ به کمک چند جمله‌ییها، و بنابراین روشی برای محاسبهٔ $f(x)$ به‌دست می‌دهد.

قضیهٔ ۴.۱ (قضیهٔ تیلر) گیریم $f(x)$ ، $n+1$ مشتق پیوسته، $n \geq 0$، بر $[a,b]$، داشته باشد و $x, x_0 \in [a,b]$. در این صورت

$$f(x) = p_n(x) + R_{n+1}(x) \tag{۱.۱.۱}$$

$$p_n(x) = f(x_0) + \frac{(x-x_0)}{1!}f'(x_0) + \ldots + \frac{(x-x_0)^n}{n!}f^{(n)}(x_0) \tag{۲.۱.۱}$$

$$\begin{aligned} R_{n+1}(x) &= \frac{1}{n!}\int_{x_0}^{x}(x-t)^n f^{(n+1)}(t)dt \\ &= \frac{(x-x_0)^{n+1}}{(n+1)!}f^{(n+1)}(\xi) \end{aligned} \tag{۳.۱.۱}$$

که ξ مقداری است بین x_0 و x.

برهان نحوهٔ به‌دست آوردن (۱.۱.۱) در بسیاری ازکتابهای دیفرانسیل و انتگرال داده شده است. در به‌دست آوردن آن از انتگرالگیری جزء به جزء در اتحاد زیرکه به دقّت انتخاب شده، استفاده می‌شود

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(t)dt$$

از n بار تکرار این عمل، روابط (۳.۱.۱) ـ (۱.۱.۱) با صورت انتگرالی باقیماندهٔ $R_{n+1}(x)$ به‌دست می‌آیند. صورت دوم $R_{n+1}(x)$ با استفاده از قضیهٔ مقدار میانگین انتگرال، با انتخاب $w(t) = (x-t)^n$ حاصل می‌شود. ■

با استفاده از قضیهٔ تیلر، فرمولهای استانداردِ زیر را پیدا می‌کنیم

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!}e^{\xi_x} \tag{۴.۱.۱}$$

$$\begin{aligned} \cos x = 1 &- \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots + (-1)^n \frac{x^{2n}}{(2n)!} \\ &+ (-1)^{n+1}\frac{x^{2n+2}}{(2n+2)!}\cos(\xi_x) \end{aligned} \tag{۵.۱.۱}$$

$$\begin{aligned} \sin x = x &- \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^{n-1}\frac{x^{2n-1}}{(2n-1)!} \\ &+ (-1)^n \frac{x^{2n+1}}{(2n+1)!}\cos(\xi_x) \end{aligned} \tag{۶.۱.۱}$$

$$\begin{aligned} (1+x)^\alpha = 1 &+ \binom{\alpha}{1}x + \binom{\alpha}{2}x^2 + \cdots + \binom{\alpha}{n}x^n \\ &+ \binom{a}{n+1}\frac{x^{n+1}}{(1+\xi)^{n+1-\alpha}} \end{aligned} \tag{۷.۱.۱}$$

که $\binom{\alpha}{k}$ به ازای هر عدد حقیقی α، با تساوی زیر داده می‌شود

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\dots(\alpha-k+1)}{k!} \qquad k = 1, 2, 3, \dots$$

در همهٔ حالات نقطهٔ مجهول ξ_x بین ۰ و x واقع است.

یک حالت خاص مهّم (۷.۱.۱) رابطهٔ زیر است

$$\frac{1}{1-x} = 1 + x + x^2 + \dots + x^n + \frac{x^{n+1}}{1-x} \qquad x \neq 1 \qquad (8.1.1)$$

در این حالت، $\alpha = -1$، و $-x$ جایگزین x شده است. باقیمانده صورتی ساده‌تر از (۷.۱.۱) دارد؛ و این مطلب به راحتی از ضرب طرفین رابطهٔ (۸.۱.۱) در $1-x$ و سپس ساده‌کردن آن اثبات می‌شود. اگر ترتیب نوشتن رابطهٔ (۸.۱.۱) را اندکی تغییر دهیم فرمول آشنای سری هندسی متناهی زیر به‌دست می‌آید

$$1 + x + x^2 + \dots + x^n = \frac{1 - x^{n+1}}{1-x} \qquad x \neq 1 \qquad (9.1.1)$$

نمایش سریهای نامتناهی برای توابع طرف چپ روابط (۸.۱.۱) ـ (۴.۱.۱) را می‌توان با میل‌دادن $n \to \infty$ به‌دست آورد. سریهای نامتناهی (۶.۱.۱) ـ (۴.۱.۱) به‌ازای همهٔ مقادیر x همگرا هستند و سریهای (۸.۱.۱) ـ (۷.۱.۱) به ازای $|x| < 1$ همگرا هستند. از فرمول (۸.۱.۱) سری هندسی نامتناهی معروف

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k \qquad |x| < 1 \qquad (10.1.1)$$

به‌دست می‌آید.

سری تیلر برای هر تابع به قدرکافی مشتقپذیر $f(x)$ را می‌توان مستقیماً از تعریف (۲.۱.۱)، با هر تعداد جملهٔ دلخواه محاسبه کرد، ولی به دلیل پیچیدگی مشتقگیریِ در بسیاری از توابع $f(x)$، اغلب بهتر است که تقریبهای چندجمله‌ییهای تیلر $p_n(x)$ یا سری تیلر آنها را با استفاده از فرمولهای قبلی (۸.۱.۱) ـ (۴.۱.۱) به‌دست آورد. ما سه مثال می‌آوریم که در آنها جملات خطا ساده‌تر از وقتی هستند که (۳.۱.۱) مستقیماً به‌کار برده می‌شود.

مثال ۱. گیریم $f(x) = e^{-x^2}$. در رابطهٔ (۴.۱.۱) به جای x مقدار $-x^2$ را قرار می‌دهیم تا به‌دست آوریم

$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2!} - \dots + (-1)^n \frac{x^{2n}}{n!} + (-1)^{n+1} \frac{x^{2x+2}}{(n+1)!} e^{\xi_x}$$

که در آن $-x^۲ \le \xi_x \le ۰$.

۲. گیریم $f(x) = \tan^{-۱}(x)$. با قراردادن $x = -u^۲$ در (۸.۱.۱) خواهیم داشت

$$\frac{۱}{۱+u^۲} = ۱ - u^۲ + u^۴ - \cdots + (-۱)^n u^{۲n} + (-۱)^{n+۱}\frac{u^{۲n+۲}}{۱+u^۲}$$

و پس از انتگرالگیری بر بازهٔ $[۰, x]$ داریم

$$\tan^{-۱}(x) = x - \frac{x^۳}{۳} + \frac{x^۵}{۵} - \cdots + (-۱)^n \frac{x^{۲n+۱}}{۲n+۱} + (-۱)^{n+۱}\int_۰^x \frac{u^{۲n+۲}}{۱+u^۲}du \qquad (۱۱.۱.۱)$$

با استفاده از قضیهٔ مقدار میانگین در انتگرالها

$$\int_۰^x \frac{u^{۲n+۲}}{۱+u^۲}du = \frac{x^{۲n+۳}}{۲n+۳}\cdot\frac{۱}{۱+\xi_x^۲}$$

که در آن ξ_x بین ۰ و x واقع است.

۳. گیریم $f(x) = \int_۰^۱ \sin(xt)dt$. با استفاده از (۶.۱.۱) داریم

$$\begin{aligned} f(x) &= \int_۰^۱ \left[xt - \frac{x^۳t^۳}{۳!} + \cdots + (-۱)^{n-۱}\frac{(xt)^{۲n-۱}}{(۲n-۱)!}\right. \\ &\qquad \left. +(-۱)^n\frac{(xt)^{۲n+۱}}{(۲n+۱)!}\cos(\xi_{xt})\right]dt \\ &= \sum_{j=۱}^{n}(-۱)^{j-۱}\frac{x^{۲j-۱}}{(۲j)!} + (-۱)^n\frac{x^{۲n+۱}}{(۲n+۱)!}\int_۰^۱ t^{۲n+۱}.\cos(\xi_{xt})dt \end{aligned}$$

که در آن ξ_{xt} بین ۰ و xt واقع است. انتگرال در باقیمانده به آسانی به وسیلهٔ $۱/(۲n+۲)$ کراندار می‌شود؛ ولی می‌توانیم آن را به شکل ساده‌تری برگردانیم. اگرچه اثبات نکرده‌ایم، ولی می‌توان نشان داد که $\cos(\xi_{xt})$ تابعی است پیوسته از t . پس با استفاده از قضیهٔ مقدار میانگین انتگرال داریم

$$\int_۰^۱ \sin(xt)dt = \sum_{j=۱}^{n}(-۱)^{j-۱}\frac{x^{۲j-۱}}{(۲j)!} + (-۱)^n\frac{x^{۲n+۱}}{(۲n+۲)!}\cos(\zeta_x)$$

که ζ_x بین ۰ و x قرار دارد.

قضیهٔ تیلر در فضای دوبعدی گیریم $f(x,y)$ تابع داده شده‌ای از دو متغیر مستقل x و y باشد. نشان خواهیم داد که چگونه قضیهٔ قبلی تیلر برای بسط $f(x,y)$ حول نقطهٔ داده‌شدهٔ $(x_۰,y_۰)$ تعمیم داده می‌شود. این نتایج به‌سادگی برای توابع بیش از دو متغیر نیز قابل تعمیم‌اند. گیریم نماد $L(x_۰,y_۰;x_۱,y_۱)$ معرف مجموعهٔ همهٔ نقاط واقع بر پاره خط واصل بین نقاط $(x_۰,y_۰)$ و $(x_۱,y_۱)$ باشد.

قضیهٔ ۵.۱ گیریم $(x_۰,y_۰)$ و $(x_۰+\xi,y_۰+\eta)$ نقاط داده‌شده‌ای باشند، و فرض می‌کنیم $f(x,y)$ برای همهٔ نقاط (x,y) در یک همسایگی $L(x_۰,y_۰;x_۰+\xi,y_۰+\eta)$ پیوسته و $n+۱$ بار مشتقپذیر باشد. در این صورت

$$\begin{aligned} f(x_۰+\xi,y_۰+\eta) &= f(x_۰,y_۰)+\sum_{j=۱}^{n}\frac{۱}{j!}\left[\xi\frac{\partial}{\partial x}+\eta\frac{\partial}{\partial y}\right]^j f(x,y)\Big|_{\substack{x=x_۰\\y=y_۰}} \\ &\quad+\frac{۱}{(n+۱)!}\left[\xi\frac{\partial}{\partial x}+\eta\frac{\partial}{\partial y}\right]^{n+۱} f(x,y)\Big|_{\substack{x=x_۰+\theta\xi\\y=y_۰+\theta\eta}} \end{aligned} \qquad (۱۲.۱.۱)$$

برای مقداری از $۰\le\theta\le۱$. نقطهٔ $(x_۰+\theta\xi,y_۰+\theta\eta)$ نقطه‌ای بر خط $L(x_۰,y_۰;x_۰+\xi,y_۰+\eta)$ است.

برهان ابتدا به معنای نماد مشتق در (۱۲.۱.۱) توجه نمایید. به‌عنوان مثال

$$\left[\xi\frac{\partial}{\partial x}+\eta\frac{\partial}{\partial y}\right]^۲ f(x,y)=\xi^۲\cdot\frac{\partial^۲f(x,y)}{\partial x^۲}+۲\xi\eta\cdot\frac{\partial^۲f(x,y)}{\partial x\partial y}+\eta^۲\cdot\frac{\partial^۲f(x,y)}{\partial y^۲}$$

معنی زیر نمایه‌های $x=x_۰$ و $y=y_۰$ آن است که مشتقات مختلف بایستی در $(x_۰,y_۰)$ محاسبه شوند.

برهان (۱۲.۱.۱) بر پایهٔ استفاده از قضیهٔ اولیهٔ تیلر برای

$$F(t)=f(x_۰+t\xi,y_۰+t\eta)\qquad ۰\le t\le۱$$

استوار است.

با استفاده از قضیهٔ ۴.۱

$$F(۱)=F(۰)+\frac{F'(۰)}{۱!}+\cdots+\frac{F^{(n)}(۰)}{n!}+\frac{F^{(n+۱)}(\theta)}{(n+۱)!}$$

به ازای $0 \leq \theta \leq 1$. روشن است که، $F(0) = f(x_0, y_0)$ و $F(1) = f(x_0 + \xi, y_0 + \eta)$.
برای مشتق اوّل داریم

$$F'(t) = \xi \frac{\partial f(x_0 + t\xi, y_0 + t\eta)}{\partial x} + \eta \frac{\partial f(x_0 + t\xi, y_0 + t\eta)}{\partial y}$$
$$= \left[\xi \frac{\partial}{\partial x} + \eta \frac{\partial}{\partial y}\right] f(x,y)\Big|_{\substack{x=x_0+t\xi \\ y=y_0+t\eta}}$$

مشتقات مراتب بالاتر به طریق مشابه محاسبه می‌شوند. ■

مثال به‌عنوان یک مثال ساده، بسط تابع $f(x,y) = x/y$ را حول نقطهٔ $(x_0, y_0) = (6, 2)$ در نظر می‌گیریم. گیریم $n = 1$. پس

$$\frac{x}{y} = f(6,2) + (x-6)\frac{\partial f(6,2)}{\partial x} + (y-2)\frac{\partial f(6,2)}{\partial y}$$
$$+ \frac{1}{2}\left[(x-6)^2 \cdot \frac{\partial^2 f(\partial,\gamma)}{\partial x^2} + 2(x-6)(y-2)\frac{\partial^2 f(\partial,\gamma)}{\partial x \partial y}\right.$$
$$\left. + (y-2)^2 \cdot \frac{\partial^2 f(\partial,\gamma)}{\partial y^2}\right]$$
$$= 3 + \frac{1}{2}(x-6) - \frac{3}{2}(y-2)$$
$$+ \frac{1}{2}\left[(x-6)^2 \cdot 0 - 2(x-6)(y-2)\frac{1}{\gamma^2} + (y-2)^2 \cdot \frac{2\delta}{\gamma^3}\right]$$

که در آن (δ, γ) نقطه‌ای است واقع بر خط $L(6, 2; x, y)$. برای (x, y) نزدیک به $(6, 2)$

$$\frac{x}{y} \doteq 3 + \frac{1}{2}x - \frac{3}{2}y$$

نمودار $z = 3 + \frac{1}{2}x - \frac{3}{2}y$ صفحهٔ مماس بر نمودار $z = x/y$، در نقطهٔ $(x, y, z) = (6, 2, 3)$ است.

بعضی نمادهای ریاضی از مفاهیم چندی در این کتاب استفاده شده است که شکلهای ساده‌تر آنها در فصلهای اولیه مورد نیازند. این مفاهیم شامل قضایایی دربارهٔ تفاضلهای منقسم توابع، فضاهای برداری و نرمهای بردار و ماتریس هستند. حداقل نمادهای لازم اکنون معرفی می‌شوند و بسط کامل آنها به جاهای مناسبتر کتاب موکول خواهد شد.

برای یک تابع داده‌شدهٔ $f(x)$ با فرض متمایزبودن x_0, x_1, x_2 تعاریف زیر را اختیار می‌کنیم

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \qquad f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} \qquad (۱۳.۱.۱)$$

این نمادها را به‌ترتیب تفاضلهای منقسم مرتبهٔ اوّل و مرتبهٔ دوّم $f(x)$ می‌نامند. روابط بین این نمادها و مشتقات $f(x)$ چنین‌اند

$$f[x_0, x_1] = f'(\xi) \qquad f[x_0, x_1, x_2] = \frac{1}{2} f''(\zeta) \qquad (۱۴.۱.۱)$$

که ξ بین x_0, x_1 و ζ بین مینیمم و ماکسیمم x_0, x_1, x_2 واقع‌اند. تفاضلهای منقسم، برخلاف آنچه که ممکن است از رابطهٔ (۱۳.۱.۱) انتظار داشت، مستقل از ترتیب شناسه‌های خود هستند. به عبارت دقیقتر

$$f[x_0, x_1] = f[x_1, x_0]$$
$$f[x_0, x_1, x_2] = f[x_i, x_j, x_k] \qquad (۱۵.۱.۱)$$

برای هر جایگشت (i, j, k) از اعداد $(۰، ۱، ۲)$. اثبات این ویژگیها و سایر ویژگیها به‌صورت مسائل به‌خواننده واگذار می‌شوند. شرح کامل تفاضلهای منقسم در بخش ۲.۳ از فصل ۳ داده شده است.

موضوعهای فضاهای برداری، ماتریسها و نرمهای بردار و ماتریس در فصل ۷، بلافاصله پس از فصلهای راجع به جبر خطی عددی مطرح شده‌اند. برخی از این مطالب را در اینجا ذکر می‌کنیم و اثبات آنها را برای فصل ۷ می‌گذاریم. دو فضای برداری هستند که کاربردهای بسیار دارند. این فضاها عبارت‌اند از:

$$\mathbf{R}^n = \left\{ x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \;\middle|\; x_n, \ldots, x_2, x_1 \text{ اعداد حقیقی} \right\}$$
$$C[a, b] = \{ f(t) \mid f(t) \text{، مقدار حقیقی و پیوسته برای } a \le t \le b \}$$

برای $x, y \in \mathbf{R}^n$ و عدد حقیقی α، عبارت $x + y$ و αx چنین تعریف می‌شوند

$$x + y = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix} \qquad \alpha x = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}$$

برای $f, g \in C[a,b]$ و عدد حقیقی α عبارت $f+g$ و αf چنین تعریف می‌شوند

$$(f+g)(t) = f(t) + g(t) \qquad (\alpha f)(t) = \alpha f(t) \qquad a \le t \le b$$

نرمهای بردارها برای اندازه‌گیری قدر بردار به‌کار برده می‌شوند. برای $\mathbf{R}^n$ مقدمتاً دو نرم مختلف را تعریف می‌کنیم

$$\| x \|_\infty = \max_{1 \le i \le n} | x_i | \qquad x \in \mathbf{R}^n \qquad (۱۶.۱.۱)$$

$$\| x \|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \qquad x \in \mathbf{R}^n \qquad (۱۷.۱.۱)$$

و برای $C[a,b]$ نرم را چنین تعریف می‌کنیم

$$\| f \|_\infty = \max_{a \le t \le b} | f(t) | \qquad f \in C[a,b] \qquad (۱۸.۱.۱)$$

می‌توان نشان داد که تعاریف بالا در سه ویژگی خاص همهٔ نرمها صدق می‌کنند.

۱. $\| v \| \ge 0$؛ $\| v \| = 0$ اگر و فقط اگر $v = 0$، یعنی v بردار صفر باشد.

۲. $\| \alpha v \| = | \alpha | \cdot \| v \|$، به ازای همهٔ بردارهای v و اعداد حقیقی α.

۳. $\| v + w \| \le \| v \| + \| w \|$ ، به ازای همهٔ بردارهای v و w.

رابطهٔ (۳) را معمولاً **نابرابری مثلث** می‌نامند. توضیح این نامگذاری و همچنین تعمیم بیشتر خواص نرمها برای $C[a,b]$ در فصل ۴ داده شده است.

نرمها را می‌توان برای ماتریسها نیز تعریف نمود. برای هر ماتریس $n \times n$

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

نرم را چنین تعریف می‌کنیم

$$\| A \|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} | a_{ij} | \qquad (۱۹.۱.۱)$$

این تعریف، در ویژگیهای نرم بردار صدق می‌کند. علاوه بر آن می‌توان نشان داد که

$$\| AB \|_\infty \leq \| A \|_\infty \| B \|_\infty \qquad (۲۰.۱.۱)$$

$$\| Ax \|_\infty \leq \| A \|_\infty \| x \|_\infty \quad \text{برای هر } x \in \mathbf{R}^n \qquad (۲۱.۱.۱)$$

که A و B ماتریسهای $n \times n$ دلخواه‌اند. اثبات این قضایا به‌عنوان مسأله واگذار می‌شود. این روابط به‌طور کلیتر در فصل ۷ نیز مورد بحث واقع شده‌اند.

مثال فضای برداری $\mathbf{R}^۲$ و ماتریسهای ۲×۲ را در نظر می‌گیریم. در حالت خاص فرض می‌کنیم

$$A = \begin{bmatrix} ۱ & -۱ \\ ۳ & ۲ \end{bmatrix} \qquad x = \begin{bmatrix} ۱ \\ ۲ \end{bmatrix} \qquad y = Ax = \begin{bmatrix} -۱ \\ ۷ \end{bmatrix}$$

پس

$$\| A \|_\infty = ۵ \qquad \| x \|_\infty = ۲ \qquad \| y \|_\infty = ۷$$

و برقراری (۲۱.۱.۱) به‌سادگی دیده می‌شود. برای آنکه نشان دهیم نابرابری (۲۱.۱.۱) را نمی‌توان بهتر کرد، بردار $x = [\,۱\;۱\,]$ را در نظر می‌گیریم. داریم

$$x = \begin{bmatrix} ۱ \\ ۱ \end{bmatrix} \qquad y = Ax = \begin{bmatrix} ۰ \\ ۵ \end{bmatrix}$$

پس

$$\| y \|_\infty = ۵ = \| A \|_\infty \| x \|_\infty$$

۲.۱ نمایش رایانه‌یی اعداد

رایانه‌های رقمی ابزار اصلی محاسبات در آنالیز عددی هستند و نتیجتاً درک شیوهٔ کار آنها بسیارمهم است. در این بخش چگونگی نمایش اعداد را در رایانه‌ها ملاحظه می‌کنیم و در سایر بخشها نتایج این گونه نمایش اعداد و حساب با رایانه را خواهیم دید.

بیشتر رایانه‌ها یک وجه اعداد صحیح و یک وجه با ممیز شناور برای نمایش اعداد دارند. وجه اعداد صحیح فقط برای نمایش اعداد صحیح به کار برده می‌شود و چندان مورد علاقهٔ ما نیست. وجه با ممیّز شناور در نمایش اعداد حقیقی مورد استفاده قرار می‌گیرد. اعداد مجاز می‌توانند

دارای اندازه‌های بسیار متفاوت باشند، ولی محدودیتهایی چه از نظر اندازه و چه از لحاظ تعداد ارقام برای آنها وجود دارد. نمایش با ممیّز شناور بستگی نزدیکی به آنچه که در بسیاری از کتابهای ریاضی دبیرستانی، نماد علمی خوانده می‌شود، دارد.

پایهٔ اعدادی که در رایانه‌ها استفاده می‌شوند به‌ندرت اعشاری است. بیشتر رایانه‌های رقمی پایهٔ دستگاه عددی ۲ (دودویی) یا شکلهای مختلفی از آن، چون پایهٔ ۸ (هشت‌هشتی) یا پایهٔ ۱۶ (شانزده شانزدهی) را به‌کار می‌برند.

مثال (الف) در پایهٔ ۲، ارقام ۰ و ۱ هستند. به‌عنوان مثال برای تبدیل یک عدد در پایهٔ ۲ به اعشاری داریم

$$(۱۱۰۱۱٫۰۱)_۲ = ۱\times۲^۴+۱\times۲^۳+۰\times۲^۲+۱\times۲^۱+۱\times۲^۰+۰\times۲^{-۱}+۱\times۲^{-۲}$$
$$= ۲۷٫۲۵$$

هنگامی که از اعداد در پایهٔ دیگری استفاده می‌کنیم معمولاً نماد $(x)_\beta$ را به‌کار می‌بریم تا نشان دهیم عدد x در دستگاه اعداد به پایهٔ β بیان شده است.

(ب) در پایهٔ ۱۶، ارقام عبارت‌اند از: ۰، ۱، ۲، ...، ۹، A، ...، F. مثلاً برای تبدیل چنین عددی به اعشاری داریم

$$(۵۶C.F)_{۱۶} = ۵\times۱۶^۲+۶\times۱۶^۱+۱۲\times۱۶^۰+۱۵\times۱۶^{-۱} = ۱۳۸۸٫۹۳۷۵$$

تبدیل اعداد اعشاری به دودویی درضمن مسائل بررسی شده است.

گیریم β عدد پایه‌ای باشد که در رایانه به‌کار گرفته شده است. پس یک عدد غیرصفر x در رایانه اساساً به شکل زیر ذخیره می‌شود

$$x = \sigma\cdot(۰٫a_۱a_۲\dots a_t)_\beta\cdot\beta^e \qquad (۱.۲.۱)$$

که در آن $\sigma = +۱$ یا -۱ و $۰\le a_i\le\beta-۱$، e یک عدد صحیح و

$$(۰٫a_۱a_۲\dots a_t)_\beta = \frac{a_۱}{\beta^۱}+\frac{a_۲}{\beta^۲}+\dots+\frac{a_t}{\beta^t}$$

حرف σ را **علامت**، e را نما و $(۰٫a_۱a_۲\dots a_t)_\beta$ را **جزء کسری** عدد با ممیّز شناور x خوانند. عدد β را همچنین پایهٔ دستگاه عددی و نقطهٔ قبل از $a_۱$ در (۱.۲.۱) را **ممیّز پایه** می‌نامند، برای

مثال ممیّز اعشاری ($\beta = ۱۰$)، ممیز دو دویی ($\beta = ۲$). عدد صحیح t تعداد ارقام در پایهٔ βرا در نمایش فوق معیّن می‌کند. همیشه فرض می‌کنیم

$$a_۱ \neq ۰$$

که نمایش با ممیّز شناورِ، به‌اصطلاح، نرمال شده را می‌دهد. همچنین فرض می‌کنیم که

$$L \leq e \leq U \qquad (۲.۲.۱)$$

که اندازهٔ ممکن x را محدود می‌نماید. عدد $x = ۰$ همیشه مجاز و مستلزم نمایش ویژه‌ای است. جدول ۱.۱ مقادیر β، t، L و U را برای شماری از رایانه‌های معمولی به‌دست می‌دهد. اساس استفاده از β، t، L و U برای معین‌کردن مشخصات حسابی از قراردادهای کتاب [فورسایت[1] و همکاران، ۱۹۷۷، ص ۱۱] گرفته شده است. بعضی رایانه‌ها (مثل، CDC CYBER) جایگذاریهایی برای ممیّز پایه به‌کار می‌برند. ما کرانهای نما را به گونه‌ای تغییر داده‌ایم که حدود برای اندازهٔ عدد با ممیّز شناور، با توجه به نظریهٔ این بخش، درست از آب درآید. نتایج نمایشهای با دقّت مضاعف را در رایانه‌هایی که در سخت‌افزار آنها این درجهٔ دقّت گنجانده شده است نیز اضافه نموده‌ایم. در جدول ۱.۱ ستونهای بیشتری هستند که بعداً توضیح داده می‌شوند.

قطع‌کردن و گردکردن بیشتر اعداد حقیقی x را نمی‌توان دقیقاً به وسیلهٔ نمایش با ممیّز شناور، که قبلاً گفته شد، نشان داد، و بنابراین، در صورت امکان، بایستی با عددی نزدیک که قابل نمایش در رایانه است تقریب زده شوند. برای یک عدد دلخواه داده‌شدهٔ x، $\mathrm{fl}(x)$ را معرّف رایانه‌یی آن، در صورت وجود، می‌گیریم. دو راه اساسی برای به‌دست آوردن $\mathrm{fl}(x)$ از x، وجود دارد: قطع‌کردن و گردکردن.

گیریم عدد حقیقی x به شکل زیر نوشته شده باشد

$$x = \sigma \cdot (۰\text{ر}a_۱a_۲ \ldots a_t a_{t+۱} \ldots)_\beta \cdot \beta^e \qquad (۳.۲.۱)$$

که $a_۱ \neq ۰$، و e در (۲.۲.۱) صدق می‌کند. نمایش قطع‌شدهٔ رایانه‌یی x، با عبارت زیر داده می‌شود

$$\mathrm{fl}(x) = \sigma \cdot (۰\text{ر}a_۱ \ldots a_t)_\beta \cdot \beta^e \qquad (۴.۲.۱)$$

1. Forsythe

جدول ۱.۱ نمایشهای با ممیز شناور در رایانه‌های گوناگون

ماشین	S/D	R/C	β	t	L	U	δ	M
CDC CYBER 170	S	R	۲	۴۸	−۹۷۶	۱۰۷۱	۳٫۵۵E − ۱۵	۲٫۸۱E۱۴
CDC CYBER 205	S	C	۲	۴۷	−۲۸،۶۲۶	۲۸،۷۱۸	۱٫۴۲E − ۱۴	۱٫۴۱E۱۴
CRAY-1	S	C	۲	۴۸	−۸۱۹۲	۸۱۹۱	۷٫۱۱E − ۱۵	۲٫۸۱E۱۴
DEC VAX	S	R	۲	۲۴	−۱۲۷	۱۲۷	۵٫۹۶E − ۸	۱٫۶۸E۷
DEC VAX	D	R	۲	۵۳	−۱۰۲۳	۱۰۲۳	۱٫۱۱E − ۱۶	۹٫۰۱E۱۵
HP-11C,15C	S	R	۱۰	۱۰	−۹۹	۹۹	۵٫۰۰E − ۱۰	۱٫۰۰E۱۰
IBM 3033	S	C	۱۶	۶	−۶۴	۶۳	۹٫۵۴E − ۷	۱٫۶۸E۷
IBM 3033	D	C	۱۶	۱۴	−۶۴	۶۳	۲٫۲۲E − ۱۶	۷٫۲۱E۱۶
Intel 8087	S	R	۲	۲۴	−۱۲۶	۱۲۷	۵٫۹۶E − ۸	۱٫۶۸E۷
Intel 8087	D	R	۲	۵۳	−۱۰۲۲	۱۰۲۳	۱٫۱۱E − ۱۶	۹٫۰۱E۱۵
PRIME 850	S	R	۲	۲۳	−۱۲۸	۱۲۷	۱٫۱۹E − ۷	۸٫۳۹E۶
PRIME 850	S	C	۲	۲۳	−۱۲۸	۱۲۷	۱٫۱۹E − ۷	۸٫۳۹E۶
PRIME 850	D	C	۲	۴۷	−۳۲،۸۹۶	۳۲،۶۳۹	۱٫۴۲E − ۱۴	۱٫۴۱E۱۴

به این نمادها توجه کنید:
S/D: دقّت معمولی یا مضاعف.
R/C: گردکردن یا قطع‌کردن.
β: عدد پایه.
t: رقمهای بعد از ممیّز [(۱.۲.۱)] را ببینید
L، U: حدود نما [(۲.۲.۱) را ببینید]
δ: واحد گردکردن [(۱۲.۲.۱) را ببینید]
M: کران دقیق اعداد صحیح [(۱۶.۲.۱) راببینید]

دلیل معرّفی قطع‌کردن آن است که در بسیاری از رایانه‌ها به جای گردکردن در هر عمل حساب، از قطع‌کردن استفاده می‌کنند.

نمایش گردشدهٔ x به‌صورت زیر است:

$$\mathrm{fl}(x)=\begin{cases}\sigma\cdot(\circ{,}a_1\dots a_t)_\beta.\beta^e & \circ\le a_{t+1}<\dfrac{\beta}{2} \quad (۵.۲.۱)\\ \sigma\cdot[(\circ{,}a_1\dots a_t)_\beta+(\circ{,}\circ\dots\circ\,1)_\beta].\beta^e & \dfrac{\beta}{2}\le a_{t+1}<\beta \quad (۶.۲.۱)\end{cases}$$

در فرمول اخیر، $(\circ{,}\circ\dots\circ\,1)_\beta$ معرّف β^{-t} است. اگرچه این تعریف $\mathrm{fl}(x)$ تا حدّی پیچیده است ولی همان تعریف استانداردِ گردکردن را، که اغلب از دستگاه اعشاری آموخته‌ایم، به‌دست می‌دهد. گاهی این تعریف به گونه‌ای دیگر به‌کار برده می‌شود تا گردکردن ناارریب به‌دست آید. در چنین حالتی، اگر

$$a_{t+1}=\frac{\beta}{2} \quad (۱) \qquad \text{و} \qquad a_j=\circ \quad (۲) \quad \text{برای } j\ge t+2$$

آنگاه گردکردن را اضافی یا نقصانی می‌گیرند بسته به اینکه به ترتیب، a_t فرد یا زوج باشد. این عمل به قاعدهٔ گردکردن نااریب می‌انجامد که بیشتر افراد برای گردکردن اعداد اعشاری آموخته‌اند، ولی ما از این پس تعریف ساده‌تر (۵.۲.۱) ـ (۶.۲.۱) را می‌پذیریم.

برای بیشتر اعداد حقیقی x، داریم $\mathrm{fl}(x) \neq x$. با نگاه کردن به خطای نسبی (یا درصد)، می‌توان نشان داد که

$$\frac{x - \mathrm{fl}(x)}{x} = -\varepsilon \qquad (۷.۲.۱)$$

$$-\beta^{-t+۱} \leq \varepsilon \leq ۰ \qquad \mathrm{fl}(x) \text{ قطع‌شدهٔ} \qquad (۸.۲.۱)$$

$$-\frac{۱}{۲}\beta^{-t+۱} \leq \varepsilon \leq \frac{۱}{۲}\beta^{-t+۱} \qquad \mathrm{fl}(x) \text{ گردشدهٔ} \qquad (۹.۲.۱)$$

ما نتیجهٔ (۸.۲.۱) را برای قطع‌کردن اثبات می‌کنیم و اثبات نتیجهٔ (۹.۲.۱) را برای گردکردن به‌عنوان یک مسئله واگذار می‌کنیم.

گیریم $\sigma = +۱$، زیرا حالت $\sigma = -۱$ علامت ε را تغییر نخواهد داد. از (۳.۲.۱) و (۴.۲.۱) به‌دست می‌آوریم

$$x - \mathrm{fl}(x) = (۰\text{ر}۰۰\ldots۰a_{t+۱}\cdots)_\beta.\beta^e$$

فرض می‌کنیم $\gamma = \beta - ۱$

$$\begin{aligned}
۰ \leq x - \mathrm{fl}(x) &\leq (۰\text{ر}۰۰\ldots۰\gamma\gamma\ldots)_\beta \cdot \beta^e \\
&= \gamma[\beta^{-t-۱} + \beta^{-t-۲} + \ldots] \cdot \beta^e \\
&= \gamma\left[\frac{\beta^{-t-۱}}{۱-\beta^{-۱}}\right] \cdot \beta^e = \beta^{-t+e} \\
۰ \leq \frac{x - \mathrm{fl}(x)}{x} &\leq \frac{\beta^{-t+e}}{(۰\text{ر}a_۱a_۲\ldots)_\beta \cdot \beta^e} \\
&\leq \frac{\beta^{-t}}{(۰\text{ر}۱۰۰\ldots)_\beta} = \beta^{-t+۱}
\end{aligned}$$

که حکم (۸.۲.۱) را ثابت می‌کند و اثبات (۹.۲.۱) به طریق مشابه‌صورت می‌گیرد.

فرمول (۷.۲.۱) معمولاً به شکل هم‌ارز آن

$$\mathrm{fl}(x) = (۱ + \varepsilon)x \qquad (۱۰.۲.۱)$$

نوشته می‌شود که ε توسط (۸.۲.۱) یا (۹.۲.۱) داده شده است. بنابراین $\mathrm{fl}(x)$ را می‌توان به‌عنوان یک اختلال کوچک نسبی x منظور کرد. این فرمول $\mathrm{fl}(x)$ ضمناً به ما امکان می‌دهد که دقیقاً نتایج خطاهای گردکردن/قطع‌کردن در عملیّات حسابِ رایانه را بررسی کنیم. مثالهایی از این دست در بخشهای بعد آورده شده است. تعریف $\mathrm{fl}(x)$ و استفاده از (۱۰.۲.۱) اقتباس از [ویلکینسن،[1] ۱۹۶۳] است، و در بسیاری از جاها در تحلیل نتایج خطای گردکردن مورد استفاده واقع شده است.

دقّت نمایش با ممیّز شناور ما اکنون دو اندازه که تصور نسبتاً روشنی از دقّت ممکن در نمایش با ممیّز شناور را به‌دست می‌دهند، معرفی می‌کنیم. اندازهٔ اوّل ارتباط نزدیکی با نتیجهٔ خطای قبلی $\mathrm{fl}(x)$ که با (۷.۲.۱) تا (۹.۲.۱) داده شده، دارد.

واحد گردکردن یک رایانه عددی است مانند δ که در شرایط زیر صدق می‌کند: (۱) δ عدد مثبتی است با ممیّز شناور، و (۲) کوچکترین عددی است که برای آن نامساوی زیر برقرار است.

$$\mathrm{fl}(1+\delta) > 1 \qquad (11.2.1)$$

بنابراین برای هر عدد $\hat{\delta} < \delta$، با ممیّز شناور داریم $\mathrm{fl}(1+\hat{\delta}) = 1$ که $1+\hat{\delta}$ و ۱ در حساب رایانه یکی هستند. این رابطه یک اندازهٔ دقیقی به ما می‌دهد تا بدانیم چند رقم دقیق ممکن است در نمایش یک عدد وجود داشته باشد. در بسیاری از برنامه‌های رایانه‌های قابل حمل و با کیفیّت بالا از این واحد گردکردن استفاده می‌کنند تا بیشترین دقتی را که ممکن است در رایانه به کار برد به‌دست آورند.

واحد گردکردن δ به‌سادگی محاسبه و به‌صورت زیر بیان می‌شود

$$\delta = \begin{cases} \beta^{-t+1} & \text{طبق تعریف قطع‌کردن برای } \mathrm{fl}(x) \\ \frac{1}{2}\beta^{-t+1} & \text{طبق تعریف گردکردن برای } \mathrm{fl}(x) \end{cases} \qquad (12.2.1)$$

ما این مطلب را برای حساب گردشده روی یک ماشین دودویی اثبات می‌کنیم. ابتدا باید نشان دهیم

$$\mathrm{fl}(1+2^{-t}) > 1 \qquad (13.2.1)$$

می‌نویسیم

$$1+2^{-t} = [(0.100\ldots)_2 + (0.00\ldots01\ldots)_2]\cdot 2^1$$

$$= (0.10\ldots010\ldots)_2 \cdot 2^1 \qquad (14.2.1)$$

در موضع $t+1$ ام

1. Wilkinson

صورت $\mathrm{fl}(۱ + ۲^{-t})$ را با توجه به اینکه در موضع $t+۱$ امِ ارقام بعد از ممیّز یک عدد ۱ وجود دارد تشکیل می‌دهیم. پس با توجه به (۶.۲.۱) داریم

$$\mathrm{fl}(۱ + ۲^{-t}) = (۰\text{ر}۱۰\ldots۰۱)_۲ \cdot ۲^{۱} = ۱ + ۲^{-t+۱}$$

در موضع t ↑

بنابراین (۱۳.۲.۱) برقرار است، اگرچه

$$\mathrm{fl}(۱ + ۲^{-t}) \neq ۱ + ۲^{-t}$$

این واقعیّت که δ نمی‌تواند از ۲^{-t} کوچکتر باشد به سادگی از بازنگری (۱۴.۲.۱) حاصل می‌شود. اگر $\hat{\delta} < \delta$، آنگاه $۱ + \hat{\delta}$ در موضع $t+۱$ ام ارقام بعد از ممیّز در (۱۴.۲.۱) یک صفر خواهد داشت: و در این حال تعریف (۵.۲.۱) برای گردکردن ایجاب می‌کند که $\mathrm{fl}(۱ + \hat{\delta}) = ۱$.

یک اندازهٔ دوّم برای بیشترین دقّت ممکن در یک نمایش با ممیّز شناور، پیداکردن بزرگترین عدد صحیح M است به طوری که برای یک عدد صحیح m

$$۰ \leq m \leq M \Rightarrow \mathrm{fl}(m) = m \qquad (۱۵.۲.۱)$$

این رابطه هم ایجاب می‌کند که $\mathrm{fl}(M + ۱) \neq M + ۱$. اثبات تساوی زیر به‌عنوان یک مسئله واگذار می‌شود

$$M = \beta^t \qquad (۱۶.۲.۱)$$

اعداد M و δ، همراه با گردکردن (R) و قطع‌کردن (C) برای رایانه‌های مختلف، چه برای گردکردن (R) و چه برای قطع کردن (C)، در جدول ۱.۱ داده شده‌اند.

مثال برای رایانه‌های PRIME با دقّت معمولی

$$M = ۲^{۲۳} = ۸۳۸۸۶۰۸$$

بنابراین تمام اعداد صحیح با شش رقم اعشاری و اکثر اعداد با هفت رقم اعشاری را دقیقاً می‌توان در نمایش با ممیز شناور با دقّت معمولی ذخیره کرد. برای واحد گردکردن

با حساب قطع‌شده $\delta = ۲^{-۲۳} \doteq ۱\text{ر}۱۹ \times ۱۰^{-۷}$

با حساب گردشده $\delta = ۲^{-۲۴} \doteq ۵\text{ر}۹۶ \times ۱۰^{-۸}$

کاربَرهای PRIME برای دقت معمولی هم حساب قطع‌شده و هم حساب گردشده را در اختیار دارند.

تقریباً در همهٔ حالات، حساب گردشده بر حساب قطع شده ارجحیّت بسیار دارد. این امر با جزئیات بیشتر در بخشهای بعدی بررسی شده است، ولی دلیل اصلی آن علامت اریبی ε در (۸.۲.۱)، در مقایسه با نااریبی آن در (۹.۲.۱) نهفته است.

پی‌ریز و سرریز وقتی‌که کرانهای نما در (۲.۲.۱) نقض شوند عدد متناظر x در (۱.۲.۱) نمی‌تواند در رایانه نشان داده شود. اکنون می‌بینیم که در این صورت در بارهٔ بردِ ممکنِ طول x چه می‌توانیم بگوییم.

کوچکترین عدد مثبت با ممیّز شناور عبارت است از

$$x_L = (۰٫۱۰\ldots۰)_\beta \cdot \beta^L = \beta^{L-۱}$$

با استفاده از $\gamma = \beta - ۱$، بزرگترین عدد مثبت با ممیّز شناور چنین خواهد شد

$$x_U = (۰٫\gamma\gamma\ldots\gamma)_\beta \cdot \beta^U = (۱ - \beta^{-t}) \cdot \beta^U$$

بنابراین تمام اعدادِ با ممیّزِ شناور x باید در رابطهٔ زیر صدق کنند

$$x_L \leq | x | \leq x_U \qquad (۱۷.۲.۱)$$

در زبان فورترَن اغلب رایانه‌ها، اگر یک عمل حساب به عددی چون x که $|x| > x_U$، منجر شود آنگاه خطای مخرّبی بروز خواهد کرد که موجب توقف برنامه می‌شود. این خطا را **خطای سرریز** می‌خوانند. برعکس، اگر

$$۰ < | x | < x_L$$

آنگاه معمولاً $\mathrm{fl}(x)$ صفر گرفته می‌شود و محاسبات ادامه می‌یابد. این خطا را **خطای پی‌ریز** می‌نامند.

مثال محاسبهٔ $x = s^{۱۰}$ را در یک رایانهٔ بزرگ IBM در نظر بگیرید. در این صورت یک خطای پی‌ریز وجود دارد اگر

$$s^{۱۰} < ۱۶^{-۶۵}$$

بنابراین اگر $| s | < ۱٫۴۹ \times ۱۰^{-۸}$، آنگاه $x = s^{۱۰}$ صفر گرفته می‌شود. همچنین اگر نامساوی

$$s^{۱۰} > ۱۶^{۶۳}$$

یا هم‌ارز آن، نامساوی

$$|\, s \,| > ۳٫۹ \times ۱۰^{۷}$$

برقرار باشد یک خطای سرریز پیش خواهد آمد.

۳.۱ تعاریف و منشأهای خطا

اکنون به ذکر یک رده‌بندی کلی از راههای مهمّی که خطاها در جواب یک مسأله وارد می‌شوند و بعضی راهها که بیرون از حیطهٔ ریاضیات قرار می‌گیرند، می‌پردازیم. مطلب را با چند تعریف ساده دربارهٔ خطا شروع می‌کنیم.

هنگام حل یک مسأله، ما دنبال جواب دقیق یا درستی هستیم که آن را با x_T نشان می‌دهیم. معمولاً تقریبهایی در حل مسأله پدید می‌آیند که یک جواب تقریبی x_A از آن نتیجه می‌شود. خطای x_A را با

$$\text{Error}(x_A) \equiv x_A \text{خطای} = x_T - x_A$$

تعریف می‌کنیم. در بسیاری از مسائل ترجیح می‌دهیم که درصد یا خطای نسبی x_A را مطالعه کنیم

$$\text{Rel}(x_A) \equiv x_A \text{ خطای نسبی} = \frac{x_T - x_A}{x_T}$$

به شرطی که $x_T \neq ۰$. به این مطلب قبلاً در اندازه‌گیری خطا در $\text{fl}(x)$، در (۷.۲.۱) اشاره کرده‌ایم.

مثال $x_T = e = ۲٫۷۱۸۲۸۱۸\ldots, \qquad x_A = \frac{۱۹}{۷} = ۲٫۷۱۴۲۸۵۷\ldots,$

$$\text{Error}(x_A) = ۰٫۰۰۳۹۹۶\ldots \qquad \text{Rel}(x_A) = ۰٫۰۰۱۴۷\ldots$$

به جای خطای نسبی، اغلب مفهوم ارقام بامعنی را به‌کار می‌بریم. گوییم x_A دارای m رقم (اعشاری) با معنی نسبت به x_T است اگر اندازهٔ خطای $x_T - x_A$ قدرمطلقی نابزرگتر از ۵ در $(m+۱)$ امین رقم x_T داشته باشد، وقتی شمارش از اولین رقم غیرصفر x_T از چپ به راست صورت گیرد.

مثال (الف) $x_T = \frac{۱}{۳}$، $x_A = ۰٫۳۳۳$، $|\, x_T - x_A \,| \doteq ۰٫۰۰۰۳۳$.

چون خطا در چهارمین رقم از چپ به راست، یعنی در اولین رقم غیرصفر x_T، کوچکتر از ۵ است، گوییم x_A دارای سه رقم با معنی نسبت به x_T است.

(ب) $x_T = ۲۳٫۴۹۶$ ، $x_A = ۲۳٫۴۹۴$ ، $|x_T - x_A| = ۰٫۰۰۲$.

عبارت x_A چهار رقم با معنی نسبت به x_T دارد، زیرا خطا در پنجمین رقم از چپ به راستِ اولین رقم غیرصفر در x_T، کوچکتر از ۵ است. توجه داشته باشید که اگر x_A به چهار رقم گرد شود، یک خطای دیگر پیش می‌آید و x_A دیگر دارای چهار رقم با معنی نخواهد بود.

(ج) $x_T = ۰٫۲۱۳۸$ ، $x_A = ۰٫۲۱۴۴$ ، $|x_T - x_A| = ۰٫۰۰۰۰۶$.

عدد x_A دو رقم با معنی نسبت به x_T دارد، نه سه رقم.

آنچه در زیر می‌آید گاهی در اندازه‌گیری ارقام با معنی به‌کار می‌رود. اگر

$$\left|\frac{x_T - x_A}{x_T}\right| \leq ۵ \times ۱۰^{-m-۱} \qquad (۱.۳.۱)$$

آنگاه x_A نسبت به x_T، m رقم با معنی دارد. برای نشان دادن این موضوع حالتی را در نظر می‌گیریم که $|x_T| < ۱$ ، $|x_T| \geq ۰٫۱$، در این صورت از (۱.۳.۱) نتیجه می‌شود که

$$|x_T - x_A| \leq ۵ \times ۱۰^{-m-۱} \, |x_T| < ۰٫۵ \times ۱۰^{-m}$$

چون $۰٫۱ \leq x_T < ۱$، در نتیجه x_A دارای m رقم با معنی است. اثبات برای حالت کلی x_T اساساً به همین گونه است، با استفاده از $x_T = \hat{x}_T \times ۱۰^e$ و حالت $|x_T| < ۱$ ، $|x_T| \geq ۰٫۱$، e یک عدد صحیح است. توجه داشته باشید که (۱.۳.۱) شرط کافی برای داشتن m رقم با معنی x_A است و نه یک شرط لازم برای آن. مثالهای (الف) و (ب) که در بالا داده شدند، یک رقم با معنی بیشتر از آنچه که آزمون (۱.۳.۱) بیان می‌کند، دارند.

منشأهای خطا اکنون به ذکر یک رده‌بندی تقریبی از منشأهای مهم خطا می‌پردازیم.

(م۱) مدل‌سازی ریاضی از یک مسألهٔ فیزیکی. منظور از تهیهٔ مدل ریاضی برای یک وضعیت فیزیکی کوشش برای به‌دست آوردن روابط ریاضی بین کمیتهایی است که از لحاظ فیزیکی مورد توجه هستند. به علت پیچیدگی واقعیتهای فیزیکی، برای ساختن یک مدل ریاضی که راحت‌تر بتوان به‌کاربرد، از فرضهای ساده‌کنندهٔ گوناگونی استفاده می‌کنند. براثر این فرضها، مدل حاصل از لحاظ دقت محدودیتهایی خواهد داشت که برحسب موارد استفاده از مدل، ممکن است این محدودیتها دردسرآفرین باشند یا نباشند. در حالتی که مدل به اندازهٔ کافی دقیق نباشد، حل عددی مدل نمی‌تواند این نبود اساسی دقّت را مرتفع سازد.

مثال پرتابه‌ای به جرم m به هوا پرتاب شده است که مسیر حرکتش همیشه نزدیک به سطح زمین باقی

می‌ماند. دستگاه مختصاتی مانند xyz در نظر می‌گیریم که مرکز آن بر سطح زمین و محور z عمود بر سطح زمین و جهت مثبت آن به طرف بالا انتخاب شده است. گیریم موقعیت پرتابه در لحظهٔ t با استفاده از نمادگذاری استانداردِ نظریهٔ میدان برداری به‌صورت $\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$ باشد. یک مدل برای حرکت پرتابه توسط قانون دوّم نیوتن به‌صورت زیر داده می‌شود

$$m\frac{d^2\mathbf{r}(t)}{dt^2} = -mg\mathbf{k} - b\frac{d\mathbf{r}(t)}{dt} \qquad (۲.۳.۱)$$

که $b > 0$ مقداری ثابت و g شتاب حاصل از گرانش است. این معادله می‌گوید تنها نیروهایی که بر پرتابه اثر می‌کنند عبارت‌اند از: (۱) نیروی گرانش زمین، (۲) نیروی اصطکاک که مستقیماً با قدر مطلق سرعت $|d\mathbf{r}(t)/dt| = |\mathbf{v}(t)|$ متناسب است و در جهت عکس مسیر حرکت عمل می‌کند.

در بعضی حالات، این یک مدل عالی است، و حتی لزومی به وارد کردن جملهٔ اصطکاک ندارد. ولی این مدل، نیروهای مقاومت را که به‌صورت عمود بر صفحهٔ حرکت، عمل می‌کنند در نظر نمی‌گیرد. مثلاً جهت عرضی باد را که اثر پدیدهٔ کوریولیس[1] را خنثی می‌کند نادیده می‌گیرد. همچنین ممکن است نیروی اصطکاک در (۲.۳.۱) متناسب با $|\mathbf{v}(t)|^\alpha$ با $\alpha \neq 1$ باشد.

اگر مدلی برای منظورهای فیزیکی از لحاظ دقت مناسب باشد، می‌خواهیم یک طرح عددی برای حفظ این دقت به‌کار بریم. ولی اگر مدل مناسب نباشد، آنالیز عددی نمی‌تواند دقت آن را بیشتر کند مگر به تصادف. از سوی دیگر پسندیده نیست که با وارد کردن مطالبی که در مقایسه با پدیدهٔ مورد مطالعه، نسبتاً کم اهمیّت هستند، مدلی بسازیم که بیش از حد لازم پیچیده باشد. گاهی یک مدل پیچیده‌تر ممکن است مشکلات آنالیز عددی دیگری ایجاد کند، بی آنکه دقّت بیشتر قابل توجهی عاید سازد. برای مطالعهٔ کتابهایی که صراحتاً در زمینهٔ مدل‌یابی ریاضی در علوم مطالبی دارند می‌توانید به بندر[2] (۱۹۷۸)، لین و سگال[3] (۱۹۷۴)، مَکی و تامسن[4] (۱۹۷۳) و روبینو[5] (۱۹۷۵) مراجعه کنید.

(م۲) اشتباهات. تا پیش از پیدایش رایانه‌ها، خطاهای اتفاقی حساب همواره مسألهٔ جدی بودند. طرحهایی برای رسیدگی، بعضاً خیلی پرزحمت، ابداع کرده بودند تا چنین خطاهایی را در صورت یافت‌شدن، پیش از آنکه از موضع خطا زیاد دور شوند، کشف و اصلاح کنند. برای مثالی از این نوع در طرحهای رسیدگی، که هنگام حل یک دستگاه معادلات خطی به‌کار رفته‌اند، به اثر فاده‌یوا[6] مراجعه کنید.

---

1. Coriolis 2. Bender 3. Lin and Segal
4. Maki and Thompson 5. Rubinow 6. Fadeeva

با پیدایش رایانه‌های رقمی، نوع اشتباهات تغییر کرده است. خطاهای اتفاقی محاسباتی (مثلاً بدعمل کردن رایانه) اکنون نسبتاً نادر، و معمولاً خطاهای برنامه‌نویسی مشکل اساسی شده‌اند. اغلب یک خطای برنامه چندین بار ضمن اجرای برنامه تکرار می‌شود و وجود آن با خروجی عددی ناموجّه آشکار می‌شود (اگرچه پیدا کردن منشأ خطا باز ممکن است دشوار باشد). ولی وقتی برنامه‌های رایانه‌ایی رفته رفته پیچیده‌تر و طولانیتر می‌شوند وجود یک خطای کوچک برنامه و کشف و اصلاح آن ممکن است دشوار باشد، ولو اینکه این خطا ممکن است یک اختلاف ظریف ولی اساسی در نتایج به بار آورد. بدین جهت خطازدایی خوب برنامه خیلی مهمّ است، اگرچه در نگاه اول قابل توجه به نظر نیاید.

برای کشف خطاهای برنامه‌نویسی، داشتن راههایی که بتوان دقّت خروجی برنامه را رسیدگی نمود حائز اهمّیت است. وقتی برای اوّلین بار برنامه را اجرا می‌کنید باید، در صورت امکان، حالتهایی را که جواب درست آنها را می‌دانید به‌کار برید. برای یک برنامهٔ پیچیده، آن را به زیر برنامه‌های کوچکتری خُرد کنید تا بتوانید هر کدام را جداگانه مورد آزمایش قرار دهید. وقتی که کل برنامه رسیدگی و از درستی آن اطمینان حاصل شد خروجی را زیر نظر داشته باشید و ببینید که قابل قبول است یا نیست.

(م ۳) **عدم قطعیّت در داده‌های فیزیکی.** بیشتر داده‌های حاصل از تجربیات فیزیکی متضمن یک نوع خطا یا عدم قطعیّت هستند. این عدم قطعیّت، بر دقّت هر محاسبه‌ای که به این داده‌ها بستگی داشته باشد اثر می‌گذارد و در دقّت جوابها محدودیتی ایجاد می‌کند. تکنیکهای تحلیل نتایج در محاسباتِ دیگرِ این خطا، کاملاً مشابه تکنیکهای تحلیل نتایج خطاهای گردکردن است، اگرچه خطا در داده‌ها معمولاً بسیار بزرگتر از خطاهای گردکردن است. در بخشهای بعد، مطالب این موضوع بیشتر مورد بحث قرار خواهندگرفت.

(م ۴) **خطاهای ماشین.** منظور از خطاهای ماشین خطاهایی هستند که ذاتاً در اثر به‌کار بردن شکلِ ممیزِ شناورِ اعداد ایجاد می‌شوند. بالاخص، منظور ما خطاهای قطع‌کردن/گردکردن و خطاهای پی‌ریز/سرریز هستند. خطاهای قطع‌کردن و گردکردن نتیجهٔ محدودیّت طول مانتیس ممیّز شناور هستند، و این خطاها در تمام عملیّات حساب رایانه‌یی پیش می‌آیند. کلیّهٔ شکلهای این خطاهای ماشینی در بخش ۲.۱ مورد بحث قرارگرفته‌اند. در بخشهای آینده بعضی از پی‌آمدهای این خطاها مورد توجه قرار می‌گیرند. همچنین، برای سادگی نمادگذاری، از این پس هرجا که بتوانیم، از اصطلاح **خطای گردکردن** که شامل قطع‌کردن نیز خواهد بود استفاده خواهیم کرد.

(م ۵) **خطای برشِ ریاضی.** این نام به خطای تقریب در حل عددی یک مسألهٔ ریاضی اطلاق می‌شود، و خطایی است که عمدتاً با موضوع آنالیز عددی در ارتباط است. این خطا شامل

تقریب فرایندهای نامحدود از راه فرایندهای محدود، جایگزین‌کردن مسائل محاسبه‌ناپذیر با مسائل محاسبه‌پذیر است. برای روشن‌شدن بیشتر موضوع چند مثال می‌آوریم.

مثال (الف) با استفاده از دو جملهٔ اوّل سری تیلر (۷.۱.۱)، داریم

$$\sqrt{۱+x} \doteq ۱ + \frac{۱}{۲}x \qquad (۳.۳.۱)$$

که وقتی x کوچک باشد تقریبِ خوبی است. برای مطالعهٔ مبحث کلی تقریب توابع، به فصل ۴ مراجعه کنید.

(ب) برای محاسبهٔ یک انتگرال روی $[۰, ۱]$ از فرمول زیر استفاده می‌کنیم

$$\int_{۰}^{۱} f(x)dx \doteq \frac{۱}{n}\sum_{j=۱}^{n} f\left(\frac{۲j-۱}{۲n}\right) \qquad n = ۱, ۲, ۳, \ldots \qquad (۴.۳.۱)$$

این عمل، قاعدهٔ انتگرالگیری عددی نقطهٔ میانی خوانده می‌شود. برای توضیح بیشتر به آخرین قسمت بخش ۲.۵ مراجعه کنید. موضوع کلی انتگرالگیری عددی در فصل ۵ آمده است.

(ج) برای مسألهٔ معادلهٔ دیفرانسیل

$$Y'(t) = f(t, Y(t)) \qquad Y(t_۰) = Y_۰ \qquad (۵.۳.۱)$$

از تقریب مشتق

$$Y'(t) \doteq \frac{Y(t+h) - Y(t)}{h}$$

که در آن h مقدار کوچکی است استفاده می‌کنیم. گیریم $t_j = t_۰ + jh$، $j \geq ۰$، یک تابع جواب تقریبی $y(t_j)$ را با تساوی

$$\frac{y(t_{j+۱}) - y(t_j)}{h} = f(t_j, y(t_j))$$

تعریف می‌کنیم. بنابراین داریم

$$y(t_{j+۱}) = y(t_j) + hf(t_j, y(t_j)) \qquad j \geq ۰ \qquad y(x_۰) = Y_۰$$

این روش همان **روش اویلر** برای حل مسألهٔ مقدار اوّلیه در معادلات دیفرانسیل معمولی است. بحث جامع و تحلیل آن در ۲.۶ آمده است. در فصل ۶ بسط کامل روشهای عددی برای مسألهٔ مقدار اولیه (۵.۳.۱) داده خواهد شد.

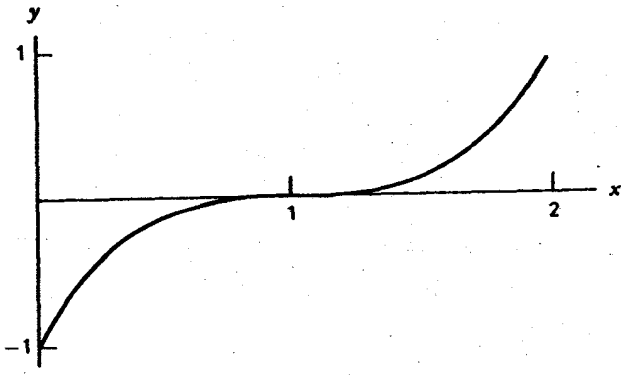
اغلب مسائل آنالیز عددی در فصلهای آینده عمدتاً شامل خطاهای برشی هستند. مورد استثنای عمده، حل دستگاههای معادلات خطی است که در آن خطاهای گردکردن منشأ اصلی خطا هستند.

نوفه در محاسبهٔ تابع. یکی از نتایج بلافصل خطاهای گردکردن آن است که محاسبهٔ تابع $f(x)$ با رایانه به یک تابع تقریبی $\hat{f}(x)$ که پیوسته نیست می‌انجامد. هر چند این امر تنها وقتی آشکار می‌شود که نمودار $\hat{f}(x)$ را روی مقیاسی به اندازهٔ کافی کوچک بررسی کنیم. پس از هر عمل حساب که در محاسبهٔ $f(x)$ بکار برده می‌شود معمولاً یک خطای گردکردن وجود دارد. وقتی نتیجهٔ این خطاهای گردکردن در نظر گرفته شود، مقدار محاسبه شدهٔ $\hat{f}(x)$ به‌دست می‌آید که خطای آن، $f(x) - \hat{f}(x)$، وقتی x تغییر کند به‌صورت یک عدد کوچک تصادفی به نظر می‌رسد. این خطا در $\hat{f}(x)$ را **نوفه** نامند. هنگامی که نمودار $\hat{f}(x)$ روی مقیاسی به اندازهٔ کافی کوچک بررسی شود، به‌صورت نوار تیره‌ای از نقاط دیده می‌شود که در آن مقادیر x تمام اعداد با ممیّز شناور قابل قبول در ماشین را در بر می‌گیرد. این امر پیامدهایی برای بسیاری از برنامه‌های بعدی که در آنها از $\hat{f}(x)$ استفاده می‌کنند خواهد داشت. مثلاً، محاسبهٔ ریشهٔ $f(x)$ با استفاده از $\hat{f}(x)$ به عدم قطعیّت در محل ریشه منجر می‌شود، زیرا محتملاً جای ریشه در محل تلاقی محور xها و نمودار نوار تیره‌رنگ $\hat{f}(x)$ خواهد بود. مثال زیر نشان می‌دهد که این امر ممکن است موجب عدم قطعیّت قابل ملاحظه‌ای در جای ریشه شود.

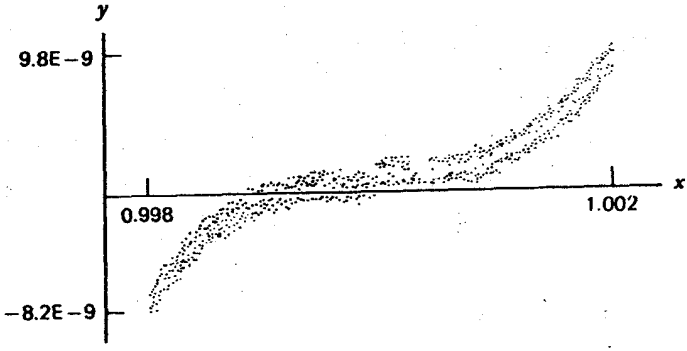
مثال گیریم

$$f(x) = x^3 - 3x^2 + 3x - 1 \qquad (۶.۳.۱)$$

که همان $(x-1)^3$ است. ما (۶.۳.۱) را با BASIC با دقت معمولی در یک ریزرایانهٔ معمولی، با استفاده از حساب گردشدهٔ دودویی و واحد گردکردن $\delta \doteq 2^{-24} = 5{,}96 \times 10^{-8}$ محاسبه نموده‌ایم. نمودار $\hat{f}(x)$ روی بازهٔ $[0, 2]$ در شکل ۱.۱ نشان داده شده است، و همان گونه که انتظار می‌رود یک منحنی پیوسته و هموار به نظر می‌آید. ولی نمودار روی یک بازهٔ کوچکتر $[1{,}998, 1{,}002]$ ماهیّت ناپیوستهٔ $f(x)$ را، همان گونه که در شکل ۲.۱ دیده می‌شود، نشان می‌دهد. در این حالت اخیر، $\hat{f}(x)$ در ۶۴۰ نقطه متساوی‌الفاصله برای مقادیر x در $[1{,}998, 1{,}002]$ محاسبه شده است که یک نوار تیره رنگ نمودار $f(x)$ را به وجود آورده است. از نمودار اخیر می‌توان دید که یک بازهٔ بزرگ عدم قطعیت برای محلّ تقاطع $\hat{f}(x)$ با محور xها وجود دارد. در بخش ۷.۲، فصل دوّم، به این مبحث باز خواهیم گشت.



شکل ۱.۱ گراف (۶.۳.۱).



شکل ۲.۱ جزئیات گراف (۶.۳.۱).

خطاهای پی‌ریز/سرریز در محاسبات. پی‌آمد دیگری از خطاهای ماشین را در نظر می‌گیریم. حدود بالا و پایین اعداد با ممیّز شناور، که در (۱۷.۲.۱) داده شده، ممکن است به خطاهایی در محاسبات بینجامد. گاهی این خطاها اجتناب‌ناپذیر، ولی بیشتر محصول طریقهٔ محاسبات‌اند.

برای نشان دادن این موضوع، محاسبهٔ قدرمطلق یک عدد مختلط را در نظر بگیرید

$$| x + iy | = \sqrt{x^2 + y^2} \qquad (۷.۳.۱)$$

این محاسبه ممکن است به پی‌ریز یا سرریز منجر شود، ولو اینکه قدرمطلق $| x+iy |$ در محدودهٔ ماشین باشد. مثلاً اگر در رابطهٔ (۱۷.۲.۱) $x_U = ۱٫۷ \times ۱۰^{۳۸}$ اختیار شود، آنگاه برای $x = y = ۱۰^{۲۰}$ رابطهٔ (۷.۳.۱) سرریز می‌دهد ولو اینکه $| x + iy | \doteq ۱٫۴ \times ۱۰^{۲۰}$. برای احتراز از این مشکل

بزرگترین مقدار x یا y، مثلاً x را معین می‌کنیم. رابطهٔ (۷.۳.۱) را دوباره به شکل زیر می‌نویسیم

$$| x + iy | = | x | \cdot \sqrt{1 + \alpha^2}, \quad \alpha = \frac{y}{x} \qquad (۸.۳.۱)$$

باید $\sqrt{1+\alpha^2}$ را در حالت $0 \leq \alpha \leq 1$، محاسبه کنیم. این کار مشکل پی‌ریز و سرریز را برای مسائل (۷.۳.۱) از بین خواهد برد.

۴.۱ انتشار خطاها

در این بخش و بخشهای بعدی آثار محاسبات با اعدادی را که خطا دارند مد نظر قرار می‌دهیم. ابتدا اعمال اصلی حساب را در نظر می‌گیریم. فرض کنید ω یکی از اعمال جمع، تفریق، ضرب یا تقسیم حساب باشد؛ و گیریم $\hat{\omega}$ شکل رایانه‌یی همان عمل حساب، که معمولاً شامل خطای گردکردن است، باشد. گیریم x_A و y_A اعدادی باشند که برای محاسبه به‌کار رفته‌اند و فرض کنید که این اعداد دارای خطا باشند و مقدار درست آنها چنین باشد

$$x_T = x_A + \varepsilon \qquad y_T = y_A + \eta$$

پس $x_A \hat{\omega} y_A$ عددی است که عملاً محاسبه شده است و برای خطای آن داریم

$$x_T \omega y_T - x_A \hat{\omega} y_A = [x_T \omega y_T - x_A \omega y_A] + [x_A \omega y_A - x_A \hat{\omega} y_A] \quad (۱.۴.۱)$$

کمیّت کروشهٔ اول را خطای منتشره گویند و کمیّت کروشهٔ دوّم معمولاً خطای قطع‌کردن یا گردکردن است. برای این کمیّت دوّم، معمولاً داریم

$$x_A \hat{\omega} y_A = \mathrm{fl}(x_A \omega y_A) \qquad (۲.۴.۱)$$

که بدین معنی است که $x_A \omega y_A$ دقیقاً محاسبه و سپس گردشده است. از ترکیب (۹.۲.۱) و (۲.۴.۱)، داریم

$$| x_A \omega y_A - x_A \hat{\omega} y_A | \leq \frac{\beta}{2} | x_A \omega y_A | \beta^{-t} \qquad (۳.۴.۱)$$

به شرطی که از گردکردن درست استفاده شده باشد.

برای خطای منتشره، حالتهای خاص را بررسی می‌کنیم.

حالت الف. ضرب. برای خطا در $x_A y_A$

$$x_T y_T - x_A y_A = x_T y_T - (x_T - \varepsilon)(y_T - \eta) = x_T \eta + y_T \varepsilon - \varepsilon\eta$$

$$\mathrm{Rel}(x_A y_A) = \frac{x_T y_T - x_A y_A}{x_T y_T} = \frac{\eta}{y_T} + \frac{\varepsilon}{x_T} - \frac{\varepsilon}{x_T} \cdot \frac{\eta}{y_T}$$
$$= \mathrm{Rel}(x_A) + \mathrm{Rel}(y_A) - \mathrm{Rel}(x_A) \cdot \mathrm{Rel}(y_A)$$

برای $|\mathrm{Rel}(y_A)| \ll ۱$ و $|\mathrm{Rel}(x_A)| \ll ۱$

$$\mathrm{Rel}(x_A y_A) \doteq \mathrm{Rel}(x_A) + \mathrm{Rel}(y_A) \qquad (۴.۴.۱)$$

نماد «$\ll$» به معنای «بسیار کوچکتر است از» است.

حالت ب. تقسیم. با استدلالی مشابه

$$\mathrm{Rel}\frac{x_A}{y_A} = \frac{\mathrm{Rel}(x_A) - \mathrm{Rel}(y_A)}{۱ - \mathrm{Rel}(y_A)} \qquad (۵.۴.۱)$$

برای $|\mathrm{Rel}(y_A)| \ll ۱$؛

$$\mathrm{Rel}(\frac{x_A}{y_A}) \doteq \mathrm{Rel}(x_A) - \mathrm{Rel}(y_A) \qquad (۶.۴.۱)$$

برای ضرب و تقسیم، خطاهای نسبی به سرعت افزایش نمی‌یابند.

حالت ج. جمع و تفریق

$$(x_T \pm y_T) - (x_A \pm y_A) = (x_T - x_A) \pm (y_T - y_A) = \varepsilon \pm \eta$$

$$\mathrm{Error}(x_A \pm y_A) = \mathrm{Error}(x_A) \pm \mathrm{Error}(y_A) \qquad (۷.۴.۱)$$

این امر کاملاً خوب و منطقی به نظر می‌رسد، ولی ممکن است گمراه کننده باشد. خطای نسبی $x_A \pm y_A$ در مقایسه با $\mathrm{Rel}(x_A)$ و $\mathrm{Rel}(y_A)$ ممکن است خیلی ناچیز باشد.

مثال گیریم $x_T = \pi$ و $x_A = ۳٫۱۴۱۶$ و $y_T = \frac{۲۲}{۷}$ و $y_A = ۳٫۱۴۲۹$ پس

$$x_T - x_A \doteq -۷٫۳۵ \times ۱۰^{-۶} \qquad \mathrm{Rel}(x_A) \doteq -۲٫۳۴ \times ۱۰^{-۶}$$

$$y_T - y_A \doteq -۴٫۲۹ \times ۱۰^{-۵} \qquad \mathrm{Rel}(y_A) \doteq -۱٫۳۶ \times ۱۰^{-۵}$$

$$(x_T - y_T) - (x_A - y_A) \doteq -۰٫۰۰۱۲۶۴۵ - (-۰٫۰۰۱۳) \doteq ۳٫۵۵ \times ۱۰^{-۵}$$

$$\mathrm{Rel}(x_A - y_A) \doteq -۰٫۰۲۸$$

اگرچه خطای $x_A - y_A$ کاملاً کوچک است ولی خطای نسبی $x_A - y_A$ بسیار بزرگتر از خطای نسبی x_A یا خطای نسبی y_A است.

خطاهای کاهش ارقام بامعنی. مثال اخیر نشان می‌دهد که از لحاظ خطای نسبی، از تفریق دو کمیّت نزدیک به‌هم، ممکن است کاهش زیادی در دقّت پیش آید. این عمل ممکن است از راههای مهّم از دست دادن دقّت هنگام انتشار خطا در یک محاسبه باشد.

اکنون چند مثال برای این پدیده می‌آوریم، توصیه‌هایی را برای چگونگی جلوگیری از این مشکلات در بعضی موارد، پیشنهاد می‌کنیم.

مثال حل معادلهٔ $ax^2 + bx + c = 0$ را وقتی $4ac$ در مقایسه با b^2 مقدار کوچکی است در نظر گرفته و از فرمولی معمول برای پیداکردن ریشه‌ها استفاده می‌کنیم

$$r_T^{(1)} = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad r_T^{(2)} = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \quad (8.4.1)$$

برای روشن‌کردن مطلب، معادلهٔ $x^2 - 26x + 1 = 0$ را در نظر می‌گیریم. از فرمولهای (8.4.1) نتیجه می‌گیریم

$$r_T^{(1)} = 13 + \sqrt{168} \qquad r_T^{(2)} = 13 - \sqrt{168} \qquad (9.4.1)$$

اکنون فرض کنید از یک ماشین با پنج رقم ده‌دهی استفاده می‌کنیم. در این ماشین $\sqrt{168} \doteq 12{,}961$. بنابراین

$$r_A^{(1)} = 13 + 12{,}961 = 25{,}961 \quad r_A^{(2)} = 13 - 12{,}961 = 0{,}039 \quad (10.4.1)$$

اگر جوابهای دقیق را بکار بریم

$$\mathrm{Rel}(r_A^{(1)}) \doteq 1{,}85 \times 10^{-5} \quad \mathrm{Rel}(r_A^{(2)}) \doteq 1{,}25 \times 10^{-2} \quad (11.4.1)$$

برای داده‌هایی که در محاسبهٔ (10.4.1) آمده‌اند، با استفاده از نمادگذاری (7.4.1) داریم

$$x_T = x_A = 13 \quad y_T = \sqrt{168} \qquad y_A = 12{,}961$$
$$\mathrm{Rel}(x_A) = 0 \qquad \mathrm{Rel}(y_A) \doteq 3{,}71 \times 10^{-5}$$

دقت در $r_A^{(۲)}$ بسیارکمتر از دقت در داده‌های x_A و y_A است که در محاسبه وارد شده‌اند. گوییم ارقام با معنی در تفاضل $r_A^{(۲)} = x_A - y_A$ از بین رفته‌اند، یا در محاسبهٔ $r_A^{(۲)}$ دچار **خطای کاهش ارقام بامعنی** شده‌ایم. در $r_A^{(۱)}$، دقّتی با پنج رقم با معنی داریم، حال آنکه در $r_A^{(۲)}$ فقط دو رقم با معنی داریم.

برای رفع این نقیصه در این حالت خاصِ محاسبهٔ $r_A^{(۲)}$، (۹.۴.۱) را به شکل زیر می‌نویسیم

$$r_T^{(۲)} = \frac{۱۳ - \sqrt{۱۶۸}}{۱} \cdot \frac{۱۳ + \sqrt{۱۶۸}}{۱۳ + \sqrt{۱۶۸}} = \frac{۱}{۱۳ + \sqrt{۱۶۸}}$$

سپس

$$\frac{۱}{۱۳ + \sqrt{۱۶۸}} \doteq \frac{۱}{۲۵٫۹۶۱} \doteq ۰٫۰۳۸۵۱۹ \equiv r_A^{(۲)} \qquad (۱۲.۴.۱)$$

در اینجا دو خطا وجود دارد، یکی مربوط به $\sqrt{۱۶۸} \doteq ۱۲٫۹۶۱$ و دیگری مربوط به تقسیم نهایی. ولی هر یک از آنها خطاهای نسبی کوچکی دارند [(۶.۴.۱) را ببینید] و مقدار جدید $r_A^{(۲)}$ از مقدار قبلی بسیار دقیقتر است. با محاسبات دقیق، اکنون داریم

$$\mathrm{Rel}(r_A^{(۲)}) \doteq -۱٫۰۳ \times ۱۰^{-۵},$$

که از (۱۱.۴.۱) خیلی بهتر است.

این محاسبهٔ جدید $r_A^{(۲)}$ نشان می‌دهد که خطای کاهش ارقام با معنی به شکل محاسبه بستگی دارد، نه به خطاهای موجود در داده‌های محاسبه. در این مثال به آسانی توانستیم راه دیگری برای محاسبات پیدا کنیم که خطای کاهش ارقام با معنی را مرتفع سازد، ولی این کار همیشه ممکن نخواهد بود. برای یک بحث کامل از روش عملی محاسبهٔ ریشه‌های یک چند جمله‌یی درجه ۲، فورسایت (۱۹۶۹) را ببینید.

مثال در بسیاری از محاسبات کاهش ارقام بامعنی، تقریبهای چندجمله‌یی تیلر می‌توانند مشکل را حل کنند. ما این مطلب را با محاسبهٔ

$$f(x) = \int_۰^۱ e^{xt} dt = \frac{e^x - ۱}{x} \qquad x \neq ۰ \qquad (۱۳.۴.۱)$$

نشان می‌دهیم. برای $x = ۰$، $f(۰) = ۱$؛ و به‌سادگی می‌توان دید که $f(x)$ در $x = ۰$ پیوسته است.

برای دیدن مسألهٔ کاهش ارقام بامعنی، وقتی x کوچک است، $f(x)$ را با یک ماشین‌حساب دستی مناسب ده رقمی در $x = ۱٫۴ \times ۱۰^{-۹}$ محاسبه می‌کنیم. نتایج چنین‌اند

$$e^x \doteq ۱٫۰۰۰۰۰۰۰۰۱$$
$$\frac{e^x - ۱}{x} \doteq \frac{۱۰^{-۹}}{۱٫۴ \times ۱۰^{-۹}} \doteq ۰٫۷۱۴ \qquad (۱۴.۴.۱)$$

سمت راست تساویها، نتایج ماشین حساب را می‌دهد و جواب درست که با ده رقم گرد شده عبارت است از

$$f(x) = ۱٫۰۰۰۰۰۰۰۰۱$$

در محاسبهٔ (۱۴.۴.۱) در صورت کسر نه رقم اول ارقام بامعنی حذف شده است.

برای اجتناب از خطای کاهش ارقام بامعنی از تقریب درجهٔ دوّم تیلر برای e^x استفاده و سپس $f(x)$ را ساده می‌کنیم

$$e^x = ۱ + x + \frac{x^۲}{۲} + \frac{x^۳}{۶}e^\xi \qquad ۰ \le \xi \le x \le ۱$$
$$f(x) = ۱ + \frac{x}{۲} + \frac{x^۲}{۶}e^\xi \qquad (۱۵.۴.۱)$$

به‌ازای مقدار قبلیِ $x = ۱٫۴ \times ۱۰^{-۹}$، مقدار $f(x)$ با خطایی کمتر از $۱۰^{-۱۸}$ چنین خواهد شد

$$f(x) \doteq ۱ + ۷ \times ۱۰^{-۱۰}$$

در حالت کلی (۱۵.۴.۱) را برای بازه‌ای چون $[۰, \delta]$ به‌کار می‌بریم و δ را طوری انتخاب می‌کنیم که مطمئن باشیم خطای

$$f(x) \doteq ۱ + \frac{x}{۲}$$

به اندازهٔ کافی کوچک است. البته، می‌توانستیم از تقریب با درجهٔ بالاتر برای e^x استفاده کنیم و مقدار δ را بزرگتر بگیریم.

به طورکلی، تقریبهای تیلر برای اجتناب از خطای کاهش ارقام با معنی در محاسبات مفیدند. ولی در بعضی موارد خطای کاهش ارقام بامعنی مستلزم دقّت بیشتر است.

مثال محاسبهٔ مجموعی به‌صورت

$$S = \sum_{j=1}^{n} x_j \qquad (۱۶.۴.۱)$$

را در نظر می‌گیریم با جمله‌های مثبت و منفی x_j، که هریک مقداری است تقریبی. بعلاوه فرض می‌کنیم که مجموع S خیلی کوچکتر از ماکسیمم قدرمطلق x_j ها باشد. در محاسبهٔ چنین مجموعی با رایانه، احتمالاً دچار خطای کاهش ارقام بامعنی می‌شویم. ما این مطلب را با یک مثال نشان می‌دهیم.

با استفاده از فرمول تیلر (۴.۱.۱) برای e^x، مقدار $e^{-۵}$ را محاسبه می‌کنیم:

$$e^{-۵} \doteq ۱ + \frac{(-۵)}{۱!} + \frac{(-۵)^۲}{۲!} + \frac{(-۵)^۳}{۳!} + \cdots + \frac{(-۵)^n}{n!} \qquad (۱۷.۴.۱)$$

تصور کنید که از رایانه‌ای با حساب ممیز شناور که با چهار رقم اعشار گرد می‌کند استفاده می‌کنیم، پس هر جمله از سری بالا را باید با چهار رقم با معنی گرد کرد. در جدول ۲.۱، مقادیر گردشدهٔ x_j همراه با مجموع دقیق جملات تا درجهٔ داده‌شده آمده است. مقدار درست $e^{-۵}$ تا چهار رقم بامعنی

جدول ۲.۱ محاسبهٔ (۱۷.۴.۱) با استفاده از حساب چهاررقم اعشاری

درجه	جمله	حاصل جمع	درجه	جمله	حاصل جمع
۰	۱٫۰۰	۱٫۰۰۰	۱۳	−۰٫۱۹۶۰	−۰٫۰۴۲۳۰
۱	−۵٫۰۰۰	−۴٫۰۰۰	۱۴	۰٫۷۰۰۱E − ۱	۰٫۰۲۷۷۱
۲	۱۲٫۵۰	۸٫۵۰۰	۱۵	−۲۳۳۴E − ۱	۰٫۰۰۴۳۷۰
۳	−۲۰٫۸۳	−۱۲٫۳۳	۱۶	۰٫۷۲۹۳E − ۲	۰٫۱۱۶۶
۴	−۲۶٫۰۴	۱۳٫۷۱	۱۷	−۰٫۲۱۴۵E − ۲	۰٫۰۰۹۵۱۸
۵	۲۶٫۰۴	−۱۲٫۳۳	۱۸	۰٫۵۹۵۸E − ۳	۰٫۰۱۰۱۱
۶	۲۱٫۷۰	۹٫۳۷۰	۱۹	−۰٫۱۵۶۸E − ۳	۰٫۰۰۹۹۵۷
۷	−۱۵٫۵۰	−۶٫۱۳۰	۲۰	۰٫۳۹۲۰E − ۴	۰٫۰۰۹۹۹۶
۸	۹٫۶۸۸	۳٫۵۵۸	۲۱	−۰٫۹۳۳۳E − ۵	۰٫۰۰۹۹۸۷
۹	−۵٫۳۸۲	−۱٫۸۲۴	۲۲	۰٫۲۱۲۱E − ۵	۰٫۰۰۹۹۸۹
۱۰	۲٫۶۹۱	۰٫۸۶۷۰	۲۳	−۰٫۴۶۱۱E − ۶	۰٫۰۰۹۹۸۹
۱۱	−۱٫۲۲۳	−۰٫۳۵۶۰	۲۴	۰٫۹۶۰۷E − ۷	۰٫۰۰۹۹۸۹
۱۲	۰٫۵۰۹۷	۰٫۱۵۳۷	۲۵	−۰٫۱۹۲E − ۷	۰٫۰۰۹۹۸۹

برابر ۰٫۰۰۶۷۳۸ است و با آخرین مجموع در جدول کاملاً متفاوت است. همچنین اگر (۱۷.۴.۱) را دقیقاً برای $n = ۲۵$ محاسبه کنیم مقدار صحیح $e^{-۵}$ تا چهار رقم بامعنی به‌دست می‌آید.

در این مثال، جمله‌های x_j نسبتاً بزرگ می‌شوند، ولی مجموع آنها عدد بسیار کوچکتر $e^{-۵}$ را تشکیل می‌دهد. این امر به معنای وجود خطاهای کاهش ارقام بامعنی در مجموعیابی است. احتراز از این مشکل در این حالت نسبتاً آسان است. یا می‌نویسیم

$$e^{-۵} = \frac{۱}{e^{۵}}$$

و $e^{۵}$ را با یک سری که حذف جملات مثبت و منفی درآن وجود ندارد محاسبه می‌کنیم؛ یا $e^{-۱} = \frac{۱}{e}$ را مثلاً با استفاده از یک سری حساب می‌کنیم، و آن را در خودش چند بار ضرب می‌کنیم تا $e^{-۵}$ حاصل شود. در مورد سریهای دیگر، ممکن است که این‌گونه راه حل ساده‌ای وجود نداشته باشد.

انتشار خطا در محاسبهٔ توابع. گیریم تابع $f(x)$ داده شده و $\hat{f}(x)$ معرّف مقدار محاسبه‌شدهٔ $f(x)$ در رایانه باشد. اگر $f(x_T)$ معرّف مقدار تابع مطلوب و $\hat{f}(x_A)$ مقداری است که درواقع حساب شده باشد. برای خطا، می‌نویسیم

$$f(x_T) - \hat{f}(x_A) = [f(x_T) - f(x_A)] + [f(x_A) - \hat{f}(x_A)] \quad (۱۸.۴.۱)$$

کمیّت اولین کروشه را **خطای منتشره** می‌نامند، و کمیّت دوّمین کروشه، خطای محاسبهٔ $f(x_A)$ در رایانه است. این خطای دوّم معمولاً یک عدد تصادفی کوچک است که از خطاهای گردکردن در اعمال حساب تعریف‌کنندهٔ $f(x)$ به‌وجود می‌آید. ما قبلاً در بخش ۳.۱ به آن به‌عنوان نوفه در محاسبهٔ $f(x)$ اشاره کرده‌ایم.

برای خطای منتشره، از قضیهٔ مقدار میانگین داریم

$$f(x_T) - f(x_A) \doteq f'(x_T)(x_T - x_A) \quad (۱۹.۴.۱)$$

در اینجا فرض می‌شود که x_A و x_T نسبتاً به یکدیگر نزدیک‌اند و $f'(x)$ وقتی x بین x_A و x_T است خیلی تغییر نمی‌کند.

مثال ۲۶ $\sin(\frac{\pi}{۵}) - \sin(۰٫۶۲۸) \doteq \cos(\frac{\pi}{۵})[(\frac{\pi}{۵}) - ۰٫۶۲۸] \doteq ۰٫۰۰۰۲۶$ که یک برآورد عالی برای خطاست.

با استفاده از قضیهٔ تیلر (۱۲.۱.۱) برای توابع دو متغیره، می‌توانیم بحث اخیر را برای انتشار خطا در توابع دو متغیره تعمیم دهیم.

$$f(x_T, y_T) - f(x_A, y_A) \doteq f_x(x_T, y_T)(x_T - x_A) + f_y(x_T, y_T)(y_T - y_A) \tag{۲۰.۴.۱}$$

که در آن $f_x \equiv \partial f/\partial x$. در اینجا فرض می‌کنیم که $f_x(x,y)$ و $f_y(x,y)$ برای (x,y) بین (x_T, y_T) و (x_A, y_A) خیلی تغییر نمی‌کنند.

مثال برای $f(x,y) = x^y$، داریم $f_x = yx^{y-1}$، $f_y = x^y \log(x)$. پس از (۲۰.۴.۱) به‌دست می‌آوریم

$$x_T^{y_T} - x_A^{y_A} \doteq \varepsilon \cdot y_T x_T^{y_T - 1} + \eta[\log(x_T)]x_T^{y_T}$$
$$\mathrm{Rel}(x_A^{y_A}) \doteq y_T[\mathrm{Rel}(x_A) + \mathrm{Rel}(y_A)\log(x_T)] \tag{۲۱.۴.۱}$$

خطای نسبی در $x_A^{y_A}$ ممکن است بزرگ باشد، اگرچه $\mathrm{Rel}(x_A)$ و $\mathrm{Rel}(y_A)$ کوچک‌اند. به‌عنوان یک مثال دیگر گیریم $y_T = y_A = ۵۰۰$، $x_T = ۱٫۲$، $x_A = ۱٫۲۰۰۱$. در این صورت $x_T^{y_T} = ۳٫۸۹۶۰۴ \times ۱۰^{۳۹}$ و $x_A^{y_A} = ۴٫۰۶۱۷۹ \times ۱۰^{۳۹}$، $\mathrm{Rel}(x_A^{y_A}) = -۰٫۰۴۲۵$. این مقدار را با $\mathrm{Rel}(x_A) = ۸٫۳ \times ۱۰^{-۵}$ مقایسه کنید.

خطای داده‌ها. اگر داده‌های ورودی یک الگوریتم فقط r رقم دقّت داشته باشند، گاهی توصیه می‌شود که در چنین حالت محاسباتی که در آنها این داده‌ها دخالت دارند، فقط حساب rرقمی را به‌کار برند. این کار معقولی نیست. البته قبول داریم که دقت محدود داده‌ها بر نتایج نهایی محاسبات الگوریتمی اثر می‌گذارد و جوابهایی به‌دست می‌دهد که دارای خطا هستند، مع‌هذا دلیلی وجود ندارد که وضعیت را با خطاهای گردکردنی که در اثر محاسبات با r رقم ناشی می‌شوند، بدتر کنیم. به‌عکس باید محاسبات را با دقت بیشتری انجام دهیم تا از تنزّل بیشتر در دقت نتایج الگوریتم جلوگیری‌شود. این امر موجب می‌شود که خطاهای گردکردن حسابی به مراتب ناچیزتر از خطای داده‌ها شوند و کمک می‌کند که دقّت در حد دقّت داده‌ها حفظ شود.

۵.۱ خطاهای مجموعیابی

در بسیاری از روشهای عددی، به‌ویژه در جبر خطی، مجموعیابی وجود دارد. در این بخش، وجوه متفاوتی از مجموعیابی را مد نظر قرار می‌دهیم، بویژه وقتی در حساب با ممیّز شناور انجام گرفته باشند.

محاسبهٔ مجموع زیر را در نظر می‌گیریم

$$S = \sum_{j=1}^{m} x_j \qquad (۱.۵.۱)$$

که x_1، ... ، x_m اعداد با ممیز شناورند. تعریف می‌کنیم

$$S_2 = \mathrm{fl}(x_1 + x_2) = (x_1 + x_2)(1 + \varepsilon_2) \qquad (۲.۵.۱)$$

که در آن از (۲.۴.۱) و (۱۰.۲.۱) استفاده کرده‌ایم. به‌طور بازگشتی

$$S_{r+1} = \mathrm{fl}(S_r + x_{r+1}) \qquad r = 2, \ldots, m-1$$

تعریف می‌شود. پس

$$S_{r+1} = (S_r + x_{r+1})(1 + \varepsilon_{r+1}) \qquad (۳.۵.۱)$$

کمیتهای ε_2، ... ، ε_m برحسب آنکه از قطع‌کردن یا گردکردن استفاده شده باشد، در (۸.۲.۱) یا (۹.۲.۱) صدق می‌کنند.

با بسط چند مجموع اوّلیه، به‌دست می‌آوریم:

$$\begin{aligned}
S_2 - (x_1 + x_2) &= \varepsilon_2(x_1 + x_2) \\
S_3 - (x_1 + x_2 + x_3) &= (x_1 + x_2)\varepsilon_2 + (x_1 + x_2)(1 + \varepsilon_2)\varepsilon_3 + x_3\varepsilon_3 \\
&\doteq (x_1 + x_2)\varepsilon_2 + (x_1 + x_2 + x_3)\varepsilon_3 \\
S_4 - (x_1 + x_2 + x_3 + x_4) &\doteq (x_1 + x_2)\varepsilon_2 + (x_1 + x_2 + x_3)\varepsilon_3 \\
&\quad + (x_1 + x_2 + x_3 + x_4)\varepsilon_4
\end{aligned}$$

ما جملات حاصلضرب $\varepsilon_i\varepsilon_j$ را نادیده گرفته‌ایم زیرا اندازه‌های بسیار کوچکتری دارند. به روش استقرا به‌دست می‌آوریم

$$\begin{aligned}
S_m - \sum_{1}^{m} x_i &\doteq (x_1 + x_2)\varepsilon_2 + \cdots + (x_1 + x_2 + \cdots + x_m)\varepsilon_m \\
&= x_1(\varepsilon_2 + \varepsilon_3 + \cdots + \varepsilon_m) + x_2(\varepsilon_2 + \varepsilon_3 + \cdots + \varepsilon_m) \\
&\quad + x_3(\varepsilon_3 + \cdots + \varepsilon_m) + \cdots + x_m\varepsilon_m \qquad (۴.۵.۱)
\end{aligned}$$

از این فرمول نتیجه می‌گیریم که بهترین نحوهٔ جمع‌کردن، اضافه‌کردن از کوچکترین عنصر به بزرگترین عنصر است. البته مثالهای نقض می‌توان ساخت، ولی برای تعداد زیادی مجموعیابی، قاعدهٔ فوق بایستی بهترین نحوه باشد. این، به‌ویژه وقتی اعداد x_i همگی دارای یک علامت باشند، درست است، زیرا در محاسبهٔ مجموعهای میانی $x_1 + x_2 + \cdots + x_m$، $m = 1, \ldots, n$، هیچ‌گونه حذفی صورت نمی‌گیرد. در این حالت، اگر به جای گردکردن، قطع‌کردن به‌کار برده شود و اگر تمام x_iها مثبت باشند آنگاه هیچ‌گونه حذفی در مجموعهای ε_i اتفاق نمی‌افتد. با نحوهٔ جمع‌کردن از کوچکترین‌ها تا بزرگترین‌ها، اثر این خطاهای قطع کردن را به کمترین مقدار می‌رسانیم.

مثال جمله‌های x_j از مجموع S را به شکل زیر تعریف می‌کنیم: کسر $1/j$ را به کسر اعشاری برگردانده و آن را تا چهاررقم بامعنی گرد می‌کنیم و آن را با x_j نشان می‌دهیم. برای روشنترکردن خطای محاسبهٔ S، حساب اعشاری چهاررقمی با ممیز شناور را به‌کار می‌بریم. جداول ۳.۱ و ۴.۱ نتایج محاسبهٔ S را به چهار طریق متفاوت نشان می‌دهند. جمع‌کردن S از بزرگتر تا کوچکتر با LS و جمع‌کردن از کوچکتر تا بزرگتر با SL نشان داده شده است. در جدول ۳.۱ از حساب قطع شده با

$$-۰٫۰۰۱ \leq \varepsilon_j \leq ۰ \qquad (۵.۵.۱)$$

و در جدول ۴.۱ از حساب گرد شده با

$$-۰٫۰۰۰۵ \leq \varepsilon_j \leq ۰٫۰۰۰۵ \qquad (۶.۵.۱)$$

استفاده شده است. اعداد ε_j به (۴.۵.۱) اشاره دارند و کران آنها از (۸.۲.۱) و (۹.۲.۱) به‌دست می‌آیند.

جدول ۳.۱ محاسبهٔ S در یک ماشین با استفاده از قطع کردن

n	مقدارواقعی	SL	خطا	LS	خطا
۱۰	۲٫۹۲۹	۲٫۹۲۸	۰٫۰۰۱	۲٫۹۲۷	۰٫۰۰۲
۲۵	۳٫۸۱۶	۳٫۸۱۳	۰٫۰۰۳	۳٫۸۰۶	۰٫۰۱۰
۵۰	۴٫۴۹۹	۴٫۴۹۱	۰٫۰۰۸	۴٫۴۷۹	۰٫۰۲۰
۱۰۰	۵٫۱۸۷	۵٫۱۷۰	۰٫۰۱۷	۵٫۱۴۲	۰٫۰۴۵
۲۰۰	۵٫۸۷۸	۵٫۸۴۱	۰٫۰۳۷	۵٫۷۸۶	۰٫۰۹۲
۵۰۰	۶٫۷۹۳	۶٫۶۹۲	۰٫۱۰۱	۶٫۵۶۹	۰٫۲۲۴
۱۰۰۰	۷٫۴۸۶	۷٫۲۸۴	۰٫۲۰۲	۷٫۰۶۹	۰٫۴۱۷

جدول ۴.۱ محاسبهٔ S در یک ماشین با استفاده از گردکردن

n	مقدارواقعی	SL	خطا	LS	خطا
۱۰	۲٫۹۲۹	۲٫۹۲۹	۰٫۰	۲٫۹۲۹	۰٫۰
۲۵	۳٫۸۱۶	۳٫۸۱۶	۰٫۰	۳٫۸۱۷	−۰٫۰۰۱
۵۰	۴٫۴۹۹	۴٫۵۰۰	−۰٫۰۰۱	۴٫۴۹۸	۰٫۰۰۱
۱۰۰	۵٫۱۸۷	۵٫۱۸۷	۰٫۰	۵٫۱۸۷	۰٫۰
۲۰۰	۵٫۸۷۸	۵٫۸۷۸	۰٫۰	۵٫۸۷۶	۰٫۰۰۲
۵۰۰	۶٫۷۹۳	۶٫۷۹۴	−۰٫۰۰۱	۶٫۷۸۳	۰٫۰۱۰
۱۰۰۰	۷٫۴۸۶	۷٫۴۸۶	۰٫۰	۷٫۴۴۹	۰٫۰۳۷

در هر دو جدول فوق‌الذکر روشن است که استراتژی جمع‌کردن S ازکوچکترین عنصر به بزرگترین برجمع‌کردن S از بزرگترین عنصر به کوچکترین ارجحیّت دارد. معهذا آنچه بیشتر اهمیت دارد خطای بسیارکوچکترگردکردن نسبت به قطع کردن است. تفاوت خیلی بیشتر از ۲ برابر است که از آندازهٔ نسبی کرانها از (۵.۵.۱) و (۶.۵.۱) به‌دست می‌آید. ما بعداً تحلیلی از این موضوع می‌آوریم.

یک تحلیل آماری از انتشار خطا مجموع معمولی خطای

$$E = \sum_{j=1}^{n} \varepsilon_j \qquad (۷.۵.۱)$$

از نوعی را که در مجموعیابی خطای (۴.۵.۱) پیش می‌آید در نظر می‌گیریم. یک کران ساده برای آن چنین است:

$$| E | \leq n\delta \qquad (۸.۵.۱)$$

که δ کرانِ $\varepsilon_1, \ldots, \varepsilon_n$ است. پس، در مثال قبلی، برحسب آنکه قطع‌کردن یا گردکردن به‌کار برده شده باشد $\delta = ۰٫۰۰۱$ یا $\delta = ۰٫۰۰۰۵$ است. کران (۸.۵.۱) مربوط به بدترین حالت است که تمام خطاهای ε_j در حد ممکن بزرگ و همه دارای یک علامت‌اند.

هنگامی که گردکردن را به‌کار می‌بریم، رفتار متقارن علامت ε_j، همان‌گونه که در (۹.۲.۱) نشان داده شد، تفاوت فاحشی در اندازهٔ E ایجاد خواهد کرد. در چنین حالتی یک مدل بهتر این است که فرض کنیم خطاهای ε_j متغیرهای تصادفی هستند که در بازهٔ $[-\delta, \delta]$ به شکل یکنواخت

توزیع شده و مستقل‌اند. پس

$$E = n\left(\frac{1}{n}\sum_{1}^{n}\varepsilon_j\right) = n\bar{\varepsilon}$$

میانگین نمونه‌یی $\bar{\varepsilon}$ یک متغیر تصادفی جدید است، که دارای توزیع احتمال با میانگین ۰ و واریانس $\delta^2/3n$ است. برای محاسبهٔ احتمال عباراتی که شامل $\bar{\varepsilon}$ هستند، لازم است توجه داشته باشیم که توزیع احتمال $\bar{\varepsilon}$ با توزیع نرمال با همان میانگین و واریانس، به خوبی تقریب زده می‌شود حتی برای مقادیر کوچکی چون $n \geq 10$. این امر از **قضیهٔ حد مرکزی** در نظریهٔ احتمال به‌دست می‌آید [مثلاً هاگ و کریگ (۱۹۷۸، فصل ۵) را ببینید.]، با استفاده از توزیع نرمال تقریبی، احتمال برقراری

$$|\bar{\varepsilon}| \leq 0.39\delta/\sqrt{n} \qquad |E| \leq 0.39\delta\sqrt{n}$$

برابر ۱/۲ و احتمال برقراری

$$|\bar{\varepsilon}| \leq 1.49\delta/\sqrt{n} \qquad |E| \leq 1.49\delta\sqrt{n} \qquad (9.5.1)$$

برابر ۰٫۹۹ است. نتیجهٔ (۹.۵.۱) در مقایسه با (۸.۵.۱)، وقتی n تا حدی بزرگ باشد بهبود قابل‌توجهی است.

این تحلیل را می‌توان برای خطای قطع‌کردن نیز به‌کار برد. ولی در آن حالت، $-\delta \leq \varepsilon_j \leq 0$. میانگین نمونه‌یی $\bar{\varepsilon}$ اکنون میانگین $\delta/2$ دارد در حالی که واریانس همان $\delta^2/3n$ باقی می‌ماند. بنابراین احتمالی برابر ۰٫۹۹ برای برقراری

$$\left(\frac{n}{2} - 1.49\sqrt{n}\right)\delta \leq -E \leq \left(\frac{n}{2} + 1.49\sqrt{n}\right)\delta \qquad (10.5.1)$$

وجود دارد. برای مقدار بزرگ n، این رابطه به ما اطمینان می‌دهد که E تقریباً برابر $n\delta/2$ است، که بسیار بزرگتر از(۹.۵.۱)، برای حالت گردکردن، است.

هنگامی که این نتایج، یعنی (۹.۵.۱) و (۱۰.۵.۱) برای مجموعیابی معمولی (۴.۵.۱)، به‌کار برده شوند، علّت احتمالی رفتار بسیار متفاوت خطای قطع‌کردن و گردکردن در جدولهای ۳.۱ و ۴.۱ ملاحظه خواهد شد. به طور کلی حساب گردکردن تقریباً همیشه بر حساب قطع‌کردن ارجحیّت دارد.

گرچه تحلیلهای آماری کرانهای واقعیتری را به‌دست می‌دهند، ولی محاسبهٔ آنها بسیار مشکلتر است. برای یک مثال پیچیده‌تر، صفحات ۵۹.۴۱ هنریچی[1] (۱۹۶۲) را برای یک تحلیل آماری

1. Henrici

خطای جواب عددی معادلات دیفرانسیل ببینید. یک مثال در جدول ۳.۶، فصل ۶ این کتاب، داده شده است.

حاصلضربهای داخلی. برای دو بردار $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{R}^m$،

$$x^T y = \sum_{j=1}^{m} x_j y_j \qquad (۱۱.۵.۱)$$

را **حاصلضرب داخلی** x و y می‌نامیم. (نماد x^T ترانهادهٔ ماتریس x را نشان می‌دهد). ویژگیهای حاصلضرب داخلی در فصل ۷ بررسی شده‌اند، ولی ما در اینجا اشاره می‌کنیم که

$$\| x \|_2 = \sqrt{x^T x} = \sqrt{\sum_{j=1}^{m} x_j^2} \qquad (۱۲.۵.۱)$$

$$\| x^T y \| \leq \| x \|_2 \| y \|_2 \qquad (۱۳.۵.۱)$$

نابرابری اخیر **نابرابری کوشی ـ شوارتس** خوانده می‌شود و حالت کلّیتر آن در فصل ۴ اثبات شده است. مجموعهایی به شکل (۱۱.۵.۱) به کرّات در مسائل جبر خطی پیش می‌آیند (برای مثال، ضرب ماتریسها). ما اکنون محاسبات عددی این گونه مجموعها را مورد بررسی قرار می‌دهیم.

فرض می‌کنیم x_i و y_i، $i = 1, \ldots, m$، اعدادی با ممیز شناورند. تعریف می‌کنیم

$$S_1 = \mathrm{fl}(x_1 y_1)$$
$$S_{k+1} = \mathrm{fl}(S_k + \mathrm{fl}(x_k y_k)) \qquad k = 1, 2, \ldots, m-1 \qquad (۱۴.۵.۱)$$

پس مانند قبل، با استفاده از (۱۰.۲.۱)

$$S_1 = x_1 y_1 (1 + \varepsilon_1)$$
$$S_2 = [S_1 + x_2 y_2 (1 + \varepsilon_2)](1 + \eta_2)$$
$$\vdots$$
$$S_m = [S_{m-1} + x_m y_m (1 + \varepsilon_m)](1 + \eta_m)$$

که عبارات ε_j و η_j، برحسب آنکه از قطع‌کردن یا گردکردن استفاده شده باشد به ترتیب در (۸.۲.۱)

و (۹.۲.۱) صدق می‌کنند. از ترکیب و مرتب‌کردن مجدد فرمولهای بالا، به‌دست می‌آوریم

$$S_m = \sum_{j=1}^{m} x_j y_j (1 + \gamma_j) \qquad (۱۵.۵.۱)$$

با

$$1 + \gamma_j = (1 + \varepsilon_j)(1 + \eta_j)(1 + \eta_{j+1}) \dots (1 + \eta_m) \qquad \eta_1 = 0$$
$$\doteq 1 + \varepsilon_j + \eta_j + \eta_{j+1} + \dots + \eta_m \qquad (۱۶.۵.۱)$$

این تقریب آخری بر این اساس است که از حاصلضربهای عبارات کوچک $\eta_i \eta_k$ و $\varepsilon_i \eta_k$ صرف‌نظر شده است. این امر ما را به تحلیل مشابهی که قبلاً برای مجموع (۱.۵.۱) انجام شد می‌رساند. تحلیل آماری خطا که بعد از (۷.۵.۱) آمده نیز معتبر است. برای یک کران مطمئن، می‌توان نشان داد که اگر $m\delta < 0.01$، آنگاه

$$| \gamma_j | \leq 1.01(m + 1 - j)\delta \qquad j = 1, \dots, m \qquad (۱۷.۵.۱)$$

که δ واحد گردکردن است که در (۱۲.۲.۱) داده شد [فورسایت و مولر[1] (۱۹۶۷، ص ۹۲ را ببینید] با به‌کارگرفتن این رابطه در (۱۵.۵.۱) و استفاده از (۱۳.۵.۱)،

$$| S - S_m | \leq \sum_{j=1}^{m} | x_j y_j \gamma_j |$$
$$\leq 1.01m \cdot \delta \, \| x \|_2 \| y \|_2 \qquad (۱۸.۵.۱)$$

این عبارت هیچ چیز در مورد خطای نسبی بیان نمی‌کند، زیرا $x^T y$ می‌تواند صفر باشد ولو اینکه تمام x_i ها و y_i ها مخالف صفر باشند.

این نتایج می‌گویند که خطای مطلق در $S_m \doteq x^T y$ خیلی به سرعت افزایش نمی‌یابد، به‌ویژه اگر از گردکردن واقعی استفاده شده باشد و تحلیل آماری قبلی (۷.۵.۱) را درنظر بگیریم. بهرحال اغلب ممکن است که این خطا را به سادگی و بدون صرف وقت زیاد به مقدار زیادی تقلیل داد، و این معمولاً در مسائل جبر خطی بسیار مهم است.

هر حاصلضرب $x_j y_j$ را با حسابی با دقت بالاتر محاسبه می‌کنیم و مجموعیابی را با همین حساب با دقت بالا انجام می‌دهیم. وقتی که محاسبهٔ مجموع کامل شد، نتیجه را با حساب با دقت معمولی گرد یا قطع می‌کنیم. برای مثال، وقتی، x_i و y_i با دقت معمولی حساب شده باشند، حاصلضربها و مجموعها را با دقت مضاعف محاسبه می‌کنیم. [در بیشتر رایانه‌ها دقت معمولی و

1. Forsythe & Moler

مضاعف از نظر مدّت زمان اجرا نسبتاً بهم نزدیک‌اند؛ معهذا در سخت‌افزار بعضی رایانه‌ها دقت مضاعف گنجانده نشده بلکه فقط در نرم‌افزار آنها گنجانده شده است، که سرعت کمتری دارد.] مجموع به‌دست آمدهٔ S_m در

$$S - S_m \doteq \delta S \qquad (۱۹.۵.۱)$$

صدق می‌کند که بهبود قابل ملاحظه‌ای نسبت به (۱۸.۵.۱) یا (۱۵.۵.۱) است. این کار را می‌توان در بخشی از محاسبات با دقت معمولی به کار برد و دقت را تا حد قابل‌ملاحظه‌ای بهبود بخشید بدون آنکه تمام محاسبات را با دقّت مضاعف انجام داد. در مسائل جبر خطی این امر ممکن است نیاز به حافظه را، در مقایسه با آنچه برای کلّ محاسبات با دقت مضاعف لازم است، نصف سازد.

۶.۱ پایداری در آنالیز عددی

شماری از مسائل ریاضی دارای جوابهایی هستند که در مقابل خطاهای محاسبه، مثلاً خطای گردکردن، کاملاً حساس‌اند. برای مواجهه با این پدیده مفاهیم **پایداری** و **ضریب وضعیت**[۱] را وارد می‌کنیم. ضریب وضعیت یک مسأله رابطهٔ نزدیکی با ماکسیمم دقّتی دارد که در حل آن مسأله با استفاده از اعداد با طول متناهی و حساب رایانه‌یی ممکن است به آن دست یافت. این مفاهیم را سپس برای روشهای عددی که در محاسبهٔ جواب به‌کار می‌روند بسط می‌دهیم. به‌طور کلی علاقه‌مندیم از آن دسته روشهای عددی استفاده کنیم که در مقابل خطاهای کوچک، حساسیت بیشتری، نسبت به آنچه در خود مسألهٔ ریاضی وجود دارد، نداشته باشند.

برای ساده کردن موضوع، بحث را به مسائلی که دارای شکل یک معادلهٔ

$$F(x, y) = ۰ \qquad (۱.۶.۱)$$

هستند محدود می‌کنیم. متغیر x مجهولی است که باید پیدا شود و متغیّر y داده‌ای است که جواب به آن بستگی دارد. این معادله می‌تواند معرّف انواع مختلف مسائل باشد. به‌عنوان مثال، (۱) F می‌تواند تابعی حقیقی ـ مقدار از متغیر حقیقی x باشد و y می‌تواند بردار ضرایب در تعریف F را نشان دهد؛ یا (۲) معادله می‌تواند یک معادلهٔ دیفرانسیل یا انتگرالی باشد، که در آن x تابع مجهول و y یک تابع داده شده یا مقادیر مرزی داده شده است.

مسألهٔ (۱.۶.۱) را **پایدار** گویند اگر جواب x به‌طور پیوسته به متغیّر y بستگی داشته باشد. این بدان معناست که اگر $\{y_n\}$ دنبالهٔ مقادیری باشد که به نحوی به y میل کند، آنگاه مقادیر

۱. در ترجمه، اغلب کتب آنالیز عددی این واژه را با عدد شرط یا عدد وضعیت مترادف دانسته‌اند ولی ما ضریب وضعیت را مناسبتر تشخیص می‌دهیم(م.)

جوابهای متناظر $\{x_n\}$ نیز بایستی به طریقی به x میل نمایند. به عبارتی دیگر، اگر تغییرات جزیی در y ایجاد کنیم، این تغییرات باید تغییرات متناظر کوچکی در x ایجاد کنند. مِلاک کوچک‌بودن این تغییرات به نُرمی بستگی دارد که برای اندازه‌گیری بزرگی بردارهای x و y استفاده شده است؛ انتخابهای مختلفی برای این نُرم‌ها وجود دارند که بسته به مسأله تغییر می‌کنند. مسائل پایدار را **مسائل خوش‌حالت** نیز می‌گویند، و ما از هر دو عبارت به‌عنوان مترادف یکدیگر استفاده می‌کنیم. اگر یک مسأله پایدار نباشد ناپایدار یا بدحالت خوانده می‌شود.

مثال (الف) جوابهای معادلهٔ زیر را در نظر می‌گیریم

$$ax^2 + bx + c = 0 \qquad a \neq 0$$

هر جواب x یک عدد مختلط است. برای داده‌ها در این حالت، از بردار ضرایب $y = (a, b, c)$ استفاده می‌کنیم. از فرمول جواب معادلهٔ درجهٔ دوّم

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

به وضوح دیده می‌شود که هر دو جواب x با داده‌های $y = (a, b, c)$ به‌طور پیوسته تغییر می‌یابند.

(ب) مسألهٔ معادلهٔ انتگرالی زیر را در نظر می‌گیریم

$$\int_0^1 \frac{0.75x(t)dt}{1.25 - \cos(2\pi(s+t))} = y(s) \qquad 0 \leq s \leq 1 \qquad (۲.۶.۱)$$

این یک مسألهٔ ناپایدار است. اختلالاتی چون $\delta_n(s) = y_n(s) - y(s)$ وجود دارند که برای آنها، وقتی که $n \to \infty$،

$$\max_{0 \leq s \leq 1} | \delta_n(s) | \to 0 \qquad n \to \infty \qquad (۳.۶.۱)$$

و جواب متناظر $x_n(s)$، به ازای همهٔ مقادیر $n \geq 1$، در رابطهٔ زیر صدق می‌کند

$$\max_{0 \leq s \leq 1} | x_n(s) - x(s) | = 1 \qquad n \geq 1 \qquad (۴.۶.۱)$$

به‌ویژه، تعریف می‌کنیم $y_n(s) = y(s) + \delta_n(s)$

$$\delta_n(s) = \frac{1}{2^n} \cos(2n\pi s) \qquad 0 \leq s \leq 1 \qquad n \geq 1$$

در این صورت می‌توان نشان داد

$$x_n(s) - x(s) = \cos(۲n\pi s)$$

که (۴.۶.۱) را ثابت می‌کند.

اگر مسألهٔ (۱.۶.۱) ناپایدار باشد مشکلات جدّی برای حل آن وجود دارد. معمولاً ممکن نیست چنین مسائلی را، پیش از پی‌بردن بیشتر به ویژگیهای جواب، حل کرد، که این امر معمولاً از توجه به زمینه‌ای که مسألهٔ ریاضی در آن طرح شده حاصل می‌شود. این موضوع اخیراً زمینهٔ بسیار فعالّی در تحقیقات ریاضیات کاربردی شده است. [برای مثال تیخونوف و ارسنین[1] (۱۹۷۷) و واها[2] (۱۹۸۰) را ببینید].

مسائل زیادی وجود دارند که از لحاظ عملی به تعبیر فوق‌الذکر، پایدارند ولی، از نظر محاسبات آنالیز عددی، کماکان پردردسرند. برای مواجهه با این مشکل، اندازه‌ای برای پایداری معرفی می‌کنیم که **ضریب وضعیت** نامیده می‌شود. این اندازه نشان می‌دهد که از نظر پایداری عملی بعضی از مسائل رفتار بهتری دارند.

ضریب وضعیّت سعی دارد بدترین اثر ممکن روی x، جواب (۱.۶.۱)، را وقتی اختلالات کوچکی در متغیر y ایجاد می‌شود اندازه‌گیری کند. گیریم δy اختلال y و $x + \delta x$ جواب معادلهٔ اختلال‌یافتهٔ

$$F(x + \delta x, y + \delta y) = ۰ \qquad (۵.۶.۱)$$

باشد. تعریف می‌کنیم

$$K(x) = \operatorname*{Sup}_{\delta y} \frac{\| \delta x \| / \| x \|}{\| \delta y \| / \| y \|} \qquad (۶.۶.۱)$$

برای اندازهٔ بزرگی از نماد $\| \cdot \|$ استفاده کرده‌ایم. تعاریف (۱۶.۱.۱) ـ (۱۸.۱.۱) را برای بردارهایی از $\boldsymbol{R^n}$ و $\boldsymbol{C[a, b]}$ یادآوری می‌نمائیم. در مثال (۲.۶.۱) برای اندازه‌گیری اختلالات هردوی x و y از نرم (۱۸.۱.۱) استفاده شده است. معمولاً x و y ممکن است متغیرهایی از انواع مختلف و بنابراین نرم‌های آنها متفاوت باشند. سوپریمم (۶.۶.۱)، روی تمام اختلالات کوچک δy، که برای آنها هنوز مسألهٔ (۵.۶.۱) معنی دارد، گرفته شده است. مسائلی که ناپایدارند به $K(x) = \infty$ می‌انجامند.

1. Tikhonov & Arsenin 2. Wahba

عدد $K(x)$ ضریب وضعیّت (۱.۶.۱) خوانده می‌شود. این عدد اندازهٔ حساسیت جواب x در مقابل تغییرات کوچک در داده‌های y است. اگر $K(x)$ خیلی بزرگ باشد، تغییرات نسبی کوچکی مانند δy برای y وجود دارند که تغییرات نسبی بزرگ δx را برای x موجب می‌شوند. ولی اگر $K(x)$ کوچک باشد، مثلاً $K(x) \leq ۱۰$، آنگاه تغییرات نسبی کوچک در y همیشه تغییرات نسبی کوچکی در x ایجاد خواهند نمود. از آنجا که محاسبات عددی تقریباً همیشه دارای خطاهای کوچک محاسباتی مختلف می‌باشند، مسائل با ضریب وضعیت بزرگ را نمی‌پسندیم. چنین مسائلی بدوضع خوانده می‌شوند و حل دقیق آنها معمولاً بسیار مشکل است.

مثال حل معادلهٔ

$$x - a^y = ۰ \qquad a > ۰ \tag{۷.۶.۱}$$

را در نظر می‌گیریم. با ایجاد اختلالی به اندازه δy در y، داریم

$$\frac{\delta x}{x} = \frac{a^{y+\delta y} - a^y}{a^y} = a^{\delta y} - ۱$$

برای ضریب وضعیتِ (۷.۶.۱)

$$K(x) = \operatorname*{Sup}_{\delta y} \left| \frac{\delta x/x}{\delta y/y} \right| = \operatorname*{Sup}_{\delta y} \left| y \left(\frac{a^{\delta y} - ۱}{\delta y} \right) \right|$$

اگر δy محدود به مقادیر کوچک باشد، داریم

$$K(x) \doteq | \, y \cdot \ln(a) \, | \tag{۸.۶.۱}$$

بدون توجه به اینکه چگونه x را از (۷.۶.۱) محاسبه می‌کنیم، اگر $K(x)$ بزرگ باشد، تغییرات نسبی کوچک در y تغییرات نسبی بزرگی را در x ایجاد می‌نمایند. اگر $K(x) = ۱۰^۴$، و اگر مقدار y مورد استفاده دارای خطای نسبی $۱۰^{-۷}$ باشد، که به لحاظ حساب رایانه‌یی با طول متناهی و گردکردن ایجاد شده است، آنگاه احتمالاً مقدار به‌دست آمدهٔ x خطایی نسبی در حدود $۱۰^{-۳}$ خواهد داشت. این یک افت زیادی در دقت است و به ندرت می‌توان از آن احترازکرد جز اینکه شاید تمام محاسبات را با حساب رایانه‌یی با دقت طولانیتر انجام داد، به شرطی که y نیز با دقت بیشتری به‌دست آید.

مثال ماتریس ناتکین $n \times n$

$$Y = \begin{bmatrix} ۱ & \frac{۱}{۲} & \frac{۱}{۳} & \cdots & \frac{۱}{n} \\ \frac{۱}{۲} & \frac{۱}{۳} & \frac{۱}{۴} & \cdots & \frac{۱}{n+۱} \\ \vdots & & & & \vdots \\ \frac{۱}{n} & \frac{۱}{n+۱} & & \cdots & \frac{۱}{۲n-۱} \end{bmatrix} \qquad (۹.۶.۱)$$

را که **ماتریس هیلبرت** خوانده می‌شود در نظر می‌گیریم. مسألهٔ محاسبهٔ معکوس Y یا معادل آن، حل $YX = I$ که I ماتریس یکه است، یک مسألهٔ خوش حالت است. جواب X با تعداد متناهی عملیات سادهٔ حساب به‌دست می‌آید. ولی مسألهٔ محاسبهٔ X وقتی n بزرگ می‌شود، بتدریج بدوضع می‌شود.

بدوضع بودن معکوس عددی Y به‌طور عملی نشان داده خواهد شد. گیریم $\hat{Y}$ نتیجهٔ واردکردن ماتریس Y در یک رایانهٔ IBM-۳۷۰ باشد و درایه‌های ماتریس با شکل ممیز شناور با دقت معمولی ذخیره شده‌اند. درایه‌های کسری ماتریس در سیستم شانزده‌شانزدهی (پایهٔ ۱۶) بسط داده شده و سپس بعد از ۶ رقم شانزده‌شانزدهی (در حدود ۷ رقم اعشاری) قطع شده‌اند. چون بیشتر درایه‌های Y دارای بسط متناهی شانزده‌شانزدهی نیستند، خطای نسبی هر درایهٔ $\hat{Y}$ در حدود $۱۰^{-۶}$ خواهد بود.

با استفاده از حساب با دقت بالاتر، می‌توانیم مقدار دقیق $\hat{Y}^{-۱}$ را محاسبه کنیم. وارون $Y^{-۱}$ به‌طور تحلیلی معلوم است، و می‌توانیم آن‌را با $\hat{Y}^{-۱}$ مقایسه نماییم. برای $n = ۶$، بعضی درایه‌های $\hat{Y}^{-۱}$ با درایه‌های متناظر خود در $Y^{-۱}$ در اوّلین رقم غیر صفر خود تفاوت دارند. برای مثال، درایه‌های سطر ۶ ستون ۲ چنین‌اند

$$(Y^{-۱})_{۶,۲} = ۸۳۱۶۰٫۰۰ \qquad (\hat{Y}^{-۱})_{۶,۲} = ۷۳۸۶۶٫۳۴$$

این امر، محاسبهٔ $Y^{-۱}$ را به یک مسألهٔ (بدوضع) بدل می‌کند، و این بدوضعی با بزرگ‌شدن n بیشتر می‌شود. بر اثر دقت ضعیفِ $\hat{Y}^{-۱}$ در مقایسه با $Y^{-۱}$، ضریب وضعیّت در (۲.۶.۱) دستِ‌کم $۱۰^{۶}$ خواهد شد. نباید این گونه تصوّر شود که این یک مثال نامعقول غیرعادی است که در عمل رخ نخواهد داد. این مثال خاص در نظریهٔ تقریبِ کمترین مربعات پیش می‌آید (برای مثال بخش ۳.۴ را ببینید). زمینهٔ کلی مسائل بدوضع در دستگاههای خطی و وارونِ ماتریسها با شرح بیشتری در فصل ۸ بررسی شده است.

پایداری الگوریتم‌های عددی. یک روش عددی، برای حل یک مسألهٔ ریاضی هنگامی پایدار خوانده می‌شود که حساسیت جواب عددی نسبت به داده‌ها از آنچه در مسألهٔ ریاضی اصلی وجود

دارد، بیشتر نباشد. ما این موضوع را، مجدداً با استفاده از (۱.۶.۱) به‌عنوان یک مدل برای این مسأله، دقیقتر بیان می‌کنیم. یک روش عددی برای حل (۱.۶.۱) معمولاً به یک دنباله از مسائل تقریبی

$$F_n(x_n, y_n) = ۰ \qquad (۱۰.۶.۱)$$

منجر می‌شود که به پارامتری چون n بستگی دارد. داده‌های y_n وقتی $n \to \infty$ به y میل می‌کنند؛ مقادیر تابع $F_n(z,w)$ برای تمام مقادیر (z,w) نزدیکِ به (x,y)، با $n \to \infty$ به $F(z,w)$ میل خواهند کرد. برای مثال، (۱.۶.۱) را می‌توان یک معادلهٔ دیفرانسیل با مقدار اولیه مجسّم نمود و (۱۰.۶.۱) ممکن است یک دنباله از تقریبات تفاضل ـ متناهی وابسته به $h = \frac{۱}{n}$ را نشان دهد، همان‌گونه که در (۵.۳.۱) و به دنبال آن آمده است. حالت دیگر وقتی است که n نشان‌دهندهٔ تعداد ارقام استفاده‌شده در محاسبات باشد و بخواهیم $F(x,y) = ۰$ را با حساب، در حد ممکن با این دقت متناهی، حل کنیم.

برای هر یک از مسائل (۱۰.۶.۱) می‌توانیم یک ضریب وضعیّت $K_n(x_n)$ درست مانند (۶.۶.۱) تعریف کنیم. با استفاده از این ضرایب وضعیّت، تعریف می‌کنیم

$$\hat{K}(x) = \underset{n\to\infty}{\text{Lim}}\,\underset{k\geq n}{\text{Sup}} K_k(x_k) \qquad (۱۱.۶.۱)$$

گوییم روشِ عددی پایدار است اگر $\hat{K}(x)$ در حدود مقدار $K(x)$ در (۶.۶.۱) باشد، مثلاً اگر

$$\hat{K}(x) \leq ۲K(x)$$

اگر این درست باشد، حساسیّت (۱۰.۶.۱) در مقابل تغییرات داده‌ها در حدود حساسیت مسألهٔ اصلی (۱.۶.۱) خواهد بود.

بعضی مسائل و روشهای عددی به سادگی در چارچوب (۱.۶.۱)، (۶.۶.۱)، (۱۰.۶.۱) و (۱۱.۶.۱) قرار نمی‌گیرند، ولی یک صورت کلّی از مسائل پایدار و ضرایبِ وضعیت را می‌توان تعریف کرد که معنی مشابهی به‌دست خواهد داد. کاربرد اصلی این مفاهیم در این کتاب، در زمینه‌های: (۱) ریشه‌یابی برای معادلات چندجمله‌یی (۲) حل معادلات دیفرانسیل؛ و (۳) مسائلی در جبر خطی عددی است. به‌طورکلی مسألهٔ کمی در مورد روشهای عددی ناپایدار در این کتاب وجود دارد. مشکل اصلی حل مسائل بدوضع خواهد بود.

مثال محاسبهٔ تابع بسلِ

$$x = J_m(y) = \left(\frac{۱}{۲}y\right)^m \sum_{k=۰}^{\infty} \frac{(-\frac{۱}{۴}y^۲)^k}{k!(m+k)!} \qquad m \geq ۰ \qquad (۱۲.۶.۱)$$

را در نظر می‌گیریم. این سری به سرعت همگراست، و محاسبهٔ x به سادگی نشان می‌دهد که این مسأله در وابستگییی که به y دارد خوش‌حالت است.

اکنون محاسبهٔ $J_m(y)$را با استفاده از رابطهٔ بازگشتی سه تایی

$$J_{m+1}(y) = \frac{2m}{y} J_m(y) - J_{m-1}(y) \qquad m \geq 1 \qquad (13.6.1)$$

با فرض معلوم‌بودن $J_0(y)$ و $J_1(y)$، در نظر می‌گیریم. اکنون به طور عددی نشان خواهیم داد که این روش برای محاسبهٔ $J_m(y)$، حتی برای m های نه خیلی بزرگ، یک روش عددی ناپایدار است. y را مساوی ۱ می‌گیریم، که (۱۳.۶.۱) به‌صورت زیر درمی‌آید:

$$J_{m+1}(1) = 2mJ_m(1) - J_{m-1}(1) \qquad m \geq 1 \qquad (14.6.1)$$

از مقادیر $J_0(1)$ و $J_1(1)$ که تا ده رقمِ با معنی دقیق‌اند استفاده کرده‌ایم. مقادیر بعدی $J_m(1)$ از (۴.۶.۱) با حسابِ دقیق محاسبه و نتایج در جدول ۵.۱ داده شده‌اند. مقادیر واقعی برای مقایسه داده شده‌اند، و واگرایی سریع مقادیر تقریبی نسبت به این مقادیر واقعی را نشان می‌دهند. تنها خطاهایی که وجود داشتند مربوط به گردکردن مقادیر $J_0(1)$ و $J_1(1)$ هستند که اختلال بزرگِ فزاینده‌ای در $J_m(1)$ وقتی m بزرگ می‌شود، ایجاد می‌کنند.

استفاده از رابطهٔ بازگشتیِ سه‌جمله‌یی

$$f_{m+1}(x) = a_m(x)f_m(x) - b_m(x)f_{m-1}(x) \qquad m \geq 1$$

جدول ۵.۱ مقادیر محاسبه‌شدهٔ $J_m(1)$

m	مقادیر محاسبه‌شده	مقادیر واقعی
۰	۰٫۷۶۵۱۹۷۶۸۶۶	۰٫۷۶۵۱۹۷۶۸۶۶
۱	۰٫۴۴۰۰۵۰۵۸۵۷	۰٫۴۴۰۰۵۰۵۸۵۷
۲	۰٫۱۱۴۹۰۳۴۸۴۸	۰٫۱۱۴۹۰۳۴۸۴۹
۳	۰٫۱۹۵۶۳۳۵۳۵E − ۱	۰٫۱۹۵۶۳۳۵۳۹۸E − ۱
۴	۰٫۲۴۷۶۶۳۶۲E − ۲	۰٫۲۴۷۶۶۳۸۹۶۴E − ۲
۵	۰٫۲۴۹۷۳۶۱E−۳	۰٫۲۴۹۷۵۷۷۳۰۲E−۳
۶	۰٫۲۰۷۲۴۸E − ۴	۰٫۲۰۹۳۸۳۳۸۰۰E − ۴
۷	−۰٫۱۰۳۸۵E − ۵	۰٫۱۵۰۲۳۲۵۸۱۷E − ۵

یک ابزار متداول در ریاضیات کاربردی و آنالیز عددی است. ولی همان گونه که قبلاً نشان داده شد، ممکن است به روشهای عددی ناپایدار منتهی شوند. برای یک تحلیل کلی از روابط بازگشتی سه‌جمله‌یی به گاوْچی[1] (۱۹۶۷) نگاه کنید. در حالتهای (۱۳.۶.۱) و (۱۴.۶.۱)، خطای از دست‌دادن تعداد زیادی از ارقام با معنی رخ می‌دهد.

بحثِ در آثار خواندنی

دانستن حسابِ رایانه‌یی برای برنامه‌نویسانی که با دقّت عددی سروکار دارند حائز اهمیّت است، به‌ویژه در نوشتن برنامه‌هایی که به طور گسترده مورد استفاده قرار می‌گیرند. همچنین در نوشتن برنامه‌هایی که با رایانه‌های متفاوت اجرا می‌شوند، ویژگیهای مختلف ممیّز شناور در رایانه‌ها بایستی در نظر گرفته شوند. نحوهٔ قدیمیِ برخورد با حساب ممیّزشناور در کنوث[2] (فصل ۴، ۱۹۸۱) و اشتربنتس[3] (۱۹۷۴) داده شده است.

بحث دقیق در موضوع انتشار خطا که سودمند باشد، به‌ویژه در مورد خطای گردکردن/قطع‌کردن، مشکل بوده است. بعضی مقالات قدیمی مهم در این مورد وجود دارند، ولی گرایشهای جاری به این موضوع به مقدار زیادی مدیون کارهای شادروان ویلکینسن است. بیشتر کارهای او در جبر خطی عددی بوده است ولی سهم مهمّی در بسیاری از زمینه‌های آنالیز عددی داشته است. برای یک آشنایی کلی با روشهای او در تحلیل انتشار خطا، همراه با کاربرد در چندین مسألهٔ مهمّ به [ویلکینسن (۱۹۶۳)، (۱۹۶۵)، (۱۹۸۴)] نگاه کنید.

روش دیگر کنترل خطا، **تحلیل بازه‌یی** خوانده می‌شود. در این روش، به جای اینکه محاسبات را در یک عدد x_A، دنبال کنیم آن را در طول یک بازهٔ $[x_l, x_u]$ دنبال می‌کنیم و اعداد x_l و x_u حدود مقدار واقعی x_T را تضمین می‌کنند. مشکل این روش آن است که $x_u - x_l$ معمولاً بسیار بزرگتر از $|x_T - x_A|$ است، عمدتاً بدین‌علّت که حذف خطاهای مختلف‌العلامه که ممکن است در محاسبهٔ x_l و x_u پیش بیایند در نظر گرفته نمی‌شوند. برای آشنایی با این زمینه، و اینکه چگونه می‌توان این کرانها را در حالتهای خاص بهبود بخشید به مور[4] (۱۹۶۶) نگاه کنید. اخیراً، این زمینه با زمینهٔ حسابِ رایانه‌یی بهم آمیخته شده است تا یک چارچوب نظری کلّی برای عرضهٔ الگوریتمهایی با کرانهای دقیق خطا را ممکن نماید. برای مثالهایی در این زمینه کتابهای درسی آلِفلت و هرتسبرگر[5] (۱۹۸۳) و کولیش و میرانکر[6] (۱۹۸۱) و گزارشهای سمپوزیوم آلِفلت و گریگوریف[7] (۱۹۸۰) و برآوردهای مور (۱۹۷۹) را ببینید.

1. Gautschi 2. Knuth 3. Sterbenz 4. Moore
5. Alefeld and Herzberger 6. Kulisch and Miranker
7. Alefeld and Grigoreiff

به مسائل بدحالت در بخش ۶.۱ اشاره‌ای شد، ولی این موضوع در سالهای اخیر مورد توجه قرار گرفته است. مسائل زیادی در اندازه‌گیریهای غیرمستقیم فیزیکی وجود دارند که به مسائل بدحالت منجر می‌شوند که در این شکل، **مسائل معکوس** خوانده می‌شوند. کتاب لاورنتیف[1] (۱۹۶۷) یک مقدمهٔ کلی برای این مسائل به دست می‌دهد، اگر چه عمدتاً دربارهٔ (۱) ادامهٔ تحلیلی توابع تحلیلی با یک متغیر مختلط، و (۲) مسائل معکوس در معادلات دیفرانسیل، بحث می‌کند. یکی از ابزارهای اصلی عددی که در مواجهه با مسائل بدحالت مورد استفاده قرار می‌گیرد، **مرتب‌سازی** خوانده می‌شود، و بسط جامعی از این روش درکتاب تیخونوف و آرسنین[2] (۱۹۷۷) داده شده است. برای رایجترین نوشته‌های روشهای عددی در مسائل بدحالت مثلاً به گروچ[3] (۱۹۸۴) و وابا (۱۹۸۰) نگاه کنید.

دو نوع جدید رایانه در ده پانزده سال اخیر پیدا شده‌اند که اکنون نقش بیسابقه و فزاینده‌ای در آنالیز عددی پیدا کرده‌اند. اینها **ریزرایانه‌ها** و **زَبَررایانه‌ها** هستند. همه از ریزرایانه‌ها مطّلع‌اند؛ دانشمندان و مهندسان به مقدار زیادی در محاسبات عددی خود از آنها استفاده می‌کنند. در ابتدا طرح حسابی در ریزرایانه‌ها نسبتاً ضعیف بود؛ و خطاهایی در عملیات حساب اصلی آنها وجود داشت. اخیراً استانداردِ جدید بسیار خوبی برای حساب در ریزرایانه‌ها تولید شده است، که با آن می‌توان برنامه‌هایی عددی در سطح بالا و کارآمد نوشت. این استاندارد یعنی **استاندارد برای حساب با ممیز شناور دودویی** IEEE، در IEEE (۱۹۸۵) توضیح داده شده است. پیاده‌سازی این استاندارد در بیشتر ریزپردازها شروع شده است، برای مثال، پالمر و مورس[4] (۱۹۸۴) را نگاه کنید.

نام زَبَررایانه به نوعی از ماشینهای طرّاحی‌شده اطلاق می‌شود که همگی قادرند محاسبات عددی را با سرعت بسیار زیاد مثلاً بیشتر از ۲۰ میلیون عمل با ممیز شناور در ثانیه را انجام دهند. این زمینه در حال گسترش است و خیلی سریع تغییر می‌کند، و لذا ما می‌توانیم فقط چند مرجع را که اشاره‌ای به اثر این ماشینها روی طراحی الگوریتم‌های عددی دارند نام ببریم. هاکنی و جسهوپ[5](۱۹۸۱) و کویین[6] (۱۹۸۷) کتابهای عمومی در معماری زَبَررایانه‌ها و طراحی الگوریتم‌های عددی در آنها، هستند. پارتر[7] (۱۹۸۴) یک گزارش سمپوزیوم است که کاربردهایی از زَبررایانه‌ها را در مسائل گوناگون فیزیک به‌دست می‌دهد؛ و اورتگا و فویکت[8] (۱۹۸۵) دربارهٔ زَبررایانه‌ها وقتی برای حل معادلات دیفرانسیل جزئی به‌کار برده می‌شوند، بحث می‌کند. این ماشینها بتدریج در تمام زمینه‌های محاسبات، اهمیّت فزاینده‌ای پیدا می‌کنند و معماری آنها محتملاً بر روی رایانه‌های معمولی کوچکتر که در حال حاضر بسیار مورد استفاده قرار می‌گیرند، اثر خواهد گذاشت.

1. Lavrentiev 2. Tikhonov and Arsenin 3. Groetsch
4. Palmer& Morse 5. Hockney and Jesshope 6. Quinn
7. Parter 8. Ortega and Voigt

ریاضیات نمادی. زمینه‌ای است که به سرعت رشد می‌کند و با آن می‌توان به جای ریاضیات عددی از روش ریاضیات تحلیلی، مثلاً در پیداکردن تابع اولیهٔ دقیق، اگر وجود داشته باشد استفاده کرد. این زمینه تاکنون اثر قابل ملاحظه‌ای روی آنالیز عددی نداشته است، ولی به نظر می‌آید که این وضعیّت در حال تغییر است، در بسیاری حالات، ریاضیات نمادی در بخشی از محاسبات به‌کار برده می‌شود و برای بقیهٔ محاسبات از روشهای عددی استفاده می‌شود. یکی از پیچیده‌ترین زبانهای برنامه‌نویسی برای اجرای ریاضیات نمادی زبان MACSYMA است که در رَند (۱۹۸۴) توضیح داده شده است. برای یک بازنگری و بررسی تاریخی زبانهای برنامه‌نویسی در این زمینه به وَن هولتسن و کالمت نگاه کنید.

ما این مبحث را با بحث دربارهٔ **نرم‌افزار ریاضی** خاتمه می‌دهیم. این زمینه به بحث دربارهٔ به‌کارگیری الگوریتمهای عددی به‌صورت برنامه‌های کامپیوتری می‌پردازد، با توجهی دقیق به مسائل دقت، کارایی، انعطاف‌پذیری، قابلیّت حمل، و سایر ویژگیهایی که سودمندی برنامه‌ها را تأمین می‌نماید. یک مجلهٔ مهم در این زمینه ACM در نرم‌افزار ریاضی است. برای یک مطالعهٔ همه‌جانبه در این موضوع، از جمله نوشته‌های مهم برنامه‌یی که در سالهای اخیر توسعه یافته‌اند، به کاول (۱۹۸۴) نگاه کنید. در پیوست این کتاب دربارهٔ بعضی بسته‌های نرم‌افزاری که اخیراً برای آنالیز عددی در دسترس قرار گرفته‌اند، توضیح بیشتری داده‌ایم.

مراجع

Alefeld, G., and R. Grigorieff, eds. (1980). *Fundamentals of Numerical Computation (Computer-oriented Numerical Analysis)*. Computing Supplementum 2, Springer-Verlag, Vienna.

Alefeld, G., and J. Herzberger (1983). *Introduction to Interval Computations*. Academic Press, New York.

Bender, E. (1978). *An Introduction to Mathematical Modelling*. Wiley, New York.

Cowell, W., ed. (1984). *Sources and Development of Mathematical Software*. Prentice-Hall, Englewood Cliffs, N.J.

Fadeeva, V. (1959). *Computational Methods of Linear Algebra*. Dover, New York.

Forsythe, G. (1969). What is a satisfactory quadratic equation solver? In B. Dejon and P. Henrici (eds.), *Constructive Aspects of the Fundamental Theorem of Algebra*, pp. 53–61, Wiley, New York.

Forsythe, G., and C. Moler (1967). *Computer Solution of Linear Algebraic Systems*. Prentice-Hall, Englewood Cliffs, N.J.

Forsythe, G., M. Malcolm, and C. Moler (1977). *Computer Methods for Mathematical Computations*. Prentice-Hall, Englewood Cliffs, N.J.

Gautschi, W. (1967). Computational aspects of three term recurrence relations.

SIAM Rev., 9, 24–82.

Groetsch, C. (1984). *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. Pitman, Boston.

Henrici, P. (1962). *Discrete Variable Methods in Ordinary Differential Equations*. Wiley, New York.

Hockney, R., and C. Jesshope (1981). *Parallel Computers: Architecture, Programming, and Algorithms*. Adam Hilger, Bristol, England.

Hogg, R., and A. Craig (1978). *Introduction to Mathematical Statistics*, 4th ed. Macmillan, New York.

Institute of Electrical and Electronics Engineers (1985). Proposed standard for binary floating-point arithmetic. (IEEE Task P754), Draft 10.0. IEEE Society, New York.

Knuth, D. (1981). *The Art of Computer Programming*, vol. 2, *Seminumerical Algorithms*, 2nd ed. Addison-Wesley, Reading, Mass.

Kulisch, U., and W. Miranker (1981). *Computer Arithmetic in Theory and Practice*. Academic Press, New York.

Lavrentiev, M. (1967). *Some Improperly Posed Problems of Mathematical Physics*. Springer-Verlag, New York.

Lin, C., and L. Segal (1974). *Mathematics Applied to Deterministic Problems in the Natural Sciences*. Macmillan, New York.

Maki, D., and M. Thompson (1973). *Mathematical Models and Applications*. Prentice-Hall, Englewood Cliffs, N.J.

Moore, R. (1966). *Interval Analysis*. Prentice-Hall, Englewood Cliffs, N.J.

Moore, R. (1979). *Methods and Applications of Interval Analysis*. Society for Industrial and Applied Mathematics, Philadelphia.

Ortega, J., and R. Voigt (1985). Solution of partial differential equations on vector and parallel computers. *SIAM Rev.*, 27, 149–240.

Parter, S., ed. (1984). *Large Scale Scientific Computation*. Academic Press, New York.

Palmer, J., and S. Morse (1984). *The 8087 Primer*. Wiley, New York.

Quinn, M. (1987). *Designing Efficient Algorithms for Parallel Computers*. McGraw-Hill, New York.

Rand, R. (1984). *Computer Algebra in Applied Mathematics: An Introduction to MACSYMA*. Pitman, Boston.

Rubinow, S. (1975). *Introduction to Mathematical Biology*. Wiley, New York.

Sterbenz, P. (1974). *Floating-Point Computation*. Prentice-Hall, Englewood Cliffs, N.J.

Tikhonov, A., and V. Arsenin (1977). *Solutions of Ill-posed Problems*. Wiley, New York.

Van Hulzen, J., and J, Calmet (1983). Computer algebra systems. In B. Buchberger, G. Collins, R. Loos (eds.), *Computer Algebra: Symbolic and Algebraic Computation*, 2nd ed. Springer-Verlag, Vienna.

Wahba, G. (1980). Ill-posed problems: Numerical and statistical methods for mildly, moderately, and severely ill-posed problems with noisy data. Tech. Rep. #595, Statistics Department, Univ. of Wisconsin, Madison. Prepared for the *Proc. Int. Symp. on Ill-Posed Problems*, Newark, Del., 1979.

Wahba, G. (1984). Cross-validated spline methods for the estimation of multivariate functions from data on functionals, in *Statistics: An Appraisal*, H. A. David and H. T. David (eds.), Iowa State Univ. Press, Ames, pp. 205–235.

Wilkinson, J. (1963). *Rounding Errors in Algebraic Processes*. Prentice-Hall, Englewood Cliffs, N.J.

Wilkinson, J. (1965). *The Algebraic Eigenvalue Problem*. Oxford, England.

Wilkinson, J. (1984). The perfidious polynomial. In Golub (ed.), *Studies in Numerical Analysis*, pp. 1–28, Mathematical Association of America.

مسائل

۱. (الف) گیریم $f(x)$ بر $a \le x \le b$ پیوسته باشد و میانگین

$$S = \frac{1}{n}\sum_{j=1}^{n} f(x_j)$$

را برای تمام نقاط x_j در بازهٔ $[a,b]$ در نظر می‌گیریم. نشان دهید برای مقداری چون $\zeta \in [a,b]$،

$$S = f(\zeta)$$

راهنمایی: از قضیهٔ مقدار میانی استفاده کنید و برد مقادیر $f(x)$ و بنابراین برد S را در نظر بگیرید.

(ب) قسمت (الف) را برای مجموع

$$S = \sum_{j=1}^{n} w_j f(x_j)$$

به ازای تمام x_jها در $[a,b]$ و تمام $w_j \ge 0$ها تعمیم دهید.

۲. نابرابریهای زیر را نتیجه بگیرید

(الف) برای تمام مقادیر $x, z \le 0$ $\quad |\, e^x - e^z \,| \le |\, x - z \,|$

(ب) برای $-\frac{\pi}{2} < x, z \le \frac{\pi}{2}$ $\quad |\, x - z \,| \le |\, \tan(x) - \tan(z) \,|$

(ج) برای $p \ge 1$ و $0 \le y \le x$ $\quad py^{p-1}(x-y) \le x^p - y^p \le px^{p-1}(x-y)$

۳. (الف) کرانِ خطا در تقریب

$$\sin(x) \doteq x \qquad |\, x \,| \leq \delta$$

را به‌دست آورید.

(ب) برای مقادیر کوچک δ، خطای نسبی در $\sin(x) \doteq x$ را با استفاده از

$$\frac{\sin(x) - x}{\sin(x)} \doteq \frac{\sin(x) - x}{x} \qquad x \neq ۰$$

اندازه بگیرید. کران این خطای نسبی تعدیل یافته را برای $|\, x \,| \leq \delta$ به‌دست آورید. δ را به گونه‌ای پیدا کنید که این خطا کوچکتر از ۰٫۰۱ شود که متناظر است با خطای یک درصد.

۴. فرض می‌کنیم $g \in C[a,b]$، نشان دهید برای مقداری از ζ در $[a,b]$

$$\int_{۰}^{h} x^{۲}(h-x)^{۲} g(x)dx = \frac{h^{۵}}{۳۰} g(\zeta)$$

۵. برای توابع زیر یک سری تیلر بسازید و کران خطای برشی پس از n جمله را به‌دست آورید.

(الف) $\sin^{-۱}(x), \quad |\, x \,| < ۱$

(ب) $\frac{۱}{x}\int_{۰}^{x} e^{-t^{۲}} dt$

(ج) $\cos(x) + \sin(x)$

(د) $\frac{۱}{x}\int_{۰}^{x} \frac{\tan^{-۱} t dt}{t}$

(ه) $\log\left[\frac{۱+x}{۱-x}\right], \quad -۱ < x < ۱$

(و) $\log(۱-x) \quad -۱ < x < ۱$

۶. (الف) با استفاده از نتیجهٔ (۱۱.۱.۱)، می‌توانیم نشان دهیم

$$\frac{\pi}{۴} = \tan^{-۱}(۱) = \sum_{j=۰}^{\infty} \frac{(-۱)^{۲j+۱}}{۲j+۱}$$

و می‌توانیم π را با ضرب آن در ۴ بدست آوریم. چرا این یک راه عملی برای محاسبهٔ π نیست؟

(ب) با به‌کار بردن تقریب چندجمله‌یی تیلر یک راه عملی برای محاسبهٔ π معرفی کنید.

۷. با به‌کار بردن قضیهٔ تیلر برای تابعهای دو متغیره، تقریبات خطی و درجهٔ دوّم را برای تابعهای $f(x,y)$ داده‌شده به ازای مقادیر کوچک x و y پیدا کنید. تابع صفحهٔ مماس $z = p(x,y)$ را که نمودار آن در نقطهٔ $(۰,۰,f(۰,۰))$ بر $z = f(x,y)$ مماس است معین کنید

(الف) $\sqrt{۱ + ۲x - y}$

(ب) $\frac{(۱+x)}{(۱+y)}$

(ج) $x \cdot \cos(x-y)$

(د) $\cos(x + \sqrt{\pi^{۲} + y})$

۸. تفاضل منقسم مرتبهٔ دوّم $f[x_۰, x_۱, x_۲]$ را که در (۱۳.۱.۱) تعریف شد در نظر می‌گیریم.

(الف) ویژگی (۱۵.۱.۱) را که می‌گوید، مرتبهٔ شناسه‌های x_0، x_1، x_2 بر مقدار تفاضل منقسم اثری ندارد، ثابت کنید.
(ب) فرمول (۱۴.۱.۱)

$$f[x_0, x_1, x_2] = \frac{1}{2} f''(\zeta)$$

را که در آن ζ مقداری بین مینیمم و ماکسیمم x_0، x_1، x_2 است ثابت کنید.
راهنمایی: در قسمت (الف) بدون کاستن از کلیّت آن می‌توان فرض کرد $x_0 < x_1 < x_2$. از قضیهٔ تیلر برای تبدیل $f[x_0, x_1, x_2]$ که حول x_1 بسط داده شده و از قضیهٔ مقدار میانی برای ساده‌کردن جملهٔ خطا استفاده کنید.
(ج) با فرض آنکه $f(x)$ دوبار پیوسته ـ مشتقپذیر باشد، نشان دهید که $f[x_0, x_1, x_2]$ می‌تواند به طور پیوسته به حالتی تعمیم یابد که بعضی یا همهٔ نقاط x_0، x_1، x_2 بر هم منطبق شوند. برای مثال نشان دهید که

$$f[x_0, x_1, x_0] \equiv \lim_{x_2 \to x_0} f[x_0, x_1, x_2]$$

موجود است و فرمولی برای محاسبهٔ آن پیدا کنید.

۹. (الف) نشان دهید که نرمهای برداری (۱۶.۱.۱) و (۱۸.۱.۱) در سه ویژگی عمومی نرمها که به دنبال (۱۸.۱.۱) بیان شده‌اند صدق می‌کنند.
(ب) نشان دهید $\| x \|_2$ در (۱۷.۱.۱) یک نرم برداری است، خود را به $n = 2$ محدود کنید.
(ج) نشان دهید نرم ماتریسی (۱۹.۱.۱) در روابط (۲۰.۱.۱) و (۲۱.۱.۱) صدق می‌کند. برای سادگی فقط ماتریسهای 2×2 را در نظر بگیرید.

۱۰. اعداد زیر را به هم‌ارزهای اعشاری آنها برگردانید.

(الف) $(1010.101)_2$ (ب) $(2A30.FF)_{16}$
(ج) $(0.1010101\ldots)_2$ (د) $(0.AAAA\ldots)_{16}$
(ه) $(0.0001100110011\ldots)_2$
(و) $(11,\ldots1)_2$ که n تا یک در پرانتز باشد.

۱۱. برای تبدیل یک عدد صحیح اعشاری مثبت x به هم‌ارز دودویی آن،

$$x = (a_n a_{n-1} \ldots a_1 a_0)_2$$

چنین می‌نویسیم

$$x = a_n \cdot ۲^n + a_{n-۱} \cdot ۲^{n-۱} + \cdots + a_۱ \cdot ۲^۱ + a_۰ \cdot ۲^۰$$

بر این اساس الگوریتم زیر را به‌کار برید

(i) $x_۰ := x$؛ $j := ۰$

(ii) تا وقتی $x_j \neq ۰$، عملیّات زیر را انجام دهید:

a_j را برابر باقیماندهٔ عدد صحیح $x_j/۲$ قرار دهید

$x_{j+۱}$ را برابر خارج قسمت عدد صحیح $x_j/۲$ قرار دهید

j را برابر $j + ۱$ قرار دهید

پایان حلقهٔ تا وقتی

زبان الگوریتم باید نیاز به توضیح نداشته باشد. از این الگوریتم برای تبدیل اعداد صحیح زیر به هم‌ارز دودویی آنها استفاده کنید.

(الف) ۴۹ **(ب)** ۱۲۷ **(ج)** ۱۲۹

۱۲. برای تبدیل یک کسر مثبت اعشاری $x < ۱$ به هم‌ارز دودویی آن

$$x = (۰.a_۱a_۲a_۳\ldots)_۲$$

آن‌را به شکل زیر بنویسید

$$x = a_۱ \cdot ۲^{-۱} + a_۲ \cdot ۲^{-۲} + a_۳ \cdot ۲^{-۳} + \ldots$$

بر این اساس الگوریتم زیر را به‌کار برید

(i) $x_۱ := x$؛ $j := ۱$

(ii) تا وقتی $x_j \neq ۰$، عملیات زیر را انجام دهید:

a_j را برابر قسمت صحیح $x_j \cdot ۲۰$ قرار دهید

$x_{j+۱}$ را برابر جزء کسری $x_j \cdot ۲۰$ قرار دهید

j را برابر $j + ۱$ قرار دهید

پایان حلقهٔ تا وقتی

این الگوریتم را برای تبدیل اعداد اعشاری زیر به هم‌ارز دودویی آنها به‌کار برید

(الف) ۰٫۸۱۲۵ **(ب)** ۱۲٫۰۶۲۵ **(ج)** ۰٫۱ **(د)** ۰٫۲

(ه) ۰٫۴ **(و)** $\frac{۱}{۷} = ۰٫۱۴۲۸۵۷۱۴۲۸۵۷\ldots$

۱۳. مسائل ۱۱ و ۱۲ را برای تبدیل یک عدد صحیح اعشاری به هم‌ارز شانزده‌شانزدهی آنها به‌کار برید.

۱۴. خروجی قطعه برنامهٔ فورترن زیر را با یک رایانهٔ دودویی با استفاده از حساب قطع‌شده، پیش‌بینی کنید

```
      I = ۰
      X = ۰٫۰
      H = ۰٫۱
۱۰    I=I+1
      X = X + H
      PRINT*, I, X
      IF(X .LT. ۱٫۰) Go To ۱۰
```

اگر عبارت "$X = X + H$" با عبارت "$X = I * X$" عوض شود آیا خروجی تغییر می‌کند؟

۱۵. کرانهای (۹.۲.۱) را برای خطای نسبی در نمایش گرد شده با ممیز شناورِ (۵.۲.۱) ـ (۶.۲.۱) به‌دست آورید.

۱۶. نتیجهٔ کران بالای $M = \beta^t$ در (۱۶.۲.۱) را به‌دست آورید.

۱۷. (الف) برنامه‌ای بنویسید که یک خطای سرریز در رایانهٔ شما تولید کند. برای مثال، عدد x را به‌عنوان ورودی بگیرید و آن‌را چندبار به توان دو برسانید.

(ب) برنامه‌ای بنویسید که به طور تجربی بزرگترین عدد با ممیز شناور مجاز را معین کند.

۱۸. (الف) یک مدل ساده برای افزایش جمعیت چنین است

$$\frac{dN}{dt} = kN \qquad t \geq t_0, \qquad N(t_0) = N_0$$

که $N(t)$ جمعیت در زمان t است و $k > 0$. نشان دهید که این رابطه مستلزم یک نرخ هندسی افزایش در جمعیت است:

$$N(t+1) = CN(t) \qquad t \geq t_0$$

فرمولی برای C پیدا کنید.

(ب) یک مدل پیچیده‌تر برای رشد جمعیت چنین است

$$\frac{dN}{dt} = kN[1 - bN] \qquad t \geq t_0, \qquad N(t_0) = N_0$$

که $b, k > ۰$ و $۱ - bN_۰ > ۰$. جواب این معادله دیفرانسیل را پیدا کنید. این جواب را با جواب قسمت (الف) مقایسه کنید. تفاوتهایی را که در افزایش جمعیت در دو مدل پیش می‌آید برای هر دو حالت مقادیر بزرگ و کوچک t شرح دهید.

۱۹. دو تابع زیر را با رایانهٔ خود محاسبه کنید

(الف) $f(x) = x^۳ - ۳x^۲ + ۳x - ۱$

(ب) $f(x) = x^۳ + ۲x^۲ - x - ۲$

آنها را برای نمونهٔ بزرگی از مقادیر x حول ۱ محاسبه کنید و سپس نوع رفتار آنها را آن‌گونه که در شکل ۲.۱ نشان داده شده به‌دست آورید. نتایج به‌دست آمده برای دو تابع را با هم مقایسه کنید.

۲۰. برنامه‌ای بنویسید که به طور تجربی حد زیر را محاسبه کند

$$\lim_{p \to \infty} (x^p + y^p)^{۱/p}$$

در اینجا x و y اعداد مثبت‌اند. اوّلاً محاسبات را در شکلی که هم‌اکنون نشان داده شد انجام دهید. ثانیاً محاسبات را به روشی که در (۸.۳.۱) به‌کار رفته تکرار کنید. برنامه را برای انواع مقادیر بزرگ و کوچک x و y، مثلاً $x = y = ۱۰^{۱۰}$ و $x = y = ۱۰^{-۱۰}$ اجرا کنید.

۲۱. برای اعداد x_A و x_T در زیر، چند رقم با معنی در x_A با توجّه به x_T وجود دارد؟

(الف) $x_A = ۴۵۱٫۰۲۳$، $x_T = ۴۵۱٫۰۱$

(ب) $x_A = -۰٫۰۴۵۱۱۳$، $x_T = -۰٫۰۴۵۱۸$

(ج) $x_A = ۲۳٫۴۲۱۳$، $x_T = ۲۳٫۴۶۰۴$

۲۲. گیریم تمام اعداد زیر به ارقامی که در اعداد نشان داده شده دقیقاً گرد شده باشند

(الف) ۰٫۹۴۷ + ۱٫۱۰۶۲، (ب) ۱۲٫۷۵۳ − ۲۳٫۴۶، (ج) (۶٫۸۳) × (۲٫۷۴۷)

(د) ۸٫۴۷۳/۰٫۰۶۴. برای هر محاسبه، کوچکترین بازه‌ای را معین کنید که جواب در آن قرار بگیرد، وقتی از مقادیر درست به جای مقادیر تقریبی استفاده شود.

۲۳. فرمول (۵.۴.۱) را برای $\text{Rel}\,(x_A/y_A)$ اثبات کنید.

۲۴. معادلهٔ $x^۲ - ۴۰x + ۱ = ۰$ داده شده است، ریشه‌های آن را تا ۵ رقم با معنی پیدا کنید.

۲۵. راههای درست جلوگیری از خطای از دست دادن ارقام بامعنی را در محاسبات زیر مشخص کنید

(الف) $\log(۱ + x) - \log(x)$ مقادیر بزرگ x

(ب) $\sin(x) - \sin(y)$ $\quad x \doteq y$

(ج) $\tan(x) - \tan(y)$ $\quad x \doteq y$

(د) $\dfrac{1 - \cos(x)}{x^2}$ $\quad x \doteq 0$

(ه) $\sqrt[3]{1+x} - 1$ $\quad x \doteq 0$

۲۶. برای احتراز از خطای از دست‌دادن ارقام بامعنی در محاسبات زیر از تقریبهای تیلر استفاده کنید.

(الف) $f(x) = \dfrac{e^x - e^{-x}}{2x}$

(ب) $f(x) = \dfrac{\log(1-x) + xe^{x/2}}{x^3}$

در هر دو حالت مقدار $\text{Lim}_{x\to 0} f(x)$ چیست؟

۲۷. ارزیابی $\cos(x)$ را برای مقادیر بزرگ x با استفاده از تقریب تیلر (۵.۱.۱)

$$\cos(x) \doteq 1 - \frac{x^2}{2!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!}$$

در نظر می‌گیریم. برای مشاهدهٔ مشکلاتی که در استفاده از این تقریب پیش می‌آید، آن را برای محاسبهٔ $\cos(2\pi) = 1$ به‌کار برید. n را به گونه‌ای بیابید که خطای تقریب تیلر کمتر از ۰٫۰۰۰۵ باشد. نوع محاسبه‌ای را که در (۱۷.۴.۱) و جدول ۲.۱ به‌کار رفته تکرار کنید. چگونه $\cos(x)$ برای مقادیر بزرگ x باید ارزیابی شود؟

۲۸. فرض کنید می‌خواهید مقادیر:

(الف) $\cos(1{,}473)$، (ج) $(1{,}471)$،

(ب) $\tan^{-1}(2{,}621)$ ، (د) $e^{2{,}653}$

را محاسبه کنید. در هر حالت، فرض می‌کنیم فقط جدولی از مقادیر تابع برحسب شناسهٔ x که با نموهای ۰٫۰۱ داده شده‌اند در دست باشد. مقداری از جدول را که شناسهٔ آن نزدیکترین مقدار را به شناسه‌های شما داشته باشد انتخاب کنید. خطای نتیجه را تخمین بزنید.

۲۹. فرض کنید $x_A = 0{,}937$ سه رقم صحیح با معنی با توجّه به x_T دارد. کران خطای نسبی x_A را به‌دست آورید. برای $f(x) = \sqrt{1-x}$، کران خطا و خطای نسبی در $f(x_A)$ را با توجّه به $f(x_T)$ پیدا کنید.

۳۰. اعداد زیر تا ارقامی که نشان داده شده‌اند به‌طور دقیق گرد شده‌اند. خطاهای مقادیر توابع را برحسب خطاهای شناسه‌های آنها تخمین بزنید. خطاهای نسبی را تخمین بزنید. خطاهای نسبی را کراندار کنید.

(الف) $\sin[(3{,}14)(2{,}685)]$ (ب) $\ln(1{,}712)$ (ج) $(1{,}56)^{3{,}414}$

۳۱. یک برنامهٔ رایانه‌یی برای مجموع

$$S = \sum_{۱}^{n} a_i$$

به سه طریق بنویسید (۱) ازکوچکترین مقدار به بزرگترین مقدار، (۲) از بزرگترین مقدار به کوچکترین مقدار (۳) با دقت مضاعف، با دقت معمولی خطای گردکردن/قطع کردن در نتیجهٔ مجموعیابی. نتیجهٔ دقت مضاعف را برای پیداکردن خطا در دو نتیجهٔ با دقت معمولی فوق به‌کار برید. نتایج را چاپ کنید. همچنین یک برنامهٔ کلی بنویسید که سری زیر را با دقت معمولی تولید کند، و برنامه‌ای را که هم‌اکنون داده شد برای محاسبهٔ مجموع سریهای زیر به‌کار برید.

راهنمایی: در نوشتن برنامه، به خاطر سادگی فرض کنید که جملات سری از بزرگ به کوچک مرتب شده‌اند.

(الف) $\sum_{۱}^{n} \frac{۱}{j}$، (ب) $\sum_{۱}^{n} \frac{۱}{j^۲}$، (ج) $\sum_{۱}^{n} \frac{۱}{j^۳}$، (د) $\sum_{۱}^{n} \frac{(-۱)^j}{j}$

۳۲. حاصلضرب $a_۰a_۱ \dots a_m$ را در نظر بگیرید که در آن $a_۰$، $a_۱$، ...، a_m، $(m+۱)$ عدد ذخیره‌شده در رایانه‌ای هستند که از حساب n رقمی در پایهٔ β استفاده می‌کند. تعریف می‌کنیم $p_۱ = \mathrm{fl}(a_۰a_۱)$، $p_۲ = \mathrm{fl}(a_۲p_۱)$، $p_۳ = \mathrm{fl}(a_۳p_۲)$، ...، $p_m = \mathrm{fl}(a_mp_{m-۱})$. اگر بنویسیم $p_m = a_۰a_۱ \dots a_m(۱+w)$، یک برآورد برای w تعیین و فرض کنید $a_i = \mathrm{fl}(a_i)$، به ازای $i = ۰, ۱, \dots, m$. یک کران دقیق برای w چیست؟ برآورد آماری برای اندازهٔ w چیست؟

۲

ریشه‌یابی معادله‌های غیرخطی

پیدا کردن یک یا چند ریشهٔ یک معادلهٔ

$$f(x) = ۰ \qquad (۱.۰.۲)$$

یکی از متداولترین مسائلی است که در ریاضیات کاربردی پیش می‌آید. در بیشتر موارد جوابهای صریح در دست نیست و باید به همین قانع باشیم که بتوانیم یک ریشه را با هر درجه از دقت مشخصی حساب کنیم. روشهای عددی در یافتن ریشه‌ها، **روشهای بارستی** (تکراری) خوانده می‌شوند که موضوع اصلی این فصل هستند.

کار را با روشهای بارستی برای حل (۱.۰.۲)، هنگامی که $f(x)$ یک تابع پیوسته مشتقپذیر حقیقی‌ـمقدار از متغیر حقیقی x باشد، آغاز می‌کنیم. روشهای بارستی در حل این رده از معادلات کاملاً کلّی نیاز به یک یا چند حدس اولیهٔ $x_۰$ برای ریشهٔ مطلوب α از $f(x)$ دارند. حدس اولیهٔ $x_۰$ را می‌توان معمولاً، با به‌کارگیری شرایطی که اولین بار مسأله در آن ظاهر شده، پیدا کرد؛ در غیر این صورت یک نمودار سادهٔ $y = f(x)$ اغلب برای برآورد $x_۰$ کافی خواهد بود.

مسألهٔ مهم دیگری که در این فصل مورد بحث قرار می‌گیرد یافتن یک یا چند ریشه از معادلهٔ

چندجمله‌ای زیر است

$$p(x) \equiv a_0 + a_1 x + \ldots + a_n x^n = 0 \qquad a_n \neq 0 \qquad (۲.۰.۲)$$

روشهای مسألهٔ اول اغلب اختصاص به ۲.۰.۲ یافته است و روش ما نیز همین خواهد بود. ولی نوشته‌های زیادی در مورد روشهایی وجود دارند که به ویژه برای معادله‌های چندجمله‌ای با استفاده از ویژگیهای خاص آنها از راهی طبیعی به‌وجود آمده‌اند. این روشها مهمترین روشهایی هستند که در ایجاد برنامه‌های خودکار رایانه‌ای برای حل ۲.۰.۲ به‌کار برده می‌شوند و ما در اینجا به بعضی از آنها اشاره خواهیم کرد.

سومین ردهٔ مسائلی که در این فصل مورد بحث قرار می‌گیرند حل دستگاه معادله‌های غیرخطی‌هستند. این دستگاهها به شکلهای گوناگون هستند و آنالیز عددی مربوط به آنها هم مفصل وهم پیچیده است. ما فقط اشاره‌ای به این موضوع خواهیم کرد و بعضی از روشهای موفق را که از نظر کاربرد نسبتاً کلی هستند ذکر خواهیم کرد. مطالعهٔ مفصل این موضوع نیاز به‌دانش کافی در جبر خطی نظری و عددی دارد که تا فصلهای ۷ تا ۹ به این مطالب پرداخته‌نشده است.

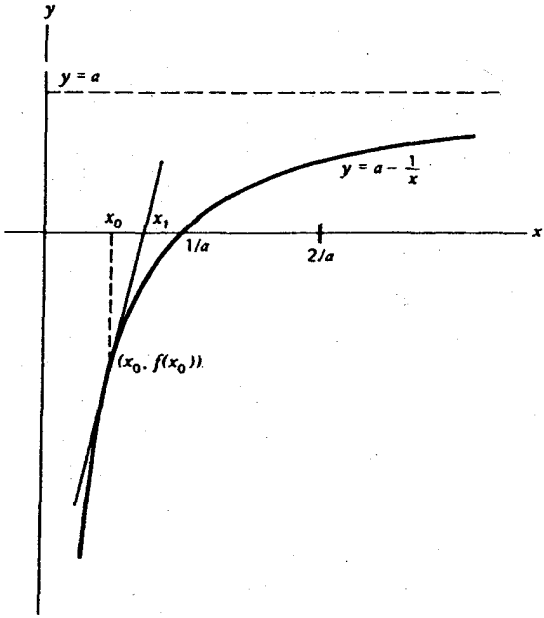
آخرین ردهٔ مسائل مورد بحث در این فصل مسائل بهینه‌سازی هستند. در این حالت درپی ماکسیمم‌سازی یا مینیمم‌سازی یک تابع حقیقی‌ـمقدار $f(x_1, \ldots, x_n)$ و پیدا کردن نقطهٔ $(x_1, \ldots, x_n)$ هستیم که در آن نقطه مقدار بهینه به‌دست می‌آید. این مسائل اغلب به یک دستگاه معادله‌های غیرخطی بدل می‌شوند، ولی معمولاً بهتر است که روشهای ویژه‌ای مستقیماً برای بهینه‌سازی تهیه شود. موضوع بهینه‌سازی به نحوی گسترده پیشرفت کرده و تکامل یافته است، ما فقط به اختصار آن را معرفی و مطالعه می‌کنیم.

برای روشن‌کردن مفهوم روش بارستی در پیدا کردن ریشهٔ ۱.۰.۲، مطلب را با یک مثال آغازمی‌کنیم.

حل معادلهٔ

$$f(x) \equiv a - \frac{1}{x} = 0 \qquad (۳.۰.۲)$$

را به ازای مقدار داده شدهٔ $a > 0$ در نظر می‌گیریم. این مسأله یک کاربرد عملی در رایانه‌های بدون عمل تقسیم دارد. این امر در بعضی رایانه‌های اولیه واقعیّت داشت، و بعضی رایانه‌های جدید نیز از الگوریتمی که ذیلاً داده می‌شود برای عمل تقسیم استفاده می‌کنند.



شکل ۱.۲ حل بارستی $a - (1/x) = 0$

گیریم x_0 یک جواب تقریبی معادله باشد. در نقطهٔ $(x_0, f(x_0))$ خط مماس بر نمودار $y = f(x)$ را رسم می‌کنیم (شکل ۱.۲). گیریم x_1 نقطه‌ای باشد که خط مماس محور xها را قطع می‌کند. این نقطه باید یک تقریب بهتری برای ریشهٔ α باشد.

برای به‌دست آوردن یک معادله برای x_1، شیب خط مماس و مشتق $f(x)$ در x_0 را برابر قرار می‌دهیم

$$f'(x_0) = \frac{f(x_0) - 0}{x_0 - x_1}$$

با توجه به ۳.۰.۲ و اندکی عملیات به‌دست می‌آوریم

$$x_1 = x_0(2 - ax_0)$$

فرمول بارستی کلی از تکرار این روند با قراردادن x_1 به جای x_0 و ادامهٔ این کار تا بینهایت، حاصل می‌شود

$$x_{n+1} = x_n(2 - ax_n) \qquad n \geq 0 \qquad (4.0.2)$$

یک شکل مناسبتر برای منظورهای نظری با وارد کردن شکل مقیاس شده مانده

$$r_n = 1 - ax_n \qquad (5.0.2)$$

به‌دست می‌آید. با استفاده از آن خواهیم داشت:

$$x_{n+1} = x_n(1 + r_n) \qquad n \geq 0 \qquad (6.0.2)$$

پس خطا برابر است با

$$e_n = \frac{1}{a} - x_n = \frac{r_n}{a} \qquad (7.0.2)$$

اکنون به تحلیل همگرایی این روش، سرعت و وابستگی آن به x_0 می‌پردازیم. ابتدا

$$r_{n+1} = 1 - ax_{n+1} = 1 - ax_n(1 + r_n) = 1 - (1 - r_n)(1 + r_n)$$

$$r_{n+1} = r_n^2 \qquad (8.0.2)$$

از راه استقرار

$$r_n = r_0^{2^n} \qquad n \geq 0 \qquad (9.0.2)$$

با توجه به ۷.۰.۲، خطای e_n وقتی $n \to \infty$، به صفر می‌گراید اگر و تنها اگر r_n به صفر میل کند. با توجّه به (۹.۰.۲)، r_n به صفر می‌گراید اگر و تنها اگر $|r_0| < 1$، یا هم‌ارز با آن

$$-1 < 1 - ax_0 < 1$$

$$0 < x_0 < \frac{2}{a} \qquad (10.0.2)$$

برای آنکه x_n به $1/a$ میل کند لازم و کافی است که x_0 طوری انتخاب شود که در (۱۰.۰.۲) صدق کند.

برای بررسی سرعت همگرایی، وقتی که (۱۰.۰.۲) برقرار باشد، فرمولهایی برای خطا و خطای نسبی به‌دست می‌آوریم. برای بررسی سرعت همگرایی در حالتی که (۱۰.۰.۲) برقرار باشد

$$e_{n+1} = \frac{r_{n+1}}{a} = \frac{r_n^2}{a} = \frac{e_n^2 a^2}{a}$$

$$e_{n+1} = a e_n^2 \qquad (11.0.2)$$

$$\frac{e_{n+1}}{1/a} = e_n^2 a^2 = \left[\frac{e_n}{1/a}\right]^2$$

$$\mathrm{Rel}(x_{n+1}) = \mathrm{Rel}(x_n)^2 \qquad n \geq 0 \qquad (12.0.2)$$

نماد $\mathrm{Rel}(x_n)$ معرّف خطای نسبی در x_n است. با توجه به معادلهٔ (۱۱.۰.۲)، می‌گوییم e_n **به طور مربعی به صفر همگرا است**. برای نشان‌دادن سرعت کاهش خطا، فرض می‌کنیم $\mathrm{Rel}(x_۰) = ۰.۱$، در این صورت $\mathrm{Rel}(x_۴) = ۱۰^{-۱۶}$. با هر بارِست شمار ارقام با معنی دو برابر می‌شود.

این مثال نحوهٔ ساختن یک روش بارستی را برای حل یک معادله نشان می‌دهد؛ و یک تحلیل‌کامل همگرایی در آن داده شده است. این تحلیل شامل اثبات همگرایی، تعیین **بازهٔ همگرایی** برای انتخاب $x_۰$، و تعیین سرعت همگرایی است. این مفاهیم با استفاده از روشهای کلّیتری در حل (۱.۰.۲) در بخشهای بعد بررسی شده‌اند.

تعریف یک دنباله از بارِستهای $\{x_n \mid n \geq ۰\}$ را به نقطهٔ α همگرا از مرتبهٔ $p \geq ۱$ خوانند اگر به ازای مقداری مانند $c > ۰$،

$$| \alpha - x_{n+۱} | \leq c \, | \alpha - x_n |^p \qquad n \geq ۰ \qquad (۱۳.۰.۲)$$

اگر $p = ۱$، دنباله را **همگرای خطی** به α خوانند. در این حالت باید $c < ۱$. ثابت c **نرخ همگرایی خطی** x_n به α نامیده می‌شود.

با استفاده از این تعریف، مثال قبلی (۶.۰.۲) (۵.۰.۲)-دارای همگرایی مرتبهٔ ۲ است که **همگرایی مربعی** نیز نامیده می‌شود. این تعریف مرتبه، همیشه برای بعضی روشهای بارستی همگرای خطی مناسب نیست. با استفاده از استقرا در (۱۳.۰.۲) با $p = ۱$، به‌دست می‌آوریم

$$| \alpha - x_n | \leq c^n \, | \alpha - x_۰ | \qquad n \geq ۰ \qquad (۱۴.۰.۲)$$

این رابطه مستقیماً همگرایی x_n به α را نشان می‌دهد. برای بعضی روشهای بارستی، می‌توانیم (۱۴.۰.۲) را مستقیماً ثابت کنیم، حال آنکه ممکن است (۱۳.۰.۲) به ازای هیچ مقدار $c < ۱$ درست نباشد. در چنین حالتی، این روش بازهم همگرای خطی با نرخ c خوانده می‌شود.

۱.۲ روش نیم‌سازی (تنصیف)

فرض کنید $f(x)$ بر بازهٔ داده شدهٔ $[a, b]$ پیوسته باشد و در رابطهٔ زیر نیز صدق کند

$$f(a)f(b) < ۰ \qquad (۱.۱.۲)$$

با استفاده از قضیهٔ مقدار میانی ۱.۱ از فصل ۱، تابع $f(x)$ باید حداقل یک ریشه‌در $[a, b]$ داشته

باشد. معمولاً $[a,b]$ به گونه‌ای انتخاب می‌شود که تنها یک ریشهٔ α را شامل شود، ولی الگوریتم زیر، برای روش نیم‌سازی به موجب (۱.۱.۲) همیشه به ریشه‌ای چون α در $[a,b]$ همگراست.

الگوریتم نیم‌ساز (f، a، b، ریشه، ε)

۱. تعریف کنید $c := (a+b)/۲$

۲. اگر $b-c \le \varepsilon$، آنگاه ریشه $:= c$ را بپذیرید و خارج شوید.

۳. اگر $۰ \le$ (علامت $f(c)$) $=$ (علامت $f(b)$)، آنگاه $a := c$؛ در غیر این صورت $b := c$.

۴. به گام اول برگردید.

بازهٔ $[a,b]$ در هرگذر از الگوریتم نصف می‌شود. به موجب گام ۳، $[a,b]$ همیشه یک ریشهٔ $f(x)$ را دربردارد. چون یک ریشهٔ α در $[a,b]$ است، بایستی یا در $[a,c]$ باشد یا در $[c,b]$ و درنتیجه

$$|\, c-\alpha \,| \le b-c = c-a$$

این توجیه آزمون گام ۲ است. پس از تکمیل الگوریتم، c یک مقدار تقریبی ریشه با

$$|\, c-\alpha \,| \le \varepsilon$$

خواهد بود.

مثال بزرگترین ریشهٔ حقیقی α از معادلهٔ

$$f(x) \equiv x^۶ - x - ۱ = ۰ \qquad (۲.۱.۲)$$

را پیدا کنید. به سادگی می‌توان نشان داد که $۱ < \alpha < ۲$، و ما این بازه را به‌عنوان بازهٔ اولیهٔ $[a,b]$ به‌کار می‌بریم. الگوریتم نیم‌ساز با $\varepsilon = ۰٫۰۰۰۰۵$ به‌کارگرفته شده‌است. نتایج در جدول ۱.۲ نشان داده شده‌اند. دوبارست اول بازهٔ اولیهٔ شامل α را به‌دست می‌دهند و سایر مقادیر c_n، $n \ge ۱$، معرّف نقاط میانی متوالی هستند که با استفاده از نیم‌سازی پیدا شده‌اند. آخرین مقدار $c_{۱۵} = ۱٫۱۳۴۷۴$ به‌عنوان تقریبی برای α با $|\, \alpha - c_{۱۵} \,| \le ۰٫۰۰۰۰۴$ پذیرفته شده است.

جواب درست چنین است:

$$\alpha \doteq ۱٫۱۳۴۷۲۴۱۳۸۴۰۱۵۲ \qquad (۳.۱.۲)$$

خطای واقعی در $c_{۱۵}$ عبارت است از

$$\alpha - c_{۱۵} = -۰٫۰۰۰۰۱۶$$

جدول ۱.۲ مثال روش نیم‌سازی

n	c_n	$f(c_n)$	n	c_n	$f(c_n)$
	۲٫۰ = b	۶۱٫۰	۸	۱٫۱۳۶۷۲	۰٫۰۲۰۶۲
	۱٫۰ = a	−۱٫۰	۹	۱٫۱۳۴۷۷	۰٫۰۰۰۴۳
۱	۱٫۵	۸٫۸۹۰۶۳	۱۰	۱٫۱۳۳۷۹	−۰٫۰۰۹۶۰
۲	۱٫۲۵	۱٫۵۶۴۷۰	۱۱	۱٫۱۳۴۲۸	−۰٫۰۰۴۵۹
۳	۱٫۱۲۵	−۰٫۰۹۷۷۱	۱۲	۱٫۱۳۴۵۲	−۰٫۰۰۲۰۸
۴	۱٫۱۸۷۵	۰٫۶۱۶۶۵	۱۳	۱٫۱۳۴۶۴	−۰٫۰۰۰۸۳
۵	۱٫۱۵۶۲۵	۰٫۲۳۳۲۷	۱۴	۱٫۱۳۴۷۰	−۰٫۰۰۰۲۰
۶	۱٫۱۴۰۶۳	۰٫۰۶۱۵۸	۱۵	۱٫۱۳۴۷۴	۰٫۰۰۰۱۶
۷	۱٫۱۳۲۸۱	−۰٫۱۹۵۸			

این خطا بسیار کوچکتر از کران خطای پیش‌بینی شده‌است. ممکن است چنین به نظر آید که می‌توانستیم با توقف در یکی از بارستهای قبلی قدری از محاسبات بکاهیم. ولی راهی برای پیش‌بینی دقت بهتر احتمالی در بارست قبلی نیست، و بنابراین راهی وجود ندارد که بتوانیم تشخیص دهیم که این بارست به اندازهٔ کافی دقیق هست یا نیست. برای مثال c_9 به اندازهٔ کافی دقیق است، ولی هیچ راهی که به ما این حقیقت را در حین محاسبات بگوید وجود ندارد.

برای بررسی سرعت همگرایی، گیریم c_n معرّف n امین مقدار c در الگوریتم باشد. لذا به سادگی می‌توان دید که

$$\alpha = \lim_{n\to\infty} c_n$$

$$| \alpha - c_n | \leq \left[\frac{۱}{۲}\right]^n (b-a) \qquad (۴.۱.۲)$$

که $b-a$ معرّف طول بازهٔ اوّلیه‌ای است که در **نیم‌سازی** داده شده است. با استفاده از صورت (۱۴.۰.۲) برای تعریف همگرایی خطی، می‌گوییم روش نیم‌سازی همگرای خطی با نرخ ۱/۲ است. ممکن است خطای واقعی در هرگام با نرخ ۱/۲ کاهش نیابد، ولی بر مبنای (۴.۱.۲) نرخ متوسط کاهش ۱/۲ است. مثال قبلی نتیجهٔ (۴.۱.۲) را نشان می‌دهد.

الگوریتم **نیم‌ساز** چند عیب دارد. اوّل، حدود دقت ماشین را آن‌گونه که در بخش ۱ـ۲ فصل یک توضیح داده شد، رعایت نمی‌کند. یک برنامهٔ عملی واحدگردکردن ماشین را در نظر می‌گیرد و اگر لازم باشد εی داده شده را تعدیل می‌نماید [(۱۲.۲.۱)]. مشکل عمدهٔ دیگر در روش **نیم‌سازی** آن است که درمقایسه با سایر روشهایی که در بخشهای بعدی تعریف می‌شوند، به‌کندی همگرا می‌شود. امتیازهای

عمدهٔ روش نیم‌سازی عبارت‌اند از: (۱) همگرایی آن تضمین شده است (به شرطی که f روی $[a,b]$ پیوسته و (۱.۱.۲) برقرار باشد) و (۲) یک کران خطای قابل قبول در دسترس است. روشهایی که در هر گام کرانهای بالایی و پایین ریشهٔ α را به‌دست‌می‌دهند **روشهای حصارکشی** (حَصر) خوانده می‌شوند. در بخش ۸.۲ ما به تشریح یک‌الگوریتم حَصر می‌پردازیم که مزایای روش نیم‌سازی مذکور در قبل و همگرایی سریعتر روش خط قاطع را (که در بخش ۲ـ۳ توضیح داده می‌شود) توأماً دارد.

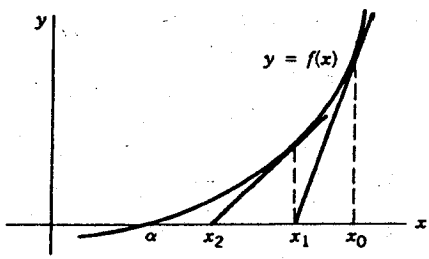
۲.۲ روش نیوتن

فرض کنید که یک برآورد اولّیهٔ x_0 برای α، ریشهٔ مطلوب $f(x)=0$، معلوم باشد. روش نیوتن دنباله‌ای از بارستهای $\{x_n : n \geq 1\}$ را تولید می‌کند، که امیدواریم به α همگرا شود. چون x_0 نزدیک به α فرض شده است، نمودار $y=f(x)$ را در مجاورت ریشهٔ α با رسم خط مماس بر نمودارش در $(x_0, f(x_0))$ تقریب می‌زنیم. سپس ریشهٔ این خط مماس[1] را برای تقریب زدن α به‌کار می‌بریم؛ این تقریب جدید را x_1 می‌نامیم. این روند را به دفعات بسیار زیاد تکرار می‌کنیم تا دنباله‌ای از بارستهای x_n به‌دست آید. مانند مثال (۳.۰.۲) در ابتدای این فصل، این کار ما را به فرمول بارستی زیر می‌رساند

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad n \geq 0 \qquad (۱.۲.۲)$$

این روند در شکل ۲.۲ برای بارستهای x_1 و x_2 نشان داده شده است.

روش نیوتن معروفترین شیوه برای یافتن ریشه‌های یک معادله است. این روش به راههای گوناگونی برای حل مسائل غیرخطی مشکلتر، مثلاً برای دستگاههای معادلات غیرخطی و معادلات انتگرالی و دیفرانسیل غیرخطی، تعمیم داده شده‌است. این روش همیشه بهترین روش برای یک مسألهٔ داده شده نیست، ولی سادگی صوری و سرعت زیاد آن اغلب موجَب می‌شود که اولین روشی باشد که برای حل یک مسألهٔ غیرخطی به‌کار می‌رود.



شکل ۲.۲ روش نیوتن

۱. یعنی محل تقاطع خط مماس با محور xها

به‌عنوان یک روش دیگر دستیابی به (۱.۲.۲)، از بسط سری تیلر استفاده می‌کنیم. تابع $f(x)$ را حول x_n بسط می‌دهیم،

$$f(x) = f(x_n) + (x - x_n)f'(x_n) + \frac{(x - x_n)^2}{2} f''(\xi)$$

که ξ بین x و x_n قرار دارد. x را مساوی α قرار می‌دهیم و با استفاده از $f(\alpha) = 0$ این عبارت را نسبت به α حل می‌کنیم تا به‌دست آوریم

$$\alpha = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{(\alpha - x_n)^2}{2} \cdot \frac{f''(\xi_n)}{f'(x_n)}$$

که ξ_n بین x_n و α واقع است. می‌توانیم جملهٔ خطا (آخرین جمله) را حذف کنیم تا یک تقریب بهتر از x_n برای α به‌دست آوریم، و ما این تقریب را به‌صورت x_{n+1} در (۱.۲.۲) داریم. در این صورت

$$\alpha - x_{n+1} = -(\alpha - x_n)^2 \cdot \frac{f''(\xi_n)}{2f'(x_n)} \qquad n \geq 0 \qquad (۲.۲.۲)$$

از این فرمول استفاده می‌کنیم تا نشان دهیم که روش نیوتن همگرایی مرتبهٔ دوّم، $p = 2$ در (۱۳.۰.۲)، دارد.

مثال باز هم معادلهٔ

$$f(x) \equiv x^6 - x - 1 = 0$$

را برای پیدا کردن بزرگترین ریشه حل می‌کنیم. روش نیوتن (۱.۲.۲) به‌کار برده شده و نتایج در جدول ۲.۲ نشان داده شده‌اند. محاسبات با حساب تقریبی با ۱۶ رقم ممیز شناور انجام شده

جدول ۲.۲ مثال روش نیوتن

n	x_n	$f(x_n)$	$\alpha - x_n$	$x_{n+1} - x_n$
۰	۲٫۰	۶۱٫۰	−۸٫۶۵۳E − ۱	
۱	۱٫۶۸۰۶۲۸۲۷۳	۱۹٫۸۵	−۵٫۴۵۹E − ۱	−۰٫۴۹۹E − ۱
۲	۱٫۴۳۰۷۳۸۹۸۹	۶٫۱۴۷	−۲٫۹۶۰E − ۱	−۱٫۷۵۸E − ۱
۳	۱٫۲۵۴۹۷۰۹۵۷	۱٫۶۵۲	−۱٫۲۰۲E − ۱	−۹٫۳۴۳E − ۲
۴	۱٫۱۶۱۵۳۸۴۳۳	۲٫۹۴۳E − ۱	−۲٫۶۸۱E − ۲	−۲٫۵۱۹E − ۲
۵	۱٫۱۳۶۳۵۳۲۷۴	۱٫۶۸۳E − ۲	−۱٫۶۲۹E − ۳	−۱٫۶۲۳E − ۳
۶	۱٫۱۳۴۷۳۰۵۲۸	۶٫۵۷۴E − ۵	−۶٫۳۹۰E − ۶	−۶٫۳۹۰E − ۶
۷	۱٫۱۳۴۷۲۴۱۳۹	۱٫۰۱۵E − ۹	−۹٫۸۷۰E − ۱۱	−۹٫۸۷۰E − ۱۱

و جدول بارستها ازگردکردن این مقادیر دقیقتر به‌دست آمده است. آخرین ستون، $x_{n+1} - x_n$، برآوردی برای $\alpha - x_n$ است؛ این مطلب بعداً در این بخش مورد بحث واقع شده است.

هرگاه یک بارست تا حدّی به α نزدیک باشد، روش نیوتن به سرعت همگرا می‌شود. این موضوع در بارستهای $x_۴$، $x_۵$، $x_۶$، $x_۷$ نمایش داده شده است. بارستهای $x_۰$، $x_۱$، $x_۲$، $x_۳$ همگرایی اولیهٔ آهسته را نشان می‌دهند که با حدس اوّلیهٔ ضعیف $x_۰$ امکان دارد. اگر حدس اولیه $x_۰ = ۱$ انتخاب شده بود، $x_۴$ تا ۷ رقم با معنی و $x_۵$ تا ۱۴ رقم بامعنی دقیق می‌شدند. این نتایج اگر با نتایجی که از روش نیم‌سازی در جدول ۱.۲ داده شده مقایسه شوند، سرعت زیادتر روش نیوتن بلافاصله آشکار می‌شود.

تحلیل همگرایی. یک قضیهٔ همگرایی داده خواهد شد که سرعت و همچنین بازه‌ای را که از آن حدسهای اولیه را می‌توان انتخاب کرد نشان می‌دهد.

قضیهٔ ۱.۲ فرض کنید $f(x)$، $f'(x)$ و $f''(x)$ به ازای جمیع مقادیر x در یک همسایگی α پیوسته باشند و فرض می‌کنیم $f(\alpha) = ۰, f'(\alpha) \neq ۰$. اگر $x_۰$ به اندازهٔ کافی نزدیک به α انتخاب شود، بارستهای x_n، $n \geq ۰$، در (۱.۲.۲) به سمت α همگرا می‌شوند. به علاوه

$$\lim_{n\to\infty} \frac{\alpha - x_{n+۱}}{(\alpha - x_n)^۲} = -\frac{f''(\alpha)}{۲f'(\alpha)} \qquad (۳.۲.۲)$$

که اثبات می‌کند بارستها همگرایی مرتبهٔ $p = ۲$ دارند.

برهان یک بازهٔ به اندازهٔ کافی کوچک $I = [\alpha - \varepsilon, \alpha + \varepsilon]$ را که در آن $f'(x) \neq ۰$، برمی‌گزینیم [با توجه به پیوستگی $f'(x)$، این بازه وجود دارد]، و سپس فرض می‌کنیم

$$M = \frac{\max_{x\in I} | f''(x) |}{۲\min_{x\in I} | f'(x) |}$$

با توجه به (۲.۲.۲) داریم

$$| \alpha - x_۱ | \leq M | \alpha - x_۰ |^۲$$

$$M | \alpha - x_۱ | \leq (M | \alpha - x_۰ |)^۲$$

می‌گیریم $| \alpha - x_۰ | \leq \varepsilon$ و $M | \alpha - x_۰ | < ۱$. پس $M | \alpha - x_۱ | < ۱$، و $| \alpha - x_۱ | \leq M | \alpha - x_۰ |$، که در نتیجه $| \alpha - x_۱ | \leq \varepsilon$. می‌توانیم از راه استقرا

همین استدلال را برای x_1، x_2، ... به‌کار برده نشان دهیم که به‌ازای جمیع مقادیر $n \geq 1$ روابط $|\alpha - x_n| \leq \varepsilon$ و $M|\alpha - x_n| < 1$ برقرارند.

برای نشان دادن همگرایی، (۲.۲.۲) را به‌کار می‌بریم تا به‌دست آوریم

$$|\alpha - x_{n+1}| \leq M|\alpha - x_n|^2$$

$$M|\alpha - x_{n+1}| \leq (M|\alpha - x_n|)^2 \qquad (۴.۲.۲)$$

با استفاده از روش استقراء

$$M|\alpha - x_n| \leq (M|\alpha - x_0|)^{2^n}$$

$$|\alpha - x_n| \leq \frac{1}{M}(M|\alpha - x_0|)^{2^n} \qquad (۵.۲.۲)$$

برقراری رابطهٔ $M|\alpha - x_0| < 1$ نشان می‌دهد که وقتی $n \to \infty$، آنگاه $x_n \to \alpha$.

در فرمول (۲.۲.۲) نقطهٔ مجهول ξ_n بین x_n و α واقع است و در نتیجه‌وقتی $n \to \infty$، آنگاه $\xi_n \to \alpha$. بنابراین

$$\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{(\alpha - x_n)^2} = -\lim_{n\to\infty} \frac{f''(\xi_n)}{2f'(x_n)} = \frac{-f''(\alpha)}{2f'(\alpha)}$$

■

ستون خطا در جدول ۲.۲ را می‌توان برای نمایش (۳.۲.۲) به‌کار برد. به‌ویژه در این مثال:

$$-\frac{f''(\alpha)}{\alpha f'(\alpha)} = -2{,}417 \qquad \frac{\alpha - x_6}{(\alpha - x_5)^2} = -2{,}41$$

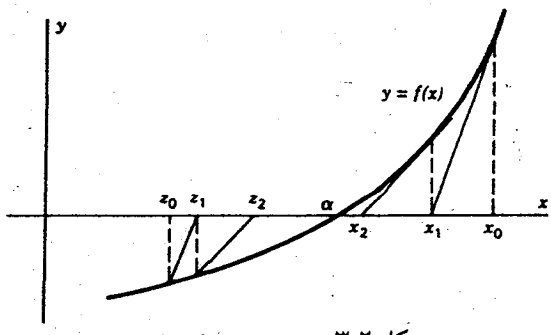
گیریم M معرّف حد طرف راست (۲.۲.۲) باشد. پس اگر x_n نزدیک به α باشد، (۲.۲.۲) نتیجه می‌دهد که

$$M(\alpha - x_{n+1}) \doteq [M(\alpha - x_n)]^2$$

برای اینکه x_n به α همگرا شود، این عبارت می‌گوید که احتمالاً باید داشته‌باشیم

$$|\alpha - x_0| < \frac{1}{M} \qquad (۶.۲.۲)$$

بنابراین M میزانی است برای آنکه بدانیم چه اندازه x_0 باید به α نزدیک باشد تا همگرایی به α تضمین شود. چند مثال با مقادیر بزرگ M در مسائل آخر این فصل داده شده‌اند.



شکل ۳.۲ روش نیوتن ـ فوریه

یک راه دیگر برای تحلیل خطای روش نیوتن استفاده از ترسیم و قضیهٔ زیر است. گیریم $f(x)$ روی بازهٔ $[a,b]$ که شامل α است دوبار پیوسته مشتقپذیر باشد. به علاوه فرض می‌کنیم $f(a)<0$، $f(b)>0$ و همچنین

$$f'(x)>0 \qquad f''(x)>0 \qquad \text{برای } a\le x\le b \tag{۷.۲.۲}$$

در این صورت $f(x)$ در $[a,b]$ اکیداً صعودی است و یک ریشهٔ یکتای α در $[a,b]$ وجود دارد. همچنین به ازای $a\le x<\alpha$، $f(x)<0$، و به‌ازای $\alpha<x\le b$، $f(x)>0$.

گیریم $x_0=b$ و بارستهای نیوتنِ x_n را به‌صورت (۲ـ۲ـ۱) تعریف می‌کنیم. سپس یک دنبالهٔ جدید از بارستها را به شکل زیر تعریف می‌کنیم.

$$z_{n+1}=z_n-\frac{f(z_n)}{f'(x_n)} \qquad n\ge 0 \tag{۸.۲.۲}$$

با $z_0=a$. بارستهای حاصله در شکل ۳.۲ نمایش داده شده‌اند. با استفاده از $\{z_n\}$ کرانهای بالا و پایین عالی برای α به‌دست می‌آوریم. استفاده از (۸.۲.۲) با روش نیوتن، **روش نیوتن ـ فوریه** نامیده می‌شود.

قضیهٔ **۲.۲** مانند گذشته، فرض می‌کنیم $f(x)$ بر بازهٔ $[a,b]$ دوبار پیوسته مشتقپذیر باشد و $f(a)<0$، $f(b)>0$ و شرطهای (۲ـ۲ـ۷) برقرار باشند. در این صورت بارستهای x_n اکیداً به α کاهش می‌یابند و بارستهای z_n اکیداً به α افزایش پیدا می‌کنند. به علاوه،

$$\lim_{n\to\infty}\frac{x_{n+1}-z_{n+1}}{(x_n-z_n)^2}=\frac{f''(\alpha)}{2f'(\alpha)} \tag{۹.۲.۲}$$

که نشان می‌دهد فاصلهٔ بین x_n و z_n با افزایش n کاهش درجهٔ ۲ دارد.

برهان ابتدا نشان می‌دهیم که

$$z_0 < z_1 < \alpha < x_1 < x_0 \qquad (۱۰.۲.۲)$$

با توجه به تعاریف (۱.۲.۲) و (۸.۲.۲)

$$x_1 - x_0 = \frac{-f(x_0)}{f'(x_0)} < 0$$
$$z_1 - z_0 = \frac{-f(z_0)}{f'(x_0)} > 0$$

از فرمول خطای (۲.۲.۲)

$$\alpha - x_1 = -(\alpha - x_0)^2 \frac{f''(\xi_0)}{2f'(x_0)} < 0$$

بالاخره

$$\begin{aligned}\alpha - z_1 &= \alpha - z_0 + \frac{f(z_0)}{f'(x_0)} = \alpha - z_0 + \frac{f(z_0) - f(\alpha)}{f'(x_0)} \\ &= \alpha - z_0 - \frac{f'(\zeta_0)(\alpha - z_0)}{f'(x_0)} \qquad z_0 < \zeta_0 < \alpha \text{ به ازای مقداری چون} \\ &= (\alpha - z_0)\left[\frac{f'(x_0) - f'(\zeta_0)}{f'(x_0)}\right] > 0\end{aligned}$$

زیرا $f'(x)$ در $[a, b]$ یک تابع صعودی است. از ترکیب این نتایج اثبات (۱۰.۲.۲) حاصل می‌شود. این برهان را می‌توان به‌طور استقرایی تکرار کرد تا ثابت شود که

$$z_n < z_{n+1} < \alpha < x_{n+1} < x_n \qquad n \geq 0 \qquad (۱۱.۲.۲)$$

دنبالهٔ $\{x_n\}$ توسط α از پایین کراندار است، پس یک اینفیمم $\overline{x}$ دارد؛ همچنین دنبالهٔ $\{z_n\}$ یک سوپرمم $\overline{z}$ دارد:

$$\lim_{n\to\infty} x_n = \overline{x} \geq \alpha \qquad \lim_{n\to\infty} z_n = \overline{z} \leq \alpha$$

با حد گرفتن در (۱.۲.۲) و (۸.۲.۲) به‌دست می‌آوریم

$$\overline{x} = \overline{x} - \frac{f(\overline{x})}{f'(\overline{x})} \qquad \overline{z} = \overline{z} - \frac{f(\overline{z})}{f'(\overline{z})}$$

که ما را به $f(\overline{x}) = ۰ = f(\overline{z})$ می‌رساند. چون α ریشهٔ یکتای $f(x)$ در $[a, b]$ است، این ثابت می‌کند که $\{x_n\}$ و $\{z_n\}$ به α همگرا هستند.

برهان (۹.۲.۲) پیچیده‌تر است و خوانندگان را به اثر آستروفسکی (۱۹۷۳، ص۷۰) ارجاع می‌دهیم. با توجه به قضیهٔ ۱.۲ و فرمول (۳.۲.۲)، دنبالهٔ $\{x_n\}$ همگرای درجهٔ ۲ به α است. قضیهٔ (۹.۲.۲) نشان می‌دهد که

$$| \alpha - x_n | \leq | z_n - x_n |$$

یک کران خطاست که کاهش درجهٔ ۲ دارد. ■

فرضهای قضیهٔ ۲.۲ بدل می‌شود به اینکه $f(x)$ دوبار پیوسته مشتقپذیر در یک همسایگی α باشد و

$$f'(\alpha)f''(\alpha) \neq ۰ \qquad (۱۲.۲.۲)$$

پس، یک بازهٔ $[a, b]$ حول α با $f'(x)$ و $f''(x)$ مخالف صفر در این بازه وجود دارد. لذا مسألهٔ ریشه‌یابی $f(x) = ۰$ در (۲ـ۲ـ۷) صدق می‌کند یا به آسانی به یک مسألهٔ هم‌ارزی که در آن صدق می‌کند بدل می‌شود. برای مثال، اگر $f'(\alpha) < ۰$ و $f''(\alpha) > ۰$، مسألهٔ ریشه‌یابی $g(x) = ۰$ را با شرط $g(x) \equiv f(-x)$ در نظر می‌گیریم. ریشهٔ g برابر $-\alpha$ خواهدبود و $g(x)$ در شرایط (۷.۲.۲) در بازه‌ای حول $-\alpha$ صدق می‌کند. مثال عددی قضیهٔ ۲.۲ به بخش مسائل واگذار خواهد شد.

برآورد خطا. روند قبلی، کرانهای بالا و پایین ریشه را به‌دست می‌دهد با فاصلهٔ $x_n - z_n$ که کاهش درجهٔ ۲ دارد. ولی در بیشتر کاربردها، روش نیوتن به تنهایی، بدون (۸.۲.۲)، به‌کار برده می‌شود. در آنجا، از روش زیر استفاده می‌کنیم.

با استفاده از قضیهٔ مقدار میانگین،

$$f(x_n) = f(x_n) - f(\alpha) = f'(\xi_n)(x_n - \alpha)$$

$$\alpha - x_n = \frac{-f(x_n)}{f'(\xi_n)}$$

که ξ_n بین x_n و α است. اگر $f'(x)$ در بین x_n و α به‌سرعت تغییر نکند، داریم $f'(\xi_n) \doteq f'(x_n)$، و

$$\alpha - x_n \doteq \frac{-f(x_n)}{f'(x_n)} = x_{n+۱} - x_n$$

که تساوی آخری از تعریف روش نیوتن به‌دست می‌آید. برای روش نیوتن برآورد خطای استانده عبارت است از

$$\alpha - x_n \doteq x_{n+1} - x_n \qquad (۱۳.۲.۲)$$

و این نکته در جدول ۲.۲ نشان داده شده است. برای خطای نسبی می‌نویسیم

$$\frac{\alpha - x_n}{\alpha} \doteq \frac{x_{n+1} - x_n}{x_{n+1}}$$

الگوریتم نیوتن. با به‌کارگیری فرمول (۱.۲.۲) و برآورد خطای (۱۳.۲.۲)، الگوریتم زیر را می‌دهیم.

الگوریتم نیوتن (f، df، x_0، ε، ریشه، itmax، ier)

۱. توجه: df تابع مشتق $f'(x)$ است، itmax ماکسیمم تعداد بارستهاست که‌بایستی محاسبه شود و ier یک اعلام خطا برای استفاده کننده است.

۲. itnum := ۱

۳. denom := $df(x_0)$

۴. اگر denom = ۰ آنگاه ier := ۲ و خارج شوید.

۵. $x_1 := x_0 - f(x_0)/\text{denom}$

۶. اگر $|x_1 - x_0| \le \varepsilon$، آنگاه قرار دهید $ier := 0$، ریشه := x_1، و خارج شوید.

۷. اگر itnum = itmax، قرار دهید ier := ۱ و خارج شوید.

۸. در غیر اینصورت itnum := itnum + ۱، $x_0 := x_1$، و به گام ۳ برگردید.

همانند الگوریتم نیم‌ساز قبلی، هیچ محدودیتی برای حساب رایانه در نظر گرفته نشده است، مع‌هذا یک برنامهٔ عملی برای انجام آن مورد نیاز است. همچنین روش نیوتن تضمینی برای همگرایی نمی‌کند، و لذا یک آزمون روی تعداد بارستها (گام۷) لازم است.

وقتی روش نیوتن همگرا باشد، معمولاً این همگرایی خیلی سریع خواهد بود، که خود مزّیتی است نسبت به روش نیم‌سازی. ولی ممکن است همگرا نباشد. یک منبع دیگر اشکال در بعضی موارد لزوم دانستن $f'(x)$ به‌طور صریح است. در بعضی مسائلِ ریشه‌یابی این امکان‌پذیر نیست. در روش بخش بعد، این نقص، به بهای اندک کاهشِ سرعتِ همگرایی، برطرف خواهد شد.

۳.۲ روش خط قاطع

همانند روش نیوتن، نمودار $y = f(x)$ در مجاورت ریشهٔ α با یک خط مستقیم تقریب زده می‌شود. در این حالت، فرض می‌کنیم که x_0 و x_1 دو برآورد اولیه از ریشهٔ α باشند. نمودار $y = f(x)$ را با **خط قاطعی** که توسط $(x_0, f(x_0))$ و $(x_1, f(x_1))$ تعیین می‌شود تقریب می‌زنیم. ریشهٔ آن را با x_2 نشان می‌دهیم که امیدواریم یک تقریب بهتری برای α باشد. این مطلب در شکل ۴.۲ نمایش داده شده است. با استفاده از فرمول شیب خط قاطع داریم

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_1) - 0}{x_1 - x_2}$$

از حل آن برحسب x_2 ؛ خواهیم داشت:

$$x_2 = x_1 - f(x_1).\frac{x_1 - x_0}{f(x_1) - f(x_0)}$$

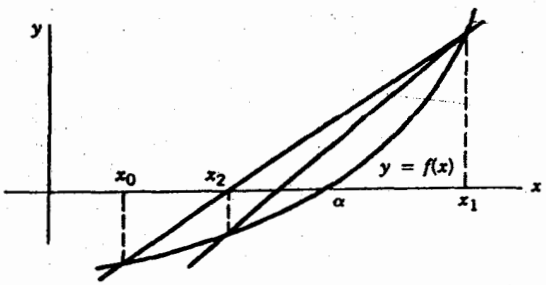
با استفاده از x_1 و x_2 این فرآیند را تکرار می‌کنیم تا x_3 و غیره به‌دست آیند. فرمول عمومی بر این پایه چنین است

$$x_{n+1} = x_n - f(x_n).\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \qquad n \geq 1 \qquad (۱.۳.۲)$$

این **روش خط قاطع** است. همانند روش نیوتن همگرایی این روش تضمین نمی‌شود، ولی وقتی همگرا شود، سرعت آن معمولاً بیشتر از سرعت روش نیم‌سازی است.

مثال باز پیداکردن بزرگترین ریشهٔ

$$f(x) \equiv x^6 - x - 1 = 0$$



شکل ۴.۲ روش خط قاطع

جدول ۳.۲ مثال روش خط قاطع

n	x_n	$f(x_n)$	$\alpha - x_n$	$x_n - x_{n-1}$
۰	۲٫۰	۶۱٫۰	۸٫۶۵E − ۱	
۱	۱٫۰	−۱٫۰	۱٫۳۵E − ۱	۱٫۶۱E − ۲
۲	۱٫۰۱۶۱۲۹۰۳۲	−۹٫۱۵۴E − ۱	۱٫۱۹E − ۱	۱٫۷۴E − ۱
۳	۱٫۱۹۰۵۷۷۷۶۹	۶٫۵۷۵E − ۱	−۵٫۵۹E − ۲	−۷٫۲۹E − ۲
۴	۱٫۱۱۷۶۵۵۸۳۱	−۱٫۶۸۵E − ۱	−۱٫۷۱E − ۲	۱٫۴۹E − ۲
۵	۱٫۱۳۲۵۳۱۵۵۰	−۲٫۲۴۴E − ۲	۲٫۱۹E − ۳	۲٫۲۹E − ۳
۶	۱٫۱۳۴۸۱۶۸۰۸	۹٫۵۳۶E − ۴	−۹٫۲۷E − ۵	−۹٫۳۲E − ۵
۷	۱٫۱۳۴۷۲۳۶۴۶	−۵٫۰۶۶E − ۶	۴٫۹۲E − ۷	۴٫۹۲E − ۷
۹	۱٫۱۳۴۷۲۴۱۳۸	−۱٫۱۳۵E − ۹	۱٫۱۰E − ۱۰	۱٫۱۰E − ۱۰

را در نظر می‌گیریم. روش خط قاطع (۱.۳.۲) به‌کار برده شده و بارست ادامه یافته‌است تا جایی که تفاضلهای متوالیِ $x_n - x_{n-1}$ به اندازهٔ کافی کوچک شده‌اند. نتایج عددی در جدول ۳.۲ داده شده‌اند. محاسبات با یک ماشین دودویی تقریباً با دقّت ۱۶ رقم اعشاری انجام گرفته است و نتایج جدول از گردکردن نتایج رایانه به‌دست آمده‌اند.

همگرایی با افزایش n به‌طور فزاینده‌ای سریع می‌شود. یک راه اندازه‌گیری آن محاسبهٔ نسبت زیر است

$$\frac{\alpha - x_{n+1}}{\alpha - x_n} \qquad n \geq 0$$

برای روشهای خطی، این عبارت وقتی که x_n به α همگرا می‌شود، معمولاً ثابت است. ولی در این مثال، این نسبتها با افزایش n کوچکتر می‌شوند. یک توضیح شهودی این است که خط مستقیمی که نقاط $(x_{n-1}, f(x_{n-1}))$ و $(x_n, f(x_n))$ را به هم وصل می‌کند، در مجاورت $x = \alpha$ به‌تدریج تقریب بسیار دقیقی برای نمودار $y = f(x)$ می‌شود، و در نتیجه ریشهٔ x_{n+1} خط مستقیم، یک برآورد بسیار بهبودیافته‌ای برای α خواهد شد. همچنین توجه کنید که بارستهای x_n با ازدیاد n، به نحوی ظاهراً تصادفی نسبت به α بالا و پایین می‌روند. توضیح این مطلب از فرمول خطای (۳.۳.۲) داده می‌شود که در پایین آمده است.

تحلیل خطا. اگر طرفین تساوی (۱.۳.۲) را در ۱− ضرب و سپس α را به طرفین آن اضافه کنیم، به‌دست می‌آوریم

$$\alpha - x_{n+1} = \alpha - x_n + f(x_n).\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}$$

با اندکی دستکاری جبری در طرف دست راست فرمول زیر به‌دست می‌آید

$$\alpha - x_{n+1} = -(\alpha - x_{n-1})(\alpha - x_n)\frac{f[x_{n-1}, x_n, \alpha]}{f[x_{n-1}, x_n]} \qquad (۲.۳.۲)$$

کمیتهای $f[x_{n-1}, x_n]$ و $f[x_{n-1}, x_n, \alpha]$ تفاضلات منقسم اول و دوّم نیوتن هستند که در (۱۳.۱.۱) فصل ۱ تعریف شدند. خواننده بایستی با قراردادن‌این مقادیر از (۱۳.۱.۱) و سپس ساده کردن آن، صحّت (۲.۳.۲) راکنترل کند. با استفاده از (۱۴.۱.۱)، فرمول (۲.۳.۲) چنین خواهد شد

$$\alpha - x_{n+1} = -(\alpha - x_{n-1})(\alpha - x_n).\frac{f''(\zeta_n)}{۲f'(\xi_n)} \qquad (۳.۳.۲)$$

که ξ_n بین x_{n-1} و x_n قرار دارد و ζ_n بین x_{n-1} و α یا بین x_nو α واقع است. با استفاده از این فرمول خطا، می‌توانیم همگرایی روش خط قاطع را امتحان کنیم.

قضیهٔ ۳.۲ فرض می‌کنیم $f(x)$، $f'(x)$ و $f''(x)$ به ازای جمیع مقادیر x در بازه‌ای شامل α پیوسته باشند، و $f'(\alpha) \neq ۰$. در این صورت اگر حدسهای اولیهٔ $x_۰$ و $x_۱$ به اندازهٔ کافی نزدیک به α انتخاب شوند، بارستهای x_n از (۲ـ۳ـ۱) به α همگرا می‌شوند. مرتبهٔ همگرایی چنین خواهد شد: $p = (۱ + \sqrt{۵})/۲ \doteq ۱٫۶۲$.

برهان در همسایگی $I = [\alpha - \varepsilon, \alpha + \varepsilon]$ با مقداری از $\varepsilon > ۰$،در همه‌جا روی I، $f'(x)$ مخالف صفر است. پس تعریف می‌کنیم

$$M = \frac{\underset{x \in I}{\text{Max}} \mid f''(x) \mid}{۲\underset{x \in I}{\text{Min}} \mid f'(x) \mid}$$

در این صورت به ازای جمیع مقادیر $x_۰$ و $x_۱$ متعلق به $[\alpha - \varepsilon, \alpha + \varepsilon]$، با استفاده از (۳.۳.۲)

$$\mid e_۲ \mid \leq \mid e_۱ \mid . \mid e_۰ \mid M,$$

$$M \mid e_۲ \mid \leq M \mid e_۱ \mid .M \mid e_۰ \mid$$

به علاوه فرض می‌کنیم که x_0 و x_1 به گونه‌ای انتخاب شده باشند که

$$\delta \equiv \text{Max}\{M \mid e_0 \mid, M \mid e_1 \mid\} < 1 \qquad (۴.۳.۲)$$

در این صورت $M \mid e_2 \mid < 1$ زیرا

$$M \mid e_2 \mid \leq \delta^2$$

همچنین $M \mid e_2 \mid \leq \delta^2 < \delta$ ایجاب می‌کند که

$$\mid e_2 \mid < \frac{\delta}{M} = \text{Max}\{\mid e_1 \mid, \mid e_0 \mid\} \leq \varepsilon$$

و بنابراین $x_2 \in [\alpha - \varepsilon, \alpha + \varepsilon]$. می‌توانیم این استدلال را به‌طور استقرایی به‌کار بریم تا نشان دهیم که $x_n \in [\alpha - \varepsilon, \alpha + \varepsilon]$ و به ازای $n \geq r$، $M \mid e_n \mid \leq \delta$. برای اثبات همگرایی و به دست آوردن مرتبهٔ همگرایی، استفاده از (۳.۳.۲) را ادامه می‌دهیم تا به‌دست آوریم

$$M \mid e_3 \mid \leq M \mid e_2 \mid . M \mid e_1 \mid \leq \delta^2 . \delta = \delta^3$$
$$M \mid e_4 \mid \leq M \mid e_3 \mid . M \mid e_2 \mid \leq \delta^5$$

برای

$$M \mid e_n \mid \leq \delta^{q_n} \qquad (۵.۳.۲)$$
$$M \mid e_{n+1} \mid \leq M \mid e_n \mid . M \mid e_{n-1} \mid \leq \delta^{q_n + q_{n-1}} = \delta^{q_{n+1}}$$

بنابراین

$$q_{n+1} = q_n + q_{n-1} \qquad n \geq 1 \qquad (۶.۳.۲)$$

با $q_0 = q_1 = 1$. این دنبالهٔ فیبوناتچی اعداد است و یک فرمول صریح برای آن می‌توان داد؛

$$q_n = \frac{1}{\sqrt{5}}[r_0^{n+1} - r_1^{n+1}] \qquad n \geq 0 \qquad (۷.۳.۲)$$
$$r_0 = \frac{1 + \sqrt{5}}{2} \doteq ۱٫۶۱۸ \qquad r_1 = \frac{1 - \sqrt{5}}{2} \doteq -۰٫۶۱۸$$

بنابراین

$$q_n \doteq \frac{1}{\sqrt{5}}(۱٫۶۱۸)^{n+1} \qquad \text{به ازای مقدار بزرگ } n \qquad (۸.۳.۲)$$

برای مثال، $q_5 = 8$، و فرمول (۸.۳.۲) عدد ۲۵ ۰ر۸ را می‌دهد. به فرمول (۵.۳.۲) برمی‌گردیم، کران خطای

$$| e_n | \leq \frac{1}{M}\delta^{q_n} \qquad n \geq 0 \qquad (۹.۳.۲)$$

را به‌دست می‌آوریم که q_n به وسیلهٔ (۷.۳.۲) داده شده است. چون وقتی $n \to \infty$، $q_n \to \infty$، خواهیم داشت $x_n \to \alpha$.

با یک استخراج دقیقتر، در واقع عملاً می‌توانیم نشان دهیم که مرتبهٔ همگرایی چنین است: $p = (1+\sqrt{5})/2$. برای ساده کردن بیان به جای آن نشان می‌دهیم که این نرخی است که کران (۹.۳.۲) با آن کاهش می‌یابد. گیریم B_n کران بالای (۹.۳.۲) را نشان دهد. پس

$$\frac{B_{n+1}}{B_n^{r_0}} = \frac{\frac{1}{M}\delta^{q_{n+1}}}{[\frac{1}{M}]^{r_0}.\delta^{r_0 q_n}} = M^{r_0-1}\delta^{q_{n+1}-r_0 q_n}$$

$$\leq \delta^{-1}M^{r_0-1} \equiv c$$

زیرا $q_{n+1} - r_0 q_n = r_1^{n+1} > -1$. پس

$$B_{n+1} \leq cB_n^{r_0}$$

که مستلزم داشتن مرتبهٔ همگرایی $p = r_0 = (1+\sqrt{5})/2$ است. یک نتیجهٔ مشابه برای خطاهای واقعی e_n برقرار است؛ به علاوه

$$\lim_{n\to\infty} \frac{| e_{n+1} |}{| e_n |^{r_0}} = \left| \frac{f''(\alpha)}{2f'(\alpha)} \right|^{(\sqrt{5}-1)/2} \qquad (۱۰.۳.۲)$$

■

با استفاده از فرمول خطای (۳.۳.۲) می‌توان رفتار نوسانی بارستهای x_n حول α در آخرین مثال را توضیح داد. برای x_n و x_{n-1} نزدیک به α (۳.۳.۲) ایجاب می‌کند که داشته باشیم

$$\alpha - x_{n+1} \doteq -(\alpha - x_n)(\alpha - x_{n-1}).\frac{f''(a)}{2f'(\alpha)} \qquad (۱۱.۳.۲)$$

علامت $\alpha - x_{n+1}$ از روی علامت دو خطای قبلی، همراه با علامت $f''(\alpha)/f'(\alpha)$ مشخص می‌شود.

شرط (۴.۳.۲) به ما اطلاع می‌دهد که مقادیر اولیهٔ x_0 و x_1 چقدر باید به α نزدیک باشند تا همگرایی داشته باشیم. اگر کمیّت M خیلی بزرگ باشد یا بالاخص، اگر

$$\left| \frac{f''(\alpha)}{2f'(\alpha)} \right|$$

خیلی بزرگ باشد، $\alpha - x_0$ و $\alpha - x_1$ باید به همین قیاس کوچکتر باشند. همگرایی ممکن است بدون (۴.۳.۲) اتفاق افتد، ولی احتمالاً چنین حالتی کاملاً اتفاقی است.

به‌عنوان یک آزمون خطا، همین برآورد خطای (۱۳.۲.۲) را که برای روش نیوتن استفاده شده به‌کار می‌بریم، یعنی

$$\alpha - x_n \doteq x_{n+1} - x_n$$

استفاده از آن در جدول ۳.۲، در آخرین مثال نشان داده شده است. چون روش خط قاطع ممکن است همگرا نباشد، در برنامه‌های اجرای آن باید تعداد بارستها یک کران بالا داشته باشد، همانند الگوریتم نیوتن در بخش اخیر.

یک مسألهٔ دیگر که ممکن است با روش خط قاطع حل شود، محاسبهٔ مشتق تقریبی

$$a_n = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} \qquad (۱۲.۳.۲)$$

است که در این صورت روش خط قاطع (۱.۳.۲) چنین نوشته می‌شود

$$x_{n+1} = x_n - \frac{f(x_n)}{a_n} \qquad (۱۳.۳.۲)$$

محاسبهٔ a_n متضمّن کاهش ارقام بامعنی هم در صورت و هم در مخرج است. بنابراین وقتی $x_n \to \alpha$، a_n با دقت کمتری تقریب مشتق f خواهدبود. با وجود این به‌دست آوردن بهبود در دقّت x_n را ادامه می‌دهیم تا اینکه به ترازنوفهٔ $f(x)$ به ازای مقدار x نزدیک به α برسیم. در این موقع، a_n ممکن است با $f'(\alpha)$ بسیار تفاوت داشته باشد، و x_{n+1} ممکن است از ریشه به سرعت دور شود. بدین دلیل، دنیس و اشنابل (۱۹۸۳ صص ۳۱ و ۳۲) استفاده از (۱۲.۳.۲) را توصیه می‌کنند تا اینکه $x_n - x_{n-1}$ به اندازهٔ کافی کوچک شود. سپس اینان تقریب دیگری را برای $f'(x)$ توصیه می‌کنند:

$$f'(x_n) \doteq a_n = \frac{f(x_n + h) - f(x_n)}{h}$$

با مقدار ثابت h. توصیهٔ آنها برای h چنین است:

$$h = \sqrt{\delta}.T_\alpha$$

که T_α یک تقریب غیر صفر قابل قبولی برای α، مثلاً x_n، است و δ واحد گردکردن رایانه است [به (۱ـ۲ـ۱۲) نگاه کنید]. توصیهٔ آنها استفاده از h است وقتی که $|x_n - x_{n-1}|$ کوچکتر از h باشد.

هزینهٔ روش خط قاطع در محاسبه‌های تابعی کمی‌افزایش می‌یابد ولی احتمالاً نه بیش از هزینهٔ یک یا دوبارست.

روش خط قاطع به‌عنوان یک روند ریشه‌یابی کارا و آسان در عمل، برای انواع بسیاری از مسائل کاملاً توصیه می‌شود. این مزیّت را هم دارد که برعکس روش نیوتن نیازی به دانستن $f'(x)$ ندارد. در بخش ۸.۲، روش خط قاطع یک جزء مهم از یک الگوریتم دیگر ریشه‌یابی را تشکیل می‌دهد که همگرایی آن تضمین شده است.

مقایسهٔ روش نیوتن و روش خط قاطع. روش نیوتن و روش خط قاطع خیلی بهم نزدیک‌اند. اگر از تقریب

$$f'(x_n) \doteq \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

در فرمول (۱.۲.۲) نیوتن استفاده کنیم، فرمول خط قاطع (۱.۳.۲) را به‌دست آوریم. شرایط همگرایی تقریباً یکی هستند [برای مثال، (۶.۲.۲) و (۴.۳.۲) را برای شرایط خطای اولیه ببینید]، و فرمولهای خطا متشابه‌اند [(۲.۲.۲) و (۳.۳.۲) را ببینید]. با این حال، دو تفاوت عمده وجود دارد. روش نیوتن‌در هر بارست به دو محاسبهٔ تابعی نیاز دارد، که $f(x_n)$ و $f'(x_n)$ هستند، در حالی که روش خط قاطع فقط یک محاسبهٔ تابعی در هر بارست می‌طلبد که $f(x_n)$ است [به شرطی که مقدار تابع $f(x_{n-1})$ مورد نیاز از آخرین بارست حفظ‌شده باشد]. روش نیوتن در هر بارست پرخرجتر است. از طرف دیگر روش نیوتن سریعتر همگرامی‌شود [مرتبهٔ همگرایی در روش نیوتن $p = 2$ و در روش خط قاطع $p \doteq 1.62$]، و در نتیجه برای یک دقت مطلوب نیاز به بارستهای کمتری دارد. یک تحلیل از تأثیر این دو تفاوت در روشهای خط قاطع و نیوتن در زیر داده شده است.

اکنون زمان لازم برای دستیابی به ریشهٔ مطلوب α در یک بازهٔ تحمّل ε را در نظر می‌گیریم. برای ساده کردن تحلیل خود، فرض می‌کنیم که حدسهای اولیه به ریشهٔ مطلوب کاملاً نزدیک باشند. تعریف می‌کنیم

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad n \geq 0$$

$$\overline{x}_{n+1} = \overline{x}_n - f(\overline{x}_n).\frac{\overline{x}_n - \overline{x}_{n-1}}{f(\overline{x}_n) - f(\overline{x}_{n-1})} \qquad n \geq 1$$

و می‌گیریم $\overline{x}_0 = x_0$. $\overline{x}_1$ را بر پایهٔ فرمول همگرایی زیرتعریف می‌کنیم. از (۲ـ۲ـ۳) و (۲ـ۳ـ۱۰)،

به ترتیب به‌دست می‌آوریم:

$$| \alpha - x_{n+1} | \doteq c \, | \alpha - x_n |^{2} \qquad n \geq 0, \quad c = \left| \frac{f''(\alpha)}{2 f'(\alpha)} \right|$$

$$| \alpha - \overline{x}_{n+1} | \doteq d \, | \alpha - \overline{x}_n |^{r} \qquad r = \frac{1 + \sqrt{5}}{2}, \quad d = c^{r-1}$$

برای خطا در بارستهای نیوتن به طریق استقرا

$$c \, | \alpha - x_{n+1} | \doteq (c \, | \alpha - x_n |)^{2}$$

$$c \, | \alpha - x_n | \doteq (c \, | \alpha - x_0 |)^{2^n}$$

$$| \alpha - x_n | \doteq \frac{1}{c} (c \, | \alpha - x_0 |)^{2^n} \qquad n \geq 0$$

همچنین برای بارستهای روش قاطع داریم:

$$| \alpha - \overline{x}_n | \doteq d \, | \alpha - \overline{x}_{n-1} |^{r}$$

$$\doteq d^{1+r+\cdots+r^{n-1}} \, | \alpha - \overline{x}_0 |^{r^n}$$

با استفاده از فرمول (۱ـ۱ـ۹) برای یک سری متناهی هندسی، چنین به‌دست می‌آوریم

$$d^{1+r+\cdots+r^{n-1}} = d^{(r^n-1)/(r-1)} = c^{r^n-1}$$

و بنابراین

$$| \alpha - \overline{x}_n | \doteq c^{r^n-1} \, | \alpha - \overline{x}_0 |^{r^n} = \frac{1}{c} [c \, | \alpha - x_0 |]^{r^n}$$

برای اینکه نابرابری $| \alpha - x_n | \leq \varepsilon$ برقرار باشد، برای بارستهای نیوتن، باید داشته باشیم

$$(c \, | \alpha - x_0 |)^{2^n} \leq c\varepsilon$$

$$n \geq \frac{K}{\log 2} \qquad K = \log \left[\frac{\log \varepsilon c}{\log c \, | \alpha - x_0 |} \right]$$

گیریم m زمان محاسبهٔ $f(x)$ باشد، و $s.m$ زمان محاسبهٔ $f'(x)$. در این صورت حداقل زمان لازم برای به‌دست آوردن دقّت مورد نظر با روش نیوتن عبارت‌است از

$$T_N = (m + ms)n = \frac{(1+s)mK}{\log 2} \qquad (۱۴.۳.۲)$$

برای روش قاطع، یک محاسبهٔ مشابه نشان می‌دهد که $|\alpha - \overline{x}_n| \leq \varepsilon$، اگر

$$n \geq \frac{K}{\log r}$$

بنابراین کمترین زمان لازم برای به‌دست آوردن دقت مورد نظر عبارتست از

$$T_S = mn = \frac{mK}{\log r} \qquad (۱۵.۳.۲)$$

برای مقایسهٔ زمانهای روش خط قاطع و روش نیوتن داریم

$$\frac{T_S}{T_N} = \frac{\log ۲}{(۱ + s)\log r}$$

روش خط قاطع سریعتر از روش نیوتن است اگر این کسر کوچکتر از واحد باشد،

$$\frac{T_S}{T_N} < ۱$$

$$s > \frac{\log ۲}{\log r} - ۱ \doteq ۰ر۴۴ \qquad (۱۶.۳.۲)$$

اگر زمان برای محاسبهٔ $f'(x)$ بیش از ۴۴ درصد زمان لازم برای محاسبهٔ $f(x)$ باشد، روش خط قاطع کاراتر است. در عمل، عوامل خیلی زیادتری در هزینه‌های نسبی دو روش اثر می‌گذارند، لذا باید با احتیاط از عامل ۴۴ر۰ استفاده شود.

استدلال قبلی در نشان دادن اینکه سرعت ریاضی همگرایی تصویرکامل مسأله نیست، مفید است. زمان کلی محاسبات، کاربرد آسان یک الگوریتم، پایداری، و عوامل دیگر نیز موجب‌برتری نسبی یک الگوریتم بر دیگری می‌شود.

۴.۲ روش مولر

روش مولر برای پیدا کردن ریشه‌های حقیقی و ریشه‌های مختلط یک تابع مفید، و از نظر اجرای رایانه‌ای، معقول و ساده است. آن را به‌دست می‌آوریم، در همگرایی آن بحث می‌کنیم و چند مثال عددی می‌آوریم.

روش مولر تعمیم راهی است که به روش خط قاطع منجر شده است. سه نقطهٔ $x_۰$، $x_۱$ و $x_۲$ داده شده‌اند، چند جمله‌ای درجهٔ دوّمی می‌سازیم که از سه نقطهٔ $(x_i, f(x_i))$، $i = ۰, ۱, ۲$ بگذرد؛ یکی از ریشه‌های این چند جمله‌ای به عنوان یک برآورد بهبود یافته برای ریشهٔ αی $f(x)$ به‌کار برده می‌شود.

این چند جمله‌ای درجهٔ دوم با رابطهٔ زیر داده می‌شود.

$$p(x)=f(x_۲)+(x-x_۲)f[x_۲,x_۱]+(x-x_۲)(x-x_۱)f[x_۲,x_۱,x_۰] \quad (۱.۴.۲)$$

تفاضلات منقسم $f[x_۲,x_۱]$ و $f[x_۲,x_۱,x_۰]$ در (۱۳.۱.۱) در فصل ۱ تعریف شده‌اند. برای کنترل برقراری

$$p(x_i)=f(x_i) \qquad i=۰,۱,۲$$

تنها x_i را در (۱.۴.۲) قرار می‌دهیم و عبارت نتیجه را با استفاده از (۱۳.۱.۱) خلاصه می‌کنیم. فرمولهای دیگری برای $p(x)$ وجود دارند که در فصل ۳ داده شده‌اند، ولی شکلی که در (۱.۴.۲) نشان داده شده مناسبترین شکل برای تعریف روش مولر است. فرمول (۱.۴.۲) را شکل تفاضلات منقسم نیوتن در درون‌یابی چند جمله‌ای گویند و در بخش ۳ـ۲ی فصل ۳ به‌طور کلی بحث شده است.

برای پیدا کردن صفرهای (۱.۴.۲) ما آن را به شکل مناسبتری بازنویسی می‌کنیم

$$\begin{aligned} y &= f(x_۲)+w(x-x_۲)+f[x_۲,x_۱,x_۰](x-x_۲)^۲ \\ w &= f[x_۲,x_۱]+(x_۲-x_۱)f[x_۲,x_۱,x_۰] \\ &= f[x_۲,x_۱]+f[x_۲,x_۰]-f[x_۰,x_۱] \end{aligned}$$

می‌خواهیم کوچکترین مقدار $x-x_۲$ را پیدا کنیم که در معادلهٔ $y=۰$ صدق کند، بنابراین ریشه‌ای از (۱.۴.۲) را پیدا می‌کنیم که نزدیکترین مقدار به $x_۲$ باشد. جواب چنین است

$$x-x_۲=\frac{-w\pm\sqrt{w^۲-۴f(x_۲)f[x_۲,x_۱,x_۰]}}{۲f[x_۲,x_۱,x_۰]}$$

با انتخاب علامتی که صورت را در حدممکن کوچک نماید. به علّت خطاهای کاهش ارقام بامعنی‌که به‌طور ضمنی در این فرمول وجود دارد، صورت را گویا می‌کنیم، تا فرمول بارستی تازه‌ای به‌دست آوریم

$$x_۳=x_۲-\frac{۲f(x_۲)}{w\pm\sqrt{w^۲-۴f(x_۲)f[x_۲,x_۱,x_۰]}} \qquad (۲.۴.۲)$$

با انتخاب علامتی که مقدار مخرج را حداکثر نماید.

برای تعریف دنبالهٔ بارستهای $\{x_n;n\le ۰\}$، (۲.۴.۲) را بطور بازگشتی تکرار می‌کنیم. اگر این بارستها به یک نقطهٔ α همگرا شوند و اگر $f'(\alpha)\neq ۰$، آنگاه α یک ریشهٔ $f(x)$

خواهد بود. برای مشاهدهٔ این امر از (۱۴.۱.۱) فصل ۱ و (۲.۴.۲) استفاده می‌کنیم تا به‌دست آوریم

$$\text{وقتی} \quad n \to \infty \qquad w \to f'(\alpha)$$

$$\alpha = \alpha - \frac{2f(\alpha)}{f'(\alpha) \pm \sqrt{[f'(\alpha)]^2 - 2f(\alpha)f''(\alpha)}}$$

که نشان می‌دهد کسر طرف راست باید صفر باشد. چون بنا به فرض $f'(\alpha) \neq 0$، روش انتخاب علامت در مخرج مستلزم این است که مخرج مخالف صفر باشد. پس صورت باید صفر باشد، که نشان می‌دهد $f(\alpha) = 0$. فرض $f'(\alpha) \neq 0$ می‌گوید که α یک ریشهٔ ساده است (بخش ۷.۲ را برای بحث دربارهٔ ریشه‌های ساده و چندگانه ببینید).

با استدلالی مشابه استدلالی که در روش خط قاطع به‌کار برده شد، می‌توان نشان داد که

$$\lim_{n\to\infty} \frac{|\alpha - x_{n+1}|}{|\alpha - x_n|^p} = \left|\frac{f^{(2)}(\alpha)}{6f'(\alpha)}\right|^{(p-1)/2} \qquad p \doteq 1.84 \qquad (۳.۴.۲)$$

به شرطی که $f(x)$ در یک همسایگی α سه بار پیوسته مشتقپذیر باشد و $f'(\alpha) \neq 0$. مرتبهٔ p ریشهٔ مثبت

$$x^3 - x^2 - x - 1 = 0$$

است.

با روش خط قاطع، انتخاب مقادیر حقیقی برای x_0 و x_1 به مقدار حقیقی x_2 می‌انجامد. ولی با روش مولر، انتخاب مقادیر حقیقی x_0، x_1، x_2، می‌تواند و باید به ریشه‌های مختلط $f(x)$ منتهی شود. این یک جنبهٔ مهم روش مولر است و یک دلیل برای استفاده از آن.

مثالهای ذیل با استفاده از یک برنامهٔ بازرگانی محاسبه شده‌اند که به خودی خود یک پیاده‌سازی روش مولر را به‌دست می‌دهد. بدون اینکه حدسهای اولیه داده شده باشند، با این برنامه ریشه‌های $f(x)$ را تقریباً با یک ترتیب افزایشی پیدا می‌کند. پس از آنکه تقریبهای z_1، z_2، ... به‌عنوان ریشه‌ها پیدا می‌شوند، از تابع

$$g(x) = \frac{f(x)}{(x - z_1)\dots(x - z_r)} \qquad (۴.۴.۲)$$

برای پیدا کردن بقیهٔ ریشه‌های $f(x)$ استفاده می‌شود. [برای بحث در خطاهای این استفاده از $g(x)$، به پیترز و ویلکینسن (۱۹۷۱) مراجعه کنید]. برای آنکه ریشهٔ تقریبی z در برنامه قابل قبول باشد، باید در یکی از شرایط زیر (که استفاده کننده مشخص می‌کند) صدق کند:

جدول ۴.۲ روش مولر، مثال ۲

IT	ریشه	f(ریشه)
۹	۱٫۱۵۷۲۲۱۱۷۳۶E − ۱	۵٫۹۶E − ۸
۱۰	۶٫۱۱۷۵۷۴۸۴۵۲E − ۱ + ۹٫۰۱E − ۲۰i	−۲٫۹۸E − ۷ + ۹٫۰۶E − ۱۱i
۱۴	۲٫۸۳۳۷۵۱۳۳۷۷E۰ − ۵٫۰۵E − ۱۷i	۲٫۵۵E − ۵ − ۴٫۷۸E − ۸i
		۷٫۱۳E − ۵ + ۹٫۳۷E − ۶i
۱۳	۴٫۵۹۹۲۲۷۶۳۹۴E۰ − ۵٫۹۵E − ۱۵i	۷٫۱۳E − ۵ + ۹٫۳۷E − ۶i
۸	۱٫۵۱۲۶۱۰۲۶۹۸E۰ + ۲٫۹۸E − ۱۶i	۳٫۳۴E − ۶ − ۲٫۳۵E − ۷i
۱۹	۱٫۳۰۰۶۰۵۴۹۹eE۱ + ۹٫۰۴E − ۱۸i	۲٫۳۲E − ۱ + ۴٫۱۵E − ۷i
۱۶	۹٫۶۲۱۳۱۶۸۴۲۵E۰ − ۴٫۹۷E − ۱۷i	−۳٫۶۶E − ۲ − ۵٫۳۸E − ۷i
۱۴	۱٫۷۱۱۶۸۵۵۱۸۷E۱ − ۸٫۴۸E − ۱۷i	−۱٫۶۸E + ۰ + ۲٫۴۰E − ۵i
۱۳	۲٫۲۱۵۱۰۹۰۳۷۹E۱ + ۹٫۳۵E − ۱۸i	۸٫۶۱E − ۱ + ۲٫۶۰E − ۵i
۷	۶٫۸۴۴۵۲۵۴۵۳۱E۰ − ۳٫۴۳E − ۲۸i	−۴٫۴۹E − ۳ − ۱٫۲۲E − ۱۸i
۴	۲٫۸۴۸۷۹۶۷۲۵۱E۱ + ۵٫۷۷E − ۲۵i	−۶٫۳۴E + ۱ − ۲٫۹۶E − ۱۱i
۴	۳٫۷۰۹۹۱۲۱۰۴۴E۱ + ۲٫۸۰E − ۲۴i	۲٫۱۲E + ۳ + ۷٫۷۲E − ۹i

جدول ۵.۲ روش مولر، مثال ۳

IT	ریشه	f(ریشه)
۴۱	۲٫۹۹۸۷۵۲۶ − ۶٫۹۸E − ۴i	−۳٫۳۳E − ۱۱ + ۶٫۷۰E − ۱۱i
۱۷	۲٫۹۹۹۷۵۹۱ − ۲٫۶۸E − ۴i	۵٫۶۸E − ۱۴ − ۶٫۴۸E − ۱۴i
۳۱	۳٫۰۰۰۳۰۹۵ − ۳٫۱۷E − ۴i	−۳٫۴۱E − ۱۳ + ۳٫۲۲E − ۱۴i
۱۰	۳٫۰۰۰۳۰۴۶ + ۳٫۱۴E − ۴i	۳٫۹۸E − ۱۳ − ۳٫۸۳E − ۱۴i
۶	۵٫۹۷E − ۱۵ + ۳٫۰۰۰۰۰۰۰۰۰i	۴٫۳۸E − ۱۱ − ۱٫۱۹E − ۱۱i
۳	۵٫۹۷E − ۱۵ − ۳٫۰۰۰۰۰۰۰۰۰i	۴٫۳۸E − ۱۱ + ۱٫۱۹E − ۱۱i

۱. $| f(x) | \leq ۱۰^{-۱۰}$

۲. دقت z تا ۸ رقم بامعنی باشد.

در جدولهای ۴.۲ و ۵.۲، ریشه‌ها به همان ترتیبی که پیدا شدند داده شده‌اند. ستون IT تعداد بارستها را که برای هر ریشه محاسبه شده‌اند می‌دهد. مثالها همگی برای یک‌چند جمله‌ای $f(x)$ داده شده‌اند، ولی برنامه برای حالت کلی $f(x)$ طرح‌ریزی شده و x می‌تواند مختلط باشد.

مثال ۱. $f(x) = x^{۲۰} - ۱$. همهٔ ۲۰ ریشه با دقّت ده رقم بامعنی یابیشتر پیدا شده‌اند. در همهٔ حالات، ریشهٔ تقریبی z در $|f(z)| < ۱۰^{-۱۰}$ و به‌طورکلّی بسیارکمتر صدق می‌کرده است. تعداد بارستها از ۱ تا ۱۸ تغییرکرده و دارای میانگینی برابر ۸٫۵ بوده است.

۲. $f(x) =$ چند جمله‌ای درجهٔ ۱۲ لاگر[1]. اجزای حقیقی ریشه‌ها، همان‌گونه‌که نشان داده شده است، به تعداد ارقامی که گرد شده‌اند، درست بوده‌اند، ولی اجزای انگاری باید صفر می‌بودند. نتایج عددی در جدول ۴.۲ داده شده‌اند. توجه نمایید که برای بسیاری از ریشه‌های تقریبی، $f(x)$ کاملاً بزرگ است.

۳.

$$f(x) = x^{۶} - ۱۲x^{۵} + ۶۳x^{۴} - ۲۱۶x^{۳} + ۵۶۷x^{۲} - ۹۷۲x + ۷۲۹$$
$$= (x^{۲} + ۹)(x - ۳)^{۴}$$

نتایج عددی در جدول ۵.۲ داده شده‌اند. به عدم دقت در اولین چهارریشه توجه نمایید، که به لحاظ نوفهٔ $f(x)$ مربوط به مکرر بودن ریشهٔ $\alpha = ۳$، ذاتی است. بخش ۷.۲ را برای یک بحث کامل در مسألهٔ محاسبهٔ ریشه‌های مکرّر ببینید.

دو مثال اخیر نشان می‌دهند که چرا دو آزمون خطا لازم‌اند، و اشاره می‌کنند که چرا در این روش کار برای محاسبهٔ هر ریشه به یک تعداد حداکثر مجاز برای تعداد بارستها نیاز است. شکل (۲.۴.۲) روش مولر مربوط به تراوب[2] (۱۹۶۴، صص ۲۱۰ تا ۲۱۳) است. برای بحث محاسباتی، ویتلی[3] (۱۹۶۸) را ببینید.

۵.۲ یک نظریهٔ کلی برای روشهای بارستی تک نقطه‌ای

اکنون حل یک معادلهٔ $x = g(x)$ را برای پیدا کردن یک ریشهٔ α از راه بارستی

$$x_{n+۱} = g(x_n) \qquad n \geq ۰ \qquad (۱.۵.۲)$$

در نظر می‌گیریم که در آن $x_۰$ یک حدس اولیه برای α است. روش نیوتن برای این الگو چنین می‌شود

$$g(x) \equiv x - \frac{f(x)}{f'(x)} \qquad (۲.۵.۲)$$

هر جواب $x = g(x)$ یک نقطهٔ ثابت g خوانده می‌شود. اگرچه ما معمولاً علاقه‌مند به حل معادلهٔ $f(x) = ۰$ هستیم، ولی راههای گوناگونی برای نوشتن‌این معادله به شکل یک مسألهٔ نقطهٔ ثابت وجود دارد. در اینجا این فرایند بازنویسی را تنها با چند مثال نشان خواهیم داد.

1. Laguerre 2. Traub 3. Whitley

جدول ۶.۲ مثالهای بارستی برای $x^۲ - ۳ = ۰$

	حالت(۱)	حالت(۲)	حالت (۳)
n	x_n	x_n	x_n
۰	۲٫۰	۲٫۰	۲٫۰
۱	۳٫۰	۱٫۵	۱٫۷۵
۲	۹٫۰	۲٫۰	۱٫۷۳۲۱۴۳
۳	۸۷٫۰	۱٫۵	۱٫۷۳۲۰۵۱

مثال حل معادلهٔ $x^۲ - a = ۰$ را برای $a > ۰$ در نظر می‌گیریم.

(i) $x = x^۲ + x - a$، یا به‌صورت کلّیتر، به ازای مقداری چون $c \neq ۰$، $x = x + c(x^۲ - a)$.

(ii) $x = a/x$.

(iii)

$$x = \frac{۱}{۲}\left(x + \frac{a}{x}\right) \qquad (۳.۵.۲)$$

یک مثال عددی با $a = ۳$، $x_۰ = ۲$ و $\alpha = \sqrt{۳} = ۱٫۷۳۲۰۵۱$ می‌آوریم. با $x_۰ = ۲$، نتایج عددی برای (۱.۵.۲) در این حالتها در جدول ۶.۲ داده شده است.

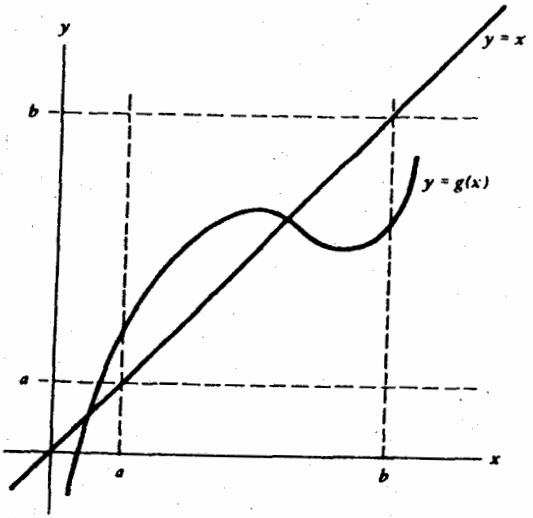
طبیعی است که سؤال شود چه چیزی موجب می‌شود که رفتار برنامه‌های بارستی مختلف به گونه‌ای باشد که دراین مثال دیده می‌شود. یک نظریهٔ کلی به‌دست می‌دهیم که این رفتار را توضیح دهد و به تحلیل روشهای بارستی جدید کمک نماید.

لم ۴.۲ گیریم $g(x)$ در بازهٔ $a \leq x \leq b$ پیوسته باشد و فرض می‌کنیم برای هر $a \leq x \leq b$ داشته باشیم $a \leq g(x) \leq b$. (گوییم g، $[a, b]$ را به خود $[a, b]$ می‌فرستد و آن را با $g([a, b]) \subset [a, b]$ نشان می‌دهیم.) در این صورت $x = g(x)$ حداقل یک جواب در $[a, b]$ دارد.

برهان تابع پیوستهٔ $g(x) - x$ را در نظر می‌گیریم. در $x = a$ این تابع مثبت و در $x = b$ منفی است. پس طبق قضیهٔ مقدار میانی، باید حداقل یک ریشه در بازهٔ $[a, b]$ داشته باشد. در شکل ۵.۲، ریشه‌ها نقاط برخورد $y = x$ با $y = g(x)$ هستند. ■

لم ۵.۲ گیریم $g(x)$ روی $[a, b]$ پیوسته باشد و فرض می‌کنیم $g([a, b] \subset [a, b])$. به علاوه فرض می‌کنیم ثابتی چون $۰ < \lambda < ۱$ وجود دارد که

$$| g(x) - g(y) | \leq \lambda | x - y | \qquad x, y \in [a, b] \text{ به ازای جمیع مقادیر } \quad (۴.۵.۲)$$



شکل ۵.۲ مثال لم ۴.۲

پس $x = g(x)$ یک جواب یکتای α در $[a, b]$ دارد. همچنین بارستهای

$$x_n = g(x_{n-1}) \qquad n \geq 1$$

برای هر انتخابی از x_0 در $[a, b]$ به α همگراست و

$$| \alpha - x_n | \leq \frac{\lambda^n}{1 - \lambda} | x_1 - x_0 | \qquad (5.5.2)$$

برهان فرض می‌کنیم $x - g(x)$ دو جواب α و β در $[a, b]$ دارد. پس

$$| \alpha - \beta | = |g(\alpha) - g(\beta)| \leq \lambda | \alpha - \beta |$$
$$(1 - \lambda) | \alpha - \beta | \leq 0$$

از آنجا که $0 < \lambda < 1$، نتیجه می‌شود که $\alpha = \beta$. همچنین با توجه به لم قبلی می‌دانیم که حداقل یک ریشهٔ α در $[a, b]$ موجود است.

برای آزمودن همگرایی بارستهای x_n، ابتدا توجه نمایید که آنها همگی در $[a, b]$ باقی می‌مانند. برای مشاهدهٔ این امر، توجه داشته باشید که می‌توان با استفاده از این که با استقراء ریاضی اثبات

نمود که به ازای همهٔ مقادیر n، $x_n \in [a,b]$. برای همگرایی،

$$| \alpha - x_{n+1} | = | g(\alpha) - g(x_n) | \leq \lambda | \alpha - x_n | \qquad (۶.۵.۲)$$

و با استقرا

$$| \alpha - x_n | \leq \lambda^n | \alpha - x_0 | \qquad n \geq 0 \qquad (۷.۵.۲)$$

که وقتی $n \to \infty$، $\lambda^n \to 0$ ؛ بنابراین $x_n \to \alpha$.

برای اثبات کران (۵.۵.۲)، با

$$| \alpha - x_0 | \leq | \alpha - x_1 | + | x_1 - x_0 | \leq \lambda | \alpha - x_0 | + | x_1 - x_0 |$$

$x_{n+1} = g(x_n) \in [a,b]$ ایجاب می‌کند $x_n \in [a,b]$

آغاز می‌کنیم که در آخرین مرحله از (۶.۵.۲) استفاده شده‌است. سپس آن را نسبت به $| \alpha - x_0 |$ حل می‌کنیم، داریم

$$| \alpha - x_0 | \leq \frac{1}{1-\lambda} | x_1 - x_0 | \qquad (۸.۵.۲)$$

از ترکیب این رابطه با (۷.۵.۲) اثبات کامل می‌شود. ■

کران (۶.۵.۲) نشان می‌دهد که دنبالهٔ $\{x_n\}$ به‌طور خطی همگراست، و بر پایهٔ تعریف (۱۳.۰.۲) نرخ همگرایی با کران λ محدود می‌شود. همچنین از برهان فوق، می‌توانیم یک کران خطا شاید دقیقتر از (۵.۵.۲) پیدا کنیم. از تکرار برهانی که به (۸.۵.۲) انجامید، به‌دست می‌آوریم

$$| \alpha - x_n | \leq \frac{1}{1-\lambda} | x_{n+1} - x_n |$$

وانگهی، استفاده از (۶.۵.۲) کران زیر را به‌دست می‌دهد

$$| \alpha - x_{n+1} | \leq \frac{\lambda}{1-\lambda} | x_{n+1} - x_n | \qquad (۹.۵.۲)$$

وقتی λ قابل محاسبه باشد، این رابطه یک کران عملی در بسیاری از حالات ارائه می‌دهد. کرانهای دیگر خطا و تخمینها در بخش بعدی مورد بحث قرار گرفته‌اند.

اگر $g(x)$ در $[a,b]$ مشتقپذیر باشد، آنگاه برای تمام $x,y\in[a,b]$

$$g(x)-g(y)=g'(\xi)(x-y) \qquad \xi \text{ بین } y,x$$

تعریف می‌کنیم

$$\lambda=\underset{a\le x\le b}{\mathrm{Max}}\mid g'(x)\mid$$

پس برای تمام مقادیر $x,y\in[a,b]$

$$\mid g(x)-g(y)\mid\le\lambda\mid x-y\mid$$

قضیهٔ ۶.۲ فرض کنید که $g(x)$ بر $[a,b]$ به‌طور پیوسته مشتقپذیر باشد، و $g([a,b])\subset[a,b]$، و اینکه

$$\lambda=\underset{a\le x\le b}{\mathrm{Max}}\mid g'(x)\mid<1 \qquad (۱۰.۵.۲)$$

در این صورت

(i) $x=g(x)$ یک جواب یکتای α در $[a,b]$ دارد.

(ii) برای هر انتخاب x_0 در $[a,b]$ با $x_{n+1}=g(x_n)$، $n\ge0$

$$\underset{n\to\infty}{\mathrm{Lim}}x_n=\alpha$$

(iii)
$$\mid\alpha-x_n\mid\le\lambda^n\mid\alpha-x_0\mid\le\frac{\lambda^n}{1-\lambda}\mid x_1-x_0\mid$$

$$\underset{n\to\infty}{\mathrm{Lim}}\frac{\alpha-x_{n+1}}{\alpha-x_n}=g'(\alpha) \qquad (۱۱.۵.۲)$$

برهان همهٔ نتیجه‌ها از لم‌های قبلی به‌دست می‌آیند مگر برای نرخ همگرایی (۱۱.۵.۲). برای آن، رابطهٔ زیر را به‌کار می‌بریم

$$\alpha-x_{n+1}=g(\alpha)-g(x_n)=g'(\xi_n)(\alpha-x_n) \qquad n\ge0 \quad (۱۲.۵.۲)$$

که ξ_n نقطهٔ مجهولی بین α و x_n است. چون $x_n\to\alpha$، باید داشته‌باشیم $\xi_n\to\alpha$، و بنابراین

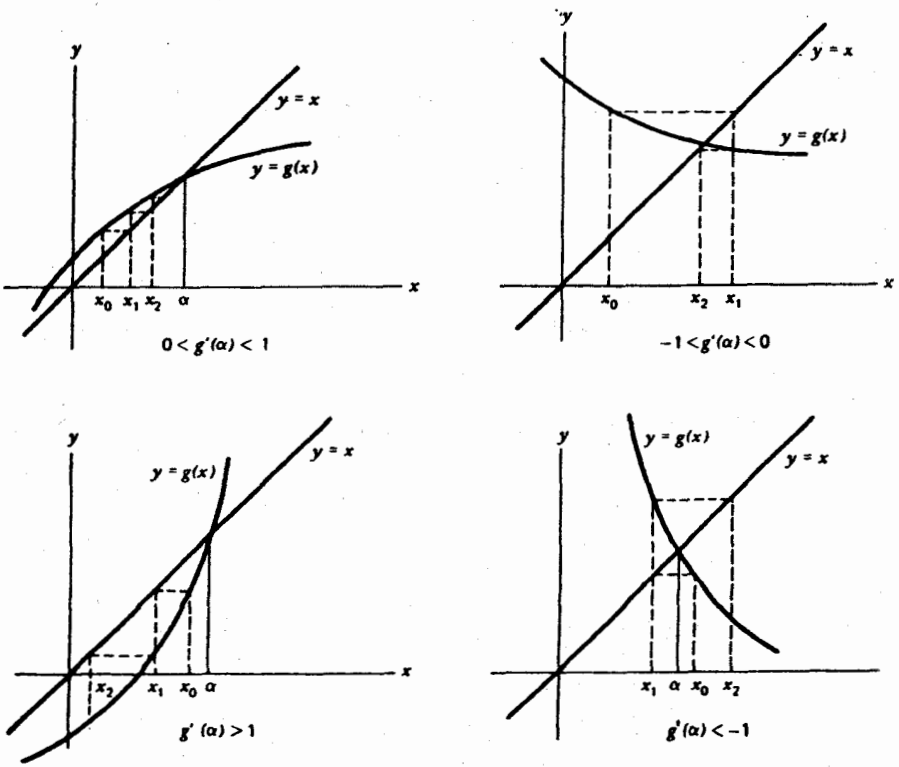
$$\underset{n\to\infty}{\mathrm{Lim}}\frac{\alpha-x_{n+1}}{\alpha-x_n}=\underset{n\to\infty}{\mathrm{Lim}}g'(\xi_n)=g'(\alpha)$$

اگر $g'(\alpha)\neq0$، آنگاه دنبالهٔ $\{x_n\}$ با مرتبهٔ دقیقاً $p=1$، یعنی همگرایی خطی، به α همگرا می‌شود. ■

این قضیه برای دستگاههای m معادلهٔ غیر خطی با m مجهول تعمیم می‌یابد. کافی است x را یک عنصر $\mathbf{R}^m$، $g(x)$ را تابعی از $\mathbf{R}^m$ به $\mathbf{R}^m$ بگیریم و به‌جای قدر مطلقها نرمهای برداری و ماتریسی و به جای $g'(x)$ ماتریس ژاکوبی $g(x)$ را قرار دهیم. فرض $g([a,b]) \subset [a,b]$ باید با فرض قویتری جایگزین شود و برای انتخاب ناحیه‌ای که $[a,b]$ را تعمیم دهد باید به دقت عمل شود. لِم‌ها تعمیم می‌یابند ولی اثبات آنها آسان نخواهد بود. در این مورد در بخش ۲ـ۱۰ بیشتر توضیح داده شده است.

برای ملاحظهٔ اهمیّت فرض (۱۰.۵.۲) در مورد اندازهٔ $g'(x)$، فرض کنید $|g'(\alpha)| > 1$. در این صورت اگر دنباله‌ای از بارستهای $x_{n+1} = g(x_n)$ می‌داشتیم و $\alpha = g(\alpha)$ یک ریشه بود، (۱۲.۵.۲) را می‌داشتیم. اگر x_n به اندازهٔ کافی به α نزدیک شود، آنگاه $|g'(\xi_n)| < 1$ و خطای $|\alpha - x_{n+1}|$ از $|\alpha - x_n|$ بزرگتر می‌شود. پس اگر $|g'(\alpha)| > 1$، همگرایی میّسر نخواهد بود. ما محاسبهٔ بارستها را در چهار حالت با نمودار نشان داده‌ایم (شکل ۶.۲ را ملاحظه کنید).

برای ساده کردن کاربرد قضیهٔ قبلی، قضیهٔ زیر را می‌آوریم.



شکل ۶.۲ مثالهایی از دنباله‌های همگرا و ناهمگرای $x_{n+1} = g(x_n)$

قضیهٔ ۷.۲ گیریم α یک جواب $x = g(x)$ باشد و فرض می‌کنیم $g(x)$ در بازه‌ای از همسایگی α به‌طور پیوسته مشتقپذیر با $| g'(\alpha) | < ۱$ باشد در این صورت نتایج قضیهٔ ۶.۲ کماکان درست‌اند به شرطی که $x_۰$ به اندازهٔ کافی به α نزدیک انتخاب شود.

برهان یک عدد λ که $| g'(\alpha) | < \lambda < ۱$ را برقرار کند انتخاب می‌کنیم، سپس یک بازهٔ $I = [\alpha - \varepsilon, \alpha + \varepsilon]$ با

$$\underset{x \in I}{\text{Max}} \mid g'(x) \mid \leq \lambda < ۱$$

را در نظر می‌گیریم. داریم $g(I) \subset I$، زیرا $| \alpha - x | \leq \varepsilon$ ایجاب می‌کند که داشته باشیم

$$| \alpha - g(x) | = | g(\alpha) - g(x) | = | g'(\xi) | \cdot | \alpha - x | \leq \lambda | \alpha - x | \leq \varepsilon$$

اکنون قضیهٔ قبلی را با استفاده از $[a, b] = [\alpha - \varepsilon, \alpha + \varepsilon]$ به‌کار می‌بریم. ■

مثال با مراجعه به مثال پیشین در این بخش، $g'(\alpha)$ را محاسبه می‌کنیم.

(i) $g(x) = x^۲ + x - ۳ \quad g'(\alpha) = g'(\sqrt{۳}) = ۲\sqrt{۳} + ۱ > ۱$

(ii) $g(x) = \frac{۳}{x} \quad g'(\sqrt{۳}) = \frac{-۳}{(\sqrt{۳})^۲} = -۱$

(iii) $g(x) = \frac{۱}{۲}(x + \frac{۳}{x}) \quad g'(x) = \frac{۱}{۲}(۱ - \frac{۳}{x^۲}) \quad g'(\sqrt{۳}) = ۰$

مثال برای $x = x + c(x^۲ - ۳)$ مقدار c را به گونه‌ای انتخاب می‌کنیم که همگرایی تضمین شود. چون جواب $\alpha = \sqrt{۳}$ است و چون $g'(x) = ۱ + ۲cx$، c باید طوری باشد که

$$-۱ < ۱ + ۲c\sqrt{۳} < ۱$$

برای یک نرخ خوب همگرایی، c را طوری می‌گیریم که

$$۱ + ۲c\sqrt{۳} \doteq ۰$$

و این انتخاب

$$c \doteq \frac{-۱}{۲\sqrt{۳}}$$

جدول ۷.۲ مثال عددی از بارستیِ (۱۳.۵.۲)

n	x_n	$\alpha - x_n$	نسبت
۰	۲٫۰	$-۲٫۶۸\mathrm{E}-۱$	
۱	۱٫۷۵	$-۱٫۷۹\mathrm{E}-۲$	۰٫۰۶۶۸
۲	۱٫۷۳۴۳۷۵۰	$-۲٫۳۲\mathrm{E}-۳$	۰٫۱۳۰
۳	۱٫۷۳۲۳۶۰۸	$-۳٫۱۰\mathrm{E}-۴$	۰٫۱۳۴
۴	۱٫۷۳۲۰۹۲۳	$-۴٫۱۵\mathrm{E}-۵$	۰٫۱۳۴
۵	۱٫۷۳۲۰۵۶۴	$-۵٫۵۶\mathrm{E}-۶$	۰٫۱۳۴
۶	۱٫۷۳۲۰۵۱۶	$-۷٫۴۵\mathrm{E}-۷$	۰٫۱۳۴
۷	۱٫۷۳۲۰۵۰۹	$-۱٫۰۰\mathrm{E}-۷$	۰٫۱۳۴

را به‌دست می‌دهد. از $c = -\frac{۱}{۴}$ استفاده می‌کنیم. در این صورت $g'(\sqrt{۳}) = ۱ - \frac{\sqrt{۳}}{۲} \doteq ۰٫۱۳۴$ که برنامهٔ بارستی زیر را می‌دهد

$$x_{n+۱} = x_n - \frac{۱}{۴}(x_n^۲ - ۳) \qquad n \geq ۰ \qquad (۱۳.۵.۲)$$

نتایج عددی در جدول ۷.۲ داده شده‌اند. ستون نسبت‌ها مقادیر

$$\frac{\alpha - x_n}{\alpha - x_{n-۱}} \qquad n \geq ۱$$

را به‌دست می‌دهد. نتایج با مقدار نظری $g'(\sqrt{۳})$ به خوبی توافق دارد.

روشهای تک نقطه‌ای مرتبهٔ بالاتر. مبحث نظریهٔ روشهای بارستی تک نقطه‌ای را با در نظر گرفتن روشهایی که دارای مرتبهٔ همگرایی بزرگتر از یک باشند، مثلاً روش نیوتن، کامل می‌کنیم.

قضیهٔ ۸.۲ گیریم α یک ریشهٔ $x = g(x)$ باشد و $g(x)$ به ازای تمام xهای نزدیک به α و $p \geq ۲$، p بار پیوسته مشتقپذیر باشد. به علاوه گیریم

$$g'(\alpha) = \cdots = g^{(p-۱)}(\alpha) = ۰ \qquad (۱۴.۵.۲)$$

در این صورت اگر حدس اولیهٔ $x_۰$ به اندازه کافی نزدیک به α انتخاب شده باشد،بارستی

$$x_{n+۱} = g(x_n) \qquad n \geq ۰$$

دارای همگرایی مرتبهٔ p بوده و

$$\lim_{n\to\infty}\frac{\alpha - x_{n+1}}{(\alpha - x_n)^p} = (-1)^{p-1}.\frac{g^{(p)}(\alpha)}{p!}$$

برهان $g(x)$ را حول α بسط می‌دهیم؛ برای یک مقدار از ξ_n بین x_n و α داریم:

$$x_{n+1} = g(x_n) = g(\alpha) + (x_n - \alpha)g'(\alpha) + \cdots + \frac{(x_n-\alpha)^{p-1}}{(p-1)!}g^{(p-1)}(\alpha) + \frac{(x_n-\alpha)^p}{p!}g^{(p)}(\xi_n)$$

با استفاده از (۱۴.۵.۲) و $\alpha = g(\alpha)$،

$$\alpha - x_{n+1} = -\frac{(x_n-\alpha)^p}{p!}g^{(p)}(\xi_n)$$

با استفاده از قضیهٔ ۷.۲ و $x_n \to \alpha$ برهان کامل می‌شود. ■

روش نیوتن را می‌توان با این قضیه تجزیه و تحلیل نمود

$$g(x) = x - \frac{f(x)}{f'(x)} \qquad g'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}$$

$$g'(\alpha) = 0 \qquad g''(\alpha) = \frac{f''(\alpha)}{f'(\alpha)}$$

این رابطه و روابط (۱۴.۵.۲) همان نتیجهٔ همگرایی (۳.۲.۲) قبلی را برای روش نیوتن به‌دست می‌دهند. برای کاربرد قضیهٔ ۸.۲ در مثالهای دیگر، مسائل آخر این فصل را ملاحظه کنید.

نظریهٔ این بخش تنها برای روشهای بارستی تک نقطه‌ای است، بنابراین روش خط قاطع و روش مولر مورد توجه قرار نگرفته‌اند. یک نظریهٔ نقطهٔ ثابت متناظر برای روشهای چند مرحله‌ای نقطهٔ ثابت نیز وجود دارد که می‌توان در تراوب[1] (۱۹۶۲) یافت. ما در اینجا از آن صرف‌نظر کرده‌ایم، اساساً به این دلیل که در فصلهای بعدی فقط نظریهٔ بارستی نقطهٔ ثابت تک نقطه‌ای مورد نیازند.

1. Traub

۶.۲ برون‌یابی ایتکن[1] برای دنباله‌های خطی ـ همگرا

از (۱۱.۵.۲)، قضیهٔ ۶.۲ برای یک بارست همگرای $x_{n+1} = g(x_n)$، $n \geq 0$ داریم

$$\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{\alpha - x_n} = g'(\alpha) \qquad (۱.۶.۲)$$

در این فصل، توجه خود را فقط به حالت همگرایی خطی معطوف می‌سازیم. بنابراین فرض می‌کنیم

$$0 < |g'(\alpha)| < 1 \qquad (۲.۶.۲)$$

برآورد خطا را در بارستها آزمایش می‌کنیم و راهی برای شتاب دادن به همگرایی دنبالهٔ $\{x_n\}$ بیان می‌کنیم.

مسأله را با در نظر گرفتن نسبتهای زیر آغاز می‌کنیم

$$\lambda_n = \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}} \qquad n \geq 2 \qquad (۳.۶.۲)$$

ادعا می‌کنیم که:

$$\lim_{n\to\infty} \lambda_n = g'(\alpha) \qquad (۴.۶.۲)$$

برای اثبات آن، می‌نویسیم

$$\lambda_n = \frac{(\alpha - x_{n-1}) - (\alpha - x_n)}{(\alpha - x_{n-2}) - (\alpha - x_{n-1})}$$

با استفاده از (۱۲.۵.۲)

$$\lambda_n = \frac{(\alpha - x_{n-1}) - g'(\xi_{n-1})(\alpha - x_{n-1})}{(\alpha - x_{n-1})/[g'(\xi_{n-2})] - (\alpha - x_{n-1})} = \frac{1 - g'(\xi_{n-1})}{1/[g'(\xi_{n-2})] - 1}$$

$$\lim_{n\to\infty} \lambda_n = \frac{1 - g'(\alpha)}{1/[g'(\alpha)] - 1} = g'(\alpha)$$

کمیّت λ_n قابل محاسبه بوده و هرگاه عملاً به مقداری چون λ همگرا شود، می‌گیریم $\lambda = g'(\alpha)$.
از $\lambda_n \doteq g'(\alpha)$ برای برآورد خطا در بارستهای x_n استفاده می‌کنیم. گیریم

$$\alpha - x_n \doteq \lambda_n(\alpha - x_{n-1})$$

---

1. Aitken

پس

$$\begin{aligned}\alpha - x_n &= (\alpha - x_{n-1}) + (x_{n-1} - x_n)\\ &\doteq \frac{1}{\lambda_n}(\alpha - x_n) + (x_{n-1} - x_n)\\ \alpha - x_n &\doteq \frac{\lambda_n}{1-\lambda_n}(x_n - x_{n-1}) \qquad (۵.۶.۲)\end{aligned}$$

این **فرمول خطای ایتکن** برای x_n است، و دقّت آن با همگرایی $\{\lambda_n\}$ به $g'(\alpha)$ افزایش می‌یابد.

مثال بارستی زیر را در نظر می‌گیریم.

$$x_{n+1} = ۶٫۲۸ + \sin(x_n) \qquad n \geq ۰ \qquad (۶.۶.۲)$$

ریشهٔ دقیق آن $\alpha \doteq ۶٫۰۱۵۵۰۳۰۷۲۹۷$ است. نتایج بارستی، همراه با مقادیر λ_n، $\alpha - x_n$، $x_n - x_{n-1}$ و برآورد خطای (۵.۶.۲)، در جدول ۸.۲ داده شده است. مقادیر λ_n به

$$g'(\alpha) = \cos(\alpha) \doteq ۰٫۹۶۴۴$$

همگراست و برآورد (۵.۶.۲)، یک نشانگر دقیق از خطای واقعی است. اندازهٔ $g'(\alpha)$ نیز نشان می‌دهد که بارستها خیلی به کندی همگرا می‌شوند، و در این حالت، $x_{n+1} - x_n$ نشانگر دقیقی از $\alpha - x_n$ نخواهد بود.

جدول ۸.۲ بارست (۶.۶.۲)

n	x_n	$x_n - x_{n-1}$	λ_n	$\alpha - x_n$	نسبت (۵.۶.۲)
۰	۶٫۰۰۰۰۰۰۰			۱٫۵۵E − ۲	
۱	۶٫۰۰۰۵۸۴۵	۵٫۸۴۵E − ۴		۱٫۴۹E − ۲	
۲	۶٫۰۰۱۱۴۵۸	۵٫۶۱۳E − ۴	۰٫۹۶۰۳	۱٫۴۴E − ۲	۱٫۳۶E − ۲
۳	۶٫۰۰۱۶۸۴۸	۵٫۳۹۰E − ۴	۰٫۹۶۰۴	۱٫۳۸E − ۲	۱٫۳۱E − ۲
۴	۶٫۰۰۲۲۰۲۶	۵٫۱۷۸E − ۴	۰٫۹۶۰۶	۱٫۳۳E − ۲	۱٫۲۶E − ۲
۵	۶٫۰۰۲۷۰۰۱	۴٫۹۷۴E − ۴	۰٫۹۶۰۷	۱٫۲۸E − ۲	۱٫۲۲E − ۲
۶	۶٫۰۰۳۱۷۸۰	۴٫۷۸۰E − ۴	۰٫۹۶۰۹	۱٫۲۳E − ۲	۱٫۱۷E − ۲
۷	۶٫۰۰۳۶۳۷۴	۴٫۵۹۳E − ۴	۰٫۹۶۱۰	۱٫۱۸E − ۲	۱٫۱۳E − ۲

فرمول برون‌یابی ایتکن. همان (۵.۶.۲) است که به‌صورت برآوردی از α نوشته شده است:

$$\alpha \doteq x_n + \frac{\lambda_n}{1-\lambda_n}(x_n - x_{n-1}) \qquad (۷.۶.۲)$$

طرف راست را با $\hat{x}_n$ برای $n \geq ۲$ نشان می‌دهیم. با گذاردن (۳.۶.۲) در (۷.۶.۲) فرمول برای $\hat{x}_n$ را می‌توان به شکل زیر نوشت

$$\hat{x}_n = x_n - \frac{(x_n - x_{n-1})^۲}{(x_n - x_{n-1}) - (x_{n-1} - x_{n-2})} \qquad n \geq ۲ \qquad (۸.۶.۲)$$

همان فرمولی که در بسیاری از کتابهای درسی داده شده است.

مثال نتایج جدول ۸.۲ را برای فرمول بارستی (۶.۶.۲) به‌کار می‌بریم. با $n = ۷$ از (۷.۶.۲) یا از (۸.۶.۲) استفاده می‌کنیم،

$$\hat{x}_۷ = ۶\text{٫}۰۱۴۹۵۱۸ \qquad \alpha - \hat{x}_۷ = ۵\text{٫}۵۱E-۴$$

بنابراین برون‌یاب $\hat{x}_۷$ به‌طور قابل ملاحظه‌ای از $x_۷$ دقیقتر است.

اکنون بارست خطی و برون‌یابی ایتکن را در یک الگوریتم ساده ترکیب می‌نمائیم.

الگوریتم ایتکن $(g, x_۰, \varepsilon, \text{root})$

۱. توجه: فرض شده است که $|g'(\alpha)| < ۱$ و بارست خطی معمولی با استفاده از $x_۰$ به α همگراست.

۲. $x_۱ := g(x_۰)$ و $x_۲ := g(x_۱)$

۳. $\hat{x}_۲ := x_۲ - \dfrac{(x_۲ - x_۱)^۲}{(x_۲ - x_۱) - (x_۱ - x_۰)}$

۴. اگر $|\hat{x}_۲ - x_۲| \leq \varepsilon$، آنگاه $\text{root} := \hat{x}_۲$ و خارج شوید.

۵ـ $x_۰ := \hat{x}_۲$ قرار داده و به مرحلهٔ ۲ بروید.

این الگوریتم، به شرطی که مفروضات مرحلهٔ اول برقرار باشند، معمولاً همگراست.

مثال برای نمایاندن الگوریتم ایتکن، مثال قبلی (۶.۶.۲) را تکرار می‌کنیم. نتایج عددی در جدول ۹.۲ داده شده‌اند. مقادیر $x_۳$، $x_۶$ و $x_۹$ مقادیر برون‌یاب ایتکن هستند که با (۷.۶.۲) تعریف شده‌اند. مقادیر λ_n فقط برای حالتهای $n = ۲, ۵, ۸$ داده شده‌اند، زیرا فقط در این صورت، خطاهای $\alpha - x_n$، $\alpha - x_{n-۱}$ و $\alpha - x_{n-۲}$ بطور خطی کاهش می‌یابند همانگونه که برای $\lambda_n \doteq g'(\alpha)$ لازم است.

جدول ۹.۲ الگوریتم ایتکن برای (۶.۶.۲)

n	x_n	λ_n	$\alpha - x_n$
۰	۶٫۰۰۰۰۰۰۰۰۰۰۰		۱٫۵۵E − ۲
۱	۶٫۰۰۰۵۸۴۵۰۱۸۰۱		۱٫۴۹E − ۲
۲	۶٫۰۰۱۱۴۵۷۷۰۷۶۱	۰٫۹۶۰۲۵	۱٫۴۴E − ۲
۳	۶٫۰۱۴۷۰۵۱۴۷۵۴۳		۷٫۹۸E − ۴
۴	۶٫۰۱۴۷۳۳۶۴۸۷۲۰		۷٫۶۹E − ۴
۵	۶٫۰۱۴۷۶۱۱۲۸۹۵۵	۰٫۹۶۴۱۸	۷٫۴۲E − ۴
۶	۶٫۰۱۵۵۰۰۸۰۲۰۶۰		۲٫۲۷E − ۶
۷	۶٫۰۱۵۵۰۰۸۸۲۹۳۵		۲٫۱۹E − ۶
۸	۶٫۰۱۵۵۰۰۹۶۰۹۳۱	۰٫۹۶۴۳۹	۲٫۱۱E − ۶
۹	۶٫۰۱۵۵۰۳۰۷۲۹۴۷		۲٫۰۵E − ۱۱

برون‌یابی اغلب با روشهای بارستی خطیِ کُند همگرا در حل دستگاههای بزرگ معادلات خطی به‌کار برده می‌شود. روشهای واقعی که به‌کار می‌روند با روشی که در بالا توضیح داده شد، تفاوت دارند، ولی آنها نیز بر پایهٔ ایدهٔ کلّی یافتن رفتار کیفی خطا، مانند (۱.۶.۲)، و سپس به‌کار بردن آن برای برآورد بهتری برای جواب، استوارند. این ایده همچنین در بسط روشهای عددی انتگرالگیری، حل معادلات دیفرانسیل و سایر مسائل ریاضی دنبال می‌شود.

۷.۲ محاسبهٔ عددی ریشه‌های چندگانه

گوئیم $f(x)$ یک ریشهٔ α از مرتبهٔ $p > ۱$ دارد اگر

$$f(x) = (x - \alpha)^p h(x) \qquad (۱.۷.۲)$$

که در آن $h(\alpha) \neq ۰$ و $h(x)$ در $x = \alpha$ پیوسته است. p را یک عدد صحیح مثبت فرض می‌کنیم، اگرچه بعضی از مطالب زیر برای مقادیر غیر صحیح p نیز معتبرند. اگر $h(x)$ در $x = \alpha$ به اندازهٔ کافی مشتقپذیر باشد، آنگاه (۲ـ۷ـ۱) هم‌ارز است با

$$f(\alpha) = f'(\alpha) = \cdots = f^{(p-۱)}(\alpha) = ۰ \qquad f^{(p)}(\alpha) \neq ۰ \qquad (۲.۷.۲)$$

وقتی ریشهٔ تابعی را با رایانه پیدا می‌کنیم، همیشه یک بازهٔ عدم قطعیّت حول ریشه وجود دارد، و اگر ریشه چندگانه باشد، این وضع بدتر خواهد شد. برای آنکه این موضوع بهتر روشن شود، محاسبهٔ دو تابع $f_۱(x) = x^۲ - ۳$ و $f_۲(x) = ۹ + x^۲(x^۲ - ۶)$ را در نظر می‌گیریم. بنابراین

$\alpha = \sqrt{۳}$ ریشهٔ یگانه تابع $f_۱$ و ریشهٔ دوگانهٔ تابع $f_۲$ است. اگر حساب اعشاری چهاررقمی را به‌کار بریم، برای $x \leq ۱٫۷۳۱$ داریم $f_۱(x) < ۰$ و $f_۱(۱٫۷۳۲) = ۰$ و برای $x > ۱٫۷۳۳$ داریم $f_۱(x) > ۰$. ولی برای $۱٫۷۳۸ \leq x \leq ۱٫۷۲۶$ داریم $f_۲(x) = ۰$. پس مقدار دقت در یافتن ریشهٔ $f_۲(x)$ محدود شده است. مثال دوّمی برای اثر نوفه در محاسبهٔ یک ریشهٔ چندگانه برای تابع $f(x) = (x - ۱)^۳$ در شکل‌های ۱.۱ و ۲.۱ از بخش ۳.۱ فصلِ یک نشان داده شده است و برای یک مثال نهایی مثال زیر را در نظر می‌گیریم.

مثال تابع

$$f(x) = (x - ۱٫۱)^۳(x - ۲٫۱)$$
$$= x^۴ - ۵٫۴x^۳ + ۱۰٫۵۶x^۲ - ۸٫۹۵۴x + ۲٫۷۹۵۱ \qquad (۳.۷.۲)$$

را روی یک ریزکامپیوتر IBM با استفاده از حساب با دقت مضاعف (در BASIC) محاسبه می‌کنیم. ضرایب دارای بسط دودویی محدود نیستند بنابراین دقیقاً در محاسبه وارد نمی‌شوند (غیر از ضریب جملهٔ $x^۴$). محاسبهٔ چند جمله‌ای $f(x)$ به شکل بسط داده شدهٔ (۳.۷.۲) انجام شده و طرح ضرب **تودرتو** به‌کار گرفته شده است:

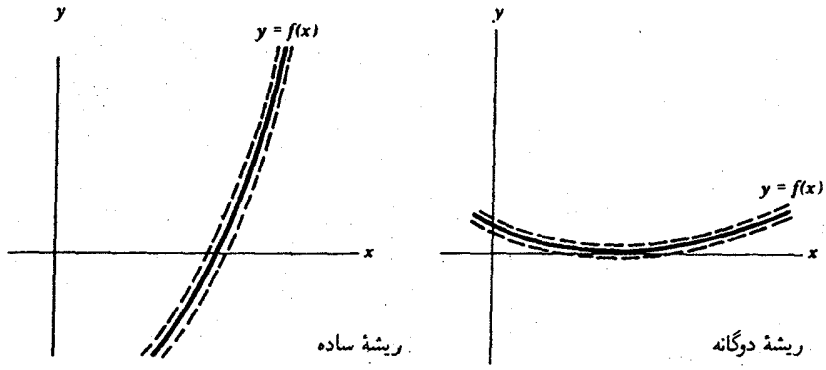
$$f(x) = ۲٫۷۹۵۱ + x(-۸٫۹۵۴ + x(۱۰٫۵۶ + x(-۵٫۴ + x))) \qquad (۴.۷.۲)$$

نتایج عددی در جدول ۱۰.۲ داده شده‌اند. توجه داشته باشید حساب به‌کار برده شده در نمایش با ممیز شناور در حدود ۱۶ رقم اعشاری دارد. پس همان گونه که از نتایج عددی جدول مشهود است، نمی‌توان دقّتی بیش از ۶ رقم اعشار در محاسبهٔ ریشهٔ $f(x)$ یعنی $\alpha = ۱٫۱$، انتظار داشت. همچنین به نتیجهٔ کاربرد دو نمایش مختلف (۳.۷.۲) و (۴.۷.۲) توجه کنید.

در محاسبهٔ هر تابع $f(x)$ به لحاظ استفاده از حساب با دقت محدود در اثر گردکردن یا قطع کردن، یک عدم قطعیّت پیدا می‌شود. این موضوع در بخش ۳ فصل یک تحت نام نوفه در محاسبهٔ تابع بحث شده است. برای ریشه‌های چندگانه، این امر به عدم قطعیّت قابل ملاحظه‌ای برای محل ریشه می‌انجامد. در شکل ۷.۲، خط پر نمودار $y = f(x)$ و خط‌چین‌ها ناحیهٔ عدم قطعیّت در محاسبهٔ $f(x)$ را که در اثر خطای گردکردن و حساب با ارقام‌محدود به‌وجود می‌آید، نشان می‌دهند. بازهٔ عدم قطعیّت در پیداکردن ریشهٔ $f(x)$، با تقاطع نوار عدم قطعیّت و محور xها داده می‌شود. روشن است که این بازه برای ریشه چندگانه بزرگتر و برای ریشهٔ ساده کوچکتر است، اگرچه پهنای قائم نوارها حول $y = f(x)$ یکی است.

جدول ۱۰.۲ ارزیابی $f(x) = (x - ۱٫۱)^۳(x - ۲٫۱)$

x	$f(x)$: (۳.۷.۲)	$f(x)$: (۴.۷.۲)
۱٫۰۹۹۹۹۲	۳٫۸۶E − ۱۶	۵٫۵۵E − ۱۶
۱٫۰۹۹۹۹۴	۳٫۸۶E − ۱۶	۲٫۷۶E − ۱۶
۱٫۰۹۹۹۹۶	۲٫۷۶E − ۱۶	۰٫۰
۱٫۰۹۹۹۹۸	−۵٫۵۵E − ۱۷	۱٫۱۱E − ۱۶
۱٫۱۰۰۰۰۰	۵٫۵۵E − ۱۷	۰٫۰
۱٫۱۰۰۰۰۲	۵٫۵۵E − ۱۷	۵٫۵۵E − ۱۷
۱٫۱۰۰۰۰۴	−۵٫۵۵E − ۱۷	۰٫۰
۱٫۱۰۰۰۰۶	−۱٫۶۷E − ۱۶	−۱٫۶۷E − ۱۶
۱٫۱۰۰۰۰۸	−۶٫۱۱E − ۱۶	−۵٫۰۰E − ۱۶



شکل ۷.۲ نوار عدم قطعیّت در محاسبهٔ یک تابع

روش نیوتن و ریشه‌های چندگانه. مسألهٔ دیگر در ریشه‌های چندگانه این است که روشهای ریشه‌یابی قبلی، وقتی که ریشهٔ مورد جستجو چندگانه باشد، به خوبی ریشهٔ ساده انجام نمی‌گیرند. اکنون این را برای روش نیوتن بررسی می‌کنیم.

روش نیوتن را به‌عنوان یک روش نقطهٔ ثابت، مانند (۲.۵.۲)، با $f(x)$ که در (۱.۷.۲) صدق می‌کند، در نظر می‌گیریم:

$$x_{n+۱} = g(x_n) \qquad g(x) = x - \frac{f(x)}{f'(x)} \qquad x \neq \alpha$$

پیش از به‌دست آوردن $g'(\alpha)$، با استفاده از (۱.۷.۲)، $g(x)$ را ساده می‌کنیم:

$$f'(x) = (x - \alpha)^p h'(x) + p(x - \alpha)^{p-۱} h(x)$$

$$g(x) = x - \frac{(x-\alpha)h(x)}{ph(x) + (x-\alpha)h'(x)}$$

با مشتقگیری

$$g'(x) = ۱ - \frac{h(x)}{ph(x) + (x-\alpha)h'(x)} - (x-\alpha).\frac{d}{dx}\left[\frac{h(x)}{ph(x) + (x-\alpha)h'(x))}\right]$$

و

$$g'(\alpha) = ۱ - \frac{۱}{p} \neq ۰ \qquad p > ۱ \text{ برای} \qquad (۵.۷.۲)$$

بنابراین روش نیوتن یک روش خطی با نرخ همگرایی $(p-۱)/p$ است.

مثال کوچکترین ریشهٔ معادلهٔ زیر را با استفاده از روش نیوتن پیدا کنید.

$$f(x) = -۴٫۶۸۹۹۹ + x(۹٫۱۳۸۹ + x(-۵٫۵۶ + x)) \qquad (۶.۷.۲)$$

نتایج عددی در جدول ۱۱.۲ نشان داده شده‌اند. محاسبات با یک ریز کامپیوتر IBM PC با دقت مضاعف (با BASIC) انجام شده‌اند. فقط قسمتهایی از نتایج آورده شده‌اند، که حالت کلی محاسبات را نشان می‌دهند. ستونی که با نسبت مشخص شده‌است، نرخ همگرایی خطی را که با λ_n در (۳.۶.۲) اندازه‌گیری می‌شود، نشان می‌دهد.

جدول ۱۱.۲ روش نیوتن برای (۶.۷.۲)

n	x_n	$f(x_n)$	$\alpha - x_n$	نسبت
۰	۱٫۲۲	−۱٫۸۸E − ۴	۱٫۰۰E − ۲	
۱	۱٫۲۲۴۹۸۶۷۳۷۴	−۴٫۷۱E − ۵	۵٫۰۱E − ۳	
۲	۱٫۲۲۷۴۹۰۰۲۲۲	−۱٫۱۸E − ۵	۲٫۵۱E − ۳	۰٫۵۰۲
۳	۱٫۲۲۸۷۴۴۱۷۰۵	−۲٫۹۵E − ۶	۱٫۲۶E − ۳	۰٫۵۰۱
۴	۱٫۲۲۹۳۷۱۸۷۴۶	−۷٫۳۸E − ۷	۶٫۲۸E − ۴	۰٫۵۰۱
۵	۱٫۲۲۹۶۸۵۸۸۴۶	−۱٫۸۵E − ۷	۳٫۱۴E − ۴	۰٫۵۰۰
۱۸	۱٫۲۲۹۹۹۹۹۶۲۱	−۲٫۸۹E − ۱۵	۳٫۸۰E − ۸	۰٫۵۰۵
۱۹	۱٫۲۲۹۹۹۹۹۸۲۳	−۶٫۶۶E − ۱۶	۱٫۷۷E − ۸	۰٫۵۲۵
۲۰	۱٫۲۲۹۹۹۹۹۹۲۴	−۱٫۱۱E − ۱۶	۷٫۵۸E − ۹	۰٫۴۹۶
۲۱	۱٫۲۲۹۹۹۹۹۹۶۳	۰٫۰	۳٫۶۶E − ۹	۰٫۳۸۳

واضح است که در این حالت روش نیوتن برای پیدا کردن ریشهٔ (۶.۷.۲)، خطی با نرخ $g'(x) = \frac{1}{2}$، با (۵.۷.۲) سازگار است، زیرا $\alpha = 1.23$ یک ریشه با چندگانی $p = 2$ است. بارستهای پایانی در جدول تحت تأثیر نوفهٔ محاسبهٔ رایانه‌ای $f(x)$ واقع شده‌اند. اگرچه نمایش با ممیز شناور تقریباً شامل ۱۶ رقم است، ولی در این حالت دقّت فقط در حدود هشت رقم است.

برای بهبود روش نیوتن، علاقه‌مندیم $g(x)$ به گونه‌ای باشد که $g'(\alpha) = 0$. با توجه به (۵.۷.۲)، تعریف می‌کنیم

$$g(x) = x - p\frac{f(x)}{f'(x)}$$

در این صورت به‌سادگی ملاحظه می‌شود که $g'(\alpha) = 0$ ؛ بنابراین

$$\begin{aligned}\alpha - x_{n+1} &= g(\alpha) - g(x_n) \\ &= -g'(\alpha)(x_n - \alpha) - \frac{1}{2}(x_n - \alpha)^2 g''(\xi_n)\end{aligned}$$

که ξ_n بین x_n و α است. بنابراین

$$\alpha - x_{n+1} = -\frac{1}{2}(\alpha - x_n)^2 g''(\xi_n)$$

که نشان می‌دهد روش

$$x_{n+1} = x_n - p\frac{f(x_n)}{f'(x_n)} \qquad n = 0, 1, 2, \ldots \qquad (۷.۷.۲)$$

همانند روش اصلی نیوتن برای ریشه‌های ساده دارای همگرایی مرتبهٔ ۲ است.

مثال رابطهٔ (۷.۷.۲) را برای مثال قبلی (۶.۷.۲)، با $P = 2$ برای یک ریشهٔ مضاعف به‌کارمی‌بریم. نتایج در جدول ۱۲.۲، با استفاده از همان رایانهٔ قبلی، نشان داده‌شده‌اند. بارستها به سرعت همگرا می‌شوند، و سپس حول ریشه نوسان می‌کنند. دقت (یا عدم آن)، نوفه در $f(x)$ و چندگانی ریشه را منعکس می‌کند.

روش نیوتن را می‌توان برای تعیین چندگانی p به‌کار برد، همانند جدول ۱۱.۲ که با (۵.۷.۲) ترکیب شده باشد، و سپس (۷.۷.۲) را می‌توان به‌کار برد و سرعت همگرایی را افزایش داد. ولی عدم قطعیّت ذاتی در ریشه که از چندگانی ریشه نتیجه می‌شود کماکان باقی می‌ماند. این نقص

جدول ۱۲.۲ روش اصلاح‌شدهٔ (۷.۷.۲) نیوتن، در حل (۶.۷.۲)

n	x_n	$f(x_n)$	$\alpha - x_n$
۰	۱٫۲۲	−۱٫۸۸E − ۴	۱٫۰۰E − ۲
۱	۱٫۲۲۹۹۷۳۴۷۴۸	−۱٫۳۱E − ۹	۲٫۶۵E − ۵
۲	۱٫۲۲۹۹۹۹۹۹۹۸	−۱٫۱۱E − ۱۶	۱٫۸۵E − ۱۰
۳	۱٫۲۳۰۰۰۰۳۲۰۸	−۱٫۹۲E − ۱۳	−۳٫۲۱E − ۷
۴	۱٫۲۳۰۰۰۰۰۰۰۱	−۱٫۱۱E − ۱۶	−۸٫۵۴E − ۱۱
۵	۱٫۲۲۹۹۹۹۳۰۴۶	−۹٫۰۴E − ۱۳	۶٫۹۵E − ۷

را می‌توان فقط با تغییر صورت مسأله ریشه‌یابی از راه تحلیلی به صورتی که در آن α ریشهٔ ساده باشد رفع کرد. ساده‌ترین راه انجام آن، تشکیل مشتق مرتبهٔ $(p-1)$ام $f(x)$ و سپس حل

$$f^{(p-1)}(x) = 0 \qquad (۸.۷.۲)$$

است.

مثال مثال قبلی دارای ریشه با چندگانی $p = 2$ است، که یک ریشهٔ ساده برای

$$f'(x) = 3x^2 - 11.12x + 9.1389$$

است. با استفاده از آخرین بارست در جدول ۱۱.۲ به‌عنوان حدس اولیه، و به‌کار بردن روش نیوتن برای پیداکردن ریشهٔ $f'(x)$ که هم‌اکنون داده شد، برای محاسبهٔ α با دقت کامل رایانه، فقط یک بارست لازم است.

۸.۲ الگوریتم ریشه‌یابی برنت

الگوریتمی را توضیح می‌دهیم که هم مزایای روش نیم‌سازی و هم روش خط قاطع را دارد، و در عین حال نقاط ضعفِ هیچ یک از آنها را ندارد. این الگوریتم از (فصل ۴، برنت[1] (۱۹۷۳)) گرفته شده و تکمیل یافتهٔ الگوریتم دیگری است که (دِکر[2](۱۹۶۹)) ارائه کرده است. الگوریتم به بازهٔ کوچکی می‌انجامد که ریشه را در بردارد. چنانچه تابع حول ریشهٔ ξ، به اندازهٔ کافی هموار باشد، آنگاه مرتبهٔ همگرایی، همانند روش خط قاطع، زَبَر خطی است.

در توضیح الگوریتم، نمادهای (برنت (۱۹۷۳، ص ۴۷)) را به‌کار می‌بریم. با دو مقدار a_0 و b_0 به برنامه وارد می‌شویم که برای این دو مقدار، (۱) حداقل یک ریشهٔ ξ از $f(x)$ بین a_0 و b_0

1. Brent 2. Dekker

وجود دارد و $f(a_0)f(b_0) \leq 0$ (۲). در برنامه همچنین یک تحمل مطلوب t داده می‌شود که از آن تحمّل حدّ توقف δ ناشی می‌شود:

$$\delta = t + 2\varepsilon \mid b \mid \qquad (۱.۸.۲)$$

که ε واحد گردکردن رایانه است [(۱ـ۲ـ۱۱) از فصل ۱ را ببینید].

در یک مرحلهٔ نمونهٔ این الگوریتم، b بهترین برآورد جاری ریشهٔ ζ، a مقدار قبلی b، c بارست گذشته است که طوری انتخاب شده است که ریشهٔ ζ بین b و c قرارگیرد (در ابتدا $c = a$). تعریف می‌کنیم $m = \frac{1}{2}(c - b)$.

اگر $f(b) = 0$ (۱)، یا $\mid m \mid \leq \delta$ (۲) الگوریتم را متوقف نمایید. در هر یک از دو حالت، مقدار تقریبی ریشه را $\hat{\zeta} = b$ قرار دهید. برای حالت (۲)، چون b معمولاً از راه روش خط قاطع به‌دست آمده است، ریشهٔ ζ معمولاً به b نزدیکتر از c است. از این رو معمولاً

$$\mid \zeta - \hat{\zeta} \mid \leq \delta$$

گرچه تنها چیزی که می‌تواند تضمین شود رابطهٔ زیر است:

$$\mid \zeta - \hat{\zeta} \mid \leq 2\delta$$

اگر آزمون خطا برقرار نشد، قرار دهید

$$i = b - f(b)\frac{b - a}{f(b) - f(a)} \qquad (۲.۸.۲)$$

سپس قراردهید

$$b'' = \begin{cases} i & \text{اگر } i \text{ بین b و} b + m = \frac{b + c}{2} \text{ واقع باشد} \\ b + m & \text{در غیر این صورت (که روش خط قاطع است)} \end{cases}$$

در حالتی که a، b و c متمایز باشند، روش خط قاطع در تعریف i، با یک روش عکس درون‌یابی درجهٔ دو، جایگزین می‌شود. این عمل به یک همگرایی اندکی سریعتر برای کل الگوریتم می‌انجامد. به دنبال تعریف b''، تعریف می‌کنیم

$$b' = \begin{cases} b'' & \text{اگر } \mid b - b'' \mid > \delta \\ b + \delta.\text{sign}(m) & \text{اگر } \mid b - b'' \mid \leq \delta \end{cases} \qquad (۳.۸.۲)$$

اگر از ریشه فاصله داشته باشید، آنگاه $b' = b''$. با این انتخاب، روش چنین است: (۱) درون‌یابی خطی (یا درجهٔ دو) است، یا: (۲) روش نیم‌سازی است؛ معمولاً برای توابع هموار$f(x)$، حالت (۱) است. این حالت معمولاً به مقداری از m می‌رسد که کوچک نمی‌شود. برای به‌دست آوردن بازهٔ کوچکی که شامل ریشهٔ ζ باشد، وقتی که به آن نزدیک هستیم، از $b' := b + \delta \cdot \text{sign}\ (m)$ استفاده می‌کنیم، که مرحله‌ای از δ در جهت c است. به علّت راهی که برای تعیین مقدار جدید c انتخاب کردیم، اکنون معمولاً بازهٔ کوچک جدیدی حول ζ به‌دست می‌آید. برِنت قبل از انتخاب bی جدید، معمولاً b' که اینجا گفته شد، یک قدم مهم دیگر که تکنیکی است برمی‌دارد.

وقتی که b' به‌دست آمد، b را مساوی b' و a را مساوی مقدار قدیمی b قرار می‌دهیم. اگر علامت $f(b)$ با استفاده از مقدار جدید b، با علامت مقدار قدیمی b یکی باشد، مقدار c عوض نمی‌شود؛ در غیر این صورت، c را برابر مقدار قدیمی b قرار می‌دهیم، که بازهٔ کوچکتری حول ζ به‌دست می‌دهد. دقّت مقدار b، همانگونه که قبلاً توضیح داده شد، اکنون امتحان می‌شود.

برِنت، خیلی مواظب بوده‌است که از مشکلات پی‌ریز و سرریز در روش خود احتراز کند و نتیجتاً تا حدودی خواندن برنامه پیچیده شده است.

مثال هر یک از حالات زیر با یک IBM PC، با همپردازندهٔ حساب ۸۰۸۷ و حساب با دقّت ساده که در استاندارد IEEE برای حساب با ممیز شناور صدق می‌کند، حل شده است. تحمل $t = ۱۰^{-۵}$ بوده و چون در تمام حالات ریشهٔ ζ برابر با ۱ بوده است،

$$\delta = ۱۰^{-۵} + ۲ \mid b \mid \times(۵٫۹۶ \times ۱۰^{-۸}) \doteq ۱٫۰۱ \times ۱۰^{-۵}$$

توابع به همان شکلی محاسبه شده‌اند که در اینجا داده شده‌اند و بازهٔ اولیّه $[a, b] = [۰, ۳]$ بوده است. در جدول مقادیر b و c تا ۷ رقم اعشاری گردشده‌اند.

حالت ۱. $f(x) = (x-۱)[۱+(x-۱)^۲]$. نتایج عددی در جدول ۱۳.۲ داده شده‌اند. لزوم به‌کار بردن $b' = b + \delta(x-۱)^۲ \cdot \text{sign}(m)$ برای به دست آوردن یک بازهٔ کوچک که شامل ξ باشد این مثال را روشن می‌کند.

حالت ۲. $f(x) = x^۲ - ۱$. نتایج عددی در جدول ۱۴.۲ داده شده‌اند.

حالت ۳. $f(x) = -۱ + x(۳ + x(-۳ + x))$. ریشهٔ $\zeta = ۱$ سه‌گانی است، ۵۰ بارست برای همگرایی به ریشهٔ تقریبی ۱٫۰۰۰۰۰۱ به‌کار رفته است. با مقادیر اوّلیهٔ $a = ۰$، $b = ۳$، روش نیم‌سازی فقط ۱۹ بارست برای همین دقّت را می‌طلبد. اگر روش برِنت با روش نیم‌سازی،

جدول ۱۳.۲ مثال ۱ روش برنت

b	$f(b)$	c
۰٫۰	−۲٫۰۰E + ۰	۳٫۰
۰٫۵	−۶٫۲۵E − ۱	۳٫۰
۰٫۷۱۳۹۰۳۸	−۳٫۱۰E − ۱	۳٫۰
۰٫۹۱۵۴۵۰۷	−۸٫۵۲E − ۲	۳٫۰
۰٫۹۹۰۱۷۷۹	−۹٫۸۲E − ۳	۳٫۰
۰٫۹۹۹۸۵۶۷	−۱٫۴۳E − ۴	۳٫۰
۰٫۹۹۹۹۹۹۹	−۱٫۱۹E − ۷	۳٫۰
۰٫۹۹۹۹۹۹۹	−۱٫۱۹E − ۷	۱٫۰۰۰۰۱۰

جدول ۱۴.۲ مثال ۲ روش برنت

b	$f(b)$	c
۰٫۰	−۱٫۰۰	۳٫۰
۰٫۳۳۳۳۳۳۳	−۸٫۸۹E − ۱	۳٫۰
۰٫۳۳۳۳۳۳۳	−۸٫۸۹E − ۱	۱٫۶۶۶۶۶۷
۰٫۷۷۷۷۷۷۸	−۴٫۰۰E − ۱	۱٫۶۶۶۶۶۷
۱٫۰۶۸۶۸۷	۱٫۴۲E − ۱	۰٫۷۷۷۷۷۷۸
۰٫۹۹۱۷۳۳۶	−۱٫۶۵E − ۲	۱٫۰۶۸۶۸۷
۰٫۹۹۹۷۲۴۴	−۵٫۵۱E − ۴	۱٫۰۶۸۶۸۷
۱٫۰۰۰۰۰۰	۲٫۳۸E − ۷	۰٫۹۹۹۷۲۴۴
۱٫۰۰۰۰۰۰	۲٫۳۸E − ۷	۰٫۹۹۹۹۹۰۰

در ردهٔ همهٔ توابع پیوسته مقایسه شود، آنگاه تعداد بارستها برای تحمّل خطای δ تقریباً چنین است:

$$\log_2\left(\frac{b-1}{\delta}\right) \quad \text{برای روش نیم‌سازی}$$

$$\left[\log_2\left(\frac{b-a}{\delta}\right)\right]^2 \quad \text{برای الگوریتم برنت}$$

پس همان‌گونه که مثال ما نشان می‌دهد، حالتهایی وجود دارند که روش نیم‌سازی بهتر است. ولی برای توابع به اندازهٔ کافی هموار با $f'(\alpha) \neq 0$، الگوریتم برنت تقریباً همیشه سریعتر است.

حالت ۴. $f(x) = (x-1)\exp[-1/(x-1)^2]$. ریشهٔ $x = 1$ چندگانی بی‌نهایت دارد، زیرا برای همهٔ مقادیر $r \geq 0$، $f^{(r)}(1) = 0$. نتایج عددی در جدول ۱۵.۲ داده شده‌اند.

جدول ۱۵.۲ مثال ۴ روش برنت

b	$f(b)$	c
۰٫۰	−۳٫۶۸E − ۱	۳٫۰
۰٫۵۷۳۱۷۵۴	−۱٫۷۶E − ۳	۳٫۰
۰٫۵۹۵۹۳۳۱	−۱٫۶۳E − ۳	۳٫۰
۰٫۶۰۹۸۴۴۳	−۵٫۴۷E − ۴	۳٫۰
۰٫۶۱۳۶۳۵۴	−۴٫۷۶E − ۴	۱٫۸۰۴۹۲۲
۰٫۶۳۸۹۲۵۸	−۱٫۶۸E − ۴	۱٫۸۰۴۹۲۲
۱٫۲۲۱۹۲۴	۳٫۳۷E − ۱۰	۰٫۶۳۸۹۲۵۸
۱٫۲۲۱۹۱۴	۳٫۳۷E − ۱۰	۰٫۶۳۸۹۲۵۸
۱٫۲۱۶۵۸۵	۱٫۲۰E − ۱۰	۰٫۶۳۸۹۲۵۸
۰٫۹۲۷۷۵۵۳	۰٫۰	۱٫۲۱۶۵۸۵

توجه کنید که برنامه یک ریشهٔ دقیق برای شکل ماشینی $f(x)$ یافته است. این به خاطر عدم دقت ذاتی در محاسبهٔ تابع $f(x)$ است؛ بخش قبلی دربارهٔ ریشه‌های چندگانه را ببینید. این ریشه البته خیلی نادقیق است، ولی این چیزی نیست که برنامه بتواند آن را اصلاح کند.

برنامهٔ اصلی برنت، که در سال ۱۹۷۳ منتشر شده هنوز عمومیّت دارد و زیاد از آن استفاده می‌شود. معذالک اصلاح و توسعهٔ آن هنوز ادامه دارد. برای یکی از آنها، و برای مرور بقیه، لِ[1] (۱۹۸۳) را ببینید.

۹.۲ ریشه‌های چندجمله‌ییها

ما اکنون حل معادلهٔ چندجمله‌ای

$$p(x) \equiv a_0 + a_1 x + \cdots + a_n x^n = 0 \qquad a_n \neq 0 \qquad (۱.۹.۲)$$

را در نظر می‌گیریم. این مسأله از راههای گوناگونی ظاهر می‌شود و نوشته‌های زیادی برای پرداختن به آن، پدید آمده‌است. گاهی یک ریشهٔ خاص خواسته می‌شود و یک حدس اولیهٔ خوب در دست است. در چنین حالتی، بهترین راه، اصلاح یکی از روشهای بارستی برای بهره‌گیری از شکل خاص چندجمله‌ای است. در حالات دیگر، ممکن است اطلاّع کمی دربارهٔ جای ریشه‌ها داشته باشیم، پس روشهای دیگری باید به‌کار برده شوند، که انواع آنها زیاد است. در این بخش، ما فقط یک گشت مختصری، بدون ادعا برکامل بودن آن، در قسمت ریشه‌یابی چندجمله‌ییها، خواهیم زد. بر اصلاح جزیی روشهای بخشهای

1. Le

گذشته تکیه خواهیم کرد و مسائل مربوط به پایداری عددی را مورد ملاحظه قرار خواهیم داد. با مروری بر بعضی قضایای معیّن محدودهٔ ریشه یا جایابی تقریبی ریشه‌های (۱.۹.۲) بحث را شروع خواهیم کرد.

قضیّه‌های جایابی. چون $p(x)$ یک چندجمله‌ای است، قضیّه‌های زیادی دربارهٔ ریشه‌های آن می‌توان بیان کرد، که دربارهٔ توابع دیگر درست نخواهند بود. معروفترین آنها، **قضیهٔ اساسی جبر** است، که به ما اجازه می‌دهد $p(x)$ را به‌صورت حاصلضربی یکتا (جز از لحاظ ترتیب) که شامل ریشه‌ها باشد بنویسیم،

$$p(x) = a_n(x - z_1)\dots(x - z_n) \qquad (۲.۹.۲)$$

z_1، ...، z_n ریشه‌های $p(x)$ اند که به تعداد درجهٔ چندگانی خود تکرار می‌شوند. اکنون چند قضیهٔ معروف جایابی تقریبی و محدود کردن این ریشه‌ها را ذکر می‌کنیم.

قاعدهٔ علائم دکارت. که به‌کار برده می‌شود تا ریشه‌های حقیقی و مثبت $p(x)$ را محدود نماید، به شرطی که ضرایب a_0، ...، a_n تماماً حقیقی باشند.

گیریم ν تعداد تغییر علامتها در ضرایب $p(x)$ در (۱.۹.۲) باشد وقتی جملات صفرنادیده گرفته شوند. گیریم k تعداد ریشه‌های حقیقی مثبت $p(x)$ باشد که چندگانی آنها به‌حساب آمده‌اند. در این صورت $k \le \nu$ و $\nu - k$ زوج است.

یک برهان این قضیه در هنریچی [1] (۱۹۷۴، ص ۴۴۲) و هاوسهولدِر [2] (۱۹۷۰، ص ۸۲) داده شده‌است.

مثال عبارت $p(x) = x^6 - x - 1$ دارای $\nu = 1$ تغییر علامت در ضرایب است. بنابراین $k = 1$؛ زیرا در غیر این صورت $k = 0$ و $\nu - k = 1$ یک عدد صحیح زوج نیست که یک تناقض است.

قاعدهٔ تغییر علامات دکارت برای محدود کردن تعداد ریشه‌های منفی $p(x)$ نیز به‌کار برده می‌شود. آن را برای چندجمله‌ای

$$q(x) = p(-x)$$

به‌کار می‌بریم. ریشه‌های مثبت این چندجمله‌ای، همان ریشه‌های منفی $p(x)$ هستند. این قاعده را برای مثال اخیر در $q(x) = x^6 + x - 1$ به‌کار می‌بریم. باز هم یک ریشهٔ حقیقی مثبت برای $q(x)$ وجود دارد، پس $p(x)$ دارای یک ریشهٔ حقیقی منفی است.

1. Henrici 2. Householder

یک کران بالا برای تمام ریشه‌های $p(x)$ با رابطهٔ زیر داده می‌شود

$$| z_i | \leq R \equiv ۱ + \underset{۰ \leq i \leq n-۱}{\text{Max}} \left| \frac{a_i}{a_n} \right| \qquad (۳.۹.۲)$$

این قضیّه از اگوستن کوشی[1] (۱۸۲۹) است و اثبات آن در هاوسهولدر (۱۹۷۰، ص ۷۱) داده شده است. قضیّهٔ مشابه دیگری از کوشی هست که بر مبنای درنظرگرفتن چندجمله‌ییهای زیر قرار دارد

$$| a_n | x^n + | a_{n-۱} | x^{n-۱} + \cdots + | a_۱ | x - | a_۰ | = ۰ \qquad (۴.۹.۲)$$

$$| a_n | x^n - | a_{n-۱} | x^{n-۱} - \cdots - | a_۱ | x - | a_۰ | = ۰ \qquad (۵.۹.۲)$$

فرض می‌کنیم $a_۰ \neq ۰$، که هم‌ارز است با آنکه فرض کنیم $x = ۰$ یک ریشهٔ $p(x)$ نیست. در این صورت با استفاده از قاعدهٔ علائم دکارت، هر یک از چندجمله‌ییهای فوق، دقیقاً یک ریشهٔ مثبت دارد؛ آنها را به ترتیب $\rho_۱$ و $\rho_۲$ می‌نامیم. پس تمام ریشه‌های z_j از $p(x)$ در رابطهٔ زیر صدق می‌کنند.

$$\rho_۱ \leq | z_j | \leq \rho_۲ \qquad (۶.۹.۲)$$

اثبات کران بالا در هنریچی (۱۹۷۴، ص ۴۵۸) و هاوسهولدر (ص ۷۰، ۱۹۷۰) داده شده است. اثبات کران پایین را می‌توان براساس روش زیر، که برای ساختن کران پایین (۳.۹.۲) نیز به‌کار می‌رود، بنانهاد. چند جمله‌ای

$$q(x) = x^n p\left(\frac{۱}{x}\right) = a_n + a_{n-۱}x + \cdots + a_۱ x^{n-۱} + a_۰ x^n \qquad a_۰ \neq ۰ \quad (۷.۹.۲)$$

را در نظر می‌گیریم. پس ریشه‌های $q(x)$ برابر $۱/z$ هستند، که z یک ریشهٔ $p(x)$ است. اگر قضیّهٔ کران بالای (۶.۹.۲) برای (۷.۹.۲) به‌کار برده شود کران پایین (۶.۹.۲) به‌دست می‌آید. این کاربرد را در ضمن یک مسأله نشان خواهیم داد.

چون هر یک از چندجمله‌ییهای (۴.۹.۲) و (۵.۹.۲) یک ریشهٔ یگانهٔ سادهٔ مثبت دارد، روش نیوتن را می‌توان به‌سادگی برای ترسیم $R_۱$ و $R_۲$ به‌کار برد. به‌عنوان یک حدس اولیه، از کران بالای (۳.۹.۲) استفاده می‌کنیم یا با حدسهای اولیهٔ مثبت کوچکتری تجربه می‌کنیم. مثالهای این قضیّه‌ها را در مسائل مطرح می‌کنیم.

قضیّه‌های بسیار دیگری از این نوع وجود دارند، که هم هنریچی (۱۹۷۴، فصل ۶) و هم هاوسهولدر (۱۹۷۰) مقالات عالی در این موضوع نوشته‌اند.

---

1. Augustin Cauchy

ضربِ تودرتو. یک راه بسیار کارآمد برای محاسبهٔ چندجمله‌ای $p(x)$، که در (۱.۹.۲) داده شده، استفاده از ضرب تودرتو است:

$$p(x) = a_0 + x(a_1 + x(a_2 + \cdots + x(a_{n-1} + a_n x)\ldots)) \qquad (۸.۹.۲)$$

در فرمول (۱.۹.۲)، n جمع و $2n-1$ ضرب و در فرمول (۲ـ۹ـ۸)، n جمع و n ضرب وجود دارد که یک صرفه‌جویی قابل ملاحظه‌ای است.

برای کارهای بعدی، مناسبتر است که ضرایب کمکی زیر را معرفی کنیم؛ گیریم $b_n = a_n$

$$b_k = a_k + z b_{k+1}, \qquad k = n-1, n-2, \ldots, 0 \qquad (۹.۹.۲)$$

با ملاحظهٔ (۸.۹.۲) به‌سادگی ملاحظه می‌شود که

$$p(z) = b_0 \qquad (۱۰.۹.۲)$$

چندجمله‌ای زیر را معرفی می‌کنیم

$$q(x) = b_1 + b_2 x + \cdots + b_n x^{n-1} \qquad (۱۱.۹.۲)$$

در این صورت

$$\begin{aligned} b_0 + (x-z)q(x) =& b_0 + (x-z)[b_1 + b_2 x + \cdots + b_n x^{n-1}] \\ =& (b_0 - b_1 z) + (b_1 - b_2 z)x + \ldots \\ &+ (b_{n-1} - b_n z)x^{n-1} + b_n x^n \\ =& a_0 + a_1 x + \cdots + a_n x^n = p(x) \\ p(x) =& b_0 + (x-z)q(x) \end{aligned} \qquad (۱۲.۹.۲)$$

که $q(x)$ خارج‌قسمت و b_0 باقیماندهٔ تقسیم $p(x)$ بر $(x-z)$ است. استفاده از (۹.۹.۲) برای محاسبهٔ $p(z)$ و تشکیل چندجمله‌ای خارج‌قسمت $q(x)$ را **روش هورنر**[1] نیز می‌خوانند.

اگر z یک ریشهٔ $p(x)$ باشد، آنگاه $b_0 = 0$ و $p(x) = (x-z)q(x)$. برای پیدا کردن ریشه‌های دیگر $p(x)$، می‌توانیم جستجوی خود را به ریشه‌های $q(x)$ محدود کنیم. این فرآیند کاهش را **تقلیل** می‌نامند و آن را باید با احتیاط به کار برد، نکته‌ای که بعداً به آن باز خواهیم گشت.

1. Horner

روش نیوتن. اگر بخواهیم برای یافتن یک ریشهٔ $p(x)$، روش نیوتن رابه‌کار بریم، باید بتوانیم هم $p(x)$ و هم $p'(x)$ را در هر نقطهٔ z محاسبه کنیم. با توجّه به (۱۲.۹.۲) داریم:

$$p'(x) = (x - z)q'(x) + q(x)$$

$$p'(z) = q(z) \qquad (۱۳.۹.۲)$$

از (۱۰.۹.۲) و (۱۳.۹.۲) در توجیه روش نیوتن برای ریشه‌یابی چندجمله‌ییها که ذیلاً آمده است استفاده می‌کنیم.

الگوریتم $Polynew\ (a, n, x_\circ, \varepsilon, itMaxroot, b, ier)$

۱. توجه: a بردار ضرایب است، itmax بیشترین تعداد بارستهایی است که باید محاسبه شود، b بردار ضرایب چندجمله‌ای تقلیل‌یافته است، و ier یک شاخص برای‌خطاست.

۲. itnum:= ۱

۳. $b_n := c = a_n$ ، $z := x_\circ$.

۴. برای $k := n - ۱, \ldots, ۱$، $b_k := a_k + zb_{k+۱}$ و $c := b_k + zc$

۵. $b_\circ := a_\circ + zb_۱$

۶. اگر $c = \circ$، آنگاه $ier = ۲$ قرار داده خارج شوید.

۷. $x_۱ = x_\circ - b_\circ/c$

۸. اگر $|x_۱ - x_\circ| \leq \varepsilon$، آنگاه $ier := \circ$ و $root := x_۱$ قرار داده خارج شوید.

۹. اگر itnum=itmax، آنگاه $ier := ۱$ قرار داده و خارج شوید.

۱۰. در غیر این صورت، $itnum := itnum + ۱$ و $x_\circ := x_۱$، و به مرحلهٔ ۳ بروید.

مسائل پایداری. چندجمله‌ییهای بسیاری هستند که ریشه‌های آنها نسبت به‌تغییرات کوچک ضرایب حساسیت زیادی دارند. در بعضی از این مسائل که دارای ریشه‌های چندگانه‌اند، تعجب‌آور نیست که این ریشه‌ها نسبت به تغییرات کوچک ضرایب حساسیت زیادی نشان دهند. ولی چندجمله‌ییهای زیادی هستند که فقط ریشه‌های سادهٔ کاملاً از هم جدا دارند و در آنها باز هم ریشه‌ها نسبت به اختلالات کوچک بسیار حساس‌اند. فرمولهایی ذیلاً داده شده‌اند که این حساسیت را توضیح می‌دهند. و مثالهای عددی نیز آورده شده‌اند.

برای این بحث نظری، توابع زیر را معرفی می‌کنیم

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n \qquad a_n \neq 0$$
$$q(x) = b_0 + b_1 x + \cdots + b_n x^n \qquad (۱۴.۹.۲)$$

و اختلال $p(x)$ را با

$$p(x;\varepsilon) = p(x) + \varepsilon q(x) \qquad (۱۵.۹.۲)$$

تعریف می‌نماییم. ریشه‌های $p(x;\varepsilon)$ را با $z_1(\varepsilon)$، . . .، $z_n(\varepsilon)$ نشان می‌دهیم که برحسب چندگانگی تکرار می‌شوند، گیریم $z_j = z_j(0)$، $j = 1, \ldots, n$، ریشهٔ متناظر $p(x) = p(x, 0)$ باشند. روشن است که صفرهای یک چندجمله‌ای توابع پیوسته‌ای از ضرایب چندجمله‌ای هستند [برای مثال هنریچی (۱۹۷۴، ص ۲۸۱) را ببینید]. در نتیجه، $z_j(\varepsilon)$ تابعی پیوسته از ε است. آنچه ما می‌خواهیم معلوم کنیم این است که وقتی ε نزدیک صفر است، ریشهٔ $z_j(\varepsilon)$ با چه تندی ε تغییر می‌کند.

مثال

$$p(x;\varepsilon) = (x-1)^3 - \varepsilon \qquad p(x) = (x-1)^3 \qquad \varepsilon > 0$$

پس ریشه‌های $p(x)$ عبارت‌اند از $z_1 = z_2 = z_3 = 1$. ریشه‌های $p(x;\varepsilon)$ عبارت‌اند از:

$$z_1(\varepsilon) = 1 + \sqrt[3]{\varepsilon} \qquad z_2(\varepsilon) = 1 + w\sqrt[3]{\varepsilon} \qquad z_3(\varepsilon) = 1 + w^2\sqrt[3]{\varepsilon}$$

با $w = \frac{1}{2}(-1 + i\sqrt{3})$. برای هر سه ریشهٔ $p(x;\varepsilon)$:

$$| z_j(\varepsilon) - 1 | = \sqrt[3]{\varepsilon}$$

برای روشن شدن مطلب، گیریم $\varepsilon = 0{,}001$. در این صورت

$$p(x;\varepsilon) = x^3 - 3x^2 + 3x - 1{,}001$$

که تغییر نسبتاً کوچکی در $p(x)$ است. ولی برای ریشه‌ها

$$| z_j(\varepsilon) - 1 | = 0{,}1$$

که تغییر نسبتاً بزرگی در ریشه‌های $z_j = ۱$ است.
ما اکنون برآوردهای کلّیتری برای $z_j(\varepsilon) - z_j$ می‌دهیم.

حالت ۱. z_j یک ریشهٔ سادهٔ $p(x)$ است، پس $p'(z_j) \neq ۰$. با استفاده از نظریهٔ توابع با متغیر مختلط، می‌دانیم که $z_j(\varepsilon)$ را می‌توان به‌صورت یک سری توانی نوشت؛

$$z_j(\varepsilon) = z_j + \sum_{l=۱}^{\infty} \gamma_l \varepsilon^l \qquad (۱۶.۹.۲)$$

برای برآورد $z_j(\varepsilon) - z_j$، یک فرمول برای اولین جملهٔ $\gamma_۱\varepsilon$ سری به‌دست می‌آوریم. برای شروع، به راحتی می‌توان دید که

$$\gamma_۱ = z_j'(۰)$$

برای محاسبهٔ $z_j'(\varepsilon)$، از اتحاد زیر، که برای کلّیهٔ مقادیر به اندازهٔ کافی کوچک ε برقرار است، مشتق می‌گیریم

$$p(z_j(\varepsilon)) + \varepsilon q(z_j(\varepsilon)) = ۰$$

به‌دست می‌آوریم

$$p'(z_j(\varepsilon))z_j'(\varepsilon) + q(z_j(\varepsilon)) + \varepsilon q'(z_j(\varepsilon))z_j'(\varepsilon) = ۰$$

$$z_j'(\varepsilon) = \frac{-q(z_j(\varepsilon))}{p'(z_j(\varepsilon)) + \varepsilon q'(z_j(\varepsilon))} \qquad (۱۷.۹.۲)$$

با قراردادن $\varepsilon = ۰$، به‌دست می‌آوریم

$$\gamma_۱ = z_j'(۰) = -\frac{q(z_j)}{p'(z_j)}$$

به رابطهٔ (۱۶.۹.۲) برمی‌گردیم

$$z_j(\varepsilon) = z_j - \frac{q(z_j)}{p'(z_j)}\varepsilon + \sum_{l=۲}^{\infty} \gamma_l \varepsilon^l$$

$$\left| z_j(\varepsilon) - \left[z_j - \frac{q(z_j)}{p'(z_j)}\varepsilon \right] \right| \leq K\varepsilon^۲ \qquad |\varepsilon| \leq \varepsilon_۰ \qquad (۱۸.۹.۲)$$

برای بعضی ثابتهای $K > ۰$ و $\varepsilon_۰ > ۰$. در برآورد $z_j(\varepsilon)$ برای مقدار کوچک ε، از رابطهٔ زیر استفاده می‌کنیم

$$z_j(\varepsilon) \doteq z_j - \frac{q(z_j)}{p'(z_j)}\varepsilon \qquad (۱۹.۹.۲)$$

ضریب ε نشان می‌دهد که تغییر $z_j(\varepsilon)$ نسبت به ε چه‌اندازه سریع است؛ اگر این نسبت زیاد باشد ریشهٔ z_j را بدوضع گویند.

حالت ۲. m، چندگانگی z_j از ۱ بزرگتر است. با به‌کاربردن تکنیکهای مربوط به حالت (۱)، برای یک مقدار $\varepsilon_۰ > ۰$ و $K > ۰$ می‌توانیم به‌دست آوریم

$$| z_j(\varepsilon) - [z_j + \gamma_۱\varepsilon^{۱/m}] | \leq K | \varepsilon |^{۲/m} \qquad | \varepsilon | \leq \varepsilon_۰ \qquad (۲۰.۹.۲)$$

m مقدار ممکن برای $\gamma_۱$ وجود دارد که m ریشهٔ مختلط

$$\gamma_۱^m = \frac{-m!q(z_j)}{p^{(m)}(z_j)}$$

هستند.

مثال چندجمله‌ای سادهٔ زیر را در نظر می‌گیریم

$$\begin{aligned} p(x) =& (x-۱)(x-۲)\dots(x-۷) \\ =& x^۷ - ۲۸x^۶ + ۳۲۲x^۵ - ۱۹۶۰x^۴ + ۶۷۶۹x^۳ - ۱۳۱۳۲x^۲ \\ &+ ۱۳۰۶۸x - ۵۰۴۰ \end{aligned} \qquad (۲۱.۹.۲)$$

برای اختلال، می‌گیریم

$$q(x) = x^۶ \qquad \varepsilon = -۰٫۰۰۲$$

پس برای ریشهٔ $z_j = j$،

$$p'(z_j) = \prod_{l \neq j}(j-l) \qquad q(z_j) = j^۶$$

از (۱۹.۹.۲)، برآورد زیر را داریم

$$z_j(\varepsilon) \doteq j + \frac{۰٫۰۰۲j^۶(-۱)^{j-۱}}{(j-۱)!(۷-j)!} = j + \delta(j) \qquad (۲۲.۹.۲)$$

جدول ۱۶.۲ مقادیر $\delta(j)$ از (۲۲.۹.۲)

j	$\delta(j)$
۱	۲٫۷۸E − ۶
۲	−۱٫۰۷E − ۳
۳	۳٫۰۴E − ۲
۴	−۲٫۲۸E − ۱
۵	۶٫۵۱E − ۱
۶	−۷٫۷۷E − ۱
۷	۳٫۲۷E − ۱

جدول ۱۷.۲ ریشه‌های $p(x;\varepsilon)$ برای (۲ـ۹ـ۲۱)

j	$z_j(\varepsilon)$	$z_j(\varepsilon) - z_j(۰)$
۱	۱٫۰۰۰۰۰۲۸	۲٫۸۰E − ۶
۲	۱٫۹۹۸۹۳۸۲	−۱٫۰۶E − ۳
۳	۳٫۰۳۳۱۲۵۳	۳٫۳۱E − ۲
۴	۳٫۸۱۹۵۶۹۲	−۱٫۸۰E − ۱
۵	۵٫۴۵۸۶۷۵۸ + ۰٫۵۴۰۱۲۵۷۸i	
۶	۵٫۴۵۸۶۷۵۸ − ۰٫۵۴۰۱۲۵۷۸i	
۷	۷٫۲۳۳۰۱۲۸	۲٫۳۳E − ۱

مقادیر عددی $\delta(j)$ در جدول ۱۶.۲، داده شده‌اند. خطای نسبی در ضریب $x^۶$ برابر است با $۷٫۱E-۵ = \frac{۰٫۰۰۲}{۲۸}$، ولی خطاهای نسبی در ریشه‌ها بسیار بزرگتر است. در واقع، اندازهٔ بعضی اختلالات $\delta(j)$ ما را به شک می‌اندازد که آیا برآورد خطّی (۲۲.۹.۲) معتبر است؟ ریشه‌های واقعی $p(x) + \varepsilon q(x)$ در جدول ۱۷.۲ داده شده‌اند، که به اختلالات پیش‌بینی شده خیلی نزدیک‌اند. بیشترین انحراف در ریشه‌ها، به ازای $j = ۵$ و $j = ۶$ رخ داده است. این ریشه‌ها مختلط‌اند که با برآورد خطی (۲۲.۹.۲) پیش‌بینی نشده بود. در این دو حالت، ε در خارج شعاع همگرایی سری توانی (۱۶.۹.۲) واقع است، زیرا این سری فقط دارای ضرایب حقیقی

$$\gamma_l = \frac{۱}{l!} z_j^{(l)}(۰)$$

است که از مشتق‌گیری (۱۷.۹.۲) به‌دست می‌آیند.

یک چندجمله‌ای راکه ریشه‌های آن نسبت به تغییرات نسبی کوچک ضرایب ناپایدارند، بد وضع گوئیم. خیلی از این چندجمله‌ییها به‌طور طبیعی در کاربردها، پدید می‌آیند. مثال قبلی، دشواری تعیین بدوضع بودن یا نبودن یک چندجمله‌ای را فقط با یک بررسی سطحی نشان می‌دهد.

تقلیل چندجمله‌ای. مسألهٔ دیگری که رخ می‌دهد تقلیل یک چندجمله‌ای به یک چندجمله‌ای از درجهٔ کمتر است، فرایندی که به دنبال (۱۲.۹.۲) تعریف شده است. چون صفرها، دقیقاً پیدا نمی‌شوند، چندجمله‌ای از درجهٔ پایینتر (۱۱.۹.۲) که با خارج‌کردن آخرین ریشه به‌دست آمده، معمولاً در تمام ضرایب خود دارای خطاست. از مثال اخیرروشن می‌شود، که این امر می‌تواندیک اختلال قابل توجهی در ریشه‌های رده‌هایی از چندجمله‌ییها ایجاد نماید. ویلکینسن[1] (۱۹۶۳) اثرات این تقلیل را تحلیل نموده و استراتژی کلّی زیر را توصیه نموده است: (۱) مسأله را با پیدا کردن ریشه‌های با اندازهٔ کوچکتر شروع، و به پیدا کردن ریشه‌های بزرگتر ختم کنید؛ (۲) پس از اینکه تقریب تمام ریشه‌ها را یافتید، با استفاده از مقادیر یافته شده به‌عنوان حدسهای اولیه، و به‌کاربردن چندجمله‌ای اصلی، حل را دوباره تکرار نمایید. بحث کامل را می‌توان در ویلکینسن (۱۹۶۳، ۵۵ـ۶۵) پیدا کرد.

مثال یافتن ریشه‌های چندجمله‌ای از درجهٔ ۶ لاگر را در نظر می‌گیریم.

$$p(x) = ۷۲۰ - ۴۳۲۰x + ۵۴۰۰x^۲ - ۲۴۰۰x^۳ + ۴۵۰x^۴ - ۳۶x^۵ + x^۶$$

الگوریتم نیوتن بخش اخیر برای یافتن ریشه‌ها به‌کارگرفته شده و با پذیرفتن هر ریشهٔ‌جدید عمل تقلیل انجام شده است. ریشه‌ها به دو طریق محاسبه شده‌اند: (۱) از بزرگتر به‌کوچکتر و (۲) از کوچکتر به بزرگتر. محاسبات با حساب با دقت ساده با رایانهٔ IBM-360 انجام پذیرفته و نتایج عددی در جدول ۱۸.۲ داده شده‌اند. مقایسهٔ ستونهایی که با روش‌های (۱) و (۲) مشخص شده‌اند، بروشنی برتری روشی را که ریشه‌ها را درجهت صعودی محاسبه می‌کند نشان می‌دهد. اگر نتایج روش (۱) به عنوان حدسهای اولیه برای بارستِ بعدی با چندجمله‌ای اصلی به‌کاربرده شوند ریشه‌های تقریبی که پیدا می‌شوند دقیقتر از جوابهای روش (۲) هستند؛ ستون جدول راکه با روش (۳) در بالای ستون مشخص شده، ببینید. این جدول مجدداً اهمیّت بارستن با چندجمله‌ای اصلی را برای از بین بردن اثرات فرایند تقلیل نشان می‌دهد.

روشهای دیگری برای تقلیل یک چندجمله‌ای وجود دارند که یکی از آنها پیدا کردن ریشهٔ بزرگتر در ابتدای امر است. برای یک بحث کامل پیترز[2] و ویلکینسن (۱۹۷۱، بخش ۵) را ببینید. یک الگوریتم برای تقلیل مرکب داده شده‌است که احتیاج به پیدا کردن ریشه‌ها به ترتیبی خاص را

1. Wilkinson 2. Peters

جدول ۱۸.۲ مثال از تقلیل کثیرالجمله

واقعی	روش (۱)	روش (۲)	روش (۳)
۱۵٫۹۸۲۸۷	۱۵٫۹۸۲۸۷	۱۵٫۹۸۲۷۹	۱۵٫۹۸۲۸۷
۹٫۸۳۷۴۶۷	۹٫۸۳۷۴۷۱	۹٫۸۳۷۴۶۹	۹٫۸۳۷۴۶۷
۵٫۷۷۵۱۴۴	۵٫۷۷۵۷۶۴	۵٫۷۷۵۲۰۷	۵٫۷۷۵۱۴۴
۲٫۹۹۲۷۳۶	۲٫۹۹۱۰۸۰	۲٫۹۹۲۷۱۰	۲٫۹۹۲۷۳۶
۱٫۱۸۸۹۳۲	۱٫۱۹۰۹۳۷	۱٫۱۸۸۹۳۲	۱٫۱۸۸۹۳۲
۰٫۲۲۲۸۴۶۶	۰٫۲۲۱۹۴۲۹	۰٫۲۲۲۸۴۶۶	۰٫۲۲۲۸۴۶۶

منتفی می‌سازد. در این مقاله، مولفان، قوانین به‌کاربردن تقلیل ضمنی

$$q(x) = \frac{p(x)}{(x - z_1) \dots (x - z_r)}$$

را برای حذف ریشه‌های z_1، ...، z_r که قبلاً محاسبه شده‌اند، نیز مورد بحث قرار می‌دهند. ما این بحث را قبلاً، در (۴.۴.۲)، در رابطه با روش مولر آورده‌ایم.

روشهای کلی ریشه‌یابی چندجمله‌ییها. الگوریتم‌های ریشه‌یابی زیادی به‌ویژه برای چندجمله‌ییها طراحی شده‌اند. تعداد زیادی از آنها مشروحاً درکتابهای دژون[1] و هنریچی (۱۹۶۹) و هنریچی (۱۹۷۴، فصل ۶) و هاوسهولدر (۱۹۷۰) آمده است. انواع این روشها آنقدر زیادند که نمی‌توانیم در اینجا به شرح آنها بپردازیم.

در یک ردهٔ وسیع از این روشهای مهم قضایای جایابی را که در (۳.۹.۲) ـ (۶.۹.۲) شرح داده شده‌اند به‌کار می‌برند، تا به‌طور بارستی، ریشه‌ها را به ناحیه‌های مجزا و همیشه کوچکتر، که اغلب دایره‌اند، جدا کنند. معروفترین این روشها، احتمالاً روش لمر- شور [2] است [هاوسهولدر (۱۹۷۰، بخش ۲ـ۷) را ببینید]. این روشها بطور خطی همگرا هستند، و به همین دلیل، اغلب آنها را با روشهای همگرای سریعتر، مانند روش نیوتن، ترکیب می‌کنند. وقتی که ریشه‌ها به ناحیه‌های مجزّا، از هم جدا شدند، روشهای سریعتری برای پیدا کردن ریشه در هر ناحیه به‌کار گرفته می‌شود. برای بحث کلی در مورد این روشهای ریشه‌یابی به هنریچی (۱۹۷۴، بخش ۶ـ۱۰) مراجعه کنید.

روشهای دیگری که صورت الگوریتم‌های گسترده ـ کاربردی پیدا کرده‌اند روش جنکینز و تراوب[3] و روش لاگر است. برای اولی هاوسهولدر (۱۹۷۰، ص ۱۷۳)، جنکینز و تراوب (۱۹۷۰)، (۱۹۷۲) را ببینید. برای روش لاگر، به هاوسهولدر (۱۹۷۰، بخش ۴ـ۵) و کاهان [4] (۱۹۶۷) مراجعه کنید.

1. Dejon 2. Lehmer-Schur 3. Jenkins and Traub 4. Kahan

یک روش عددی، آسان درکاربرد، براین پایه بنا شده است که بتوان ویژه مقدارهای ماتریس را محاسبه کرد. برای یک چندجمله‌ای داده شدهٔ $p(x)$، ممکن است بتوان به‌سادگی ماتریسی ساخت که $p(x)$ چندجمله‌ای مشخصّهٔ آن باشد (مسألهٔ ۲ از فصل ۹ را ببینید). چون نرم‌افزار عالی برای حل مسألهٔ ویژه مقدار وجود دارد، این نرم‌افزار را می‌توان برای پیداکردن ریشه‌های چندجمله‌ای $p(x)$ به‌کار برد.

۱۰.۲ دستگاه معادلات غیرخطی

در این بخش و بخش بعد سروکار ما با حل عددی دستگاه معادلات غیرخطی چند متغیره است. این مسائل کاربردهای گسترده‌ای دارند و به شکلهای گوناگونی هستند. روشهای متنوّعی برای حل چنین دستگاهها وجود دارد، لذا ما فقط خود موضوع را معرفی می‌کنیم و نظریه‌های کلّی و بعضی روشهای عددی را که به‌سادگی قابل برنامه‌نویسی هستند ارائه خواهیم داد. برای بسط کامل تحلیل عددی حل دستگاههای غیرخطی، به چند قضیّه از جبر خطی عددی نیاز داریم که تا فصل ۷ـ۹ آورده نشده‌اند.

برای سادگی بیان و بهتر فهمیدن، این نظریه فقط برای دو معادله ارائه شده است:

$$f_1(x_1, x_2) = 0 \qquad f_2(x_1, x_2) = 0 \qquad (۱.۱۰.۲)$$

وقتی که مفاهیم اساسی روشن شد، تعمیم مطلب برای n معادلهٔ n مجهولی ساده خواهد بود. برای کمک بیشتر صورت و جواب (۱.۱۰.۲) را با نماد برداری در نظر می‌گیریم:

$$\mathbf{f}(\mathbf{x}) = 0 \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix} \qquad (۲.۱۰.۲)$$

به حل (۱.۱۰.۲) می‌توان مانند یک فرایند دو مرحله‌ای نگاه کرد: (۱) خمهای صفر رویه‌های $z = f_1(x_1, x_2)$ و $z = f_2(x_1, x_2)$ را در صفحهٔ x_1x_2 پیدا می‌کنیم، و (۲) نقاط برخورد این خمهای صفر را در صفحه x_1x_2 به‌دست می‌آوریم. این دیدگاه دربخش بعد برای تعمیم روش نیوتن برای حل (۱.۱۰.۲) به‌کار رفته است.

نظریهٔ نقطهٔ ثابت. با تعمیم قسمتی از نظریهٔ بارستیِ نقطهٔ ثابت، بخش ۲ـ۵، مطلب را آغاز می‌کنیم. فرض می‌کنیم مسألهٔ ریشه‌یابی (۱.۱۰.۲) به شکل هم‌ارز آن به‌صورت زیر دوباره نوشته شده است

$$x_1 = g_1(x_1, x_2) \qquad x_2 = g_2(x_1, x_2) \qquad (۳.۱۰.۲)$$

جواب آن را با نماد

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

نشان می‌دهیم. بارستیِ نقطهٔ ثابت

$$x_{1,n+1} = g_1(x_{1,n}, x_{2,n}) \qquad x_{2,n+1} = g_2(x_{1,n}, x_{2,n}) \qquad (۴.۱۰.۲)$$

را مطالعه می‌کنیم. با استفاده از نماد برداری می‌نویسیم

$$\mathbf{x}_{n+1} = \mathbf{g}(\mathbf{x}_n) \qquad (۵.۱۰.۲)$$

که در آن

$$\mathbf{x}_n = \begin{bmatrix} x_{1,n} \\ x_{2,n} \end{bmatrix} \qquad \mathbf{g}(\mathbf{x}) = \begin{bmatrix} g_1(x_1, x_2) \\ g_2(x_1, x_2) \end{bmatrix}$$

مثال حل دستگاه

$$f_1 \equiv 3x_1^2 + 4x_2^2 - 1 = 0 \qquad f_2 \equiv x_2^2 - 8x_1^3 - 1 = 0 \qquad (۶.۱۰.۲)$$

برای جواب $\boldsymbol{\alpha}$ نزدیک به $(x_1, x_2) = (-0.5, 0.25)$ را در نظر می‌گیریم. این دستگاه را با بارست زیر حل می‌کنیم

$$\begin{bmatrix} x_{1,n+1} \\ x_{2,n+1} \end{bmatrix} = \begin{bmatrix} x_{1,n} \\ x_{2,n} \end{bmatrix} - \begin{bmatrix} 0.016 & -0.17 \\ 0.52 & -0.26 \end{bmatrix} \begin{bmatrix} 3x_{1,n}^2 + 4x_{2,n}^2 - 1 \\ x_{2,n}^2 - 8x_{1,n}^3 - 1 \end{bmatrix} \qquad (۷.۱۰.۲)$$

منشأ این صورت جدید (۶.۱۰.۲) بعداً داده خواهد شد. نتایج عددی (۷.۱۰.۲) در جدول ۱۹.۲ داده شده است. واضح است که این بارستها به سرعت همگرا هستند.

برای تحلیل همگرایی (۵.۱۰.۲)، دو معادلهٔ (۴.۱۰.۲) را از معادلات متناظر

$$\alpha_1 = g_1(\alpha_1, \alpha_2) \qquad \alpha_2 = g_2(\alpha_1, \alpha_2)$$

جدول ۱۹.۲ مثال (۷.۱۰.۲) از بارست نقطهٔ ـ ثابت

n	$x_{1,n}$	$x_{2,n}$	$f_1(x_{1,n}, x_{2,n})$	$f_2(x_{1,n}x_{2,n})$
۰	−۰٫۵	۰٫۲۵	۰٫۰	۱٫۵۶E − ۲
۱	−۰٫۴۹۷۳۴۳۷۵۰	۰٫۲۵۴۰۶۲۵۰۰	۲٫۴۳E − ۴	۵٫۴۶E − ۴
۲	−۰٫۴۹۷۲۵۴۷۹۴	۰٫۲۵۴۰۷۷۹۲۲	۹٫۳۵E − ۶	۲٫۱۲E − ۵
۳	−۰٫۴۹۷۲۵۱۳۴۳	۰٫۲۵۴۰۷۸۵۶۶	۳٫۶۴E − ۷	۸٫۲۶E − ۷
۴	−۰٫۴۹۷۲۵۱۲۰۸	۰٫۲۵۴۰۷۸۵۹۲	۱٫۵۰E − ۸	۳٫۳۰E − ۸

که شامل جوابهای درست α هستند، کم می‌کنیم. قضیهٔ مقدار میانی برای توابع دو متغیره (قضیّهٔ ۱ـ۵ با $n = 1$) را برای این تفاضلات به‌کار می‌بریم تا به‌دست‌آوریم

$$\alpha_i - x_{i,n+1} = \frac{\partial g_i(\xi_{1,n}^{(i)}, \xi_{2,n}^{(i)})}{\partial x_1}(\alpha_1 - x_{1,n}) + \frac{\partial g_i(\xi_{1,n}^{(i)}, \xi_{2,n}^{(i)})}{\partial x_2}(\alpha_2 - x_{2,n})$$

برای $i = 1, 2$. نقاط $\boldsymbol{\xi}_n^{(i)} = (\xi_{1,n}^{(i)}, \xi_{2,n}^{(i)})$ روی قطعه خطی است که $\boldsymbol{\alpha}$ را به $\mathbf{x}_n$ متصّل می‌سازد.

در شکل ماتریسی، این معادلات خطا چنین خواهند شد

$$\begin{bmatrix} \alpha_1 - x_{1,n+1} \\ \alpha_2 - x_{2,n+1} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial g_1(\xi_n^{(1)})}{\partial x_1} & \dfrac{\partial g_1(\xi_n^{(1)})}{\partial x_2} \\ \dfrac{\partial g_2(\xi_n^{(2)})}{\partial x_1} & \dfrac{\partial g_2(\xi_n^{(2)})}{\partial x_2} \end{bmatrix} \begin{bmatrix} \alpha_1 - x_{1,n} \\ \alpha_2 - x_{2,n} \end{bmatrix} \qquad (۸.۱۰.۲)$$

گیریم G_n معرّفِ ماتریس (۸.۱۰.۲) باشد. در این صورت این معادله را می‌توان دوباره چنین نوشت

$$\boldsymbol{\alpha} - \mathbf{x}_{n+1} = G_n(\boldsymbol{\alpha} - \mathbf{x}_n) \qquad (۹.۱۰.۲)$$

در اینجا بجاست که ماتریس ژاکوبی را برای توابع g_1 و g_2 معرفی کنیم:

$$G(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial g_1(\mathbf{x})}{\partial x_1} & \dfrac{\partial g_1(\mathbf{x})}{\partial x_2} \\ \dfrac{\partial g_2(\mathbf{x})}{\partial x_1} & \dfrac{\partial g_2(\mathbf{x})}{\partial x_2} \end{bmatrix} \qquad (۱۰.۱۰.۲)$$

در (۹.۱۰.۲)، اگر $\mathbf{x}_n$ نزدیک به $\boldsymbol{\alpha}$ باشد، G_n نزدیک به $G(\boldsymbol{\alpha})$ می‌شود. این امر موجب می‌شود اندازه یا نرم ماتریس $G(\boldsymbol{\alpha})$ در تحلیل همگرایی (۹.۱۰.۲) نقش اساسی داشته باشد. ماتریس $G(\boldsymbol{\alpha})$ نقش $g'(\alpha)$ را در نظریهٔ بخش ۵.۲، بازی می‌کند. برای اندازه‌گیری اندازهٔ خطاهای $\boldsymbol{\alpha} - \mathbf{x}_n$ و ماتریس‌های G_n و $G(\boldsymbol{\alpha})$ از نرم‌های برداری و ماتریسی (۱۶.۱.۱) و (۱۹.۱.۱) فصل ۱ استفاده می‌کنیم.

قضیهٔ ۹.۲ گیریم D مجموعه‌ای بسته، کراندار و محدب در صفحه باشد. (D را محدب گوییم، اگر برای هر دو نقطه در D، قطعه خطی هم که آنها را به هم وصل می‌کند در D باشد). فرض کنید مؤلفه‌های $\mathbf{g}(\mathbf{x})$ در تمام نقاط D پیوسته مشتقپذیر باشند. همچنین فرض کنید

۱.

$$\mathbf{g}(D) \subset D \qquad (۱۱.۱۰.۲)$$

۲.

$$\lambda = \underset{x \in D}{\text{Max}} \parallel \mathbf{G}(\mathbf{x}) \parallel_\infty < ۱ \qquad (۱۲.۱۰.۲)$$

آنگاه

(**الف**) $\mathbf{x} = \mathbf{g}(\mathbf{x})$ یک ریشهٔ یکتای $\boldsymbol{\alpha} \in D$ دارد.

(**ب**) برای هر نقطهٔ اولیهٔ داده شدهٔ $\mathbf{x}_۰ \in D$، بارست (۵.۱۰.۲) در D به $\boldsymbol{\alpha}$ می‌گراید.

(**ج**)

$$\parallel \boldsymbol{\alpha} - \mathbf{x}_{n+۱} \parallel_\infty \leq (\parallel G(\boldsymbol{\alpha}) \parallel_\infty + \epsilon_n) \parallel \boldsymbol{\alpha} - \mathbf{x}_n \parallel_\infty \qquad (۱۳.۱۰.۲)$$

$$\text{با } \varepsilon_n \to ۰ \text{ هرگاه } n \to \infty$$

برهان (**الف**) وجود نقطهٔ ثابت $\boldsymbol{\alpha}$ را می‌توان با اثبات اینکه دنبالهٔ بارستهای $\{\mathbf{x}_n\}$ از (۵.۱۰.۲) در D همگراست، نشان داد. ما اثبات وجود را به‌عنوان یک مسأله به خواننده واگذار می‌کنیم و فقط یکتایی $\boldsymbol{\alpha}$ را ثابت می‌نماییم.

فرض کنید $\boldsymbol{\alpha}$ و $\boldsymbol{\beta}$ هر دو، نقطه‌های ثابت $\mathbf{g}(\mathbf{x})$ در D باشند. در این صورت

$$\boldsymbol{\alpha} - \boldsymbol{\beta} = \mathbf{g}(\boldsymbol{\alpha}) - \mathbf{g}(\boldsymbol{\beta}) \qquad (۱۴.۱۰.۲)$$

قضیهٔ مقدار میانی را برای مؤلفهٔ iام به‌کار می‌بریم تا به‌دست آوریم

$$g_i(\boldsymbol{\alpha}) - g_i(\boldsymbol{\beta}) = \nabla g_i(\boldsymbol{\xi}^{(i)})(\boldsymbol{\alpha} - \boldsymbol{\beta}) \qquad i = ۱, ۲ \quad (۱۵.۱۰.۲)$$

که در آن

$$\nabla g_i(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial g_i}{\partial x_۱} & \dfrac{\partial g_i}{\partial x_۲} \end{bmatrix}$$

و $\xi^{(i)} \in D$ روی پاره‌خطی که $\boldsymbol{\alpha}$ را به $\boldsymbol{\beta}$ وصل می‌کند قرار دارد. چون $\| \mathbf{G}(\mathbf{x}) \| \leq \lambda < 1$، با توجّه به تعریف نرم خواهیم داشت،

$$\left|\frac{\partial g_i(\mathbf{x})}{\partial x_1}\right| + \left|\frac{\partial g_i(\mathbf{x})}{\partial x_2}\right| \leq \lambda < 1, \qquad \mathbf{x} \in D, \quad i = 1, 2$$

از ترکیب این رابطه با (۱۵.۱۰.۲)،

$$\begin{aligned} |g_i(\boldsymbol{\alpha}) - g_i(\boldsymbol{\beta})| &\leq \lambda ||\boldsymbol{\alpha} - \boldsymbol{\beta}||_\infty \\ ||\mathbf{g}(\boldsymbol{\alpha}) - \mathbf{g}(\boldsymbol{\beta})||_\infty &\leq \lambda ||\boldsymbol{\alpha} - \boldsymbol{\beta}||_\infty \end{aligned} \qquad (۱۶.۱۰.۲)$$

این رابطه وقتی با (۱۴.۱۰.۲) ترکیب شود نتیجه می‌دهد؛

$$\| \boldsymbol{\alpha} - \boldsymbol{\beta} \|_\infty \leq \lambda \| \boldsymbol{\alpha} - \boldsymbol{\beta} \|_\infty$$

که تنها به‌ازای $\boldsymbol{\alpha} = \boldsymbol{\beta}$ امکان دارد، و یکتایی $\boldsymbol{\alpha}$ را در D نشان می‌دهد.

(ب) شرط (۱۱.۱۰.۲) تضمین می‌کند که اگر $\mathbf{x}_0 \in D$، هر $\mathbf{x}_n \in D$. اکنون $\mathbf{x}_{n+1} = \mathbf{g}(\mathbf{x}_n)$ را از $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\alpha})$ کم می‌کنیم و به‌دست می‌آوریم

$$\boldsymbol{\alpha} - \mathbf{x}_{n+1} = \mathbf{g}(\boldsymbol{\alpha}) - \mathbf{g}(\mathbf{x}_n)$$

نتیجهٔ (۱۶.۱۰.۲) برای هر دو نقطه‌ای از D صادق است. با به کار بردن آن

$$\| \boldsymbol{\alpha} - \mathbf{x}_{n+1} \|_\infty \leq \lambda \| \boldsymbol{\alpha} - \mathbf{x}_n \|_\infty \qquad (۱۷.۱۰.۲)$$

و به‌طور استقرایی

$$||\boldsymbol{\alpha} - \mathbf{x}_n||_\infty \leq \lambda^n ||\boldsymbol{\alpha} - \mathbf{x}_0||_\infty \qquad (۱۸.۱۰.۲)$$

چون $\lambda < 1$، رابطهٔ اخیر نشان می‌دهد که وقتی $n \to \infty$، آنگاه $\mathbf{x}_n \to \boldsymbol{\alpha}$.

(ج) با توجّه به (۹.۱۰.۲) و استفاده از (۲۱.۱.۱)،

$$||\boldsymbol{\alpha} - \mathbf{x}_{n+1}||_\infty \leq ||\mathbf{G}_n||_\infty ||\boldsymbol{\alpha} - \mathbf{x}_n||_\infty \qquad (۱۹.۱۰.۲)$$

وقتی $n \to \infty$، نقاط $\xi_n^{(i)}$ که در محاسبهٔ G_n به‌کار رفته‌اند همگی به $\boldsymbol{\alpha}$ میل خواهند کرد، زیرا روی پاره‌خط واصل $\mathbf{x}_n$ و $\boldsymbol{\alpha}$ قرار دارند. پس وقتی $n \to \infty$، آنگاه $||G_n||_\infty \to ||G(\boldsymbol{\alpha})||_\infty$. نتیجهٔ (۱۳.۱۰.۲) را می‌توان از (۱۹.۱۰.۲) با فرض $\varepsilon_n = ||G_n||_\infty - ||G(\boldsymbol{\alpha})||_\infty$ به‌دست‌آورد. ■

قضیهٔ قبل، تعمیم قضیهٔ ۶.۲ به توابع دو متغیره است. فرع ذیل تعمیم قضیهٔ ۷.۲ است.

فرع ۱۰.۲ گیریم α یک نقطهٔ ثابت $\mathbf{g}(\mathbf{x})$ باشد و فرض می‌کنیم مؤلفه‌های $\mathbf{g}(\mathbf{x})$ در یک همسایگی α پیوسته مشتقپذیر باشند. به‌علاوه فرض می‌کنیم

$$\| G(\alpha) \|_\infty < ۱ \qquad (۲۰.۱۰.۲)$$

در این صورت برای $\mathbf{x}_۰$ که به اندازهٔ کافی نزدیک به α انتخاب شده باشد، بارست $\mathbf{x}_{n+۱} = \mathbf{g}(\mathbf{x}_n)$ به سمت α می‌گراید و نتایج قضیهٔ ۹.۲ در یک ناحیهٔ بسته، کراندار و محدب حول α معتبر است. ■

ما اثبات این فرع را به‌عنوان یک مسأله واگذار می‌کنیم. بر پایهٔ نتایج فصل ۷، همگرایی خطی $\mathbf{x}_n$ به α باز هم درست است اگر قدر مطلق همهٔ ویژه مقدارهای $G(\alpha)$ کوچکتر از یک باشند، که می‌توان نشان داد فرض ضعیفتری از (۲۰.۱۰.۲) است.

مثال مثال قبلی (۷.۱۰.۲) را ادامه می‌دهیم. به‌سادگی

$$G(\alpha) \doteq \begin{bmatrix} ۰٫۰۳۸۹۲۰ & ۰٫۰۰۰۴۰۱ \\ ۰٫۰۰۸۵۲۹ & -۰٫۰۰۶۶۱۳ \end{bmatrix}$$

حساب می‌شود. بنابراین

$$||G(\alpha)||_\infty \doteq ۰٫۰۳۹۳$$

پس شرط (۲۰.۱۰.۲) از قضیه برقرار است. از (۱۳.۱۰.۲)، رابطهٔ

$$\frac{||\alpha - \mathbf{x}_{n+۱}||_\infty}{||\alpha - \mathbf{x}_n||_\infty} \leq ||G_n||_\infty \doteq ۰٫۰۳۹۳$$

برای تمام مقادیر به اندازهٔ کافی بزرگ n، تقریباً درست است.

فرض کنید A یک ماتریس ثابت و ناتکین ۲×۲ باشد. پس می‌توانیم (۱.۱۰.۲) را به شکل زیر دوباره بنویسیم

$$\mathbf{x} = \mathbf{x} + A\mathbf{f}(\mathbf{x}) \equiv \mathbf{g}(\mathbf{x}) \qquad (۲۱.۱۰.۲)$$

مثال (۷.۱۰.۲) این روند را نشان می‌دهد. برای آنکه ببینیم A چه شرایطی باید داشته باشد، ماتریس ژاکوبی را تشکیل می‌دهیم. به‌سادگی

$$G(\mathbf{x}) = I + AF(\mathbf{x})$$

به‌دست می‌آید. که $F(\mathbf{x})$ ماتریس ژاکوبی f_1 و f_2 است.

$$F(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial f_1(\mathbf{x})}{\partial x_1} & \dfrac{\partial f_1(\mathbf{x})}{\partial x_2} \\ \dfrac{\partial f_2(\mathbf{x})}{\partial x_1} & \dfrac{\partial f_2(\mathbf{x})}{\partial x_2} \end{bmatrix} \qquad (۲۲.۱۰.۲)$$

می‌خواهیم A را به‌گونه‌ای انتخاب کنیم که (۲۰.۱۰.۲) برقرار شود. و برای همگرایی سریع، می‌خواهیم $\| G(\alpha) \|_\infty \doteq 0$، یا

$$A \doteq -F(\boldsymbol{\alpha})^{-1}$$

ماتریس در (۷.۱۰.۲) با استفاده از

$$A \doteq -F(\mathbf{x}_0)^{-1}$$

بدین طریق انتخاب شده‌است. از این رابطه به فکر استفاده از پیوسته نو کردن A، یعنی $A = -F(\mathbf{x}_n)^{-1}$، می‌افتیم. روشی که به‌دست می‌آید چنین است

$$\mathbf{x}_{n+1} = \mathbf{x}_n - F(\mathbf{x}_n)^{-1}\mathbf{f}(\mathbf{x}_n) \qquad n \geq 0 \qquad (۲۳.۱۰.۲)$$

این روش را در بخش بعد مطالعه می‌کنیم.

۱۱.۲ روش نیوتن برای دستگاههای غیرخطی

همانند روش نیوتن برای حل یک معادله، برای حل دستگاه معادلات غیر خطی بیش از یک راه برای نگریستن و پیدا کردن روش نیوتن وجود دارد. ما با یک روش تحلیلی آغاز می‌کنیم و سپس یک نگرش هندسی ارائه خواهیم داد.

قضیهٔ تیلر برای توابع دومتغیره را برای هر یک از معادله‌های $f_i(x_1, x_2) = 0$ با بسط $f_i(\boldsymbol{\alpha})$، $i = 1, 2$، حول $\mathbf{x}_0$ به‌کار می‌بریم:

$$0 = f_i(\alpha) = f_i(\mathbf{x}_0) + (\alpha_1 - x_{1,0})\frac{\partial f_i(\mathbf{x}_0)}{\partial x_1} + (\alpha_2 - x_{2,0})\frac{\partial f_i(\mathbf{x}_0)}{\partial x_2}$$

$$+\frac{1}{2}\left[(\alpha_1 - x_{1,0})\frac{\partial}{\partial x_1} + (\alpha_2 - x_{2,0})\frac{\partial}{\partial x_2}\right]^2 f_i(\boldsymbol{\xi}^{(i)}) \qquad (۱.۱۱.۲)$$

$\xi^{(i)}$ روی پاره خطّ و اصل بین $\mathbf{x}_0$ و $\boldsymbol{\alpha}$ واقع است. اگر از جمله‌های درجهٔ دوّم صرف‌نظر کنیم تقریب زیر را به‌دست می‌آوریم

$$0 \doteq f_1(\mathbf{x}_0) + (\alpha_1 - x_{1,0})\frac{\partial f_1(\mathbf{x}_0)}{\partial x_1} + (\alpha_2 - x_{2,0})\frac{\partial f_1(\mathbf{x}_0)}{\partial x_2}$$

$$0 \doteq f_2(\mathbf{x}_0) + (\alpha_1 - x_{1,0})\frac{\partial f_2(\mathbf{x}_0)}{\partial x_1} + (\alpha_2 - x_{2,0})\frac{\partial f_2(\mathbf{x}_0)}{\partial x_2} \qquad (۲.۱۱.۲)$$

در شکل ماتریسی

$$0 \doteq \mathbf{f}(\mathbf{x}_0) + F(\mathbf{x}_0)(\boldsymbol{\alpha} - \mathbf{x}_0) \qquad (۳.۱۱.۲)$$

که $F(\mathbf{x}_0)$ ماتریس ژاکوبی f است که در (۲۲.۱۰.۲) داده شده است.

اگر رابطهٔ فوق را نسبت به $\boldsymbol{\alpha}$ حل کنیم داریم:

$$\boldsymbol{\alpha} \doteq \mathbf{x}_0 - F(\mathbf{x}_0)^{-1}\mathbf{f}(\mathbf{x}_0) \equiv \mathbf{x}_1$$

تقریب $\mathbf{x}_1$ بایستی یک بهبود $\mathbf{x}_0$ باشد، به شرطی که $\mathbf{x}_0$ به اندازهٔ کافی به $\boldsymbol{\alpha}$ نزدیک انتخاب شده باشد. این عمل ما را به روشی بارستی می‌رساند که ابتدا در پایان بخش قبلی به‌دست‌آمده بود

$$\mathbf{x}_{n+1} = \mathbf{x}_n - F(\mathbf{x}_n)^{-1}\mathbf{f}(\mathbf{x}_n) \qquad n \geq 0 \qquad (۴.۱۱.۲)$$

این روش نیوتن برای حل دستگاه غیرخطی $\mathbf{f}(\mathbf{x}) = 0$ است.

در کاربرد واقعی، $F(\mathbf{x}_n)$ را وارون نمی‌کنیم، به‌ویژه برای دستگاههای بیش از دو معادله. به جای آن، یک دستگاه خطی را برای یک جملهٔ اصلاحی $\mathbf{x}_n$ حل می‌کنیم:

$$\mathbf{F}(\mathbf{x}_n)\boldsymbol{\delta}_{n+1} = -\mathbf{f}(\mathbf{x}_n)$$

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \boldsymbol{\delta}_{n+1} \qquad (۵.۱۱.۲)$$

این عمل زمان محاسبات را بسیار کوتاهتر می‌کند، زیرا حدود یک سوم عملیات وارون‌سازی $F(\mathbf{x}_n)$کار می‌برد. بخش‌های ۱.۸ و ۲.۸ را برای بحث در حل عددی دستگاه معادلات خطی ملاحظه کنید.

یک راه هندسی برای پیداکردن روش نیوتن، شبیه به تقریب خط مماسی که برای یک معادلهٔ غیرخطی در بخش ۲.۲ به‌کار رفت، وجود دارد. نمودار معادلهٔ

$$z = f_i(\mathbf{x}_0) + (x_1 - x_{1,0})\frac{\partial f_i(\mathbf{x}_0)}{\partial x_1} + (x_2 - x_{2,0})\frac{\partial f_i(\mathbf{x}_0)}{\partial x_2} \equiv p_i(x_1, x_2)$$

در فضا صفحه‌ای است مماس بر نمودار $z = f_i(x_1, x_2)$ در نقطهٔ $\mathbf{x}_0$ برای $i = 1, 2$. اگر $\mathbf{x}_0$ نزدیک به $\boldsymbol{\alpha}$ باشد، آنگاه این صفحات مماس باید تقریبات خوبی برای رویه‌های متناظر $z = f_i(x_1, x_2)$ در $\mathbf{x} = (x_1, x_2)$ نزدیک به $\boldsymbol{\alpha}$، باشند. در این صورت محل برخورد خم‌های صفر صفحات مماس $z = p_i(x_1, x_2)$ باید تقریب خوبی برای محل برخورد متناظر α از خم‌های صفر رویه‌های اصلی $z = f_i(x_1, x_2)$ باشد. این به نتیجهٔ (۲.۱۱.۲) می‌انجامد. محل برخورد خمهای صفر $z = p_i(x_1, x_2)$، $i = 1, 2$، نقطهٔ $\mathbf{x}_1$ است.

مثال دستگاه زیر را در نظر می‌گیریم

$$f_1 \equiv 4x_1^2 + x_2^2 - 4 = 0 \qquad f_2 \equiv x_1 + x_2 - \sin(x_1 - x_2) = 0$$

دستگاه فقط دو ریشه دارد، یکی نزدیک به $(1, 0)$ و دیگری قرینهٔ آن نسبت به مرکزنزدیک به $(-1, 0)$. با استفاده از (۴.۱۱.۲) با $\mathbf{x}_0 = (1, 0)$، نتایج جدول ۲۰.۲ را به‌دست می‌آوریم.

تحلیل همگرایی. برای تحلیل همگرایی روش نیوتن، (۴.۱۱.۲)، آن را مانندیک روش بارستی نقطهٔ ثابت با

$$g(\mathbf{x}) = \mathbf{x} - F(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x}) \qquad (۶.۱۱.۲)$$

در نظر می‌گیریم. همچنین فرض می‌کنیم

دترمینان$F(\boldsymbol{\alpha})$مخالف صفر است

که مشابه این فرض است که $\boldsymbol{\alpha}$ یک ریشهٔ ساده است وقتی با یک معادلهٔ تنها نظیر معادلهٔ مربوط به قضیهٔ ۱.۲ سروکار داریم. در این صورت می‌توان نشان داد که ژاکوبی $G(\mathbf{x})$ از (۶.۱۱.۲) در $\mathbf{x} = \boldsymbol{\alpha}$ صفر است (مسألهٔ ۵۳ را ببینید)؛ در نتیجه، شرط (۲۰.۱۰.۲) برقرار است.

جدول ۲۰.۲ مثال از روش نیوتن

n	$x_{1,n}$	$x_{2,n}$	$f_1(\mathbf{x}_n)$	$f_2(\mathbf{x}_n)$
۰	۱٫۰	۰٫۰	۰٫۰	۱٫۵۹E − ۱
۱	۱٫۰	−۰٫۱۰۲۹۲۰۷۱۵۴	۱٫۰۶E − ۲	۴٫۵۵E − ۳
۲	۰٫۹۹۸۶۰۸۷۵۹۸	−۰٫۱۰۵۵۳۰۷۲۳۹	۱٫۴۶E − ۵	۶٫۶۳E − ۷
۳	۰٫۹۹۸۶۰۶۹۴۴۱	−۰٫۱۰۵۵۳۰۴۹۲۳	۱٫۳۲E − ۱۱	۱٫۸۷E − ۱۲

در این صورت فرع ۲ـ۱۰ ایجاب می‌کند که $\mathbf{x}_n$ به $\boldsymbol{\alpha}$ همگرا باشد به شرطی که $\mathbf{x}_0$ به اندازهٔ کافی به $\boldsymbol{\alpha}$ نزدیک باشد. به علاوه می‌توان نشان داد که این بارستی از درجهٔ دو است. به‌ویژه فرمولهای (۱.۱۱.۲) و (۴.۱۱.۲) را می‌توان ترکیب کرده نتیجهٔ

$$||\boldsymbol{\alpha} - \mathbf{x}_{n+1}||_\infty \leq B||\boldsymbol{\alpha} - \mathbf{x}_n||_\infty^2 \qquad n \geq 0 \qquad (۷.۱۱.۲)$$

را برای مقدار ثابتی مانند $B > 0$ به‌دست‌آورد.

شکلهای گوناگون روش نیوتن. روش نیوتن در مقایسه با سایر روشهای حل دستگاه معادلات غیرخطی، دارای محاسن و معایبی است. از جملهٔ محاسن آن دارا بودن شکل ساده و انعطاف‌پذیری زیاد برای کاربرد آن در مسائل بسیار متنوع است. اگر نخواهیم زحمت محاسبهٔ مشتقات جزیی در برنامهٔ رایانه‌یی را به خود بدهیم می‌توانیم از تفاضلات تقریبی استفاده کنیم. به‌عنوان مثال، معمولاً از

$$\frac{\partial f_i(x_1, x_2)}{\partial x_1} \doteq \frac{f_i(x_1 + \varepsilon, x_2) - f_i(x_1, x_2)}{\varepsilon} \qquad (۸.۱۱.۲)$$

با عدد خیلی کوچک ε استفاده می‌کنیم. برای بحث مفصّل در انتخاب ε، دنیس و اشنابل[1] (۱۹۸۳، صص ۹۴ـ۹۹) را ببینید.

اولین عیب روش نیوتن این است که روشهای دیگری هستند که (۱) کار با آنها ارزانتراست، و یا (۲) برای بعضی از رده‌های مسائل کار با آنها آسانتر است. برای یک دستگاه m معادلهٔ m مجهولی غیرخطّی، هر بارست برای روش نیوتن بطورکلی مستلزم $m^2 + m$ محاسبهٔ تابع است. به علاوه، روش نیوتن برای هر بارست به حل یک دستگاه خطی m معادلهٔ m مجهولی نیاز دارد، که به بهای حدود $\frac{2}{3}m^3$ عمل حساب، برای هر دستگاه خطی تمام می‌شود. روشهای دیگری وجود دارند که در همگرایی ریاضی، یک اندازه یا تقریباً یک اندازه سرعت دارند ولی به تعداد کمتری محاسبهٔ تابع و عملیات حساب رایانه‌ای نیاز دارند. این روشها را معمولاً روشهای نیوتنی‌نما، شبه‌نیوتنی یا نیوتنی اصلاح شده می‌گویند. برای یک توضیح کلی دربارهٔ بسیاری از این روشها دنیس و اشنابل (۱۹۸۳) را ببینید.

یک اصلاح سادهٔ روش نیوتن تثبیت ماتریس ژاکوبی برای چند، مثلاً k مرحله است؛

$$\mathbf{x}_{rk+j+1} = \mathbf{x}_{rk+j} - F(\mathbf{x}_{rk})^{-1}\mathbf{f}(\mathbf{x}_{rk+j}) \qquad j = 0, 1, \ldots, k-1 \qquad (۹.۱۱.۲)$$

برای $r = 0, 1, 2, \ldots$. این بدان معنی است که دستگاه خطی

$$F(\mathbf{x}_{rk})\boldsymbol{\delta}_{rk+j+1} = -\mathbf{f}(\mathbf{x}_{rk+j})$$

1. Dennis & Schanel

$$\mathbf{x}_{rk+j+1}=\mathbf{x}_{rk+j}+\boldsymbol{\delta}_{rk+j+1}, \qquad j=0,1,\ldots,k-1, \quad (10.11.2)$$

خیلی بهتر از روش اصلی نیوتن (5.11.2) حل می‌شود. حل دستگاه خطی از مرتبهٔ m در حدود $\frac{2}{3}m^3$ عمل در اولین حالت، $j=0$، نیاز دارد. ولی در هر یک از حالتهای بعدی، $j=1,\ldots,k-1$، فقط $2m^2$ عمل حساب برای حل آن لازم است. بخش ۱.۸ را برای توضیح کامل ببینید. سرعت همگرایی (9.11.2) آهسته‌تر از سرعت همگرایی روش اصلی (4.11.2) است، ولی زمان محاسبهٔ واقعی روش اصلاح شده غالباً خیلی کمتر از زمان روش اصلی‌است. برای توضیحات بیشتر این مسأله پوترا و پتاک[1] (۱۹۸۴، ص ۱۱۹) را ببینید.

یک مشکل دوّم با روش نیوتن و بسیاری از روشهای دیگر، این است که اغلب $\mathbf{x}_0$ باید بطور قابل قبولی به $\boldsymbol{\alpha}$ نزدیک باشد تا همگرایی حاصل شود. اصلاحاتی از روش نیوتن وجود دارند که برای انتخابهای بد $\mathbf{x}_0$، همگرایی را تسریع می‌کنند. برای مثال، تعریف می‌کنیم

$$\mathbf{x}_{n+1}=\mathbf{x}_n+s\mathbf{d}_n \qquad \mathbf{d}_n=-F(\mathbf{x}_n)^{-1}\mathbf{f}(\mathbf{x}_n) \quad (11.11.2)$$

و $s>0$ را طوری انتخاب می‌کنیم که

$$\| \mathbf{f}(\mathbf{x}_n+s\mathbf{d}_n) \|_2^2=\sum_{j=1}^{m}[f_j(\mathbf{x}_n+s\mathbf{d}_n)]^2 \quad (12.11.2)$$

می‌نیمم شود. انتخاب $s=1$ در (11.11.2) روش نیوتن را به‌دست می‌دهد، ولی این مقدار ممکن است بهترین انتخاب نباشد. در بعضی حالات، لازم است s بسیار کوچکتر از یک باشد، حداقل در ابتدا، تا همگرایی تضمین شود. برای بحث مفصّلتر دنیس و اشنابل (۱۹۸۳، فصل ۶) را ببینید.

برای تحلیلی از بعضی برنامه‌های حل دستگاههای غیرخطی، هیبرت[2] (۱۹۸۲) را ببینید. نامبرده دربارهٔ مشکلات چنین نرم‌افزارهایی نیز بحث نموده است.

۱۲.۲ بهینه‌سازی نامقیّد

بهینه‌سازی به پیدا کردن ماکسیمم یا مینیمم یک تابع پیوستهٔ $f(x_1,\ldots,x_n)$ گفته می‌شود. این مسأله‌ای بسیار مهم است، به ویژه در مهندسی صنایع جدید، علم مدیریت و سایر زمینه‌ها. در این بخش از بعضی روشها بحث می‌شود و محاسبهٔ مینیمم یا ماکسیمم یک تابع $f(x_1,\ldots,x_n)$ نشان داده می‌شود. در اینجا هیچ الگوریتم دقیقی داده نشده است زیرا این کار مفصلی را می‌طلبد.

1. Potra and Ptak 2. Hiebert

برای عرضهٔ بسیاری از مسائل از نماد بردار استفاده شده است تا به‌طور کلّی نتایج برای تعداد m متغیر به‌دست آید. ما فقط بهینه‌سازی غیر مقیّد را مورد توجّه قرار می‌دهیم، که در آن هیچ محدودیتی برای $(x_1, \ldots, x_m)$ وجود ندارد. همچنین فقط برای سادگی فرض می‌کنیم $f(x_1, \ldots, x_m)$ برای تمام مقادیر $(x_1, \ldots, x_m)$ معین است.

چون رفتار تابع $f(\mathbf{x})$ممکن است خیلی تغییر کند، مسأله باید محدودتر شود. یک نقطهٔ α **یک مینیمم موضعی** اکید f خوانده می‌شود اگر برای همهٔ مقادیر $\mathbf{x}$ نزدیک به α و $\mathbf{x} \neq \alpha$ نابرابری $\mathbf{f(x) > f(\alpha)}$ برقرار باشد. ما خود را به یافتن مینیمم موضعی اکید $f(\mathbf{x})$ محدود می‌نماییم. معمولاً یک حدس اولیهٔ $\mathbf{x}_0$ از α معلوم است و تابع $f(\mathbf{x})$ دوبار پیوسته مشتق‌پذیر نسبت به متغیرهای $x_1, \ldots, x_m$ فرض می‌شود.

بیانی دیگر به‌صورت دستگاه غیرخطی. با فرض مشتقپذیری، یک شرط لازم برای اینکه α مینیمم موضعی اکید باشد

$$\frac{\partial f(\alpha)}{\partial x_i} = 0 \qquad i = 1, \ldots, m \qquad (۱.۱۲.۲)$$

است. پس دستگاه غیرخطی

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = 0 \qquad i = 1, \ldots, m \qquad (۲.۱۲.۲)$$

را می‌توان حل کرد و بررسی نمود که هر جواب محاسبه شده ماکسیمم یا مینیمم موضعی است یا هیچکدام. برای نمادگذاری، بردار گرادیان را معرفی می‌کنیم

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \vdots \\ \dfrac{\partial f}{\partial x_m} \end{bmatrix}$$

با استفاده از این بردار، دستگاه (۲.۱۲.۲) به شکل فشرده‌تر زیر نوشته می‌شود

$$\nabla f(\mathbf{x}) = 0 \qquad (۳.۱۲.۲)$$

برای حل (۳.۱۲.۲)، می‌توان از روش (۴.۱۱.۲) نیوتن و همچنین سایر روشهای ریشه‌یابی استفاده کرد. با به‌کار بردن دستگاههای غیرخطی روش نیوتن رابطهٔ زیر به‌دست می‌آید

$$\mathbf{x}_{n+1} = \mathbf{x}_n - H(\mathbf{x}_n)^{-1} \nabla f(\mathbf{x}_n) \qquad n \geq 0 \qquad (۴.۱۲.۲)$$

که در آن $H(\mathbf{x})$ ماتریس هسه‌ای تابع f است،

$$\boldsymbol{H}(\mathbf{x})_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}, \qquad 1 \le i, j \le m$$

اگر $\boldsymbol{\alpha}$ یک مینیمم موضعی اکید f باشد، با استفاده از قضیهٔ (۱۲.۱۱.۱) تیلر می‌توان نشان داد که $\boldsymbol{H}(\boldsymbol{\alpha})$ یک ماتریس ناتکین است؛ پس $\boldsymbol{H}(\mathbf{x})$ برای $\mathbf{x}$ های نزدیک به $\boldsymbol{\alpha}$ ناتکین خواهد بود. برای همگرایی، تحلیل روش نیوتن در بخش قبل را می‌توان به‌کار برد و همگرایی مرتبهٔ دوم $\mathbf{x}_n$ به $\boldsymbol{\alpha}$ را مشروط برآنکه $\mathbf{x}_0$ به اندازهٔ کافی نزدیک به $\boldsymbol{\alpha}$ انتخاب شده باشد، ثابت نمود.

معایب عمدهٔ روش بارستی (۴.۱۲.۲) همانهایی هستند که برای روش نیوتن در حل دستگاههای غیرخطی، در آخرین قسمت بخش قبل داده شد. روشهای بهینه‌سازی کاراتری وجود دارند که‌در آنها برای پیداکردن تقریب $\boldsymbol{\alpha}$، فقط از $f(\mathbf{x})$ و $\nabla f(\mathbf{x})$ استفاده می‌شود. ممکن است این روشها به تعداد بارستهای بیشتری نیاز داشته باشند ولی معمولاً زمان کلی محاسبات بسیار کمتر از زمان محاسبه در روش نیوتن خواهد بود. به علاوه این روشها درپی یافتن همگرایی، در مجموعهٔ بزرگتری از مقادیر اولیهٔ $\mathbf{x}_0$ هستند.

روشهای کاهشی. فرض کنید می‌کوشیم یک تابع $f(\mathbf{x})$ را مینیمم کنیم. بیشتر روشها برای این کار، بر پایهٔ فرایند بارستیِ دو مرحله‌ای کلّی زیر بنا شده‌اند.

مرحلهٔ ک ۱: در نقطهٔ $\mathbf{x}_n$، یک امتداد $\mathbf{d}_n$ می‌گیریم به‌گونه‌ای که وقتی $\mathbf{x}$ در امتداد $\mathbf{d}_n$ از $\mathbf{x}_n$ دور می‌شود، $f(\mathbf{x})$ کاهش یابد.

مرحلهٔ ک ۲: گیریم $\mathbf{x}_{n+1} = \mathbf{x}_n + s\mathbf{d}_n$، که s به گونه‌ای انتخاب شده است که تابعِ

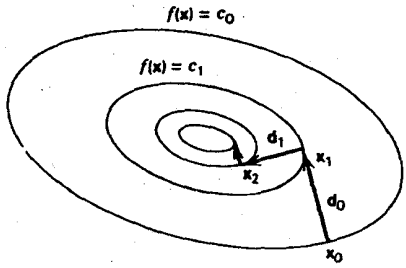
$$\varphi(s) = f(\mathbf{x}_n + s\mathbf{d}_n) \qquad s \ge 0 \qquad (۵.۱۲.۲)$$

را مینیمم سازد. معمولاً، s کوچکترین مینیمم نسبی مثبت $\varphi(s)$ انتخاب‌می‌شود. این گونه روشها، **روشهای کاهشی** خوانده می‌شوند. با هر بارست

$$f(\mathbf{x}_{n+1}) < f(\mathbf{x}_n)$$

روشهای کاهشی در شرایطی کلّیتر از شرایط در روش (۴.۱۲.۲) نیوتن، ضامن همگرایی هستند. رویه تراز،

$$C = \{\mathbf{x} \mid f(\mathbf{x}) = f(\mathbf{x}_0)\}$$



شکل ۸.۲ نمایش روش تندترین کاهش

را در نظر می‌گیریم، فقط تکهٔ همبندی از آن، مثلاً C' را که شامل $\mathbf{x}_0$ باشد اختیار می‌کنیم. اگر C' کراندار و $\boldsymbol{\alpha}$ را در درون خود داشته باشد، آنگاه روشهای کاهشی در شرایط بسیار کلی، همگرا خواهند بود. این امر در شکل ۸.۲ برای حالت دو متغیره نشان داده شده است. چندین خم تراز $f(x_1, x_2) = c$ برای مجموعه‌ای از مقادیر c که به $f(\boldsymbol{\alpha})$ نزدیک می‌شوند، نشان داده شده است. بردارهای $\mathbf{d}_n$ امتدادهایی هستند که در امتداد آنها $f(\mathbf{x})$ کاهش می‌یابد.

راههای زیادی برای انتخاب امتدادهای $\mathbf{d}_n$ وجود دارند و شناخته‌ترین آنها در زیر می‌آیند:

۱. روش تندترین کاهش. در اینجا $\mathbf{d}_n = -\nabla f(\mathbf{x}_n)$. این امتدادی است که در جهت آن $f(\mathbf{x})$ تندترین کاهش را وقتی از $\mathbf{x}_n$ دور می‌شود، خواهد داشت. این خطّ مشی خوبی در نزدیک $\mathbf{x}_n$ است ولی غالباً به یک خط ّمشی ضعیف برای همگرایی سریع به $\boldsymbol{\alpha}$ تبدیل می‌شود.

۲. روشهای شبه‌نیوتنی. به این روشها می‌توان به‌عنوان تقریبهای روش (۴.۱۲.۲) نیوتن نگاه کرد. در این روشها از تقریبهای $H(\mathbf{x}_n)$ و $H(\mathbf{x}_n)^{-1}$ که محاسبهٔ آنها آسان است، استفاده می‌شود و این روشها نیز کاهشی هستند. معروفترین مثال، روش دیویدسن ـ فلچر ـ پاوئل[1] و روشهای بروید‌ن[2] هستند.

۳. روش گرادیان مزدوج. در این روش برای تولید امتدادهای $\mathbf{d}_n$، از تعمیم مفهوم در یک فضای برداری استفاده می‌شود. که $\mathbf{d}_n$ها به طریق بهینه‌ای به تابع $f(x)$ که باید بهینه شود مربوط می‌شوند. در فصل ۸ روش گرادیان مزدوج برای حل دستگاههای معادله‌های خطی به‌کار گرفته شده است.

روشهای دیگری برای مینیمم کردن یک تابع وجود دارند، ولی تعداد آنها آن قدر زیاد است که نمی‌توان در اینجا آورد. به‌عنوان منابع کلی برای مفاهیم بالا، مراجعه کنید به: دنیس اشنابل (۱۹۸۳) و فلچر (۱۹۸۰)، لوئنبرگر[3] (۱۹۸۴) و گیل[4] و همکاران (۱۹۸۰). یک روش مهم و کاملاً متفاوت برای مینیمم کردن یک تابع روش سارکی است که در نلدر[5] و مِد[6] (۱۹۶۵) داده

1. Davidson-Flecher-Powell
2. Broyden
3. Luenberger
4. Gilletal
5. Nelder
6. Mead

شده و بحث آن درگیل و همکاران (۱۹۸۱، ص ۹۴) و وودز[1] (۱۹۸۵،فصل ۲) داده شده است. در این روش فقط از مقادیر تابع (نه مقادیر مشتق) استفاده می‌شود، و ظاهراً برای توابع نوفه‌ای بسیار مناسب است.

یک طرح مهم برای به وجود آوردن برنامه‌هایی برای حل مسائل بهینه‌سازی و دستگاههای غیرخطی در Argonne National Laboraltory در جریان است. این بسته برنامه MINPACK خوانده می‌شود و نسخهٔ اول آن در دسترس است [موره و همکاران (۱۹۸۰) و موره و همکاران (۱۹۸۴) را ببینید]. این بسته شامل برنامه‌هایی برای دستگاههای غیرخطی و مسائل کمترین مربعات غیرخطی است. نسخه‌های‌دیگر قرار است شامل برنامه‌هایی برای مسائل بهینه‌سازی مقیّد و نامقید باشد.

بحث در آثار خواندنی

نوشته‌های زیادی در مورد محاسبهٔ ریشه‌های یک معادله وجود دارد. برای توضیحاتی بیشتر از آنچه که در این کتاب آمده است، به کتابهای هاوسهولدر (۱۹۷۰)، آستروفسکی (۱۹۷۳)، تراؤب (۱۹۶۴) مراجعه کنید. روش نیوتن یکی از روشهایی است که بسیار به‌کارگرفته‌شده‌است و گسترش آن مرهون افراد زیادی است. برای یک گزارش تاریخی در سهم نیوتن، رافسون، و کوشی در این روش به گولداستاین (۱۹۷۷، صص ۶۴ و ۲۷۸) مراجعه کنید.

برای برنامه‌های رایانه‌یی، اغلب افراد، هنوز برنامه‌ای را که شخصاً تهیه دیده و مناسب برای کار خودشان است به‌کار می‌برند. ولی، باید یکی از برنامه‌هایی را که برای منظورهای کلّی در سالهای اخیر نوشته و توسعه داده شده است و در کتابخانه‌های نرم‌افزاری بازرگانی در دسترس هستند در نظر گرفت. این برنامه‌ها معمولاً دقیق و کارا هستند و استفاده از آنها آسان است. در میان این برنامه‌های کلی آنها که از برِنت (۱۹۷۳) و دکر (۱۹۶۹) گرفته شده‌اند، مشهورتر از همه هستند، و توسعهٔ بیشتر آنها هنوز ادامه دارد، مانند له[2] (۱۹۸۵) در کتابخانه‌های رایانه‌ای IMSL و NAG این برنامه‌ها و برنامه‌های عالی ریشه‌یابی دیگر موجود است.

ریشه‌یابی چندجمله‌ییها موضوعی بسیار قدیمی است که به زمانهای یونان‌باستان می‌رسد. روشها و نوشته‌های زیادی در مورد آنها وجود دارد، و در دو سه دههٔ اخیر روشهای جدید زیادی برای آنها یافته‌اند. برای آشنایی با این موضوع کتابهای مقدّماتی در این زمینه، کتابهای هنریچی[3] و دوژون[4] (۱۹۶۹)، هنریچی (۱۹۷۴، فصل ۶)، هاوسهولدر (۱۹۷۰)، تراب (۱۹۶۴) وکتابنامه‌های آنها را ملاحظه نمائید. مقالهٔ ویلکینسن (۱۹۸۴) بعضی مشکلات عملی حل مسألهٔ ریشه‌یابی چند جمله‌ییها را با رایانه نشان می‌دهد. برنامه‌های رایانه‌ای دقیق، کارا، خودکار و قابل اعتماد برای

1. Woods 2. Le 3. Henrici 4. Dejon

ریشه‌یابی چندجمله‌ییها نوشته شده‌اند. از این جمله‌اند (الف) برنامه‌های جنکینز[1] (۱۹۷۵) و جنکینز و تراوب (۱۹۷۰)،(۱۹۷۲)، و (ب) برنامهٔ ZERPOL اثر اسمیت (۱۹۶۷) که بر مبنای روش لاگر[2] نوشته شده است. هاوسهولدر [(۱۹۷۰، ص ۱۷۶)، کاهان (۱۹۶۷) را ببینید.]. این برنامه‌های خودکار، هم از نظر ریاضی و هم از نظر الگوریتمی، آن قدر پیچیده‌اند که نمی‌توان در یک کتاب مقدماتی، چون این کتاب به معرفی آنها پرداخت. مع هذا کار با آنها باارزش است. بسیاری از افراد نمی‌توانند برنامه‌ای بنویسند که از نظر سرعت و دقت بتواند با آنها رقابت کند. به‌ویژه دقّت، بسیار اهمیّت دارد، زیرا مسألهٔ ریشه‌یابی چندجمله‌ای، می‌تواند نسبت به خطای گردکردن، همان گونه که در مثالهای قبلی نشان داده شد بسیار حساس باشد.

مطالعهٔ روشهای عددی برای حل دستگاههای معادله‌های غیرخطی و مسائل بهینه‌سازی، اخیراً یک زمینهٔ بسیار متداول تحقیق گردیده است. برای آشنایی با روشهای عددی حل دستگاههای غیرخطی، بیکر و فیلیپس[3] (۱۹۸۱، فصل ۱)، اورتگا و راینبولت[4] (۱۹۷۰) و راینبولت (۱۹۷۴) را ببینید. برای تعمیم این روشها به معادلات دیفرانسیل و انتگرالی غیرخطی، بیکر و فیلیپس (۱۹۸۱)، کانتوروویچ (۱۹۴۸) [که مقاله‌ای کلاسیک در این زمینه است]، کانتوروویچ و اکیلوف (۱۹۶۴) و رال (۱۹۶۹) را ببینید. برای یک بازنگری روشهای عددی در بهینه‌سازی، دنیس (۱۹۸۴) و پاول (۱۹۸۲) را ببینید. آشنایی کلّی در اشنابل و دنیس (۱۹۸۳)، فلچر (۱۹۸۰) و (۱۹۸۱)، گیل و همکاران (۱۹۸۱)، لوئنبرگر[5] (۱۹۸۴) داده شده است. به‌عنوان یک مثال از تحقیقات اخیر در نظریهٔ بهینه‌سازی و توسعهٔ نرم‌افزار، باگز و همکاران (۱۹۸۵) را ببینید. برای برنامه‌های کامپیوتری هیبرت (۱۹۸۲) و موره و همکاران (۱۹۸۴) را ببینید.

مراجع

Baker, C., and C. Phillips, eds. (1981). ***The Numerical Solution of Nonlinear Problems*****. Clarendon Press, Oxford, England.**

Boggs, P., R. Byrd, and R. Schnabel, eds. (1985). ***Numerical Optimization 1984*****. Society for Industrial and Applied Mathematics, Philadelphia.**

Brent, R. (1973). ***Algorithms for Minimization Without Derivatives*****. Prentice-Hall, Englewood Cliffs, N.J.**

Byrne, G., and C. Hall, eds. (1973). ***Numerical Solution of Systems of Nonlinear Algebraic Equations*****. Academic Press, New York.**

1. Jenkins 2. Laguerre 3. Baker and Phillips
4. Ortega and Rheinboldt 5. Luenberger

Dejon, B., and P. Henrici, eds. (1969). *Constructive Aspects of the Fundamental Theorem of Algebra*. Wiley, New York.

Dekker, T. (1969). Finding a zero by means of successive linear interpolation. In B. Dejon and P. Henrici (eds.), *Constructive Aspects of the Fundamental Theorem of Algebra*, pp. 37–51. Wiley, New York.

Dennis, J. (1984). A user's guide to nonlinear optimization algorithms. *Proc. IEEE*, **72**, 1765–1776.

Dennis, J., and R. Schnabel (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, N.J.

Fletcher, R. (1980). *Practical Methods of Optimization*, Vol. 1, *Unconstrained Optimization*. Wiley, New York.

Fletcher, R. (1981). *Practical Methods of Optimization*, Vol. 2, *Constrained Optimization*. Wiley, New York.

Forsythe, G. (1969). What is a satisfactory quadratic equation solver? In B. Dejon and P. Henrici (eds.), *Constructive Aspects of the Fundamental Theorem of Algebra*, pp. 53–61. Wiley, New York.

Gill, P., W. Murray, and M. Wright (1981). *Practical Optimization*. Academic Press, New York.

Goldstine, H. (1977). *A History of Numerical Analysis*. Springer-Verlag, New York.

Henrici, P. (1974). *Applied and Computational Complex Analysis*, Vol. 1. Wiley, New York.

Hiebert, K. (1982). An evaluation of mathematical software that solve systems of nonlinear equations. *ACM Trans. Math. Softw.*, **11**, 250–262.

Householder, A. (1970). *The Numerical Treatment of a Single Nonlinear Equation*. McGraw-Hill, New York.

Jenkins, M. (1975). Algorithm 493: Zeroes of a real polynomial. *ACM Trans. Math. Softw.*, **1**, 178–189.

Jenkins, M., and J. Traub (1970). A three state algorithm for real polynomials using quadratic iteration. *SIAM J. Numer. Anal.*, **7**, 545–566.

Jenkins, M., and J. Traub (1972). Algorithm 419—Zeros of a complex polynomial. *Commun. ACM*, **15**, 97–99.

Kahan, W. (1967). Laguerre's method and a circle which contains at least one zero of a polynomial. *SIAM J. Numer. Anal.*, **4**, 474–482.

Kantorovich, L. (1948). Functional analysis and applied mathematics. *Usp. Mat. Nauk*, **3**, 89–185.

Kantorovich, L., and G. Akilov (1964). *Functional Analysis in Normed Spaces*. Pergamon, London.

Le, D. (1985). An efficient derivative-free method for solving nonlinear equations. *ACM Trans. Math. Softw.*, **11**, 250–262.

Luenberger, D. (1984). *Linear and Nonlinear Programming*, 2nd ed. Wiley, New York.

Moré, J., B. Garbow, and K. Hillstrom (1980). *User Guide for MINPACK-1*. Argonne Nat. Lab. Rep. ANL-80-74.

Moré, J., and D. Sorenson. Newton's method. In *Studies in Numerical Analysis*, G. Golub (ed.), pp. 29–82. Math. Assoc. America, Washington, D.C.

Moré, J., D. Sorenson, B. Garbow, and K. Hillstrom (1984). The MINPACK project. In *Sources and Development of Mathematical Software*, Cowell (ed.), pp. 88–111. Prentice-Hall, Englewood Cliffs, N.J.

Nelder, A., and R. Mead (1965). A simplex method for function minimization. *Comput. J.*, 7, 308–313.

Ortega, J., and W. Rheinboldt (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York.

Ostrowski, A. (1973). *Solution of Equations in Euclidean and Banach Spaces*. Academic Press, New York.

Peters, G., and J. Wilkinson (1971). Practical problems arising in the solution of polynomial equations. *J. Inst. Math. Its Appl.* **8**, 16–35.

Potra, F., and V. Ptak (1984). *Nondiscrete Induction and Iterative Processes*. Pitman, Boston.

Powell, M., ed. (1982). *Nonlinear Optimization 1981*. NATO Conf. Ser. Academic Press, New York.

Rall, L. (1969). *Computational Solution of Nonlinear Operator Equations*. Wiley, New York.

Rheinboldt, W. (1974). *Methods for Solving Systems of Nonlinear Equations*. Society for Industrial and Applied Mathematics, Philadelphia.

Smith, B. (1967). ZERPOL: A zero finding algorithm for polynomials using Laguerre's method. Dept. of Computer Science, Univ. Toronto, Toronto, Ont., Canada.

Traub, J. (1964). *Iterative Methods for the Solution of Equations*. Prentice-Hall, Englewood Cliffs, N.J.

Whitley, V. (1968). Certification of algorithm 196: Muller's method for finding roots of an arbitrary function. *Commun. ACM* **11**, 12–14.

Wilkinson, J. (1963). *Rounding Errors in Algebraic Processes*. Prentice-Hall, Englewood Cliffs, N.J.

Wilkinson, J. (1984). The perfidious polynomial. In *Studies in Numerical Analysis*, G. Golub (ed.). Math. Assoc. America, Washington, D.C.

Woods, D. (1985). An interactive approach for solving multi-objective optimization problems. Ph.D. dissertation, William Marsh Rice Univ., Houston, Tex.

مسائل

۱. مثالهای مقدماتی برای $f(x) = a - (۱/x)$ به حاصلضرب نامتناهی زیر مربوط‌اند

$$\prod_{j=۰}^{\infty}(۱ + r^{۲^j}) \equiv \lim_{n\to\infty}[(۱ + r)(۱ + r^۲)(۱ + r^۴)\dots(۱ + r^{۲^n})]$$

با استفاده از فرمولهای (۶.۰.۲) و (۹.۰.۲) می‌توانیم مقدار حاصلضرب نامحدود فوق رامحاسبه کنیم. این مقدار چیست، و چه شرطی برای r باید برقرار باشد تا حاصلضرب فوق همگرا شود؟

راهنمایی: بگیرید $r = r_۰$ و x_n را برحسب $x_۰$ و $r_۰$ بنویسید.

۲. یک برنامه بنویسید که در آن از الگوریتم نیم‌سازی که در بخش ۱.۲ داده شده استفاده شود. برنامه را برای محاسبهٔ ریشه‌های حقیقی معادلات زیر به‌کار برید. تحمل خطا را برابر $\varepsilon = ۱۰^{-۵}$ انتخاب کنید.

(الف) $e^x - ۳x^۲ = ۰$ (ب) $x^۳ = x^۲ + x + ۱$

(ج) $e^x = \dfrac{۱}{۰\text{ر}۱ + x^۲}$ (د) $x = ۱ + ۰\text{ر}۳\cos(x)$

۳. با استفاده از برنامهٔ مسألهٔ ۲ (الف) کوچکترین ریشهٔ مثبت $x - \tan(x) = ۰$ و (ب) نزدیکترین ریشهٔ این معادله به $x = ۱۰۰$ را محاسبه کنید.

۴. الگوریتم نیوتن را که در بخش ۲.۲ داده شده است تکمیل کنید و آن را برای معادلات مسألهٔ ۲ به‌کار گیرید.

۵. روش نیوتن را برای محاسبهٔ ریشه‌هایی که در مسألهٔ ۳ خواسته شده به‌کار گیرید. کوشش کنید که تفاوتهای پیدا کردن ریشه‌ها را در (الف) و (ب) توضیح دهید.

۶. روش نیوتن را برای محاسبهٔ تنها ریشهٔ

$$x + e^{-Bx^۲}\cos(x) = ۰$$

با $B > ۰$ پارامتری به‌کار برید. مقادیر افزایشی متفاوتی را برای B مثلاً $B = ۱, ۵, ۱۰, ۲۵, ۵۰$، به‌کار گیرید. از بین انتخابهای استفاده شده برای $x_۰$، $x_۰ = ۰$ را انتخاب کرده و هر رفتار غیرعادی را توضیح دهید. از جنبهٔ نظری، روش نیوتن برای هر مقدار $x_۰$ و B همگراست. این موضوع را با محاسباتی که برای مقادیر بزرگتر B کرده‌اید، مقایسه کنید.

۷. یک مسألهٔ جالب ریشه‌یابی چندجمله‌ای، در محاسبهٔ قسط‌السنین پیش می‌آید. مبلغ $P_۱$ تومان در ابتدا سالهای $N_۱, ۲, ۱, \dots$ به حساب گذاشته شده است، وبهرهٔ آن در هر سال با نرخ مرکّب r (مثلاً $r = ۰\text{ر}۰۵$ یعنی با سودی برابر ۵ درصد) حساب شده است. در ابتدای سالهای

$N_1 + 1$، $N_1 + 2$، ...، $N_1 + N_2$ مبلغ P_2 تومان از حساب برداشته می‌شود پس از آخرین برداشت، حساب دقیقاً صفر می‌شود. رابطهٔ بین متغیرها به‌صورت زیر است:

$$P_1[(1+r)^{N_1} - 1] = P_2[1 - (1+r)^{-N_2}]$$

اگر $N_1 = 30$، $N_2 = 20$، $P_1 = 2000$، $P_2 = 8000$، r چقدر است؟ یک روش ریشه‌یابی به انتخاب خود به‌کار برید.

۸. روش نیوتن ـ فوریه را برای حل معادلات مسائل ۲ و ۶ به‌کار برید.

۹. روش خط قاطع را برای حل معادلات مسألهٔ ۲ به‌کار برید.

۱۰. روش خط قاطع را برای حل معادلهٔ مسألهٔ ۶ به‌کار برید.

۱۱. فرمول خطای (۲.۳.۲) را برای روش خط قاطع زیر ثابت کنید.

$$\alpha - c = -(\alpha - b)(\alpha - a)\frac{f[a,b,\alpha]}{f[a,b]}$$

۱۲. روش نیوتن را برای پیدا کردن ریشهٔ مثبت $a > 0$ در نظر می‌گیریم. با فرض $x_0 > 0$ و $x_0 \neq \sqrt{a}$، نتایج زیر را پیدا کنید.

(الف) $x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right)$.

(ب) $x_{n+1}^2 - a = \left[\frac{x_n^2 - a}{2x_n}\right]^2$، $n \geq 0$، و بنابراین $x_n > \sqrt{a}$ برای تمام مقادیر $n > 0$.

(ج) بارستهای $\{x_n\}$ یک دنبالهٔ نزولی اکید برای $n \geq 1$ است.

راهنمایی: علامت $x_{n+1} - x_n$ را در نظر بگیرید.

(د) $e_{n+1} = -e_n^2/(2x_n)$ با $e_n = \sqrt{a} - x_n$.

$\mathrm{Rel}(x_{n+1}) = \frac{-\sqrt{a}}{2x_n}[\mathrm{Rel}(x_n)]^2$، $n \geq 0$. که در آن $\mathrm{Rel}(x_n)$ خطای نسبی x_n است.

(ه) اگر $x_0 \geq \sqrt{a}$ و $|Rel(x_0)| \leq 0.1$، کران $\mathrm{Rel}(x_4)$ را پیدا کنید.

۱۳. روش نیوتن یک روش متداول در محاسبهٔ ریشه‌های دوم در رایانه است. برای استفاده از روش رایانه‌ای در محاسبهٔ $\sqrt{a}$ یک حدس اولیهٔ x_0 باید انتخاب شود و مناسبتر از همه این است که به جای آزمایش همگرایی از تعداد ثابتی بارست استفاده کنیم. برای قطعیّت، فرض می‌کنیم که حساب رایانه‌ای ما دودویی و جزء اعشاری شامل ۴۸ بیت دودویی باشد. می‌نویسیم

$$a = \hat{a}.2^e \qquad \frac{1}{2} \leq \hat{a} < 1$$

این را به‌سادگی می‌توان به شکل

$$a = b.۲^f \qquad \frac{۱}{۴} \le b < ۱$$

که در آن f یک عدد صحیح زوج است، بدل کرد. پس

$$\sqrt{a} = \sqrt{b}.۲^{f/۲} \qquad \frac{۱}{۲} \le \sqrt{b} < ۱$$

و هرگاه $\sqrt{b}$ معلوم باشد، $\sqrt{a}$ به شکل ممیّز شناور استانده خواهد بود.

این امر مسأله را به محاسبهٔ $\sqrt{b}$ برای $\frac{۱}{۴} \le b < ۱$ بدل می‌کند. فرمول درون‌یابی خطی

$$x_۰ = \frac{۱}{۳}(۲b + ۱) \qquad \frac{۱}{۴} \le b \le ۱$$

را به‌عنوان یک حدس اولیه برای بارست نیوتن در محاسبهٔ $\sqrt{b}$ به‌کار برید. کران خطا را $\sqrt{b} - x_۰$ بگیرید. تعداد بارستهای لازم را برای برقراری

$$۰ \le x_n - \sqrt{b} \le ۲^{-۴۸}$$

که حد درجهٔ دقت ماشین برای b در یک محاسبهٔ خاص است، تخمین بزنید. [توجه داشته باشید که تأثیر خطای گردکردن نادیده گرفته شده است]. چگونه می‌توان انتخاب $x_۰$ را بهبود بخشید؟

۱۴. بارستهای نیوتن را برای حل $x^۲ - ۱ = ۰$ به‌طور عددی محاسبه کنید و از $x_۰ = ۱۰۰۰۰۰$ استفاده کنید. سرعت همگرایی نتیجه را پیدا کنید و توضیح دهید.

۱۵. (الف) روش نیوتن را برای تابع

$$f(x) = \begin{cases} \sqrt{x} & x \ge ۰ \\ -\sqrt{-x} & x < ۰ \end{cases}$$

با ریشهٔ $\alpha = ۰$ به‌کار برید. رفتار بارستها چگونه است؟ آیا همگرا هستند، اگر جواب مثبت است، با چه نرخی؟

(ب) مانند (الف) عمل کنید ولی برای

$$f(x) = \begin{cases} \sqrt[۳]{x^۲} & x \ge ۰ \\ -\sqrt[۳]{x^۲} & x < ۰ \end{cases}$$

۱۶. یک دنبالهٔ $\{x_n\}$ را به α همگرای زَبَر خطی گویند اگر

$$| \alpha - x_{n+۱} | \le c_n | \alpha - x_n | \qquad n \ge ۰$$

با $c_n \to 0$ هرگاه $n \to \infty$. نشان دهید که در این حالت

$$\lim_{n\to\infty} \frac{|\alpha - x_n|}{|x_{n+1} - x_n|} = 1$$

بنابراین وقتی $n \to \infty$، اعتبار $|x_{n+1} - x_n| \doteq |\alpha - x_n|$ با n افزایش می‌یابد.

۱۷. در روش نیوتن برای پیدا کردن یک ریشهٔ $f(x) = 0$، مانند α، گاهی برای همگرایی نیاز است که حدس اولیهٔ x_0 خیلی به ریشهٔ α نزدیک باشد. ثابت کنید که برای $\alpha = \pi/2$، ریشهٔ از تابع

$$f(x) = \cos(x) + \sin^2(50x)$$

مسأله چنین است.

یک برآورد تقریبی از کوچکی $|x_0 - \alpha|$ برای همگرایی به α به‌دست آورید.

راهنمایی: (۶.۲.۲) را ملاحظه کنید.

۱۸. برنامه‌ای بنویسید که روش مولر را انجام دهد. آن را برای ریشه‌یابی درمسائل ۳،۲ و ۶ به‌کارگیرید.

۱۹. نشان دهید که $x = 1 + \tan^{-1}(x)$ دارای یک جواب α است. یک بازهٔ $[a, b]$ پیدا کنید که شامل α باشد به گونه‌ای که برای هر $x \in [a, b]$ بارستِ

$$x_{n+1} = 1 + \tan^{-1}(x_n) \qquad n \geq 0$$

به α همگرا شود. چند بارست اول را محاسبه و نرخ همگرایی را پیدا کنید.

۲۰. مانند مسألهٔ ۱۹ ولی برای بارست

$$x_{n+1} = 3 - 2\log(1 + e^{-x_n}) \qquad n \geq 0$$

عمل کنید.

۲۱. برای پیدا کردن ریشهٔ $f(x) = 0$ به روش بارستی، معادله را مجدداً به شکل زیر می‌نویسیم:

$$x = x + cf(x) \equiv g(x)$$

به‌ازای ثابت $c \neq 0$. اگر α یک ریشهٔ $f(x)$ و $f'(\alpha) \neq 0$ باشد، c چگونه باید انتخاب شود تا دنبالهٔ $x_{n+1} = g(x_n)$ به α همگرا شود؟

۲۲. معادلهٔ

$$x = d + hf(x)$$

را در نظر می‌گیریم که در آن d یک ثابت داده شده و $f(x)$ یک تابع پیوسته از x است. اگر $h = 0$، یک ریشه $\alpha = d$ است. نشان دهید که برای تمام مقادیر h به‌اندازهٔ کافی کوچک این معادله یک ریشهٔ $\alpha(h)$ دارد. با چه شرطی، اگر به شرطی نیاز باشد، یکتایی ریشه $\alpha(h)$ در بازه‌ای حول d تضمین می‌شود؟

۲۳. بارست $x_{n+1} = 2 - (1 + c)x_n + cx_n^3$ برای بعضی از مقادیر c به $\alpha = 1$ همگرا می‌شود [به شرطی که x_0 به اندازهٔ کافی نزدیک به α انتخاب شده‌باشد]. این مقادیر c را پیدا کنید. برای چه مقداری از c، همگرایی از درجهٔ دو است؟

۲۴. کدام یک از بارستهای زیر به نقطهٔ ثابت داده‌شدهٔ α همگراست. (به شرطی که x_0 به اندازهٔ کافی به α نزدیک باشد)؟ اگر بارست همگرا باشد مرتبهٔ همگرایی را پیدا کنید، در حالت همگرایی خطی، نرخ همگرایی را پیدا کنید.

(الف) $\alpha = 2 \qquad x_{n+1} = -16 + 6x_n + \dfrac{12}{x_n}$

(ب) $\alpha = 3^{1/3} \qquad x_{n+1} = \dfrac{2}{3}x_n + \dfrac{1}{x_n^2}$

(ج) $\alpha = 3 \qquad x_{n+1} = \dfrac{12}{1 + x_n}$

۲۵. نشان دهید که

$$x_{n+1} = \frac{x_n(x_n^2 + 3a)}{3x_n^2 + a} \qquad n \geq 0$$

برای محاسبهٔ $\sqrt{a}$ یک روش مرتبهٔ سه است. با فرض اینکه x_0 به اندازهٔ کافی‌به α نزدیک انتخاب شده باشد، حد زیر را محاسبه کنید.

$$\lim_{n\to\infty} \frac{\sqrt{a} - x_{n+1}}{(\sqrt{a} - x_n)^3}$$

۲۶. با استفاده از قضیهٔ ۸.۲، نشان دهید که فرمول (۱۱.۴.۲) یک روش بارستی همگرای مرتبهٔ ۳ است.

۲۷. یک فرمول بارستی به شکل زیر تعریف می‌کنیم

$$x_{n+1} = z_{n+1} - \frac{f(z_{n+1})}{f'(x_n)}, \qquad z_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

نشان دهید که مرتبهٔ همگرایی $\{x_n\}$ به α حداقل ۳ است.

راهنمایی: قضیهٔ ۲ـ۸ را به‌کار برید و بگیرید

$$g(x) = h(x) - \frac{f(h(x))}{f'(x)} \quad , \quad h(x) = x - \frac{f(x)}{f'(x)}$$

۲۸. یک اصلاح دیگری برای روش نیوتن، مشابه با روش خط قاطع، وجود دارد، که از تقریب متفاوتی برای مشتق $f'(x_n)$ در آن استفاده می‌شود. تعریف می‌کنیم

$$x_{n+1} = x_n - \frac{f(x_n)}{D(x_n)} \qquad D(x_n) = \frac{f(x_n + f(x_n)) - f(x_n)}{f(x_n)} \qquad n \geq 0$$

این روش تک‌ـ نقطه‌ای، روش استیفنسن[1] نامیده می‌شود. فرض می‌کنیم $f'(\alpha) \neq 0$، نشان دهید که این روش از مرتبهٔ دو است.

راهنمایی: بارست را به شکل $x_{n+1} = g(x_n)$ بنویسید. از $f(x) = (x-\alpha)h(x)$ با $h(\alpha) \neq 0$ استفاده کنید و این فرمول را برای $g(x)$ برحسب $h(x)$ محاسبه کنید. سپس، قضیهٔ ۲ـ۸ را به‌کار برید.

۲۹. جدول زیر مقادیر بارستهای یک بارست همگرای خطی $x_{n+1} = g(x_n)$ را نشان می‌دهد. خواسته‌های زیر را تخمین بزنید: (الف) نرخ همگرایی خطی، (ب) نقطهٔ ثابت α، و (ج) خطای $x_5 - \alpha$.

n	x_n
۰	۱٫۰۹۴۹۲۴۲
۱	۱٫۲۰۹۲۷۵۱
۲	۱٫۲۸۰۷۹۱۷
۳	۱٫۳۲۵۴۹۴۳
۴	۱٫۳۵۳۴۳۳۹
۵	۱٫۳۷۰۸۹۶۲

۳۰. می‌توان نشان داد که الگوریتم ایتکن که در بخش ۲ـ۶ داده شد از نظر سرعت همگرایی از درجهٔ دوم است. گیریم که بارست اصلی $x_{n+1} = g(x_n)$، $n \geq 0$ باشد. فرمول (۸.۶.۲) را می‌توان به شکل هم‌ارز زیر مجدداً نوشت

$$\alpha \doteq \hat{x}_{n-2} = x_{n-2} + \frac{(x_{n-1} - x_{n-2})^2}{(x_{n-1} - x_{n-2}) - (x_n - x_{n-1})} \qquad n \geq 2$$

برای آزمایش سرعت همگرایی برون‌یابهای ایتکن، دنبالهٔ متناظر زیر را در نظر می‌گیریم

$$z_{n+1} = z_n + \frac{[g(z_n) - z_n]^2}{[g(z_n) - z_n] - [g(g(z_n)) - g(z_n)]} \qquad n \geq 0$$

1. Steffenson

مقادیر z_n، مقادیر پی‌درپی $\hat{x}_n$ مسأله هستند که از الگوریتم ایتکن به دست‌آمده‌اند.

برای ۱ یا ۰ $\neq g'(\alpha)$ نشان دهید که z_n با سرعت درجهٔ دو به α می‌گراید. این درست است حتی اگر ۱ $< |g'(\alpha)|$ و بارست اصلی واگرا باشد.

راهنمایی: سعی نکنید که قضیهٔ ۸.۲ را مستقیماً به‌کار برید، زیرا خیلی پیچیده می‌شود و به جای آن بنویسید

$$g(x) = (x-\alpha)h(x) \qquad h(\alpha) = g'(\alpha) \neq 0$$

این رابطه را به‌کار برید تا نشان دهید که برای بعضی توابع $H(x)$ که حول $x = \alpha$ کراندار باشد

$$\alpha - z_{n+1} = H(z_n)(\alpha - z_n)^2$$

۳۱. دنبالهٔ زیر را در نظر می‌گیریم

$$x_n = \alpha + \beta\rho^n + \gamma\rho^{2n} \qquad n \geq 0, \qquad |\rho| < 1$$

با $\beta, \gamma \neq 0$، که با یک نرخ خطی ρ به α همگراست. گیریم $\hat{x}_{n-2}$ برون‌یاب ایتکن باشد:

$$\hat{x}_{n-2} = x_n - \frac{(x_n - x_{n-1})^2}{(x_n - x_{n-1}) - (x_{n-1} - x_{n-2})} \qquad n \geq 0$$

نشان دهید که

$$\hat{x}_{n-2} = \alpha + a\rho^{2n} + b\rho^{4n} + c_n\rho^{6n}$$

که در آن c_n وقتی $n \to \infty$، کراندار است. عباراتی برای a و b بیابید. دنبالهٔ $\{\hat{x}_n\}$ با یک نرخ خطی ρ^2 به α همگراست.

۳۲. گیریم $f(x)$ یک ریشهٔ چندگانهٔ α، مثلاً از مرتبهٔ $m > 1$ داشته باشد. نشان دهید که

$$K(x) = \frac{f(x)}{f'(x)}$$

دارای یک ریشهٔ سادهٔ α است. چرا این امر در دشواری اصلی محاسبهٔ عددی ریشه‌های چندگانه، یعنی بازهٔ بزرگ عدم اطمینان α، به ما کمکی نمی‌کند؟

۳۳. روش نیوتن را برای محاسبهٔ هر چه دقیقتر ریشه‌های چند جمله‌ییهای زیر به‌کار برید. چندگانی هر ریشه را برآورد کنید و در صورت لزوم، برای بهبود مقادیر محاسبه شدهٔ خود، راه دیگری را امتحان کنید.

(الف) $x^۴ - ۳٫۲x^۳ + ۰٫۹۶x^۲ + ۴٫۶۰۸x - ۳٫۴۵۶$

(ب) $x^۵ + ۰٫۹x^۴ - ۱٫۶۲x^۳ - ۱٫۴۵۸x^۲ + ۰٫۶۵۶۱x + ۰٫۵۹۰۴۹$

۳۴. از برنامهٔ مسألهٔ ۲ در حل معادلات زیر برای ریشهٔ $\alpha = ۱$ استفاده کنید. بازهٔ اولیهٔ $[۰, ۳]$ را به‌کار برید و در همهٔ حالات $\varepsilon = ۱۰^{-۵}$ را به‌عنوان تحملِ توقف برنامه در نظر بگیرید. نتایج را با آنچه در بخش ۲ـ۸ با روش برنت به‌دست‌آمده بود مقایسه نمایید.

(i) $(x - ۱)[۱ + (x - ۱)^۲] = ۰$

(ii) $x^۲ - ۱ = ۰$

(iii) $-۱ + x(۳ + x(-۳ + x)) = ۰$

(iv) $(x - ۱)\exp(-۱/(x - ۱)^۲) = ۰$

۳۵. با استفاده از کران بالای (۶.۹.۲) و توصیهٔ (۷.۹.۲)، کران پایین (۶.۹.۲) را ثابت کنید.

۳۶. گیریم $p(x)$ یک چندجمله‌ای از درجهٔ n باشد. گیریم ریشه‌های متمایز آن: $\alpha_۱$، ...، α_r به ترتیب از درجهٔ چندگانی $m_۱$، ...، m_r باشند،

(الف) نشان دهید که

$$\frac{p'(x)}{p(x)} = \sum_{j=۱}^{r} \frac{m_j}{x - \alpha_j}$$

(ب) گیریم c عددی باشد که برای آن $p'(c) \neq ۰$. نشان دهید که یک $p(x)$ ریشهٔ α دارد که در رابطهٔ زیر صدق می‌کند

$$|\alpha - c| \leq n \left| \frac{p(c)}{p'(c)} \right|$$

۳۷. برای چندجمله‌ای

$$p(x) = a_۰ + a_۱x + \cdots + a_nx^n \qquad a_n \neq ۰$$

تعریف می‌کنیم

$$R = \frac{|\, a_۰ \,| + |\, a_۱ \,| + \cdots + |\, a_{n-۱} \,|}{|\, a_n \,|}$$

نشان دهید که هر ریشه x از $p(x) = ۰$ در رابطه

$$|\, x \,| \leq \text{Max}\{R, \sqrt[n]{R}\}$$

صدق می‌کند.

۳۸. یک برنامهٔ رایانه‌ای بنویسید که چندجمله‌ییهای $p(x)$ زیر را برای مقادیر داده شدهٔ x محاسبه کند و نوفه را در مقادیر $p(x)$ به‌دست آورد. برای هر مقدار x، $p(x)$ را در حساب با دقّت ساده و دقّت مضاعف محاسبه نمایید؛ تفاضل آنها را به‌عنوان نوفه در حساب با دقّت ساده، بر اثر خطای گردکردن در محاسبهٔ $p(x)$ تلقی نمایید. هم فرمول معمولی (۱.۹.۲) و هم قاعدهٔ هورنر (۸.۹.۲) را برای محاسبهٔ هر چندجمله‌ای به‌کار برید. این محاسبات باید نشان دهند که نوفه در دو حالت فوق متفاوت است.

(الف) $p(x) = x^۴ - ۵٫۷x^۳ - ۰٫۴۷x^۲ + ۲۹٫۸۶۵x - ۲۶٫۱۶۰۲$؛ $-۳ \le x \le ۵$ با طول گام x برابر $h = ۰٫۱$.

(ب) $p(x) = x^۴ - ۵٫۴x^۳ + ۱۰٫۵۶x^۲ - ۸٫۹۵۴x + ۲٫۷۹۵۱$، $۱ \le x \le ۱٫۲$، با طول گام x برابر $h = ۰٫۰۰۱$.

توجه: برای کامپیوترهای مختلف ممکن است مقادیر کوچکتر یا بزرگتر h مناسب باشند. همچنین، برای مقایسهٔ معتبرتر، قبل از استفاده از حساب با دقّت مضاعف، ضرایب با دقّت ساده را وارد کنید.

۳۹. حساب مختلط و روش نیوتن را برای محاسبهٔ یک ریشهٔ مختلط

$$p(z) = z^۴ - ۳z^۳ + ۲۰z^۲ + ۴۴z + ۵۴$$

که نزدیک به $z_۰ = ۲٫۵ + ۴٫۵i$ باشد به‌کار برید.

۴۰. برنامه‌ای برای محاسبهٔ ریشه‌های چندجمله‌ییهای زیر، تا اندازهٔ ممکن دقیق، بنویسید.

(الف) $۶۷۶۰۳۹x^{۱۲} - ۱۹۳۹۹۳۸x^{۱۰} + ۲۰۷۸۵۰۵x^۸ - ۱۰۲۱۰۲۰x^۶ + -۱۸۰۱۸x^۲ + ۲۳۱$

$۲۲۵۲۲۵x^۴$

(ب) $x^۴ - ۴٫۰۹۶۱۵۲۴۲۲۷۰۶۶۳۱x^۳ + ۳٫۲۸۴۲۳۲۳۳۵۰۲۲۷۰۵x^۲$

$+۴٫۷۰۳۸۴۷۵۷۷۲۹۳۳۶۸x - ۵٫۷۱۵۷۶۷۶۶۴۹۷۷۲۹۴$

۴۱. از یک بسته برنامهٔ ریشه‌یابی برای چندجمله‌ییها در محاسبهٔ ریشه‌های چندجمله‌ییهای مسائل ۳۸، ۳۹ و ۴۰ استفاده کنید.

۴۲. در مثال $f(x) = (x-۱)(x-۲)\dots(x-۷)$، در (۲۱.۹.۲) بخش ۹.۲، اختلالی به اندازهٔ $\varepsilon_i x^i$ برای ضریب x^i ایجاد می‌کنیم که در آن ε_i به گونه‌ای انتخاب شده است که اختلال نسبی در ضریب x^i همان اختلال مثال داده شده در کتاب برای ضریب $x^۶$ است. نظریهٔ خطی‌سازی (۱۹.۹.۲) چه اختلالاتی را در ریشه‌ها پیشگویی می‌کند؟ تغییر در کدام ضریب بزرگترین تغییرها را در ریشه‌ها ایجاد می‌کند؟

۴۳. چندجمله‌ای $f(x) = x^۵ - ۳۰۰x^۲ - ۱۲۶x + ۵۰۰۵$ دارای یک ریشهٔ $\alpha = ۵$ است. اثر تغییر ضریب $x^۵$ را از ۱ به $۱ + \varepsilon$ روی ریشهٔ α برآورد نمایید.

۴۴. نتیجهٔ پایداری (۱۹.۹.۲) برای ریشه‌های چندجمله‌ای را می‌توان برای توابع دیگر تعمیم داد. گیریم α یک ریشهٔ سادهٔ $f(x) = ۰$ باشد و گیریم $f(x)$ و $g(x)$ حول α پیوسته مشتقپذیر باشند. تابع $F_\varepsilon(x) = f(x) + \varepsilon g(x)$ را تعریف می‌کنیم. گیریم $\alpha(\varepsilon)$ یک ریشهٔ $F_\varepsilon(x) = ۰$ متناظر با $\alpha = \alpha(۰)$ برای مقدار کوچک ε، باشد. برای اینکه ببینیم چنین $\alpha(\varepsilon)$ای وجود دارد و برای اثبات اینکه این تابع پیوسته مشتقپذیر است، **قضیهٔ تابع ضمنی** برای توابع یک متغیره را به‌کار می‌بریم. با استفاده از این قضیه، (۱۹.۹.۲) را برای حالت کلی تعمیم دهید.

۴۵. با استفاده از نتیجهٔ پایداری در مسألهٔ ۴۴، ریشهٔ $\alpha(\varepsilon)$ معادلهٔ

$$x - \tan(x) + \varepsilon = ۰$$

را برآورد نمایید. دو حالت را صریحاً برای ریشهٔ α از معادلهٔ $x - \tan(x) = ۰$ در نظر بگیرید؛

(۱) $\alpha \in (۰٫۵\pi, ۱٫۵\pi)$ (۲) $\alpha \in (۳۱٫۵\pi, ۳۲٫۵\pi)$.

۴۶. دستگاه زیر را در نظر می‌گیریم

$$x = \frac{۰٫۵}{۱ + (x+y)^۲} \qquad y = \frac{۰٫۵}{۱ + (x-y)^۲}$$

یک ناحیهٔ کراندار D بیابید که برای آن فرضهای قضیهٔ (۹.۲) برقرار باشند.

راهنمایی: علامت مؤلفه‌های ریشهٔ α چه خواهد بود؟ همچنین مقادیر ماکسیمم ممکن برای x و y در فرمولهای فوق چه خواهد بود؟

۴۷. دستگاه زیر را در نظر می‌گیریم

$$x = ۱ + h \cdot \frac{e^{-x^۲}}{۱ + y^۲} \qquad y = ۰٫۵ + h \cdot \tan^{-۱}(x^۲ + y^۲)$$

نشان دهید که اگر h به اندازهٔ کافی کوچک انتخاب شده باشد، این دستگاه دارای یک ریشهٔ یکتای α در ناحیهٔ مستطیلی شکل خواهد داشت. به‌علاوه، نشان دهید که یک بارست ساده به شکل (۴.۱۰.۲) به این جواب همگرا خواهد بود.

۴۸. فرع ۱۰.۲ را ثابت کنید.

راهنمایی: از پیوستگی مشتقات جزئی مؤلفه‌های $\mathbf{g}(\mathbf{x})$ استفاده کنید.

۴۹. ثابت کنید که بارستهای $\mathbf{x}_n$ در قضیهٔ ۹. به یک جواب $\mathbf{x} = \mathbf{g}(\mathbf{x})$ همگراست.

راهنمایی: مجموع نامتناهی

$$\mathbf{x}_۰ + \sum_{n=۰}^{\infty} [\mathbf{x}_{n+۱} - \mathbf{x}_n]$$

را در نظر بگیرید. حاصل‌جمعهای جزئی آن چنین‌اند

$$\mathbf{x}_0 + \sum_{n=0}^{N-1} [\mathbf{x}_{n+1} - \mathbf{x}_n] = \mathbf{x}_N$$

پس اگر این سری نامتناهی مثلاً به $\boldsymbol{\alpha}$ همگرا شود، آنگاه $\mathbf{x}_N$ به α همگرا می‌شود. با نشان دادن رابطهٔ

$$\| \mathbf{x}_{n+1} - \mathbf{x}_n \|_\infty \leq \lambda \| \mathbf{x}_n - \mathbf{x}_{n-1} \|_\infty$$

و استفاده از آن، نشان دهید که سری نامتناهی فوق همگرای مطلق است. سپس نشان دهید که $\boldsymbol{\alpha}$ یک نقطهٔ ثابت $\mathbf{g}(\mathbf{x})$ است،

۵۰. با استفاده از روش نیوتن برای دستگاههای غیرخطی، دستگاه غیرخطی

$$x^2 + y^2 = 4 \qquad x^2 - y^2 = 1$$

را حل کنید. به‌سادگی معلوم می‌شود که جوابهای حقیقی دستگاه، $(\pm\sqrt{2.5}, \pm\sqrt{1.5})$ است. به‌عنوان یک حدس اولیه، $(x_0, y_0) = (1.6, 1.2)$ را به‌کار برید.

۵۱. دستگاه

$$x^2 + xy^3 = 9 \qquad 3x^2y - y^3 = 4$$

را با استفاده از روش نیوتن برای دستگاههای غیرخطی، حل کنید. هر یک از حدسهای اولیهٔ $(x_0, y_0) = (1.2, 2.5), (-2, 2.5), (-1.2, -2.5), (2, -2.5)$ را به‌کار برید. ببینید کدام ریشه است که این روش به آن همگرا می‌شود، تعداد بارستها و سرعت همگرایی را نیز پیدا کنید.

۵۲. با استفاده از روش نیوتن برای دستگاههای غیرخطی، دستگاههای غیرخطی زیر را برای تمام ریشه‌های آنها حل کنید. از نمودارهای آنها برای تعیین حدسهای اولیه استفاده کنید.

(الف)

$$x^2 + y^2 - 2x - 2y + 1 = 0 \qquad x + y - 2xy = 0$$

(ب)

$$x^2 + 2xy + y^2 - x + y - 4 = 0$$

$$5x^2 - 6xy + 5y^2 + 16x - 16y + 12 = 0$$

۵۳. ثابت کنید که ژاکوبی

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} - F(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$$

در هر ریشهٔ α از معادلهٔ $\mathbf{f}(\mathbf{x}) = 0$ به شرطی که $F(\alpha)$ ناتکین باشد، برابر صفر است. از ترکیب این رابطه با فرع ۱۰.۲ بخش ۱۰.۲ اثبات همگرایی روش نیوتن حاصل می‌شود.

۵۴. روش نیوتن (۴.۱۲.۲) را برای پیداکردن مقدار مینیمم تابع

$$f(\mathbf{x}) = x_1^4 + x_1x_2 + (1 + x_2)^2$$

به‌کار برید. حدسهای اوّلیهٔ مختلفی را به‌کار برید و رفتار همگرایی را مشاهده کنید.

۳

نظریهٔ درونیابی

مفهوم درونیابی عبارت است از انتخاب یک تابع $p(x)$ از رده‌ای از توابع داده شده به‌نحوی که نمودار $y = p(x)$ از یک مجموعهٔ متناهی از نقاط داده شده بگذرد. در بیشتر این فصل تابع درونیاب $p(x)$ را به یک چندجمله‌یی محدود می‌کنیم.

نظریهٔ درونیابی چندجمله‌یی شماری کاربرد مهم دارد. در این کتاب، کاربرد نخست آن، تهیهٔ ابزارهای ریاضی برای ایجاد روشهایی است در زمینه‌های نظریهٔ تقریب، انتگرال‌گیری عددی و حل عددی معادلات دیفرانسیل. کاربرد دوّم آن، تهیهٔ وسایلی است برای کار با توابعی که به شکل جدول داده شده‌اند. به عنوان مثال، تقریباً همه با درونیابی خطی در یک جدول لگاریتم از جبر دبیرستانی آشنایی داریم. از جنبهٔ محاسباتی شکلهای مناسبی برای درونیابی چندجمله‌یی با داده‌های جدولی، به‌دست می‌آوریم و خطای حاصل را تحلیل خواهیم نمود. معلوم است که با استفادهٔ گسترده از ماشینهای حساب و رایانه امروزه به استفادهٔ خیلی کمتر از گذشتهٔ نزدیک به درونیابی جدولی، نیاز است. ما آن را از این جهت آورده‌ایم که فرمولهای به‌دست آمده در موارد دیگر هنوز مفیدند و نیز درونیابی جدولی، مثالها و تمرینات مناسبی در اختیار ما می‌گذارد.

این فصل را با آوردن دو موضوع دیگر خاتمه می‌دهیم. این دو موضوع عبارت‌اند از: (۱) توابع درونیاب چندجمله‌یی تکه‌یی، بویژه توابع برازا، و (۲) درونیابی با توابع مثلثاتی.

۱.۳ نظریهٔ درونیابی چندجمله‌یی

گیریم x_0، x_1، ...، x_n اعداد حقیقی یا مختلط متمایز باشند و y_0، y_1، ...، y_n مقادیر تابع متناظر آنها. اکنون به مطالعهٔ مسألهٔ پیداکردن یک چندجمله‌یی $p(x)$ می‌پردازیم که داده‌های معلوم

$$p(x_i) = y_i \qquad i = 0, 1, \ldots, n \qquad (۱.۱.۳)$$

را درونیابی می‌کند. آیا چنین چندجمله‌یی وجود دارد؟ اگر جواب مثبت است، درجهٔ آن چیست؟ آیا یکتاست؟ فرمول تولید $p(x)$ از داده‌های فوق چیست؟

چند جمله‌یی کلی درجهٔ m

$$p(x) = a_0 + a_1 x + \ldots + a_m x^m$$

$m+1$ پارامتر مستقل a_0، a_1، ...، a_m دارد. چون (۱.۱.۳)، $n+1$ شرط برای $p(x)$ پدید می‌آورد، منطقی است که ابتدا حالت $m = n$ را در نظر بگیریم. پس می‌خواهیم a_0، a_1، ...، a_n را به گونه‌ای بیابیم که داشته باشیم

$$\begin{aligned} &a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_n x_0^n = y_0 \\ &\vdots \\ &a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_n x_n^n = y_n \end{aligned} \qquad (۲.۱.۳)$$

این یک دستگاه $n+1$ معادلهٔ خطی با $n+1$ مجهول است، و حل آن کاملاً هم‌ارز با حل مسألهٔ درونیابی چندجمله‌یی است. با نمادهای برداری و ماتریسی، این دستگاه چنین نوشته می‌شود:

$$Xa = y$$

با

$$X = [x_i^j] \qquad i, j = 0, 1, \ldots, n \qquad (۳.۱.۳)$$
$$a = [a_0, a_1, \ldots, a_n]^T, y = [y_0, \ldots, y_n]^T$$

ماتریس X، ماتریس واندرموند[1] نامیده می‌شود.

قضیهٔ ۱.۳ برای $n+1$ نقطهٔ متمایز داده‌شدهٔ x_0، ...، x_n و $n+1$ عرض داده‌شدهٔ $y_0, \ldots, y_n$، یک چندجمله‌یی $p(x)$ از درجهٔ نابزرگتر از n هست که مقادیر y_i را در x_i، $i = 0, 1, \ldots, n$، درونیابی می‌کند. این چندجمله‌یی در مجموعهٔ تمام چندجمله‌یی‌های از درجهٔ حداکثر n، یکتاست.

1. Vandermonde

برهان سه برهان برای این قضیهٔ مهم داده شده است. هرکدام اطلاعات مورد نیازی را به‌دست می‌دهد و کاربردهای مهمی در مسائل دیگر درونیابی خواهد داشت.

(i) می‌توان نشان داد که برای ماتریس X در (۳.۱.۳)،

$$\det(X) = \prod_{0 \le j < i \le n} (x_i - x_j) \qquad (۴.۱.۳)$$

(مسألهٔ ۱ را ببینید). این نشان می‌دهد که $\det(X) \neq 0$، زیرا نقاط x_i متمایزند. پس X ناتکین است و دستگاه $Xa = y$ یک جواب یکتای a دارد. این، وجود و یکتایی چندجمله‌یی درونیاب از درجهٔ نابزرگتر از n را اثبات می‌کند.

(ii) به‌موجب یک قضیهٔ استانداردِ جبر خطی (قضیهٔ ۲.۷ فصل ۷)، دستگاه $Xa = y$ دارای یک جواب یکتاست اگر و فقط اگر دستگاه همگن $Xb = 0$ فقط جواب نمایان $b = 0$ داشته باشد. بنابراین، فرض می‌کنیم برای یک مقدار b، دستگاه $Xb = 0$. با استفاده از b تعریف می‌کنیم

$$p(x) = b_0 + b_1 x + \cdots + b_n x^n$$

با توجه به دستگاه $Xb = 0$، داریم

$$p(x_i) = 0 \qquad i = 0, 1, \ldots, n$$

چندجمله‌یی $p(x)$، $n+1$ صفر دارد و درجهٔ $p(x)$ نابزرگتر از n است. این ممکن نیست مگر اینکه $p(x) \equiv 0$. پس همهٔ ضرایب $b_i = 0$، $i = 0, 1, \ldots, n$، که اثبات را کامل می‌کند.

(iii) اکنون چندجمله‌یی درونیاب را به طور صریح ارائه می‌دهیم. برای شروع، مسألهٔ درونیابی خاصی را در نظر می‌گیریم که در آن برای یک مقدار i، $0 \le i \le n$،

$$y_i = 1 \qquad y_j = 0 \qquad i \neq j$$

یک چندجمله‌یی می‌خواهیم از درجهٔ نابزرگتر از n با n صفر x_j، $i \neq j$. پس

$$p(x) = c(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)$$

برای مقدار ثابتی مانند c. شرط $p(x_i) = 1$ نتیجه می‌دهد که

$$c = [(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)]^{-1}$$

این چندجمله‌یی خاص به شکل

$$l_i(x) = \prod_{j \neq i} \left(\frac{x - x_j}{x_i - x_j} \right) \qquad i = ۰, ۱, \ldots, n \qquad (۵.۱.۳)$$

نوشته می‌شود. برای حل مسألهٔ کلی درونیابی (۱.۱.۳)، می‌نویسیم

$$p(x) = y_۰ l_۰(x) + y_۱ l_۱(x) + \ldots + y_n l_n(x)$$

با استفاده از ویژگیهای خاص چند جمله‌ییهای $l_i(x)$، دیده می‌شود که تابع $p(x)$ در رابطهٔ (۱.۱.۳) صدق می‌کند. همچنین، درجهٔ $p(x)$ نابزرگتر از n است، زیرا تمام $l_i(x)$ ها از درجهٔ n هستند.

برای اثبات یکتایی، گیریم $q(x)$ چندجمله‌یی دیگری از درجهٔ نابزرگتر از n باشد که در رابطهٔ (۱.۱.۳) صدق می‌کند. تعریف می‌کنیم

$$r(x) = p(x) - q(x)$$

در این صورت درجهٔ $r(x)$ نابزرگتر از n است، و

$$r(x_i) = p(x_i) - q(x_i) = y_i - y_i = ۰ \qquad i = ۰, ۱, \ldots, n$$

چون $r(x)$ دارای $n + ۱$ صفر است، باید داشته باشیم $r(x) \equiv ۰$. این ثابت می‌کند که $p(x) \equiv q(x)$، و برهان کامل می‌شود. ■

از ویژگی یکتایی در بسیاری از مطالبی که بعداً خواهد آمد استفاده خواهیم کرد. فرمولهای دیگری برای مسألهٔ درونیابی (۱.۱.۳) به‌دست خواهیم آورد، و یکتایی می‌گوید که با همهٔ فرمولها همان یک چندجمله‌ای یکتا به‌دست می‌آید. همچنین، بدون یکتایی، امکان داشت دستگاه خطی (۲.۱.۳) به‌طور یکتا حلپذیر نباشد، این امر با استفاده از قضایای جبر خطی، وجود داده‌هایی چون بردارهای y را ایجاب می‌کرد که برای آنها هیچ چند جمله‌ای درونیابی با درجهٔ نابزرگتر از n وجود نداشته باشد.

فرمول

$$p_n(x) = \sum_{i=۰}^{n} y_i l_i(x) \qquad (۶.۱.۳)$$

فرمول لاگرانژ برای چندجمله‌یی درونیاب نامیده می‌شود.

مثال

$$p_1(x) = \frac{x - x_1}{x_0 - x_1} y_0 + \frac{x - x_0}{x_1 - x_0} y_1 = \frac{(x_1 - x)y_0 + (x - x_0)y_1}{x_1 - x_0}$$

$$p_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} y_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} y_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} y_2$$

چندجمله‌یی از درجهٔ نابزرگتر از ۲ که از سه نقطهٔ $(0, 1), (-1, 2)$ و $(3, 1)$ می‌گذرد عبارت است از

$$P_2(x) = \frac{(x+1)(x-1)}{(0+1)(0-1)} \cdot 1 + \frac{(x-0)(x-1)}{(-1-0)(-1-1)} \cdot 2 + \frac{(x-0)(x+1)}{(1-0)(1+1)} \cdot 3$$

$$= 1 + \frac{1}{2}x + \frac{3}{2}x^2$$

اگر تابع $f(x)$ داده شده باشد، می‌توانیم با استفاده از چندجمله‌یی درونیابِ

$$p_n(x; f) \equiv p_n(x) = \sum_{i=0}^{n} f(x_i) l_i(x) \qquad (7.1.3)$$

تقریبی برای آن بسازیم.

این چندجمله‌ای تابع $f(x)$ را در نقاط x_0، x_1، ... x_n درونیابی می‌کند. برای مثال، بعداً تابع $f(x) = \log_{10} x$ را با درونیابی خطی در نظر می‌گیریم. قضیهٔ اساسی که در تحلیل خطای درونیابی مورد استفاده قرار می‌گیرد قضیّهٔ زیر است. نماد $\mathscr{H}\{a, b, c, \ldots\}$ معرف کوچکترین بازهٔ شامل تمام اعداد حقیقی a، b، c، ... است.

قضیهٔ ۲.۳ گیریم x_0، x_1، ... x_n اعداد حقیقی متمایز باشند و f یک تابع حقیقی داده شده با $n+1$ مشتق پیوسته در بازهٔ $I_t = \mathscr{H}\{t, x_0, \ldots, x_n\}$ باشد که t مقدار حقیقی داده‌شده‌ای است. آنگاه عددی مانند $\xi \in I_t$ با ویژگی زیر وجود دارد:

$$f(t) - \sum_{j=0}^{n} f(x_j) l_j(t) = \frac{(t - x_0) \ldots (t - x_n)}{(n+1)!} f^{(n+1)}(\xi) \qquad (8.1.3)$$

برهان توجه کنید که قضیّه وقتی t یک نقطهٔ گرهی باشد درست است، زیرا در آن حالت هر دو طرف (۸.۱.۳) برابر صفر می‌شود. فرض کنید t نقطهٔ‌گرهی نباشد. تعریف می‌کنیم:

$$E(x) = f(x) - p_n(x) \qquad p_n(x) = \sum_{j=0}^{n} f(x_j) l_j(x)$$

$$G(x) = E(x) - \frac{\Psi(x)}{\Psi(t)} E(t) \qquad \text{به‌ازای هر } x \in I_t \qquad (9.1.3)$$

که در آن

$$\Psi(x) = (x - x_0) \dots (x - x_n)$$

تابع $G(x)$ در بازهٔ I_t، $n+1$ بار پیوسته مشتقپذیر است زیرا $E(x)$ و $\Psi(x)$ چنین‌اند. همچنین

$$G(x_i) = E(x_i) - \frac{\Psi(x_i)}{\Psi(t)} E(t) = 0 \qquad i = 0, 1, \dots, n$$
$$G(t) = E(t) - E(t) = 0$$

بنابراین G، $n+2$ صفر متمایز در I_tدارد. با استفاده از قضیهٔ مقدار میانی، G' دارای $n+1$ صفر متمایز است. استدلال استقرایی نشان می‌دهد که به ازای $j = 0, 1, \dots, n+1$ $G^{(j)}(x)$ دارای $n+2-j$ صفر در I_t، است. گیریم ξ یک صفر $G^{(n+1)}(x)$ باشد،

$$G^{(n+1)}(\xi) = 0$$

چون

$$E^{(n+1)}(x) = f^{(n+1)}(x)$$
$$\Psi^{(n+1)}(x) = (n+1)!$$

به‌دست می‌آوریم

$$G^{(n+1)}(x) = f^{(n+1)}(x) - \frac{(n+1)!}{\Psi(t)} E(t)$$

اگر به جای x، ξ بگذاریم و معادله را نسبت به $E(t)$ حل کنیم داریم،

$$E(t) = \frac{\Psi(t)}{(n+1)!} \cdot f^{(n+1)}(\xi)$$

که همان قضیهٔ مطلوب است.

این طریق پیداکردن فرمول خطا ممکن است «دشوار» به‌نظر آید، ولی این یک راه معمولی به‌دست آوردن بعضی از فرمولهای خطاست. ■

مثال برای $n = 1$ با استفاده از x به جای t

$$f(x) - \frac{(x_1 - x) f(x_0) + (x - x_0) f(x_1)}{x_1 - x_0} = \frac{(x - x_0)(x - x_1)}{2} f''(\xi_x) \quad (۱۰.۱.۳)$$

که در آن $\xi_x \in \mathscr{H}\{x_0, x_1, x\}$ زیرنمایهٔ x در ξ_x صریحاً نشان می‌دهد که ξ به x بستگی دارد؛ معمولاً زیرنمایه را برای راحتی نمی‌نویسیم.

اکنون حالت $n = ۱$ را برای تکنیک معمول دبیرستانی در درونیابی خطی در یک جدول لگاریتم به‌کار می‌بریم. گیریم

$$f(x) = \log_{۱۰} x$$

پس $f''(x) = -\log_{۱۰} e/x^۲$، $\log_{۱۰} e \doteq ۰٫۴۳۴$. در جدول معمولاً باید $x_۰ < x < x_۱$. پس

$$E(x) = \frac{(x - x_۰)(x_۱ - x)}{۲} \cdot \frac{\log_{۱۰} e}{\xi^۲} \qquad x_۰ \le \xi \le x_۱$$

و از این، کرانهای بالا و پایین به‌دست می‌آیند

$$\frac{\log_{۱۰} e}{x_۱^۲} \cdot \frac{(x - x_۰)(x_۱ - x)}{۲} \le E(x) \le \frac{\log_{۱۰} e}{x_۰^۲} \cdot \frac{(x - x_۰)(x_۱ - x)}{۲}$$

این نشان می‌دهد که تابع خطای $E(x)$ بسیار شبیه به یک چندجمله‌یی درجهٔ دوّم است، به‌ویژه اگر فاصلهٔ $h = x_۱ - x_۰$ نسبتاً کوچک باشد. برای یک کران یکنواخت در $[x_۰, x_۱]$، برای $x_۰ \ge ۱$، همان‌گونه که در جدول لگاریتم معمول است،

$$\max_{x_۰ \le x \le x_۱} (x_۱ - x)(x - x_۰) = \frac{h^۲}{۴}$$

$$| \log_{۱۰} x - p_۱(x) | \le \frac{h^۲}{۸} \frac{۰٫۴۳۴}{x_۰^۲} = \frac{۰٫۰۵۴۲h^۲}{x_۰^۲} \le ۰٫۰۵۴۲h^۲ \qquad (۱۱.۱.۳)$$

ملاحظه می‌کنیم که در جدول، برای x نزدیک به ۱۰ خطای درونیابی بسیار کوچکتر از x نزدیک به ۱ است. همچنین است خطای ماکسیمم نزدیک به نقطهٔ وسط $[x_۰, x_۱]$.

در یک جدول چهار رقمی، $h = ۰٫۰۱$

$$| \log_{۱۰} x - p_۱(x) | \le ۵٫۴۲ \times ۱۰^{-۶} \qquad ۱ \le x_۰ < x_۱ \le ۱۰$$

چون درایه‌های جدول تا چهار رقم داده شده‌اند (مثل $\log_{۱۰} ۲ = ۰٫۳۰۱۰$)، این نتیجه به اندازهٔ کافی دقیق است. پس اگر این نتیجهٔ به‌دست آمده تا این اندازه دقیق است، چرا به یک جدول دقیقتر پنج‌رقمی نیاز داریم؟ برای اینکه اثرات خطاهای گردکردن موجود در درایه‌های جدول را ندیده گرفته‌ایم. برای مثال، در $\log_{۱۰} ۲ \doteq ۰٫۳۰۱۰$.

$$| \log_{۱۰} ۲ - ۰٫۳۰۱۰ | \le ۰٫۰۰۰۰۵$$

این، بر خطای درونیابی وقتی $x_۰$ یا $x_۱$ برابر ۲ باشد غلبه دارد.

تحلیل خطای گردکردن در درونیابی خطی. گیریم

$$f(x_0) = f_0 + \epsilon_0 \qquad f(x_1) = f_1 + \epsilon_1$$

که f_0 و f_1 درایه‌های جدول و ϵ_0 و ϵ_1 خطاهای گردکردن هستند. فرض می‌کنیم برای یک ϵی معلوم

$$| \epsilon_0 |, | \epsilon_1 | \leq \epsilon$$

در مورد جدول لگاریتم چهار رقمی، $\epsilon = 0.00005$.

می‌خواهیم عبارت

$$\mathscr{E}(x) = f(x) - \frac{(x_1 - x) f_0 + (x - x_0) f_1}{x_1 - x_0} \qquad x_0 \leq x \leq x_1 \qquad (۱۲.۱.۳)$$

را کراندار کنیم. با استفاده از $f_i = f(x_i) - \epsilon_i$،

$$\begin{aligned} \mathscr{E}(x) =& f(x) - \frac{(x_1 - x) f(x_0) + (x - x_0) f(x_1)}{x_1 - x_0} \\ &+ \frac{(x_1 - x)\epsilon_0 + (x - x_0)\epsilon_1}{x_1 - x_0} \\ \equiv& E(x) + R(x) \end{aligned} \qquad (۱۳.۱.۳)$$

$$E(x) = \frac{(x - x_0)(x - x_1)}{2} f''(\xi) \qquad \xi \in [x_0, x_1]$$

خطای $\mathscr{E}(x)$ مجموع خطاهای نظری درونیابی $E(x)$ و تابع $R(x)$ است که $R(x)$ به ϵ_0 و ϵ_1 بستگی دارد. چون $R(x)$ یک خط مستقیم است، ماکسیمم آن روی $[x_0, x_1]$ در یکی از دو نقطهٔ انتهایی قرار می‌گیرد،

$$\underset{x_0 \leq x \leq x_1}{\text{Max}} | R(x) | = \text{Max}\{| \epsilon_0 |, | \epsilon_1 |\} \leq \epsilon \qquad (۱۴.۱.۳)$$

با $x_1 = x_0 + h$، $x_0 \leq x \leq x_1$ داریم

$$| \mathscr{E}(x) | \leq \frac{h^2}{8} \underset{x_0 \leq t \leq x_1}{\text{Max}} | f''(t) | + \text{Max}\{| \epsilon_0 |, | \epsilon_1 |\} \qquad (۱۵.۱.۳)$$

مثال در مثال لگاریتم قبلی، که از جدول چهار رقمی استفاده شده است

$$| \mathscr{E}(x) | \leq 5.42 \times 10^{-6} + 5 \times 10^{-5} \doteq 5.5 \times 10^{-5}$$

در یک جدول پنج رقمی، $h = ۰٫۰۰۱$، $\epsilon = ۰٫۰۰۰۰۰۵$ و

$$| \mathscr{E}(x) | \leq ۵٫۴۲ \times ۱۰^{-۸} + ۵ \times ۱۰^{-۶} \doteq ۵٫۰۵ \times ۱۰^{-۶} \qquad x_۰ \leq x \leq x_۱$$

تنها خطای مهم در استفاده از یک جدول پنج رقمی خطای گردکردن است. درواقع، به‌نظر مفید می‌آید که جدول پنج‌رقمی را به جدول شش‌رقمی تبدیل کنیم بدون اینکه اندازهٔ فاصلهٔ h تغییر داده شود. در این صورت یک ماکسیمم خطا برابر $۵٫۵ \times ۱۰^{-۷}$ برای $\mathscr{E}(x)$ خواهیم داشت بدون اینکه محاسبات افزایش چشمگیری داشته باشند. این دلایل در مورد خطای گردکردن برای درونیابی چند جمله‌ای از درجهٔ بالاتر نیز عمومیت می‌یابد اگرچه نتیجهٔ آن برای $\text{Max} \mid R(x) \mid$ کمی پیچیده‌تر می‌شود (مسألهٔ ۸ را ببینید).

توجه کنید که در هیچ یک از نتایج این بخش خطاهای گردکردن جدیدی که در محاسبهٔ $p_n(x)$ رخ می‌دهند به حساب نیامده است. این خطاها به کمک قضایایی که در بخش بعد می‌آیند، مینیمم می‌شوند.

۲.۳ تفاضلات منقسم نیوتن

شکل لاگرانژی چندجمله‌ای درونیابی را می‌توان برای درونیابی یک تابع که به شکل جدولی داده شده است به‌کار برد؛ از جداول موجود در آبراموویتس و استگون[1] (۱۹۶۴، فصل ۲۵) $l_i(x)$ آسانتر محاسبه می‌شود. ولی شکلهای دیگری وجود دارند که بسیار مناسبترند و در این بخش و بخش بعدی بحث می‌شوند. با شکل لاگرانژی، مشکل می‌توان از یک چندجمله‌ای درونیابی به یک چندجمله‌ای دیگر با یک درجهٔ بیشتر رسید. مقایسهٔ چندجمله‌ییهای درونیابی با درجات مختلف، برای انتخاب درجهٔ چندجمله‌ای که باید به‌کار رود تکنیک مفیدی است. فرمولهایی که در این بخش مطالعه می‌شوند برای نقاط شبکه‌یی نامتساوی‌الفاصلهٔ $\{x_i\}$ هستند. بدین لحاظ این فرمولها، برای درونیابی معکوس در جدول مناسب‌اند، نکته‌ای است که بعداً نشان خواهیم داد. این فرمولها، در بخش ۳.۳ برای نقاط شبکه‌یی متساوی‌الفاصله تخصیص یافته‌اند.

می‌نویسیم:

$$p_n(x) = p_{n-۱} + C(x) \qquad C(x) = \text{جملهٔ تصحیحی} \qquad (۱.۲.۳)$$

پس، در حالت کلی، $C(x)$ یک چندجمله‌ای از درجهٔ n است، زیرا معمولاً درجهٔ $p_{n-۱}$ برابر

1. Abramovitz and Stegun

$n-۱$ است و درجهٔ p_n مساوی n. همچنین داریم:

$$C(x_i) = p_n(x_i) - p_{n-۱}(x_i) = f(x_i) - f(x_i) = ۰ \qquad i = ۰, ۱, \dots n-۱$$

بنابراین

$$C(x) = a_n(x - x_۰) \dots (x - x_{n-۱})$$

چون $p_n(x_n) = f(x_n)$، از رابطهٔ (۱.۲.۳) به‌دست می‌آوریم:

$$a_n = \frac{f(x_n) - p_{n-۱}(x_n)}{(x_n - x_۰) \dots (x_n - x_{n-۱})}$$

به دلایلی که در زیر آمده‌است، این ضریب a_n را **تفاضل منقسم مرتبهٔ n ام نیوتن** مربوط به f می‌نامند، و آن را با نماد زیر نشان می‌دهند.

$$a_n \equiv f[x_۰, x_۱, \dots, x_n]$$

پس فرمول درونیابی ما چنین می‌شود

$$p_n(x) = p_{n-۱}(x) + (x - x_۰) \dots (x - x_{n-۱}) f[x_۰, x_۱ \dots, x_n] \qquad (۲.۲.۳)$$

برای آنکه اطلاعات بیشتری در مورد a_n به‌دست آوریم، به فرمول (۷.۱.۳) لاگرانژ برای $p_n(x)$ برمی‌گردیم. می‌نویسیم

$$\Psi_n(x) = (x - x_۰) \dots (x - x_n) \qquad (۳.۲.۳)$$

بنابراین

$$\Psi'_n(x_i) = (x_i - x_۰) \dots (x_i - x_{i-۱})(x_i - x_{i+۱}) \dots (x_i - x_n)$$

و اگر x یک نقطهٔ گرهی نباشد

$$p_n(x) = \sum_{j=۰}^{n} \frac{\Psi_n(x)}{(x - x_j)\Psi'_n(x_j)} \cdot f(x_j) \qquad (۴.۲.۳)$$

از آنجا که a_n ضریب x^n در $p_n(x)$ است، فرمول لاگرانژ را برای به‌دست آوردن ضریب x^n به‌کار می‌بریم. با نگاه‌کردن به هر جملهٔ درجهٔ n ام در فرمول (۴.۲.۳)، به‌دست می‌آوریم

$$f[x_۰, x_۱, \dots, x_n] = \sum_{j=۰}^{n} \frac{f(x_j)}{\Psi'_n(x_j)} \qquad (۵.۲.۳)$$

از این فرمول، یک ویژگی مهم برای تفاضل منقسم به‌دست می‌آوریم. گیریم $(i_0, i_1, \ldots, i_n)$ یک جایگشت $(0, 1, \ldots, n)$ باشد. به‌سادگی به‌دست می‌آوریم:

$$\sum_{j=0}^{n} \frac{f(x_j)}{\Psi'_n(x_j)} = \sum_{j=0}^{n} \frac{f(x_{i_j})}{\Psi'_n(x_{i_j})}$$

زیرا مجموع دوّم، تنها یک جابه‌جایی در مجموع اول است. از طرف دیگر برای هر جایگشت $(0, 1, \ldots, n)$، مانند $(i_0, \ldots, i_n)$،

$$f[x_0, x_1, \ldots, x_n] = f[x_{i_0}, x_{i_1}, \ldots, x_{i_n}] \qquad (۶.۲.۳)$$

یک فرمول مفید دیگر برای محاسبهٔ $f[x_0, x_1, \ldots, x_n]$ فرمول زیر است:

$$f[x_0, x_1, \ldots, x_n] = \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0} \qquad (۷.۲.۳)$$

که نامگذاری تفاضل منقسم را نیز توجیه می‌کند. این نتیجه را می‌توان از فرمول (۵.۲.۳) یا فرمول دیگری که ذیلاً برای $p_n(x)$ بیان می‌شود، ثابت کرد:

$$p_n(x) = \frac{(x_n - x)p_{n-1}^{(0,n-1)}(x) + (x - x_0)p_{n-1}^{(1,n)}(x)}{x_n - x_0} \qquad (۸.۲.۳)$$

که در آن $p_{n-1}^{(0,n-1)}(x)$ چندجمله‌ای از درجهٔ نابزرگتر از $n-1$ است که $f(x)$ را در $\{x_0, \ldots, x_{n-1}\}$ درونیابی می‌کند و $p_{n-1}^{(1,n)}(x)$ چندجمله‌ای است که $f(x)$ را در $\{x_1, \ldots, x_n\}$ درونیابی می‌کند. اثبات فرمولهای (۷.۲.۳) و (۸.۲.۳) در مسألهٔ ۱۳ آمده است.

با بازگشت به فرمول (۲.۲.۳)، فرمولهای زیر را خواهیم داشت:

$$\begin{aligned}
p_0(x) &= f(x_0) \\
p_1(x) &= f(x_0) + (x - x_0)f[x_0, x_1] \\
&\vdots \\
p_n(x) &= f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \\
&\quad + \ldots + (x - x_0)\ldots(x - x_{n-1})f[x_0, x_1, \ldots, x_n]
\end{aligned} \qquad (۹.۲.۳)$$

جدول ۱.۳ قالب برای ساختن تفاضلات منقسم $f(x)$

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}] \dots$
x_0	f_0		
		$f[x_0, x_1]$	
x_1	f_1		$f[x_0, x_1, x_2] \dots$
		$f[x_1, x_2]$	
x_2	f_2		$f[x_1, x_2, x_3] \dots$
		$f[x_2, x_3]$	
x_3	f_3		$f[x_2, x_3, x_4] \dots$
		$f[x_3, x_4]$	
x_4	f_4		$f[x_3, x_4, x_5] \dots$
		$f[x_4, x_5]$	
x_5	f_5		
$\vdots$	$\vdots$	$\vdots$	$\vdots$

این فرمول را **فرمول تفاضل منقسم نیوتن برای چندجمله‌ای درونیاب** می‌نامند. این فرمول برای محاسبات خیلی بهتر از فرمول لاگرانژ است (اگرچه گونه‌های دیگری از فرمول لاگرانژ وجود دارند که از این فرمول نیوتن کاراترند.)

برای ساختن تفاضلات منقسم، قالبی راکه در جدول ۱.۳ داده شده به‌کار می‌گیریم. صورت هرکسر تفاضل، از تفاضل دو درایهٔ مجاور در ستونهای سمت چپ ستونی که دارید می‌سازید به‌دست می‌آید.

مثال یک جدول تفاضل منقسم برای $f(x) = \sqrt{x}$ می‌سازیم که در جدول ۲.۳ نشان داده

جدول ۲.۳ مثالی برای ساختن تفاضلات منقسم

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$D^2 f[x_i]$	$D^3 f[x_i]$	$D^4 f[x_i]$
۲٫۰	۱٫۴۱۴۲۱۴				
		۰٫۳۴۹۲۴			
۲٫۱	۱٫۴۴۹۱۳۸		−۰٫۰۴۱۱۰		
		۰٫۳۴۱۰۲		۰٫۰۰۹۱۶۷	
۲٫۲	۱٫۴۸۳۲۴۰		−۰٫۰۳۸۳۵		−۰٫۰۰۲۰۸۴
		۰٫۳۳۳۳۵		۰٫۰۰۸۳۳۳	
۲٫۳	۱٫۵۱۶۵۷۵		−۰٫۰۳۵۸۵		
		۰٫۳۲۶۱۸			
۲٫۴	۱٫۵۴۹۱۹۳				

شده است. از نماد $D^r f(x_i) = f[x_i, x_{i+1}, \ldots, x_{i+r}]$ استفاده کرده‌ایم. توجه داشته باشید که درایه‌های جدول در ستون $f(x_i)$ در رقم هفتم خطای گردکردن دارند و این امر بر دقت تفاضلات منقسم حاصل اثر می‌گذارد. بحثی دربارهٔ نتایج خطای گردکردن در محاسبهٔ $p_n(x)$ برای هر دو فرمول لاگرانژ و تفاضل منقسم نیوتن، در پاول (۱۹۸۱، ص ۵۱) داده شده است.

یک الگوریتم ساده برای ساختن تفاضلات

$$f(x_0), f[x_0, x_1], f[x_0, x_1, x_2], \ldots, f[x_0, x_1, \ldots, x_n]$$

که در محاسبهٔ فرمول (۹.۲.۳) نیوتن مورد نیازند، می‌توان ارائه کرد.

الگوریتم $Divdif(d, x, n)$

۱. **توجه:** d و x بردارهایی با درایه‌های به ترتیب، $f(x_i)$ و x_i، $i = 0, 1, \ldots, n$، هستند. هنگام خروج، d_i شامل $f[x_0, \ldots, x_i]$ است.

۲. مرحلهٔ ۴ را برای $i = 1, 2, \ldots, n$ عمل کنید.

۳. مرحلهٔ ۴ را برای $j = n, n-1, \ldots, i$ عمل کنید.

۴. $d_j := (d_j - d_{j-1})/(x_j - x_{j-i})$

۵. از الگوریتم خارج شوید.

برای محاسبهٔ فرمول نیوتن چندجمله‌ای درونیاب (۹.۲.۳)، ضرب تودرتوی چندجمله‌ای (۸.۹.۲) از فصل ۲ را با تغییری ساده می‌آوریم.

الگوریتم Interp (d, x, n, t, p)

۱. **توجه:** درورودی، d و x به ترتیب بردارهایی شامل $f[x_0, \ldots, x_i]$ و x_i، $i = 0, 1, \ldots, n$، هستند. هنگام خروج، p شامل مقدار $p_n(t)$، چندجمله‌ای درجهٔ n است که f را در x درونیابی می‌کند.

۲. $p := d_n$

۳. از مرحلهٔ ۴ برای $i = n-1, n-2, \ldots, 0$ عمل کنید.

۴. $p := d_i + (t - x_i)p$

۵. از الگوریتم خارج شوید.

جدول ۳.۳ مثالی در استفاده از فرمول نیوتن

x	$p_1(x)$	$p_2(x)$	$p_3(x)$	$p_4(x)$	$\sqrt{x}$
۲٫۰۵	۱٫۴۳۱۶۷۶	۱٫۴۳۱۷۷۹	۱٫۴۳۱۷۸۲	۱٫۴۳۱۷۸۲	۱٫۴۳۱۷۸۲
۲٫۱۵	۱٫۴۶۶۶۰۰	۱٫۴۶۶۲۹۲	۱٫۴۶۶۲۸۸	۱٫۴۶۶۲۸۸	۱٫۴۶۶۲۸۸
۲٫۴۵	۱٫۵۷۱۳۷۲	۱٫۵۶۴۸۹۹	۱٫۵۶۵۲۶۰	۱٫۵۶۵۲۴۷	۱٫۵۶۵۲۴۸

مثال برای $f(x) = \sqrt{x}$، در جدول ۳.۳، مقادیر $p_n(x)$ برای مقادیر مختلفِ x و n داده شده‌اند. $p_4(x)$، چندجمله‌ای بالاترین درجه، از مقادیر تابع در نقاط شبکه از $x_0 = 2{,}0$ تا $x_4 = 2{,}4$ استفاده می‌کند. تفاضلات منقسم لازم در مثال آخر، در جدول ۲.۳ داده شده‌اند.

وقتی مقدار x خارج از $\mathscr{H}\{x_0, x_1, \ldots, x_n\}$ می‌افتد، غالباً می‌گوییم که $p_n(x)$ تابع $f(x)$ را برون‌یابی می‌کند. در مثال اخیر، به بی‌دقّتی زیاد در مقدار برون‌یابی $p_3(2{,}45)$ در مقایسه با مقادیر $p_3(2{,}05), p_3(2{,}15)$، توجه نمایید. ولی در این کتاب، واژهٔ **درونیابی** همواره، امکان بیرون افتادن x از بازهٔ $\mathscr{H}\{x_0, x_1, \ldots, x_n\}$ را نیز شامل می‌شود.

اغلب مقدار تابع $f(x)$ را داریم، و می‌خواهیم x مقدار متناظر آن را به‌دست آوریم. این را **درونیابی وارون** می‌نامند. معمولاً استفاده‌کنندگان از جداول لگاریتم برای محاسبهٔ آنتی‌لگاریتم یک عدد x، آن را به‌کار می‌برند. برای محاسبهٔ x، آن را مانند یک متغیر وابسته و $y = f(x)$ را یک متغیر مستقل می‌گیریم. اگر جدولی از مقادیر (x_i, y_i)، $i = 0, 1, \ldots, n$، داده شده باشد، یک چندجمله‌ای $p_n(y)$ می‌سازیم که مقادیر x_i را در y_i، $i = 0, 1, \ldots, n$، درونیابی کند. در واقع، تابع معکوس $g(y) \equiv f^{-1}(y)$ را درونیابی می‌کنیم؛ فرمول خطای قضیهٔ ۲.۳ با $x = f^{-1}(y)$ به شکل زیر در می‌آید:

$$x - p_n(y) = \frac{(y - y_0)\ldots(y - y_n)}{(n+1)!} g^{(n+1)}(\xi) \qquad (۱۰.۲.۳)$$

به ازای مقداری از $\xi \in \mathscr{H}\{y, y_0, y_1, \ldots, y_n\}$. چنانچه لازم باشد، مشتقات $g(y)$ را می‌توان با مشتق‌گیری از فرمول مرکب $g(f(x)) = x$ محاسبه کرد.

برای مثال:

$$g'(y) = \frac{1}{f'(x)} \qquad \text{برای } y = f(x)،$$

مثال مقادیر تابع بسل $J_0(x)$ را که از آبراموویتس و استگون، ۱۹۶۴، فصل ۹ استخراج شده در جدول ۴.۳ در نظر می‌گیریم. مقدار x را که به‌ازای آن $J_0(x) = 0{,}1$، محاسبه می‌کنیم. جدول ۵.۳، مقادیر $p_n(y)$ را برای $n = 0, 1, \ldots, 6$ با $x_0 = 2{,}0$ به‌دست می‌دهد.

جدول ۴.۳ مقادیر تابع بسل $J_0(x)$

x	$J_0(x)$
۲٫۰	۰٫۲۲۳۸۹۰۷۷۹۱
۲٫۱	۰٫۱۶۶۶۰۶۹۸۰۳
۲٫۲	۰٫۱۱۰۳۶۲۲۶۶۹
۲٫۳	۰٫۰۵۵۳۹۷۸۴۴
۲٫۴	۰٫۰۰۲۵۰۷۶۸۳۲
۲٫۵	−۰٫۰۴۸۳۸۳۷۷۶۴
۲٫۶	−۰٫۰۹۶۸۰۴۹۵۴۴
۲٫۷	−۰٫۱۴۲۴۴۹۳۷۰۰
۲٫۸	−۰٫۱۸۵۰۳۶۰۳۳۴
۲٫۹	−۰٫۲۲۴۳۱۱۵۴۵۸

چندجمله‌ای $p_n(y)$ که تابع وارون $J_0(x)$ را درونیابی می‌کند، $g(y)$ می‌نامیم. جوابِ $x =$ ۲٫۲۱۸۶۸۳۸ تا هشت رقم معنی‌دار درست است.

یک فرمول خطای درونیابی با استفاده از تفاضلات منقسم. گیریم t یک عدد حقیقی متمایز از نقاط گرهی x_0، x_1، ...، x_n باشد. چندجمله‌ای $p_{n+1}(x)$ که $f(x)$ را در نقاط x_0، x_1، ...، x_n و t درونیابی می‌کند می‌سازیم.

$$\begin{aligned} p_{n+1}(x) =& f(x_0) + (x - x_0) f[x_0, x_1] + \dots \\ &+ (x - x_0) \dots (x - x_{n-1}) f[x_0, \dots, x_n] \\ &+ (x - x_0) \dots (x - x_n) f[x_0, x_1, \dots, x_n, t] \\ =& p_n(x) + (x - x_0) \dots (x - x_n) f[x_0, \dots, x_n, t] \end{aligned}$$

جدول ۵.۳ مثال از درونیابی وارون

n	$p_n(y)$	$p_n(y) - p_{n-1}(y)$	$g[y_0, \dots, y_n]$
۰	۲٫۰		۲٫۰
۱	۲٫۲۱۶۲۷۵۴۲۵	۲٫۱۶E − ۱	−۱٫۷۴۵۶۹۴۲۸۲
۲	۲٫۲۱۸۶۱۹۶۰۸	۲٫۳۴E − ۳	۰٫۲۸۴۰۷۴۸۴۰۵
۳	۲٫۲۱۸۶۸۶۲۵۲	۶٫۶۶E − ۵	−۰٫۷۷۹۳۷۱۱۸۱۲
۴	۲٫۲۱۸۶۸۳۳۴۴	−۲٫۹۱E − ۶	۰٫۷۶۴۸۹۸۶۷۰۴
۵	۲٫۲۱۸۶۸۳۹۶۴	۶٫۲۰E − ۷	−۱٫۶۷۲۳۵۷۲۶۴
۶	۲٫۲۱۸۶۸۳۷۷۳	−۱٫۹۱E − ۷	۳٫۴۷۷۳۳۳۱۲۶

چون $p_{n+1}(t) = f(t)$، می‌گیریم $x = t$ و به‌دست می‌آوریم:

$$f(t) - p_n(t) = (t - x_0)\dots(t - x_n)f[x_0, \dots, x_n, t] \quad (۱۱.۲.۳)$$

که در آن $p_n(t)$ به سمت چپ معادله برده شده است. این رابطه فرمول دیگری را برای خطای $f(t) - p_n(t)$ به ما می‌دهد، فرمولی که در موارد بسیاری مفید است. اگر این فرمول را با فرمول خطای قبلی (۸.۱.۳) مقایسه و $\Psi_n(t)$، ضریب مشترک را حذف کنیم، برای مقداری مانند $\xi \in \mathscr{H}\{x_0, x_1, \dots, x_n, t\}$ داریم

$$f[x_0, x_1, \dots, x_n, t] = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

برای این که نتیجهٔ فوق را نسبت به شناسه‌های آن متقارن سازیم، معمولاً می‌گیریم $t = x_{n+1}$، $n = m - 1$ و فرمول زیر را برای یک مقدار $\xi \in \mathscr{H}\{x_0, \dots, x_m\}$ به‌دست می‌آوریم:

$$f[x_0, x_1, \dots, x_m] = \frac{f^{(m)}(\xi)}{m!} \quad (۱۲.۲.۳)$$

با این نتیجه، فرمول (۹.۲.۳) نیوتن، مانند سِری بریده شدهٔ تیلر $f(x)$ حول x_0 به‌نظر می‌آید، به شرطی که $x_n - x_0$ خیلی بزرگ نباشد.

مثال از جدول تفاضلات منقسم ۲.۳ برای تابع $f(x) = \sqrt{x}$، داریم:

$$f[۲٫۰, ۲٫۱, \dots, ۲٫۴] = -۰٫۰۰۲۰۸۴$$

چون $f^{(۴)}(x) = -۱۵/(۱۶x^{۳}\sqrt{x})$، به‌سادگی می‌توان نشان داد که

$$\frac{f^{(۴)}(۲٫۳۱۰۳)}{۴!} \doteq -۰٫۰۰۲۰۸۴$$

پس در این حالت در (۱۲.۲.۳)، داریم $\xi \doteq ۲٫۳۱$.

مثال اگر فرمول (۱۲.۲.۳) برای براورد مشتقات $g(y)$، تابع وارون $J_0(x)$، که در یکی از مثالهای قبل داده شد، استفاده شود، مشتقات $g^{(n)}(y)$ سریعاً با n افزایش می‌یابد. برای مثال،

$$g[y_0, \dots, y_۶] \doteq ۳٫۴۸$$
$$g^{(۶)}(\xi) \doteq (۶!)(۳٫۴۸) \doteq ۲۵۰۰$$

به ازای یک ξ در $[۰٫۲۲۳۹, ۰٫۰۹۶۸-]$. برآوردهای مشابهی را می‌توان برای مشتقات دیگر انجام داد.

برای تعمیم تعریف تفاضل منقسم نیوتن، در حالتی که بعضی یا تمام نقاط گره برهم منطبق می‌شوند، فرمول دیگری را برای تفاضل منقسم معرفی می‌کنیم.

قضیهٔ ۳.۳ ارمیت ـ جنوکی[1] گیریم $x_۰$، $x_۱$، ...، x_n متمایز باشند و $f(x)$ در بازهٔ $\mathscr{H}\{x_۰, x_۱, \ldots, x_n\}$ n بار پیوسته مشتقپذیر باشد. آنگاه

$$f[x_۰, x_۱, \ldots, x_n] = \int_{\tau_n} \ldots \int f^{(n)}(t_۰x_۰ + \ldots + t_nx_n)dt_۱ \ldots dt_n \qquad (۱۳.۲.۳)$$

که در آن

$$\tau_n = \{(t_۱, t_۲, \ldots, t_n) \mid \forall t_i \geq ۰, \sum_{۱}^{n} t_i \leq ۱\} \qquad (۱۴.۲.۳)$$

$$t_۰ = ۱ - \sum_{۱}^{n} t_i$$

توجه داشته باشید که $t_۰ \geq ۰$ و $\sum_{۰}^{n} t_i = ۱$.

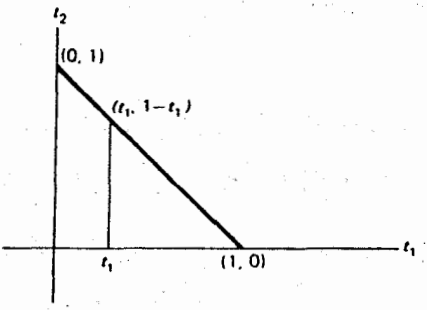
برهان نشان می‌دهیم که (۱۳.۲.۳) به ازای $n = ۱, ۲$ صحیح است و از این دو حالت، برهان استقرایی در حالت کلی روشن می‌شود.

۱. $n = ۱$. آنگاه $\tau_۱ = [۰, ۱]$

$$\begin{aligned}\int_{۰}^{۱} f'(t_۰x_۰ + t_۱x_۱)dt_۱ &= \int_{۰}^{۱} f'(x_۰ + t_۱(x_۱ - x_۰))dt_۱ \\ &= \frac{۱}{x_۱ - x_۰} f(x_۰ + t_۱(x_۱ - x_۰))\Big|_{t_۱=۰}^{t_۱=۱} \\ &= \frac{f(x_۱) - f(x_۰)}{x_۱ - x_۰} = f[x_۰, x_۱]\end{aligned}$$

۲. $n = ۲$. آنگاه $\tau_۲$ مثلثی است با رأسهای $(۰, ۰)$، $(۰, ۱)$ و $(۱, ۰)$ که در شکل ۱.۳ نشان داده شده است.

1. Hermite-Gennochi



شکل ۱.۳ ناحیه τ_2

$$\begin{aligned}
&\int_{\tau_n}\int f''(t_0x_0+t_1x_1+t_2x_2)dt_1dt_2\\
&\quad=\int_0^1\int_0^{1-t_1} f''(x_0+t_1(x_1-x_0)+t_2(x_2-x_0))dt_2dt_1\\
&\quad=\int_0^1 \frac{1}{x_2-x_0}\left[f'(x_0+t_1(x_1-x_0)+t_2(x_2-x_0))\right]_{t_2=0}^{t_2=1-t_1} dt_1\\
&\quad=\frac{1}{x_2-x_0}\left[\int_0^1 f'(x_2+t_1(x_1-x_2))dt_1\right.\\
&\qquad\qquad\left.-\int_0^1 f'(x_0+t_1(x_1-x_0))dt_1\right]\\
&\quad=\frac{1}{x_2-x_0}\{f[x_1,x_2]-f[x_0,x_1]\}=f[x_0,x_1,x_2]
\end{aligned}$$

با روندی مشابه برای حالت کلی عمل می‌کنیم. با یک بار انتگرالگیری مسأله را به یک بُعد کمتر کاهش می‌دهیم. سپس به فرضهای استقراء استناد نموده (۷.۲.۳) را به‌کار می‌بریم تا برهان کامل شود. ■

اکنون می‌توانیم (۱۳.۲.۳)، را در $f[x_0,x_1,\ldots,x_n]$ به‌کار بریم. می‌بینیم که اگر $f(x)$ در $\mathscr{H}\{x_0,\ldots,x_n\}$، n بار پیوسته مشتقپذیر باشد، آنگاه $f[x_0,\ldots,x_n]$ تابعی است پیوسته از n متغیر x_0، x_1، ...، x_n، خواه این متغیرها متمایز باشند یا نباشند. به عنوان مثال، اگر فرض کنیم که تمام نقاط به x_0 بدل شوند، آنگاه برای تفاضل منقسم مرتبهٔ n،

$$\begin{aligned}
f[x_0,\ldots,x_0]&=\int_{\tau_n}\cdots\int f^{(n)}(x_0)dt_1\ldots dt_n\\
&=f^{(n)}(x_0)\cdot \mathrm{Vol}(\tau_n)
\end{aligned}$$

که در آن $\mathrm{Vol}(\tau_n)$ حجم τ_n است. با توجه به مسألهٔ ۱۵ داریم $\mathrm{Vol}(\tau_n) = 1/n!$ و بنابراین

$$f[x_0, \ldots, x_0] = \frac{f^{(n)}(x_0)}{n!} \qquad (۱۵.۲.۳)$$

این رابطه را می‌توانستیم مستقیماً از رابطهٔ (۱۲.۲.۳) پیشبینی کنیم. ولی اگر فقط بعضی از نقاط گرهی یکی می‌شدند، باید برای توجیه وجود $f[x_0, \ldots, x_n]$ از رابطه (۱۳.۲.۳) استفاده می‌کردیم.

درکاربردهای انتگرالگیری عددی، باید بدانیم که آیا

$$\frac{d}{dx} f[x_0, \ldots, x_n, x] \qquad (۱۶.۲.۳)$$

وجود دارد یا نه. اگر f، $n+2$ بار پیوسته مشتقپذیر باشد، آنگاه می‌توانیم قضیهٔ ۳.۳ را بکار بریم. با استفاده از قضایای مشتقگیری و انتگرالگیری نسبت به یک پارامتر انتگرالده می‌توانیم وجود (۱۶.۲.۳) را نتیجه بگیریم. از راه مستقیم‌تر

$$\begin{aligned}
\frac{d}{dx} f[x_0, \ldots, x_n, x] &= \lim_{h \to 0} \frac{f[x_0, \ldots, x_n, x+h] - f[x_0, \ldots, x_n, x]}{h} \\
&= \lim_{h \to 0} \frac{f[x_0, \ldots, x_n, x+h] - f[x, x_0, \ldots, x_n]}{h} \\
&= \lim_{h \to 0} f[x, x_0, \ldots, x_n, x+h] = f[x, x_0, x_1, \ldots, x_n, x]
\end{aligned}$$

$$\frac{d}{dx} f[x_0, x_1, \ldots, x_n, x] = f[x_0, x_1, \ldots, x_n, x, x] \qquad (۱۷.۲.۳)$$

وجود و پیوستگی طرف راست، با استفاده از (۱۳.۲.۳) تضمین می‌شود.

نظریه‌ای پربار شامل درونیابی چندجمله‌ای و تفاضلات منقسم وجود دارد، ولی ما در اینجا موضوع را با یک قضیّه سر راست پایانی خاتمه می‌دهیم. اگر $f(x)$ یک چندجمله‌ای از درجهٔ m باشد، آنگاه

$$f[x_0, \ldots, x_n, x] = \begin{cases} \text{چندجمله‌ای از درجهٔ } m-n-1 & n < m-1 \\ a_m & n = m-1 \\ 0 & n > m-1 \end{cases} \qquad (۱۸.۲.۳)$$

که در آن، $f(x) = a_m x^m +$ (جملات از درجهٔ پایینتر). برای برهان قضیّه به مسألهٔ ۱۴ مراجعه کنید.

۳.۳ تفاضلات متناهی و جدول جهتدار فرمولهای درونیابی

در این بخش، تفاضلات پیشرو و پسرو را معرفی می‌کنیم، همراه با فرمولهای درونیابیی که از آنها استفاده می‌شود. این تفاضلات را کلا **تفاضلات متناهی** می‌نامند و آنها را در ساختن فرمولهای درونیابی برای جداولی که طولها، یعنی $\{x_i\}$ ها متساوی‌الفاصله‌اند، به‌کار می‌برند. این فرمولهای درونیابی در حل عددی معادلات دیفرانسیل معمولی و جزیی نیز کاربرد دارند. بعلاوه، از تفاضلات متناهی برای مسألهٔ تعیین بیشترین درجهٔ چندجمله‌ای درونیابی، که بتوان با اطمینان به‌کاربرد، بر پایهٔ دقتی که درایه‌های جدول دارند، استفاده می‌شود. و تفاضلات متناهی را برای کشف نوفه در داده‌ها، وقتی نوفه نسبت به خطای گردکردن یا خطای عدم قطعیّت اندازه‌گیری فیزیکی زیاد باشد، می‌توان به‌کار برد. این موضوع در بخش ۴.۳ بحث شده است.

برای یک $h > 0$ مفروض، تعریف می‌کنیم

$$\Delta_h f(z) = f(z+h) - f(z)$$

معمولاً h از قراین مطلب معلوم می‌شود، و می‌نویسیم

$$\Delta f(z) = f(z+h) - f(z) \qquad (۱.۳.۳)$$

این مقدار را **تفاضل پیشرو** f در z، و Δ را **عملگر تفاضل پیشرو** می‌نامند. ما همیشه با نقاط گرهی متساوی‌الفاصلهٔ $x_i = x_0 + ih$، $i = 0, 1, \ldots, n$ کار می‌کنیم. پس می‌نویسیم

$$\Delta f(x_i) = f(x_{i+1}) - f(x_i)$$

یا به گونهٔ خلاصه‌تر

$$\Delta f_i = f_{i+1} - f_i \qquad f(x_i) = f_i \qquad (۲.۳.۳)$$

برای $r \geq 0$، تعریف می‌کنیم

$$\Delta^{r+1} f(z) = \Delta^r f(z+h) - \Delta^r f(z) \qquad (۳.۳.۳)$$

با $\Delta^0 f(z) = f(z)$. $\Delta^r f(z)$ تفاضل پیشرو مرتبهٔ rام در z است. تفاضلات پیشرو به‌سادگی محاسبه می‌شوند، و مثالهایی بعداً در ارتباط با فرمول درونیابی داده خواهد شد.

ابتدا، با به کاربردن فرمول تفاضلات منقسم نیوتن، نتایجی برای عملگر تفاضل پیشرو به‌دست می‌آوریم.

لم ۱ برای $k \geq 0$

$$f[x_0, x_1, \ldots, x_k] = \frac{1}{k!h^k}\Delta^k f_0 \qquad (۴.۳.۳)$$

برهان برای $k = 0$، درستی نتیجه بدیهی است. برای $k = 1$،

$$f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0} = \frac{1}{h}\Delta f_0$$

که (۴.۳.۳) را نشان می‌دهد. گیریم حکم (۴.۳.۳) برای تمام تفاضلات پیشرو از مرتبهٔ $k \leq r$ درست است. در این صورت برای $k = r + 1$، با استفاده از (۷.۲.۳)

$$f[x_0, x_1, \ldots, x_{r+1}] = \frac{f[x_1, \ldots, x_{r+1}] - f[x_0, \ldots, x_r]}{x_{r+1} - x_0}$$

با به‌کارگرفتن فرض استقراء، این برابر است با

$$\frac{1}{(r+1)h}\left[\frac{1}{r!h^r}\Delta^r f_1 - \frac{1}{r!h^r}\Delta^r f_0\right] = \frac{1}{(r+1)!h^{r+1}}\Delta^{r+1} f_0 \quad \blacksquare$$

اکنون فرمول (۹.۲.۳) درونیابی نیوتن را به صورت فرمولی که از تفاضلات پیشرو به جای تفاضلات منقسم، استفاده می‌کند درمی‌آوریم. برای یک مقدار داده‌شدهٔ xای که به‌ازای آن چند جمله‌ای درونیاب را محاسبه خواهیم کرد. برای نشان دادن وضعیت x نسبت به x_0 تعریف می‌کنیم

$$\mu = \frac{x - x_0}{h}$$

برای مثال، $\mu = 1{,}6$ یعنی x در $\frac{6}{10}$ فاصلهٔ x_1 تا x_2 قرار دارد. به فرمولی برای

$$(x - x_0)\ldots(x - x_k)$$

برحسب متغیر μ، نیاز داریم:

$$x - x_j = x_0 + \mu h - (x_0 + jh) = (\mu - j)h$$

$$(x - x_0)\ldots(x - x_k) = \mu(\mu - 1)\ldots(\mu - k)h^{k+1} \qquad (۵.۳.۳)$$

از ترکیب (۴.۳.۳) و (۵.۳.۳) با فرمول درونیابی تفاضل منقسم (۹.۲.۳)، به‌دست می‌آوریم:

$$p_n(x) = f_0 + \mu h \cdot \frac{\Delta f_0}{h} + \mu(\mu - 1)h^2 \cdot \frac{\Delta^2 f_0}{2!h^2} + \cdots + \mu(\mu - 1)\ldots(\mu - n + 1)h^n \cdot \frac{\Delta^n f_0}{n!h^n}$$

ضرایب دوجمله‌ای را تعریف می‌کنیم،

$$\binom{\mu}{k} = \frac{\mu(\mu-۱)\dots(\mu-k+۱)}{k!} \qquad k>۰ \qquad (۶.۳.۳)$$

و $\binom{\mu}{۰} = ۱$. آنگاه

$$p_n(x) = \sum_{j=۰}^{n} \binom{\mu}{j} \Delta^j f_۰ \qquad \mu = \frac{x-x_۰}{h} \qquad (۷.۳.۳)$$

این شکل تفاضل پیشرو نیوتن برای چندجمله‌ای درونیاب است.

مثال برای $n=۱$،

$$p_۱(x) = f_۰ + \mu \Delta f_۰ \qquad (۸.۳.۳)$$

این فرمولی است که اغلب در جدول هنگام درونیابی خطی از آن استفاده می‌شود.

برای $n=۲$،

$$p_۲(x) = f_۰ + \mu \Delta f_۰ + \frac{\mu(\mu-۱)}{۲}\Delta^۲ f_۰ \qquad (۹.۳.۳)$$

که یک شکل سادهٔ محاسبه‌پذیر از چندجمله‌ای درونیاب درجهٔ دوم است.

تفاضلات پیشرو به روندی مشابه آنچه که در تفاضلات منقسم گفته‌شده ساخته می‌شوند، ولی در اینجا عملهای تقسیم وجود ندارند (جدول ۶.۳ را ببینید).

مثال تفاضلات پیشرو برای $f(x)=\sqrt{x}$ در جدول ۷.۳ داده شده‌اند. مقادیر چندجمله‌ای درونیاب همان مقادیری هستند که با استفاده از فرمول تفاضلات منقسم نیوتن پیدا شده‌اند، ولی فرمول تفاضلات پیشرو (۷.۳.۳) برای محاسبه به مراتب آسانتر است.

مثال $p_n(x)$ را با استفاده از جدول ۷.۳ برای $n=۱,۲,۳,۴$، با $x=۲٫۱۵$ محاسبه کنید. توجه کنید که $\sqrt{۲٫۱۵}=۱٫۴۶۶۲۸۷۸$؛ و $\mu=۰٫۵$.

$$p_۱(x) = ۱٫۴۱۴۲۱۴ + ۰٫۵(۰٫۰۳۴۹۲۴) = ۱٫۴۱۴۲۱۴ + ۰٫۵۲۳۸۶ = ۱٫۴۶۶۶$$

$$p_۲(x) = p_۱(x) + \frac{(۰٫۵)(۰٫۵)}{۲}(-۰٫۰۰۰۸۲۲) = ۱٫۴۶۶۶ - ۰٫۰۰۰۳۰۸۲۵$$

$$= ۱٫۴۶۶۲۹۲$$

جدول ۶.۳ قالب ساختن تفاضلات پیشرو

x_i	f_i	Δf_i	$\Delta^2 f_i$	$\Delta^3 f_i$	...
x_0	f_0				
		Δf_0			
x_1	f_1		$\Delta^2 f_0$		
		Δf_1		$\Delta^3 f_0$	
x_2	f_2		$\Delta^2 f_1$		
		Δf_2		$\Delta^3 f_1$	⋮
x_3	f_3		$\Delta^2 f_2$		
		Δf_3		$\Delta^3 f_2$	
x_4	f_4		$\Delta^2 f_3$		
		Δf_4			
x_5	f_5				
⋮	⋮				

$$p_3(x) = p_2(x) + \frac{(1{,}5)(0{,}5)(-0{,}5)}{6}(0{,}000055) = 1{,}466292 - 0{,}0000034$$
$$= 1{,}466288$$

$$p_4(x) = p_3(x) + \frac{(1{,}5)(0{,}5)(-0{,}5)(-1{,}5)}{24}(-0{,}000005) = 1{,}466288$$
$$+ 0{,}0000012 = 1{,}466288$$

جمله‌های تصحیحی به‌سادگی محاسبه می‌شوند، و با مشاهدهٔ اندازهٔ آنها، یک تصوّر دقیق کلّی در این که چه وقت درجهٔ n به‌قدر کافی بزرگ است به‌دست می‌آید. توجه کنید که هفت رقم دقت در جدول مقادیر $\sqrt{x}$ حداکثر به یک رقم دقت در تفاضل پیشرو $\Delta^4 f_0$

جدول ۷.۳ جدول تفاضلات پیشرو برای $f(x) = \sqrt{x}$

x_i	f_i	Δf_i	$\Delta^2 f_i$	$\Delta^3 f_i$	$\Delta^4 f_i$
۲٫۰	۱٫۴۱۴۲۱۴				
		۰٫۰۳۴۹۲۴			
۲٫۱	۱٫۴۴۹۱۳۸		−۰٫۰۰۰۸۲۲		
		۰٫۰۳۴۱۰۲		۰٫۰۰۰۰۵۵	
۲٫۲	۱٫۴۸۳۲۴۰		−۰٫۰۰۰۷۶۷		−۰٫۰۰۰۰۰۵
		۰٫۰۳۳۳۳۵		۰٫۰۰۰۰۵۰	
۲٫۳	۱٫۵۱۶۵۷۵		−۰٫۰۰۰۷۱۷		
		۰٫۰۳۲۶۱۸			
۲٫۴	۱٫۵۴۹۱۹۳				

می‌انجامد. تفاضلات پیشرو از مرتبهٔ بیشتر از ۳ تقریباً همه نتیجهٔ خطاهای گردکردن در تفریق درایه‌های جدول است؛ در نتیجه درونیابی در این جدول باید به چندجمله‌ییهای از درجهٔ کمتر از چهار محدود شود. این موضوع با یک توجیه نظری بیشتری در بخش بعد خواهد آمد.

شکلهای دیگری از تفاضلات و فرمولهای درونیابی متناظر با آنها وجود دارد.

تفاضل پسرو را چنین تعریف می‌کنیم:

$$\nabla f(z) = f(z) - f(z-h)$$
$$\nabla^{r+1} f(z) = \nabla^r f(z) - \nabla^r f(z-h) \quad r \geq 0 \qquad (۱۰.۳.۳)$$

نتایج کاملاً مشابهی با تفاضلات پیشرو می‌توان به‌دست آورد. و ما فرمول درونیابی تفاضل پسرو نیوتن را به‌دست می‌آوریم.

$$p_n(x) = f_0 + \binom{-\nu}{1}\nabla f_0 + \binom{-\nu+1}{2}\nabla^2 f_0 + \ldots + \binom{-\nu+n-1}{n}\nabla^n f_0 \quad (۱۱.۳.۳)$$

در این فرمول، مانند قبل، گره‌های درونیابی عبارت‌اند از $x_0, x_{-1}, x_{-2}, \ldots, x_{-n}$ با $x_{-j} = x_0 - jh$ عدد ν با

$$\nu = \frac{x_0 - x}{h}$$

داده می‌شود، و بدین معناست که وقتی از این فرمول استفاده می‌نماییم x معمولاً کوچکتر از x_0 است. نمودار تفاضلات پسرو را می‌توان به طریق مشابه با نمودار تفاضلات پیشرو ساخت. فرمول تفاضل پسرو در فصل ۶، برای مطالعهٔ خانوادهٔ فرمولهای آدامز[1] (به پاس احترام به جان کوچ آدامز ستاره‌شناس انگلیسی قرن نوزدهم) در حل عددی معادلات دیفرانسیل به‌کار برده شده است.

فرمولهای تفاضلی دیگری و فرمولهای درونیابی متناظر با آنها را می‌توان پیدا کرد. چون کاربرد آنها بسیار کمتر از آنهایی است که بیان شد، خوانندگان را به هیلدبرانت[2] ارجاع می‌دهیم.

1. John Couch Adams 2. Hildebrand (1956)

۴.۳ خطاها در داده‌ها و تفاضلات پیشرو

برای پیداکردن نوفه در داده‌های فیزیکی، مادامی که نوفه نسبت به حدود خطای تجربی بزرگ است می‌توانیم از یک جدول تفاضل پیشرو، استفاده کنیم. با چند لم مقدماتی شروع می‌کنیم.

لم ۲ $\Delta^r f(x_i) = h^r f^{(r)}(\xi_i)$، برای مقداری از $x_i \le \xi_i \le x_{i+r}$.

برهان با استفاده از لم ۱ و (۱۲.۲.۳).

$$\Delta^r f_i = h^r r! f[x_i, \ldots, x_{i+r}] = h^r r! \frac{f^{(r)}(\xi_i)}{r!} = h^r f^{(r)}(\xi_i)$$

■

لم ۳ برای هر دو تابع دلخواه f و g، و هر دو مقدار ثابت α و β،

$$\Delta^r(\alpha f(x) + \beta g(x)) = \alpha \Delta^r f(x) + \beta \Delta^r g(x) \qquad r \ge 0$$

برهان حکم برای $r = 0$ یا $r = 1$ بدیهی است. فرض کنید حکم برای هر $r \le n$ درست است، و آن را برای $r = n + 1$ ثابت کنید؛ با به کاربردن تعریف (۳.۳.۳) و فرض استقرا:

$$\begin{aligned}\Delta^{n+1}[\alpha f(x) + \beta g(x)] =& \Delta^n[\alpha f(x+h) + \beta g(x+h)] \\ & - \Delta^n[\alpha f(x) + \beta g(x)] \\ =& \alpha \Delta^n f(x+h) + \beta \Delta^n g(x+h) \\ & - \alpha \Delta^n f(x) - \beta \Delta^n g(x)\end{aligned}$$

سپس با جابه‌جاکردن جملات، داریم:

$$\begin{aligned}&\alpha[\Delta^n f(x+h) - \Delta^n f(x)] + \beta[\Delta^n g(x+h) - \Delta^n g(x)] \\ &\quad = \alpha \Delta^{n+1} f(x) + \beta \Delta^{n+1} g(x)\end{aligned}$$

■

لم ۲ می‌گوید که اگر مشتقات $f(x)$ کراندار باشند، یا اگر در مقایسه با h^{-n} به سرعت افزایش نیابند، آنگاه تفاضلات پیشرو $\Delta^n f(x)$، با افزایش n، بایستی کوچکتر شوند. ما بعد اثر خطاهای گردکردن و خطاهای دیگری با اندازهٔ بزرگتر از خطاهای گردکردن را در نظر می‌گیریم. می‌نویسیم

$$f(x_i) = \tilde{f}_i + e(x_i) \qquad i = 0, 1, 2, \ldots \qquad (۱.۴.۳)$$

که $\tilde{f}_i$ مقداری است که در جدول برای ساختن جدول تفاضل پیشرو به‌کار می‌بریم. پس:

$$\begin{aligned}\Delta^r \tilde{f}_i &= \Delta^r f(x_i) - \Delta^r e(x_i) \\ &= h^r f^{(r)}(\xi_i) - \Delta^r e(x_i) \qquad (۲.۴.۳)\end{aligned}$$

همان گونه که در جدول تفاضل پیشرو برای $f(x) = \sqrt{x}$ نشان داده شد، اولین عبارت با افزایش r کوچکتر می‌شود.

برای درک بهتر از رفتار $\Delta^r e(x_i)$، حالت سادهٔ زیر را در نظر می‌گیریم که در آن

$$e(x_i) = \begin{cases} ۰ & i \neq k \\ \varepsilon & i = k \end{cases} \qquad (۳.۴.۳)$$

تفاضل پیشرو این تابع در جدول ۸.۳ داده شده است. می‌توان ثابت کرد که ستون $\Delta^r e(x_i)$ به شکل زیر است:

$$۰, \ldots, ۰, \varepsilon, -\binom{r}{۱}\varepsilon, \binom{r}{۲}\varepsilon, -\binom{r}{۳}\varepsilon, \ldots, (-۱)^{r+۱}\varepsilon, ۰, \ldots \qquad (۴.۴.۳)$$

بنابراین اثر خطای گردکردن، وقتی تفاضلات درجات بالاتر شکل می‌گیرند، پخش می‌شود و افزایش می‌یابد.

دربارهٔ خطاهای گردکردن که یک تابع کلی خطا را مشخص می‌کنند، مانند تابع خطا در (۱.۴.۳)، به اثرات آنها می‌توان به دیدهٔ مجموع توابع (۳.۴.۳) نگاه کرد. چون مقادیر $e_i(x)$ به‌طور کلی از نظر اندازه و علامت تغییر می‌یابند، اثرات آنها ظاهراً به تصادف روی هم می‌افتند. ولی اندازهٔ تفاضلات آنها همچنان افزایش می‌یابد، و در نتیجه، تفاضلات مرتبهٔ بالای مقادیر جدول $\tilde{f}_i$ سرانجام بی‌فایده می‌شوند. هنگامی که اندازهٔ تفاضلات $\Delta^r \tilde{f}_i$ با ازدیاد r شروع به افزایش می‌کند، به احتمال زیاد، خطای گردکردن بر آنها غلبه می‌کند و این تفاضلات نباید به‌کار روند. لذا از یک فرمول درونیابی با درجهٔ کمتر از r باید استفاده شود.

کشف نوفه در داده‌ها. همین تحلیل را می‌توان برای کشف و تصحیح خطاهای منفردی که نسبت به خطاهای گردکردن بزرگ هستند، به‌کار برد. چون (۲.۴.۳) می‌گوید که *اثر خطاها سرانجام غالب* می‌شود، ما دنبال الگویی مانند (۴.۴.۳) می‌گردیم. روش کلّی در جدول ۹.۳ نشان داده شده است. از (۲.۴.۳)

$$\Delta^r e(x_i) = \Delta^r f(x_i) - \Delta^r \tilde{f}_i$$

جدول ۸.۳ تفاضلات پیشرو برای تابع خطای $e(x)$

x_i	$e(x_i)$	$\Delta e(x_i)$	$\Delta^2 e(x_i)$	$\Delta^3 e(x_i)$
$\vdots$	$\vdots$	۰	$\vdots$	۰
x_{k-2}	۰	$\vdots$	۰	$\vdots$
		۰		ε
x_{k-1}	۰		ε	
		ε		-3ε
x_k	ε		-2ε	
		$-\varepsilon$		3ε
x_{k+1}	۰		ε	
		۰		$-\varepsilon$
x_{k+2}	۰		۰	
				۰
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

جدول ۹.۳ مثالی از کشف یک خطای منفرد در داده‌ها

$\tilde{f}_i$	$\Delta\tilde{f}_i$	$\Delta^2\tilde{f}_i$	$\Delta^3\tilde{f}_i$	خطاهای تخمین	تخمین $\Delta^3 f(x_i)$
۰٫۱۰۳۹۶					
	۰٫۰۱۷۰۰				
۰٫۱۲۰۹۶		−۰٫۰۰۰۱۴			
	۰٫۰۱۶۸۶		−۰٫۰۰۰۰۳	۰	−۰٫۰۰۰۰۳
۰٫۱۳۷۸۲		−۰٫۰۰۰۱۷			
	۰٫۰۱۶۶۹		−۰٫۰۰۰۰۲	۰	−۰٫۰۰۰۰۲
۰٫۱۵۴۵۱		−۰٫۰۰۰۱۹			
	۰٫۰۱۶۵۰		۰٫۰۰۰۰۶	ϵ	−۰٫۰۰۰۰۲
۰٫۱۷۱۰۱		−۰٫۰۰۰۱۳			
	۰٫۰۱۶۳۷		−۰٫۰۰۰۲۵	-3ϵ	−۰٫۰۰۰۰۲
۰٫۱۸۷۳۸		−۰٫۰۰۰۳۸			
	۰٫۰۱۵۹۹		۰٫۰۰۰۲۱	3ϵ	−۰٫۰۰۰۰۲
۰٫۲۰۳۳۷		−۰٫۰۰۰۱۷			
	۰٫۰۱۵۸۲		−۰٫۰۰۰۱۰	$-\epsilon$	−۰٫۰۰۰۰۲
۰٫۲۱۹۱۹		−۰٫۰۰۰۲۷			
	۰٫۰۱۵۵۵				
۰٫۲۳۴۷۴					

با استفاده از $r = 3$ و انتخاب یکی از درایه‌های خطا به‌طور دلخواه، مثلاً اولین، داریم

$$\varepsilon = -۰٫۰۰۰۰۲ - (۰٫۰۰۰۰۶) = -۰٫۰۰۰۰۸$$

این خطا را امتحان می‌کنیم تا ببینیم چگونه ستون $\Delta^r\tilde{f}_i$ را تغییر می‌دهد. (جدول ۱۰.۳ را ببینید).

جدول ۱۰.۳ تصحیح یک خطای داده‌ها

$\Delta^{۳}f(x_i)$	$\Delta^{۳}e(x_i)$	$\Delta^{۳}\tilde{f}_i$
−۰٫۰۰۰۰۲	۰٫۰	−۰٫۰۰۰۰۲
−۰٫۰۰۰۰۲	−۰٫۰۰۰۰۸	۰٫۰۰۰۰۶
−۰٫۰۰۰۰۱	۰٫۰۰۰۲۴	−۰٫۰۰۰۲۵
−۰٫۰۰۰۰۳	−۰٫۰۰۰۲۴	۰٫۰۰۰۲۱
−۰٫۰۰۰۰۲	۰٫۰۰۰۰۸	−۰٫۰۰۰۱۰

با یک انتخاب دیگر ϵ، مثلاً $\epsilon = -۰٫۰۰۰۰۷$، بهبودی حاصل نمی‌شود، اگرچه نتایج ممکن است به همان خوبی بماند.

به عقب برمی‌گردیم، درایهٔ $\tilde{f}_i = ۰٫۱۸۷۳۸$ بایستی چنین باشد:

$$f(x_i) = \tilde{f}_i + e(x_i) = ۰٫۱۸۷۳۸ + (-۰٫۰۰۰۰۸) = ۰٫۱۸۷۳۰$$

در جدولی که دو یا سه خطای منفرد وجود داشته باشد، تفاضلات مراتب بالاتر آنها ممکن است تداخل کنند و کشف خطاها را مشکلتر سازند (مسألهٔ ۲۲ را ببینید).

۵.۳ نتایج دیگری دربارهٔ خطای درونیابی

فرمول خطا را دوباره در نظر می‌گیریم.

$$f(x) - p_n(x) = \frac{(x-x_۰)\dots(x-x_n)}{(n+۱)!} f^{(n+۱)}(\xi_x) \qquad \xi_x \in \mathscr{H}\{x_۰, \dots, x_n, x\} \tag{۱.۵.۳}$$

فرض می‌کنیم که $f(x)$ در یک بازهٔ I، که شامل $\mathscr{H}\{x_۰, x_۱, \dots, x_n, x\}$ است، به ازای جمیع مقادیر x مورد نظر، $n+۱$ بار پیوسته مشتق‌پذیر است. چون ξ_x مجهول است، باید برای محاسبهٔ (۱.۵.۳)، به‌جای $f^{(n+۱)}(\xi_x)$ قرار دهیم

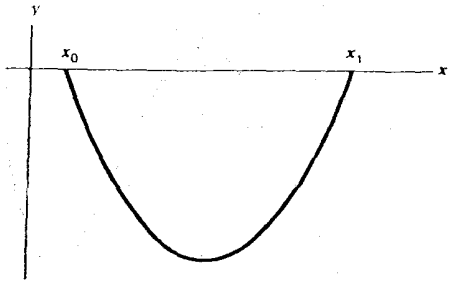
$$c_{n+۱} = \underset{t\in I}{\text{Max}} \mid f^{(n+۱)}(t) \mid \tag{۲.۵.۳}$$

توجه خود را به کراندارنمودن چندجمله‌ای

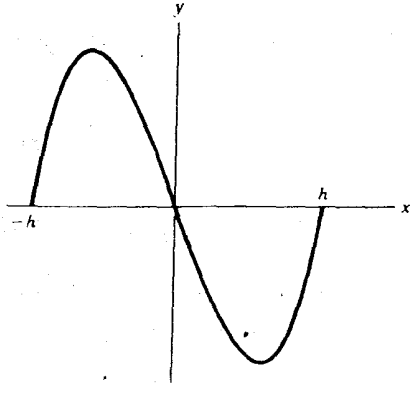
$$\Psi_n(x) = (x-x_۰)\dots(x-x_n) \tag{۳.۵.۳}$$

متمرکز می‌کنیم. از (۱.۵.۳)، به‌دست می‌آوریم

$$\underset{x\in I}{\text{Max}} \mid f(x) - p_n(x) \mid \leq \frac{c_{n+۱}}{(n+۱)!} \underset{x\in I}{\text{Max}} \mid \Psi_n(x) \mid \tag{۴.۵.۳}$$



شکل ۲.۳ $y = \Psi_1(x)$



شکل ۳.۳ $y = \Psi_2(x)$

ما فقط حالتی را که گره‌ها متساوی‌الفاصله هستند در نظر می‌گیریم: $x_j = x_0 + jh$، $j = 0, 1, \ldots, n$. ابتدا حالتهای ویژه‌ای از مقادیر n را درنظر می‌گیریم و سپس برای حالت کلی n اظهار نظر می‌کنیم.

حالت ۱. $n = 1$. $\Psi_1(x) = (x - x_0)(x - x_1)$. به‌سادگی دیده می‌شود

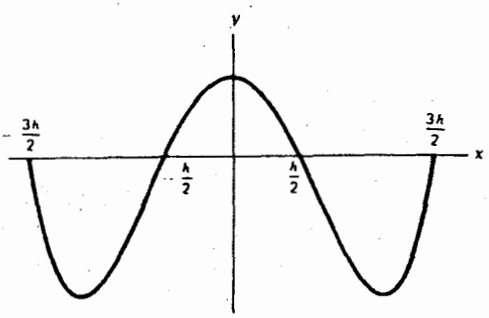
$$\max_{x_0 \le x \le x_1} | \Psi_1(x) | = \frac{h^2}{4}$$

برای توضیح به شکل ۲.۳ نگاه کنید.

حالت ۲. $n = 2$. برای کراندار نمودن $\Psi_2(x)$ در $[x_0, x_2]$، آن را بر محور xها انتقال می‌دهیم تا چند جمله‌ای هم‌ارز زیر را پیدا کنیم

$$\hat{\Psi}_2(x) = (x + h)x(x - h)$$

که نمودار آن در شکل ۳.۳ نشان داده شده است. شکل و اندازهٔ آن درست مانند چند جملهٔ اصلی



شکل ۴.۳ $y = \hat{\Psi}_۴(x)$

$\Psi_۲(x)$ است، ولی از نظر تحلیلی ساده‌تر می‌توان آن را کراندار کرد. با استفاده از $\hat{\Psi}(x)$ به‌سادگی به‌دست می‌آوریم

$$\underset{x_۱-h/۲\le x\le x_۱+h/۲}{\text{Max}} \mid \Psi_۲(x) \mid = ۰\text{ر}۳۷۵h^۳$$

$$\underset{x_۰\le x\le x_۲}{\text{Max}} \mid \Psi_۲(x) \mid = \frac{۲\sqrt{۳}}{۹}h^۳ \doteq ۰\text{ر}۳۸۵h^۳ \qquad (۵.۵.۳)$$

بنابراین قرارگرفتن x در بازهٔ $[x_۰, x_۲]$ نزدیک $x_۱$ مهم نیست، اگرچه برای درونیابیهای درجهٔ بالاتر، این یک تفاوتی به بار خواهد آورد. از ترکیب (۵.۵.۳) و (۴.۵.۳) داریم

$$\underset{x_۰\le x\le x_۲}{\text{Max}} \mid f(x) - p_۲(x) \mid \le \frac{\sqrt{۳}}{۲۷}h^۳ \cdot \underset{x_۰\le x\le x_۲}{\text{Max}} \mid f^{(۳)}(t) \mid \qquad (۶.۵.۳)$$

و $\sqrt{۳}/۲۷ \doteq ۰\text{ر}۰۶۴$.

حالت ۳. $n = ۳$. مانند گذشته، چندجمله‌ای را برای متقارن کردن گره‌ها نسبت به مرکز انتقال می‌دهیم، به‌دست می‌آوریم:

$$\hat{\Psi}_۳(x) = (x^۲ - \frac{۹}{۴}h^۲)(x^۲ - \frac{۱}{۴}h^۲)$$

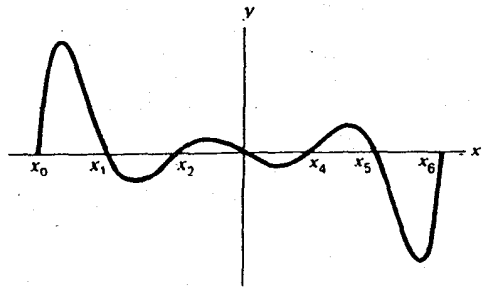
نمودار $\hat{\Psi}_۳(x)$ در شکل ۴.۳ نشان داده شده است. با این تغییر

$$\underset{x_۱\le x\le x_۲}{\text{Max}} \mid \Psi_۳(x) \mid = \frac{۹}{۱۶}h^۴ \doteq ۰\text{ر}۵۶h^۴$$

$$\underset{x_۰\le x\le x_۳}{\text{Max}} \mid \Psi_۳(x) \mid = h^۴$$

بنابراین در درونیابی $f(x)$ در نقطهٔ x، گره‌ها باید به گونه‌ای انتخاب شوند که $x_۱ < x < x_۲$. در این صورت از (۱.۵.۳)،

$$\underset{x_۱\le x\le x_۲}{\text{Max}} \mid f(x) - p_۳(x) \mid \le \frac{۳h^۴}{۱۲۸} \cdot \underset{x_۰\le t\le x_۳}{\text{Max}} \mid f^{(۴)}(t) \mid \qquad (۷.۵.۳)$$



شکل ۵.۳ $y = \Psi_6(x)$

حالت ۴. برای حالت کلی $n > 3$، رفتاری که هم‌اکنون برای $n = 3$ نشان داده شده، تشدید می‌شود. برای مثال، نمودار $\Psi_6(x)$ را که در شکل ۵.۳ نشان داده شده است در نظر می‌گیریم. مانند قبل، می‌توانیم نشان دهیم

$$\max_{x_2 \le x \le x_4} |\Psi_6(x)| \doteq ۱۲٫۳۶h^7$$

$$\max_{x_0 \le x \le x_6} |\Psi_6(x)| \doteq ۹۵٫۸h^7$$

برای آنکه خطای درونیابی به حداقل برسد، گره‌های درونیابی باید به گونه‌ای انتخاب شوند که نقطهٔ درونیابی x در حد ممکن به نقطهٔ وسط $[x_0, x_n]$ نزدیک باشد. وقتی n افزایش می‌یابد، برای درونیابی در گرههای متساوی‌الفاصله، تأکید زیادی می‌شود که نقاط طوری انتخاب شوند که x نزدیک وسط $[x_0, x_n]$ قرار گیرد.

مثال درونیابی درجهٔ پنج $J_0(x)$ را در $x = ۲٫۴۵$ در نظر می‌گیریم، با مقادیر $J_0(x)$ که در جدول ۴.۳، بخش ۲.۳ داده شده‌اند. اوّل $x_0 = ۲٫۴$ و $x_5 = ۲٫۹$ را به‌کار می‌بریم. به‌دست می‌آوریم:

$$P_5(۲٫۴۵) = -۰٫۲۳۲۲۶۷۳۸۴ \quad \text{خطا} = -۴٫۹ \times ۱۰^{-9}$$

ثانیاً $x_0 = ۲٫۲$ و $x_5 = ۲٫۷$ را به‌کار می‌بریم. آنگاه

$$P_5(۲٫۴۵) = -۰٫۲۳۲۲۶۷۴۲۳ \quad \text{خطا} = ۱٫۰ \times ۱۰^{-9}$$

در اثر نزدیکی وضعیت x به وسط $[x_0, x_5]$، خطا پنج بار کوچکتر است.

در جداول آبراموویتس و استگون[1] (۱۹۶۴) تعداد زیادی ارقام معنی‌دار داده شده‌اند، و h،

1. Abramowitz and Stegun

فاصلهٔ شبکه به همان نسبت کوچک نیست. در نتیجه، درونیابی با درجهٔ بالا باید به‌کار گرفته شود. اگرچه این کار بیشتری برای استفاده‌کننده از جدول ایجاد می‌نماید، ولی موجب می‌شود که جدول در شکل فشرده‌تر جای بسیار کمتر بگیرد و می‌توان جداول بیشتری برای توابع بیشتری در یک حجم جای داد. وقتی که با استفاده از این جداول، درونیابی با درجهٔ بالا انجام می‌گیرد، گره‌ها باید طوری انتخاب شوند که در صورت امکان x نزدیک به $(x_0 + x_n)/2$ قرار گیرد.

مسألهٔ تقریب. وقتی از رایانه استفاده می‌کنیم، معمولاً ترجیح می‌دهیم که یک تقریب تحلیلی تابع را ذخیره کنیم نه یک جدول مقادیری که با آن درونیابی می‌کنیم. تقریب‌یابی یک تابع داده‌شدهٔ $f(x)$ در یک بازهٔ داده‌شدهٔ $[a, b]$ را با استفاده از چندجمله‌ییهای درونیابی در نظر می‌گیریم. به‌ویژه، حالتی را در نظر می‌گیریم که چندجمله‌ای $p_n(x)$ با درونیابی $f(x)$ در زیرتقسیمهای متساوی‌الفاصلهٔ $[a, b]$ تولید شده است.

برای هر $n \geq 1$، تعریف می‌کنیم $h = (b-a)/n$، $x_j = a + jh$، $j = 0, 1, \ldots, n$. گیریم $p_n(x)$ چندجمله‌ای درونیاب $f(x)$ در x_0، ...، x_n است. سپس می‌پرسیم که آیا

$$\max_{a \leq x \leq b} | f(x) - p_n(x) | \qquad (۸.۵.۳)$$

با $n \to \infty$ به صفر میل می‌کند یا نه؟ جواب این است که، لزوماً نه. برای توابع بسیاری، مثلاً e^x در $[0, 1]$، خطا در (۸.۵.۳) وقتی $n \to \infty$ به صفر همگراست (مسألهٔ ۲۴ را ببینید). ولی توابع دیگری وجود دارند، که کاملاً خوشرفتارند و همگرا نیستند.

معروفترین مثال عدم همگرایی مثالی است که کارل رونگه[1] آورده است. گیریم

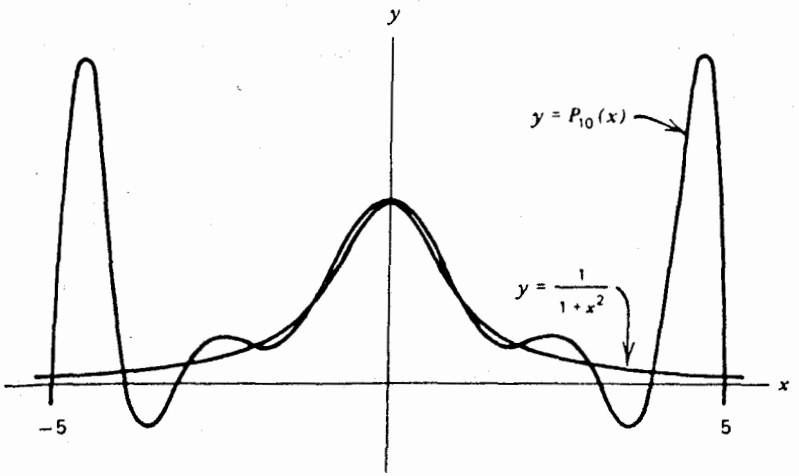
$$f(x) = \frac{1}{1 + x^2} \qquad -5 \leq x \leq 5 \qquad (۹.۵.۳)$$

در کتاب آیزکسون و کِلر[2] (۱۹۶۶، صص ۲۷۹.۲۷۵) نشان داده شده است که برای هر $|x| > 3.64$، $5 > |x|$،

$$\sup_{n \geq k} | f(x) - p_n(x) | = \infty, \qquad k \geq 0 \qquad (۱۰.۵.۳)$$

بنابراین $p_n(x)$ برای چنین xی هرگاه $n \to \infty$، به $f(x)$ همگرا نمی‌شود. در ابتدا این امر مغایر با احساس ما به‌نظر می‌رسد، ولی اساس این امر رفتار چندجمله‌ای $y = \Psi_n(x)$ نزدیک به دو سربازهٔ $[a, b] = [x_0, x_n]$ است. این موضوع در نمودار $p_{10}(x)$ و $f(x)$ در شکل ۶.۳

1. Carl Runge 2. Isaacson and Keller



شکل ۶.۳ درونیابی $1/(1+x^2)$

به خوبی نمایان است. اگرچه درونیابی با نقاط متساوی‌الفاصله ممکن است دنباله‌ای همگرا از چندجمله‌ییهای درونیابی ایجاد ننماید، مجموعه‌های مناسبتری از نقاط شبکه‌ای $\{x_j\}$ وجود دارند که برای تمام توابع پیوسته مشتقپذیر تقریبات خوبی به بار می‌آورند. این شبکه در بخش ۷.۴ از فصل ۴ دربارهٔ نظریهٔ تقریب، توضیح داده شده است.

۶.۳ درونیابی ارمیت

در بسیاری از کاربردها، مناسب است که چندجمله‌ای $p(x)$ چنان در نظر گرفته شود که تابع $f(x)$ را درونیابی نماید، و بعلاوه $p'(x)$ چندجمله‌ای $f'(x)$، مشتق تابع، را نیز درونیابی کند. در این کتاب، کاربرد ابتدایی این موضوع، به‌عنوان یک ابزار ریاضی در مبحث انتگرالگیری عددی گاوسی در فصل ۵، آمده است. این‌گونه درونیابی برای مطالعهٔ روشهای عددی در حل بعضی مسائل معادلات دیفرانسیل نیز مناسب‌اند.

موضوع را با در نظر گرفتن یک قضیهٔ وجود برای مسألهٔ اساسی درونیابی شروع می‌کنیم

$$p(x_i) = y_i \qquad p'(x_i) = y'_i \qquad i = ۱, \ldots, n \qquad (۱.۶.۳)$$

که در آن $x_۱$، ...، x_n گره‌های متمایز (حقیقی یا مختلط) و $y_۱$، ...، y_n، $y'_۱$، ...، y'_n داده‌ها هستند (نمادگذاری از $n+۱$ نقطهٔ $\{x_۰, \ldots, x_n\}$ به $2n$ نقطهٔ $\{x_۱, \ldots, x_n\}$ تغییر یافته است تا مانند کاربرد در فصل ۵ باشد). $2n$ شرط برای (۱.۶.۳) گذاشته شده‌است. پس دنبال یک چندجمله‌ای $p(x)$ حداکثر از درجهٔ $2n-۱$ هستیم.

برای بحث دربارهٔ وجود و یکتایی $p(x)$، سوّمین برهان قضیهٔ ۱.۳ را تعمیم می‌دهیم. برای هماهنگی با نمادگذاری بخش ۱.۳، می‌گیریم

$$\begin{aligned}
\Psi_n(x) &= (x-x_1)\dots(x-x_n)\\
l_i(x) &= \frac{(x-x_1)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)}{(x_i-x_1)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x-x_n)}\\
&= \frac{\Psi_n(x)}{(x-x_i)\Psi_n'(x_i)}\\
\tilde{h}_i(x) &= (x-x_i)[l_i(x)]^2\\
h_i(x) &= [1-2l_i'(x_i)(x-x_i)][l_i(x)]^2
\end{aligned} \tag{۲.۶.۳}$$

در این صورت به ازای $i,j=1,\dots,n$،

$$h_i'(x_j)=\tilde{h}_i(x_j)=0 \qquad 1\le i,j\le n$$

$$h_i(x_j)=\tilde{h'}_i(x_j)=\begin{cases}0 & i\neq j\\ 1 & i=j\end{cases} \tag{۳.۶.۳}$$

چندجمله‌ای درونیاب برای (۱.۶.۳) با رابطهٔ زیر داده می‌شود

$$H_n(x)=\sum_{i=1}^{n} y_i h_i(x)+\sum_{i=1}^{n} y_i'\tilde{h}_i(x) \tag{۴.۶.۳}$$

برای نشان دادن یکتایی $H_n(x)$، فرض می‌کنیم چندجمله‌ای دیگری با درجهٔ نابزرگتر از $2n-1$ باشد که در (۱.۶.۳) صدق می‌کند. تابع $R=H-G$ را تعریف می‌کنیم. لذا با توجه به (۱.۶.۳)

$$R(x_i)=R'(x_i)=0 \qquad i=1,2,\dots,n$$

که R یک چندجمله‌ای از درجهٔ نابزرگتر از $2n-1$ است با n ریشهٔ دوگانهٔ $x_1, x_2, \dots, x_n$. این مطلب فقط وقتی درست است که به‌ازای یک چندجمله‌ای $q(x)$ داشته باشیم

$$R(x)=q(x)(x-x_1)^2\dots(x-x_n)^2$$

اگر $q(x)\not\equiv 0$، آنگاه درجهٔ $R(n)\le 2n$ که یک تناقض است. پس باید $R(x)\equiv 0$.

برای پیدا کردن یک شکل محاسبه پذیرتر از (۴.۶.۳) و یک عبارت خطا، ابتدا چندجمله‌ای درونیاب $f(x)$ را در نقاط z_1، z_2، ...، z_{2n} درنظر می‌گیریم که به شکل تفاضل منقسم نیوتن نوشته شده است:

$$p_{2n-1}(x) = f(z_1) + (x - z_1)f[z_1, z_2] + \dots$$
$$+(x - z_1)\dots(x - z_{2n-1})f[z_1, \dots, z_{2n}] \qquad (۵.۶.۳)$$

برای خطا

$$f(x) - p_{2n-1}(x) = (x - z_1)\dots(x - z_{2n})f[z_1, \dots, z_{2n}, x] \qquad (۶.۶.۳)$$

در فرمول (۵.۶.۳) می‌توانیم گره‌ها را منطبق برهم فرض کنیم و باز فرمول برقرار خواهد بود. به‌ویژه فرض کنید

$$z_1, z_2 \to x_1, \qquad z_3, z_4 \to x_2, \dots, z_{2n-1}, z_{2n} \to x_n$$

تا به‌دست آوریم

$$p_{2n-1}(x) = f(x_1) + (x - x_1)f[x_1, x_1] + (x - x_1)^2 f[x_1, x_1, x_2] + \dots$$
$$+(x - x_1)^2 \dots (x - x_{n-1})^2 (x - x_n) f[x_1, x_1, \dots, x_n, x_n] \qquad (۷.۶.۳)$$

این یک چندجمله‌ای است با درجهٔ نابزرگتر از $2n - 1$. برای خطای آن، از (۶.۶.۳) وقتی که $z_1, z_2 \to x_1, \dots, z_{2n-1}, z_{2n} \to x_n$ حد می‌گیریم. چون تفاضل منقسم با فرض آنکه f به اندازه کافی مشتق‌پذیر باشد پیوسته است

$$f(x) - p_{2n-1}(x) = (x - x_1)^2 \dots (x - x_n)^2 f[x_1, x_1, \dots, x_n, x_n, x] \qquad (۸.۶.۳)$$

ادعا می‌کنیم ـ $p_{2n-1}(x) = H_n(x)$. برای اثبات، گیریم $f(x)$ تابعی $2n + 1$ بار پیوسته مشتق‌پذیر باشد، در این صورت ملاحظه می‌کنیم که:

$$f(x_i) - p_{2n-1}(x_i) = 0 \qquad i = 1, 2, \dots, n$$

همچنین

$$f'(x) - p'_{2n-1}(x) = (x - x_1)^2 \dots (x - x_n)^2 \frac{d}{dx} f[x_1, x_1, \dots, x_n, x_n, x]$$
$$+ 2f[x_1, x_1, \dots, x_n, x_n, x] \sum_{i=1}^{n} \left[(x - x_i) \prod_{\substack{j=1 \\ j \neq i}}^{n} (x - x_j)^2 \right]$$

و

$$f'(x_i) - p'_{2n-1}(x_i) = 0 \qquad i = 1, \ldots, n$$

بنابراین درجهٔ p_{2n-1} نابزرگتر از $2n-1$ است و p_{2n-1} و p'_{2n-1} در رابطهٔ (۱.۶.۳) در داده‌های $y_i = f(x_i)$ و $y'_i = f'(x_i)$ صدق می‌کنند. از یکتایی چندجمله‌ای درونیاب ارمیت داریم $p_{2n-1} = H_n$. بنابراین (۷.۶.۳) یک فرمول تفاضل منقسم برای محاسبهٔ $H_n(x)$ است و (۸.۶.۳) یک فرمول خطا به صورت

$$f(x) - H_n(x) = [\Psi_n(x)]^2 f[x_1, x_1, \ldots, x_n, x_n, x] \qquad (۹.۶.۳)$$

به‌دست می‌دهد.

با استفاده از رابطهٔ (۱۳.۲.۳) برای تعمیم (۱۲.۲.۳)، به‌دست می‌آوریم:

$$f(x) - H_n(x) = [\Psi_n(x)]^2 \frac{f^{(2n)}(\xi_n)}{(2n)!} \qquad \xi_x \in \mathscr{H}\{x_1, \ldots, x_n, x\} \quad (۱۰.۶.۳)$$

مثال صورت درونیابی ارمیت که در بیشترین جاها مورد استفاده قرار می‌گیرد شاید چندجمله‌ای درجهٔ ۳ی ارمیت باشد، که مسألهٔ زیر را حل می‌کند:

$$\begin{aligned} p(a) &= f(a) & p'(a) &= f'(a) \\ p(b) &= f(b) & p'(b) &= f'(b) \end{aligned} \qquad (۱۱.۶.۳)$$

فرمول (۴.۶.۳) چنین می‌شود:

$$H_3(x) = \left[1 + 2\frac{x-a}{b-a}\right]\left[\frac{b-x}{b-a}\right]^2 \cdot f(a) + \left[1 + 2\frac{b-x}{b-a}\right]\left[\frac{x-a}{b-a}\right]^2 \cdot f(b)$$
$$+ \frac{(x-a)(b-x)^2}{(b-a)^2} f'(a) - \frac{(x-a)^2(b-x)}{(b-a)^2} f'(b) \qquad (۱۲.۶.۳)$$

فرمول تفاضل منقسم (۷.۶.۳) چنین می‌شود:

$$H_3(x) = f(a) + (x-a)f'(a) + (x-a)^2 f[a, a, b]$$
$$+ (x-a)^2(x-b) f[a, a, b, b] \qquad (۱۳.۶.۳)$$

که در آن

$$f[a,a,b] = \frac{f[a,b] - f'(a)}{b-a}$$
$$f[a,a,b,b] = \frac{f'(b) - ۲f[a,b] + f'(a)}{(b-a)^۲}$$

فرمول (۱۳.۶.۳) را می‌توان با الگوریتم ضرب تودرتو، که مشابه با $Interp$ است که در بخش ۲.۳ آمده است، محاسبه کرد.

فرمول خطا برای (۱۲.۶.۳) یا (۱۳.۶.۳) چنین است:

$$f(x) - H_۲(x) = (x-a)^۲(x-b)^۲ f[a,a,b,b,x] \qquad (۱۴.۶.۳)$$
$$= \frac{(x-a)^۲(x-b)^۲}{۲۴} \cdot f^{(۴)}(\xi_x) \quad \xi_x \in \mathscr{H}\{a,b,x\}$$
$$\max_{a\le x\le b} \mid f(x) - H_۲(x) \mid \le \frac{(b-a)^۴}{۳۸۴} \cdot \max_{a\le t\le b} \mid f^{(۴)}(t) \mid \qquad (۱۵.۶.۳)$$

کاربرد دیگری از چند جمله‌ای درجهٔ سوم ارمیت در بخش بعد خواهد آمد و یک مثال عددی در جدول ۱۲.۳ آن بخش داده شده است.

مسألهٔ درونیابی کلی ارمیت. مسألهٔ سادهٔ (۱.۶.۳) ارمیت را به مسألهٔ زیر تعمیم می‌دهیم: یک چندجمله‌ای $p(x)$ پیدا کنید که در روابط زیر صدق کند.

$$\begin{aligned} p^{(i)}(x_۱) &= y_۱^{(i)} \qquad i = ۰, ۱, \dots, \alpha_۱ - ۱ \\ &\vdots \\ p^{(i)}(x_n) &= y_n^{(i)} \qquad i = ۰, ۱, \dots, \alpha_n - ۱ \end{aligned} \qquad (۱۶.۶.۳)$$

اعداد $y_j^{(i)}$ داده‌های معلوم‌اند و تعداد شرایط برای $p(x)$ در x_j برابر α_j است، $j = ۱, \dots, n$. تعریف می‌کنیم

$$N = \alpha_۱ + \cdots + \alpha_n$$

آنگاه در میان چندجمله‌یهای با درجهٔ نابزرگتر از $N-۱$ چندجمله‌ای یکتایی وجود دارد که در (۱۶.۶.۳) صدق می‌کند. برهان به‌عنوان مسألهٔ ۲۵، واگذار شده است. همهٔ نتایج قبلی، مانند

(۴.۶.۳) و (۹.۶.۳) را می‌توان تعمیم داد. به‌عنوان یک حالت خاص جالب، فرض می‌کنیم $\alpha_1 = N$ و $n = 1$. این بدان معناست که $p(x)$ باید در

$$p^{(i)}(x_1) = f^{(i)}(x_1) \qquad i = 0, 1 \ldots, N-1$$

صدق کند. برای نمادگذاری مناسبتر، به جای $y_1^{(i)}$، $f^{(i)}(x_i^1)$ گذارده‌ایم. در این صورت با استفاده از (۱۵.۲.۳)، شکل تفاضل منقسم نیوتن برای چندجمله‌ای درونیاب ارمیت چنین می‌شود

$$\begin{aligned} p(x) =& f(x_1)(x-x_1)f[x_1,x_1] + (x-x_1)^2 f[x_1,x_1,x_1] + \ldots \\ &+ (x-x_1)^{N-1} f[x_1,\ldots,x_1] \\ =& f(x_1) + (x-x_1)f'(x_1) + \frac{(x-x_1)^2}{2!} f''(x_1) + \ldots \\ &+ \frac{(x-x_1)^{N-1}}{(N-1)!} f^{(N-1)}(x_1) \end{aligned}$$

که چندجمله‌ای تیلر f حول x_1 نیز هست.

۷.۳ درونیابی چندجمله‌یی تکّه‌یی

از اوایل دههٔ ۱۹۶۰، موضوع توابع چندجمله‌یی تکه‌یی، به‌ویژه توابع برازا، عمومیّت روزافزون یافته است. این توابع چندجمله‌یی به انحاء گوناگون در نظریهٔ تقریب، گرافیک رایانه‌یی، برازش داده‌ها، انتگرالگیری و مشتقگیری عددی و حل عددی معادلات انتگرالی، دیفرانسیل، و مشتقات جزیی به‌کار رفته‌اند. ما فقط از دیدگاه نظریهٔ درونیابی به توابع چندجمله‌ای تکّه‌ای نگاه می‌کنیم، ولی کاربردهای زیادی در چند زمینه دارند که درونیابی به‌طور جنبی در آن زمینه‌ها ظاهر می‌شود.

برای یک تابع چندجمله‌ای تکه‌ایِ $p(x)$، یک شبکه از نقاط گرهی وابسته وجود دارد:

$$-\infty < x_0 < x_1 < \cdots < x_n < \infty \qquad (۱.۷.۳)$$

نقاط x_j اغلب **گره** یا **نقطهٔ شکستگی** نامیده می‌شوند. تابع $p(x)$ یک چندجمله‌ای در هر یک از زیربازه‌های:

$$(-\infty, x_0], [x_0, x_1], \ldots, [x_n, \infty) \qquad (۲.۷.۳)$$

است، اگرچه زیربازه‌های $(-\infty, x_0]$ و $[x_n, \infty)$ اغلب در آنها نیستند. $p(x)$ را **یک چندجمله‌ای تکّه‌ای مرتبهٔ** r گویند اگر درجهٔ $p(x)$ در هر یک از زیربازه‌های (۲.۷.۳) کوچکتر از r باشد. هیچ قید پیوستگی برای $p(x)$ و مشتقات آن وجود ندارد، اگرچه معمولاً $p(x)$ پیوسته است. در این بخش، خود را بیشتر به توابع چندجمله‌ای تکّه‌ای درجهٔ ۳ (مرتبهٔ چهار) محدود می‌کنیم. این عمومیترین مورد درکاربردهاست، و معین بودن مرتبه، بیان مطلب را آسان می‌نماید.

یک راه رده‌بندی مسائل درونیابی چندجمله‌ای تکّه‌ای تقسیم آنها به **موضعی** یا **کلّی** است. برای یک مسألهٔ موضعی، چندجمله‌ای $p(x)$ در هر زیربازهٔ $[x_{i-1}, x_i]$ به وسیلهٔ داده‌های درونیابی، در نقاط گرهی داخل و مجاور $[x_{i-1}, x_i]$، کاملاً معین می‌شود. ولی برای یک مسألهٔ کلّی انتخاب $p(x)$ در هر $[x_{i-1}, x_i]$ به همهٔ داده‌های درونیابی بستگی دارد. مطالعهٔ مسائل کلّی تا اندازه‌ای پیچیده‌تر است؛ بهترین مثالها، توابع بَرازا هستند که بعداً تعریف می‌شوند.

مسائل درونیابی موضعی. فرض کنید که می‌خواهیم $f(x)$ را در بازهٔ $[a, b]$ تقریب بزنیم. با انتخاب یک شبکهٔ

$$a = x_0 < x_1 < \cdots < x_n = b \qquad (۳.۷.۳)$$

که اغلب متساوی‌الفاصله است شروع می‌کنیم. اولین حالت تابع درونیابی چندجمله‌ای تکه‌ای ما بر پایهٔ استفاده از درونیابی چندجمله‌ای معمولی در هر زیربازهٔ $[x_{i-1}, x_i]$ استوار است.

گیریم چهارگره درونیابی در هر زیر بازهٔ $[x_{i-1}, x_i]$ داده شده است،

$$x_{i-1} \le z_{i,1} < z_{i,2} < z_{i,3} < z_{i,4} \le x_i \qquad i = 1, \ldots, n \quad (۴.۷.۳)$$

$p(x)$ را یک چندجمله‌ای از درجهٔ نابزرگتر از ۳ در $x_{i-1} < x < x_i$ تعریف می‌کنیم که $f(x)$ را در $z_{i,1}, \ldots, z_{i,4}$ درونیابی کند. اگر

$$z_{i,1} = x_{i-1} \qquad z_{i,4} = x_i \qquad i = 1, 2, \ldots, n \qquad (۵.۷.۳)$$

آنگاه $p(x)$ در $[a, b]$ پیوسته است. درگره‌های (۴.۷.۳)، این تابع درونیابی را، **تابع چندجمله‌ای تکّه‌ای لاگرانژ** می‌نامیم و آن را با $L_n(x)$ نشان می‌دهیم.

با توجه به فرمول خطای (۸.۱.۳) در درونیابی چندجمله‌ای، فرمول خطا در $L_n(x)$ در رابطهٔ زیر صدق می‌کند

$$f(x) - L_n(x) = \frac{(x - z_{i,1})(x - z_{i,2})(x - z_{i,3})(x - z_{i,4})}{4!} f^{(4)}(\xi_i) \quad (۶.۷.۳)$$

به ازای $x_{i-1} \le x \le x_i$، $x_{i-1} < \xi_i < x_i$ با $i = 1, \ldots, n$. حالت خاص تساوی فاصله‌ها را در نظر بگیرید. گیریم

$$\delta_i = \frac{(x_i - x_{i-1})}{3} \qquad z_{i,j} = x_{i-1} + (j-1)\delta_i \qquad j = 1, 2, 3, 4 \qquad (۷.۷.۳)$$

در این حالت (۷.۵.۳) و (۶.۷.۳) به ازای $i = 1, 2, \ldots, n$ ، نتیجه می‌دهند

$$| f(x) - L_n(x) | \le \frac{\delta_i^4}{24} \cdot \underset{x_{i-1} \le t \le x_i}{\text{Max}} | f^{(4)}(t) | \qquad x_{i-1} < x < x_i \qquad (۸.۷.۳)$$

از این رابطه می‌توان دید، برای آنکه خطا در $a \le x \le b$ در یک سطح باقی بماند، فاصله‌گذاری δ_i باید براساس اندازهٔ مشتق $f^{(4)}(t)$ در نقاط $[x_{i-1}, x_i]$، انتخاب شود. بنابراین اگر تابع $f(x)$ رفتار متغیری در بازهٔ $[a, b]$ داشته باشد، تابع چندجمله‌ای تکّه‌ای $L_n(x)$ را می‌توان با تنظیم مناسب شبکهٔ (۳.۷.۳)، به نحوی انتخاب کرد که همان رفتار را دنبال کند. درونیابی چندجمله‌ای معمولی در $[a, b]$، این انعطاف‌پذیری را ندارد، و این یک دلیل به کاربردن درونیابی چندجمله‌ای تکه‌ای است. برای مواردی که نقاط شبکهٔ متساوی‌الفاصلهٔ (۳.۷.۳) را به‌کار می‌بریم، نتیجهٔ (۸.۷.۳) همگرایی را تضمین می‌نماید، حال آنکه در درونیابی چندجمله‌ای معمولی ممکن است، مانند مثال (۹.۵.۳) رونگ همگرایی برقرار نباشد.

مثال برای $f(x) = e^x$ در $[0, 1]$، فرض کنید می‌خواهیم فاصله‌گذاریها برابر باشند و خطای ماکسیمم از 10^{-6} کوچکتر باشد. با استفاده از $\delta = \delta_i$ در (۸.۷.۳)، می‌خواهیم داشته باشیم:

$$\frac{\delta^4}{24} e \le 10^{-6}$$

$$\delta \le 0.055, \qquad n = \frac{1}{3\delta} \ge 6.12$$

شاید به لحاظ ماهیتِ پایداری (۸.۷.۳)، استفاده از $n = 6$ کافی باشد. در اینجا شش زیربازهٔ درونیابی درجهٔ سوّم داریم.

از نظر نیاز به ذخیره‌سازی درونیابی چندجمله‌ای تکّه‌ای لاگرانژ، با فرض (۷.۷.۳)، باید چهار تکه اطلاعات در هر زیربازهٔ $[x_{i-1}, x_i]$ نگهداری کنیم. این امر درکلّ ذخیره‌سازی به $4n$ جا در حافظه برای $L_n(x)$ و همچنین $n - 1$ جای اضافی برای نقاط شکستگی (۳.۷.۳) نیاز دارد. انتخاب اینکه چگونه اطلاعات نگهداری شوند بستگی به چگونگی محاسبهٔ $L_n(x)$ دارد. اگر

مشتقات $L_n(x)$ خواسته شده باشد، مناسبترین شکل ذخیره‌سازی $L_n(x)$، شکل چندجمله‌ای تیلر در هر زیربازهٔ $[x_{j-1}, x_j]$ است:

$$L_n(x) = a_j + b_j(x - x_{j-1}) + c_j(x - x_{j-1})^2 + d_j(x - x_{j-1})^3$$
$$x_{j-1} \le x \le x_j \qquad (۹.۷.۳)$$

این محاسبه بهتر است به شکل تودرتو انجام گیرد. ضرایب $\{a_j, b_j, c_j, d_j\}$ از شکلهای استانداردِ چندجمله‌ای درونیابی درجهٔ سوم به‌سادگی به‌دست می‌آیند.

یک شکل دوّم تابع درونیابی چندجمله‌ای تکّه‌ای که به‌طور گسترده‌ای از آن استفاده می‌شود، بر پایهٔ چندجمله‌ای درجهٔ ۳ ارمیت (۱۲.۶.۳) ـ (۱۵.۶.۳) است. در هر زیربازهٔ $[x_{i-1}, x_i]$، گیریم $Q_n(x)$، چندجمله‌ای درجهٔ ۳ ارمیت باشد که $f(x)$ را در x_{i-1} و x_i درونیابی می‌کند. تابع $Q_n(x)$ یک چندجمله‌ای درجهٔ سوم تکّه‌ای بر شبکهٔ (۳.۷.۳) است، و چون در شرایط درونیابی، هم $Q_n(x)$ و هم $Q'_n(x)$ در $[a, b]$ پیوسته هستند، پس $Q_n(x)$ معمولاً هموارتر از $L_n(x)$ است.

برای خطای، $Q_n(x)$ در $[x_{i-1}, x_i]$، (۱۵.۶.۳) را به‌کار می‌بریم تا به‌دست آوریم

$$| f(x) - Q_n(x) | \le \frac{h_i^4}{۳۸۴} \cdot \max_{x_{i-1} \le t \le x_i} | f^{(4)}(t) | \qquad x_{i-1} \le x \le x_i \quad (۱۰.۷.۳)$$

با $h_i = x_i - x_{i-1}$، $1 \le i \le n$. وقتی این رابطه با رابطهٔ (۸.۷.۳) برای خطا در $L_n(x)$ مقایسه شود، ممکن است چنین به‌نظر آید که درونیابی درجهٔ ۳ تکه‌ای ارمیت بهتر است. ولی این موضوعی گمراه‌کننده است. برای روشنترشدن موضوع، گیریم شبکهٔ (۳.۷.۳) متساوی‌الفاصله با $x_i - x_{i-1} = h$ به ازای تمام iها باشد. گیریم $L_n(x)$ بر پایهٔ (۳.۷.۳) درست شده باشد و در (۸.۷.۳) داشته باشیم $\delta = h/3$. توجه داشته باشید که $Q_n(x)$ بر پایهٔ $2n + 2$ تکّه از داده‌ها دربارهٔ $f(x)$، یعنی $\{f(x_i), f'(x_i) \mid i = 0, 1, \ldots, n\}$ درست شده است. و $L_n(x)$ بر پایهٔ $3n + 1$ تکّه از داده‌ها دربارهٔ $f(x)$ درست شده است. برای مقایسهٔ خطای $L_{n_1}(x)$ با خطای $Q_{n_2}(x)$ با فرض $n_2 \doteq 1.5 n_1$ این دو را معادل می‌گیریم. در این صورت کرانهای خطا که از (۸.۷.۳) و (۱۰.۷.۳) به‌دست می‌آیند دقیقاً یکی می‌شوند.

چون تفاوتی در خطا وجود ندارد، شکل تابع چندجمله‌ای تکه‌ای بستگی به کاربردی دارد که روش برای آن بکارگرفته می‌شود. درکاربردهای انتگرال‌گیری عددی، تابع تکه‌ای لاگرانژ مناسبترین است؛ همچنین این تابع در حل بعضی معادلات انتگرال تکین، توسط روشهای انتگرال‌گیری حاصلضرب بخش ۶.۵ از فصل پنج به‌کار برده می‌شود. تابع تکه‌ای ارمیت برای حل بعضی مسائل

معادلات دیفرانسیل مفید است. برای مثال، این یک تابع رایج در روش عنصر متناهی، برای حل مسائل مقدار مرزی معادلات دیفرانسیل مرتبه دوّم است؛ استرنگ و فیکس[1] (۱۹۷۳، فصل ۱) را ببینید. مثالهای عددی برای مقایسهٔ $L_n(x)$ و $Q_n(x)$ در جدول‌های ۱۱.۳ و ۱۲.۳، پس از معرفی توابع برازا داده شده‌اند.

توابع برازا. مانند قبل، شبکهٔ

$$a = x_0 < x_1 < \cdots < x_n = b$$

را در نظر می‌گیریم. $s(x)$ را **تابع برازا** از مرتبهٔ $m \geq 1$ می‌نامیم اگر در دو ویژگی زیر صدق کند.

(و-۱) $s(x)$ یک چندجمله‌ای از درجهٔ کوچکتر از m در هر زیربازهٔ $[x_{i-1}, x_i]$ باشد.

(و-۲) $s^{(r)}(x)$ در $[a, b]$ برای $0 \leq r \leq m-2$، پیوسته باشد.

مشتق تابع بر ازای مرتبهٔ m یک تابع برازا از مرتبهٔ $m-1$ است و به همین گونه است برای انتگرالهای آن. اگر پیوستگی در (و-۱) تا $s^{(m-1)}$ ادامه یابد، آنگاه می‌توان ثابت کرد که $s(x)$ یک چندجمله‌ای از درجهٔ نابزرگتر از $m-1$ در $[a, b]$ است (مسألهٔ ۳۳ را ببینید). توابع برازای درجهٔ ۳ (از مرتبهٔ $m = 4$) به دلایل گوناگون رایجترین توابع برازا هستند. اینها توابع هموار برای برازش داده‌ها هستند، و هنگامی که در درونیابی به‌کار برده می‌شوند، رفتار نوسانی را که مشخصهٔ درونیابی چندجمله‌ای با درجهٔ بالاست ندارند. یک دلیل دیگر برای شکلهای خاص درونیابی برای درجهٔ ۳ در (۲۷.۲.۳) و در مسألهٔ ۳۸ داده شده است.

برای مسألهٔ درونیابی، می‌خواهیم یک تابع برازای $s(x)$ از درجهٔ ۳ پیدا کنیم که داشته باشیم.

$$s(x_i) = y_i \qquad i = 0, 1, \ldots, n \qquad (۱۱.۷.۳)$$

بررسی را شروع می‌کنیم با اینکه ببینیم پس از قیدهای (۱۱.۷.۳) چند درجهٔ آزادی برای انتخاب $s(x)$ باقی می‌ماند. تکنیکی که به‌کار برده می‌شود مستقیماً به یک وسیلهٔ عملی برای محاسبهٔ $s(x)$ نمی‌انجامد، ولی بینش بیشتری به‌دست می‌دهد. می‌نویسیم

$$s(x) = a_i + b_i x + c_i x^2 + d_i x^3 \quad x_{i-1} \leq x \leq x_i \quad i = 1, \ldots, n \quad (۱۲.۷.۳)$$

در اینجا $4n$ ضریب مجهول $\{a_i, b_i, c_i, d_i\}$ وجود دارند. قیدهای مربوط به $s(x)$ عبارت‌اند از (۱۱.۷.۳) و محدودیتهای پیوستگی (و-۲)،

$$s^{(j)}(x_i + 0) = s^{(j)}(x_i - 0) \quad i = 1, \ldots, n-1, \quad j = 0, 1, 2 \quad (۱۳.۷.۳)$$

1. Strang and Fix

این قیود بر روی هم تعداد

$$n+۱+۳(n-۱)=۴n-۲$$

شرط ایجاد می‌نمایند. پس از $۴n$ مجهول حداقل دو درجه آزادی در انتخاب ضرایب (۱۲.۷.۳) وجود دارد. باید انتظار داشته باشیم که با تحمیل شرایط اضافی به $s(x)$ بتوانیم یک تابع برازای درونیاب یکتای $s(x)$ به‌دست آوریم.

اکنون روشی برای ساختن $s(x)$ ارائه می‌نماییم. ابتدا نماد زیر را معرفی می‌کنیم

$$M_i=s''(x_i) \qquad i=۰,۱,\ldots,n \qquad (۱۴.۷.۳)$$

چون $s(x)$ در $[x_i,x_{i+۱}]$ از درجهٔ ۳ است، $s''(x)$ خطی است و بنابراین

$$s''(x)=\frac{(x_{i+۱}-x)M_i+(x-x_i)M_{i+۱}}{h_i} \qquad i=۰,۱,\ldots,n-۱ \quad (۱۵.۷.۳)$$

که در آن $h_i=x_{i+۱}-x_i$. با این فرمول، $s''(x)$ در $[x_۰,x_n]$ پیوسته است. دوبار انتگرال می‌گیریم تا به‌دست آوریم

$$s(x)=\frac{(x_{i+۱}-x)^۳M_i+(x-x_i)^۳M_{i+۱}}{۶h_i}+C(x_{i+۱}-x)+D(x-x_i)$$

که C و D دلخواه‌اند. شرط درونیاب (۱۱.۷.۳) ایجاب می‌کند:

$$C=\frac{y_i}{h_i}-\frac{h_iM_i}{۶} \qquad D=\frac{y_{i+۱}}{h_i}-\frac{h_iM_{i+۱}}{۶}$$

$$s(x)=\frac{(x_{i+۱}-x)^۳M_i+(x-x_i)^۳M_{i+۱}}{۶h_i}+\frac{(x_{i+۱}-x)y_i+(x-x_i)y_{i+۱}}{h_i}$$
$$-\frac{h_i}{۶}[(x_{i+۱}-x)M_i+(x-x_i)M_{i+۱}]$$
$$x_i\le x\le x_{i+۱} \qquad ۰\le i\le n-۱ \qquad (۱۶.۷.۳)$$

این فرمول پیوستگی $s(x)$ را در $[a,b]$، و نیز شرط درونیاب (۱۱.۷.۳) را نتیجه می‌دهد. برای تعیین مقادیر ثابت $M_۰,\ldots,M_n$ فرض می‌کنیم $s'(x)$ در $x_۱,\ldots,x_{n-۱}$ پیوسته است

$$\lim_{x\searrow x_i}s'(x)=\lim_{x\nearrow x_i}s'(x) \qquad i=۱,\ldots,n-۱ \qquad (۱۷.۷.۳)$$

در $[x_i, x_{i+1}]$ داریم

$$s'(x) = \frac{-(x_{i+1}-x)^2 M_i + (x-x_i)^2 M_{i+1}}{2h_i} + \frac{y_{i+1}-y_i}{h_i} - \frac{(M_{i+1}-M_i)h_i}{6} \qquad (۱۸.۷.۳)$$

و در $[x_{i-1}, x_i]$ داریم

$$s'(x) = \frac{-(x_i-x)^2 M_{i-1} + (x-x_{i-1})^2 M_i}{2h_{i-1}} + \frac{y_i - y_{i-1}}{h_{i-1}} - \frac{(M_i - M_{i-1})h_{i-1}}{6}$$

با استفاده از (۱۷.۷.۳) و اندکی دستکاری، به ازای $i = 1, \ldots, n-1$ خواهیم داشت:

$$\frac{h_{i-1}}{6} M_{i-1} + \frac{h_i + h_{i-1}}{3} M_i + \frac{h_i}{6} M_{i+1} = \frac{y_{i+1}-y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}} \qquad (۱۹.۷.۳)$$

این روابط $n-1$ معادله برای $n+1$ مجهول M_0، ...، M_n به‌دست می‌دهند. معمولاً دو شرط انتهایی در x_0 و x_n تعیین می‌کنیم تا دو درجه آزادی موجود در (۱۹.۷.۳) را از میان برداریم.

حالت ۱. شرایط مشتق در نقاط انتهایی. می‌خواهیم $s(x)$ در شرایط

$$s'(x_0) = y'_0, \qquad s'(x_n) = y'_n \qquad (۲۰.۷.۳)$$

که y'_0، y'_n ثابتهای داده شده‌اند، صدق کنند. با استفاده از این شرایط همراه با (۱۸.۷.۳)، برای $i = 0$ و $i = n-1$، معادلات اضافی زیر را به‌دست می‌آوریم

$$\frac{h_0}{3} M_0 + \frac{h_0}{6} M_1 = \frac{y_1 - y_0}{h_0} - y'_0$$
$$\frac{h_{n-1}}{6} M_{n-1} + \frac{h_{n-1}}{3} M_n = y'_n - \frac{y_n - y_{n-1}}{h_{n-1}}$$

از ترکیب با (۱۹.۷.۳) یک دستگاه معادلات خطی خواهیم داشت به صورت:

$$AM = D \qquad (۲۱.۷.۳)$$

با

$$D^T = \left[\frac{y_1-y_0}{h_0} - y'_0, \frac{y_2-y_1}{h_1} - \frac{y_1-y_0}{h_0}, \ldots, \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1}-y_{n-2}}{h_{n-2}}, y'_n - \frac{y_n - y_{n-1}}{h_{n-1}}\right]$$

$$M^T = [M_\circ, M_1, \ldots, M_n]$$

$$A = \begin{bmatrix} \frac{h_\circ}{3} & \frac{h_\circ}{6} & \circ & \circ & & \cdots & \circ \\ \frac{h_\circ}{6} & \frac{h_\circ + h_1}{3} & \frac{h_1}{6} & & & & \\ \circ & \frac{h_1}{6} & \frac{h_1 + h_2}{3} & \frac{h_2}{6} & & & \vdots \\ \circ & & & \ddots & & & \\ \vdots & & & & \frac{h_{n-2}}{6} & \frac{h_{n-2}+h_{n-1}}{3} & \frac{h_{n-1}}{6} \\ \circ & & \cdots & & & \frac{h_{n-1}}{6} & \frac{h_{n-1}}{3} \end{bmatrix} \qquad (۲۲.۷.۳)$$

این ماتریس، متقارن، معین مثبت و غالب قطری است، و دستگاه خطی $AM = D$ به‌طور یکتا حل پذیر است. این دستگاه به‌سادگی و به‌سرعت با تقریباً $8n$ عمل حساب حل می‌شود (مطلب مربوط به دستگاههای سه قطری در بخش ۳.۸ را ببینید.)

$s(x)$، تابع برازای درجهٔ ۳ که به‌دست آوردیم، گاهی **درونیاب برازای درجهٔ ۳ کامل** خوانده می‌شود، که آن را با $s_c(x)$ نشان می‌دهیم. تحلیل خطای آن نیاز به بحث دامنه‌داری دارد، ما فقط قضایای دبور[1] (۱۹۷۸، صص ۶۹.۶۸)، را نقل می‌کنیم.

قضیهٔ ۴.۳ گیریم $f(x)$ در $a \le x \le b$، چهاربار پیوسته مشتقپذیر باشد و دنبالهٔ افرازهای زیر داده شده باشد

$$\tau_n : a = x_\circ^{(n)} < x_1^{(n)} < \cdots < x_n^{(n)} = b$$

و تعریف می‌کنیم

$$| \tau_n | = \max_{1 \le i \le n} (x_i^{(n)} - x_{i-1}^{(n)})$$

گیریم $s_{c,n}(x)$ درونیاب برازای درجهٔ ۳ کامل $f(x)$ در افراز τ_n باشد:

$$s_{c,n}(x_i^{(n)}) = f(x_i^{(n)}) \qquad i = \circ, 1, \ldots, n$$
$$s'_{c,n}(a) = f'(a) \qquad s'_{c,n}(b) = f'(b)$$

در این صورت به ازای مقادیر ثابت مناسب c_j،

$$\max_{a \le x \le b} | f^{(j)}(x) - s_{c,n}^{(j)}(x) | \le c_j | \tau_n |^{4-j} \cdot \max_{a \le x \le b} | f^{(4)}(x) | \qquad (۲۳.۷.۳)$$

1. de Boor

برای $j = ۰, ۱, ۲$. با فرض اضافی

$$\sup_n \left\{ \frac{| \tau_n |}{\underset{۱\le i\le n}{\text{Min}} (x_i^{(n)} - x_{i-۱}^{(n)})} \right\} < \infty$$

نتیجهٔ (۲۳.۷.۳) برای $j = ۳$ نیز برقرار است. ثابتهای قابل قبول عبارت‌اند از:

$$c_۰ = \frac{۵}{۳۸۴} \qquad c_۱ = \frac{۱}{۲۴} \qquad c_۲ = \frac{۳}{۸} \qquad (۲۴.۷.۳)$$

برهان برهانهای بسیاری از این قضایا همراه با قضایای دیگری مربوط به $s_{c,n}(x)$ را می‌توان در دبور (۱۹۷۸)، صص ۶۸.۶۹) یافت. ■

اگر در رابطهٔ (۲۳.۷.۳)، بگیریم $j = ۰$، خواهیم دید که برای یک شبکهٔ یکنواخت τ_n، نرخ همگرایی متناسب با $۱/n^۴$ است. این را درونیابی درجهٔ سوّم تکه‌ای لاگرانژ و ارمیت هم داشتیم، ولی ضریب ثابت $c_۰$ تقریباً سه برابر کوچکتر است. پس درونیاب برازای درجهٔ ۳ کامل باید تا اندازه‌ای تقریب برتری باشد، همان‌گونه که نتایج جداول ۱۱.۳ تا ۱۴.۳ این موضوع را نشان می‌دهند.

دلیل دیگر برای استفاده از $s_c(x)$، ویژگی بهینگی زیر است. گیریم $g(x)$ تابع دلخواهی دوبار پیوسته مشتقپذیر در $[a, b]$ باشد، و به علاوه در شرایط درونیابیهای (۱۱.۷.۳) و (۲۰.۷.۳) صدق کند. بنابراین

$$\int_a^b | s''_e(x) |^۲ \, dx \le \int_a^b | g''(x) |^۲ \, dx \qquad (۲۵.۷.۳)$$

که تساوی فقط در حالت $g(x) = s_c(x)$ برقرار است. بنابراین $s_c(x)$، در بین تمام توابعی که در شرایط درونیابی (۱۱.۷.۳) و (۲۰.۷.۳) صدق می‌کنند «کمترین نوسان» را دارد. برای اثبات این قضیّه، گیریم $k(x) = s_c(x) - g(x)$، و می‌نویسیم

$$\begin{aligned} \int_a^b | g''(x) |^۲ \, dx &= \int_a^b | s''_c(x) - k''(x) |^۲ \, dx \\ &= \int_a^b | s''_c(x) |^۲ \, dx - ۲ \int_a^b s''_c(x) k''(x) dx \\ &\quad + \int_a^b | k''(x) |^۲ \, dx \end{aligned}$$

با انتگرالگیری جزءبه‌جزء، و استفاده از شرایط درونیابی و ویژگیهای $s_c(x)$ می‌توانیم نشان دهیم

$$\int_a^b s''_c(x)k''(x)dx = ۰ \qquad (۲۶.۷.۳)$$

و بنابراین

$$\int_a^b \mid g''(x) \mid^۲ dx = \int_a^b \mid s''_c(x) \mid^۲ dx + \int_a^b \mid s''_c(x) - g''(x) \mid^۲ dx \qquad (۲۷.۷.۳)$$

این رابطه (۲۵.۷.۳) را ثابت می‌کند. تساوی در (۲۵.۷.۳) فقط وقتی اتفاق می‌افتد که در $[a,b]$، $s''_c(x) - g''(x) \equiv ۰$ یا هم‌ارز با آن، $s_c(x) - g(x)$ خطی باشد. در این صورت شرایط درونیاب ایجاب می‌نماید $s_c(x) - g(x) \equiv ۰$. ما بحث بیشتر در این موضوع را به مسألهٔ ۳۸ احاله می‌کنیم.

حالت ۲. شرطِ «نبود یک گره». وقتی مشتقات $f'(a)$ و $f'(b)$ در دست نباشند، به شرایط انتهایی دیگری برای $s(x)$ نیاز داریم تا دستگاه معادله‌های (۱۹.۷.۳) کامل شود. این امر با شرط پیوسته بودن $s^{(۳)}(x)$ در نقاط $x_۱$ و $x_{n-۱}$، انجام می‌شود. این شرط هم‌ارز با این است که خواسته شود $s(x)$ یک تابع برازای درجهٔ ۳ با نقاط گرهی $\{x_۰, x_۲, x_۳, \ldots, x_{n-۲}, x_n\}$ باشد، حال آنکه هنوز هم به درونیابی در نقاط گرهی $\{x_۰, x_۱, x_۲, \ldots, x_{n-۱}, x_n\}$ نیاز داریم. این کار دستگاه (۱۹.۷.۳) را به $n-۳$ معادله بدل می‌کند و درونیابی در نقاط $x_۱, x_{n-۱}$ دو معادلهٔ جدید به‌دست می‌دهند (پیداکردن آنها را به مسألهٔ ۳۴ احاله می‌کنیم). مجدداً یک دستگاه خطی سه قطری $AM = D$ به‌دست می‌آوریم، اگرچه ماتریس A بعضی ویژگیهای خوب (۲۲.۷.۳) را نخواهد داشت. تابع برازای به‌دست آمده در اینجا را با $s_{nk}(x)$ نشان می‌دهند که زیرنمایهٔ آن به معنای شرط «نبود یک گره» است. یک تحلیل همگرایی مشابه آنچه در قضیهٔ ۴.۳ آمد، می‌توان برای s_{nk} نوشت. برای این بحث دبور (۱۹۷۸، ص ۲۱۱)، (۱۹۸۵) را ببینید.

روشهای دیگری برای معرفی شرایط نقطهٔ انتهایی، وقتی $f'(a)$ و $f'(b)$ مجهول باشند، وجود دارد. بحث دربارهٔ بعضی از اینها را می‌توان در دبور (۱۹۷۸، ص ۵۶) یافت. ولی، طرح ارائه شده در اینجا ساده‌ترین شکل در کاربرد است و به‌طور گسترده‌ای مورداستفاده قرار می‌گیرد. در حالات خاص بعضی شرایط نقاط انتهایی ساده‌تری وجود دارند که می‌توان از آنها استفاده نمود، یکی از اینها را در مسألهٔ ۳۸ مطرح کرده‌ایم. ولی در حالت کلی، شرایط نقاط انتهایی به شکلی که قبلاً آمد، لازم است، تا نرخ همگرایی را که در قضیهٔ ۴.۳ داده شد تأمین کند.

مثالهای عددی. گیریم $f(x) = tan^{-۱}x$ و $۰ \le x \le ۵$. در جدول ۱۱.۳ خطاهای

$$E_i = \underset{۰ \le x \le ۵}{\text{Max}} \mid f^{(i)}(x) - L_n^{(i)}(x) \mid \qquad i = ۰, ۱, ۲, ۳ \qquad (۲۸.۷.۳)$$

جدول ۱۱.۳ درونیابی درجهٔ سوّم تکه‌ای لاگرانژ: $L_n(x)$

n	E_0	نسبت	E_1	نسبت	E_2	نسبت	E_3	نسبت
۲	$۱٫۲۰E-۲$		$۱٫۲۲E-۱$		$۷٫۸۱E-۱$		۲٫۳۲	
		۳٫۳		۲٫۱		۱٫۵		۱٫۲
۴	$۳٫۶۲E-۳$		$۵٫۸۳E-۲$		$۵٫۲۴E-۱$		۱٫۹۵	
		۱۱٫۴		۶٫۱		۳٫۲		۱٫۶
۸	$۳٫۱۸E-۴$		$۹٫۷۵E-۳$		$۱٫۶۴E-۱$		۱٫۱۹	
		۱۶٫۹		۸٫۱		۳٫۹		۱٫۷
۱۶	$۱٫۸۸E-۵$		$۱٫۱۱E-۳$		$۴٫۲۱E-۲$		۰٫۶۸۲	
		۱۴٫۵		۷٫۳		۳٫۷		۱٫۹
۳۲	$۱٫۳۰E-۶$		$۱٫۶۱E-۴$		$۱٫۱۴E-۲$		۰٫۳۵۹	

جدول ۱۲.۳ درونیابی درجهٔ سوّم تکه‌ای ارمیت: $Q_n(x)$

n	E_0	نسبت	E_1	نسبت	E_2	نسبت	E_3	نسبت
۳	$۲٫۶۴E-۲$		$۵٫۱۸E-۲$		$۴٫۹۲E-۱$		۲٫۰۶	
		۵٫۶		۳٫۰		۲٫۳		۱٫۵
۶	$۴٫۷۳E-۳$		$۱٫۷۴E-۲$		$۲٫۱۴E-۱$		۱٫۳۳	
		۱۶٫۰		۸٫۰		۳٫۶		۱٫۵
۱۲	$۲٫۹۵E-۴$		$۲٫۱۷E-۳$		$۵٫۹۱E-۲$		۰٫۸۹۱	
		۱۳٫۱		۶٫۷		۳٫۶		۱٫۹
۲۴	$۲٫۲۶E-۵$		$۳٫۲۵E-۴$		$۱٫۶۶E-۲$		۰٫۴۷۵	
		۱۶٫۰		۸٫۰		۴٫۰		۲٫۰
۴۸	$۱٫۴۱E-۶$		$۴٫۰۶E-۵$		$۴٫۱۸E-۳$		۰٫۲۴۱	

داده شده‌اند که در آنها $L_n(x)$ تابع درجهٔ ۳ تکه‌ای لاگرانژ است، که $f(x)$ را در نقاط $x_j = a+jh$، $h = (b-a)/n$ و $j = ۰, ۱, \ldots, n$، درونیابی می‌کند. ستونهایی که بالای آنها نرخ همگرایی نوشته شده است، نرخ کاهش خطا را وقتی n دو برابر می‌شود، مشخص می‌نمایند. توجه کنید که نرخ همگرایی $L_n^{(i)}$ به $f^{(i)}$، با $h^{۴-i}$، $i = ۰, ۱, ۲, ۳$، متناسب است. این را می‌توان دقیقاً اثبات کرد و یک راهنمایی برای اثبات آن در مسألهٔ ۳۲ داده شده است.

در جدول ۱۲.۳، خطاهای مشابه برای تابع درجهٔ ۳ تکّه‌ای ارمیت $Q_n(x)$ که $f(x)$ را درونیابی می‌کند داده شده است.

جدول ۱۳.۳ درونیابی برازای درجهٔ سوّم کامل: $s_{c,n}(x)$

n	$E_۰$	نسبت	$E_۱$	نسبت	$E_۲$	نسبت	$E_۳$	نسبت
۶	$۷٫۰۹E-۳$		$۲٫۴۵E-۲$		$۱٫۴۰E-۱$		$۱٫۰۶E۰$	
		۲۱٫۹		۱۰٫۷		۴٫۸		۲٫۶
۱۲	$۳٫۲۴E-۴$		$۲٫۲۸E-۳$		$۲٫۹۰E-۲$		$۴٫۰۹E-۱$	
		۱۰٫۶		۵٫۶		۲٫۹		۱٫۶
۲۴	$۳٫۰۶E-۵$		$۴٫۰۹E-۴$		$۹٫۸۴E-۳$		$۲٫۵۳E-۱$	
		۲۰٫۷		۹٫۷		۴٫۶		۲٫۱
۴۸	$۱٫۴۸E-۶$		$۴٫۲۲E-۵$		$۲٫۱۳E-۳$		$۱٫۲۲E-۱$	
		۱۶٫۴		۸٫۱		۴٫۰		۲٫۰
۹۶	$۹٫۰۴E-۸$		$۵٫۱۹E-۶$		$۵٫۳۰E-۴$		$۶٫۰۹E-۲$	

توجه کنید که در اینجا هم خطاها با

$$\max_{a\le x\le b} \mid f^{(i)}(x) - Q_n^{(i)}(x) \mid \le ch^{۴-i} \qquad i = ۰, ۱, ۲, ۳$$

سازگار است، که باز می‌توان آن را برای بعضی از مقادیر $c > ۰$ ثابت کرد.

همان‌گونه که قبلاً، به دنبال (۱۰.۷.۳) بیان شد، توابع $L_n(x)$ و $Q_m(x)$، $m = ۱٫۵n$، دارای دقّتی قابل مقایسه، در تقریب $f(x)$ هستند، و جداول ۱۱.۳ و ۱۲.۳ این امر را تائید می‌کنند. برعکس، $Q'_m(x)$ در مقایسه با $L'_n(x)$ تقریبی دقیقتر از $f'(x)$ است. در مسألهٔ ۳۲ توضیحی برای آن داده شده است.

در جدول ۱۳.۳ نتایج استفاده از $s_c(x)$ درونیاب برازای درجهٔ سوّم کامل را داده‌ایم. برای مقایسهٔ $L_n(x)$ با $Q_m(x)$ برای مقادیر قابل مقایسهٔ داده‌های $f(x)$، همان تعداد نقاط درونیابی متساوی‌الفاصله را که در $L_n(x)$ به‌کار رفته است به‌کار می‌بریم.

مثال یک مثال آموزندهٔ دیگر انتخاب $f(x) = x^۴$، $۰ \le x \le ۱$، است. در همهٔ فرمولهای درونیابی قبلی، $f^{(۴)}(x)$ به صورت یک ضریب در فرمولهای خطا وجود دارد. چون $f^{(۴)}(x) = ۲۴$، ثابت است، خطا در هر سه شکل درونیابی در رابطهٔ زیر صدق می‌کند

$$\max_{۰\le x\le ۱} \mid x^۴ - f_n(x) \mid = c_j h^{۴-j} \qquad j = ۰, ۱, ۲, ۳ \qquad (۲۹.۷.۳)$$

ثابتهای c_j با شکل درونیابی که به‌کار رفته است تغییر می‌یابند. در محاسبات فعلی، خطاها درست مانند (۲۹.۷.۳) رفتار می‌کنند، و بنابراین وسیلهٔ دیگری برای مقایسهٔ روشها به‌دست می‌دهند. نتایج جدول ۱۴.۳ را فقط برای دقیقترین حالات آورده‌ایم.

جدول ۱۴.۳ مقایسهٔ سه شکل درونیابی درجهٔ سوّم تکه‌یی

روش	E_0	E_1	E_2	E_3
لاگرانژ $n=32$	$1.10E-8$	$6.78E-6$	$2.93E-3$	۰٫۳۷۵
ارمیت $n=48$	$1.18E-8$	$1.74E-6$	$8.68E-4$	۰٫۲۵۰
برازا $n=96$	$7.36E-10$	$2.12E-7$	$1.09E-4$	۰٫۰۶۲۵

این مثالها نشان می‌دهند که درونیابی برازای درجهٔ سوّم در بعضی حالات به‌طور چشمگیری دقیقتر است. این مثالها همچنین نشان می‌دهند که تمام این روشها، از نظر دقّت، احتمالاً مناسب‌اند، و همهٔ آنها با یک نرخ همگرا هستند. بنابراین، تصمیم اینکه کدام یک از روشها به‌کارگرفته شود، به عوامل دیگری که معمولاً در زمینهٔ کاربرد پیش می‌آیند بستگی دارد. ثابت شده است که توابع برازا برای مسائل برازش داده‌ها، و برازش منحنی بسیار مفیدند، و توابع لاگرانژ و ارمیت، به ترتیب، برای تقریب تحلیلی در حل معادلات انتگرالی و دیفرانسیل. تمام این شکلهای تقریب چندجمله‌ای تکه‌یی برای تمام این کاربردها سودمندند و شخص باید براساس نیازهایی که در حل مسألهٔ مورد نظر دارد، شکل تقریب را انتخاب کند.

B- برازاها. یک راه نمایش توابع برازای درجهٔ سوّم در (۱۲.۷.۳)، (۱۳.۷.۳) داده شده است که در آن یک چندجمله‌ای درجهٔ سه در هر زیربازه داده شده است. این روش برای مسائل درونیابی، مانند آنچه در (۱۶.۷.۳) داده شده، کافی است، ولی برای بیشتر کاربردها، روشهای بهتری برای نمایش توابع برازای درجهٔ سه وجود دارد. مانند گذشته، توجّه ما به توابع برازای درجهٔ سه با نقاط گرهی $\{x_0, x_1, \ldots, x_n\}$ است.

تعریف می‌کنیم

$$x_+^r = \begin{cases} 0 & x < 0 \\ x^r & x \geq 0 \end{cases} \qquad (۳۰.۷.۳)$$

این یک تابع برازا از مرتبهٔ $r+1$ است و فقط یک گره $x = 0$ دارد. این تابع را می‌توان برای به‌دست آوردن یک نمایش دوم توابع برازا به‌کار برد. گیریم $s(x)$ یک تابع برازا از مرتبهٔ m با گره‌های

$\{x_0, \ldots, x_n\}$ باشد. آنگاه برای $x_0 \leq x \leq x_n$.

$$s(x) = p_{m-1}(x) + \sum_{j=1}^{n-1} \beta_j (x - x_j)_+^{m-1} \qquad (۳۱.۷.۳)$$

که $p_{m-1}(x)$ یک چند جمله‌ای یکتا انتخابی از درجهٔ نابزرگتر از $m-1$ است و β_1، ...، β_{n-1} ضرایب تعیین شدهٔ یکتا هستند. اثبات این فرمول به مسألهٔ ۳۷ محوّل شده است. اگر این نمایش برای حل مسائل دیگر به‌کار برده شود، چندین اشکال وجود دارند. جدیترین اشکال این است که این نمایش به طرحهای عددی بدوضع می‌انجامد. بدین دلیل، نمایش دیگری از $s(x)$ ارائه می‌دهیم، که ویژگیهای عددی خیلی بهتری دارد. برای ساده‌کردن موضوع، ما فقط برازاهای درجهٔ سه را در نظر می‌گیریم.

با افزایش گره‌های $\{x_0, \ldots, x_n\}$ شروع می‌کنیم. نقاط گرهی دیگری به شکل دلخواه انتخاب می‌کنیم:

$$x_{-3} < x_{-2} < x_{-1} < x_0 \qquad x_n < x_{n+1} < x_{n+2} < x_{n+3} \qquad (۳۲.۷.۳)$$

برای $i = -3, -2, \ldots, n-1$ یک تفاضل منقسم مرتبهٔ چهار

$$B_i(x) = (x_{i+4} - x_i) f_x[x_i, x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4}] \qquad (۳۳.۷.۳)$$

را برای تابع

$$f_x(t) = (t - x)_+^3 \qquad (۳۴.۷.۳)$$

تعریف می‌کنیم. تابع $B_i(x)$ یک B ـ برازا نامیده می‌شود، که مخفف تابع برازای پایه است. به‌عنوان شکل دیگر (۳۳.۷.۳)، فرمول (۵.۲.۳) تفاضلات منقسم را به‌کار می‌بریم و به‌دست می‌آوریم

$$B_i(x) = (x_{i+4} - x_i) \sum_{j=i}^{i+4} \frac{(x_j - x)_+^3}{\Psi'_i(x_j)}$$

$$\Psi_i(x) = (x - x_i)(x - x_{i+1})(x - x_{i+2})(x - x_{i+3})(x - x_{i+4}) \qquad (۳۵.۷.۳)$$

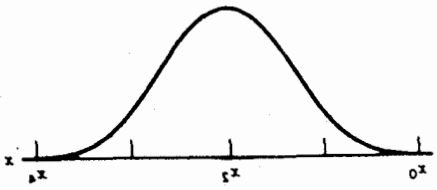
این نشان می‌دهد که $B_i(x)$، یک برازای درجهٔ سه با گره‌های x_i، ...، x_{i+4} است. یک نمودار B ـ برازا در شکل ۷.۳ داده شده است.

بعضی از ویژگیهای مهم B ـ بَرازا را ذیلاً خلاصه می‌کنیم.

قضیهٔ ۵.۳ B ـ برازای درجهٔ سه در روابط زیر صدق می‌کند:

(الف)

$$B_i(x) = 0 \text{، خارج از } x_i < x < x_{i+4} \qquad (۳۶.۷.۳)$$



شکل ۷.۳ B ـ بَرازای $B_0(x)$

(ب)

$$0 \le B_i(x) \le 1 \quad \text{برای همهٔ مقادیر } x. \tag{۳۷.۷.۳}$$

(ج)

$$\sum_{i=-3}^{n-1} B_i(x) = 1 \qquad x_0 \le x \le x_n \tag{۳۸.۷.۳}$$

(د)

$$\int_{x_i}^{x_{i+4}} B_i(x)dx = \frac{(x_{i+4} - x_i)}{4} \tag{۳۹.۷.۳}$$

(ه) اگر $s(x)$ یک تابع برازای درجهٔ سه باگرههای $\{x_0, \ldots, x_n\}$ باشد، آنگاه برای $x_0 \le x \le x_n$

$$s(x) = \sum_{i=-3}^{n-1} \alpha_i B_i(x) \tag{۴۰.۷.۳}$$

که انتخاب $\alpha_{-3}, \ldots, \alpha_{n-1}$ یکتاست.

برهان (الف) برای $x \le x_i$، تابع $f_x(t)$ یک چندجمله‌ای درجهٔ سه در بازهٔ $x_i \le t \le x_{i+4}$ است. بنابراین تفاضل منقسم مرتبهٔ چهار آن صفر است. به ازای $x \ge x_{i+4}$، تابع $f_x(t)$ در بازهٔ $[x_i, x_{i+4}]$ متحد با صفر، و بنابراین $B_i(x) = 0$

(ب) دبور (۱۹۷۸، ص ۱۳۱) را ببینید.

(ج) با استفاده از رابطهٔ بازگشتی برای تفاضلات منقسم

$$B_i(x) = f_x[x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4}] - f_x[x_i, x_{i+1}, x_{i+2}, x_{i+3}] \tag{۴۱.۷.۳}$$

سپس فرض می‌کنیم $x_k \le x \le_{k+1}$. در این صورت $B_{k-3}(x)$، $B_{k-2}(x)$، ...، $B_k(x)$ تنها B ـ برازاهایی هستند که ممکن است در x ناصفر باشند. با استفاده از (۴۱.۷.۳)

$$\sum_{i=-۳}^{n-۱} b_i(x) = \sum_{i=k-۳}^{k} B_i(x)$$
$$= \sum_{i=k-۳}^{k} \left(f_x[x_{i+۱}, x_{i+۲}, x_{i+۳}, x_{i+۴}] - f_x[x_i, x_{i+۱}, x_{i+۲}, x_{i+۳}]\right)$$
$$= f_x[x_{k+۱}, x_{k+۲}, x_{k+۳}, x_{k+۴}] - f_x[x_{k-۳}, x_{k-۲}, x_{k-۱}, x_k]$$
$$= ۱ - ۰ = ۱$$

در مرحلهٔ آخر از این استفاده می‌کنیم که اولاً: $f_x(t)$ در $[x_{k+۱}, x_{k+۴}]$ از درجهٔ ۳ است، پس تفاضل منقسم براساس (۱۸.۲.۳)، برابر ۱ است؛ و ثانیاً: $f_x(t) \equiv ۰$ در $[x_{k-۳}, x_k]$.

(د) دبور (۱۹۷۸، ص ۱۵۱) را ببینید.

(ه) مفهوم B ـ برازاها را شونبرگ[1] پایه‌گذاری کرده، و فرمول (۴۰.۷.۳) از اوست. برای اثبات، دبور (۱۹۷۸، ص ۱۱۳) را ببینید. ■

به‌دلیل (۳۶.۷.۳)، مجموع (۴۰.۷.۳) حداکثر چهار جملهٔ ناصفر دارد. برای $x_k \le x < x_{k+۱}$

$$s(x) = \sum_{i=k-۳}^{k} \alpha_i B_i(x) \qquad (۴۲.۷.۳)$$

به علاوه با استفاده از (۳۷.۷.۳) و (۳۸.۷.۳)،

$$\min\{\alpha_{k-۳}, \alpha_{k-۲}, \alpha_{k-۱}, \alpha_k\} \le s(x) \le \max\{\alpha_{k-۳}, \alpha_{k-۲}, \alpha_{k-۱}, \alpha_k\}$$

که نشان می‌دهد مقدار $s(x)$ با ضرایب B ـ برازاهای نزدیک به x، کراندار شده است. از این جهت، (۴۰.۷.۳) یک نمایش موضعی $s(x)$، برای هر $x \in [x_۰, x_n]$ است.

یک برداشت کلّیتر از B ـ بَرازا و همچنین بعضی ویژگیهایی که اینجا حذف‌شده، در (فصلهای ۹ـ۱۱، دبور، ۱۹۷۸) داده شده است. در آنجا برنامه‌هایی برای محاسبهٔ B ـ برازاها نیز داده شده است.

یک تعمیم مهم بَرازاها وقتی به‌وجود می‌آید که گرهها بتوانند یکی شوند. به‌ویژه، گیریم بعضی از گرهها در (۳۳.۷.۳) یکی شده‌اند. در این حالت، مادام که $x_i < x_{i+۴}$، تابع $B_i(x)$ یک چندجمله‌ای درجهٔ سه تکه‌ای است. اگر دو گره یکی شوند، تعداد مشتقات پیوسته در نقطهٔ گرهی چندگانه، از دو به یک کاهش می‌یابد. اگر سه نقطهٔ گرهی یکی شوند، معنی آن این خواهد بود که $B_i(x)$ فقط پیوسته خواهد بود. با این کار، (۴۰.۷.۳) نمایشی می‌شود برای همهٔ

1. I. J. Schoenberg

چندجمله‌ییهای درجهٔ سه تکه‌ای. در این صورت، تمام توابع چندجمله‌ای تکه‌ای توابع برازا هستند و بالعکس. این امر در دبور (۱۹۷۸) کاملاً تحقیق شده است.

۸.۳ درونیابی مثلثاتی

توابع دوره‌ای یک ردهٔ بسیار مهم از توابع، هستند. تابع $f(t)$ را **دوره‌ای** با دورهٔ تناوب τ خوانند اگر

$$f(t+\tau) = f(t) \qquad -\infty < t < \infty$$

و این رابطه برای هیچ مقدار مثبت کوچکتر از τ درست نباشد. معروفترین توابع دوره‌ای، توابع مثلثاتی هستند. توابع دوره‌ای درکاربردها بسیار فراوانند. این امر توجه ما را به درونیابی مناسبی برای داده‌هایی که از این توابع به‌دست آمده‌اند، توجیه می‌کند. علاوه بر آن، ما از این موضوع برای معرفی تبدیل سریع فوریه، که برای حل بسیاری از مسائل که شامل داده‌های توابع دوره‌ای هستند، استفاده می‌کنیم. با انتخاب یک مقیاس مناسب برای متغیر مستقل همیشه می‌توان دورهٔ نوسان τ را 2π گرفت:

$$f(t+2\pi) = f(t) \qquad -\infty < t < \infty \qquad (۱.۸.۳)$$

این تابعهای $f(t)$ را با استفاده از چندجمله‌ییهای مثلثاتی تقریب می‌زنیم

$$p_n(t) \leftarrows a_0 + \sum_{j=1}^{n} a_j \cos(jt) + b_j \sin(jt) \qquad (۲.۸.۳)$$

اگر $|a_n| + |b_n| \neq 0$، آنگاه این تابع $p_n(t)$ را **چندجمله‌ای مثلثاتی از درجهٔ n** خوانند. می‌توان با استفاده از فرمولهای جمع مثلثاتی، نشان داد که یک صورت هم‌ارز آن به شکل زیر است:

$$p_n(t) = \alpha_0 + \sum_{j=1}^{n} \alpha_j [\cos(t)]^j + \beta_j [\sin(t)]^j \qquad (۳.۸.۳)$$

که تا اندازه‌ای، استفاده از کلمهٔ چندجمله‌ای را برای چنین تابعی توجیه می‌کند. چندجمله‌ای $p_n(t)$ دارای دورهٔ تناوب 2π یا کسر صحیحی از آن است.

در مطالعهٔ مسائل درونیابی با جوابی به صورت $p_n(t)$، باید $2n+1$ شرط درونیابی برقرار کنیم، زیرا $p_n(t)$ شامل $2n+1$ ضریب α_j و β_j است. چون تابع $f(t)$ و چندجمله‌ای $p_n(t)$ دوره‌ای هستند گره‌های درونیابی را، در بازهٔ $0 \leq t < 2\pi$ (یا هم‌ارز آن)، $(-\pi \leq t < \pi$ یا $0 < t \leq 2\pi)$ می‌گیریم. بنابراین گرههای درونیابی را به شکل زیر فرض می‌کنیم،

$$0 \leq t_0 < t_1 < \cdots < t_{2n} < 2\pi \qquad (۴.۸.۳)$$

و می‌خواهیم که $p_n(t)$ به گونه‌ای انتخاب شود که در روابط زیر صدق نماید

$$p_n(t_i) = f(t_i) \qquad i = ۰, ۱, \ldots, ۲n \qquad (۵.۸.۳)$$

بعداً نشان می‌دهیم که این مسأله دارای جوابی یکتاست.

این مسألهٔ درونیابی یک جواب صریح دارد که با فرمول لاگرانژ (۶.۱.۳) در درونیابی چندجمله‌ای، قابل مقایسه است؛ این موضوع در مسألهٔ ۴۱ نشان داده شده است. به جای آنکه از این راه آغاز کنیم، ابتدا (۴.۸.۳) و (۵.۸.۳) را به مسألهٔ هم‌ارز آن که شامل چندجمله‌ییها و توابعی از یک متغیر مختلط است، برمی‌گردانیم. این صورت جدید، بهترین قالب ریاضی برای درونیابی چندجمله‌ییهای مثلثاتی است.

با استفاده از **فرمول اویلر**

$$e^{i\theta} = \cos\theta + i\sin\theta \qquad i = \sqrt{-۱} \qquad (۶.۸.۳)$$

به‌دست می‌آوریم

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{۲} \qquad \sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{۲i} \qquad (۷.۸.۳)$$

با استفاده از این روابط در (۲.۸.۳)، خواهیم داشت

$$p_n(t) = \sum_{j=-n}^{n} c_j e^{ij_t} \qquad (۸.۸.۳)$$

ضرایب با روابط زیر به هم بستگی دارند

$$c_۰ = a_۰ \qquad c_{-j} = \frac{۱}{۲}(a_j - ib_j) \qquad c_j = \frac{۱}{۲}(a_j + ib_j) \qquad ۱ \le j \le n$$

اگر $\{c_j\}$ها داده شده باشند، ضرایب $\{a_j, b_j\}$ به‌سادگی از حل معادلات اخیر به‌دست می‌آیند.

اگر بگیریم $z = e^{it}$، می‌توانیم (۸.۸.۳) را به صورت یک تابع مختلط بنویسیم

$$P_n(z) = \sum_{j=-n}^{n} c_j z^j \qquad (۹.۸.۳)$$

تابع $z^n P_n(z)$ یک چندجمله‌ای از درجهٔ نابزرگتر از $۲n$ است.

برای بیان مسألهٔ درونیابی (۴.۸.۳) و (۵.۸.۳) به صورتی دیگر، گیریم $z_j = e^{it_j}$، $j = ۰, \ldots, ۲n$. با قید (۴.۸.۳)، اعداد z_j نقاط متمایزی روی دایرهٔ واحد $|z| = ۱$ در صفحهٔ

مختلط هستند. مسألهٔ درونیابی چنین است:

$$P_n(z_j) = f(t_j) \qquad j = ۰, \ldots, ۲n \qquad (۱۰.۸.۳)$$

برای دیدن یکتایی جواب این مسأله، توجه نمایید که با تعریف $Q(z) = z^n P_n(z)$ این معادله هم‌ارز است با

$$Q(z_j) = z_j^n f(t_j) \qquad j = ۰, \ldots, ۲n$$

که یک مسألهٔ درونیابی چندجمله‌ای با $۲n+۱$ نقطهٔ گرهی متمایز $z_۰$، ...، $z_{۲n}$ است. قضیهٔ ۱.۳ نشان می‌دهد که یک جواب یکتا وجود دارد. همچنین فرمول (۶.۱.۳) لاگرانژ به $Q(z)$ و از آنجا به $P(z)$ تعمیم می‌یابد.

دلایل چندی، هم نظری و هم عملی، برای برگرداندن درونیابی مثلثاتی به شکل متغیر مختلط وجود دارند. مهمترین آنها از نظر ما این است که درونیابی و تقریب با چندجمله‌ییهای مثلثاتی کاملاً به موضوع توابع مشتقپذیر متغیر مختلط مربوط است، و قسمت بیشتر نظریه از این چشم‌انداز بهتر فهمیده می‌شود. ما این نظریه را بسط نمی‌دهیم، یک بحث کامل آن را می‌توان در هنریچی (۱۹۸۶، فصل ۱۳) و زیگموند (۱۹۵۹، فصل ۱۰) یافت.

درونیابی متساوی‌الفاصله. جالبترین شکل درونیابی در کاربرد، آن است که نقاط شبکهٔ t_j متساوی الفاصله انتخاب شوند. به عبارت دقیقتر، تعریف می‌کنیم

$$t_j = j\frac{۲\pi}{۲n+۱} \qquad j = ۰, \pm۱, \pm۲, \ldots \qquad (۱۱.۸.۳)$$

نقاط $t_۰, \ldots, t_{۲n}$ در (۴.۸.۳) صدق می‌کنند، و نقاط $z_j = e^{it_j}$، $j = ۰, \ldots, ۲n$، نقاط متساوی‌الفاصله روی دایرهٔ واحد $|z| = ۱$ هستند. همچنین توجه داریم که نقاط z_j وقتی j به اندازهٔ $۲n+۱$ افزایش یابد، تکرار می‌شوند.

ما اکنون صورت دیگری از شکل لاگرانژی $p_n(t)$ را، هنگامی که گرههای $\{t_j\}$ در (۱۱.۸.۳) صدق می‌کنند، ارائه می‌دهیم. با لم زیر شروع می‌کنیم.

لم ۴ به ازای همهٔ اعداد صحیح k،

$$\sum_{j=۰}^{۲n} e^{ikt_j} = \begin{cases} ۲n+۱ & e^{it_k} = ۱ \\ ۰ & e^{it_k} \neq ۱ \end{cases} \qquad (۱۲.۸.۳)$$

شرط $e^{it_k} = ۱$ هم‌ارز است با k مضرب صحیحی از $۲n+۱$ است.

برهان گیریم $z = e^{it_k}$، با استفاده از (۱۱.۸.۳)، $e^{ikt_j} = e^{ijt_k}$، و مجموع در (۱۲.۸.۳) چنین می‌شود

$$S = \sum_{j=0}^{2n} z^j$$

اگر $z = 1$، آنگاه $S = 2n + 1$. اگر $z \neq 1$، آنگاه از فرمول سری هندسی (۸.۱.۱) داریم

$$S = \frac{z^{2n+1} - 1}{z - 1}$$

با استفاده از (۱۱.۸.۳)، $z^{2n+1} = e^{2\pi ki} = 1$؛ بنابراین $S = 0$. ■

شرط درونیابی (۱۰.۸.۳) را می‌توان به شکل زیر نوشت

$$\sum_{k=-n}^{n} c_k e^{ikt_j} = f(t_j) \qquad j = 0, 1, \ldots, 2n \qquad (۱۳.۸.۳)$$

برای پیداکردن ضرایب c_k، از لم ۴ استفاده می‌کنیم. معادلهٔ j ام را در e^{-ilt_j} ضرب، سپس نسبت به j جمع می‌کنیم و l را مقیّد می‌کنیم که در $-n \leq l \leq n$ صدق کند. این عمل نتیجه می‌دهد

$$\sum_{j=0}^{2n} \sum_{k=-n}^{n} c_k e^{i(k-l)t_j} = \sum_{j=0}^{2n} e^{-ilt_j} f(t_j) \qquad (۱۴.۸.۳)$$

ترتیب جمع را وارونه و سپس از لم ۴ استفاده می‌کنیم تا به‌دست آوریم

$$\sum_{j=0}^{2n} e^{i(k-l)t_j} = \begin{cases} 0 & k \neq l \\ 2n+1 & k = l \end{cases}$$

با گذاردن این مقدار در (۱۴.۸.۳)، خواهیم داشت

$$c_l = \frac{1}{2n+1} \sum_{j=0}^{2n} e^{-ilt_j} f(t_j) \qquad l = -n, \ldots, n \qquad (۱۵.۸.۳)$$

ضرایب $\{c_n، \ldots، c_{-n}\}$ تبدیل متناهی فوریهٔ داده‌های $\{f(t_0), \ldots, f(t_{2n})\}$ خوانده می‌شوند. این ضرایب یک فرمول صریح برای چندجمله‌ای درونیاب مثلثاتی $sp_n(t)$ (۸.۸.۳) به‌دست می‌دهند. فرمول (۱۵.۸.۳) به ضرایب فوریهٔ $f(t)$ مربوط است:

$$\gamma_l = \frac{1}{2\pi} \int_0^{2\pi} e^{-ilt} f(t) dt \qquad -\infty < l < \infty \qquad (۱۶.۸.۳)$$

اگر از قاعدهٔ انتگرالگیری عددی ذوزنقه‌ای [بخش ۱.۵ را ببینید] برای انتگرالهای فوق استفاده شود، با استفاده از $2n+1$ تقسیم جزیی $[0, 2\pi]$، (۱۵.۸.۳) نتیجه می‌شود، مشروط بر آنکه $f(t)$ در $[0, 2\pi]$ متناوب باشد. ما اکنون در همگرایی $p_n(t)$ به $f(t)$ بحث می‌کنیم.

قضیهٔ ۶.۳ گیریم $f(t)$ یک تابع پیوسته، دوره‌یی و 2π مضربی صحیح از دورهٔ تناوب آن باشد. تعریف می‌کنیم

$$\rho_n(f) = \inf_{\deg(q)\le n} [\operatorname*{Max}_{0\le t\le 2\pi} | f(t) - q(t) |] \qquad (۱۷.۸.۳)$$

که در آن $q(t)$ یک چندجملهٔ مثلثاتی است. آنگاه تابع درونیاب $p_n(t)$ از (۸.۸.۳) و (۱۵.۸.۳) در رابطهٔ زیر صدق می‌کند

$$\operatorname*{Max}_{0\le t\le 2\pi} | f(t) - p_n(t) | \le c[\ln(n+2)]\rho_n(f) \qquad n \ge 0 \qquad (۱۸.۸.۳)$$

ثابت c مستقل از f و n است.

برهان چون اثبات نسبتاً پیچیده است به زیگموند (۱۹۵۹، ص ۱۹ فصل ۱۰) مراجعه کنید. ■

کمیّت $\rho_n(f)$ خطای مینیماکس نامیده می‌شود (فصل ۴ را ببینید)، و از راههای گوناگون می‌توان آن را برآورد کرد. مهمترین قید $\rho_n(f)$ شاید همان قیدی است که جکسن[1] نهاده است. فرض می‌کنیم $f(t)$ در بازهٔ $[0, 2\pi]$، k بار پیوسته مشتقپذیر باشد، $k \ge 0$، و بعلاوه فرض می‌کنیم، به ازای مقداری از $0 < \alpha \le 1$، $f^{(k)}(t)$ در شرط

$$| f^{(k)}(t_1) - f^{(k)}(t_2) | \le c_f | t_1 - t_2 |^{\alpha} \qquad 0 \le t_1, t_2 \le 2\pi$$

صدق می‌کند. (این را شرط هولدر[2] گویند). پس

$$\rho_n(f) \le \frac{c_k(f)}{n^{k+\alpha}} \qquad n \ge 1 \qquad (۱۹.۸.۳)$$

که $c_k(f)$ مستقل از n است. برای اثبات مینارداوس[3] (۱۹۶۷، ص ۵۵) را ببینید.

یک شقّ دیگر فرمول خطای (۱۸.۸.۳) را هنریچی (۱۹۸۶، فرع ۱۳٫۶c) داده است که از ضرایب سری فوریه (۱۶.۸.۳) برای $f(t)$ استفاده کرده است.

---

1. D. Jackson 2. Hölder 3. Meinardus

جدول ۱۵.۳ خطا در درونیابی چندجمله‌ای مثلثاتی

n	E_n	n	E_n
۱	$۵٫۳۹E-۱$	۶	$۶٫۴۶E-۶$
۲	$۹٫۳۱E-۲$	۷	$۴٫۰۱E-۷$
۳	$۱٫۱۰E-۲$	۸	$۲٫۲۲E-۸$
۴	$۱٫۱۱E-۳$	۹	$۱٫۱۰E-۹$
۵	$۹٫۱۱E-۵$	۱۰	$۵٫۰۰E-۱۱$

مثال تقریب زدن $f(t)=e^{\sin(t)}$ را با استفاده از تابع درونیاب $p_n(t)$ در نظر می‌گیریم. خطای ماکسیمم

$$E_n = \underset{۰\le t\le ۲\pi}{\text{Max}} \mid f(t)-p_n(t) \mid$$

به ازای مقادیر مختلفn در جدول ۱۵.۳ داده شده است. همگرایی سریع است.

تبدیل سریع فوریه. تقریب $f(t)$ با $p_n(t)$ در مثال قبل، برای مقادیرکوچک n بسیار دقیق بود. امّا، محاسبهٔ تبدیل متناهی فوریه (۱۵.۸.۳) درکاربردهای دیگر، اغلب نیاز به مقادیر بزرگ n دارد. اکنون روشی را معرفی می‌کنیم که درکاهش هزینهٔ محاسبهٔ $\{c_t\}$، وقتی n بزرگ باشد، بسیار سودمند است.

به جای استفاده از فرمول (۱۵.۸.۳)، فرمول هم‌ارز آن، یعنی

$$d_k = \frac{۱}{m}\sum_{j=۰}^{m-۱} w_m^{jk}.f_j \qquad w_m = e^{-۲\pi i/m} \qquad k=۰,۱,\dots,m-۱ \qquad (۲۰.۸.۳)$$

را با داده‌های مفروض $\{f_۰,\dots,f_{m-۱}\}$ در نظر می‌گیریم. این فرمول را تبدیل متناهی فوریه از مرتبهٔ m می‌نامند. در فرمول (۱۵.۸.۳)، می‌گیریم $m=۲n+۱$. k مجاز است هر عدد صحیحی باشد، با توجه به

$$d_{k+m}=d_k \qquad -\infty<k<\infty \qquad (۲۱.۸.۳)$$

بنابراین کافی است $d_۰,\dots,d_{m-۱}$ یا هر m ضریب پیاپی دیگر d_k را محاسبه کنیم.

برای مقایسهٔ فرمول (۲۰.۸.۳) با شکل دیگری که در زیر خواهد آمد، هزینهٔ محاسبهٔ $d_۰,\dots,d_{m-۱}$ را، با استفاده از (۲۰.۸.۳) حساب می‌کنیم. برای محاسبهٔ d_k، گیریم $z_k=w_m^k$. پس

$$d_k=\frac{۱}{m}\sum_{j=۰}^{m-۱} f_j z_k^j \qquad (۲۲.۸.۳)$$

با استفاده از ضرب تودرتو، در اینجا به $m-1$ ضرب و $m-1$ جمع نیاز داریم. تقسیم بر m را ندیده می‌گیریم زیرا اغلب عوامل دیگری به کاربرده می‌شوند. محاسبهٔ z_k فقط به یک ضرب نیاز دارد زیرا $z_k = w_m z_{k-1}$، $k \geq 2$. هزینهٔ کلی محاسبهٔ $d_0, \ldots, d_{m-1}$ برابر m^2 ضرب و $m(m-1)$ جمع است.

برای معرفی منظور اصلی از تبدیل سریع فوریه، گیریم $m = pq$، که p و q اعدادی صحیح و مثبت بزرگتر از یک هستند. تعریف (۲۰.۸.۳) را به شکل هم‌ارز زیر می‌نویسیم

$$d_k = \frac{1}{p} \sum_{l=0}^{p-1} \frac{1}{q} \sum_{g=0}^{q-1} w_m^{k(l+pg)} f_{l+pg}$$

از $w_m^p = \exp(-2\pi i/q) = w_q$ استفاده می‌کنیم. پس

$$d_k = \frac{1}{p} \sum_{l=0}^{p-1} w_m^{kl} \left[\frac{1}{q} \sum_{g=0}^{q-1} w_q^{kg} f_{l+pg} \right] \qquad k = 0, 1, \ldots, m-1$$

می‌نویسیم

$$e_k^{(l)} = \frac{1}{q} \sum_{g=0}^{q-1} w_q^{kg} f_{l+pg} \qquad 0 \leq l \leq p-1 \qquad (۲۳.۸.۳)$$

$$d_k = \frac{1}{p} \sum_{l=0}^{p-1} w_m^{kl} e_k^{(l)} \qquad 0 \leq k \leq m-1 \qquad (۲۴.۸.۳)$$

وقتی $\{e_k^{(l)}\}$ معلوم شد، هر مقدار d_k، با استفاده از قاعدهٔ ضرب تودرتوی (۲۲.۸.۳)، به p ضرب نیاز خواهد داشت. اگر فرض کنیم همهٔ $e_k^{(l)}$ها قبلاً حساب شده‌اند، به‌دست آوردن (۲۴.۸.۳) به mp ضرب نیاز خواهد داشت. تعدادی در همین حدود هم عمل جمع وجود خواهد داشت.

اکنون نظر خود را به محاسبهٔ $e_k^{(l)}$ معطوف می‌داریم. زیرنمایهٔ k از ۰ تا $m-1$ تغییر می‌کند، ولی به محاسبهٔ همهٔ آنها نیاز نیست. توجه کنید که

$$e_{k+q}^{(l)} = \frac{1}{q} \sum_{g=0}^{q-1} w_q^{(k+q)g} f_{l+pq} = e_k^{(l)}$$

زیرا $w_q^q = 1$. بنابراین فقط لازم است $e_k^{(l)}$ برای $k = 0, 1, \ldots, q-1$ محاسبه شود و سپس این مقادیر خود تکرار می‌شوند. برای هر l، $\{e_0^{(l)}, \ldots, e_q^{(l)}\}$ یک تبدیل متناهی فوریه به ازای داده‌های $\{f_l, f_{l+p}, \ldots, f_{l+p(q-1)}\}$، $0 \leq l \leq p-1$، است. بنابراین محاسبهٔ $\{e_k^{(l)}\}$ به محاسبهٔ p تبدیل متناهی فوریه از مرتبهٔ q (یعنی برای داده‌های به‌طول q) تبدیل می‌شود.

تبدیل سریع فوریه استفادهٔ مکرّر از این نحوهٔ عمل است برای تبدیل محاسبه به تبدیلهای متناهی فوریه با مراتب پایین و پایینتر. برای توضیح بیشتر، فرض کنید r عدد صحیح و $m = ۲^r$. در مرحلهٔ اول، گیریم $p = ۲$ و $q = ۲^{r-۱}$. پس محاسبهٔ (۲۴.۸.۳) به $۲m$ ضرب به اضافهٔ هزینهٔ محاسبهٔ دو تبدیل متناهی فوریه از مرتبهٔ $q = ۲^{r-۱}$، نیاز دارد. برای هر یک از اینها، فرایند را به‌طور بازگشتی تکرار می‌کنیم. در این فرآیند، r ردیف وجود دارد که سرانجام به محاسبهٔ تبدیل متناهی فوریه از مرتبه یک می‌انجامد. تعداد کل ضربها با رابطهٔ زیر داده می‌شود

$$۲m + ۲\left[۲\left(\frac{m}{۲}\right)\right] + ۴\left[۲\left(\frac{m}{۴}\right)\right] + \ldots ۲^r\left[۲\left(\frac{m}{۲^r}\right)\right]$$

که برابر است با

$$۲rm = ۲m \cdot \log_۲ m$$

بنابراین تعداد عملیات متناسب است با $m \cdot \log_۲ m$، در مقابل $m^۲$ عمل در الگوریتم ضرب تودرتوی (۲۲.۸.۳). وقتی m بزرگ، مثلاً $۲^{۱۰}$ یا بیشتر باشد، در زمان صرفه‌جویی هنگفتی می‌شود. برای حالت خاص $m = ۲^r$، یک حساب دقیقتر نشان می‌دهد که درواقع فقط به $m \log_۲ m$ ضرب نیاز است، و یک روند دیگری وجود دارد که فقط به نصف این تعداد نیاز دارد.

برای سایر مقادیر m همان‌گونه که قبلاً ذکر شد، تعمیمهایی از روند قبلی وجود دارند که باز هم نیاز به عملیاتی متناسب با $m \log_۲ m$ دارند، ولی حالت $m = ۲^r$ به بزرگترین صرفه‌جویی می‌انجامد. برای بحث بیشتر در این موضوع، خواننده را به هنریچی (۱۹۸۶، فصل ۱۳) ارجاع می‌دهیم. در آنجا همچنین بحثی دارد در مورد پایداری الگوریتم وقتی خطای گردکردن به حساب آمده باشد. استفاده از تبدیل سریع فوریه انقلابی در بسیاری از موضوعات به وجود آورده است و محاسباتی را که قبلاً عملی نبوده‌اند ممکن ساخته است.

بحث در آثار خواندنی

همان‌گونه که در مقدمه گفته شد، نظریهٔ درونیابی پایه‌ای برای توسعهٔ روشها، در انتگرالگیری عددی و مشتقگیری، نظریهٔ تقریب، و حل عددی معادلات دیفرانسیل است. دربارهٔ هر یک از این موضوعها در فصلهای آتی بحث شده است و نوشتارهای مربوط به آنها در جای خود آورده شده است. مطالب بیشتری از نظریهٔ درونیابی در دبور (۱۹۷۸)، دیویس (۱۹۶۳)، هنریچی (۱۹۸۲، فصل ۵ و۷) و هیلدبرانت (۱۹۵۶) داده شده است. برای گزارشهای تاریخی بسیاری از موضوعات این فصل به گولدشتاین (۱۹۷۷) مراجعه نمایید.

عرضهٔ کامپیوترهای رقمی، انقلابی در آنالیز عددی، از جمله نظریهٔ درونیابی، ایجادنمود. قبل

از استفاده از کامپیوترهای رقمی، محاسبات دستی الزامی بود. بدین‌معنی که از روشهایی عددی استفاده می‌شد که نیاز کمتری به محاسبه داشت. این روشها، غالباً پیچیده‌تر از روشهایی بودند که امروزه در کامپیوتر استفاده می‌شوند، و از ویژگیهای ریاضی منحصر به هر مسأله، بهره‌برداری خاص می‌نمایند. در این روشها از جداول استفادهٔ بسیاری می‌کردند تا از تکرار محاسباتی که توسط دیگران انجام شده، جلوگیری شود؛ فرمولهای درونیابی بر پایهٔ تفاضلات متناهی به‌طور وسیعی به‌کار گرفته می‌شدند. یک مبحث گسترده‌ای به‌نام **حساب تفاضلات متناهی** ابداع شده بود که در حل مسائل در زمینه‌های گوناگون آنالیز عددی و ریاضیات کاربردی به‌کار می‌رفت. برای معرفی کلی این روش در آنالیز عددی هیلد برانت (۱۹۵۹) و مراجعی را که در آن داده شده ملاحظه کنید.

استفاده از کامپیوترهای رقمی نیاز به زمینه‌های دیگر را برای نظریهٔ درونیابی تغییر داده و به‌طور وسیعی نیاز به فرمولهای درونیابی بر پایهٔ تفاضلات متناهی را تقلیل داده است. ولی هنوز یک جای مهم هم برای محاسبات دستی و هم برای استفاده از جداول ریاضی وجود دارد، بویژه برای توابع پیچیده‌تر در فیزیک‌ریاضی. هر کس که با آنالیز عددی کار می‌کند باید یک کتاب مقدماتی از این جداول مثل جدول معروف CRC را داشته باشد. دفتر ملی جدولهای استانداردِ آبراموویچ و استگان (۱۹۶۴) مرجع بسیار عالی برای توابع غیرمقدماتی است. دسترسی به ماشین حسابهای دستی پیشرفته و ریزکامپیوترها، محاسبات دستی (یا شخصی) جدیدی به‌وجود آورده است.

نظریهٔ تقریب چندجمله‌ای تکه‌ای از دههٔ ۱۹۶۰ به بعد بسیار متداول‌شده، و در بسیاری از زمینه‌ها از آن استفاده می‌شود. برای مثال، استرَنگ و فیکس (۱۹۷۳، فصل ۱) را برای کاربرد در حل مسائل مقادیر مرزی در معادلات دیفرانسیل معمولی ببینید، و پاولیدیس[1] (۱۹۸۲، فصل ۱۲.۱۰) را برای کاربرد در گرافیک رایانه‌یی ملاحظه کنید. بیشترین استفاده از توابع چندجمله‌ای تکه‌ای در حول توابع بَرازا دور می‌زند. آغاز نظریهٔ توابع برازا را عمدتاً به شونبرک[2] در مقاله‌های ۱۹۴۶ او نسبت می‌دهند و کارهای او برای کمک به توسعهٔ این موضوع برجسته بوده است (مثلاً شونبرک (۱۹۷۳) را ببینید). اکنون نوشته‌های مفصّلی دربارهٔ توابع بَرازا وجود دارد که شامل کارهای افراد و گروههاست. برای یک مطالعهٔ کلّی، آلبرگ و همکاران[3] (۱۹۶۷) و دبور (۱۹۷۸) و شومیکر[4] (۱۹۸۱) را ببینید. بعضی از معروفترین نرم‌افزارهای رایانه‌یی، برای استفاده از توابع بَرازا، بر پایهٔ برنامه‌های دبور (۱۹۷۸) تهیه شده‌اند. نسخه‌هایی از آنها در کتابخانه‌های آنالیز عددی IMSL و NAG وجود دارند.

تبدیلات متناهی فوریه، درونیابی مثلثاتی و مباحث وابسته به آنها، نسبتاً قدیمی‌اند؛ برای مثال گولدشتاین (۱۹۷۷، ص ۲۳۸) را برای کارهای گاوس در زمینهٔ درونیابی مثلثاتی، ببینید. از آنجا که سری فوریه و تبدیلات فوریه ابزارهای مهمی در ریاضیات کاربردی هستند، جای تعجب نیست

1. Pavlidis 2. Schoenberg 3. Ahlberg et al 4. Schumaker

که توجه بسیار زیادی به تقریبهای خاص آنها وجود دارد. به دنبال مقالهٔ مشهور کولی[1] و تیوکی[2] (۱۹۶۵) دربارهٔ تبدیل سریع فوریه، کاربرد تبدیلات سریع فوریه و موضوعات وابسته به آنها فوق‌العاده افزایش یافته است. به‌عنوان مثال از این راه روشهایی بسیار سریع برای حل معادلات دیفرانسیل با مشتقات جزیی لاپلاس روی ناحیه‌های مستطیلی پیدا شده‌اند، که ما در فصل ۸ بیشتر دربارهٔ آنها بحث خواهیم کرد. برای یک توضیح کلاسیک از درونیابی مقدماتی زیگموند (۱۹۵۹، فصل ۱۰)، و برای یک بازنگری جدید در تمام زمینهٔ آنالیز متناهی فوریه هنریچی (۱۹۷۶، فصل ۱۳) را ببینید.

نظریهٔ درونیابی چندجمله‌ای چند متغیره، زمینه‌ای است که به‌سرعت توسعه می‌یابد، به لحاظ کمبود جا، از این کتاب حذف شده است. در روش عنصر متناهی برای حل معادلات دیفرانسیل با مشتقات جزیی استفادهٔ وسیعی از نظریهٔ درونیابی چند متغیره می‌شود و معرفیهای خوبی از این نظریه درکتابهایی در زمینهٔ روش عنصر متناهی یافت می‌شوند. برای مثال جین[3] (۱۹۸۴)، لاپیدوس[4] و پیندر[5] (۱۹۸۲)، میچل[6] و وِیت[7] (۱۹۷۷) و استرنگ[8] و فیکس[9] (۱۹۷۳) را ببینید. اخیراً، کار درگرافیک رایانه‌یی موجب پیشرفتهای جدیدی شده است [بارن‌هیل[10] (۱۹۷۷) و پاولیدیس (۱۹۸۲، فصل ۱۳) را ببینید].

مراجع

Abramowitz, M., and I. Stegun, eds. (1964). *Handbook of Mathematical Functions*. National Bureau of Standards, Washington, D.C. (Available now from Dover, New York.)

Ahlberg, J., E. Nilson, and J. Walsh (1967). *The Theory of Splines and Their Applications*. Academic Press, New York.

Barnhill, R. (1977). Representation and approximation of surfaces. In *Mathematical Software III*, J. Rice, ed., pp. 69–120. Academic Press, New York.

De Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York.

De Boor, C. (1985). Convergence of cubic spline interpolation with the not-a-knot condition. Math. Res. Ctr. Tech. Rep. 1876, Madison, Wis.

CRC Standard Mathematical Tables. Chem. Rubber Publ. Co., Cleveland. (Published yearly).

Cooley, J., and J. Tukey (1965). An algorithm for the machine calculation of

1. Cooley 2. Tukey 3. Jain 4. Lapidus
5. Pinder 6. Mitchel 7. Wait 8. Strang
9. Fix 10. Barnhill

complex Fourier series. *Math. Comput.*, **19**, 297–301.

Davis, P. (1963). *Interpolation and Approximation*. Ginn (Blaisdell), Boston.

Goldstine, H. (1977). *A History of Numerical Analysis from the 16th through the 19th Century*. Springer-Verlag, New York.

Henrici, P. (1982). *Essentials of Numerical Analysis*. Wiley, New York.

Henrici, P. (1986). *Applied and Computational Complex Analysis*, Vol. 3. Wiley, New York.

Hildebrand, F. (1956). *Introduction to Numerical Analysis*. McGraw-Hill, New York.

Isaacson, E., and H. Keller (1966). *Analysis of Numerical Methods*. Wiley, New York.

Jain, M. (1984). *Numerical Solution of Differential Equations*, 2nd ed., Wiley, New Delhi.

Lapidus, L., and G. Pinder (1982). *Numerical Solution of Partial Differential Equations in Science and Engineering*. Wiley, New York.

Lorentz, G., K. Jetter, and S. Riemenschneider (1983). *Birkhoff Interpolation*. In *Encylopedia of Mathematics and Its Applications*, Vol. 19. Addison-Wesley, Reading, Mass.

Meinardus, G. (1967). *Approximation of Functions: Theory and Numerical Methods* (transl. L. Schumaker). Springer-Verlag, New York.

Mitchell, A., and R. Wait (1977). *The Finite Element Method in Partial Differential Equations*. Wiley, London.

Pavlidis, T. (1982). *Algorithms for Graphics and Image Processing*. Computer Science Press, Rockville, Md.

Powell, M. (1981). *Approximation Theory and Methods*. Cambridge Univ. Press, Cambridge, England.

Schoenberg, I. (1946). Contributions to the approximation of equidistant data by analytic functions. *Quart. Appl. Math*, **4**(Part A), 45–99; (Part B), 112–141.

Schoenberg, I. (1973). *Cardinal Spline Interpolation*. Society for Industrial and Applied Mathematics, Philadelphia.

Schumaker, L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

Stoer, J., and R. Bulirsch (1980). *Introduction to Numerical Analysis*. Springer-Verlag, New York.

Strang, G., and G. Fix (1973). *An Analysis of the Finite Element Method*. Prentice-Hall, Englewood Cliffs, N.J.

Zygmund, A. (1959). *Trigonometric Series*, Vols. 1 and 2. Cambridge Univ. Press, Cambridge.

مسائل

۱. ماتریس واندرموند X را که در (۳.۱.۳) داده شده است یادآوری کرده تعریف می‌کنیم.

$$V_n(x) = \det \begin{bmatrix} ۱ & x_۰ & x_۰^۲ & \dots & x_۰^n \\ \vdots & & & & \vdots \\ ۱ & x_{n-۱} & x_{n-۱}^۲ & & x_{n-۱}^n \\ ۱ & x & x^۲ & \dots & x^n \end{bmatrix}$$

(الف) نشان دهید $V_n(x)$ یک چندجمله‌ای درجهٔ n است، و ریشه‌های آن $x_{n-۱}, \dots, x_۰$ هستند. فرمول زیر را به‌دست آورید

$$V_n(x) = (x - x_۰) \dots (x - x_{n-۱}) V_{n-۱}(x_{n-۱})$$

راهنمایی: آخرین سطر $V_n(x)$ را توسط دترمینان‌های جزئی بسط داده نشان دهید که $V_n(x)$ یک چندجمله‌ای درجهٔ n است و ضریب جملهٔ x^n را پیدا کنید.

(ب) نشان دهید

$$\det(X) \equiv V_n(x_n) = \prod_{۰ \le j < i \le n} (x_i - x_j)$$

۲. برای توابع پایهٔ $l_{j,n}(x)$ که در (۵.۱.۳) داده شده است، ثابت کنید برای هر $n \ge t$

$$\sum_{j=۰}^{n} l_{j,n}(x) = ۱ \qquad \text{برای تمام مقادیر } x$$

۳. توابع لاگرانژ $l_n(x), \dots, l_۰(x)$ را که در (۵.۱.۳) تعریف شده و سپس به حالت کمی متفاوت (۴.۲.۳) با استفاده از رابطهٔ

$$\Psi_n(x) = (x - x_۰) \dots (x - x_n)$$

بازنویس شده است به‌یاد می‌آوریم. گیریم $w_j = [\Psi_n'(x_j)]^{-۱}$. نشان دهید که چندجمله‌ای $p_n(x)$، درونیاب $f(x)$ را می‌توان به شکل زیر نوشت

$$p_n(x) = \frac{\sum_{j=۰}^{n} [w_j f(x_j)]/(x - x_j)}{\sum_{j=۰}^{n} w_j/(x - x_j)}$$

به شرطی که x یک نقطهٔ گرهی نباشد. این عبارت را نمایش **گرانیگاهی** $p_n(x)$ گویند، که به صورت یک مجموع وزن‌دار از مقادیر $\{f(x_0), \ldots, f(x_n)\}$ داده شده است. برای بحث دربارهٔ استفاده از این نمایش، هنریچی (۱۹۸۲،ص ۲۳۷) را ببینید.

۴. درونیابی خطی در جدول مقادیر e^x، $0 \le x \le 2$ با $h = 0.01$ را در نظر می‌گیریم. گیریم که مقادیر جدول با دقت ۵ رقم بامعنی مانند جدول CRC داده شده باشند. کران خطای درونیابی خطی را، شامل قسمتی که به خطای گردکردن در داده‌های جدول مربوط می‌شود، به‌دست آورید.

۵. درونیابی خطی در یک جدول $\cos x$ را که x بر حسب درجه، $0^\circ \le x \le 90^\circ\mathrm{C}$، با طول گام (درجه) $h = 1' = \frac{1}{60}$ داده شده است در نظر می‌گیریم. با فرض اینکه داده‌های جدول با خطای گردکردن تا ۵ رقم بامعنی داده شده‌باشند، کران خطای کلی درونیابی را پیدا کنید.

۶. فرض کنید که می‌خواهید یک جدول مقادیر $\sin x$، $0 \le x \le \pi/2$ با طول گام h بسازید. فرض کنید که در این جدول درونیابی خطی باید به‌کار گرفته شود و خطای کلی، شامل اثرات خطای گردکردن داده‌های جدول، باید حداکثر برابر 10^{-6} باشد. h چقدر باید باشد (آن را در یک اندازهٔ مناسب برای کاربرد واقعی انتخاب کنید) و داده‌های جدول تا چند رقم معنی‌دار باید داده شوند؟

۷. مسألهٔ ۶ را با e^x و $0 \le x \le 1$ حل کنید.

۸. مطالب مربوط به اثر خطای گردکردن در داده‌های جدول بخش ۱.۳ را به درونیابی درجهٔ دوّم تعمیم دهید. گیریم $\varepsilon_i = f(x_i) - \tilde{f}_i$، $i = 0, 1, 2$، و $\mathrm{Max}\{|\varepsilon_0|, |\varepsilon_1|, |\varepsilon_2|\}$ نابزرگتر از ε باشد. نشان دهید که نتیجهٔ این خطاهای گردکردن در خطای درونیابی درجهٔ دوّم و فرض $x_0 \le x \le x_2$ و $x_1 - x_0 = x_2 - x_1 = h$، با 1.25ε کراندار می‌شود.

۹. مسأله ۶ را حل کنید، ولی درونیابی درجهٔ دوّم و نتیجهٔ مسألهٔ ۸ را به‌کار برید.

۱۰. فرض کنید می‌خواهید جدولی از مقادیر $f(x) = \log_{10} x$، $1 \le x \le 10$، بسازید و فرض کنید درونیابی درجهٔ دوّم باید به‌کار گرفته شود و خطای کلی درونیابی، شامل خطای گرد کردن در داده‌های جدول، باید کوچکتر از 10^{-6} باشد. یک گام فاصله‌گذاری h مناسب انتخاب کنید و تعداد ارقامی را که داده‌ها باید داشته باشند معلوم کنید. آیا بهتر نیست که فاصلهٔ h با تغییر x در $[1, 10]$ تغییر کند؟ اگر جواب مثبت است، یک *افراز* مناسب $[1, 10]$ متناظر با مقادیر h را پیشنهاد نمائید. نتیجهٔ مسألهٔ ۸ را برای اثر خطای گردکردن به‌کار برید.

۱۱. گیریم $x_0, \ldots, x_n$ نقاط حقیقی متمایز باشند، و مسألهٔ درونیابی زیر را درنظر می‌گیریم. یک تابع

$$P_n(x) = \sum_{j=0}^{n} c_j e^{jx}$$

انتخاب می‌کنیم به طوری که

$$P_n(x_i) = y_i \qquad i = ۰, ۱, \ldots, n$$

که $\{y_i\}$ها داده‌های معلوم‌اند. نشان دهید که یک انتخاب یکتا برای مقادیر $c_n, \ldots, c_۰$ وجود دارد.

راهنمایی: مسأله را می‌توان به مسألهٔ معمولی درونیابی چندجمله‌ای تبدیل کرد.

۱۲. پیداکردن یک تابع گویای $p(x) = (a + bx)/(۱ + cx)$ را در نظر بگیرید که در رابطهٔ

$$P(x_i) = y_i \qquad i = ۱, ۲, ۳$$

صدق می‌کند و $x_۱$، $x_۲$، $x_۳$ نقاط متمایز هستند. آیا چنین تابع $p(x)$یی وجود دارد؟ یا شرایط اضافی دیگری لازم است تا وجود و یکتایی $p(x)$ را تضمین نماید؟ برای یک نظریهٔ کلّی درونیابی گویا اشتور و بولیرش[1] (۱۹۸۰، ص۵۸) را ببینید.

۱۳. (الف) شکل بازگشتی فرمول درونیابی را که در (۸.۲.۳) داده شده است ثابت کنید.

(ب) فرمول بازگشتی (۷.۲.۳) را برای تفاضلات منقسم ثابت کنید.

۱۴. رابطه‌های (۱۸.۲.۳) مربوط به تفاضل منقسم متغیّر یک چندجمله‌ای را ثابت کنید.

۱۵. ثابت کنید که $\mathrm{Vol}(\tau_n) = ۱/n!$ که τ_n سادکی تعریف‌شده در R^n است که در (۱۴.۲.۳) از قضیهٔ ۳.۳ در بخش ۲.۳، تعریف شده است.

راهنمایی: رابطهٔ (۱۳.۲.۳) و قضایای دیگر در تفاضلات منقسم را همراه با انتخاب خاص $f(x)$، به‌کار برید.

۱۶. گیریم $p_۲(x)$ چندجمله‌ای درجهٔ دوّمی باشد که $f(x)$ را در نقاط متساوی‌الفاصلهٔ $x_۰$، $x_۱ = x_۰ + h$ و $x_۲ = x_۰ + ۲h$ درونیابی می‌کند. فرمولهای خطاهای $f'(x_i) - p'_۲(x_i)$، $i = ۰, ۱, ۲$ را پیدا کنید. فرض کنید $f(x)$ سه بار پیوسته مشتقپذیر باشد، کرانهای قابل محاسبه برای این خطاها را به‌دست آورید.

راهنمایی: فرمول خطای (۱۱.۲.۳) را به‌کار برید.

۱۷. زیر روال رایانه‌یی را برای پیاده کردن الگوریتم‌های Divdif و Interp که در بخش ۲.۳ داده شده بنویسید و سپس یک برنامهٔ اجراکنندهٔ اصلی تهیه کنید که از آن در ساختن جدول درونیابی استفاده شود. جدولی از آبراموویتس و اشتگون (۱۹۶۴) انتخاب و با در نظر گرفتن چندین درجهٔ پیاپی n برای چندجمله‌ای درونیابی، برنامه را آزمایش کنید.

۱۸. یک مسألهٔ درونیابی معکوس را با استفاده از جدول $J_۰(x)$ که در بخش ۲.۳ داده شده

1. Stoer and Bulirsch

انجام دهید. مقداری برای x پیدا کنید که $J_0(x) = 0$، یعنی یک برآورد دقیق برای ریشه پیدا کنید. دقت خود را برآورد کرده و آنرا با جواب واقعی $x = ۲٫۴۰۴۸۲۵۵۵۷۷$ مقایسه کنید.

۱۹. مشابه لم ۱ در بخش ۳.۳ را برای تفاضلات پسرو به‌دست آورید. از این لم و صورت منقسم چندجمله‌ای درونیاب (۹.۲.۳) نیوتن برای پیداکردن فرمول درونیابی تفاضل پسرو (۱۱.۳.۳) استفاده کنید.

۲۰. جدول زیر برای مقادیر

$$j_0(x) = \sqrt{\frac{\pi}{2x}} \cdot J_{1/2}(x)$$

را که از آبراموویتس و اشتگون (۱۹۶۴، فصل ۱۰) استخراج شده در نظر می‌گیریم.

x	$j_0(x)$	x	$j_0(x)$
۰٫۰	۱٫۰۰۰۰۰	۰٫۷	۰٫۹۲۰۳۱
۰٫۱	۰٫۹۹۸۳۳	۰٫۸	۰٫۸۹۶۷۰
۰٫۲	۰٫۹۹۳۳۵	۰٫۹	۰٫۸۷۰۳۶
۰٫۳	۰٫۹۸۵۰۷	۱٫۰	۰٫۸۴۱۴۷
۰٫۴	۰٫۹۷۳۵۵	۱٫۱	۰٫۸۱۰۱۹
۰٫۵	۰٫۹۵۸۸۵	۱٫۲	۰٫۷۷۶۷۰
۰٫۶	۰٫۹۴۱۰۷	۱٫۳	۰٫۷۴۱۲۰

بر پایهٔ خطاهای گردکردن در داده‌های جدول، بیشترین درجهٔ درونیابی چندجمله‌ای که در این جدول استفاده‌شده چه باید باشد؟

راهنمایی: از جدول تفاضلات پیشرو برای پیداکردن نتایج خطای گردکردن استفاده کنید.

۲۱. داده‌های زیر از یک چندجمله‌ای با درجه نابیشتر از ۵ استخراج شده است. درجهٔ چندجمله‌ای چه بوده است؟

x	-۲	-۱	۰	۱	۲	۳
$P(x)$	-۵	۱	۱	۱	۷	۲۵

۲۲. داده‌های زیر نوفه‌ای دارند که نسبت به خطای گردکردن بزرگ است. این نوفه را پیدا کنید و داده‌ها را به طریق مناسب تغییر دهید. فقط مقادیر تابع داده شده‌اند، زیرا نقاط گرهی برای محاسبهٔ

جدول تفاضلات پیشرو لازم نیستند.

۳۰۴۳۱۹	۴۱۹۳۲۷	۵۴۵۸۱۱	۶۸۳۱۰۰
۳۲۶۳۱۳	۴۴۳۶۵۵	۵۷۲۴۳۳	۷۱۱۷۰۹
۳۴۸۸۱۲	۴۶۸۵۲۹	۵۹۹۴۷۵	۷۴۰۷۵۶
۳۷۱۸۰۶	۴۹۳۸۵۲	۶۲۶۹۰۹	۷۷۰۱۸۸
۳۹۵۲۸۵	۵۱۹۶۱۵	۶۵۴۷۹۰	۸۰۰۰۰۰

۲۳. برای $f(x) = 1/(1+x^2)$، $-5 \le x \le 5$، با استفاده از $n+1$ نقطهٔ گرهی متساوی‌الفاصله در $[-5, 5]$، $P_n(x)$ را بسازید. $P_n(x)$ را در تعداد زیادی نقطه محاسبه و نمودار آن یا نمودار خطا را در $[-5, 5]$، مانند شکل ۶.۳، رسم کنید.

۲۴. تابع e^x و تقریب آن توسط یک چندجمله‌ای درونیاب روی $[0, b]$ را در نظر می‌گیریم. برای $n \ge 1$ گیریم $h = b/n$، $x_j = jh$ و $j = 0, 1, \ldots, n$، و گیریم $p_n(x)$ چندجمله‌ای درجهٔ n باشد، که e^x را در نقاط گرهی $x_0, \ldots, x_n$ درونیابی کند. ثابت کنید که وقتی $n \to \infty$،

$$\max_{0 \le x \le b} | e^x - p_n(x) | \to 0$$

راهنمایی: نشان دهید که $| \Psi_n(x) | \le n! h^{n+1}$، $0 \le x \le b$؛ به هر زیربازهٔ $[x_{j-1}, x_j]$ جداگانه نگاه کنید.

۲۵. ثابت کنید که مسألهٔ کلّی (۱۶.۶.۳) ارمیت، در بین چندجمله‌ییهایی از درجهٔ نابزرگتر از $N-1$، یک جواب یکتای $p(x)$ دارد.

راهنمایی: نشان دهید که مسألهٔ همگن برای دستگاه خطی متناظر فقط دارای جواب صفر است.

۲۶. مسألهٔ ارمیت زیر را در نظر می‌گیریم

$$p^{(j)}(x_i) = y_i^{(j)} \qquad i = 1, 2 \qquad j = 0, 1, 2$$

که در آن $p(x)$ یک چندجمله‌ای از درجهٔ نابزرگتر از ۵ است.

(الف) برای $p(x)$ فرمولی از نوع فرمول لاگرانژ به‌دست آورید که تعمیمی از درونیابی درجه سوم (۱۲.۶.۳) ارمیت باشد.

راهنمایی: برای توابع پایه که در $l(x_2) = l'(x_2) = l''(x_2) = 0$ صدق می‌کنند، از $l(x) = (x - x_2)^3 g(x)$، با $g(x)$ از درجهٔ نابزرگتر از ۲ استفاده کنید. $g(x)$ را پیدا کنید.

(ب) یک فرمول تفاضلات منقسم نیوتن پیدا کنید که تعمیم فرمول (۱۳.۶.۳) باشد.
(ج) یک فرمول خطا پیدا کنید که تعمیم فرمول (۱۴.۶.۳) باشد.

۲۷. گیریم $p(x)$ یک چندجمله‌ای باشد که مسألهٔ درونیابی ارمیت زیر را حل کند.

$$p^{(j)}(a) = f^{(j)}(a) \qquad p^{(j)}(b) = f^{(j)}(b) \qquad j = ۰, ۱, \ldots, n-۱$$

وجود آن با استدلال مسألهٔ ۲۵ تضمین می‌شود. فرض کنید که $f(x)$ دارای $۲n$ مشتق پیوسته روی $[a, b]$ باشد، نشان دهید که برای $a \le x \le b$،

$$f(x) - p_n(x) = \frac{(x-a)^n (x-b)^n}{(۲n)!} f^{(۲n)}(\xi_x)$$

با $a \le \xi_x \le b$.

راهنمایی: استدلال در قضیهٔ ۲.۳ را تعمیم دهید.

۲۸. (الف) یک چندجمله‌ای $p(x)$ از درجهٔ نابزرگتر از ۲ پیدا کنید که در

$$p(x_۰) = y_۰ \qquad p'(x_۰) = y'_۰ \qquad p'(x_۱) = y'_۱$$

صدق کند. فرمولی به شکل زیر به‌دست آورید

$$p(x) = y_۰ l_۰(x) + y'_۰ l_۱(x) + y'_۱ l_۲(x)$$

(ب) فرمولی برای مسألهٔ درونیابی چندجمله‌ای زیر بیابید.
گیریم $x_i = x_۰ + ih$، $i = ۰, ۱, ۲$، یک چندجمله‌ای $p(x)$ از درجهٔ نابزرگتر از ۴ بیابید که برای آن

$$p(x_i) = y_i \qquad i = ۰, ۱, ۲$$
$$P'(x_۰) = y'_۰ \qquad p'(x_۲) = y'_۲$$

که مقادیر y داده شده‌اند.

۲۹. مسألهٔ پیداکردن یک چندجمله‌ای درجهٔ دو $p(x)$ که برای آن

$$p(x_۰) = y_۰ \qquad p'(x_۱) = y'_۱ \qquad p(x_۲) = y_۲$$

را در نظر بگیرید به‌طوری که $x_۰ \ne x_۲$ و $\{y_۰, y'_۱, y'_۲\}$ داده‌های معلوم باشند. با فرض حقیقی‌بودن گره‌های $x_۰$، $x_۱$، $x_۲$ چه شرایطی باید برقرار باشد تا چنین $p(x)$ای موجود و یکتا باشد؟ این مسأله، مسألهٔ ۲۸ (الف) و مسألهٔ بعدی مثالهایی از مسائل **درونیابی ارمیت ـ برکاف**[1] هستند [لورنتس و همکاران (۱۹۸۳) را ببینید].

1. Hermite-Birkhoff

۳۰. (الف) نشان دهید که یک چندجمله‌ای درجهٔ سه‌ی یکتای $p(x)$ وجود دارد که برای آن

$$p(x_0) = f(x_0) \qquad p(x_2) = f(x_2)$$
$$p'(x_1) = f'(x_1) \qquad p''(x_1) = f''(x_1)$$

که $f(x)$ یک تابع داده شده است و $x_2 \neq x_0$. فرمولی برای $p(x)$ به‌دست آورید.
(ب) گیریم $x_0 = -1$، $x_1 = 0$، $x_2 = 1$، با فرض اینکه $f(x)$ روی $[-1, 1]$ چهاربار پیوسته مشتقپذیر باشد، نشان دهید که به ازای $-1 \leq x \leq 1$

$$f(x) - p(x) = \frac{x^4 - 1}{4!} f^{(4)}(\xi_x)$$

برای مقداری $\xi_x \in [-1, 1]$.

راهنمایی: از برهان قضیه ۲.۳ پیروی کنید.

۳۱. برای تابع $f(x) = \sin x$، $0 \leq x \leq \pi/2$، تابع درجهٔ سه‌ی تکه‌ای $Q_n(x)$ ارمیت و تابع درجهٔ سه‌ی تکه‌ای $L_m(x)$ لاگرانژ را برای $n = 3, 6, 12$، $m = \frac{2}{3}n$ پیدا کنید. خطاهای ماکسیمم $f(x) - L_m(x)$، $f'(x) - L'_m(x)$، $f(x) - Q_n(x)$ و $f'(x) - Q'_n(x)$ را محاسبه کنید. این را می‌توان با دقت قابل قبول در محاسبهٔ خطاها در $8n$ نقطهٔ متساوی‌الفاصله در $[0, \pi/2]$ به‌دست آورد.

۳۲. (الف) گیریم $p_3(x)$ یک چندجمله‌ای درجهٔ سه باشد که $f(x)$ را در نقاط متساوی‌الفاصلهٔ $x_j = x_0 + jh$، $j = 0, 1, 2, 3$، درونیابی می‌کند. با فرض اینکه $f(x)$ به اندازهٔ کافی مشتقپذیر است، کران خطای استفاده از $p'_3(x)$ را به‌عنوان تقریبی برای $f'(x)$، $x_0 \leq x \leq x_1$، پیدا کنید.

(ب) گیریم $H_2(x)$ چندجمله‌ای درجهٔ سه‌ی ارمیت باشد که $f(x)$ و $f'(x)$ را در x_0 و $x_1 = x_0 + h$ درونیابی می‌کند. کران خطای $f'(x) - H'_2(x)$ را برای $x_0 \leq x \leq x_1$ پیدا کنید.

(ج) توابع چندجمله‌ای تکه‌ای $L_m(x)$ و $Q_n(x)$ بخش ۷.۳ را در نظر بگیرید، $m = \frac{2n}{3}$، و کران خطاهای $f'(x) - L'_m(x)$ و $f'(x) - Q'_n(x)$ را پیدا کنید. آن را برای حالت خاص $f(x) = x^4$ به‌کار برید و جوابهای خود را با نتایج عددی که در جدول ۱۴.۳ داده شده‌اند مقایسه کنید.

۳۳. گیریم $s(x)$ یک تابع برازا از مرتبهٔ m باشد. گیریم b یک گره و $s(x)$ یک چندجمله‌ای از درجهٔ نابزرگتر از $m - 1$ روی $[a, b]$ و $[b, c]$ باشد. نشان دهید که اگر $s^{(m-1)}(x)$ در $x = b$ پیوسته باشد، آنگاه $s(x)$ برای $a \leq x \leq c$ یک چندجمله‌ای از درجهٔ نابزرگتر از $m - 1$ است.

۳۴. شرایطی را برای برازای درونیاب درجهٔ سه‌ی $s(x)$ پیدا کنید که از شرایط «نه یک گره»ی نقاط انتهایی نتیجه می‌شوند. به بحث حالت (۲) پس از (۲۷.۷.۳) رجوع کنید.

۳۵. پیداکردن تابع درونیاب برازای درجهٔ ۳ برای داده‌های

x	۰	۱	۲	۲٫۵	۳	۴
y	۱٫۴	۰٫۶	۱٫۰	۰٫۶۵	۰٫۶	۱٫۰

را در نظر می‌گیریم. شرط «نه یک گره» را برای به‌دست آورن شرایط مرزی مکمل (۱۹.۷.۳) به‌دست آورید. تابع نتیجهٔ $s(x)$ را رسم کنید. آن را با درونیابی خطی تکه‌ای، که نقاط پشت سرهم (x_i, y_i) را با پاره‌خطها به‌هم متصل می‌کند، مقایسه کنید.

۳۶. برنامه‌ای بنویسید که نرخ همگرایی درونیابی برازای درجهٔ ۳ (مانند جدول ۱۳.۳) را با شرایط مرزی گوناگون، بررسی کند. بسیاری از مراکز رایانه‌یی بسته‌هایی برای تولید این درونیابها دارند که شرائط مرزی را مشخص می‌کنند. اگر این بسته را ندارید، یک دستگاه خطی حلاّل با (۱۹.۷.۳) و شرایط مرزی اضافی به‌کار برید. شرایط مرزی زیر را بررسی کنید؛ (الف) مشتقات مانند (۲۰.۷.۳) داده شده‌اند؛ (ب) شرط «نه یک گره» که در مسألهٔ ۳۸ داده شده است و (ج) شرایط برازای طبیعی $M_0 = M_n = 0$ در مسألهٔ ۳۸. این برنامه را برای مطالعهٔ همگرایی $s(x)$ به $f(x)$ در حالات زیر به‌کار برید: (۱) $f(x) = e^x$ در $[0, 1]$، (۲) $f(x) = \sin(x)$ در $[0, \pi/2]$، و (۳) $f(x) = x\sqrt{x}$ در $[0, 1]$. به رفتار خطا نزدیک نقاط انتهایی توجه کنید و آنها را مقایسه نمایید.

۳۷. **(الف)** گیریم $q(x)$ یک برازای درجهٔ ۳ با تنها یک گره $x = a$ باشد. به علاوه فرض کنید که برای $x \le a$، $q(x) \equiv 0$. نشان دهید که برای مقداری از c داریم $q(x) = c(x-a)_+^3$.

راهنمایی: برای $x \ge a$ چندجمله‌ای درجهٔ سه‌ی $s(x)$ را به شکل زیر بنویسید

$$s(x) \equiv c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3$$

سپس مفروضات دربارهٔ $s(x)$ را به‌کار برید.

(ب) با استفاده از قسمت (الف)، رابطهٔ (۳۱.۷.۳) را برای توابع برازای درجهٔ سوّم ($m = 4$) ثابت کنید.

۳۸. محاسبهٔ یک برازای درونیاب درجهٔ سه را براساس (۱۹.۷.۳) با شرایط اضافی زیر در نظر می‌گیریم

$$s''(x_0) = M_0 = 0 \qquad s''(x_n) = M_n = 0$$

این مسأله دارای جواب یکتای $\hat{s}(x)$ است. نشان دهید

$$\int_{x_0}^{x_n} [\hat{s}''(x)]^2 dx \leq \int_{x_0}^{x_n} [g''(x)]^2 dx$$

که در آن $g(x)$ یک تابع دوبار پیوسته مشتقپذیر است که در شرایط درونیاب $g(x_i) = y_i$، $i = 0, 1, \ldots, n$ صدق می‌کند.

راهنمایی: نشان دهید که (۲۷.۷.۳) برای $\hat{s}(x)$ معتبر است. [این برازای درونیاب را **برازای درونیاب درجهٔ سه طبیعی** گویند. این یک درونیاب هموار برای داده‌هاست، ولی معمولاً نزدیک نقاط انتهایی به کندی همگراست. برای اطلاعات بیشتر دربارهٔ این مطلب به دبور (۱۹۷۸، ص ۵۵) مراجعه کنید.]

۳۹. برای تعریف یک B-برازای از مرتبهٔ m که در بازهٔ $[x_i, x_{i+m}]$ ناصفر است، تعریف می‌کنیم

$$B_i^{(m)}(x) = (x_{i+m} - x_i) f_x[x_i, x_{i+1}, \ldots, x_{i+m}]$$

با $f_x(t) = (t - x)_+^{m-1}$. یک رابطهٔ بازگشتی برای $B_i^{(m)}(x)$ برحسب B-برازاهای مرتبهٔ $m - 1$ پیدا کنید.

۴۰. با استفاده از تعریف (۳۳.۷.۳) رفتار B-برازاهای درجهٔ سه را وقتی گرهها مجاز باشند بر هم منطبق شوند، تحقیق کنید. نشان دهید که اگر دوتا از گره‌ها در $\{x_i, x_{i+1}, \ldots, x_{i+4}\}$ بر هم منطبق شوند، $B_i(x)$ فقط یک مشتق پیوسته در گره منطبق شده دارد. همچنین، اگر سه گره بر هم منطبق شوند، نشان دهید که $B_i(x)$ در آن نقطه پیوسته است ولی مشتقپذیر نیست.

۴۱. گیریم $0 \leq t_0 < t_1 < \cdots < t_{2n} < 2\pi$، و مسألهٔ درونیابی چندجمله‌ای مثلثاتی (۵.۸.۳) را در نظر می‌گیریم. تعریف می‌کنیم

$$l_j(t) = \Pi_{\substack{k=0 \\ k \neq j}}^{2n} \frac{\sin \frac{1}{2}(t - t_k)}{\sin \frac{1}{2}(t_j - t_k)}$$

برای $j = 0, 1, \ldots, 2n$، $l_j(t_i) = \delta_{ij}$ به‌سادگی معلوم می‌شود که $0 \leq i, j \leq 2n$. نشان دهید $l_j(t)$ یک چندجمله‌ای مثلثاتی از درجهٔ نابزرگتر از n است. در این صورت جواب (۵.۸.۳) با رابطهٔ زیر داده می‌شود

$$P_n(t) = \sum_{j=0}^{2n} f(t_j) l_j(t)$$

راهنمایی: استقرا بر n را همراه با اتحادهای مثلثاتی معمولی به‌کار برید.

۴۲. (الف) فرمولهای زیر را ثابت کنید:

$$\sum_{j=1}^{m-1} \sin\left(\frac{2\pi jk}{m}\right) = 0, \quad k \text{ برای همهٔ اعدادصحیح } m \geq 2$$

$$\sum_{j=0}^{m-1} \cos\left(\frac{2\pi jk}{m}\right) = \begin{cases} m & \text{k مضربی از m باشد} \\ 0 & \text{k مضربی از m نباشد} \end{cases}$$

(ب) با استفاده از فرمولهای فوق فرمولهایی برای عبارات زیر بیابید

$$\sum_{j=0}^{m-1} \cos\left(\frac{2\pi jk}{m}\right)\cos\left(\frac{2\pi jl}{m}\right) \qquad \sum_{j=1}^{m-1} \sin\left(\frac{2\pi jk}{m}\right)\sin\left(\frac{2\pi jl}{m}\right)$$

$$\sum_{j=0}^{m-1} \cos\left(\frac{2\pi jk}{m}\right)\sin\left(\frac{2\pi jl}{m}\right)$$

به‌ازای $0 \leq k$ و $l \leq m-1$. به فرمولهای به‌دست آمده **روابط متعامد گسسته** گویند و نظیر روابط متعامد انتگرالی برای $\{\cos(kx)$ و $\sin(kx)\}$ هستند.

۴۳. تبدیل متناهی فوریه از مرتبهٔ m را برای دنباله‌های زیر پیدا کنید

(الف) $x_k = 1 \qquad 0 \leq k \leq m-1$

(ب) $x_k = (-1)^k \qquad 0 \leq k \leq m-1 \qquad m$ زوج

(ج) $x_k = k \qquad 0 \leq k \leq m-1$

۴

تقریب توابع

برای تعیین مقدار اغلب توابع ریاضی ابتدا باید تقریبهای محاسبه‌پذیری برای آنها تهیّه کنیم. توابع در کاربردها از راههای گوناگونی تعریف می‌شوند، انتگرال‌ها و سری‌های نامتناهی معمولیترین فرمول‌هایی هستند که در این تعریفها به‌کار می‌روند. این گونه تعریفها برای مشخص‌کردن ویژگیهای تابع سودمندند، ولی معمولاً راه کارآمدی برای تعیین مقدار تابع نیستند. در این فصل ما استفاده از چندجمله‌ییها را در تقریب یک تابع داده‌شده بررسی می‌کنیم. ابزارهای مختلف تهیهٔ تقریبهای چندجمله‌یی را توضیح می‌دهیم و آنها را برحسب دقت نسبی‌شان باهم مقایسه می‌کنیم.

برای محاسبهٔ یک تابع $f(x)$ در رایانه، معمولاً داشتن یک تقریب تحلیلی $f(x)$، از لحاظ مکان و زمان، کاراتر از ذخیره‌کردن یک جدول و استفاده از درونیابی است. همچنین بهتر است، در بین چندجمله‌ییهایی که دقت مطلوب در تقریب $f(x)$ را به‌دست می‌دهند، پایینترین درجهٔ ممکن به‌کار برده شود. بخشهای زیر برخی از روشهای تولید تقریب را به‌دست می‌دهند، و معمولاً تقریبهای بهتر آنهایی هستند که تهیّهٔ آنهادشوارتر است. مقدار زمان و کوششی که برای تهیّهٔ یک تقریب به‌کار می‌رود باید با مقدار استفاده از تقریب نسبت مستقیم داشته باشد. اگر فقط چند مرتبه به‌کار رود اغلب یک سری بریده شدهٔ تیلر کافی است. ولی اگر تقریب را باید میلیونها بار افراد زیادی به‌کار برند، آنگاه دقت زیادی در تهیهٔ آن تقریب باید به عمل آید.

غیر از چندجمله‌ییها شکلهای دیگری برای تقریب توابع وجود دارند. توابع گویا خارج‌قسمتهای چندجمله‌ییها هستند، و معمولاً تا حدّی صورتهای تقریب کاراتری هستند. ولی از آنجاکه چندجمله‌ییها شکل مناسب و کارای تقریب را در اختیار می‌گذارند، و به دلیل اینکه نظریهٔ تقریب توابع گویا پیچیده‌تر از نظریّهٔ تقریب چندجمله‌ییهاست، ما برای بررسی فقط چندجمله‌ییها را انتخاب کرده‌ایم. قضایای این فصل را می‌توان در پیداکردن تقریبهای چندجمله‌ییهای تکّه‌یی، که تا حدّی مشابه توابع درونیاب چندجمله‌ییهای تکّه‌یی بخش ۷.۳ فصل قبل هستند، به‌کار برد.

۱.۴ قضیهٔ وایرشتراس و قضیهٔ تیلر

برای توجیه استفاده از چندجمله‌ییها در تقریب توابع پیوسته، قضیّهٔ زیر را معرفی می‌کنیم.

قضیهٔ ۱.۴ (وایرشتراس) گیریم $f(x)$ برای $a \le x \le b$ پیوسته باشد و $\epsilon > 0$. در این صورت یک چندجمله‌یی $p(x)$ وجود دارد که برای آن

$$| f(x) - p(x) | \le \epsilon \qquad a \le x \le b$$

برهان برهانهای زیادی برای این قضیّه و تعمیمهای آن وجود دارد. چون این امر برای مطالعهٔ آنالیز عددی، نکتهٔ مهمّی نیست، ما فقط به یک برهان ساختنی اشاره می‌کنیم. برای برهانهای دیگر، فصل ۶، دیویس (۱۹۶۳) را ببینید.

برای سادگی فرض کنید، $[a, b] = [0, 1]$: با یک تعویض متغیر مناسب، همیشه می‌توان در صورت لزوم بازه را به این حالت تبدیل کرد. تعریف می‌کنیم

$$p_n(x) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k} \qquad 0 \le x \le 1$$

گیریم $f(x)$ بر $[0, 1]$ کراندار باشد. پس در هر نقطهٔ x که f پیوسته باشد

$$\lim_{n\to\infty} p_n(x) = f(x)$$

اگر $f(x)$ در هر نقطهٔ x از $[0, 1]$ پیوسته باشد، همگرایی p_n به f بر $[0, 1]$ یکنواخت است، یعنی

$$\max_{0\le x\le 1} | f(x) - p_n(x) | \to 0 \qquad \text{وقتی } n \to \infty \qquad (1.1.4)$$

این مطلب برای پیداکردن یک چندجمله‌یی که در حکم قضیه صدق کند، راه روشنی به‌دست می‌دهد. اثبات این نتایج را همراه با ویژگیهای دیگری از چندجمله‌ییهای تقریب $p_n(x)$ که

چندجمله‌یی برنشتاین نامیده می‌شوند، می‌توان در صفحات ۱۰۸ تا ۱۱۸ دیویس (۱۹۶۳)، یافت. این چندجمله‌ییها از لحاظ رفتار کیفی شباهت خیلی زیادی به تابع $f(x)$ دارند. برای مثال، اگر $f(x)$، r بار پیوسته مشتقپذیر بر $[۰, ۱]$ باشد، آنگاه

$$\underset{۰\le x\le ۱}{\text{Max}} \mid f^{(r)}(x) - p_n^{(r)}(x) \mid \to ۰ \qquad \text{وقتی} \qquad n \to \infty$$

گرچه یک چنین ویژگی کلّی تقریب در حد خودش باارزش است ولی در این حالت، همگرایی (۱.۱.۴) عموماً خیلی کند است. برای مثال، اگر $f(x) = x^۲$، آنگاه

$$\lim_{n\to\infty} n[p_n(x) - f(x)] = x(۱ - x)$$

و بنابراین برای مقادیر بزرگ n

$$p_n(x) - x^۲ \doteq \frac{۱}{n} x(۱ - x)$$

حتی برای تقریب حالت نمایانی چون $f(x) = x^۲$، خطا به سرعت کاهش نمی‌یابد. ■

قضیهٔ تیلر قضیهٔ تیلر قبلاً، در قضیهٔ ۴.۱، بخش ۱.۱ از فصل ۱ ارائه گردید. این اولین ابزار مهم برای تقریب زدن یک تابع است، و اغلب به عنوان یک تقریب اولیه در محاسبهٔ تقریبهای کاراتر به‌کار می‌رود. برای اینکه ببینیم چرا تقریب تیلر آنقدرها کارا نیست، مثال زیر را در نظر می‌گیریم.

مثال خطای تقریب e^x را با استفاده از چندجمله‌یی $p_۳(x)$ درجهٔ سوم تیلر بر بازهٔ $[-۱, ۱]$، با بسط در حول $x = ۰$ پیدا کنید.

$$p_۳(x) = ۱ + x + \frac{۱}{۲}x^۲ + \frac{۱}{۶}x^۳$$

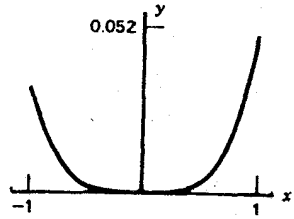
$$e^x - p_۳(x) = R_۴(x) = \frac{۱}{۲۴}x^۴ e^{\xi} \qquad (۲.۱.۴)$$

با ξ بین ۰ و x.

برای اینکه خطا را با دقت آزمایش کنیم، آن را از بالا و پایین محدود می‌کنیم:

$$\frac{۱}{۲۴}x^۴ \le e^x - p_۳(x) \le \frac{e}{۲۴}x^۴ \qquad ۰ \le x \le ۱$$

$$\frac{e^{-۱}}{۲۴}x^۴ \le e^x - p_۳(x) \le \frac{۱}{۲۴}x^۴ \qquad -۱ \le x \le ۰$$



شکل ۱.۴ منحنی خطا برای $p_۲(x) \doteq e^x$.

خطا با افزایش $| x |$، افزایش می‌یابد، و با محاسبهٔ مستقیم خواهیم داشت:

$$\underset{-۱\leq x\leq ۰}{\text{Max}} \mid e^x - P_۲ \mid \doteq ۰٫۰۵۱۶ \qquad (۳.۱.۴)$$

این خطا در بازهٔ $[۱, ۱-]$ به‌طور هموار توزیع نشده است (شکل ۱.۴ را ببینید). خطا در نزدیکی مبدأ بسیار کوچکتر از خطا نزدیک به نقاط انتهایی ۱− و ۱+ است. این توزیع نامساوی خطا، که نوعاً معرّف باقیماندهٔ تیلر است، بدین معناست که معمولاً چندجمله‌ییهای تقریب بسیار بهتری از همان درجه یافت می‌شوند. مثالهای بیشتری در بخش بعد داده شده است.

فضای تابعی $C[a, b]$ مجموعهٔ $C[a, b]$ ی توابع پیوسته و حقیقی مقدار بر بازهٔ $[a, b]$ در بخش ۱.۱ از فصل ۱ معرفی شده است. در این مجموعه معمولاً نُرم زیر را به‌کار می‌بریم

$$\| f \|_\infty = \underset{a\leq x\leq b}{\text{Max}} \mid f(x) \mid \qquad f \in C[a, b] \qquad (۴.۱.۴)$$

این نرم به نامهای مختلفِ **نرم ماکسیمم**، نرم چبیشف، نرم بینهایت، و نرم یکنواخت نیز نامیده شده است. این اندازهٔ طبیعی برای استفاده در نظریّهٔ تقریب است، زیرا ما می‌خواهیم

$$\|f - p\|_\infty = \max_{a\leq x\leq b} \mid f(x) - p(x) \mid \qquad (۵.۱.۴)$$

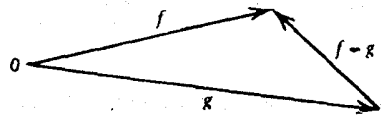
را برای چندجمله‌ییهای گوناگون $p(x)$، معین و مقایسه کنیم. نرم دیگری برای $C[a, b]$ در بخش ۳.۴ معرفی شده‌است، که آن نیز در اندازه‌گیری بزرگی $f(x) - p(x)$ مفید است.

همان‌گونه که در بخش ۱.۱ اشاره کردیم، نرم ماکسیمم در ویژگیهای مخصوص نرم که ذیلاً می‌آید صدق می‌کند:

$$\| f \| = ۰ \quad \text{اگر و فقط اگر} \quad f \equiv ۰ \qquad (۶.۱.۴)$$

$$\| \alpha f \| = | \alpha | \| f \| \quad \text{برای تمام اسکالرهای } \alpha \text{ و } f \in C[a, b] \text{ تمام} \qquad (۷.۱.۴)$$

$$\| f + g \| \leq \| f \| + \| g \| \quad \text{برای تمام مقادیر } f, g \in C[a, b] \qquad (۸.۱.۴)$$



شکل ۲.۴ نمایش تعریف فاصلهٔ $D(f,g)$.

اثبات این ویژگیها، برای (۴.۱.۴) کاملاً سرراست است. این ویژگیها نشان می‌دهند که نرم باید تعمیم قدرمطلق یک عدد تلقی شود.

اغلب سودمند است که $C[a,b]$ را یک فضای برداری تلقی کنیم، اگرچه هیچ استفادهٔ مهمی از این ایده نخواهیم کرد. در اینجا بردارها، توابع $f(x)$، $a \leq x \leq b$ هستند. فاصلهٔ بردار f از بردار g را چنین تعریف می‌کنیم:

$$D(f,g) = \| f - g \| \qquad (۹.۱.۴)$$

که با احساس ما از مفهوم فاصله در فضاهای برداری ساده‌تر مطابقت می‌نماید. این مطلب در شکل ۲.۴ نشان داده شده است.

با استفاده از نابرابری (۸.۱.۴)،

$$\| f - g \| = \| (f - h) + (h - g) \| \leq \| f - h \| + \| h - g \|$$

$$D(f,g) \leq D(f,h) + D(h,g) \qquad (۱۰.۱.۴)$$

این نابرابری به علّت تعبیر روشن آن در اندازه‌گیری طولهای اضلاع مثلثی با رئوس f و g و h، نابرابری مثلثی خوانده می‌شود. فرمول هم‌ارز (۸.۱.۴) را نیز نابرابری مثلثی می‌گویند.

یک قضیهٔ مفید دیگر، عکس نابرابری مثلثی است

$$| \| f \| - \| g \| | \leq \| f - g \| \qquad (۱۱.۱.۴)$$

برای اثبات آن از (۸.۱.۴) استفاده می‌کنیم تا به‌دست آوریم

$$\| f \| \leq \| f - g \| + \| g \|$$

$$\| f \| - \| g \| \leq \| f - g \|$$

همچنین داریم

$$\| g \| - \| f \| \leq \| g - f \| = \| f - g \|$$

که تساوی اخیر از (۷.۱.۴) با $\alpha = -1$ به‌دست می‌آید. از ترکیب این دو نابرابری (۱۱.۱.۴) اثبات می‌شود.

یک معرفی کاملتر از فضاهای برداری (فقط برای ابعاد متناهی) و نرم‌های برداری در فصل ۷ داده شده است. و یک هندسهٔ دیگری برای $C[a,b]$، همراه با معرفی یک نرم دیگر غیر از (۴.۱.۴)، در بخشهای ۳.۴ و ۴.۴ مطرح شده است. برای مواردی که می‌خواهیم دربارهٔ توابعی که چندین مشتق پیوسته دارند، صحبت کنیم، فضای تابعی $C^r[a,b]$ را معرفی می‌نماییم، که متشکّل از توابع $f(x)$ با r مشتق پیوسته بر $[a,b]$ است. این فضای برداری فایدهٔ جداگانه‌ای دارد ولی ما آن‌را فقط به عنوان یک وسیلهٔ ساده‌کردن نمادی تلقی می‌کنیم.

۲.۴ مسألهٔ تقریب مینیماکس

گیریم $f(x)$ بر $[a,b]$ پیوسته باشد. برای مقایسهٔ چندجمله‌ییهای $p(x)$، تقریبهای $f(x)$ که از روشهای گوناگونی به‌دست آمده‌اند، طبیعی است بپرسیم بهترین دقتی که ممکن است از به‌کاربردن چندجمله‌یی از هر درجهٔ $n \geq 0$ به‌دست آورد چیست؟ بنابراین به معرفی **خطای مینیماکس** می‌رسیم

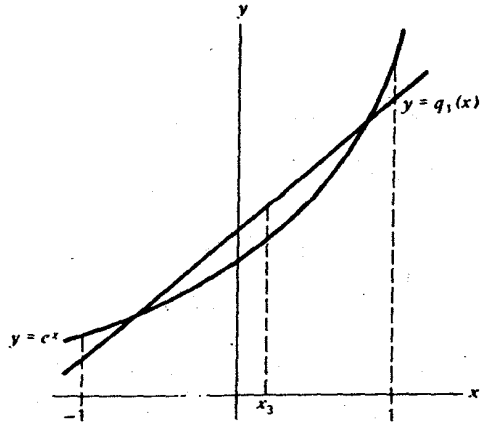
$$\rho_n(f) = \inf_{\deg(q)\leq n} \| f - q \|_\infty \qquad (۱.۲.۴)$$

یک چندجمله‌یی $q(x)$ از درجهٔ نابزرگتر از n وجود ندارد که $f(x)$ را با خطای ماکسیممی کمتر از $\rho_n(f)$ تقریب زند. با معرفی $\rho_n(f)$، می‌خواهیم ببینیم آیا یک چندجمله‌یی چون $q_n^*(x)$ وجود دارد به‌طوری که

$$\rho_n(f) = \| f - q_n^* \|_\infty \qquad (۲.۲.۴)$$

و اگر چنین است، آیا این چندجمله‌یی یکتاست؟ ویژگیهای آن چیست و چگونه می‌توان آن را ساخت؟ تقریب $q_n^*(x)$ را **تقریب مینیماکس** برای $f(x)$ بر $[a,b]$ خوانند و نظریهٔ آن در بخش ۶.۴ مطرح شده است.

مثال چندجمله‌یی تقریب مینیماکس $q_1^*(x)$ را برای e^x در $-1 \leq x \leq 1$ محاسبه کنید. گیریم $q_1^*(x) = a_0 + a_1 x$. برای پیداکردن a_0 و a_1، باید از یک بینش هندسی استفاده کنیم. نمودار $y = e^x$ را با نمودار یک تقریب ممکنِ $y = q_1(x)$، مانند شکل ۳.۴ در نظر می‌گیریم.



شکل ۳.۴ تقریب مینیماکس خطی برای e^x

فرض می‌کنیم

$$\epsilon(x) = e^x - [a_0 + a_1 x] \qquad (۳.۲.۴)$$

روشن است که $q_1^*(x)$ و e^x باید در دو نقطهٔ بازهٔ $[-1, 1]$، مثلاً در $-1 < x_1 < x_2 < 1$، با هم برابر باشند. وگرنه می‌توانیم با جابه‌جاکردن مناسب نمودار $y = q_1^*$ این تقریب را بهبود بخشیم. همچنین

$$\rho_1 = \underset{-1 \le x \le 1}{\text{Max}} \mid \epsilon(x) \mid$$

و $\epsilon(x_1) = \epsilon(x_2) = 0$. با استدلال دیگری که بر پایهٔ جابه‌جاکردن نمودار $y = q_1^*(x)$ استوار است، می‌توانیم نتیجه بگیریم که ماکسیمم خطای ρ_1 دقیقاً در سه نقطه حاصل می‌شود.

$$\epsilon(-1) = \rho_1 \qquad \epsilon(1) = \rho_1 \qquad \epsilon(x_3) = -\rho_1 \qquad (۴.۲.۴)$$

که $x_1 < x_3 < x_2$. چون $\epsilon(x)$ یک مینیمم نسبی در x_3 دارد، داریم $\epsilon'(x_3) = 0$. از ترکیب این چهار معادله، داریم

$$e^{-1} - [a_0 - a_1] = \rho_1 \qquad e - [a_0 + a_1] = \rho_1$$
$$e^{x_3} - [a_0 + a_1 x_3] = -\rho_1 \qquad e^{x_3} - a_1 = 0 \qquad (۵.۲.۴)$$

این معادلات دارای جوابهای زیرند

$$a_1 = \frac{e - e^{-1}}{2} \doteq ۱٫۱۷۵۲ \qquad x_2 = \ln_e(a_1) \doteq ۰٫۱۶۱۴$$
$$\rho_1 = \frac{1}{2}e^{-1} + \frac{x_2}{4}(e - e^{-1}) \doteq ۰٫۲۷۸۸$$
$$a_0 = \rho_1 + (1 - x_2)a_1 \doteq ۱٫۲۶۴۳$$

بنابراین

$$q_1^*(x) = ۱٫۲۶۴۳ + ۱٫۱۷۵۲x \qquad (۶.۲.۴)$$

و $\rho_1 \doteq ۰٫۲۷۸۸$.

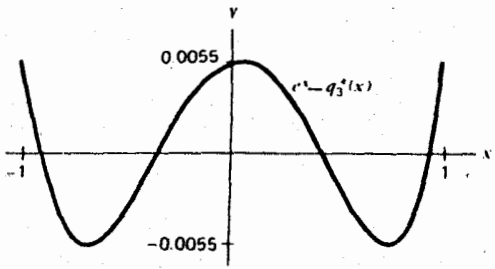
با استفاده از آنچه که الگوریتم رِمز[1] خوانده می‌شود، می‌توانیم $q_3^*(x)$ را برای e^x در $[-۱, ۱]$ بسازیم:

$$q_3^*(x) = ۰٫۹۹۴۵۷۹ + ۰٫۹۹۵۶۶۸x + ۰٫۵۴۲۹۷۳x^2 + ۰٫۱۷۹۵۳۳x^3 \qquad (۷.۲.۴)$$

نمودار خطای آن در شکل ۴.۴ داده شده است، و برخلاف تقریبهای تیلر (شکل ۱.۴ را ببینید) خطا در سراسر بازهٔ تقریب به‌طور هموار توزیع شده است.

مثال خود را با بیان خطاهای تقریب مینیماکس $q_n^*(x)$ و تقریبهای $p_n(x)$ تیلر برای $f(x) = e^x$ در $-۱ \le x \le ۱$، خاتمه می‌دهیم. خطاهای ماکسیمم برای مقادیر گوناگون n در جدول ۱.۴ داده شده است.

دقت $q_n^*(x)$ به‌طور چشمگیری از دقت $p_n(x)$ بیشتر است و تفاوت با افزایش n زیاد می‌شود. باید توجه داشت که e^x تابعی است که سری تیلر آن در مقایسه با $\log x$ و $\tan^{-1} x$ به سرعت همگرا می‌شود. دربارهٔ توابع اخیر، مینیماکس در مقایسه با سری تیلر، بهتر به‌نظر می‌رسد.



شکل ۴.۴ خطا برای تقریب مینیماکس درجهٔ ۳ برای e^x

1. Remes

جدول ۱.۴ خطای تیلر مینیماکس مقدار برای e^x

n	$\| f - p_n \|_\infty$	$\| f - q_n^* \|_\infty$
۱	۷٫۱۸E − ۱	۲٫۷۹E − ۱
۲	۲٫۱۸E − ۱	۴٫۵۰E − ۲
۳	۵٫۱۶E − ۲	۵٫۵۳E − ۳
۴	۹٫۹۵E − ۳	۵٫۴۷E − ۴
۵	۱٫۶۲E − ۳	۴٫۵۲E − ۵
۶	۲٫۲۶E − ۴	۳٫۲۱E − ۶
۷	۲٫۷۹E − ۵	۲٫۰۰E − ۷
۸	۳٫۰۶E − ۶	۱٫۱۱E − ۸
۹	۳٫۰۱E − ۷	۵٫۵۲E − ۱۰

۳.۴ مسألهٔ تقریب کمترین مربعات

به علّت دشواری محاسبهٔ تقریب مینیماکس، معمولاً، یک تقریب بینابینی در نظر می‌گیریم که تقریب کمترین مربعات خوانده می‌شود. نماد

$$\| g \|_2 = \sqrt{\int_a^b | g(x) |^2 \, dx} \qquad g \in C[a,b] \qquad (۱.۳.۴)$$

را معرفی می‌کنیم. این یک نرم تابعی است که، مانند نرم ماکسیمم (۴.۱.۴)، در ویژگیهای (۶.۱.۴) ـ (۸.۱.۴) صدق می‌کند. این تعمیم نرم معمولی اقلیدسی برای R^n است که در (۱۷.۱.۱) تعریف شده بود. در بخش بعد، هنگامی که تعریف بالا را تعمیم خواهیم داد، به اثبات نابرابری مثلثی (۸.۱.۴) برمی‌گردیم.

برای تابع مفروض $f \in C[a,b]$ و $n \geq 0$، تعریف می‌کنیم

$$M_n(f) = \inf_{\deg(r) \leq n} \| f - r \|_2 \qquad (۲.۳.۴)$$

مانند قبل، می‌خواهیم بدانیم آیا یک چندجمله‌یی r_n^* وجود دارد که عبارت

$$M_n(f) = \| f - r_n^* \|_2 \qquad (۳.۳.۴)$$

را مینیمم سازد؟ آیا این چندجمله‌یی یکتاست؟ آیا می‌توانیم آن را حساب کنیم؟

برای توجیه بیشترِ (۲.۳.۴)، محاسبهٔ خطای متوسط در تقریب $f(x)$ توسط $r(x)$ را در نظر می‌گیریم. برای عدد صحیحی مانند $m \geq ۱$، نقاط گرهی x_j را چنین تعریف می‌کنیم

$$x_j = a + \left(j - \frac{۱}{۲}\right)\left(\frac{b-a}{m}\right) \qquad j = ۱, ۲, \ldots, m$$

اینها نقاط میانی m زیر بازهٔ متساوی‌الفاصلهٔ $[a, b]$ هستند. پس یک **خطای متوسط** تقریب‌زدن $f(x)$ توسط $r(x)$ بر $[a, b]$ عبارت است از

$$\begin{aligned} E &= \lim_{m\to\infty} \left\{ \frac{۱}{m} \sum_{j=۱}^{m} [f(x_j) - r(x_j)]^۲ \right\}^{۱/۲} \\ &= \lim_{m\to\infty} \left\{ \frac{۱}{b-a} \sum_{j=۱}^{m} [f(x_j) - r(x_j)]^۲ \left(\frac{b-a}{m}\right) \right\}^{۱/۲} \\ &= \frac{۱}{\sqrt{b-a}} \sqrt{\int_a^b | f(x) - r(x) |^۲ \, dx} \\ E &= \frac{\| f - r \|_۲}{\sqrt{b-a}} \end{aligned} \qquad (۴.۳.۴)$$

بنابراین، تقریب کمترین مربعات باید خطای متوسط کوچکی بر $[a, b]$ داشته باشد. کمیت E جذر **میانگین مربع خطاها** در تقریب $f(x)$ توسط $r(x)$ نامیده می‌شود.

مثال گیریم $f(x) = e^x$، $-۱ \leq x \leq ۱$، و گیریم $r_۱(x) = b_۰ + b_۱ x$. مینیمم عبارت

$$\| f - r_۱ \|_۲^۲ = \int_{-۱}^{۱} [e^x - b_۰ - b_۱ x]^۲ dx \equiv F(b_۰, b_۱) \qquad (۵.۳.۴)$$

را پیدا کنید. اگر عبارت زیر علامت انتگرال را بسط دهیم و انتگرال را به انتگرال‌های کوچکتر بشکنیم، $F(b_۰, b_۱)$ یک چندجمله‌یی درجهٔ دوّم از دو متغیر $b_۰$ و $b_۱$ خواهد شد،

$$F = \int_{-۱}^{۱} \{e^{۲x} + b_۰^۲ + b_۱^۲ x^۲ - ۲b_۰ e^x - ۲b_۱ x e^x + ۲b_۰ b_۱ x\} dx$$

برای پیداکردن مینیمم، شرط لازم در یک نقطهٔ مینیمم را می‌نویسیم

$$\frac{\partial F}{\partial b_۰} = ۰ \qquad \frac{\partial F}{\partial b_۱} = ۰$$

به جای مشتق‌گیری از انتگرال اخیر از عبارت زیر علامت انتگرال (۵.۳.۴) مشتق می‌گیریم؛

$$\circ = \frac{\partial F}{\partial b_\circ} = \int_{-1}^{1} \frac{\partial}{\partial b_\circ}[e^x - b_\circ - b_1 x]^2 dx = 2\int_{-1}^{1}[e^x - b_\circ - b_1 x](-1)dx$$
$$\circ = \frac{\partial F}{\partial b_1} = 2\int_{-1}^{1}[e^x - b_\circ - b_1 x](-x)dx$$

بنابراین

$$b_\circ = \frac{1}{2}\int_{-1}^{1} e^x dx = \sinh(1) \doteq 1{,}1752$$
$$b_1 = \frac{3}{2}\int_{-1}^{1} x e^x dx = 3e^{-1} \doteq 1{,}1036$$
$$r_1^*(x) = 1{,}1752 + 1{,}1036x \qquad (۶.۳.۴)$$

با بررسی مستقیم

$$\| e^x - r_1^* \|_\infty \doteq \circ{,}44$$

این یک مقدار بین تقریبهای q_1^* و $p_1(x)$ است که قبلاً به‌دست آمده بودند. معمولاً تقریب کمترین مربعات یک تقریب یکنواخت نسبتاً خوب، برتر از تقریبهای سری تیلر است.

به عنوان یک مثال دیگر، تقریب کمترین مربعات درجهٔ سوّم برای e^x بر $[-1, 1]$ چنین است

$$r_3^*(x) = \circ{,}996294 + \circ{,}997955x + \circ{,}536722x^2 + \circ{,}176139x^3 \qquad (۷.۳.۴)$$

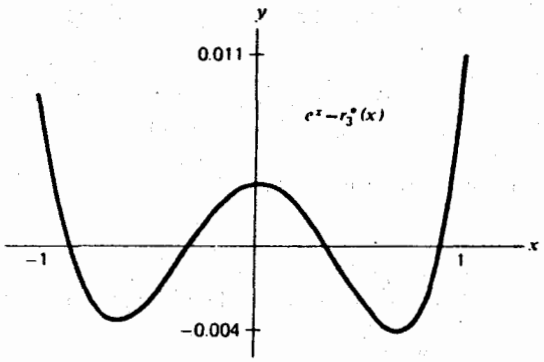
و

$$\| e^x - r_3^* \|_\infty = \circ{,}0112$$

نمودار خطا در شکل ۵.۴ داده شده است. توجه کنید که خطا در $[-1, 1]$ به اندازهٔ مینیماکس $q_3^*(x)$ در شکل ۴.۴ به طور هموار توزیع نشده است.

مسألهٔ کلی کمترین مربعات مسألهٔ کمترین مربعات (۲.۳.۴) را با دخالت‌دادن خطاهای میانگین وزین در تقریب $f(x)$ توسط یک چندجمله‌یی $r(x)$ تعمیم می‌دهیم. نظریهٔ کلی وجود، یکتایی، و ساختن تقریبهای کمترین مربعات در بخش ۵.۴ داده شده است. در آنجا از نظریهٔ چندجمله‌ییهای متعامد که در بخش بعد خواهد آمد، استفاده شده است.

گیریم $w(x)$ یک تابع وزن نامنفی بر بازهٔ (a, b) باشد، که ممکن است نامتناهی باشد، و برای آن ویژگیهای زیر را فرض می‌کنیم:



شکل ۵.۴ خطا در تقریب کمترین مربعات درجهٔ سوم برای e^x

.۱

$$\int_a^b |x|^n w(x)dx \qquad (۸.۳.۴)$$

انتگرالپذیر و برای تمام مقادیر $n \geq ۰$ متناهی است؛

۲. فرض می‌کنیم برای یک تابع پیوسته و نامنفی $g(x)$ داشته باشیم؛

$$\int_a^b w(x)g(x)dx = ۰ \qquad (۹.۳.۴)$$

در این صورت، در (a,b)، $g(x) \equiv ۰$.

مثال توابع وزنی که در این کتاب از آنها بیشتر استفاده شده به شکل زیرند؛

$$w(x) \equiv ۱ \qquad a \leq x \leq b$$

$$w(x) = \frac{۱}{\sqrt{۱-x^۲}} \qquad -۱ \leq x \leq ۱$$

$$w(x) = e^{-x} \qquad ۰ \leq x < \infty$$

$$w(x) = e^{-x^۲} \qquad -\infty < x < \infty$$

اکنون می‌توان برای یک بازهٔ متناهی $[a,b]$، مسألهٔ کلی کمترین مربعات را بیان کرد. $g \in C[a,b]$ داده شده است، آیا در میان تمام چندجمله‌ییهای $r(x)$ از درجهٔ نابزرگتر از n، یک چندجمله‌یی $r_n^*(x)$ وجود دارد که

$$\int_a^b w(x)[f(x)-r(x)]^۲dx \qquad (۱۰.۳.۴)$$

را مینیمم سازد؟

تابع $w(x)$ به خطاهای نقاط مختلف بازهٔ $[a,b]$ اهمیتهای متفاوتی می‌دهد. این امر در مطالعهٔ تقریبهای نزدیک مینیماکس مفید خواهد بود. برای محاسبهٔ (۱۰.۳.۴) برای یک چندجمله‌یی اختیاری $r(x)$ از درجهٔ نابزرگتر از n تعریف می‌کنیم

$$F(a_0,a_1,\ldots,a_n)=\int_a^b w(x)\left[f(x)-\sum_{j=0}^{n}a_jx^j\right]^2dx \qquad (۱۱.۳.۴)$$

ما می‌خواهیم وقتی که ضرایب $\{a_j\}$ همهٔ اعداد حقیقی را اختیار می‌کنند، F را مینیمم کنیم. یک شرط لازم برای اینکه F در نقطهٔ $(a_0,\ldots,a_n)$ مینیمم باشد این است که

$$\frac{\partial F}{\partial a_i}=0 \qquad i=0,1,\ldots,n \qquad (۱۲.۳.۴)$$

با مشتق‌گیری از انتگرال (۱۱.۳.۴) و استفاده از (۱۲.۳.۴)، دستگاه خطی زیر را به‌دست می‌آوریم

$$\sum_{j=0}^{n}a_j\int_a^b w(x)x^{i+j}dx=\int_a^b w(x)f(x)x^i dx \qquad i=0,1,\ldots,n \qquad (۱۳.۳.۴)$$

برای آنکه ببینیم چرا این جواب مسألهٔ کمترین مربعات رضایتبخش نیست، حالت خاص $w(x)\equiv 1$، $[a,b]=[0,1]$، را در نظر می‌گیریم. در این صورت دستگاه خطی چنین می‌شود

$$\sum_{j=0}^{n}\frac{a_j}{i+j+1}=\int_0^1 f(x)x^i dx \qquad i=0,1,\ldots,n \qquad (۱۴.۳.۴)$$

ماتریس ضرایب، ماتریس هیلبرت از مرتبهٔ $n+1$ است که در (۹.۶.۱) معرفی شده است. جواب دستگاه خطی (۱۴.۳.۴) نسبت به تغییرات کوچک ضرایب یا ثابتهای طرف راست، بسیار حساس است. پس این راه خوبی برای رسیدن به مسألهٔ کمترین مربعات نیست. درحساب با دقت معمولی در رایانهٔ IBM 3033، حالات $n\geq 4$ اصلاً رضایتبخش نخواهد بود.

۴.۴ چندجمله‌ییهای متعامد

همان‌گونه که از نمودارهای x^n بر $[0,1]$، $n\geq 0$، مشاهده می‌شود، این یک‌جمله‌ییها رفتاری خیلی نزدیک به رفتار خطی ـ وابسته دارند و خیلی شبیه یکدیگرند، و این امر علت ناپایداری دستگاه خطی (۱۴.۳.۴) است. برای اجتناب از این مسأله، پایهٔ دیگری برای چندجمله‌ییها در نظر می‌گیریم که بر چندجمله‌ییهای متعامد در یک فضای تابعی که در زیر می‌آوریم مبتنی هستند. این نتایج در تقریب توابع، و در بسیاری از کارها در ریاضیات کاربردی، از اهمیت اساسی برخوردارند.

گیریم $w(x)$ همان $w(x)$ در (۸.۳.۴) و (۹.۳.۴) باشد، و **حاصلضرب داخلی** دو تابع پیوستهٔ f و g را چنین تعریف می‌کنیم

$$(f,g)=\int_a^b w(x)f(x)g(x)dx \qquad f,g\in C[a,b] \qquad (۱.۴.۴)$$

در این صورت ویژگیهای سادهٔ زیر به آسانی ثابت می‌شوند.

۱. $(\alpha f,g)=(f,\alpha g)=\alpha(f,g)$ به ازای همهٔ اسکالرهای α

۲. $(f_۱+f_۲,g)=(f_۱,g)+(f_۲,g)$

$(f,g_۱+g_۲)=(f,g_۱)+(f_۱,g_۲)$

۳. $(f,g)=(g,f)$

۴. به ازای هر $f\in C[a,b]$، $(f,f)\geq ۰$ و $(f,f)=۰$، اگر وتنها اگر، برای $a\leq x\leq b$، $f(x)=۰$.

نرم دو یا نرم اقلیدسی را با رابطهٔ زیر تعریف می‌کنیم

$$\| f \|_۲=\sqrt{\int_a^b w(x)[f(x)]^۲dx}=\sqrt{(f,f)} \qquad (۲.۴.۴)$$

این تعریف، درویژگیهای نرم (۶.۱.۴) ـ (۸.۱.۴) صدق می‌کند. ولی اثبات نابرابری مثلثی (۸.۱.۴) دیگر بدیهی نیست، و به نابرابری معروف زیر بستگی دارد.

لم (نابرابری کوشی ـ شوارتس) برای $f,g\in C[a,b]$،

$$|(f,g)|\leq \| f \|_۲ \| g \|_۲ \qquad (۳.۴.۴)$$

برهان اگر $g=۰$، درستی قضیه بدیهی است. حال $g\neq ۰$ را در نظر می‌گیریم. برای هر عدد حقیقی α،

$$۰\leq (f+\alpha g,f+\alpha g)=(f,f)+۲\alpha(f,g)+\alpha^۲(g,g)$$

چندجمله‌یی طرف راست حداکثر یک ریشهٔ حقیقی دارد، و بنابراین مبیّن معادله نمی‌تواند مثبت باشد.

$$۴|(f,g)|^۲-۴(f,f)(g,g)\leq ۰$$

و از این رابطه (۳.۴.۴) نتیجه می‌شود. توجه داشته باشید که در (۳.۴.۴) تساوی تنها وقتی برقرار خواهد بود که مبین صفر باشد. ولی این امر ایجاب می‌کند که یک α^* وجود داشته باشد که به‌ازای آن چندجمله‌یی صفر شود. پس

$$(f+\alpha^* g, f+\alpha^* g) = 0$$

و بنابراین $f = -\alpha^* g$. بنابراین تساوی (۴.۴.۳) برقرار است اگر و تنها اگر (۱): یا f مضربی از g باشد، (۲): یا f یا g متحد با صفر باشد. ■

برای اثبات نابرابری مثلثی (۸.۱.۴)،

$$\| f+g \|_2^2 = (f+g, f+g) = (f,f) + 2(f,g) + (g,g)$$
$$\leq \| f \|_2^2 + 2 \| f \|_2 \| g \|_2 + \| g \|_2^2 = (\| f \|_2 + \| g \|_2)^2$$

از طرفین نامساوی جذر می‌گیریم، به‌دست می‌آید

$$\| f+g \|_2 \leq \| f \|_2 + \| g \|_2 \qquad (۴.۴.۴)$$

می‌خواهیم برای چندجمله‌ییها پایه‌ای غیر از پایهٔ معمول تک‌جمله‌ییهای $\{1, x, \ldots, x^n\}$ بیابیم. یک پایهٔ متعامد می‌سازیم که تعمیم پایهٔ متعامد در فضای R^n است (بخش ۱.۷ را ببینید). گوییم f و g **متعامدند** اگر

$$(f,g) = 0 \qquad (۵.۴.۴)$$

قضیّهٔ زیر یک قضیّهٔ وجودی سازنده برای چندجمله‌ییهای متعامد است.

قضیهٔ ۲.۴ (گرام‌ـاشمیت)[1] یک دنباله از چندجمله‌ییهای $\{\varphi_n(x) \mid n \geq 0\}$ با درجهٔ $n = \varphi(n)$، به ازای همهٔ مقادیر n، وجود دارد به طوری که به ازای همهٔ مقادیر m و n ناکوچکتر از صفر و $m \neq n$

$$(\varphi_n, \varphi_m) = 0 \qquad (۶.۴.۴)$$

به علاوه ما می‌توانیم دنباله را با ویژگیهای اضافی زیر بسازیم

1. Gram-Schmidt

(۱) $(\varphi_n, \varphi_n) = ۱$ به ازای هر n؛ (۲) ضریب x^n در $\varphi_n(x)$ مثبت باشد. با این ویژگیهای اضافی، دنبالهٔ $\{\varphi_n\}$ یکتاست.

برهان یک روش سازنده و بازگشتی برای پیدا کردن عناصر این دنباله ارائه می‌دهیم (این روند، رَوَندِ گرام ـ شمیت نامیده می‌شود). گیریم

$$\varphi_۰(x) = c$$

c یک ثابت است. c را طوری می‌گیریم که $\| \varphi_۰ \|_۲ = ۱$ و $c > ۰$. در این صورت

$$(\varphi_۰, \varphi_۰) = c^۲ \int_a^b w(x)dx = ۱$$

$$c = [\int_a^b w(x)dx]^{-۱/۲}$$

برای ساختن $\varphi_۱(x)$ بارابطهٔ زیر شروع می‌کنیم

$$\psi_۱(x) = x + a_{۱,۰}\varphi_۰(x)$$

در این صورت

$$(\psi_۱, \phi_۰) = ۰ \Rightarrow ۰ = (x, \varphi_۰) + a_{۱,۰}(\varphi_۰, \varphi_۰)$$

$$a_{۱,۰} = -(x, \phi_۰) = \frac{-\int_a^b xw(x)dx}{\left[\int_a^b w(x)dx\right]^{۱/۲}}$$

تعریف می‌کنیم

$$\varphi_۱(x) = \frac{\psi_۱(x)}{\| \psi_۱ \|_۲}$$

و توجه داریم که

$$\| \varphi_۱ \|_۲ = ۱ \quad (\varphi_۱, \varphi_۰) = ۰$$

و ضریب x مثبت است.

برای ساختن $\varphi_n(x)$ ابتدا تعریف می‌کنیم

$$\psi_n(x) = x^n + a_{n,n-۱}\varphi_{n-۱}(x) + \cdots + a_{n,۰}\varphi_۰(x) \qquad (۷.۴.۴)$$

و ثابتها را طوری انتخاب کنیم که ψ_n بر φ_j، به ازای $j = ۰, \ldots, n-۱$، عمود شود. در این صورت از $(\psi_n, \varphi_j) = ۰$ نتیجه می‌شود

$$a_{n,j} = -(x^n, \varphi_j) \qquad j = ۰, ۱, \ldots, n-۱ \qquad (۸.۴.۴)$$

$\varphi_n(x)$ مورد نظر عبارت است از

$$\varphi_n(x) = \frac{\psi_n(x)}{\| \psi_n \|_۲} \qquad (۹.۴.۴)$$

و بقیّه را با استقرا ادامه می‌دهیم. ■

مثال برای حالت خاص $w(x) \equiv ۱$ و $[a, b] = [-۱, ۱]$، داریم

$$\varphi_۰(x) = \sqrt{\frac{۱}{۲}} \qquad \varphi_۱(x) = \sqrt{\frac{۳}{۲}}x \qquad \varphi_۲(x) = \sqrt{\frac{۵}{۲}}(۳x^۲ - ۱)$$

و چندجمله‌ییهای بعدی را می‌توان با فرایند بالا ساخت.

نوشتارهای بسیار زیادی در بارهٔ این‌گونه چندجمله‌ییها از جمله فرمول‌های گوناگونی برای این چندجمله‌ییها موجود است. ما الزاماً یک نگاه اجمالی به آنها می‌اندازیم. چندجمله‌ییها معمولاً به شکلی داده شده‌اند که برای آنها $\| \varphi_n \|_۲ \neq ۱$.

حالتهای خاص

حالت ۱. چندجمله‌ییهای لژاندر. گیریم $w(x) \equiv ۱$ در $[-۱, ۱]$. تعریف می‌کنیم

$$P_n(x) = \frac{(-۱)^n}{۲^n n!} \cdot \frac{d^n}{dx^n}[(۱ - x^۲)^n] \qquad n \geq ۱ \qquad (۱۰.۴.۴)$$

با $P_۰(x) \equiv ۱$. این چند جمله‌ییها در $[-۱, ۱]$ متعامدند، درجهٔ $P_n(x)$ برابر n است و به ازای همهٔ مقادیر n، $P_n(۱) = ۱$. همچنین

$$(P_n, P_n) = \frac{۲}{۲n + ۱}$$

$$\varphi_n(x) = \sqrt{\frac{۲n + ۱}{۲}} P_n(x) \qquad (۱۱.۴.۴)$$

حالت ۲. چندجمله‌ییهای چبیشف. گیریم $w(x) = ۱/\sqrt{۱ - x^۲}$، $-۱ \leq x \leq ۱$. در این صورت

$$T_n(x) = \cos(n \cos^{-۱} x) \qquad n \geq ۰ \qquad (۱۲.۴.۴)$$

یک خانوادهٔ متعامد از چندجمله‌ییهای با درجهٔ (T_n) برابر n است. برای اینکه ببینیم $T_n(x)$ یک چندجمله‌یی است، می‌گیریم $\theta = \cos^{-1} x$، $0 \leq \theta \leq \pi$. در این صورت

$$T_{n\pm 1}(x) = \cos(n \pm 1)\theta = \cos(n\theta)\cos\theta \mp \sin(n\theta)\sin\theta$$

$$T_{n+1}(x) + T_{n-1}(x) = 2\cos(n\theta)\cos\theta = 2T_n(x)x$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \qquad n \geq 1 \qquad (۱۳.۴.۴)$$

همچنین با محاسبهٔ مستقیم در (۱۲.۴.۴)،

$$T_0(x) \equiv 1 \qquad T_1(x) = x$$

با استفاده از (۱۳.۴.۴)

$$T_2(x) = 2x^2 - 1 \qquad T_3(x) = 2x(2x^2 - 1) - x = 4x^3 - 3x$$

این چند جمله‌ییها در شرایط: $T_n(1) = 1$، $n \geq 1$ و شرایط زیر نیز صدق می‌کنند

$$(T_n, T_m) = \begin{cases} 0 & n \neq m \\ \pi & n = m = 0 \\ \frac{\pi}{2} & n = m > 0 \end{cases} \qquad (۱۴.۴.۴)$$

چندجمله‌ییهای چبیشف در نظریهٔ تقریب بسیار مهم‌اند، و در بسیاری از زمینه‌های دیگر ریاضیات کاربردی نیز پدید می‌آیند. برای یک بحث کاملتری در مورد آنها، ریولین[1] (۱۹۷۴) و فاکس و پارکر[2] (۱۹۶۸) را ببینید. ویژگیهای دیگر چندجمله‌ییهای چبیشف را در بخش بعد ارائه خواهیم داد.

حالت ۳. چندجمله‌ییهای لاگر. گیریم $w(x) = e^{-x}$، $[a, b] = [0, \infty)$. در این صورت

$$L_n(x) = \frac{1}{n! e^{-x}} \cdot \frac{d^n}{dx^n}\{x^n e^{-x}\} \qquad n \geq 0 \qquad (۱۵.۴.۴)$$

به ازای جمیع مقادیر n، $\| L_n \|_2 = 1$، و $\{L_n\}$ها در $[0, \infty)$ با تابع وزن e^{-x}، متعامدند.

یک خانواده از توابع را متعامد گوییم اگر هر عضو خانواده بر تک‌تک اعضای دیگر خانواده عمود باشد. خانواده را **خانوادهٔ یکا متعامد** خوانیم اگر متعامد باشد و طول هر عضو خانواده یک باشد، یعنی، $\| f \|_2 = 1$. برای مثالهای دیگر از چندجمله‌ییهای متعامد، آبراموویتس و اشتگون (۱۹۶۴، فصل ۲۲)، (دیویس (۱۹۶۳، پیوست)، سگو[3] (۱۹۶۸) را ببینید.

1. Riulin 2. Fox and Parker 3. Szego

بعضی ویژگیهای چندجمله‌ییهای متعامد این قضایا در این فصل و برای فصلهای بعد سودمندند.

قضیهٔ **۳.۴** گیریم $\{\varphi_n(x) \mid n \geq 0\}$ یک خانواده از چندجمله‌ییهای متعامد بر (a,b) با تابع وزن $w(x)$ باشد. با چنین خانواده‌ای، همیشه به‌طور ضمنی فرض می‌کنیم که درجهٔ φ_n برابر n و $n \geq 0$. اگر $f(x)$ یک چندجمله‌یی از درجهٔ m باشد، آنگاه

$$f(x) = \sum_{n=0}^{m} \frac{(f,\varphi_n)}{(\varphi_n,\varphi_n)}\varphi_n(x) \qquad (۱۶.۴.۴)$$

برهان ابتدا نشان می‌دهیم که هر چندجمله‌یی را می‌توان به صورت ترکیبی خطی از چندجمله‌ییهای متعامد از درجهٔ نابیشتر نوشت. چون درجهٔ $\varphi_0 = 0$، داریم $\varphi_0(x) = c$، یک ثابت، و بنابراین

$$1 \equiv \frac{1}{c}\varphi_0(x)$$

چون درجهٔ $(\varphi_1) = 1$، از فرایند گرام ـ شمیت داریم،

$$\varphi_1(x) = c_{1,1}x + c_{1,0}\varphi_0(x) \qquad c_{1,1} \neq 0$$

$$x = \frac{1}{c_{1,1}}[\varphi_1(x) - c_{1,0}\varphi_0(x)]$$

با استقرا در فرایند گرام ـ شمیت

$$\varphi_r(x) = c_{r,r}x^r + c_{r,r-1}\varphi_{r-1}(x) + \cdots + c_{r,0}\varphi_0(x) \qquad c_{r,r} \neq 0$$

و

$$x^r = \frac{1}{c_{r,r}}[\varphi_r(x) - c_{r,r-1}\varphi_{r-1}(x) - \cdots - c_{r,0}\varphi_0(x)]$$

پس هریک جمله‌یی را می‌توان به صورت ترکیبی از چندجمله‌ییهای متعامد درجهٔ نابیشتر بیان کرد. از این امر به سادگی نتیجه می‌شود که یک چندجمله‌یی دلخواه $f(x)$ از درجهٔ m را می‌توان به ازای انتخاب مقادیری مانند $b_m, \ldots, b_0$ به شکل زیر نوشت.

$$f(x) = b_m\varphi_m(x) + \cdots + b_0\varphi_0(x)$$

برای محاسبهٔ هر b_i، طرفین را در $w(x)$ و $\varphi_i(x)$ ضرب نموده روی (a,b) انتگرال می‌گیریم. لذا

$$(f,\varphi_i) = \sum_{j=0}^{m} b_j(\varphi_j,\varphi_i) = b_i(\varphi_i,\varphi_i) \quad ,$$

$$b_i = \frac{(f,\varphi_i)}{(\varphi_i,\varphi_i)}$$

که (۱۶.۴.۴) و قضیه را اثبات می‌کند. ■

فرع اگر $f(x)$ یک چندجمله‌یی از درجهٔ نابزرگتر از $m-1$ باشد، آنگاه

$$(f, \varphi_m) = 0 \qquad (17.4.4)$$

و $\varphi_m(x)$ و $f(x)$ متعامدند. ■

برهان این مطلب به سادگی از (۱۶.۴.۴) و تعامد خانوادهٔ $\{\varphi_n(x)\}$ نتیجه می‌شود. ■

قضیّهٔ بعدی درک مستقیمی از شکل نمودارهای چندجمله‌ییهای متعامد به‌دست می‌دهد و برای کار با تربیع گاوسی که در فصل ۵ خواهد آمد نیز بسیار مهم است.

قضیهٔ ۴.۴ گیریم $\{\varphi_n(x) \mid n \geq 0\}$ یک خانواده از چندجمله‌ییهای متعامد بر $[a, b]$ با تابع وزن $w(x)$ باشد. در این صورت چند جمله‌یی $\varphi_n(x)$ دقیقاً n ریشهٔ حقیقی متمایز در بازهٔ باز (a, b) دارد.

برهان گیریم $x_1, x_2, \ldots, x_m$ همهٔ صفرهای $\varphi_n(x)$ باشند که

۱. $a < x_i < b$

۲. $\varphi_n(x)$ در x_i تغییر علامت می‌دهد.

چون درجهٔ $\varphi_n) = n$، بدیهی است، داریم $m \leq n$. فرض می‌کنیم $m < n$ و به تناقض می‌رسیم.

تعریف می‌کنیم

$$B(x) = (x - x_1) \ldots (x - x_m)$$

به موجب تعریفِ نقاط $x_1, \ldots, x_m$، چندجمله‌یی

$$\varphi_n(x)B(x) = (x - x_1) \ldots (x - x_m)\varphi_n(x)$$

در (a, b) تغییر علامت نمی‌دهد. برای آنکه این امر را روشنتر ببینیم، توجه می‌کنیم که فرضهای مربوط به $x_1, \ldots, x_m$ رابطهٔ:

$$\varphi_n(x) = h(x)(x - x_1)^{r_1} \ldots (x - x_m)^{r_m}$$

را ایجاب می‌کنند که در آن هر r_i فرد است و $h(x)$ در (a, b) تغییر علامت نمی‌دهد. پس عبارت

$$\varphi_n(x)B(x) = (x - x_1)^{r_1+1} \ldots (x - x_m)^{r_m+1}h(x)$$

تغییرعلامت نمی‌دهد. در نتیجه

$$\int_a^b w(x)B(x)\varphi_n(x)dx \neq 0$$

زیرا روشن است که $B(x)\varphi_n(x) \not\equiv 0$. ولی چون m، درجهٔ B از n کوچکتر است، از فرع قضیهٔ ۳.۴ نتیجه می‌شود که

$$\int_a^b w(x)B(x)\varphi_n(x)dx = (B, \varphi_n) = 0$$

و این یک تناقض است، پس باید داشته باشیم $m = n$. پس حکم قضیه حاصل می‌شود؛ زیرا $\varphi_n(x)$ حداکثر n ریشه دارد و فرضهای مربوط به $x_1, \ldots, x_n$ ایجاب می‌کنند که همهٔ آنها ساده باشند، یعنی $\varphi_n'(x_i) \neq 0$. ■

مانند قبل فرض می‌کنیم $\{\varphi_n(x) \mid n \geq 0\}$ خانواده‌ای متعامد بر (a, b) با تابع وزن $w(x) \geq 0$ باشد. A_n و B_n را با

$$\varphi_n(x) = A_n x^n + B_n x^{n-1} + \ldots \qquad (۱۸.۴.۴)$$

تعریف می‌کنیم. همچنین می‌نویسیم

$$\varphi_n(x) = A_n(x - x_{n,1})(x - x_{n,2})\ldots(x - x_{n,n}) \qquad (۱۹.۴.۴)$$

گیریم

$$\gamma_n = (\varphi_n, \varphi_n) > 0 \qquad a_n = \frac{A_{n+1}}{A_n} \qquad (۲۰.۴.۴)$$

قضیهٔ ۵.۴ (رابطهٔ بازگشتی سه‌گانه) گیریم $\{\varphi_n\}$ یک خانواده از چندجمله‌ییهای متعامد بر (a, b) با تابع وزن $w(x) \geq 0$ باشد. در این صورت به ازای $n \geq 1$

$$\varphi_{n+1}(x) = (a_n x + b_n)\varphi_n(x) - c_n\varphi_{n-1}(x) \qquad (۲۱.۴.۴)$$

که در آن

$$c_n = \frac{A_{n+1}A_{n-1}}{A_n^2} \cdot \frac{\gamma_n}{\gamma_{n-1}} \qquad b_n = a_n \cdot \left[\frac{B_{n+1}}{A_{n+1}} - \frac{B_n}{A_n}\right] \qquad (۲۲.۴.۴)$$

برهان ابتدا توجه نمایید که رابطهٔ بازگشتی سه‌گانهٔ (۱۳.۴.۴) برای چندجمله‌ییهای چبیشف یک مثال از (۲۱.۴.۴) است. برای به‌دست آوردن (۲۱.۴.۴)، اثبات را با در نظرگرفتن چندجمله‌یی

زیر شروع می‌کنیم:

$$\begin{aligned} G(x) =& \varphi_{n+1}(x) - a_n x\varphi_n(x) \\ =& [A_{n+1}x^{n+1} + B_{n+1}x^n + \cdots] \\ & - \frac{A_{n+1}}{A_n} x[A_n x^n + B_n x^{n-1} + \cdots] \\ =& \left[B_{n+1} - \frac{A_{n+1}B_n}{A_n}\right] x^n + \cdots \end{aligned}$$

و درجهٔ G از n نابزرگتر است. به موجب قضیهٔ ۴.۳ برای مجموعهٔ مناسبی از d_0، ...، d_n می‌توانیم بنویسیم

$$G(x) = d_n\varphi_n(x) + \cdots + d_0\varphi_0(x)$$

d_i را حساب می‌کنیم

$$d_i = \frac{(G, \varphi_i)}{(\varphi_i, \varphi_i)} = \frac{1}{\gamma_i}[(\varphi_{n+1}, \varphi_i) - a_n(x\varphi_n, \varphi_i)] \qquad (۲۳.۴.۴)$$

به ازای $i \leq n$ داریم $(\varphi_{n+1}, \varphi_i) = 0$ و به ازای $i \leq n-2$ داریم

$$(x\varphi_n, \varphi_i) = \int_a^b w(x)\varphi_n(x)x\varphi_i(x)dx = 0$$

زیرا در این صورت درجهٔ $(x\varphi_i(x)) \leq n-1$. از ترکیب این نتایج

$$d_i = 0 \qquad 0 \leq i \leq n-2$$

و بنابراین

$$\begin{aligned} G(x) &= d_n\varphi_n(x) + d_{n-1}\varphi_{n-1}(x) \\ \varphi_{n+1}(x) &= (a_n x + d_n)\varphi_n(x) + d_{n-1}\varphi_{n-1}(x) \qquad (۲۴.۴.۴) \end{aligned}$$

این وجود یک رابطهٔ بازگشتی سه‌گانه را نشان می‌دهد و بقیهٔ کار دستکاری فرمولها برای به‌دست آوردن d_n و d_{n-1} است که مانند b_n و c_n در (۲۲.۴.۴) داده شده‌اند. این ثابتها را در اینجا به‌دست نیاورده‌ایم ولی مقادیر آنها برای بعضی کاربردها مهم‌اند (مسألهٔ ۱۸ را ببینید). ■

مثال ۱. برای چندجمله‌ییهای لاگر

$$L_{n+1}(x) = \frac{1}{n+1}[2n+1-x]L_n(x) - \frac{n}{n+1}L_{n-1}(x) \qquad (۲۵.۴.۴)$$

۲. برای چندجمله‌ییهای لژاندر

$$P_{n+1}(x) = \frac{2n+1}{n+1}xP_n(x) - \frac{n}{n+1}P_{n-1}(x) \qquad (۲۶.۴.۴)$$

قضیهٔ ۶.۴ (اتحاد کریستوفل ـ داربو) برای یک خانواده از چندجمله‌ییهای متعامد $\{\varphi_n\}$ با تابع وزن $w(x) \geq 0$،

$$\sum_{k=0}^{n} \frac{\varphi_k(x)\varphi_k(y)}{\gamma_k} = \frac{\varphi_{n+1}(x)\varphi_n(y) - \varphi_n(x)\varphi_{n+1}(y)}{a_n\gamma_n(x-y)}, x \neq y \quad (۲۷.۴.۴)$$

برهان برهان براساس عملیات روی رابطهٔ بازگشتی سه‌گانه صورت می‌گیرد سگو (۱۹۶۷، ص ۴۳) را ببینید). ■

۵.۴ مسألهٔ تقریب کمترین مربعات (ادامه)

اکنون به مسألهٔ کلی کمترین مربعات، مینیمم‌سازی (۱۰.۳.۴) در بین همهٔ چندجمله‌ییهای از درجهٔ نابزرگتر از n برمی‌گردیم. فرض می‌کنیم $\{\varphi_k(x) \mid k \geq 0\}$ یک خانواده از چندجمله‌ییهای متعامد با تابع وزن $w(x) \geq 0$ باشد، یعنی

$$(\varphi_n, \varphi_m) = \delta_{n,m} = \begin{cases} 1 & n = m \\ 0 & n \neq m \end{cases}$$

پس یک چندجمله‌یی دلخواه $f(x)$ از درجهٔ نابزرگتر از n را می‌توان چنین نوشت

$$r(x) = b_0\varphi_0(x) + \cdots + b_n\varphi_n(x) \qquad (۱.۵.۴)$$

برای تابع داده شدهٔ $f \in C[a,b]$،

$$\| f - r \|_2^2 = \int_a^b w(x)\left[f(x) - \sum_{j=0}^{n} b_j\varphi_j(x)\right]^2 dx \equiv G(b_0, \ldots, b_n) \quad (۲.۵.۴)$$

مسألهٔ کمترین مربعات را با مینیمم‌سازی G حل می‌کنیم. مانند قبل، می‌توانیم قرار دهیم

$$\frac{\partial G}{\partial b_i} = 0 \qquad i = 0, 1, \ldots, n$$

ولی برای به‌دست آوردن یک نتیجهٔ کاملتر، به راه دیگری می‌پردازیم. برای هر انتخاب $b_0, \ldots, b_n$،

$$0 \le G(b_0, \ldots, b_n) = \left(f - \sum_{j=0}^{n} b_j\varphi_j, f - \sum_{i=0}^{n} b_i\varphi_i\right)$$

$$= (f,f) - 2\sum_{j=0}^{n} b_j(f,\varphi_j) + \sum_i \sum_j b_i b_j(\varphi_i,\varphi_j)$$

$$= \| f \|_2^2 - 2\sum_{j=0}^{n} b_j(f,\varphi_j) + \sum_{j=0}^{n} b_j^2$$

$$= \| f \|_2^2 - \sum_{j=0}^{n}(f,\varphi_j)^2 + \sum_{j=0}^{n}\Big[(f,\varphi_j) - b_j\Big]^2 \qquad (3.5.4)$$

که می‌توان درستی آن را با بسط آخرین جمله بررسی کرد. بنابراین G مینیمم است اگر و تنها اگر

$$b_j = (f,\varphi_j) \qquad j = 0, 1, \ldots n$$

پس تقریب کمترین مربعات وجود دارد، یکتاست و با رابطهٔ زیر داده می‌شود

$$r_n^*(x) = \sum_{j=0}^{n}(f,\varphi_j)\varphi_j(x)$$

علاوه بر این، از (۲.۵.۴) و (۳.۵.۴)

$$\| f - r_n^* \|_2 = \left[\| f \|_2^2 - \sum_{j=0}^{n}(f,\varphi_j)^2\right]^{1/2}$$

$$= \sqrt{\| f \|_2^2 - \| r_n^* \|_2^2} \qquad (4.5.4)$$

$$\| f \|_2^2 = \| r_n^* \|_2^2 + \| f - r_n^* \|_2^2 \qquad (5.5.4)$$

به عنوان یک توضیح مفید برای به‌دست آوردن $r_{n+1}^*(x)$، از رابطهٔ زیر استفاده کنید

$$r_{n+1}^*(x) = r_n^*(x) + (f,\varphi_{n+1})\varphi_{n+1}(x) \qquad (6.5.4)$$

قضیهٔ ۷.۴ فرض می‌کنیم $[a,b]$ متناهی باشد،

$$\lim_{n\to\infty} \| f - r_n^* \|_2 = 0 \qquad (7.5.4)$$

برهان از تعریف r_n^* به عنوان یک چندجمله‌یی مینیمم‌ساز برای $\| f - r_n \|_۲$ ، داریم

$$\| f - r_۱^* \|_۲ \geq \| f - r_۲^* \|_۲ \geq \ldots \geq \| f - r_n^* \|_۲ \geq \ldots \qquad (۸.۵.۴)$$

گیریم $\varepsilon > ۰$ دلخواه باشد. پس طبق قضیهٔ وایرشتراس، یک چندجمله‌یی $Q(x)$ از درجهٔ، مثلاً m وجود دارد که برای آن

$$\max_{a \leq x \leq b} | f(x) - Q(x) | \leq \frac{\varepsilon}{c} \qquad c = \sqrt{\int_a^b w(x)dx}$$

بنابه تعریف $r_m^*(x)$،

$$\| f - r_m^* \|_۲ \leq \| f - Q \|_۲ = \left[\int_a^b w(x)[f(x) - Q(x)]^۲ \right]^{۱/۲}$$
$$\leq \left[\int_a^b w(x) \frac{\varepsilon^۲}{c^۲} dx \right]^{۱/۲} = \varepsilon$$

از ترکیب این رابطه با (۸.۵.۴)، به ازای جمیع مقادیر $n > m$

$$\| f - r_n^* \|_۲ \leq \varepsilon$$

چون ε دلخواه بود، (۷.۵.۴) اثبات می‌شود. ■

با استفاده از (۵.۵.۴) و یک محاسبهٔ ساده $\| r_n^* \|^۲$، **نامساوی بسل** را خواهیم داشت:

$$\| r_n^* \|_۲^۲ = \sum_{j=۰}^{n} (f, \varphi_j)^۲ \leq \| f \|_۲^۲ \qquad (۹.۵.۴)$$

و با استفاده از (۷.۵.۴) در (۵.۵.۴) **تساوی پارسوال** را به‌دست می‌آوریم:

$$\| f \|_۲ = \left[\sum_{j=۰}^{\infty} (f, \varphi_j)^۲ \right]^{۱/۲} \qquad (۱۰.۵.۴)$$

قضیهٔ ۷.۴ نمی‌گوید که $\| f - r_n^* \|_\infty \to ۰$. ولی اگر فرضهای مشتقپذیری بیشتری برای $f(x)$ در نظرگرفته شود، قضایای مربوط به همگرایی یکنواخت r_n^* به f را می‌توان ثابت کرد. یک مثال بعداً داده شده است.

بسطهای چندجمله‌یی لژاندر برای حل مسألهٔ کمترین مربعات بر یک بازهٔ متناهی $[a,b]$ با تابع وزن $w(x) \equiv ۱$، می‌توانیم آن را به یک مسأله بر $[-۱, ۱]$ برگردانیم. تبدیل متغیر

$$x = \frac{b+a+(b-a)t}{۲} \qquad (۱۱.۵.۴)$$

بازهٔ $-۱ \le t \le ۱$ را به بازهٔ $a \le x \le b$ بدل می‌کند. برای یک تابع داده شدهٔ $f \in C[a,b]$، تعریف می‌کنیم

$$F(t) = f\left(\frac{b+a+(b-a)t}{۲}\right) \qquad -۱ \le t \le ۱ \qquad (۱۲.۵.۴)$$

پس

$$\int_a^b [f(x) - r_n(x)]^۲ dx = \left(\frac{b-a}{۲}\right) \int_{-۱}^{۱} [F(t) - R_n(t)]^۲ dt$$

که $R_n(t)$ از $r_n(x)$ با استفاده از (۱۱.۵.۴) به‌دست می‌آید. تغییر متغیر (۱۱.۵.۴) یک تناظر یک به یک بین چندجمله‌ییهای از درجهٔ m بر $[a,b]$ و از درجهٔ m بر $[-۱, ۱]$، برای هر $m > ۰$، به‌دست می‌دهد. بنابراین مینیمم‌کردن $\| f - r_n \|_۲$ بر $[a,b]$ هم‌ارز است با مینیمم‌کردن $\| \varphi - R_n \|_۲$ بر $[-۱, ۱]$. لذا ما توجه خود را به مسألهٔ کمترین مربعات بر $[-۱, ۱]$ محدود می‌کنیم.

برای تابع مفروض $f \in [-۱, ۱]$، خانوادهٔ یکا متعامد مذکور در قضیهٔ ۲.۴ عبارت است از $\varphi_۰(x) \equiv \frac{۱}{\sqrt{۲}}$،

$$\varphi_n(x) = \sqrt{\frac{۲n+۱}{۲}} \cdot \frac{(-۱)^n}{۲^n n!} \cdot \frac{d^n}{dx^n}[(۱-x^۲)^n] \qquad n \ge ۱ \qquad (۱۳.۵.۴)$$

تقریب کمترین مربعات چنین است

$$r_n^*(x) = \sum_{j=۰}^{n} (f, \varphi_j)\varphi_j(x) \qquad (f, \varphi_j) = \int_{-۱}^{۱} f(x)\varphi_j(x)dx \qquad (۱۴.۵.۴)$$

که جواب مسألهٔ کلی کمترین مربعات است که در بخش ۳.۴ مطرح شده است. ضرایب (f, φ_j) را ضرایب لژاندر می‌خوانند.

مثال ضرایب بسط (f, φ_j) در (۱۴.۵.۴) در جدول ۲.۴ برای $f(x) = e^x$ بر $[-۱, ۱]$ داده شده‌اند. تقریب $r_۳^*(x)$ که به شکل چندجمله‌یی استاندهٔ نوشته شده‌بود، قبلاً در (۷.۳.۴) داده

جدول ۲.۴ ضرایب بسط لژاندر برای e^x

j	(f,φ_j)	j	(f,φ_j)
۰	۱٫۶۶۱۹۸۵	۳	۰٫۰۳۷۶۶۰
۱	۰٫۹۰۱۱۱۷	۴	۰٫۰۰۴۶۹۸
۲	۰٫۲۲۶۳۰۲	۵	۰٫۰۰۰۴۶۹

شده‌است. خطای متوسّط E در $r_۳^*(x)$ را از ترکیب روابط (۴.۳.۴)، (۴.۵.۴)، (۹.۵.۴) و از جدول ضرایب به‌دست می‌آوریم

$$E = \frac{۱}{\sqrt{۲}} \| e^x - r_۳^*(x) \|_۲ \doteq ۰٫۰۰۳۴$$

بسطهای چندجمله‌یی چبیشف در اینجا $w(x) = ۱/\sqrt{۱-x^۲}$ تابع وزن است و

$$\varphi_۰(x) = \frac{۱}{\sqrt{\pi}} \qquad \varphi_n(x) = \sqrt{\frac{۲}{\pi}} T_n(x) \qquad n \geq ۱ \qquad (۱۵.۵.۴)$$

جواب کمترین مربعات چنین است

$$C_n(x) = \sum_{j=۰}^{n} (f,\varphi_j)\varphi_j(x) \qquad (f,\varphi_j) = \int_{-۱}^{۱} \frac{f(x)\varphi_j(x)}{\sqrt{۱-x^۲}} dx \quad (۱۶.۵.۴)$$

با استفاده از تعریف $\varphi_n(x)$ برحسب $T_n(x)$،

$$C_n(x) = {\sum_{j=۰}^{n}}' c_j T_j(x) \qquad c_j = \frac{۲}{\pi} \int_{-۱}^{۱} \frac{f(x)T_j(x)}{\sqrt{۱-x^۲}} dx \quad (۱۷.۵.۴)$$

علامت پریم بر نماد مجموعیابی بدین معناست که قبل از مجموعیابی، اولین جمله باید نصف شود.

بسط چبیشف ارتباط نزدیکی به بسط کسینوس فوریه دارد. با استفاده از $x = \cos\theta$، $۰ \leq \theta \leq \pi$،

$$C_n(\cos\theta) = {\sum_{j=۰}^{n}}' c_j \cos(j\theta) \qquad (۱۸.۵.۴)$$

$$c_j = \frac{۲}{\pi} \int_{۰}^{\pi} \cos(j\theta) f(\cos\theta) d\theta \qquad (۱۹.۵.۴)$$

بنابراین $C_n(\cos\theta)$ برش براز $n+۱$ جملهٔ اول بسط کسینوس فوریهٔ

$$f(\cos\theta) = {\sum_{۰}^{\infty}}' c_j \cos(j\theta)$$

است. اگر بسط کسینوس فوریهٔ $f(\cos\theta)$ بر $[0,\pi]$ معلوم باشد، آنگاه با قراردادن $\theta = \cos^{-1} x$ بسط چبیشف $f(x)$ به‌دست می‌آید.

به دلایلی که بعداً ارائه خواهد شد، تقریب کمترین مربعات چبیشف مفیدتر از تقریب کمترین مربعات لژاندر است. بدین دلیل ما قضیهٔ همگرایی برای (۱۷.۵.۴) را با تفصیل بیشتری بیان خواهیم کرد.

قضیهٔ ۸.۴ گیریم $f(x)$ دارای r مشتق پیوسته بر $[-1,1]$ باشد، با $r \geq 1$. در این صورت به ازای یک ثابت B که به f و r بستگی دارد، برای $C_n(x)$، تقریب کمترین مربعات چبیشف که در (۱۷.۵.۴) تعریف شده است داریم

$$\| f - C_n \|_\infty \leq \frac{B \ln n}{n^r} \qquad n \geq 2 \qquad (۲۰.۵.۴)$$

بنابراین با $n \to \infty$، $C_n(x)$ به‌طور یکنواخت به $f(x)$ میل می‌کند به شرطی که $f(x)$ پیوسته مشتق‌پذیر باشد.

برهان برای اثبات، ریولین (۱۹۷۴، قضیهٔ ۳.۳ صفحهٔ ۱۳۴) و مِینااردوس[1] (۱۹۷۶، قضیهٔ ۴۵ صفحهٔ ۵۷) را ترکیب نمایید. ■

مثال به‌عنوان مثال برای بسط چبیشف دوباره تقریبهای $f(x) = e^x$ را در نظر می‌گیریم. برای e^x، ضرایب c_j از (۱۷.۵.۴) برابرند با

$$\begin{aligned} c_j &= \frac{2}{\pi} \int_{-1}^{1} \frac{e^x T_j(x)}{\sqrt{1-x^2}} dx \\ &= \frac{2}{\pi} \int_0^{\pi} e^{\cos\theta} \cdot \cos(j\theta) d\theta \end{aligned} \qquad (۲۱.۵.۴)$$

فرمول اخیر برای انتگرالگیری عددی بهتر است، زیرا تابع انتگرالده نقطهٔ تکین ندارد. قاعدهٔ میانگاهی (نقطه میانی) یا ذوزنقه‌یی به علّت دوره‌یی‌بودن تابع زیر علامت انتگرال روش بسیار عالی خواهد بود (فرع ۱، قضیهٔ ۵.۵، در بخش ۴.۵ را ببینید). با استفاده از انتگرالگیری عددی مقادیر جدول ۳.۴ را به‌دست می‌آوریم. با به‌کاربردن (۱۷.۵.۴) و فرمول $T_j(x)$ خواهیم داشت

$$C_1(x) = 1.266 + 1.130x$$

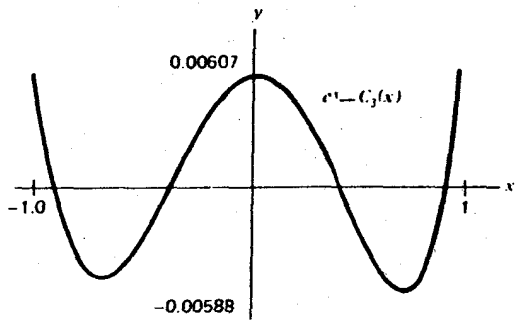
$$C_2(x) = 0.994571 + 0.9973078x + 0.5429911x^2 + 0.1773247x^3$$

$$\| e^x - C_1(x) \|_\infty = 0.32 \qquad \| e^x - C_2(x) \|_\infty = 0.00607 \qquad (۲۲.۵.۴)$$

1. Meinardus

جدول ۳.۴ ضرایب بسط چبیشف برای e^x

j	c_j
۰	۲٫۵۳۲۱۳۱۷۶
۱	۱٫۱۳۰۳۱۸۲۱
۲	۰٫۲۷۱۴۹۵۳۴
۳	۰٫۰۴۴۳۳۶۸۵
۴	۰٫۰۰۵۴۷۴۲۴
۵	۰٫۰۰۰۵۴۲۹۳



شکل ۶.۴ خطا در تقریب کمترین مربعات درجهٔ سوّم چبیشف برای e^x

نمودار $e^x - C_3(x)$ در شکل ۶.۴ داده شده است، و خیلی شبیه به خطای مینیماکس در شکل ۴.۴ است. خطاهای ماکسیمم در این تقریبهای کمترین مربعات چبیشف خیلی نزدیک به خطاهای مینیماکس‌اند و برای اغلب منظورهای عملی کفایت می‌کنند.

چندجمله‌یی $C_n(x)$ به شکل (۱۷.۵.۴) را می‌توان به‌صورت معمول آن، یعنی بر پایهٔ یک‌جمله‌ییهای x^j، درآورد و همان‌گونه که قبلاً در مثال (۲۲.۵.۴) صورت گرفت، آن را محاسبه کرد؛ ولی $C_n(x)$ به شکل (۱۷.۵.۴) سریعتر و دقیقتر محاسبه می‌شود. برای چندجمله‌ییهای $T_n(x)$ چبیشف از رابطهٔ بازگشتی سه‌گانهٔ (۱۳.۴.۴) استفاده می‌کنیم. الگوریتم زیر منسوب به کِلنشاؤ[1] است و ما آن را از صفحهٔ ۱۲۵ کتاب ریولین (۱۹۷۴) اخذ کرده‌ایم.

الگوریتم چبوال (مقدار x، n، a)chebeval

۱. توجه: با این الگوریتم مقدار $\sum_{j=0}^{n} a_j T_j(x)$ محاسبه می‌شود

1. C. W. Clenshaw

۲. $b_{n+۱} = b_{n+۲} := ۰$ ، $z = ۲x$

۳. تا مرحلهٔ ۵ عمل را برای $j = n, n-۱, \ldots, ۰$، انجام دهید.

۴. $b_j = zb_{j+۱} - b_{j+۲} + a_j$

۵. jی بعدی

۶. مقدار $:= (b_۰ - b_۲)/۲$

این الگوریتم تقریباً به همان اندازهٔ الگوریتم ضرب تودرتو (۸.۹.۲) از بخش ۲.۹ کارایی دارد. مقایسهٔ مشروحتر را به مسألهٔ ۲۵ احاله می‌کنیم. الگوریتمهای مشابهی برای سایر بسطهای چندجمله‌ییهای متعامد وجود دارند، که باز هم از رابطهٔ بازگشتی سه‌گانهٔ متناظر استفاده می‌نمایند. برای بررسی خطاهای گردکردن در روند چبوال به (صفحهٔ ۱۲۷) ریولین (۱۹۷۴) و صفحهٔ ۵۷ فکس و پارکر (۱۹۶۸) مراجعه کنید.

۶.۴ تقریبهای مینیماکس

برای یک تقریب یکنواخت خوب تابع داده شدهٔ $f(x)$، منطقی به‌نظر می‌آید که خطا در بازهٔ تقریب یکنواخت توزیع شده باشد. به‌علاوه مثالهای قبلی با $f(x) = e^x$ این نکته را روشن می‌سازند و نشان می‌دهند که خطای ماکسیمم از لحاظ علامت نوسان خواهد کرد. در جدول ۴.۴ آماری از شکلهای گوناگون تقریب e^x در $[-۱, ۱]$ خلاصه شده است که شامل بعضی از روشهای بخش ۷.۴ نیز می‌باشد. برای نشان‌دادن اهمیت توزیع یکنواخت خطا با تابع خطایی که علامت آن نوسان می‌نماید، دو قضیه بیان می‌کنیم. اولین قضیه برای برآورد $\rho_n(x)$، خطای مینیماکس، بدون اینکه مجبور باشیم تقریب مینیماکس $q_n^*(x)$ را پیدا کنیم مفید است.

جدول ۴.۴ مقایسه تقریبهای گوناگون خطی و درجهٔ سوّم برای e^x

روش تقریب	خطای ماکسیمم	
	خطی	درجهٔ سوّم
$\rho_n(x)$، چندجمله‌یی تیلر	۰٫۷۱۸	۰٫۰۵۱۶
$r_n^*(x)$، کمترین مربعات لژاندر	۰٫۴۳۹	۰٫۰۱۱۲
$C_n(x)$، کمترین مربعات چبیشف	۰٫۳۲۲	۰٫۰۰۶۰۷
$I_n(x)$، فرمول درونیابی با نقاط گرهی چبیشف	۰٫۳۷۲	۰٫۰۰۶۶۶
$F_n(x)$، فرمول نوسانی اجباری چبیشف	۰٫۲۸۶	۰٫۰۰۵۵۸
$q_n^*(x)$، مینیماکس	۰٫۲۷۹	۰٫۰۰۵۵۳

قضیهٔ ۹.۴ (دولاواله پوسن)[1] گیریم $f \in C[a,b]$ و $n \geq 0$. فرض کنید یک چندجمله‌یی $Q(x)$ از درجهٔ نابزرگتر از n داریم که در رابطهٔ زیر صدق می‌کند

$$f(x_j) - Q(x_j) = (-1)^j e_j \qquad j = 0, 1, \ldots, n+1 \qquad (۱.۶.۴)$$

که تمام e_jها مخالف صفر و دارای یک علامت‌اند و

$$a \leq x_0 < x_1 < \cdots < x_{n+1} \leq b$$

در این صورت

$$\min_{0 \leq j \leq n+1} | e_j | \leq \rho_n(f) \equiv \| f - q_n^* \|_\infty \leq \| f - Q \|_\infty \qquad (۲.۶.۴)$$

برهان کران بالای (۲.۶.۴) از تعریف $\rho_n(f)$ به‌دست می‌آید. برای اثبات کران پایین، خلاف آن را فرض می‌کنیم و به تناقض می‌رسیم.

فرض کنید

$$\rho_n(f) < \min_{0 \leq j \leq n+1} | e_j | \qquad (۳.۶.۴)$$

لذا طبق تعریف $\rho_n(f)$، یک چندجمله‌یی $\rho(x)$ با درجهٔ نابزرگتر از n وجود دارد که برای آن

$$\rho_n(f) \leq \| f - P \|_\infty < \min | e_j | \qquad (۴.۶.۴)$$

یک چندجمله‌یی با درجهٔ نابزرگتر از n به شکل زیر تعریف می‌کنیم

$$R(x) = Q(x) - P(x)$$

برای سادگی، همهٔ e_j ها را مثبت می‌گیریم؛ استدلال مشابهی برای همهٔ eهای منفی نیز برقرار است. برای هریک از jها، $R(x_j)$ را محاسبه کرده و علامت $R(x_j)$ را در نظر می‌گیریم. ابتدا با استفاده از (۴.۶.۴) داریم؛

$$\begin{aligned} R(x_0) &= Q(x_0) - P(x_0) = [f(x_0) - P(x_0)] - [f(x_0) - Q(x_0)] \\ &= [f(x_0) - P(x_0)] - e_0 < 0 \end{aligned}$$

سپس

$$R(x_1) = Q(x_1) - P(x_1) = [f(x_1) - P(x_1)] + e_1 > 0$$

1. de la Vallée-Poussin

جدول ۵.۴ ماکسیمم‌های نسبی $| e^x - C_۲(x) |$

x	$e^x - C_۲(x)$
−۱٫۰	۰٫۰۰۴۹۷
−۰٫۶۹۱۹	−۰٫۰۰۵۱۱
۰٫۰۳۱۰	۰٫۰۰۵۴۷
۰٫۷۲۲۹	−۰٫۰۰۵۸۸
۱٫۰	۰٫۰۰۶۰۷

به استقرا، علامت $R(x_j)$ برابر است با $(-۱)^{j+۱}$، $j = ۰, ۱, \ldots, n+۱$. پس $R(x)$، $(n+۲)$ بار تغییرعلامت می‌دهد و درنتیجه $R(x)$ دارای $n+۱$ ریشه است. چون درجهٔ $R(x)$ نابزرگتر از n است، این ممکن نخواهد بود مگر آنکه $R(x) \equiv ۰$. پس $P \equiv Q$ که با (۱.۶.۴) و (۴.۶.۴) متناقض است. ■

مثال تقریب کمترین مربعات درجهٔ سوّم چبیشف برای e^x بر $[۱, -۱]$ را که در (۲۲.۵.۴) با $C_n(x)$ نشان داده‌شده بود یادآور می‌شویم. ماکسیمم خطای آن در بازهٔ $[۱, -۱]$ در جدول ۵.۴ داده شده است. این خطاها در مفروضات قضیهٔ ۹.۴ صدق می‌نمایند، و بنابراین

$$۰٫۰۰۴۹۷ \leq \rho_۳(f) \leq ۰٫۰۰۶۰۷$$

مطابق این جدول می توانیم نتیجه‌گیری کنیم که $C_۳(x)$ به بهترین تقریب ممکن خیلی نزدیک بوده است. توجه داشته باشیم که با محاسبهٔ دقیق و استفاده از (۷.۲.۴)، خواهیم داشت $\rho_۳(f) = ۰٫۰۰۵۵۳$.

قضیهٔ ۱۰.۴ (قضیهٔ همنوسانی چبیشف)
گیریم $f \in C[a, b]$ و $n \geq ۰$. آنگاه یک چندجمله‌یی یکتای $q_n^*(x)$ از درجهٔ نابزرگتر از n وجود دارد که برای آن

$$\rho_n(f) = \| f - q_n^* \|_\infty$$

این چندجمله‌یی توسط ویژگی زیر به طور یکتا توصیف می‌شود: حداقل $n+۲$ نقطهٔ

$$a \leq x_۰ < x_۱ < \cdots < x_{n+۱} \leq b$$

وجود دارند که برای آنها

$$f(x_j) - q_n^*(x_j) = \sigma(-1)^j \rho_n(f) \qquad j = 0, 1, \ldots, n+2 \quad (۵.۶.۴)$$

σ برابر ± 1 است و فقط بستگی به f و n دارد.

برهان برهان کاملاً فنی است و به‌صورتی پیچیده از راه برهان خلف ثابت می‌شود. به همین دلیل در اینجا حذف شده است. برای یک بحث کامل (فصل ۷ دیویس[1] (۱۹۶۳) را ببینید). ■

مثال تقریب مینیماکس درجهٔ سوم $q_3^*(x)$ برای e^x که در (۷.۲.۴) داده شده بود، در احکام این قضیه صدق می‌نماید، چنانچه می‌توان از نمودار خطا در شکل ۴.۴ بخش ۴.۲ ملاحظه کرد.

با توجه به این قضیه می‌توانیم ببینیم که سری تیلر همیشه یک تقریب یکنواخت ضعیف است. خطای سری تیلر

$$f(x) - p_n(x) = \frac{(x - x_0)^{n+1}}{(n+1)!} f^{n+1}(\xi_x) \qquad (۶.۶.۴)$$

در بازهٔ تقریب به طور یکنواخت تغییر نمی‌کند و علامت آن هم خیلی نوسان نمی‌کند.

برای اینکه بهتر نشان دهیم که $q_n^*(x)$ چگونه، با افزایش n، $f(x)$ را به خوبی تقریب می‌زند قضیهٔ زیر از د. جکسن[2] را می‌آوریم.

قضیهٔ ۱۱.۴ (جکسن) گیریم $f(x)$ به ازای مقداری از $k \geq 0$ دارای k مشتق پیوسته باشد. به‌علاوه فرض می‌کنیم که $f^{(k)}(x)$، به ازای مقداری مانند $M > 0$ و $0 < \alpha \leq 1$، در رابطهٔ زیر صدق می‌کند

$$\operatorname*{Sup}_{a \leq x, y \leq b} | f^{(k)}(x) - f^{(k)}(y) | \leq M | x - y |^{\alpha} \qquad (۷.۶.۴)$$

[می‌گوییم $f^{(k)}(x)$ با نمای α در شرط هولدر صدق می‌کند]. در این صورت ثابتی چون d_k، مستقل از f و n موجود است که برای آن

$$\rho_n(f) \leq \frac{M d_k}{n^{k+\alpha}} \qquad n \geq 1 \qquad (۸.۶.۴)$$

برهان میناردوس (۱۹۶۷، قضیهٔ ۴۵ ص ۵۷) را ببینید. توجه نمایید که اگر بخواهیم (۷.۶.۴) و فرض k بار پیوسته مشتقپذیری تابع $f(x)$ را نادیده بگیریم، آنگاه فقط $k-1$ را به جای k در

1. Davis 2. Jackson

جدول ۶.۴ مقایسهٔ ρ_n و M_n برای $f(x) = e^x$

n	۲	۳	۴	۵	۶	۷
$M_n(f)$	۱٫۱۳E − ۱	۱٫۴۲E − ۲	۱٫۴۲E − ۳	۱٫۱۸E − ۴	۸٫۴۳E − ۶	۵٫۲۷E − ۷
$\rho_n(f)$	۴٫۵۰E − ۲	۵٫۵۳E − ۳	۵٫۴۷E − ۴	۴٫۵۲E − ۵	۳٫۲۱E − ۶	۲٫۰۰E − ۷

قضیه به‌کار می‌بریم با $\alpha = ۱$ و $M = \| f^{(k)} \|_\infty$. در این صورت حکم قضیه چنین می‌شود

$$\rho_n(f) \leq \frac{d_{k-1}}{n^k} \| f^{(k)} \|_\infty \qquad (۹.۶.۴)$$

همچنین باید توجه کرد که اگر f بینهایت بار مشتقپذیر باشد، آنگاه $q_n^*(x)$، سریعتر از هر توانی از $\frac{1}{n^k}$، $k \geq ۱$، در بازهٔ $[a, b]$ به‌طور یکنواخت به $f(x)$ می‌گراید. ■

با توجّه به قضیهٔ ۱۲.۴ در بخش بعد، می‌توانیم قضیهٔ زیر را ثابت کنیم. اگر $f(x)$ بر $[a, b]$، $n + ۱$ بار پیوسته مشتقپذیر باشد، آنگاه

$$\rho_n(f) \leq \frac{[(b-a)/۲]^{n+۱}}{(n+۱)!۲^n} \| f^{(n+۱)} \|_\infty \equiv M_n(f) \qquad (۱۰.۶.۴)$$

اقامهٔ برهان به صورت مسألهٔ ۳۸ به خواننده واگذار شده است. توابع بینهایت مشتقپذیری وجود دارند که برای آنها $M_n(f) \to \infty$. معذالک، برای بیشتر توابعی که زیاد به‌کار می‌روند، کران (۱۰.۶.۴) براورد نسبتاً دقیقی از اندازهٔ $\rho_n(f)$، به‌نظر می‌رسد. این براورد برای $f(x) = e^x$ بر بازهٔ $[-۱, ۱]$ در جدول ۶.۴ نشان داده شده است. برای براوردها و کرانهای دیگر $\rho_n(f)$، مینارودوس (۱۹۶۷، بخش ۲.۶) را ببینید.

۷.۴ تقریبهای نزدیکِ مینیماکس

با توجه به قضیهٔ همنوسانی چبیشف، می‌توانیم روشهایی به‌دست آوریم که اغلب براورد خوبی برای تقریب مینیماکس بدهند. با تقریب کمترین مربعات $C_n(x)$ از (۱۷.۵.۴) شروع می‌کنیم. این تقریب اغلب براورد خوبی برای $q_n^*(x)$ است و ویژگیهای $C_n(x)$ انگیزهٔ دیگر تقریبهای نزدیکِ مینیماکس ما هستند.

با توجّه به (۱۷.۵.۴) داریم،

$$C_n(x) = \sum_{j=۰}^{n}{}' c_j T_j(x) \qquad c_j = \frac{۲}{\pi} \int_{-۱}^{۱} \frac{f(x) T_j(x) dx}{\sqrt{۱ - x^۲}} \qquad (۱.۷.۴)$$

که پریم در نماد مجموعیابی بدین معناست که اولین جمله ($j = ۰$) قبل از جمع‌کردن سری باید نصف شود. اگر $f \in C[-۱, ۱]$، آنگاه با استفاده از (۷.۵.۴)

$$f(x) = \sum_{j=۰}^{\infty}{}' c_j T_j(x) \qquad (۲.۷.۴)$$

با برقراری همگرایی به این معنی که

$$\lim_{n\to\infty} \int_{-۱}^{۱} \frac{۱}{\sqrt{۱-x^۲}} \left[f(x) - \sum_{j=۰}^{n}{}' c_j T_j(x) \right]^۲ dx = ۰$$

برای همگرایی یکنواخت، قضیهٔ نسبتاً قوی زیر را داریم

$$\rho_n(f) \leq \| f - C_n \|_\infty \leq \left(۴ + \frac{۴}{\pi^۲} \ln(n) \right) \rho_n(f) \qquad (۳.۷.۴)$$

برای اثبات، ریولین (۱۹۷۴ ص ۱۳۴) را ببینید. از ترکیب این رابطه با قضیهٔ (۹.۶.۴) جکسن، کران همگرایی (۲۰.۵.۴) قبلی از قضیهٔ ۸.۴ به‌دست می‌آید.

اگر $f \in C^r[a, b]$، می‌توان ثابت کرد که مقدار ثابتی چون c وابسته فقط به f و r، وجود دارد که برای آن

$$| c_j | \leq \frac{c}{j^r} \qquad j \geq ۱ \qquad (۴.۷.۴)$$

اثبات این فرمول، با در نظر گرفتن c_jها به‌عنوان ضرایب فوریهٔ $f(\cos\theta)$ و سپس استفاده از قضایای سری فوریه راجع به نرخ کاهش این ضرایب انجام می‌گیرد. بنابراین وقتی r بزرگ می‌شود، ضرایب c_j سریعتر کاهش می‌یابند.

برای بسط منقطع $C_n(x)$، اگر $c_{j+۱} \neq ۰$، و ضرایب c_j به سرعت به صفر بگرایند، داریم

$$f(x) - C_n(x) = \sum_{n+۱}^{\infty} c_j T_j(x) \doteq c_{n+۱} T_{n+۱}(x) \qquad (۵.۷.۴)$$

از تعریف $T_{n+۱}$ داریم

$$| T_{n+۱}(x) | \leq ۱ \qquad -۱ \leq x \leq ۱ \qquad (۶.۷.۴)$$

همچنین برای $n+۲$ نقطهٔ

$$x_j = \cos\left(\frac{j\pi}{n+۱} \right) \qquad j = ۰, ۱, \ldots, n+۱ \qquad (۷.۷.۴)$$

داریم

$$T_{n+1}(x_j) = (-1)^j \qquad (۸.۷.۴)$$

کران (۶.۷.۴) دقیقاً به $n+2$ نقطه، یعنی ماکسیمم تعداد ممکن می‌رسد. با استفاده از این رابطه در (۵.۷.۴)، جملهٔ $c_{n+1}T_{n+1}$ دقیقاً $n+2$ ماکسیمم و مینیمم نسبی دارد که همگی اندازهٔ مساوی دارند. بنابراین، از قضیهٔ همنوسانی چبیشف می‌توانیم انتظار داشته باشیم که $C_n(x)$ خیلی نزدیک به تقریب مینیماکس $q_n^*(x)$ باشد.

مثال مثال (۲۲.۵.۴) برای $f(x) = e^x$ نزدیک به آخر بخش ۵.۴ را به یاد آورید. در آنجا، ضرایب c_j خیلی سریع کاهش می‌یافتند، و

$$\| e^x - C_3 \|_\infty = ۰٫۰۰۶۰۷ \qquad c_4T_4(x) = ۰٫۰۰۵۴۷T_4(x)$$
$$\| e^x - q_3^*(x) \|_\infty = ۰٫۰۰۵۵۳$$

توضیحات پاراگراف اخیر با این مثال روشن می‌شود.

مثال می‌توان نشان داد که

$$\tan^{-1} x = 2\left[\alpha T_1(x) - \frac{\alpha^3}{3}T_3(x) + \frac{\alpha^5}{5}T_5(x) - \dots\right] \qquad (۹.۷.۴)$$

برای $-1 \le x \le 1$ با $\alpha = \sqrt{2} - 1 \doteq ۰٫۴۴$، همگرای یکنواخت است. پس

$$C_{2n+1}(x) = 2\left[\alpha T_1(x) - \frac{\alpha^3}{3}T_3(x) + \dots + \frac{(-1)^n}{2n+1}\alpha^{2n+1}T_{2n+1}(x)\right] \qquad (۱۰.۷.۴)$$

برای خطا،

$$E_{2n+1}(x) \equiv \tan^{-1} x - C_{2n+1}(x) = \frac{2}{2n+3}(-1)^{n+1}\alpha^{2n+3}T_{2n+3}(x)$$
$$+2\sum_{j=n+2}^{\infty}\frac{(-1)^j\alpha^{2j+1}}{2j+1}T_{2j+1}(x)$$

کرانهای بالا و پایین این عبارت را پیدا می‌کنیم تا خطای $C_{2n+1}(x)$ و خطای مینیماکس $\rho_{2n+1}(f)$ را تخمین بزنیم.

با گرفتن کرانهای بالا

$$\left|2\sum_{n+2}^{\infty}\frac{(-1)^j\alpha^{2j+1}}{2j+1}T_{2j+1}(x)\right| \le 2\sum_{n+2}^{\infty}\frac{\alpha^{2j+1}}{2j+1} < \frac{2}{2n+5}\cdot\frac{\alpha^{2n+5}}{1+\alpha^2}$$

بنابراین کرانهای بالا و پایین E_{2n+1} را، با توجه به اینکه $\alpha^2/(1-\alpha^2) \doteq 0.207$، به‌دست می‌آوریم،

$$E_{2n+1}(x) \le \frac{2}{2n+3}(-1)^{n+1}\alpha^{2n+3}T_{2n+3}(x) + \frac{2\alpha^{2n+5}}{(2n+5)(1-\alpha^2)}$$
$$\le \frac{2\alpha^{2n+3}}{2n+3}\left[(-1)^{n+1}T_{2n+3}(x) + 0.207\right]$$

همچنین

$$E_{2n+1}(x) \ge \frac{\alpha^{2n+3}}{2n+3}\left[(-1)^{n+1}T_{2n+3}(x) - 0.207\right]$$

بنابراین

$$E_{2n+1}(x) \doteq \frac{\alpha^{2n+3}}{2n+3}(-1)^{n+1}T_{2n+3}(x) \qquad (11.7.4)$$

که دارای $2n+4$ ماکسیمم و مینیمم نسبی با علامتهای یک در میان مثبت و منفی است. با توجه به قضیهٔ ۹.۴ و نامساویهای قبلی

$$\frac{2(0.793)\alpha^{2n+3}}{2n+3} \le \rho_{2n+1}(f) = \rho_{2n+2}(f) \le \frac{2(1.207)\alpha^{2n+3}}{2n+3} \qquad (12.7.4)$$

برای محاسبهٔ عملی ضرایب c_j در (۱.۷.۴) از فرمول

$$c_j = \frac{2}{\pi}\int_0^\pi \cos(j\theta)f(\cos\theta)d\theta \qquad (13.7.4)$$

و قاعدهٔ انتگرالگیری عددی میانگاهی یا ذوزنقه‌یی استفاده می‌کنیم. این مطالب قبلاً در (۲۱.۵.۴) نشان داده شده است.

درونیابی در صفرهای چبیشف اگر خطا در تقریب مینیماکس تقریباً برابر $c_{n+1}T_{n+1}(x)$ باشد، همان‌گونه که در (۵.۷.۴) نشان داده شد، خطا در ریشه‌های $T_{n+1}(x)$ در $[-1, 1]$ باید نزدیک صفر باشد. این ریشه‌ها نقاط زیرند

$$x_j = \cos\left[\frac{2j+1}{2n+2}\pi\right] \qquad j = 0, 1, \ldots, n \qquad (14.7.4)$$

گیریم $I_n(x)$ یک چندجمله‌یی با درجهٔ نابزرگتر از n باشد که $f(x)$ را در این گرههای $\{x_j\}$ درونیابی نماید. چون خطای این درونیاب در x_jها، صفرهای چبیشف $T_{n+1}(x)$ برابر صفر است، پیوستگی چندجمله‌یی درونیابی نسبت به مقادیر تابع که چندجمله‌یی را تعریف می‌کنند، گویای آن است که $I_n(x)$ باید تقریباً برابر $C_n(x)$ و لذا همچنین تقریباً برابر $q_n^*(x)$ باشد. به عبارت دقیقتر،

شکل لاگرانژی $I_n(x)$ را نوشته و باعملیاتی روی آن به‌دست می‌آوریم:

$$\begin{aligned} I_n(x) &= \sum_{j=0}^{n} f(x_j) l_j(x) \\ &= \sum_{j=0}^{n} C_n(x_j) l_j(x) + \sum_{j=0}^{n} [f(x_j) - C_n(x_j)] l_j(x) \\ &\doteq C_n(x) \end{aligned}$$

$$I_n(x) \doteq C_n(x) \qquad (۱۵.۷.۴)$$

زیرا $f(x_j) - C_n(x_j) \doteq c_{n+1} T_{n+1}(x_j) = 0$. عبارت $I_n(x)$ را می‌توان با الگوریتم‌های $Divdif$ و $Interp$ بخش ۲.۳ حساب کرد. مواظب باشید: اگر $C_{n+1} = 0$، خطای $f(x) - q_n^*(x)$ احتمالاً صفرهای تقریبی یک چندجمله‌یی از درجهٔ بالاتر معمولاً $T_{n+2}(x)$ چبیشف را خواهد داشت. اگر $f(x)$ در $[-1, 1]$ یا فرد باشد یا زوج این حالت ممکن است اتفاق بیفتد (مسألهٔ ۲۹ را ببینید). تابع $f(x)$ را بر بازهٔ $[-a, a]$ زوج گویند اگر به‌ازای تمام مقادیر x در $[-a, a]$، $f(-x) = f(x)$، فرد گویند اگر $f(-x) = -f(x)$. مثال قبلی با $f(x) = \tan^{-1}(x)$ یک نمایش خوبی از موضوع است. انگیزهٔ دیگر درنظرگرفتن $I_n(x)$ به‌عنوان یک تقریب نزدیکِ مینیماکس مبتنی بر قضیهٔ مهم زیر دربارهٔ چندجمله‌ییهای چبیشف است.

قضیهٔ ۱۲.۴ مسألهٔ مینیم‌سازی

$$\tau_n = \inf_{\deg(Q) \le n-1} \left[\operatorname*{Max}_{-1 \le x \le 1} | x^n + Q(x) | \right] \qquad (۱۶.۷.۴)$$

را که در آن $Q(x)$ یک چندجمله‌یی است برای یک عدد صحیح ثابت $n > 0$ در نظر می‌گیریم. اگر $Q(x)$ را به‌طور ضمنی با

$$x^n + Q(x) = \frac{1}{2^{n-1}} T_n(x) \qquad (۱۷.۷.۴)$$

تعریف کنیم مینیمم τ_n به طور یکتا حاصل می‌شود. این مینیمم برابر است با

$$\tau_n = \frac{1}{2^{n-1}} \qquad (۱۸.۷.۴)$$

برهان نخست بعضی واقعیتها در مورد چندجمله‌ییهای چبیشف را درنظر می‌گیریم. بنابه تعریف، $T_0(x) \equiv 1$ و $T_1(x) = x$. رابطهٔ بازگشتی سه‌گانهٔ

$$T_{n+1}(x) = 2x T_n(x) - T_{n-1}(x)$$

پایهٔ یک برهان استقرایی است برای

$$T_n(x) = ۲^{n-۱}x^n + \text{جملات از درجهٔ پایینتر} \qquad n \geq ۱ \qquad (۱۹.۷.۴)$$

بنابراین

$$\frac{۱}{۲^{n-۱}}T_n(x) = x^n + \text{جملات از درجات پایینتر} \qquad n \geq ۱ \qquad (۲۰.۷.۴)$$

چون $T_n(x) = \cos n\theta$، $x = \cos\theta$، $۰ \leq \theta \leq \pi$، چندجمله‌یی $T_n(x)$ در $n+۱$ نقطه در $[-۱, ۱]$ ماکسیمم و مینیمم می‌شود:

$$x_j = \cos\left(\frac{j\pi}{n}\right) \qquad j = ۰, ۱, \ldots, n \qquad (۲۱.۷.۴)$$

چون تابع کسینوس دوره‌یی است، برای مقادیر دیگر j مقادیر جدیدی برای x_j به‌دست نمی‌آید. برای این نقاط

$$T_n(x_j) = (-۱)^j \qquad j = ۰, ۱, \ldots, n \qquad (۲۲.۷.۴)$$

و

$$-۱ = x_n < x_{n-۱} < \cdots < x_۱ < x_۰ = ۱$$

ضریب جملهٔ بزرگترین درجهٔ چندجمله‌یی $T_n(x)/۲^{n-۱}$ برابر یک است و

$$\underset{-۱\leq x\leq ۱}{\text{Max}} \mid \frac{۱}{۲^{n-۱}}T_n(x) \mid = \frac{۱}{۲^{n-۱}} \qquad (۲۳.۷.۴)$$

بنابراین $\tau_n < ۱/۲^{n-۱}$. فرض می‌کنیم که

$$\tau_n < \frac{۱}{۲^{n-۱}} \qquad (۲۴.۷.۴)$$

نشان می‌دهیم که این فرض به تناقض می‌انجامد. فرض (۲۴.۷.۴) و تعریف (۱۶.۷.۴) دلالت بر وجود یک چندجمله‌یی به صورت زیر دارند

$$M(x) = x^n + Q(x) \qquad \text{درجهٔ } Q(x) \text{ نابزرگتر از } n-۱$$

با

$$\tau_n \leq \underset{-۱\leq x\leq ۱}{\text{Max}} \mid M(x) \mid < \frac{۱}{۲^{n-۱}} \qquad (۲۵.۷.۴)$$

تعریف می‌کنیم

$$R(x) = \frac{۱}{۲^{n-۱}}T_n(x) - M(x)$$

که دارای درجهٔ نابزرگتر از $n-1$ است. علامت $R(x_j)$ را در نقاط (۲۱.۷.۴) امتحان می‌کنیم. با استفاده از (۲۲.۷.۴) و (۲۴.۷.۴)

$$R(x_0) = R(1) = \frac{1}{2^{n-1}} - M(1) > 0$$
$$R(x_1) = -\frac{1}{2^{n-1}} - M(x_1) = -\left[\frac{1}{2^{n-1}} + M(x_1)\right] < 0$$

و علامت $R(x_j)$ برابر $(-1)^j$ است. چون R دارای $n+1$ تغییر علامت است، باید حداقل n صفر داشته باشد. ولی در این صورت درجهٔ R که کوچکتر از n است، ایجاب می‌کند که $R \equiv 0$؛ پس $M \equiv (1/2^{n-1})T_n$.

برای اثبات اینکه هیچ چندجمله‌یی دیگری غیر از $1/2^{n-1}T_n(x)$، (۱۶.۷.۴) را مینیمم نمی‌کند، یک شکل دیگری از برهان قبل به‌کار برده می‌شود. ما از ذکر این برهان خودداری می‌کنیم. ■

اکنون مسألهٔ تعیین $n+1$ نقطهٔ گرهی x_j در $[-1, 1]$ را در نظر می‌گیریم، که باید در ساختن یک چندجمله‌یی درونیاب $p_n(x)$ که تابع مفروض $f(x)$ را در $[-1, 1]$ تقریب می‌زند، مورد استفاده قرار گیرد. خطا در $p_n(x)$ برابر است با

$$f(x) - p_n(x) = \frac{(x-x_0)\dots(x-x_n)}{(n+1)!} f^{(n+1)}(\xi_x) \qquad (۲۶.۷.۴)$$

مقدار $f^{n+1}(\xi_x)$ به $\{x_j\}$ بستگی دارد، ولی این بستگی چنان نیست که بتوان به صراحت به آن پرداخت. پس برای کوچک ساختن $\| f - p_n \|_\infty$ در حد ممکن، فقط کمیت

$$\underset{-1\le x\le 1}{\text{Max}} \; | (x-x_0)\dots(x-x_n) | \qquad (۲۷.۷.۴)$$

را در نظر می‌گیریم. $\{x_j\}$ را طوری انتخاب می‌کنیم که این کمیت را مینیمم سازد. چندجمله‌یی (۲۷.۷.۴) از درجهٔ $n+1$ و ضریب بزرگترین درجهٔ x در آن برابر یک است. طبق قضیهٔ قبل، (۲۷.۷.۴) وقتی مینیمم می‌شود که آن را چندجمله‌یی $T_{n+1}(x)/2^n$ انتخاب کنیم و مقدار مینیمم (۲۷.۷.۴) برابر $1/2^n$ است. گرههای $\{x_j\}$ صفرهای $T_{n+1}(x)$ اند و این نقاط در (۱۴.۷.۴) داده شده‌اند. با این انتخاب گرهها، $p_n = I_n$ و

$$\| f - I_n \|_\infty \le \frac{1}{(n+1)!2^n} \| f^{(n+1)} \|_\infty \qquad (۲۸.۷.۴)$$

مثال گیریم $f(x) = e^x$ و $x = 3$. صورت تفاضلات منقسم چندجمله‌یی درونیاب نیوتن را به‌کار می‌بریم. گره‌ها، مقادیر تابع، و تفاضلات منقسم لازم در جدول ۷.۴ داده شده‌اند. با محاسبهٔ

جدول ۷.۴ داده‌های درونیابی برای $f(x) = e^x$

i	x_i	$f(x_i)$	$f[x_0, \ldots, x_i]$
۰	۰٫۹۲۳۸۸۰	۲٫۵۱۹۰۴۴۲	۲٫۵۱۹۰۴۴۲
۱	۰٫۳۸۲۶۸۳	۱٫۴۶۶۲۱۳۸	۱٫۹۴۵۳۷۶۹
۲	−۰٫۳۸۲۶۸۳	۰٫۶۸۲۰۲۸۸	۰٫۷۰۴۷۴۲۰
۳	۰٫۹۲۳۸۸۰	۰٫۳۹۶۹۷۶۰	۰٫۱۷۵۱۷۵۷

مستقیم داریم

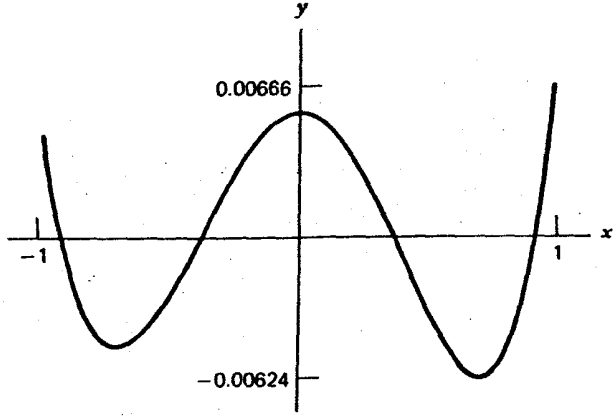
$$\underset{-1 \le x \le 1}{\text{Max}} \mid e^x - I_2(x) \mid \doteq ۰٫۰۰۶۶۶ \qquad (۲۹.۷.۴)$$

و حال آنکه کران در (۲۸.۷.۴) برابر ۰٫۰۰۱۴ است. نمودار $e^x - I_2(x)$ در شکل ۷.۴ آمده است.

خطای $\| f - I_n \|_\infty$ معمولاً خیلی بدتر از $\rho_n(f)$ نیست. یک نتیجهٔ دقیق چنین است

$$\| f - I_n \|_\infty \le \left[\frac{2}{\pi} \log(n+1) + 2\right] \rho_n(f) \qquad n \ge 0 \quad (۳۰.۷.۴)$$

برهان را در ریولین (۱۹۷۰، ص ۱۳) ببینید. همان‌گونه که (۲۹.۷.۴) در مثالهای قبلِ $f(x) = e^x$ نشان می‌دهد، نتایج واقعی عددی معمولاً بهتر از آن است که با کران (۲۹.۷.۴) پیش‌بینی می‌شود.



شکل ۷.۴ $e^x - I_2(x)$

نوسان اجباری خطا گیریم $f(x) \in C[-۱, ۱]$، و تعریف می‌کنیم

$$F_n(x) = \sum_{k=۰}^{n}{}' c_{n,k} T_k(x) \qquad -۱ \le x \le ۱ \qquad (۳۱.۷.۴)$$

گیریم

$$-۱ \le x_{n+۱} < x_n < \cdots < x_۱ < x_۰ \le ۱$$

گره‌هایی باشند که در مورد انتخاب آنها ذیلاً توضیح داده می‌شود. ضرایب $c_{n,k}$ را چنان اختیار می‌کنیم که خطای $f(x) - F_n(x)$ به طریقی نوسان کند که در قضیهٔ ۱۰.۴ لازم فرض شده است:

$$f(x_i) - F_n(x_i) = (-۱)^i E_n \qquad i = ۰, ۱, \ldots, n+۱ \qquad (۳۲.۷.۴)$$

مجهول دیگر E_n را وارد کرده‌ایم که امیدواریم صفر نباشد. $n+۲$ مجهول داریم: $c_{n,۰}$، $c_{n,۱}$، ...، $c_{n,n}$ و E_n، و $n+۲$ معادله در (۳۲.۷.۴) موجودند. پس برای اینکه یک جواب داشته باشیم، شانس زیادی وجود دارد. اگر یک جواب موجود باشد، طبق قضیه ۹.۴،

$$| E_n | \le \rho_n(f) \le \| f - F_n \|_\infty \qquad (۳۳.۷.۴)$$

برای انتخاب نقاط گرهی، توجه کنید اگر $c_{n+۱} T_{n+۱}(x)$ در (۵.۷.۴) تقریباً خطای مینیماکس باشد، آنگاه مینیمم‌ها و ماکسیمم‌های نسبی در خطای مینیماکس $f(x) - q_n^*(x)$ در مینیمم‌ها و ماکسیمم‌های نسبی $T_{n+۱}(x)$ واقع می‌شوند. این ماکسیمم‌ها و مینیمم‌ها وقتی اتفاق می‌افتند که $T_{n+۱}(x) = \pm ۱$ و با رابطهٔ زیر داده می‌شوند

$$x_i = \cos\left(\frac{i\pi}{n+۱}\right) \qquad i = ۰, ۱, \ldots, n+۱ \qquad (۳۴.۷.۴)$$

به‌نظر می‌آید که اگر $F_n(x)$ تقریب مینیماکس $q_n^*(x)$ اطلاق شود، نقاط فوق یک انتخاب عالی برای استفاده در (۳۲.۷.۴) باشند.

دستگاه (۳۲.۷.۴) چنین می‌شود

$$\sum_{k=۰}^{n}{}' c_{n,k} T_k(x_i) + (-۱)^i E_n = f(x_i) \qquad i = ۰, ۱, \ldots, n+۱ \qquad (۳۵.۷.۴)$$

توجه داشته باشید که

$$T_k(x_i) = \cos(k . \cos^{-۱}(x_i)) = \cos\left(\frac{ki\pi}{n+۱}\right)$$

E_n را برابر $c_{n,n+1}/2$ اختیار می‌کنیم. دستگاه (۳۵.۷.۴) چنین می‌شود

$$\sum_{k=0}^{n+1}{}''c_{n,k}\cos\left(\frac{ki\pi}{n+1}\right)=f(x_i)\qquad i=0,1,\ldots,n+1 \qquad (۳۶.۷.۴)$$

زیرا

$$\cos\left(\frac{ki\pi}{n+1}\right)=(-1)^i \qquad \text{برای } k=n+1$$

نماد $\sum''$ بدین معناست که اولین و آخرین جمله، قبل از شروع به مجموعیابی باید نصف شوند. برای حل (۳۶.۷.۴) به روابط زیر نیاز داریم:

$$\sum_{i=0}^{n+1}{}''\cos\left(\frac{ij\pi}{n+1}\right)\cos\left(\frac{ik\pi}{n+1}\right)=\begin{cases}n+1 & j=k=0 \text{ یا } n+1\\ \frac{(n+1)}{2} & 0<j=k<n+1\\ 0 & j\neq k \quad 0\le j \quad k\le n+1\end{cases} \qquad (۳۷.۷.۴)$$

اثبات این روابط به اثبات روابط در مسألهٔ ۴۲ فصل ۳، بستگی نزدیکی دارد.

معادلهٔ iام در (۳۶.۷.۴) را در $\cos[\frac{(ij\pi)}{(n+1)}]$، به ازای یک $0\le j\le n+1$، ضرب و سپس روی i مجموعیابی می‌کنیم، با توجه به نصف‌کردن جملات اول و آخر، این نتیجه به‌دست می‌آید:

$$\sum_{k=0}^{n+1}{}''c_{n,k}\sum_{i=0}^{n+1}{}''\cos\left(\frac{ij\pi}{n+1}\right)\cos\left(\frac{ik\pi}{n+1}\right)=\sum_{i=0}^{n+1}{}''f(x_i)\cos\left(\frac{ij\pi}{n+1}\right)$$

با استفاده از رابطهٔ (۳۷.۷.۴)، تمام جملات غیر از یکی در مجموعیابی روی k صفر می‌شوند. با امتحان کردن دو حالت $j=0,n+1$، $0<j<n+1$ همان فرمول به‌دست می‌آید:

$$c_{n,j}=\frac{2}{n+1}\sum_{i=0}^{n+1}{}''f(x_i)\cos\left(\frac{ij\pi}{n+1}\right)\qquad 0\le j\le n+1 \qquad (۳۸.۷.۴)$$

فرمول برای E_n چنین است

$$E_n=\frac{1}{n+1}\sum_{i=0}^{n+1}{}''(-1)^i f(x_i) \qquad (۳۹.۷.۴)$$

بین این تقریب $E_n(x)$ و بسط $C_n(x)$ چبیشف چند رابطه وجود دارد. مهمترین آنها، ضرایب $c_{n,j}$ تقریبهای c_j در $C_n(x)$ هستند. فرمول (۱۳.۷.۴) را برای c_j با استفاده از قاعدهٔ

ذوزنقه‌یی (۵.۱.۵) با $n+۱$ زیر بازه محاسبه می‌کنیم:

$$c_j = \frac{۲}{\pi}\int_{۰}^{\pi} f(\cos\theta)\cos(j\theta)d\theta$$
$$\doteq \frac{۲}{\pi}\sum_{i=۰}^{n+۱}{}''f\left(\cos\left(\frac{i\pi}{n+۱}\right)\right)\cdot\cos\left(\frac{ji\pi}{n+۱}\right)\cdot\frac{\pi}{n+۱} = c_{n,j} \qquad (۴۰.۷.۴)$$

به خوبی می‌دانیم که وقتی انتگرالده دوره‌یی باشد، قاعدهٔ انتگرالگیری عددی ذوزنقه‌یی خیلی دقیق است (قضیهٔ ۵.۵ بخش ۴.۵ را ببینید.) به‌علاوه می‌توان نشان داد که

$$c_{nj} = c_j + c_{۲(n+۱)-j} + c_{۲(n+۱)+j} + c_{۴(n+۱)-j} + \dots \qquad (۴۱.۷.۴)$$

اگر ضرایب چبیشف در (۲.۷.۴) سریعاً کاهش یابند، تقریب $F_n(x)$ تقریباً برابر $C_n(x)$ می‌شود و محاسبه ساده‌تر از محاسبهٔ $C_n(x)$ انجام می‌گیرد.

مثال مانند قبل، $f(x)=e^x$، $-۱\le x\le ۱$، را به‌کار می‌بریم. به ازای $n=۱$ گرهها چنین‌اند

$$\{x_i\} = \{-۱, ۰, ۱\}$$

و $E_۱ = ۰٫۲۷۲$

$$F_۱(x) = ۱٫۲۷۱۵ + ۱٫۱۷۵۲x \qquad (۴۲.۷.۴)$$

برای خطا، ماکسیمم‌های نسبی $e^x - F_۱(x)$ در جدول ۸.۴ داده شده‌اند. برای $n=۳$، $E_۳ = ۰٫۰۰۵۴۷$ و

$$F_۲(x) = ۰٫۹۹۴۵۲۶ + ۰٫۹۹۵۶۸۲x + ۰٫۵۴۳۰۸۱x^۲ + ۰٫۱۷۹۵۱۹x^۳ \qquad (۴۳.۷.۴)$$

نقاط خطای ماکسیمم در جدول ۹.۴ داده شده‌اند. از قضیهٔ ۹.۴ چنین به دست می‌آوریم:

$$۰٫۰۰۵۴۷ \le \rho_۳(f) \le ۰٫۰۰۵۵۸$$

و این رابطه می‌گوید که $F_۳(x)$ یک تقریب عالی برای $q_۴^*(x)$ است.

برای آنکه ببینیم $F_۳(x)$ یک تقریب $C_۳(x)$ است، ضرایب $c_{۳,j}$ را با ضرایب c_j که در جدول ۳.۴ در مثال (۲۲.۵.۴) در پایان بخش ۵.۴ داده‌شده، مقایسه می‌کنیم. نتایج در جدول ۱۰.۴ داده شده‌اند.

جدول ۸.۴ ماکسیمم‌های نسبی $|\, e^x - F_1(x) \,|$

x	$e^x - F_1(x)$
−۱٫۰	۰٫۲۷۲
۰٫۱۶۱۴	−۰٫۲۸۶
۱٫۰	۰٫۲۷۲

جدول ۹.۴ ماکسیمم‌های نسبی $|\, e^x - F_2(x) \,|$

x	$e^x - F_2(x)$
−۱٫۰	۰٫۰۰۵۴۷
−۰٫۶۸۳۲	−۰٫۰۰۵۵۲
۰٫۰۴۹۳	۰٫۰۰۵۵۸
۰٫۷۳۲۴	−۰٫۰۰۵۵۴
۱٫۰	۰٫۰۰۵۴۷

مانند حالت چندجمله‌یی تقریب درونیاب $I_n(x)$، اگر $f(x)$ در $[-۱, ۱]$ فرد یا زوج باشد باید مواظب بود. در چنین حالتی، n را به شکل زیر انتخاب می‌کنیم.

(۴۴.۷.۴) اگر f $\left\{\begin{matrix}\text{زوج}\\ \text{فرد}\end{matrix}\right.$ باشد، آنگاه n را $\left\{\begin{matrix}\text{فرد}\\ \text{زوج}\end{matrix}\right.$ می‌گیریم.

این انتخاب تضمین می‌کند که در (۵.۷.۴)، $c_{n+1} \neq ۰$، و گره‌های انتخابی درست خواهند بود. یک تحلیل همگرایی $F_n(x)$ به $f(x)$ در شمپاین[1] (۱۹۷۰) داده شده است، که کرانی مانند

جدول ۱۰.۴ ضرایب بسط $c_2(x)$ و $F_2(x)$ برای e^x

j	c_j	$c_{n,j}$
۰	۲٫۵۳۲۱۳۱۷۶	۲٫۵۳۲۱۳۲۱۵
۱	۱٫۱۳۰۳۱۸۲۱	۱٫۱۳۰۳۲۱۴۲
۲	۰٫۲۷۱۴۹۵۳۴	۰٫۲۷۱۵۴۰۳۲
۳	۰٫۰۴۴۳۳۶۸۵	۰٫۰۴۴۸۷۹۷۸
۴	۰٫۰۰۵۴۷۴۲۴	$E_4 = ۰٫۰۰۵۴۷۴۲۴$

1. Shampine

کران (۳۰.۷.۴) برای $I_n(x)$ به‌دست می‌دهد:

$$\| f - F_n \|_\infty \leq w(n)\rho_n(f) \qquad (۴۵.۷.۴)$$

که $w(n)$ از لحاظ تجربی تقریباً مساوی ضریب کرانی در (۳۰.۷.۴) است. $I_n(x)$ و $F_n(x)$، هر دو، تقریبهای عملی نزدیک مینیماکس هستند.

اکنون یک الگوریتم برای محاسبهٔ $F_n(x)$ می‌دهیم که می‌توان آن را با استفاده از الگوریتم چبوال بخش ۵.۴ محاسبه کرد.

الگوریتم Approx (c, E, f, n)

۱. تبصره: با این الگوریتم ضرایب c_j در

$$F_n(x) = \sum_{j=0}^{n}{}' c_j T_j(x) \qquad -1 \leq x \leq 1$$

برطبق فرمول (۳۸.۷.۴) محاسبه می‌شود، و E از (۳۹.۷.۴) به‌دست می‌آید. جملهٔ c_0 قبل از استفاده از الگوریتم چبوال باید نصف شود.

۲. $x_i := \cos(i\pi/(n+1))$ را تولید کنید.

$$f_i = f(x_i) \qquad i := 0, 1, \ldots, n+1$$

۳. برای $j = 0, 1, \ldots, n+1$ تا مرحلهٔ ۸ عمل را ادامه بدهید

۴. $\text{sum} := [f_0 + (-1)^j f_{n+1}]/2$

۵. برای $i = 1, \ldots, n$ تا مرحلهٔ ۶ عمل را انجام دهید

۶. $\text{sum} := \text{sum} + f_i \cos(ij\pi/(n+1))$

۷. حلقه را در i ختم کنید

۸. $c_j \doteq 2\text{sum}/(n+1)$

۹. حلقه را در j ختم کنید.

۱۰. $E := c_{n+1}/2$ و خارج شوید.

کسینوسهای موجود در مرحلهٔ ۶ را می‌توان با استفاده از فرمول‌های جمع مثلثاتی برای توابع سینوس و کسینوس، به‌نحو کاراتری محاسبه کرد، ولی ما آنها را به علت سادگی آموزشی مطلب، انتخاب کرده‌ایم، زیرا زمان اجرای رایانه‌یی برای الگوریتم ما خیلی کم است. به همین دلیل، ما تکنیک FFTی بخش ۸.۳ را به‌کار نبرده‌ایم.

بحث در آثار خواندنی

نظریهٔ تقریب از جنبهٔ کلاسیک یک زمینهٔ مهم ریاضیات است، و نیز ابزاری با اهمیت فزاینده در مطالعهٔ مسائل جدیدِ بسیار متنوع در ریاضیات کاربردی، مثلاً، در فیزیک‌ریاضی و ریاضیات ترکیبیاتی است. تنوع مسائل و روشهای نظریهٔ تقریب را در کتابهای، اچیزر[1] (۱۹۵۶)، اسکی[2] (۱۹۷۵b)، دیویس (۱۹۶۳)، لورنتس (۱۹۶۶)، میناردوس (۱۹۶۷)، پاوئل (۱۹۸۱)، رایسRice (۱۹۶۴)، (۱۹۶۸) می‌توان دید. کارکلاسیک در زمینهٔ چندجمله‌ییهای متعامد در سگو (۱۹۶۷)، و یک بازنگری ازکارهای تازه‌تر در اسکی (۱۹۷۵a) داده شده است. برای چندجمله‌ییهای چبیشف وکاربردهای زیاد آنها در همهٔ ریاضیات کاربردی و آنالیز عددی، فاکس و پارکر (۱۹۶۸)، ریولین (۱۹۷۴) را ببینید. در این کتاب به موضوع مربوط به سریهای فوریه و تقریب با چندجمله‌ییهای مثلثاتی فقط اشاره شده، ولی این مبحث در عدّهٔ زیادی ازکاربردها از اهمیّت کلیدی برخوردار است. یک مرجع کلاسیک زیگموند (۱۹۵۹) است. زمینه‌های زیادی در نظریهٔ تقریب وجود دارند که ما حتی تعریف هم نکرده‌ایم. برای یک بازنگری عالی در این زمینه گاوچی[3] (۱۹۷۵) راکه شامل فهرست مراجع خوبی است ببینید. یک زمینهٔ اصلی که در این کتاب حذف شده است، تقریب به وسیلهٔ توابع گویاست. برای زمینهٔ کلی، میناردوس (۱۹۶۷، فصل نهم) و رایس (۱۹۶۸، فصل نهم) را ببینید. تعمیم چندجمله‌ییهای تیلر به توابع گویا را تقریب پاده[4] می‌نامند؛ برای مقدمات بیکر (۱۹۷۵) و برزینسکی[5] (۱۹۸۰) را ببینید. در زمینهٔ مربوط به بسطهای وابستهٔ بسط کسرهای مسلسل توابع وال (۱۹۴۸) را ببینید. بسیاری از توابع که از نظر کاربردی جالب هستند نمونه‌هایی هستند که توابع خاص در فیزیک‌ریاضی نامیده می‌شوند. این نمونه‌ها شامل توابع متعالی اساسی (سینوس، لگاریتم، نمایی، ریشهٔ دوم) و علاوه بر آن، چندجمله‌ییهای متعامد، توابع بسل، تابع گاما و تابع ابرهندسی هستند. نوشته‌های مفصّلی دربارهٔ توابع خاص وجود دارند و تقریبهای ویژه‌ای برای اغلب آنها ساخته شده است. مهمترین مراجع برای توابع خاص عبارت‌اند از: آبراموویتس و استگون (۱۹۶۴)، که یک کتاب راهنماست که با حمایت ادارهٔ استانداردهای ملی آمریکا تهیه‌شده و اردلی[6] و همکاران (۱۹۵۳) و یک مجموعهٔ سه‌جلدی که غالباً به آن با عنوان «پروژهٔ بیتمن[7]» اشاره می‌شود. برای یک نگرش جامع و بازنگری روشهای تقریب توابع خاص گاوچی (۱۹۷۵) را ببینید. چکیدهٔ جامعی از قضایای نظری برای توابع خاص و روشهایی برای محاسبهٔ عددی آنها در لوک[8] (۱۹۶۹)، (۱۹۷۵)، (۱۹۷۷) آمده است. برای نمونه‌های نسبتاً متداولتر روندهای مطالعهٔ توابع خاص گزارشهای سمپوزیوم اسکی (۱۹۷۵b) را ببینید.

---

1. Achieser 2. Askey 3. Gautschi, W 4. Padé
5. Beezinski 6. Erdélyi, S. W 7. Batemenn 8. Luke

از زمان استفادهٔ وسیع از رایانه در دههٔ ۱۹۵۰، نیاز به تقریبهای با کیفیت بالای چندجمله‌یی یا تابع گویا برای توابع ریاضی اساسی و سایر توابع خاص به وجود آمده است. همان‌گونه که قبلاً اشاره شد تقریب این توابع نیاز به شناخت ویژگیهای آنها دارد. ولی در عین حال اطلاع دقیق از حساب رایانه‌های رقمی که در فصل اول بیان گردید لازم است. یک مطالعهٔ کلّی از روشهای عددی برای تولید تقریبهای چندجمله‌یی در فریزر (۱۹۶۵) داده شده است، که سازمان‌دهی این فصل را تحت تأثیر قرار داده است. برای یک بحث کامل از تقریب توابع مقدماتی همراه با الگوریتم‌های مشروح به کودی و ویت (۱۹۸۰) مراجعه کنید؛ بحثی از طرح برنامه‌ریزی مربوطه درکودی (۱۹۸۴) آمده است. برای بحث مشابهی از تقریبها، ولی بحثی که شامل بعضی از توابع خاص معمولی نیز باشد، هارت (۱۹۶۸) را ببینید. برای یک مجموعهٔ وسیع از تقریبهای توابع خاص لوک (۱۹۷۵) و (۱۹۷۷) را ببینید. برای توابع کلّی، یک برنامهٔ موفق که برای تولید تقریبهای مینیماکس بسیار مورد استفاده قرارگرفته است، درکودی و همکاران (۱۹۶۸) داده شده است. برنامه‌های کلی برای محاسبهٔ تقریبهای مینیماکس درکتابخانه‌های IMSL و NAG موجودند.

مراجع

Abramowitz, M., and I. Stegun (eds.) (1964). *Handbook of Mathematical Functions*. National Bureau of Standards, U.S. Government Printing Office, Washington, D.C. (It is now published by Dover, New York.)

Achieser, N. (1956). *Theory of Approximation* (transl. C. Hyman). Ungar, New York.

Askey, R. (1975a). *Orthogonal Polynomials and Special Functions*. Society for Industrial and Applied Mathematics, Philadelphia.

Askey, R. (ed.) (1975b). *Theory and Application of Special Functions*. Academic Press, New York.

Baker, G., Jr. (1975). *Essentials of Padé Approximants*. Academic Press, New York.

Brezinski, C. (1980). *Padé-Type Approximation and General Orthogonal Polynomials*. Birkhäuser, Basel.

Cody, W. (1984). FUNPACK–A package of special function routines. In *Sources and Development of Mathematical Software*, W. Cowell (ed.), pp. 49–67. Prentice-Hall, Englewood Cliffs, N.J.

Cody, W. and W. Waite (1980). *Software Manual for the Elementary Functions*. Prentice-Hall, Englewood Cliffs, N.J.

Cody, W., W. Fraser, and J. Hart (1968). Rational Chebyshev approximation using linear equations. *Numer. Math.*, **12**, 242–251.

Davis, P. (1963). *Interpolation and Approximation*. Ginn (Blaisdell), Boston.

Erdélyi, A., W. Magnus, F. Oberhettinger, and F. Tricomi (1953). *Higher Transcendental Functions*, Vols. I, II, and III. McGraw-Hill, New York.

Fox, L., and I. Parker (1968). *Chebyshev Polynomials in Numerical Analysis*. Oxford Univ. Press, Oxford, England.

Fraser, W. (1965). A survey of methods of computing minimax and near-minimax polynomial approximations for functions of a single independent variable. *J. ACM*, **12**, 295–314.

Gautschi, W. (1975). Computational methods in special functions—A survey. In R. Askey (ed.), *Theory and Application of Special Functions*, pp. 1–98. Academic Press, New York.

Hart, J., E. Cheney, C. Lawson, H. Maehly, C. Mesztenyi, J. Rice, H. Thacher, and C. Witzgall (1968). *Computer Approximations*. Wiley, New York. (Reprinted in 1978, with corrections, by Krieger, Huntington, N.Y.)

Isaacson, E., and H. Keller (1966). *Analysis of Numerical Methods*. Wiley, New York.

Lorentz, G. (1966). *Approximation of Functions*. Holt, Rinehart & Winston, New York.

Luke, Y. (1969). *The Special Functions and Their Applications*, Vols. I and II. Academic Press, New York.

Luke, Y. (1975). *Mathematical Functions and Their Approximations*. Academic Press, New York.

Luke, Y. (1977). *Algorithms for the Computation of Mathematical Functions*. Academic Press, New York.

Meinardus, G. (1967). *Approximation of Functions: Theory and Numerical Methods* (transl. L. Schumaker). Springer-Verlag, New York.

Powell, M. (1981). *Approximation Theory and Methods*. Cambridge Univ. Press, Cambridge, England.

Rice, J. (1964). *The Approximation of Functions: Linear Theory*. Addison-Wesley, Reading, Mass.

Rice, J. (1968). *The Approximation of Functions: Advanced Topics*. Addison-Wesley, Reading, Mass.

Rivlin, T. (1974). *The Chebyshev Polynomials*. Wiley, New York.

Shampine, L. (1970). Efficiency of a procedure for near-minimax approximation. *J. ACM*, **17**, 655–660.

Szego, G. (1967). *Orthogonal Polynomials*, 3rd ed. Amer. Math. Soc., Providence, R.I.

Wall, H. (1948). *Analytic Theory of Continued Fractions*. Van Nostrand, New York.

Zygmund, A. (1959). *Trigonometric Series*, Vols. I and II. Cambridge Univ. Press, Cambridge, England.

مسائل

۱. برای اینکه نشان دهید چندجمله‌ییهای $p_n(x)$ برنشتاین در قضیهٔ ۱.۴، تقریبهای ضعیفی هستند، تقریب چندجمله‌یی درجهٔ چهارم $p_n(x)$ را برای $f(x) = \sin(\pi x)$، $\circ \leq x \leq ۱$، حساب کنید. آن را با تقریب چندجمله‌یی درجهٔ چهارم تیلر که حول $x = ۱/۲$ بسط داده شده مقایسه کنید.

۲. گیریم $S = \sum_{۱}^{\infty}(-۱)^j a_j$ یک سری همگرا باشد و فرض کنید که همهٔ a_jها ناکوچکتر از صفر باشند و

$$a_۱ \geq a_۲ \geq \ldots \geq a_n \geq \ldots$$

ثابت کنید که

$$| S - \sum_{۱}^{n}(-۱)^j a_j | \leq a_{n+۱}$$

۳. با استفاده از مسألهٔ ۲، همگرایی سریهای زیر را بررسی کنید. کران خطا را وقتی سریها پس از n جمله بریده شوند به‌دست آورید و توجه کنید که به x بستگی دارند. مقدار n را به گونه‌ای بیابید که خطا از $۱۰^{-۵}$ کمتر باشد. این مسأله روش معمول دیگری را برای کراندارکردن خطا در سری تیلر نشان می‌دهد،

(الف) $J_\circ(x) = \sum_{j=\circ}^{\infty} \frac{(-۱)^j(\frac{۱}{۴}x^۲)^j}{(j!)^۲}$

(ب) $\sum_{j=\circ}^{\infty} \frac{(-۱)^j x^{۲j}}{j^۲}$

۴. نمودار خطاهای تقریبهای $p_n(x)$ سری تیلر را برای $f(x) = \sin[(\pi/۲)x]$ در $-۱ \leq x \leq ۱$، به ازای $n = ۱, ۳, ۵$ رسم کنید. به رفتار خطا، هم در نزدیکی مرکز و هم در نزدیکی نقاط انتهایی توجه نمایید.

۵. گیریم $f(x)$ بر بازهٔ $[-\alpha, \alpha]$ به ازای مقداری از $\alpha > \circ$ سه بار پیوسته مشتقپذیر باشد، و تقریب آن را توسط تابع گویای

$$R(x) = \frac{a + bx}{۱ + cx}$$

در نظر می‌گیریم. برای تعمیم مفهوم سری تیلر، ثابتهای a و b و c را به گونه‌ای انتخاب کنید که

$$R^{(j)}(\circ) = f^{(j)}(\circ) \qquad j = \circ, ۱, ۲$$

آیا همیشه ممکن است چنین تابع تقریب $R(x)$ی پیدا کرد؟ تابع $R(x)$ یک مثال از تقریب پاده است. بیکر (۱۹۸۰) و برزینسکی (۱۹۸۰) را ببینید.

۶. مسألهٔ ۵ را برای حالت $f(x) = e^x$ بهکار برید و $R(x)$ را پیدا کنید. خطای آن را بر $[-۱, ۱]$ تحلیل و آن را با خطای چندجملهیی درجهٔ دوم تیلر مقایسه کنید.

۷. بهوسیلهٔ اتحادهای گوناگون، اغلب میتوان یکبازهای راکه لازم است تابعی در آن تقریب شود به یک بازهٔ کوچکتر بدل کرد. نشان دهید که چگونه میتوان هریک از توابع زیر را از $-\infty < x < \infty$ به بازهٔ دادهشده تقلیل داد. معمولاً چند عمل اضافی، ولی ساده، لازم است.

(الف) $e^x \quad ۰ \le x \le ۱$

(ب) $\cos(x) \quad ۰ \le x \le \pi/۴$

(ج) $\tan^{-۱}(x) \quad ۰ \le x \le ۱$

(د) $\ln(x) \quad ۱ \le x \le ۲$. $\ln(x)$ در $۰ < x < \infty$ را تقلیل دهید.

۸. (الف) گیریم $f(x)$ بر $[a, b]$ پیوسته مشتقپذیر باشد. فرض کنید $p(x)$ یک چندجملهیی باشد که برای آن $\| f' - p \|_\infty \le \varepsilon$، و تعریف میکنیم

$$q(x) = f(a) + \int_a^x p(t)dt \qquad a \le x \le b$$

نشان دهید که $q(x)$ یک چندجملهیی است و در رابطهٔ زیر صدق میکند

$$\| f - q \|_\infty \le \varepsilon(b - a)$$

(ب) قسمت (الف) را به حالتی تعمیم دهید که $f(x)$، N بار پیوسته مشتقپذیر بر $[a, b]$ باشد، $N \ge ۲$ و $p(x)$ یک چندجملهیی باشد که در رابطهٔ زیر صدق کند

$$\| f^{(N)} - p \|_\infty \le \varepsilon$$

فرمولی برای یک چندجملهیی $q(x)$ پیدا کنید که $f(x)$ را تقریب بزند و شامل تنها یک انتگرال باشد.

(ج) فرض کنید که $f(x)$ بینهایت بار مشتقپذیر بر $[a, b]$ باشد، یعنی، $f^{(j)}(x)$ در $[a, b]$ موجود و برای همهٔ مقادیر $j \ge ۰$ پیوسته باشد. [این فرض ایجاب نمیکند که $f(x)$ دارای بسط تیلر همگرا بر $[a, b]$ باشد.] ثابت کنید یک دنباله از چندجملهییهای $\{p_n(x) \mid n \ge ۱\}$ موجود است که برای آنها بهازای جمیع مقادیر $j \ge ۰$،

$$\lim_{n\to\infty} \| f^{(j)} - p_n^{(j)} \|_\infty = ۰$$

راهنمایی: قضیهٔ وایرشتراس و قسمت (ب) را بهکار برید.

۹. قضیهٔ زیر را ثابت کنید: گیریم $f \in C^۲[a,b]$ با $f''(x) > ۰$ در $a \le x \le b$. اگر $q_۱^*(x) = a_۰ + a_۱x$ تقریب خطی مینیماکس برای $f(x)$ بر $[a,b]$ باشد، آنگاه

$$a_۱ = \frac{f(b)-f(a)}{b-a} \qquad a_۰ = \frac{f(a)+f(c)}{۲} - \left(\frac{a+c}{۲}\right)\left[\frac{f(b)-f(a)}{b-a}\right]$$

که c جواب یکتای

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

است. ρ چیست؟

۱۰. (الف) چندجمله‌ییهای خطی تیلر را برای $f(x) = \ln(x)$ در $۲ \le x \le ۱$، در حول $x_۰ = \frac{۳}{۲}$ پیدا کنید. نمودار خطا را رسم کنید.

(ب) تقریب خطی مینیماکس برای $f(x) = \ln(x)$ را در $[۱, ۲]$ پیدا کنید. نمودار خطا را رسم و آن را با تقریب تیلر مقایسه کنید.

۱۱. (الف) نشان دهید که تقریب خطی مینیماکس $\sqrt{۱+x^۲}$ در $[۰, ۱]$ چنین است

$$q_۱^*(x) = ۰٫۹۵۵ + ۰٫۴۱۴x$$

(ب) با استفاده از (الف)، تقریب

$$\sqrt{y^۲+z^۲} \doteq ۰٫۹۵۵z + ۰٫۴۱۴y \qquad ۰ \le y \le z$$

را پیدا و خطا را تعیین کنید.

۱۲. تقریب کمترین مربعات خطی برای $f(x) = \ln(x)$ را در $[۱, ۲]$ به‌دست آورید. خطا را با نتایج مسألهٔ ۱۰ مقایسه کنید.

۱۳. مقداری از α را پیدا کنید که رابطهٔ زیر را مینیمم کند

$$\int_۰^۱ | e^x - \alpha | \, dx$$

مقدار مینیمم چیست؟ این یک مثال ساده باز راه دیگری برای اندازه‌گیری خطا در یک تقریب و بهترین تقریب آن است.

۱۴. مسائل مینیمم‌سازی زیر را حل و معلوم کنید که آیا مقداری یکتا برای α که این مینیمم را به‌دست دهد وجود دارد؟ در هر یک از حالات، α می‌تواند در حوزهٔ همهٔ اعداد حقیقی تغییر کند.

تابع $f(x) = x$ را با چندجمله‌ییهایی به شکل $\alpha x^۲$ تقریب می‌زنیم.

(الف) $\underset{\alpha}{\text{Min}} \int_{-1}^{1}[x-\alpha x^2]^2 dx$

(ب) $\underset{\alpha}{\text{Min}} \int_{-1}^{1} \mid x-\alpha x^2 \mid dx$

(ج) $\underset{\alpha}{\text{Min}} \underset{-1\le x\le 1}{\text{Max}} \mid x-\alpha x^2 \mid dx$

۱۵. با استفاده از (۱۰.۴.۴) نشان دهید که $\{p_n(x)\}$ یک خانوادهٔ متعامد است و $\| p_n \|_2 = \sqrt{2/(2n+1)}$، $n \ge 0$.

۱۶. تحقیق کنید که تابعهای

$$\varphi_n(x) = \frac{(-1)^n}{n!} e^x \frac{d^n}{dx^n}(x^n e^{-x})$$

به‌ازای $n \ge 0$ در بازهٔ $[0, \infty)$ نسبت به تابع وزن $w(x) = e^{-x}$ متعامدند. (توجه کنید که $\int_0^\infty e^{-x} x^m dx = m!$، برای $m = 0, 1, 2, \ldots$).

۱۷. (الف) ماکسیمم و مینیمم نسبی $T_n(x)$ را در $[-1, 1]$ پیدا کنید و (۲۱.۷.۴) را به‌دست آورید.

(ب) صفرهای $T_n(x)$ را پیدا کنید و (۱۴.۷.۴) را به‌دست آورید.

۱۸. فرمول‌هایی را که برای b_n و c_n در رابطهٔ بازگشتی سه‌گانهٔ (۲۱.۴.۴) از قضیّهٔ ۵.۴ داده شده‌اند پیدا کنید.

۱۹. روند قضیهٔ ۲.۴ گرام ـ اشمیت را برای اجتناب از مرحلهٔ نرمال‌سازی $\varphi_n = \psi_n / \| \psi_n \|_2$ تغییر دهید:

$$\psi_n(x) = x^n + b_{n,n-1}\psi_{n-1}(x) + \cdots + b_{n,0}\psi_0(x)$$

و ضرایب $b_{n,j}$، $0 \le j \le n-1$ را بیابید.

۲۰. با استفاده از مسألهٔ ۱۹، ψ_0، ψ_1، ψ_2 را برای توابع وزن $w(x)$ در بازه‌های $[a, b]$ که در زیر تعیین شده‌اند، پیدا کنید.

(الف) $w(x) = \ln(x)$، $0 \le x \le 1$

راهنمایی: $\int_0^1 x^n \ln(x) dx = [-1/(n+1)^2]$، $n \ge 0$.

(ب) $w(x) = x$، $0 \le x \le 1$.

(ج) $w(x) = \sqrt{1-x^2}$، $-1 \le x \le 1$.

۲۱. گیریم $\{\varphi_n(x) \mid n \ge 1\}$ها در (a, b) با تابع وزن $w(x) \ge 0$ متعامد باشند. صفرهای $\varphi_n(x)$ را با

$$a < z_{n,n} < z_{n-1,n} < \cdots < z_{1,n} < b$$

نشان دهید. ثابت کنید که صفرهای $\varphi_n(x)$ صفرهای $\varphi_{n+1}(x)$ را از هم جدا می‌کنند، یعنی

$$a < z_{n+1,n+1} < z_{n,n} < z_{n,n+1} < \cdots < z_{2,n+1} < z_{1,n} < z_{n,n+1} < b$$

راهنمایی: از استقرا بر درجهٔ n استفاده کنید. بنویسید $\varphi_n(x) = A_n x^n + \ldots$، $A_n > 0$ و از رابطهٔ بازگشتی سه‌گانهٔ (۲۱.۴.۴) برای محاسبهٔ چندجمله‌ییها در صفرهای $\varphi_n(x)$ استفاده کنید. مشاهده می‌کنید که علامت برای $\varphi_{n+1}(x)$ و $\varphi_{n-1}(x)$ تغییر می‌کند.

۲۲. اتحاد (۲۷.۴.۴) کریستوفل ـ دار بو را برای حالت $x = y$ بسط دهید و فرمول زیر را پیدا کنید.

$$\sum_{k=0}^{n} \frac{[\varphi_k(x)]^2}{\gamma_k}$$

راهنمایی: حد (۲۷.۴.۴) را وقتی $y \to x$ درنظر بگیرید.

۲۳. گیریم $f(x) = \cos^{-1}(x)$ برای $-1 \le x \le 1$ (شاخهٔ اصلی $0 \le f \le \pi$) باشد. چندجمله‌یی از درجهٔ دو زیر را پیدا کنید

$$p(x) = a_0 + a_1 x + a_2 x^2$$

که عبارت زیر را مینیمم سازد.

$$\int_{-1}^{1} \frac{[f(x) - p(x)]^2}{\sqrt{1 - x^2}} dx$$

۲۴. $S_n(x)$ را با $S_n(x) = \frac{1}{n+1} T'_{n+1}(x)$، که در آن $T_{n+1}(x)$ چندجمله‌یی چبیشف از درجهٔ $n+1$ است تعریف کنید. چندجمله‌یی $S_n(x)$ را **چندجمله‌یی نوع دوم چبیشف** گویند.

(الف) نشان دهید که $\{S_n(x) \mid n \ge 0\}$ یک خانوادهٔ متعامد در $[-1, 1]$ نسبت به تابع وزن $w(x) = \sqrt{1 - x^2}$ است.

(ب) نشان دهید که خانوادهٔ $\{S_n(x)\}$ مانند $T_n(x)$ در رابطهٔ بازگشتی سه‌گانهٔ (۱۳.۴.۴) صدق می‌کند.

(ج) برای $f \in C[-1, 1]$ مفروض، مسألهٔ زیر را حل کنید

$$\text{Min} \int_{-1}^{1} \sqrt{1 - x^2}\ [f(x) - p_n(x)]^2 dx$$

که در آن $p_n(x)$ می‌تواند هر چندجمله‌یی از درجهٔ نابزرگتر از n باشد.

۲۵. تعداد عملیات محاسباتی الگوریتم چبوال بخش ۵.۴ را پیدا کنید. تعداد جمعها، تعداد ضربها را به‌دست آورید. سپس با الگوریتم ضرب معمولی تودرتو مقایسه کنید.

۲۶. نشان دهید که چارچوب بخشهای ۴.۴ و ۵.۴ برای چندجمله‌ییهای مثلثاتی از درجهٔ نابزرگتر از n نیز به‌کار می‌رود. نشان دهید که خانوادهٔ $\{1, \sin(x), \cos(x), \ldots, \sin(nx), \cos(nx)\}$ در $[0, 2\pi]$ متعامدند. تقریب کمترین مربعات برای $f(x)$ را در این بازه با استفاده از این گونه چندجمله‌ییها به‌دست آورید. [اگر $n \to \infty$، سری معروف فوریه به‌دست می‌آید (زیگموند (۱۹۵۹)) را ببینید].

۲۷. گیریم $f(x)$ تابعی پیوسته و زوج (فرد) بر $[-a, a]$ باشد. نشان دهید که تقریب مینیماکس $q_n^*(x)$ برای $f(x)$ یک تابع زوج (فرد) بر $[-a, a]$ خواهد بود خواه n زوج باشد خواه فرد.

راهنمایی: قضیهٔ ۱۰.۴ و یکتایی در قضیه را به‌کار برید.

۲۸. با استفاده از (۱۰.۶.۴) کران $\rho_n(f)$ برای توابع $f(x)$ در زیر را در بازه‌های داده‌شده پیدا کنید $n = 1, 2, \ldots, 10$.

(الف) $\sin(x)$ $\quad 0 \le x \le \pi/2$

(ب) $\ln(x)$ $\quad 1 \le x \le e$

(ج) $\tan^{-1}(x)$ $\quad 0 \le x \le \pi/4$

(د) e^x $\quad 0 \le x \le 1$

۲۹. برای بسط چبیشف (۲.۷.۴)، نشان دهید که اگر $f(x)$ در $[-1, 1]$ زوج (فرد) باشد، آنگاه $c_j = 0$ اگر j فرد (زوج) باشد.

۳۰. برای تابع $f(x) = \sin[(\pi/2)x]$، $-1 \le x \le 1$، تقریبات کمترین مربعات درجهٔ ۳ی لژاندر و چبیشف را برای $f(x)$ پیدا کنید. خطای هریک از تقریبات را معلوم و نمودار آنها را رسم کنید. قضیهٔ ۹.۴ را برای کراندارکردن خطای مینیماکس $\rho_3(f)$ به‌کار برید.

راهنمایی: انتگرالگیری عددی را برای ساختن ضرایب کمترین مربعات به‌کار برید. به توضیحاتی که به دنبال (۱۳.۷.۴) برای تقریب کمترین مربعات چبیشف آمده است توجه کنید.

۳۱. تقریب نزدیکِ مینیماکس درونیاب $I_n(x)$ را برای توابع اساسی ریاضی f که در زیر داده شده‌اند در بازه‌های داده‌شده به‌ازای $n = 1, 2, \ldots, 8$ به‌دست آورید. با استفاده از روندهای استاندهٔ رایانهٔ خود، خطا را محاسبه کنید. نمودار خطا را رسم کنید، و با استفاده از قضیهٔ ۹.۴، کرانهای بالا و پایین $\rho_n(f)$ را به‌دست آورید.

(الف) e^x $\quad 0 \le x \le 1$

(ب) $\sin(x)$ $\quad ۰ \le x \le \pi/۲$

(ج) $\tan^{-۱}(x)$ $\quad ۰ \le x \le ۱$

(د) $\ln(x)$ $\quad ۱ \le x \le ۲$

۳۲. مسألهٔ ۳۱ را برای تقریب نزدیکِ مینیماکس $F_n(x)$ تکرار کنید.

۳۳. مسألهٔ ۳۱ و ۳۲ را برای

$$f(x) = \frac{۱}{x}\int_{۰}^{x} \frac{\sin(t)}{t}dt \qquad ۰ \le x \le \frac{\pi}{۲}$$

به‌کار برید.

راهنمایی: یک تقریب درجهٔ بالای تیلر برای محاسبهٔ $f(x)$ پیدا کنید، سپس تبدیل (۱۲.۵.۴) را به‌کار برید تا یک مسألهٔ تقریب در $[-۱, ۱]$ به‌دست آید.

۳۴. برای $f(x) = e^x$ در $[-۱, ۱]$، ساختن $I_n(x)$ را در نظر بگیرید. رابطهٔ زیر را برای خطای $I_n(x)$ با مقادیر مناسب α_n و β_n به‌دست آورید.

$$\alpha_n \mid T_{n+۱}(x) \mid \le \mid f(x) - I_n(x) \mid \le \beta_n \mid T_{n+۱}(x) \mid \qquad -۱ \le x \le ۱$$

کرانهای ناصفر بالا و پایین $\rho_n(f)$ را پیدا کنید.

۳۵. (الف) تابع $\sin(x)$ در $x = ۰$ صفر می‌شود. برای آنکه یک تقریب بهتری برای آن از لحاظ خطای نسبی پیدا کنید، تابع $f(x) = \sin(x)/x$ را در نظر بگیرید. تقریب نزدیک مینیماکس $I_n(x)$ را روی $۰ \le x \le \pi/۲$ برای $n = ۱, ۲, \ldots, ۷$ محاسبه و سپس $\sin(x) - xI_n(x)$ را با نتایج مسألهٔ ۳۱ (ب) مقایسه کنید.

(ب) قسمت (الف) را برای تابع $f(x) = \tan^{-۱}(x)$، $۰ \le x \le ۱$، تکرار کنید.

۳۶. گیریم $f(x) = a_n x^n + \cdots + a_۱ x + a_۰$ و $a_n \neq ۰$. تقریب مینیماکس $f(x)$ در $[-۱, ۱]$ را با یک چندجمله‌یی از درجهٔ نابیشتر از $n-۱$ پیدا کرده و همچنین $\rho_{n-۱}(f)$ را به‌دست آورید.

۳۷. گیریم $\alpha = \text{Min}\left[\underset{|x|\le ۱}{\text{Max}} \mid x^۶ - x^۲ - p_۵(x) \mid\right]$، که مینیمم روی تمام چندجمله‌ییهای از درجهٔ نابیشتر از ۵ گرفته شده است.

(الف) α را پیدا کنید.

(ب) چندجمله‌یی $p_۵(x)$ را که به‌ازای آن مینیمم حاصل می‌شود به‌دست آورید.

۳۸. قضیهٔ (۱۰.۶.۴) را ثابت کنید. **راهنمایی:** تقریب نزدیکِ مینیماکس $I_n(x)$ را در نظر بگیرید.

۳۹. **(الف)** برای $f(x) = e^x$ در $[-۱, ۱]$ چندجمله‌یی $p_۴(x)$ از درجهٔ چهار تیلر را که حول $x = ۰$ بسط داده شده است پیدا کنید.

(ب) با استفاده از مسألهٔ ۳۶، چندجمله‌یی مینیماکس $m_{۴,۳}(x)$ از درجهٔ ۳ را که $p_۴(x)$ را تقریب می‌زند پیدا کنید. نمودار خطای $e^x - m_{۴,۳}(x)$ را رسم و آن را با خطای $e^x - p_۳(x)$ که در شکل ۱.۴ نشان داده شده مقایسه کنید. فرایند تبدیل چندجمله‌یی تیلر به یک چندجمله‌یی تیلر از درجهٔ پایینتر بدین طریق را، **صرفه‌جویی** یا **تلسکوپی‌کردن** گویند. این فرایند معمولاً چندین‌بار پی‌درپی به‌کار می‌رود تا یک چندجمله‌یی درجهٔ بالای تیلر را به یک چندجمله‌یی تقریب از درجهٔ به مراتب پایینتر تبدیل کند.

۴۰. با استفاده از برنامهٔ استانده در مرکز رایانه‌یی خود برای محاسبهٔ تقریبهای مینیماکس، تقریب مینیماکس $q_n^*(x)$ را برای توابع داده‌شدهٔ زیر در بازه‌های داده‌شده به‌دست آورید. این عمل را برای $n = ۱, ۲, ۳, \ldots, ۸$ انجام دهید. نتایج را با نتایج مسألهٔ ۳۱ مقایسه کنید.

(الف) e^x $\quad ۰ \le x \le ۱$

(ب) $\sin(x)$ $\quad ۰ \le x \le \pi/۲$

(ج) $\tan^{-۱}(x)$ $\quad ۰ \le x \le ۱$

(د) $\ln(x)$ $\quad ۱ \le x \le ۲$

۴۱. تقریبهای مینیماکس $q_n^*(x)$، $n = ۱, ۳, ۵, ۷, ۹$، را برای

$$f(x) = \frac{۱}{x}\int_۰^x e^{-t^۲}dt \qquad -۱ \le x \le ۱$$

به‌دست آورید.

راهنمایی: ابتدا یک تقریب تیلر با دقت زیاد تهیّه کنید و سپس آن را با برنامهٔ مسألهٔ ۴۰ به‌کار برید.

۴۲. مسألهٔ ۴۱ را برای

$$f(x) = \frac{۱}{x}\int_۰^x \frac{\sin(t)}{t}dt \qquad |x| \le \pi$$

تکرار کنید.

۵

انتگرالگیری عددی

در این فصل روشهای عددی محاسبهٔ انتگرالهای معین را پیدا و تحلیل می‌کنیم. این انتگرالها معمولاً به شکل زیرند

$$I(f) = \int_a^b f(x)dx \qquad (۱.۰.۵)$$

که در آنها بازهٔ $[a, b]$ متناهی است. چنین انتگرالهایی را اغلب نمی‌توان به‌طور صریح محاسبه نمود، و در بسیاری از آنها به جای محاسبهٔ دقیق و استفاده از تابع اولیهٔ پیچیدهٔ $f(x)$، سریعتر آن است که از راه انتگرالگیری عددی محاسبه شوند. به تقریب $I(f)$ معمولاً **انتگرالگیری عددی** یا **تربیع** گویند.

روشهای عددی بسیاری برای محاسبهٔ (۱.۰.۵) وجود دارند، ولی بیشتر آنها را می‌توان به گونه‌ای ساخت که در چارچوب سادهٔ زیر بگنجند. برای انتگرالده $f(x)$، یک خانوادهٔ تقریب‌زن $\{f_n(x), n \geq ۱\}$ پیدا و تعریف می‌کنیم

$$I_n(f) = \int_a^b f_n(x)dx \doteq I(f_n) \qquad (۲.۰.۵)$$

معمولاً می‌خواهیم تقریب $f_n(x)$ در رابطهٔ زیر صدق کند

$$\| f - f_n \|_\infty \to ۰ \quad \text{هرگاه} \quad n \to \infty \qquad (۳.۰.۵)$$

و شکل هر $f_n(x)$ باید به گونه‌ای انتخاب شود که $I_n(f)$ به‌سادگی محاسبه شود. برای خطا داریم

$$E_n(f) = I(f) - I_n(f) = \int_a^b [f(x) - f_n(x)]dx$$

$$| E_n(f) | \leq \int_a^b | f(x) - f_n(x) | dx \leq (b-a) \| f - f_n \|_\infty \qquad (۴.۰.۵)$$

بیشتر روشهای انتگرالگیری عددی را می‌توان در این چارچوب نگریست اگرچه بعضی از آنها از دیدگاههای دیگر بهتر مطالعه شده‌اند. یک دسته از روشهایی که در این چارچوب نمی‌گنجند آنهایی هستند که بر پایهٔ برونیابی با استفاده از براوردهای مجانبی خطا، انجام می‌شوند. این روشها در بخش ۴.۵ مطالعه خواهند شد.

بیشتر انتگرالهای عددی $I_n(f)$ در محاسبه دارای شکل زیرند

$$I_n(f) = \sum_{j=1}^{n} w_{j,n} f(x_{j,n}) \qquad n \geq 1 \qquad (۵.۰.۵)$$

ضرایب $w_{j,n}$، **وزنهای انتگرالگیری** یا **وزنهای تربیع** خوانده می‌شوند؛ و نقاط $x_{j,n}$، را **گره‌های** انتگرالگیری می‌نامند که معمولاً در $[a, b]$ انتخاب می‌شوند. اندیس وابستگی وزنها وگره‌ها به n معمولاً با نوشتن به‌صورت w_j و x_j، حذف می‌شود، زیرا به‌طور ضمنی وابستگی به n فهمیده می‌شود. روشهای استانده یا دارای وزنها وگره‌هایی هستند که فرمولهای ساده‌ای دارند، یا در غیر اینصورت در جداولی داده شده‌اند که در دسترس هستند. بنابراین نیازی به ساختن صریح توابع $f_n(x)$ در (۲.۰.۵) نخواهد بود، اگرچه به‌یادداشتن نقش آنها در تعریف $I_n(f)$ ممکن است مفید باشد.

مثال زیر یک مثال ساده از (۲.۰.۵) ـ (۴.۰.۵) است، ولی به شکل (۵.۰.۵) نیست.

مثال مطلوب است محاسبهٔ

$$I = \int_0^1 \frac{e^x - 1}{x} dx \qquad (۶.۰.۵)$$

این انتگرالده یک نقطهٔ تکین برداشتنی در مبدأ دارد. برای تعریف $f_n(x)$ سری تیلر را برای e^x به‌کار می‌بریم (۴.۱.۱)[فصل ۱ را ببینید]. سپس تعریف می‌کنیم

$$I_n = \int_0^1 \sum_{j=1}^{n} \frac{x^{j-1}}{j!} dx$$

$$= \sum_{j=1}^{n} \frac{1}{(j!)(j)} \qquad (۷.۰.۵)$$

برای خطا در I_n، فرمول (۴.۱.۱) تیلر را به‌کار می‌بریم تا به ازای مقداری چون $0 \leq \xi_x \leq x$ به‌دست آوریم

$$f(x) - f_n(x) = \frac{x^n}{(n+1)!} e^{\xi_x}$$

در این صورت

$$I - I_n = \int_0^1 \frac{x^n}{(n+1)!} e^{\xi_x} dx$$

$$\frac{1}{(n+1)!(n+1)} \leq I - I_n \leq \frac{e}{(n+1)!(n+1)} \qquad (۸.۰.۵)$$

دنبالهٔ (۷.۰.۵) سریعاً همگراست و (۸.۰.۵) به ما اجازه‌می‌دهد که خطا را با دقت محاسبه کنیم. مثلاً، برای $n = 6$

$$I_6 = ۱٫۳۱۷۸۷۰۳۷$$

و از (۸.۰.۵)

$$۲٫۸۳ \times ۱۰^{-5} \leq I - I_6 \leq ۷٫۷۰ \times ۱۰^{-5}$$

خطای واقعی برابر است با $۳٫۱۸ \times ۱۰^{-5}$.

برای انتگرالهایی که انتگرالده در آنها یک نوع رفتار بدی داشته باشد، مثلاً، در نقطه‌ای بینهایت باشد، اغلب انتگرالده را به شکل زیر در نظر می‌گیریم

$$I(f) = \int_a^b w(x) f(x) dx \qquad (۹.۰.۵)$$

فرض شده است که رفتار بد در $w(x)$ است، که **تابع وزن** خوانده می‌شود و تابع $f(x)$ خوش‌رفتار فرض شده است. برای مثال، محاسبهٔ

$$\int_0^1 (\ln x) f(x) dx$$

را برای توابع پیوستهٔ دلخواه $f(x)$ در نظر می‌گیریم. چارچوب (۲.۰.۵) ـ (۴.۰.۵) به‌سادگی برای پرداختن به (۹.۰.۵) تعمیم داده می‌شود. روشهایی برای این انتگرالها در بخش ۳.۵ و ۶.۵ در نظر گرفته شده‌اند.

بیشتر فرمولهای انتگرالگیری عددی بر پایهٔ تعریف $f_n(x)$ در (۲.۰.۵) بااستفاده از درونیابی چند جمله‌یی یا درونیابی چندجمله‌یی تکه‌یی، قرار دارند. فرمول‌هایی راکه در آنها از چنین درونیابی با نقاط

گرهی متساوی‌الفاصله، استفاده می‌شود، در بخشهای ۱.۵ و ۲.۵ به‌دست آورده و مورد بحث قرار داده‌ایم. فرمولهای انتگرالگیری گاوسی که از بعضی نظرها بهینه هستند و همگرایی خیلی سریع دارند، در بخش ۳.۵ داده شده‌اند. این فرمولها بر اساس تعریف $f_n(x)$ با استفاده از چند جمله‌یی درونیاب در نقاط گرهی که با دقت خاصی انتخاب شده‌اند و لازم نیست متساوی‌الفاصله باشند، استخراج شده‌اند.

فرمولهای مجانبی خطا برای روشهای بخشهای ۱.۵ و ۲.۵ در بخش ۴.۵ داده شده و مورد بحث قرارگرفته‌اند و بعضی فرمولهای جدید، بر پایهٔ برونیابی با این فرمول‌های خطا، به‌دست آمده‌اند. روشهایی برای کنترل خطا به طور خودکار، با حفظ کارایی، در بخش ۵.۵ داده شده‌اند. در بخش ۶.۵ یک بررسی کلی از روشهایی برای انتگرالهایی که تکین یا به تعبیری بد- رفتار هستند انجام شده است و در بخش ۷.۵ کار دشوار مشتقگیری عددی مورد بحث قرار می‌گیرد.

۱.۵ قاعدهٔ ذوزنقه‌یی و قاعدهٔ سیمپسون

مطالعهٔ انتگرالگیری عددی را با ارائهٔ دو روش عددی کاملاً معروف برای محاسبهٔ

$$I(f) = \int_a^b f(x)dx \qquad (۱.۱.۵)$$

آغاز می‌کنیم. این روشها را به طور کامل شرح داده و تحلیل می‌نماییم و از آنها به‌عنوان مقدمات مطالب بخشهای بعد استفاده خواهیم کرد. بازهٔ $[a,b]$ در این بخش همیشه‌متناهی است.

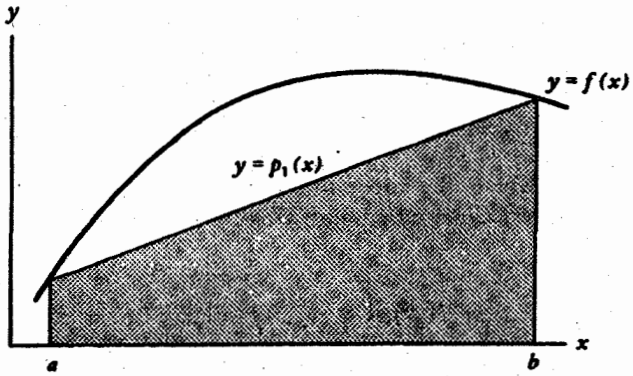
قاعدهٔ ذوزنقه‌یی قاعدهٔ ذوزنقه‌یی ساده بر پایهٔ تقریب‌زدن $f(x)$ با یک خط مستقیم، خط واصل بین دو نقطهٔ $(a,f(a))$ و $(b,f(b))$ قرار دارد. با انتگرالگیری این خط مستقیم به‌جای $f(x)$، تقریب زیر را به دست می‌آوریم

$$I_۱(f) = \left(\frac{b-a}{۲}\right)[f(a)+f(b)] \qquad (۲.۱.۵)$$

این مقدار البته مساحت ذوزنقه‌ای است که در شکل ۱.۵ نمایش داده شده است. برای پیداکردن فرمول خطا، فرمول خطای درونیابی (۱۰.۱.۳) را به‌کار می‌بریم

$$f(x) - \frac{(b-x)f(a)+(x-a)f(b)}{b-a} = (x-a)(x-b)f[a,b,x]$$

همچنین برای کار با خطا در قاعدهٔ ذوزنقه‌یی در این بخش فرض می‌کنیم که $f(x)$ در $[a,b]$



شکل ۱.۵ نمایش قاعدهٔ ذوزنقه‌یی

دوبار پیوسته مشتقپذیر است. در این صورت

$$E_1(f) = \int_a^b f(x)dx - \frac{(b-a)}{2}[f(a)+f(b)]$$
$$= \int_a^b (x-a)(x-b)f[a,b,x]dx \qquad (۳.۱.۵)$$

با استفاده از قضیهٔ مقدار میانگین برای انتگرالها (قضیهٔ ۳.۱ فصل یک) و فرمول (۱۲.۲.۳) داریم

$$E_1(f) = f[a,b,\xi]\int_a^b (x-a)(x-b)dx \qquad a \le \xi \le b$$
$$= \left[\frac{1}{2}f''(\eta)\right]\left[-\frac{1}{6}(b-a)^3\right] \qquad \eta \in [a,b]$$

بنابراین

$$E_1(f) = -\frac{(b-a)^3}{12}f''(\eta) \qquad \eta \in [a,b] \qquad (۴.۱.۵)$$

اگر $b-a$ به اندازهٔ کافی کوچک نباشد، قاعدهٔ ذوزنقه‌یی (۲.۱.۵) خیلی مفید نیست. در چنین انتگرالی، آن را به مجموع انتگرالها روی زیربازه‌های کوچک تبدیل می‌کنیم، و بعد (۲.۱.۵) را برای هر یک از این انتگرالهای کوچکتر به‌کار می‌بریم. گیریم $n \ge 1$ و $h = (b-a)/n$ و $x_j = a + jh$ برای $j = 0, 1, \ldots, n$. در این صورت

$$I(f) = \int_a^b f(x)dx = \sum_{j=1}^{n}\int_{x_{j-1}}^{x_j} f(x)dx$$
$$= \sum_{j=1}^{n}\left\{\left(\frac{h}{2}\right)[f(x_{j-1})+f(x_j)] - \frac{h^3}{12}f''(\eta_j)\right\}$$

با $x_{j-1} \le \eta_j \le x_j$. هیچ لزومی ندارد که زیربازه‌های $[x_{j-1}, x_j]$ همگی طولهای مساوی داشته باشند، ولی معمول است که ابتدا اصول کلی روش را بدین طریق معرفی نمایند. اگرچه در عمل هم این طریق متداول ، ولی مواردی هست که در آن فاصلهٔ متغیر نقاط شبکه بیشتر مطلوب است.

عبارات در مجموع را می‌توان ترکیب نموده **قاعدهٔ ذوزنقه‌یی مرکب** را به‌دست آورد،

$$I_n(f) = h[\frac{1}{2}f_0 + f_1 + f_2 + \cdots + f_{n-1} + \frac{1}{2}f_n] \qquad n \ge 1 \qquad (5.1.5)$$

که در آن $f(x_j) \equiv f_j$. برای خطا در $I_n(f)$،

$$E_n(f) = I(f) - I_n(f) = \sum_{j=1}^{n} -\frac{h^3}{12}f''(\eta_j)$$

$$= -\frac{h^3 n}{12}\left[\frac{1}{n}\sum_{j=1}^{n} f''(\eta_j)\right] \qquad (6.1.5)$$

برای عبارت داخل کروشه

$$\underset{a\le x\le b}{\text{Min}} f''(x) \le M \equiv \frac{1}{n}\sum_{j=1}^{n} f''(\eta_j) \le \underset{a\le x\le b}{\text{Max}} f''(x)$$

چون $f''(x)$ در $a \le x \le b$ پیوسته است، و M بین ماکسیمم و مینیمم $f''(x)$ در $[a,b]$ است، در نقطه‌ای در $[a,b]$ مانند η، $f''(\eta) = M$. پس می‌توانیم بنویسیم

$$E_n(f) = -\frac{(b-a)h^2}{12}f''(\eta) \qquad \eta \in [a,b] \qquad (7.1.5)$$

با استفاده از این تحلیل می‌توان براورد دیگری برای خطا به‌دست آورد. از (۶.۱.۵) داریم

$$\underset{n\to\infty}{\text{Lim}}\frac{E_n(f)}{h^2} = \underset{n\to\infty}{\text{Lim}}\left[-\frac{h}{12}\sum_{j=1}^{n} f''(\eta_j)\right]$$

$$= -\frac{1}{12}\underset{n\to\infty}{\text{Lim}}\sum_{j=1}^{n} f''(\eta_j)h$$

چون $x_{j-1} \le \eta_j \le x_j$، $j = 1, \ldots, n$ آخرین مجموع، یک مجموع ریمان است؛ بنابراین

$$\underset{n\to\infty}{\text{Lim}}\frac{E_n(f)}{h^2} = -\frac{1}{12}\int_a^b f''(x)dx = -\frac{1}{12}[f'(b) - f'(a)] \qquad (8.1.5)$$

$$E_n(f) \doteq -\frac{h^2}{12}[f'(b) - f'(a)] \equiv \tilde{E}_n(f) \qquad (9.1.5)$$

عبارت $\tilde{E}_n(f)$ یک براورد خطای مجانبی برای $E_n(f)$ نامیده می‌شود، و به تعبیر (۸.۱.۵) موجّه است.

تعریف گیریم $E_n(f)$ یک فرمول دقیق خطا و $\tilde{E}_n(f)$ براوردی از آن باشد. گوییم که $\tilde{E}_n(f)$ **برآورد خطای مجانبی** برای $E_n(f)$ است اگر

$$\lim_{n\to\infty}\frac{\tilde{E}_n(f)}{E_n(f)} = ۱ \qquad (۱۰.۱.۵)$$

یا هم‌ارز با آن

$$\lim_{n\to\infty}\frac{E_n(f)-\tilde{E}_n(f)}{E_n(f)} = ۰$$

براورد (۹.۱.۵)، بر اساس (۸.۱.۵)، در این تعریف صدق می‌کند.

قاعدهٔ ذوزنقه‌یی مرکب (۵.۱.۵) را می‌توانستیم ازگذاردن تابع درونیاب خطّی تکّه‌یی $f_n(x)$ به جای $f(x)$ در گره‌های $x_۰$، $x_۱$ ، ... ، x_n نیز به‌دست آوریم. از این پس قاعدهٔ ذوزنقه‌یی مرکب را فقط قاعدهٔ ذوزنقه‌یی می‌نامیم.

مثال قاعدهٔ ذوزنقه‌یی (۹.۱.۵) را برای محاسبهٔ

$$I = \int_۰^\pi e^x\cos(x)dx \qquad (۱۱.۱.۵)$$

به‌کار می‌بریم. مقدار صحیح آن برابر است با $I = -(e^\pi + ۱)/۲ \doteq ۰ - ۱۲٫۰۷۰۳۴۶۳۱۶۴$ مقادیر I_n در جدول ۱.۵ همراه با خطاهای درست E_n و براوردهای مجانبی $\tilde{E}_n$ که از (۹.۱.۵) به‌دست آمده، داده شده است. توجه نمایید که وقتی n دوبرابر (و بنابراین h نصف) می‌شود خطا با مضرب ۴ کاهش می‌یابد. از عامل ضرب $h^۲$ که در (۷.۱.۵) و (۹.۱.۵) وجود دارد این نتیجه قابل پیش‌بینی بود: وقتی h نصف می‌شود $h^۲$ با مضرب ۴ کاهش می‌یابد. این مثال همچنین نشان می‌دهد که قاعدهٔ ذوزنقه‌یی در مقایسه با سایر روشهایی که در این فصل بحث خواهدشد، چندان کارایی ندارد.

با استفاده از براورد خطای $\tilde{E}_n(f)$، می‌توانیم قاعدهٔ انتگرالگیری عددی بهتری را تعریف کنیم:

$$\begin{aligned} CT_n(f) &\equiv I_n(f) + \tilde{E}_n(f) \\ &= h\left[\frac{۱}{۲}f_۰ + f_۱ + \cdots + f_{n-۱} + \frac{۱}{۲}f_n\right] - \frac{h^۲}{۱۲}[f'(b) - f'(a)] \end{aligned} \qquad (۱۲.۱.۵)$$

این را **قاعدهٔ ذوزنقه‌یی تصحیح شده** می‌نامند. دقت $\tilde{E}_n(f)$ موجب می‌شود که $CT_n(f)$ به مراتب دقیقتر از قاعدهٔ ذوزنقه‌یی باشد. شکل دیگری برای به‌دست آوردن (۱۲.۱.۵) در مسالهٔ ۴

جدول ۱.۵ قاعده ذوزنقه‌یی برای محاسبه (۱۱.۱.۵)

n	I_n	E_n	Ratio	$\tilde{E}$
۲	−۱۷٫۳۸۹۲۵۹	۵٫۳۲		۴٫۹۶
۴	−۱۳٫۳۳۶۰۲۳	۱٫۲۷	۴٫۲۰	۱٫۲۴
۸	−۱۲٫۳۸۲۱۶۲	۳٫۱۲E − ۱	۴٫۰۶	۳٫۱۰E − ۱
۱۶	−۱۲٫۱۴۸۰۰۴	۷٫۷۷E − ۲	۴٫۰۲	۷٫۷۶E − ۲
۳۲	−۱۲٫۰۸۹۷۴۲	۱٫۹۴E − ۲	۴٫۰۰	۱٫۹۴E − ۲
۶۴	۱۲٫۰۷۵۱۹۴	۴٫۸۵E − ۳	۴٫۰۰	۴٫۵۸E − ۳
۱۲۸	−۱۲٫۰۷۱۵۵۸	۱٫۲۱E − ۳	۴٫۰۰	۱٫۲۱E − ۳
۲۵۶	−۱۲٫۰۷۰۶۴۹	۳٫۰۳E − ۴	۴٫۰۰	۳٫۰۳E − ۴
۵۱۲	−۱۲٫۰۷۰۴۲۲	۷٫۵۷E − ۵	۴٫۰۰	۷٫۵۷E − ۵

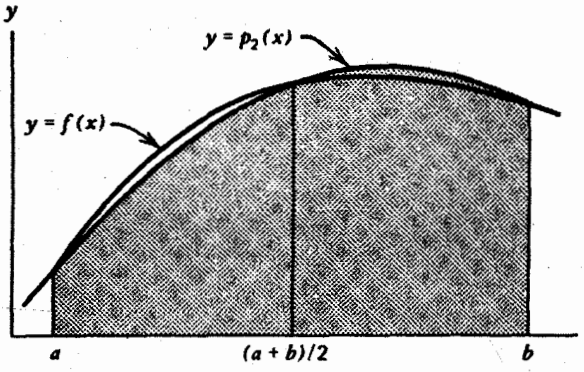
جدول ۲.۵ قاعدهٔ ذوزنقه‌یی اصلاح شده برای (۱۱.۱.۵)

n	$CT_n(f)$	خطا	نسبت	خطای ذوزنقه‌یی
۲	−۱۲٫۴۲۵۵۲۸۳۶۷	۳٫۵۵E − ۱		۵٫۳۲
۴	−۱۲٫۰۹۵۰۹۰۱۰۶	۲٫۴۷E − ۲	۱۴٫۴	۱٫۲۷
۸	−۱۲٫۰۷۱۹۲۹۲۴۵	۱٫۵۸E − ۳	۱۵٫۶	۳٫۱۲E − ۱
۱۶	−۱۲٫۰۷۰۴۴۵۸۰۴	۹٫۹۵E − ۵	۱۵٫۹	۷٫۷۷E − ۲
۳۲	−۱۲٫۰۷۰۳۵۲۵۴۳	۶٫۲۳E − ۶	۱۶٫۰	۱٫۹۴E − ۲
۶۴	−۱۲٫۰۷۰۳۴۶۷۰۶	۳٫۸۹E − ۷	۱۶٫۰	۴٫۸۵E − ۳
۱۲۸	−۱۲٫۰۷۰۳۴۶۳۴۱	۲٫۴۳E − ۸	۱۶٫۰	۱٫۲۱E − ۳

توصیه شده است، که نشان می‌دهد (۱۲.۱.۵) در چارچوب تقریب نظری (۲.۰.۵) ـ (۴.۰.۵) درمی‌آید. مشکل اساسی در استفاده از $CT_n(f)$ نیاز به $f'(a)$ و $f'(b)$ است.

مثال $CT_n(f)$ را برای مثال قبلی (۱۱.۱.۵) به‌کار می‌بریم. نتایج همراه با خطاهای قاعدهٔ ذوزنقه‌یی برای مقایسه در جدول ۲.۵ داده شده‌اند. به تجربه معلوم شده است که خطا در $CT_n(f)$ متناسب با $h^۴$ است، در حالی که در قاعدهٔ ذوزنقه‌یی متناسب با $h^۲$ بود. یک اثبات این امر در مسألهٔ ۴ داده شده است.

قاعده سیمپسون برای بهبود قاعدهٔ سادهٔ ذوزنقه‌یی (۲.۱.۵)، یک چندجمله‌یی درونیاب درجهٔ دوم $p_۲(x)$ را برای تقریب $f(x)$ روی $[a,b]$ به‌کار می‌بریم. گیریم $c = (a+b)/۲$، و تعریف می‌کنیم



شکل ۲.۵ نمایش قاعدهٔ سیمپسون

$$I_۲(f) = \int_a^b \left[\frac{(x-c)(x-b)}{(a-c)(a-b)}f(a) + \frac{(x-a)(x-b)}{(c-a)(c-b)}f(c) + \frac{(x-a)(x-c)}{(b-a)(b-c)}f(b)\right]dx$$

با انتگرالگیری، به دست می‌آوریم

$$I_۲(f) = \frac{h}{۳}\left[f(a) + ۴f\left(\frac{a+b}{۲}\right) + f(b)\right] \qquad h = \frac{b-a}{۲} \qquad (۱۳.۱.۵)$$

این دستور قاعدهٔ سیمپسون خوانده می‌شود. نمایش آن در شکل ۲.۵ داده شده است، که قسمت سایه‌دار مساحت زیر نمودار $y = p_۲(x)$ را نشان می‌دهد.

برای خطا با فرمول درونیابی خطای (۱۱.۲.۳) شروع می‌کنیم تا به دست آوریم

$$E_۲(f) = I(f) - I_۲(f)$$
$$= \int_a^b (x-a)(x-c)(x-b)f[a,b,c,x]dx \qquad (۱۴.۱.۵)$$

ما نمی‌توانیم قضیهٔ مقدار میانگین را به‌کار بریم زیرا چندجمله‌یی انتگرالده در $x = c = (a+b)/۲$ تغییر علامت می‌دهد. برای کار با قاعدهٔ سیمپسون در این فصل، فرض می‌کنیم که $f(x)$ در $[a,b]$ چهاربار پیوسته مشتقپذیر باشد. تعریف می‌کنیم

$$w(x) = \int_a^x (t-a)(t-c)(t-b)dt$$

به آسانی می‌توان نشان داد که برای $a < x < b$

$$w(x) > ۰ \qquad w(a) = w(b) = ۰$$

با انتگرالگیری جزء به جزء داریم:

$$E_۲(f) = \int_a^b w'(x) f[a,b,c,x]dx$$
$$= [w(x)f[a,b,c,x]]_{x=a}^{x=b} - \int_a^b w(x)\frac{d}{dx}f[a,b,c,x]dx$$
$$= -\int_a^b w(x)f[a,b,c,x,x]dx$$

در آخرین تساوی از رابطهٔ (۱۷.۲.۳) استفاده شده است. با استفاده از قضیهٔ مقدار میانگین انتگرالها و (۱۲.۲.۳)،

$$E_۲(f) = -f[a,b,c,\xi,\xi]\int_a^b w(x)dx \qquad a \le \xi \le b$$
$$= -\frac{f^{(۴)}(\eta)}{۲۴}\left[\frac{۴}{۱۵}h^۵\right] \qquad h = \frac{b-a}{۲} \qquad \eta \in [a,b]$$

بنابراین

$$E_۲(f) = -\frac{h^۵}{۹۰}f^{(۴)}(\eta) \qquad \eta \in [a,b] \qquad (۱۵.۱.۵)$$

از این رابطه می‌بینیم که اگر $f(x)$ یک چندجمله‌یی از درجهٔ نابزرگتر از ۳ باشد $E_۲(f) = ۰$، به رغم اینکه درونیابی درجهٔ دو وقتی دقیق است که $f(x)$ یک چندجمله‌یی حداکثر از درجهٔ ۲ باشد. حذف اتفاقی بیشتر خطاها در شکل ۲.۵ نموده شده است. این حذف موجب شده که قاعدهٔ سیمپسون بسیار دقیقتر از قاعدهٔ ذوزنقه‌یی باشد.

بازهم یک قاعدهٔ ترکیبی می‌سازیم. برای n زوج و ناکوچکتر از ۲ تعریف می‌کنیم $h = (b-a)/n$، $x_j = a + jh$ و $j = ۰, ۱, \ldots n$. در این صورت

$$I(f) = \int_a^b f(x)dx = \sum_{j=۰}^{n/۲}\int_{x_{۲j-۲}}^{x_{۲j}} f(x)dx$$
$$= \sum_{j=۱}^{n/۲}\left\{\frac{h}{۳}[f_{۲j-۲} + ۴f_{۲j-۱} + f_{۲j}] - \frac{h^۵}{۹۰}f^{(۴)}(\eta_j)\right\}$$

با $x_{۲j-۲} \le \eta_j \le x_{۲j}$. جملات در مجموع را ساده می‌کنیم، قاعدهٔ ترکیبی سیمپسون به‌دست می‌آید:

$$I_n(f) = \frac{h}{۳}[f_۰ + ۴f_۱ + ۲f_۲ + ۴f_۳ + ۲f_۴$$
$$+ \ldots + ۲f_{n-۲} + ۴f_{n-۱} + f_n] \qquad (۱۶.۱.۵)$$

مانند قبل، این را فقط قاعدهٔ سیمپسون می‌نامیم. شاید بتوان گفت که این قاعدهٔ انتگرالگیری عددی است که بیشتر مورد استفاده قرار می‌گیرد. این روش ساده، به‌کار بردن آن آسان و برای طیف وسیعی از انتگرالها نسبتاً دقیق است.

برای خطا، مانند قاعدهٔ ذوزنقه‌یی،

$$E_n(f) = I(f) - I_n(f) = -\frac{h^5(n/2)}{90} \cdot \frac{2}{n} \sum_{j=1}^{n/2} f^{(4)}(\eta_j)$$

$$E_n(f) = -\frac{h^4(b-a)}{180} f^{(4)}(\eta) \qquad \eta \in [a,b] \qquad (17.1.5)$$

می‌توانیم فرمول خطای مجانبی را نیز به‌دست آوریم

$$E_n(f) = -\frac{h^4}{180}[f^{(3)}(b) - f^{(3)}(a)] \equiv \tilde{E}_n(f) \qquad (18.1.5)$$

اثبات اساساً مانند همان اثباتی است که برای پیداکردن (9.1.5) ارائه کردیم.

مثال قاعدهٔ سیمپسون (16.1.5) را برای محاسبهٔ انتگرال (11.1.5)، یعنی

$$I = \int_0^{\pi} e^x \cos(x) dx$$

را که قبلاً به‌عنوان یک مثال برای قاعدهٔ ذوزنقه‌یی حساب کرده بودیم، به‌کار می‌بریم. نتایج عددی در جدول ۳.۵ داده شده‌اند. باز هم نرخ کاهش خطا، نتایج داده شده به وسیلهٔ (17.1.5) و (18.1.5) را تایید می‌کند. در مقایسه با نتایج پیشین جدول ۱.۵ برای قاعدهٔ ذوزنقه‌یی، روشن

جدول ۳.۵ قاعده سیمپسون برای محاسبهٔ (11.1.5)

n	I_n	E_n	Ratio	$\tilde{E}_n$
۲	−۱۱٫۵۹۲۸۳۹۵۵۳۴	−۴٫۷۸E − ۱		−۱٫۶۳
۴	−۱۱٫۹۸۴۹۴۴۰۱۹۸	−۸٫۵۴E − ۲	۵٫۵۹	−۱٫۰۲E − ۱
۸	−۱۲٫۰۶۴۲۰۸۹۵۷۲	−۶٫۱۴E − ۳	۱۴٫۹	−۶٫۳۸E − ۳
۱۶	−۱۲٫۰۶۹۹۵۱۳۲۳۳	−۳٫۹۵E − ۴	۱۵٫۵	−۳٫۹۹E − ۴
۳۲	−۱۲٫۰۷۰۳۲۱۴۵۶۱	−۲٫۴۹E − ۵	۱۵٫۹	−۲٫۴۹E − ۵
۶۴	−۱۲٫۰۷۰۳۴۴۷۵۹۹	−۱٫۵۶E − ۶	۱۶٫۰	−۱٫۵۶E − ۶
۱۲۸	−۱۲٫۰۷۰۳۴۶۲۱۹۱	−۹٫۷۳E − ۸	۱۶٫۰	−۹٫۷۳E − ۸

است که قاعدهٔ سیمپسون بهتر است. در مقایسه با جدول ۲.۵، قاعدهٔ سیمپسون کمی بدتر است، ولی سرعت همگرایی یکی است. قاعدهٔ سیمپسون این مزیت را دارد که نیاز به مقادیر مشتق ندارد.

فرمولهای خطای هسته‌یی پئانو روش دیگری برای به‌دست آوردن فرمولهای خطا وجود دارد که از محاسبهٔ مشتق در یک نقطهٔ نامعلوم η، نتیجه نمی‌شود. ابتدا قاعدهٔ ذوزنقه‌یی را در نظر می‌گیریم. فرض کنید که $f' \in C[a,b]$ و $f''(x)$ در $[a,b]$ انتگرال‌پذیر باشند، در این صورت با استفاده از قضیهٔ تیلر [قضیهٔ ۴.۱ در فصل ۱]،

$$f(x) = p_1(x) + R_2(x) \qquad p_1(x) = f(a) + (x-a)f'(a)$$
$$R_2(x) = \int_a^x (x-t)f''(t)dt$$

با توجه به (۳.۱.۵)، به ازای هر دو تابع $F, G \in C[a,b]$ داریم

$$E_1(F+G) = E_1(F) + E_1(G) \qquad (۱۹.۱.۵)$$

بنابراین

$$E_1(f) = E_1(p_1) + E_1(R_2) = E_1(R_2)$$

زیرا با توجه به (۴.۱.۵) داریم $E_1(p_1) = 0$. اگر به جای $E_1(f)$ مقدار آن را بگذاریم خواهیم داشت

$$E_1(R_2) = \int_a^b R_2(x)dx - (\frac{b-a}{2})[R_2(a) + R_2(b)]$$
$$= \int_a^b \int_a^x (x-t)f''(t)dt - (\frac{b-a}{2}) \int_a^b (b-t)f''(t)dt$$

به‌طور کلی برای هر تابع انتگرال‌پذیر $G(x,t)$،

$$\int_a^b \int_a^x G(x,t)dt\,dx = \int_a^b \int_t^b G(x,t)dx\,dt \qquad (۲۰.۱.۵)$$

بنابراین

$$E_1(R_2) = \int_a^b f''(t) \int_t^b (x-t)dxdt - \left(\frac{b-a}{2}\right) \int_a^b (b-t)f''(t)dt$$

اگر انتگرالها را ترکیب و نتایج را ساده کنیم، داریم

$$E_1(f) = \frac{1}{2} \int_a^b f''(t)(t-a)(t-b)dt \qquad (۲۱.۱.۵)$$

برای قاعدهٔ ذوزنقه‌یی مرکب (۵.۱.۵)

$$E_n(f) = \int_a^b K(t) f''(t) dt \qquad (۲۲.۱.۵)$$

$$K(t) = \frac{۱}{۲}(t - t_{j-۱})(t - t_j) \qquad t_{j-۱} \le t \le t_j \qquad j = ۱, ۲, \ldots, n \qquad (۲۳.۱.۵)$$

فرمولهای (۲۱.۱.۵) و (۲۲.۱.۵) را عبارت خطای هسته‌یی پئانو و $K(t)$ را هستهٔ پئانو می‌نامند. برای ارائهٔ کلّیتر موضوع، دیویس (۱۹۶۳، فصل ۳) را ببینید.

به‌عنوان یک مثال ساده از کاربرد آن، کرانهای (۲۲.۱.۵) را پیدا کرده به‌دست می‌آوریم

$$| E_n(f) | \le \| K \|_\infty \int_a^b | f''(t) | dt = \frac{h^۲}{۸} \int_a^b | f''(t) | dt \qquad (۲۴.۱.۵)$$

اگر $f''(t)$ خیلی قلّه‌یی باشد، رابطهٔ اخیر ممکن است کرانی بهتر از (۷.۱.۵) برای خطا باشد، زیرا در (۷.۱.۵) معمولاً به‌جای $| f''(\eta) |$ باید $\| f'' \|_\infty$ را بگذاریم.

برای قاعدهٔ سیمپسون، قضیهٔ تیلر را به‌کار برده می‌نویسیم

$$f(x) = p_۳(x) + R_۴(x)$$

$$R_۴(x) = \frac{۱}{۶} \int_a^x (x - t)^۳ f^{(۴)}(t) dt$$

مانند قبل

$$E_۲(f) = E_۲(p_۳) + E_۲(R_۴) = E_۲(R_۴)$$

سپس $E_۲(R_۴)$ را با جایگذاری مستقیم و ساده‌کردن به‌دست می‌آوریم:

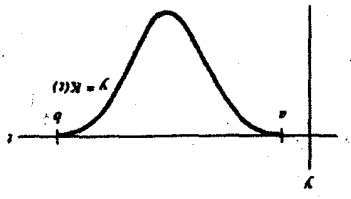
$$E_۲(f) = \int_a^b R_۴(x) dx - \frac{h}{۳}\left[R_۴(a) + ۴R_۴\left(\frac{a+b}{۲}\right) + R_۴(b)\right]$$

از اینجا نتیجه می‌شود

$$E_۲(f) = \int_a^b K(t) f^{(۴)}(t) dt \qquad (۲۵.۱.۵)$$

$$K(t) = \begin{cases} \frac{۱}{۷۲}(t-a)^۳(۳t - a - ۲b) & a \le t \le \frac{a+b}{۲} \\ \frac{۱}{۷۲}(b-t)^۳(b + ۲a - ۳t) & \frac{a+b}{۲} \le t \le b \end{cases} \qquad (۲۶.۱.۵)$$

یک نمودار $K(t)$ در شکل ۳.۵ داده شده است. با محاسبهٔ مستقیم داریم:



شکل ۳.۵ هستهٔ پئانو برای قاعدهٔ سیمپسون

جدول ۴.۵ انتگرالگیری ذوزنقه‌یی و سیمپسونی: مثال (۱)

	قاعدهٔ ذوزنقه‌یی		قاعدهٔ سیمپسون	
n	خطا	نسبت	خطا	نسبت
۲	$-۷٫۱۹۷-۲$		$-۳٫۳۷۰-۳$	
۴	$-۱٫۸۱۷-۲$	۳٫۹۶	$-۲٫۳۱۵-۴$	۱۴٫۶
۸	$-۴٫۵۵۳-۳$	۳٫۹۹	$-۱٫۵۴۳-۵$	۱۵٫۰
۱۶	$-۱٫۱۳۹-۳$	۴٫۰۰	$-۱٫۰۰۸-۶$	۱۵٫۳
۳۲	$-۲٫۸۴۸-۴$	۴٫۰۰	$-۶٫۴۸۹-۸$	۱۵٫۵
۶۴	$-۷٫۱۲۱-۵$	۴٫۰۰	$-۴٫۱۴۱-۹$	۱۵٫۷
۱۲۸	$-۱٫۷۸۰-۵$	۴٫۰۰	$-۲٫۶۲۶-۱۰$	۱۵٫۸

$$\| K \|_\infty = \frac{h^۴}{۷۲} \qquad \int_a^b K(t)dt = -\frac{h^۵}{۹۰} \qquad h = \frac{b-a}{۲}$$

مانند قاعدهٔ ذوزنقه‌یی مرکب، این نتایج به‌سادگی برای قاعدهٔ مرکب سیمپسون تعمیم داده می‌شوند. مثالهای زیر را برای توضیح کاملتر رفتار قاعده‌های سیمپسون و ذوزنقه‌یی آورده‌ایم.

مثال ۱.

$$f(x) = x^۳\sqrt{x}, [a,b] = [۰,۱], I = \frac{۲}{۹}$$

جدول ۴.۵ خطا را برای مقادیر صعودی n به‌دست می‌دهد. مشتق $f^{(۴)}(x)$ در $x = ۰$ تکین است، بنابراین فرمول (۱۷.۱.۵) را نمی‌توان در این حالت به‌کار برد. به‌عنوان راه دیگر تعمیم (۲۵.۱.۵) را برای قاعدهٔ سیمپسون به‌کار می‌بریم، تا به‌دست آوریم

$$| E_n(f) | \leq \| K \|_\infty \int_a^b | f^{(۴)}(t) | dt = \left(\frac{h^۴}{۷۲}\right)\left(\frac{۱۰۵}{۸}\right) = \frac{۳۵}{۱۹۲}h^۴$$

بنابراین وقتی که h نصف می‌شود (یعنی n دوبرابر می‌شود)، خطا باید ۱۶ بار کوچک شود. این روش نیز یک کران تا حدّی واقع بینانه برای خطا به دست می‌دهد. توجه می‌کنید که چگونه

مقادیر تجربی در ستونهای نسبت جدول ۴.۵ و مقادیر پیش‌بینی شدهٔ نظری ۴ و ۱۶ به ترتیب، تطابق نزدیک دارند.

.۲

$$f(x) = \frac{۱}{۱ + (x - \pi)^۲} \qquad [a, b] = [۰, ۵] \qquad I \doteq ۲٫۳۳۹۷۶۶۲۸۳۶۷$$

بر طبق نظریه، مشتقپذیری نامتناهی $f(x)$ ایجاب می‌کند که نسبتها در قاعده‌های ذوزنقه‌یی و سیمپسون به ترتیب ۰٫۴ و ۰٫۱۶ باشد. ولی برای چند مقدار اولیهٔ I_n ، همانگونه که جدول ۵.۵ نشان می‌دهد، لازم نیست که چنین باشد. انتگرالده نسبتاً قله‌یی است، بویژه مشتقات مراتب بالای آن، و این امر بر سرعت همگرایی اثر می‌گذارد.

.۳

$$f(x) = \sqrt{x} \qquad [a, b] = [۰, ۱] \qquad I = \frac{۲}{۳}$$

چون مقدار $f'(x)$ در $x = ۰$ بینهایت می‌شود، هیچ یک از نتایج نظری را که قبلاً داده شده در این حالت نمی‌توان بکاربرد. نتایج عددی در جدول ۶.۵ داده شده‌اند. توجه می‌کنید که هنوز هم خطا یک رفتار منظّمی دارد. در واقع، خطاهای دو روش، وقتی n دوبرابر می‌شود به یک نسبت تنزّل می‌کنند. این نسبتِ $۲٫۸۳ \doteq ۲^{۱٫۵}$ در بخش ۴.۵، فرمول (۲۴.۴.۵)، توضیح داده شده است.

.۴

$$f(x) = e^{\cos(x)} \qquad [a, b] = [۰, ۲\pi] \qquad I \doteq ۷٫۹۵۴۹۲۶۵۲۱۰۱۲۸۴$$

جدول ۵.۵ انتگرالگیری ذوزنقه‌یی و سیمپسون: مثال ۲

	قاعدهٔ ذوزنقه‌یی		قاعدهٔ سیمپسون	
n	خطا	نسبت	خطا	نسبت
۲	۱٫۷۳۱ − ۱		−۲٫۸۵۳ − ۱	
۴	۷٫۱۱۰ − ۲	۲٫۴۳	۳٫۷۰۹ − ۲	−۷٫۶۹
۸	۷٫۴۹۶ − ۳	۹٫۴۸	−۱٫۳۷۱ − ۲	−۲٫۷۱
۱۶	۱٫۹۵۳ − ۳	۳٫۸۴	۱٫۰۵۹ − ۴	−۱۳۰
۳۲	۴٫۸۹۲ − ۴	۳٫۹۹	۱٫۰۸۰ − ۶	۹٫۸۱
۶۴	۱٫۲۲۳ − ۴	۴٫۰۰	۶٫۷۴۳ − ۸	۱۶٫۰
۱۲۸	۳٫۰۵۹ − ۵	۴٫۰۰	۴٫۲۱۷ − ۹	۱۶٫۰

جدول ۶.۵ انتگرالگیری ذوزنقه‌یی و سیمپسون: مثال (۳)

	قاعدهٔ ذوزنقه		قاعدهٔ سیمپسون	
n	خطا	نسبت	خطا	نسبت
۲	۶٫۳۱۱ − ۲		۲٫۸۶۰ − ۲	
۴	۲٫۳۳۸ − ۲	۲٫۷۰	۱٫۰۱۲ − ۲	۲٫۸۲
۸	۸٫۵۳۶ − ۳	۲٫۷۴	۳٫۵۸۷ − ۳	۲٫۸۳
۱۶	۳٫۰۸۵ − ۳	۲٫۷۷	۱٫۲۶۸ − ۳	۲٫۸۳
۳۲	۱٫۱۰۸ − ۳	۲٫۷۸	۴٫۴۸۵ − ۴	۲٫۸۳
۶۴	۳٫۹۵۹ − ۴	۲٫۸۰	۱٫۵۸۶ − ۴	۲٫۸۳
۱۲۸	۱٫۴۱۰ − ۴	۲٫۸۱	۵٫۶۰۶ − ۵	۲٫۸۳

جدول ۷.۵ انتگرالگیری ذوزنقه‌یی و سیمپسون،مثال (۴)

	قاعدهٔ ذوزنقه‌یی		قاعدهٔ سیمپسون	
n	خطا	نسبت	خطا	نسبت
۲	−۱٫۷۴		۷٫۲۱E − ۱	
۴	−۳٫۴۴E − ۲	۵٫۰۶E + ۱	۵٫۳۴E − ۱	۱٫۳۵
۸	−۱٫۲۵E − ۶	۲٫۷۵E + ۴	۱٫۱۵E − ۲	۴٫۶۴E + ۱
۱۶	< ۱٫۰۰E − ۱۴	> ۱٫۲۵E + ۸	۴٫۱۷E − ۷	۲٫۷۶E + ۴
۳۲			< ۱٫۰۰E − ۱۴	> ۴٫۱۷E + ۷

نتایج در جدول ۷.۵ نشان داده شده و بسیار خوب‌اند. هر دو روش همگرای خیلی سریع هستند، بارجحان قاعدهٔ ذوزنقه‌یی بر قاعدهٔ سیمپسون. این مثال همگرایی عالی قاعدهٔ ذوزنقه‌یی را برای انتگرالده‌های دوره‌یی نشان می‌دهد، این موضوع در بخش ۴.۵ تحلیل شده است. نشانه‌ای از این رفتار را می‌توان در عبارات خطای مجانبی (۹.۱.۵) و (۱۸.۱.۵) ملاحظه کرد، زیرا هر دو براورد $f(x)$ در چنین حالتی برابر صفرند.

۲.۵ فرمولهای انتگرالگیری نیوتن ـ کوتس

قاعدهٔ ذوزنقه‌یی سادهٔ (۲.۱.۵) و قاعدهٔ سیمپسون (۱۳.۱.۵) دو مثال اول فرمول‌انتگرالگیری نیوتن‌ـکوتس[1] هستند. برای $n \geq 1$، گیریم $h = (b-a)/n$، $x_j = a + jh$، $j = 0, 1, \ldots, n$.

1. Cotes

برای تعریف $I_n(f)$، $p_n(x)$، چندجمله‌یی درونیاب $f(x)$ در نقاط $x_n, \ldots, x_1, x_0$، را به‌جای $f(x)$ می‌گذاریم:

$$I(f) = \int_a^b f(x)dx \doteq I_n(f) = \int_a^b p_n(x)dx \qquad (۱.۲.۵)$$

با استفاده از فرمول (۶.۱.۳) لاگرانژ برای $p_n(x)$،

$$I_n(f) = \int_a^b \sum_{j=0}^{n} l_{j,n}(x) f(x_j)dx = \sum_{j=0}^{n} w_{j,n} f(x_j) \qquad (۲.۲.۵)$$

با

$$w_{j,n} = \int_a^b l_{j,n}(x)dx \qquad j = 0, 1, \ldots, n \qquad (۳.۲.۵)$$

معمولاً زیرنمایهٔ n را حذف کرده فقط w_j می‌نویسیم. ماتاکنون حالتهای $n = 1$ و $n = 2$ را حساب کرده‌ایم. برای نشان دادن محاسبهٔ وزنها، حالت w_0 را برای $n = 3$ مطرح می‌کنیم.

$$w_0 = \int_a^b l_0(x)dx = \int_{x_0}^{x_3} \frac{(x - x_1)(x - x_2)(x - x_3)dx}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)}$$

یک تبدیل متغیر محاسبات را آسان می‌کند. گیریم $x = x_0 + \mu h$، $0 \le \mu \le 3$. در این صورت

$$\begin{aligned} w_0 &= -\frac{1}{6h^3} \int_{x_0}^{x_3} (x - x_1)(x - x_2)(x - x_3)dx \\ &= -\frac{1}{6h^3} \int_0^3 (\mu - 1)h.(\mu - 2)h.(\mu - 3)h.hd\mu \\ &= -\frac{h}{6} \int_0^3 (\mu - 1)(\mu - 2)(\mu - 3)d\mu \\ w_0 &= \frac{3h}{8} \end{aligned}$$

فرمول کامل برای $n = 3$ عبارت است از

$$I_3(f) = \frac{3h}{8}[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)] \qquad (۴.۲.۵)$$

که قاعدهٔ سه هشتم نامیده می‌شود.

برای خطا، قضیه زیر را ارائه می‌دهیم.

قضیهٔ ۱.۵ (الف) برای n زوج، فرض می‌کنیم $f(x)$ در $[a,b]$، $n+2$ بار پیوسته مشتقپذیر باشد. در این صورت

$$I(f)-I_n(f)=C_nh^{n+3}f^{(n+2)}(\eta)\qquad \eta\in[a,b] \tag{۵.۲.۵}$$

با

$$C_n=\frac{1}{(n+2)!}\int_0^n \mu^2(\mu-1)\dots(\mu-n)d\mu \tag{۶.۲.۵}$$

(ب) برای n فرد، فرض می‌کنیم $f(x)$ در $[a,b]$، $n+1$ بار پیوسته ـ مشتقپذیر باشد، در این صورت

$$I(f)-I_n(f)=C_nh^{n+2}f^{(n+1)}(\eta)\qquad \eta\in[a,b] \tag{۷.۲.۵}$$

که در آن

$$C_n=\frac{1}{(n+1)!}\int_0^n \mu(\mu-1)\dots(\mu-n)d\mu$$

برهان برهان را برای حالت (الف) که مهمترین حالت است به اختصار شرح می‌دهیم. برای اثبات کامل هر دو حالت (آیزکسون و کلر (۱۹۶۶، صص ۳۰۸ ـ ۳۱۴) را ببینید. از (۱۱.۲.۳)،

$$\begin{aligned}E_n(f)&=I(f)-I_n(f)\\&=\int_a^b(x-x_0)(x-x_1)\dots(x-x_n)f[x_0,x_1,\dots,x_n,x]dx\end{aligned}$$

تعریف می‌کنیم

$$w(x)=\int_a^x(t-x_0)\dots(t-x_n)dt$$

پس

$$w(a)=w(b)=0\qquad w(x)>0\qquad \text{برای } a<x<b$$

اثبات $w(x)>0$ را می‌توان در آیزکسون و کلر (۱۹۶۶، ص ۳۰۹) دید. به سادگی می‌توان نشان داد که $w(b)=0$، زیرا انتگرالده $(t-x_0)\dots(t-x_n)$ نسبت به گره میانی $x_{n/2}=(a+b)/2$، تابع فرد است.

با استفاده از انتگرالگیری جزء به جزء و (۳-۲-۱۷) داریم،

$$E_n(f) = \int_a^b w'(x) f[x_0, \ldots, x_n, x] dx$$
$$= [w(x) f[x_0, \ldots, x_n, x]]_a^b - \int_a^b w(x) \frac{d}{dx} f[x_0, \ldots, x_n, x] dx$$
$$E_n(f) = -\int_a^b w(x) f[x_0, \ldots, x_n, x, x] dx$$

با استفاده از قضیهٔ مقدار میانگین انتگرال و (۱۲.۲.۳)

$$E_n(f) = -f[x_0, \ldots, x_n, \xi, \xi] \int_a^b w(x) dx$$
$$= -\frac{f^{(n+2)}(\eta)}{(n+2)!} \int_a^b \int_a^x (t - x_0) \ldots (t - x_n) dt dx \qquad (۸.۲.۵)$$

ترتیب انتگرالگیری را عوض کرده و تبدیل متغیر $t = x_0 + \mu h$ ، $0 \leq \mu \leq n$ ، را به‌کار می‌بریم:

$$\int_a^b w(x) dx = \int_a^b \int_t^b (t - x_0) \ldots (t - x_n) dx dt$$
$$= \int_{x_0}^{x_n} (t - x_0) \ldots (t - x_{n-1})(t - x_n)(x_n - t) dt$$
$$= -h^{n+3} \int_0^n \mu(\mu - 1) \ldots (\mu - n + 1)(\mu - n)^2 d\mu$$

تبدیل متغیر $\nu = n - \mu$ را به‌کار می‌بریم تا نتیجهٔ زیر به دست آید

$$\int_a^b w(x) dx = -h^{n+3} \int_0^n (n - \nu) \ldots (1 - \nu) \nu^2 d\nu$$

با استفاده از این واقعیت که n زوج است و از ترکیب نتیجهٔ آخر با (۸.۲.۵) نتایج (۵.۲.۵) و (۶.۲.۵) به‌دست می‌آید. ■

به‌عنوان یک مرجع آسان، متداولترین فرمولهای نیوتن-کوتس در جدول ۸.۵ داده شده‌اند. برای $n = 4$، $I_4(f)$ را اغلب **قاعدهٔ بول** نامند. مانند گذشته، در جدول می‌گیریم $h = (b - a)/n$.

تعریف گویند یک فرمول انتگرالگیری $\tilde{I}(f)$ که $I(f)$ را تقریب می‌زند، دارای درجهٔ دقت m است اگر

۱. برای همهٔ چندجمله‌ییهای $f(x)$ از درجهٔ نابزرگتر از m، $\tilde{I}(f) = I(f)$

۲. برای یک چندجمله‌یی f از درجهٔ $m + 1$ ، $\tilde{I}(f) \neq I(f)$

جدول ۸.۵ متداولترین فرمولهای نیوتن ـ کوتس

$n=1$	$\int_a^b f(x)dx=\frac{h}{2}\left[f(a)+f(b)\right]-\frac{h^3}{12}f''(\varepsilon)$ قاعدهٔ ذوزنقه‌یی
$n=2$	$\int_a^b f(x)dx=\frac{h}{3}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]-\frac{h^5}{90}f^{(4)}(\xi)$ قاعدهٔ سیمپسون
$n=3$	$\int_a^b f(x)dx=\frac{3h}{8}\left[f(a)+3f(a+h)+3f(b-h)+f(b)\right]-\frac{3h^5}{80}f^{(4)}(\xi)$
$n=4$	$\int_a^b f(x)dx=\frac{2h}{45}\left[7f(a)+32f(a+h)+12f\left(\frac{a+b}{2}\right)+32f(b-h)+7f(b)\right]-\frac{8h^7}{945}f^{(6)}(\xi)$

مثال $n=1,3$ در جدول ۸.۵، درجه‌های دقّت نیز به ترتیب برابر است با $m=n=1,3$. ولی برای $n=2,4$، درجه‌های دقّت به ترتیب برابر است با $m=n+1=3,5$. این روابط این نتیجهٔ کلی را نشان می‌دهند که فرمولهای نیوتن ـ کوتس برای اندیس زوج n، در مقایسه با فرمولهای با اندیس فرد n، از درجهٔ دقت بالاتری برخوردارند. [فرمولهای (۵.۲.۵) و (۷.۲.۵) را ملاحظه کنید.]

هر یک از فرمولهای نیوتن ـ کوتس را می‌توان برای ساختن یک قاعدهٔ مرکب به‌کار برد. شاید سودمندترین آنها قاعده‌ای است که بر مبنای قاعدهٔ بول بنا شده است. (مسألهٔ ۷ را ببینید). ما از ذکر جزئیات بیشتر صرف‌نظرمی‌کنیم.

بحث همگرایی سؤال مورد توجّه دیگر این است که وقتی $n\to\infty$، آیا $I_n(f)$ بسمت $I(f)$ همگرا می‌شود یا نه. با توجه به عدم همگرایی چند جمله‌ییهای درونیابی با نقاط متساوی‌الفاصله برای بعضی از انتخابهای $f(x)$ ۱۰.۵.۳ [را ببینید]، بایستی ما انتظار مشکلاتی را داشته باشیم. جدول ۹.۵ نتایج را برای یک مثال معروف نشان می‌دهد،

$$I=\int_{-4}^{4}\frac{dx}{1+x^2}=2.\tan^{-1}(4)\doteq 2{,}6516 \qquad (9.2.5)$$

جدول ۹.۵ مثال (۹.۲.۵) نیوتن ـ کوتس

n	I_n
۲	۵٫۴۹۰۲
۴	۲٫۲۷۷۶
۶	۳٫۳۲۸۸
۸	۱٫۹۴۱۱
۱۰	۳٫۵۹۵۶

این انتگرالهای عددی نیوتن ـ کوتس واگرا هستند؛ و این بیانگر آن است که فرمولهای انتگرالگیری نیوتن‌ـکوتس $I_n(f)$ در (۲.۲.۵) لازم نیست که به $I(f)$ همگرا باشند.

برای فهمیدن دلایل عدم همگرایی فرمولهای نیوتن ـ کوتس برای (۹.۲.۵)، ما ابتدا یک بحث کلی همگرایی روش‌های انتگرالگیری را عرضه می‌کنیم.

تعریف گیریم $\mathscr{F}$ یک خانواده از توابع پیوسته در بازهٔ داده شدهٔ $[a,b]$ باشد. $\mathscr{F}$ را در $[a,b]$ **چگال** گوییم اگر برای هر $f \in C[a,b]$ و هر $\epsilon > 0$ یک تابع f_ϵ در $\mathscr{F}$ وجود داشته باشد که برای آن

$$\underset{a\le x\le b}{\text{Max}} \mid f(x) - f_\epsilon(x) \mid \le \epsilon \qquad (۱۰.۲.۵)$$

مثال ۱. با توجّه به قضیهٔ وایرشتراس [قضیهٔ ۱.۴ را ببینید]، مجموعهٔ همهٔ چند جمله‌ییها در $C[a,b]$ چگال است.

۲. گیریم $n \ge 1$، $h = (b-a)/n$، $x_j = a + jh$، $0 \le j \le n$. گیریم $f(x)$ بر هر زیر بازهٔ $[x_{j-1}, x_j]$ خطی باشد. $\mathscr{F}$ را مجموعهٔ همهٔ این گونه توابع خطی تکه‌یی $f(x)$ برای همهٔ nها تعریف می‌کنیم. اثبات چگال بودن $\mathscr{F}$ در $C[a,b]$ را به مسألهٔ ۱۱ واگذار می‌کنیم.

قضیهٔ ۲.۵ گیریم

$$I_n(f) = \sum_{j=0}^{n} w_{j,n} f(x_{j,n}) \qquad n \ge 1$$

یک دنباله از فرمولهای انتگرالگیری عددی باشد که

$$I(f) = \int_a^b f(x)dx$$

را تقریب می‌زند. گیریم $\mathscr{F}$ یک خانوادهٔ چگال در $C[a,b]$ باشد. در این صورت

$$I_n(f) \to I(f) \qquad \text{برای هر} \quad f \in C[a,b] \qquad (۱۱.۲.۵)$$

اگر و تنها اگر

۱.

$$I_n(f) \to I(f) \qquad \text{برای همهٔ} \quad f \in \mathscr{F} \qquad (۱۲.۲.۵)$$

و

.۲

$$B \equiv \underset{n\geq 1}{\text{Sup}} \sum_{j=0}^{n} \mid w_{j,n} \mid < \infty \qquad (۱۳.۲.۵)$$

برهان (الف) نمایان است که، (۱۲.۲.۵) از (۱۱.۲.۵) نتیجه می‌شود. ولی اثبات نتیجه شدن (۱۳.۲.۵) از (۱۱.۲.۵) بسیار مشکل است. این یک مثال از **اصل کرانداری یکنواخت** است و تقریباً درهمهٔ کتابهای درسی آنالیز تابعی یافت می‌شود؛ برای مثال کرایر (۱۹۸۲، ص ۱۲۱) را ببینید.
(ب) اکنون ثابت می‌کنیم که از (۱۲.۲.۵) و (۱۳.۲.۵)، (۱۱.۲.۵) نتیجه می‌شود. گیریم $f \in C[a,b]$ داده شده باشد و $\epsilon > 0$ دلخواه باشد. با استفاده از این فرض که $\mathscr{F}$ در $C[a,b]$ چگال است، $f_\epsilon \in \mathscr{F}$ را به گونه‌ای می‌گیریم که

$$\underset{a\leq x\leq b}{\text{Max}} \mid f(x) - f_\epsilon(x) \mid \leq \frac{\epsilon}{[2(b-a+B)]} \qquad (۱۴.۲.۵)$$

سپس می‌نویسیم

$$\begin{aligned} I(f) - I_n(f) =& [I(f) - I(f_\epsilon)] + [I(f_\epsilon) - I_n(f_\epsilon)] \\ &+ [I_n(f_\epsilon) - I_n(f)] \end{aligned}$$

با استفاده از (۱۳.۲.۵) و (۱۴.۲.۵) به سادگی می‌توان به‌دست آورد

$$\begin{aligned} |I(f) - I_n(f)| \leq & |I(f) - I(f_\epsilon)| + | I(f_\epsilon) - I_n(f_\epsilon) | \\ &+ | I_n(f_\epsilon) - I_n(f) | \\ \leq & \frac{\epsilon}{2} + | I(f_\epsilon) - I_n(f_\epsilon) | \end{aligned}$$

با استفاده از (۱۲.۲.۵) وقتی $n \to \infty$، $I_n(f_\epsilon) \to I(f_\epsilon)$. بنابراین برای همهٔ مقادیر n به اندازه کافی بزرگ مثلاً $n \geq n_\epsilon$

$$| I(f) - I_n(f) | \leq \epsilon$$

چون ϵ دلخواه بود، این نشان می‌دهد که وقتی $n \to \infty$، $I_n(f) \to I(f)$. ■

چون انتگرالهای عددی $I_n(f)$ نیوتن ـ کوتس برای $f(x) = 1/(1+x^2)$ بر بازهٔ $[-4,4]$ به $I(f)$ همگرا نمی‌شود، باید نتیجه شود که یا شرط (۱۲.۲.۵) یا (۱۳.۲.۵) نقض شده است. اگر $\mathscr{F}$ را مجموعهٔ چندجمله‌ییها انتخاب کنیم آنگاه (۱۲.۲.۵) برقرار است، زیرا برای هر چندجمله‌یی

p از درجهٔ نابزرگتر از n ، $I_n(p) = I(p)$. بنابراین (۱۳.۲.۵) باید نادرست باشد. برای فرمولهای (۲.۲.۵) نیوتن ـ کوتس

$$\operatorname{Sup}_n \sum_{j=0}^{n} \mid w_{j,n} \mid = \infty \qquad (۱۵.۲.۵)$$

چون برای حالت خاص $f(x) \equiv ۱$، $I(f) = I_n(f)$، برای هر n داریم

$$\sum_{j=0}^{n} w_{j,n} = b - a \qquad n \leq ۱ \qquad (۱۶.۲.۵)$$

از ترکیب این نتایج، علامت وزنهای $w_{j,n}$ وقتی n به اندازه کافی بزرگ می‌شود باید تغییر یابد. برای مثال، برای $n = ۸$،

$$\int_{x_0}^{x_۸} f(x)dx \doteq I_۸(f) = \frac{۴h}{۱۴۱۷۵}[۹۸۹(f_0 + f_۸) + ۵۸۸۸(f_۱ + f_۷) - ۹۲۸(f_۲ + f_۶) + ۱۰۴۹۶(f_۳ + f_۵) - ۴۵۴۰ f_۴]$$

چنین فرمولهایی ممکن است موجب خطاهای کاهش ارقام با معنی شوند، اگرچه بعید است تا وقتی n بزرگ نیست این یک مسألهٔ جدی باشد. ولی به علت همین مسأله، افراد معمولاً از به‌کار بردن فرمولهای نیوتن ـ کوتس برای $n \geq ۸$ احتراز می‌کنند، حتی از تشکیل فرمولهای مرکب.

مهمترین مسألهٔ (۲.۲.۵) در روش نیوتن ـ کوتس، آن است که برای انتگرالده‌های کاملاً خوشرفتار مانند (۹.۲.۵)، ممکن است انتگرال همگرا نباشد.

قاعدهٔ میانگاهی فرمول‌های نیوتن ـ کوتس دیگری وجود دارند که در آنها یک یا هر دو حدّ انتگرالگیری از نقاط شبکهٔ درونیابی (و انتگرالگیری) حذف می‌شوند. بهترین آنها، که ساده‌ترین آنها نیز هست، **قاعدهٔ میانگاهی** است. این قاعده بر پایهٔ درونیابی انتگرالده $f(x)$ با مقدار ثابت $f((a+b)/۲)$ قرار دارد؛ و فرمول انتگرالگیریِ حاصل چنین است

$$\int_a^b f(x)dx = (b-a) f\left(\frac{a+b}{۲}\right) + \frac{(b-a)^۳}{۲۴} f''(\eta) \qquad \eta \in [a,b] \quad (۱۷.۲.۵)$$

برای شکل مرکب آن، تعریف می‌کنیم

$$x_j = a + (j - \tfrac{۱}{۲})h, \qquad j = ۱, ۲, \ldots n$$

که میانگاه بازه‌های $[a+(j-1)h, a+jh]$ است. در این صورت

$$\int_a^b f(x)dx = I_n(f) + E_n(f)$$

$$I_n(f) = h[f_1 + f_2 + \ldots + f_n] \qquad (18.2.5)$$

$$E_n(f) = \frac{h^2(b-a)}{24} f''(\eta) \qquad \eta \in [a,b] \qquad (19.2.5)$$

اثبات این نتایج به‌صورت مسألهٔ ۱۰ واگذار شده است.

این فرمولهای انتگرالگیری که در آنها یک یا هر دو حدّ انتگرالگیری حذف شده‌اند، **فرمولهای باز نیوتن ـ کوتس** خوانده می‌شوند، و فرمولهای قبلی **فرمولهای بسته** نامیده می‌شوند. فرمولهای باز از مرتبهٔ بالاتر برای استخراج فرمولهای عددی در حل معادلات دیفرانسیل معمولی به‌کار برده شده‌اند.

۳.۵ انتگرالگیری گاوسی

قاعده‌های ذوزنقه‌یی مرکب و سیمپسون بر پایهٔ استفاده از تقریب چندجمله‌یی از درجهٔ پایین برای انتگرالده $f(x)$ در زیربازه‌هایی با اندازهٔ نزولی، قرار دارند. دراین بخش رده‌ای از روشها را بررسی می‌کنیم که در آنها از تقریبهای چندجمله‌یی $f(x)$ از درجهٔ صعودی استفاده می‌شود. فرمولهای انتگرالگیری به‌دست آمده، در بیشتر حالات بسیار دقیق‌اند، و اگر فردی با محاسبهٔ انتگرالهای زیادی روبه‌رو باشد باید به این فرمولها جداً توجه کند.

برای عمومیت بیشتر، فرمولهای زیر را مورد توجّه قرار می‌دهیم

$$I_n(f) = \sum_{j=1}^{n} w_{j,n} f(x_{j,n}) \doteq \int_a^b w(x)f(x)dx = I(f) \qquad (1.3.5)$$

تابع وزن $w(x)$ بر $[a,b]$ غیرمنفی و انتگرالپذیر فرض می‌شود و همچنین باید در مفروضات (۸.۳.۴) و (۹.۳.۴) از بخش ۳.۴ نیز صدق کند. نقاط گرهی $\{x_{j,n}\}$ و وزنهای $\{w_{j,n}\}$ باید طوری انتخاب شوند که $I_n(f)$، برای چندجمله‌یی $f(x)$ از درجهٔ در حد امکان بالا، برابر $I(f)$ شود. امید می‌رود که این شرایط به فرمول $I_n(f)$ای بینجامد، که برای انتگرالده‌های $f(x)$ که توسط چندجمله‌ییها به خوبی تقریب‌پذیرند، تقریباً دقیق باشد. در بخش ۲.۵ فرمولهای نیوتن ـ کوتس، وقتی n بزرگ می‌شود، درجهٔ دقت افزایش می‌یابد، ولی با وجود این برای بسیاری از انتگرالده‌های خوشرفتار همگرا نیستند. مشکل در فرمولهای نیوتن ـ کوتس این است که نقاط شبکهٔ $\{x_{j,n}\}$

باید متساوی‌الفاصله باشند. باحذف این محدودیت ما قادریم فرمولهای جدید $I_n(f)$ ای به‌دست آوریم که برای هر تابع $f \in C[a,b]$ همگرا باشند.

برای آنکه درک مستقیمی از پیداکردن $I_n(f)$ به دست آوریم، حالت خاص زیر را در نظر می‌گیریم

$$\int_{-1}^{1} f(x)dx \doteq \sum_{j=1}^{n} w_j f(x_j) \qquad (۲.۳.۵)$$

که در آن $w(x) \equiv 1$ و وابستگی صریح $\{w_j\}$ و $\{x_j\}$ به n حذف شده است. وزنهای $\{w_j\}$ و نقاط شبکه‌یی $\{x_j\}$ باید تعیین شوند، که خطای

$$E_n(f) = \int_{-1}^{1} f(x)dx - \sum_{j=1}^{n} w_j f(x_j) \qquad (۳.۳.۵)$$

برای چندجمله‌یی $f(x)$ از درجهٔ هرچه ممکن بالا، برابر صفر شود. برای پیداکردن معادلات گرهها و وزنها، ابتدا توجه می‌کنیم که

$$E_n(a_0 + a_1x + \cdots + a_mx^m) = a_0E_n(1) + a_1E_n(x) + \cdots + a_mE_n(x^m) \qquad (۴.۳.۵)$$

بنابراین $E_n(f)$ برای چندجمله‌یی از درجهٔ نابزرگتر از m برابر صفر است اگر و تنها اگر

$$E_n(x^i) = 0 \qquad i = 0, 1, \ldots, m \qquad (۵.۳.۵)$$

حالت ۱. $n = 1$. چون دو پارامتر w_1 و x_1 وجود دارند، نیاز داریم

$$E_n(1) = 0 \qquad E_n(x) = 0$$

که نتیجه می‌دهد

$$\int_{-1}^{1} 1dx - w_1 = 0 \qquad \int_{-1}^{1} xdx - w_1x_1 = 0$$

از اینجا لازم می‌آید $w_1 = 2$ و $x_1 = 0$. پس فرمول (۲.۳.۵) چنین می‌شود

$$\int_{-1}^{1} f(x)dx \doteq 2f(0)$$

که قاعدهٔ میانگاهی است.

حالت ۲. $n = ۲$. چهار پارامتر موجودند، $w_۱$، $w_۲$، $x_۱$، $x_۲$، و بنابراین چهار شرط برای این پارامترها می‌گذاریم

$$E_n(x^i) = \int_{-۱}^{۱} x^i dx - [w_۱ x_۱^i + w_۲ x_۲^i] = ۰ \qquad i = ۰, ۱, ۲, ۳$$

یا

$$\begin{aligned} w_۱ + w_۲ &= ۲ \\ w_۱ x_۱ + w_۲ x_۲ &= ۰ \\ w_۱ x_۱^۲ + w_۲ x_۲^۲ &= \frac{۲}{۳} \\ w_۱ x_۱^۳ + w_۲ x_۲^۳ &= ۰ \end{aligned}$$

این معادلات به یک فرمول یکتای زیر می‌انجامند

$$\int_{-۱}^{۱} f(x)dx \doteq f\left(-\frac{\sqrt{۳}}{۳}\right) + f\left(\frac{\sqrt{۳}}{۳}\right) \qquad (۶.۳.۵)$$

که درجهٔ دقت آن ۳ است. این را با قاعده سیمپسون (۱۳.۱.۵) مقایسه کنید که برای رسیدن به این دقت از سه نقطهٔ گرهی استفاده می‌کند.

حالت ۳. برای حالت کلی n، $۲n$ پارامتر آزاد $\{x_i\}$ و $\{w_i\}$ موجودند و باید حدس بزنیم که یک فرمول (۲.۳.۵) وجود دارد که با استفاده از n گره درجهٔ دقتی برابر $۲n - ۱$ خواهد داشت. معادلاتی که باید حل شوند عبارت‌اند از

$$E_n(x^i) = ۰ \qquad i = ۰, ۱, \ldots ۲n - ۱$$

یا

$$\sum_{j=۱}^{n} w_j x_j^i = \begin{cases} ۰ & i = ۱, ۳, \ldots, ۲n - ۱ \\ \dfrac{۲}{i+۱} & i = ۰, ۲, \ldots, ۲n - ۲ \end{cases} \qquad (۷.۳.۵)$$

این معادلات، غیرخطی‌اند و حلپذیری آنها به هیچ وجه ساده نیست. به علّت دشواری استفاده از این دستگاه غیرخطی، از روش دیگری برای این نظریه برای (۲.۳.۵) استفاده می‌کنیم، که تا اندازه‌ای پیچ در پیچ است.

گیریم $\{\varphi_n(x) \mid n \geq ۰\}$ نسبت به تابع وزن $w(x) \geq ۰$ بر (a, b) چند جمله‌ییهای متعامد باشند. صفرهای $\varphi_n(x)$ را با

$$a < x_۱ < \cdots < x_n < b \qquad (۸.۳.۵)$$

نشان می‌دهیم. همچنین از نمادگذاری (۴-۴-۱۸)-(۴-۴-۲۰) داریم

$$\varphi_n(x) = A_n x^n + \cdots \qquad a_n = \frac{A_{n+1}}{A_n}$$

$$\gamma_n = \int_a^b w(x)[\varphi_n(x)]^2 dx \qquad (۹.۳.۵)$$

قضیهٔ ۳.۵ برای هر $n \geq 1$، یک فرمول انتگرالگیری عددی یکتای (۱.۳.۵) از درجهٔ دقت $2n-1$ وجود دارد. با فرض اینکه $f(x)$ در بازهٔ $[a,b]$، $2n$ بار پیوسته مشتقپذیر باشد، فرمول $I_n(f)$ و خطای آن با عبارت زیر داده می‌شود

$$\int_a^b w(x)f(x)dx = \sum_{j=1}^{n} w_j f(x_j) + \frac{\gamma_n}{A_n^2(2n)!} f^{(2n)}(\eta) \qquad (۱۰.۳.۵)$$

به‌ازای یک $a < \eta < b$. نقاط گرهی $\{x_j\}$ صفرهای $\varphi_n(x)$ هستند و وزنهای $\{w_j\}$ چنین‌اند؛

$$w_j = \frac{-a_n\gamma_n}{\varphi_n'(x_j)\varphi_{n+1}(x_j)} \qquad j = 1, \ldots n \qquad (۱۱.۳.۵)$$

برهان برهان به سه قسمت تقسیم شده است. ابتدا با استفاده از نقاط گرهی (۸.۳.۵) یک فرمول با درجهٔ دقت $2n-1$ پیدا می‌کنیم. سپس نشان می‌دهیم که این فرمول یکتاست. سرانجام پیدا کردن فرمول خطا و وزنها را مطرح می‌کنیم.

(الف) ساختن فرمول. درونیابی ارمیت به‌عنوان ابزار برای ساختن به‌کار برده شده است (بخش ۶.۳ را برای یادآوری نمادها و نتایج ببینید). برای نقاط گرهی (۸.۳.۵)، چندجمله‌یی درونیاب ارمیت $f(x)$ و $f'(x)$ عبارت است از

$$H_n(x) = \sum_{j=1}^{n} f(x_j)h_j(x) + \sum_{j=1}^{n} f'(x_j)\tilde{h}_j(x) \qquad (۱۲.۳.۵)$$

با $h_j(x)$ و $\tilde{h}_j(x)$ که در (۲.۶.۳) بخش ۶.۳ تعریف شده‌اند. خطای درونیابی چنین است

$$\mathscr{E}_n(x) = f(x) - H_n(x) = [\psi_n(x)]^2 f[x_1, x_1, \ldots x_n, x_n, x]$$
$$= \frac{[\psi_n(x)]^2}{(2n)!} f^{(2n)}(\xi) \qquad \xi \in [a,b] \qquad (۱۳.۳.۵)$$

با

$$\psi_n(x) = (x - x_1)\ldots(x - x_n)$$

توجه کنید که

$$\psi_n(x) = \frac{\varphi_n(x)}{A_n} \qquad (۱۴.۳.۵)$$

زیرا $\psi_n(x)$ و $\varphi_n(x)$ از درجهٔ n و صفرهای آنها یکی هستند.

با استفاده از (۱۲.۳.۵)، اگر $f(x)$ پیوسته مشتقپذیر باشد، آنگاه

$$\int_a^b w(x)f(x)dx = \int_a^b w(x)H_n(x)dx + \int_a^b w(x)\mathscr{E}_n(x)dx$$
$$\equiv I_n(f) + E_n(f) \qquad (۱۵.۳.۵)$$

درجهٔ دقت حداقل برابر $۲n-۱$ است، زیرا از (۱۳.۳.۵) نتیجه می‌شود که اگر $f(x)$ چندجمله‌یی از درجهٔ کوچکتر از $۲n$ باشد، $\mathscr{E}_n(x) = ۰$. همچنین از (۱۳.۳.۵)،

$$E_n(x^{۲n}) = \int_a^b w(x)\mathscr{E}_n(x)dx = \int_a^b w(x)[\psi_n(x)]^۲dx > ۰ \qquad (۱۶.۳.۵)$$

بنابراین درجهٔ دقت $I_n(f)$ دقیقاً برابر $۲n-۱$ است.

برای پیدا کردن یک فرمول ساده‌تر برای $I_n(f)$،

$$I_n(f) = \sum_{j=۱}^{n} f(x_j)\int_a^b w(x)h_j(x)dx + \sum_{j=۱}^{n} f'(x_j)\int_a^b w(x)\tilde{h}_j(x)dx \qquad (۱۷.۳.۵)$$

نشان می‌دهیم که تمام انتگرالها در دومین مجموع صفرند. یادآوری می‌کنیم که از (۲.۶.۳)،

$$\tilde{h}_j(x) = (x-x_j)[l_j(x)]^۲$$
$$l_j(x) = \frac{\psi_n(x)}{(x-x_j)\psi_n'(x_j)} = \frac{\varphi_n(x)}{(x-x_j)\varphi_n'(x_j)}$$

در آخرین مرحله از (۱۴.۳.۵) استفاده می‌کنیم. بنابراین

$$\tilde{h}_j(x) = (x-x_j)l_j(x)l_j(x) = \frac{\varphi_n(x)l_j(x)}{\varphi_n'(x_j)} \qquad (۱۸.۳.۵)$$

چون درجهٔ (l_j) برابر $n-۱$ است و چون $\varphi_n(x)$ بر همهٔ چندجمله‌ییهای با درجهٔ کوچکتر از n متعامد است، داریم

$$\int_a^b w(x)\tilde{h}_j(x)dx = \frac{۱}{\varphi_n'(x_j)}\int_a^b w(x)\varphi_n(x)l_j(x)dx = ۰ \quad j = ۱, \ldots, n \qquad (۱۹.۳.۵)$$

فرمول انتگرالگیری (۱۵.۳.۵) به‌صورت زیر در می‌آید

$$\int_a^b w(x)f(x)dx = \sum_{j=1}^{n} w_j f(x_j) + E_n(f)$$

$$w_j = \int_a^b w(x)h_j(x)dx \qquad j = 1, \ldots, n \qquad (۲۰.۳.۵)$$

(ب) یکتایی فرمول (۱۹.۳.۵). فرض می‌کنیم که یک فرمول انتگرالگیری عددی

$$\int_a^b w(x)f(x)dx \doteq \sum_{j=1}^{n} v_j f(z_j) \qquad (۲۱.۳.۵)$$

داریم با درجهٔ دقت ناکوچکتر از $2n-1$. فرمول درونیابی ارمیت را برای $f(x)$ در گرههای $z_1, \ldots, z_n$ می‌سازیم. در این صورت برای هر چندجمله‌یی $f(x)$ از درجهٔ نابزرگتر از $2n-1$،

$$f(x) = \sum_{j=1}^{n} f(z_j)h_j(x) + \sum_{j=1}^{n} f'(z_j)\tilde{h}_j(x) \qquad \deg(f) \le 2n-1 \qquad (۲۲.۳.۵)$$

که $h_j(x)$ و $\tilde{h}_j(x)$ با استفاده از $\{z_j\}$ تعریف شده‌اند. رابطهٔ (۲۲.۳.۵) را در $w(x)$ ضرب کرده و از فرض دربارهٔ درجهٔ دقت (۲۱.۳.۵) استفاده می‌کنیم و انتگرال می‌گیریم تا به دست آوریم

$$\sum_{j=1}^{n} v_j f(z_j) = \sum_{j=1}^{n} f(z_j) \int_a^b w(x)h_j(x)dx + \sum_{j=1}^{n} f'(z_j) \int_a^b w(x)\tilde{h}_j(x)dx \qquad (۲۳.۳.۵)$$

برای هر چندجمله‌یی $f(x)$ از درجهٔ نابزرگتر از $2n-1$.

گیریم $f(x) = \tilde{h}_i(x)$. ویژگی (۳.۶.۳) از $\tilde{h}_i(x)$ را به‌کار می‌بریم تا از (۲۳.۳.۵) پیدا کنیم که

$$0 = \int_a^b w(x)\tilde{h}_i(x)dx \qquad i = 1, \ldots, n \qquad (۲۴.۳.۵)$$

مانند (۱۸.۳.۵)، می‌توانیم بنویسیم

$$\tilde{h}_i(x) = (x - z_i)[l_i(x)]^2 = \frac{l_i(x)\omega_n(x)}{\omega_n'(z_i)}$$

$$\omega(x) = (x - z_1)\ldots(x - z_n)$$

در این صورت (۲۴.۳.۵) چنین می‌شود

$$\int_a^b w(x)\omega_n(x)l_i(x)dx = ۰ \qquad i = ۱, ۲, \dots n$$

چون همهٔ چندجمله‌ییهای از درجهٔ نابزرگتر از $n-۱$ را می‌توان به‌صورت ترکیبی از $l_۱(x), \dots, l_n(x)$ نوشت، $w_n(x)$ بر همهٔ چندجمله‌ییهای از درجهٔ نابزرگتر از $n-۱$ متعامد است. با استفاده از یکتایی چندجمله‌ییهای متعامد [از قضیهٔ ۲.۴]، $\omega_n(x)$ مضرب ثابتی از $\varphi_n(x)$ است. بنابراین باید صفرهای آنها یکی باشند، و

$$z_i = x_i \qquad i = ۱, \dots, n$$

برای کامل نمودن اثبات یکتایی، باید نشان دهیم که $w_i = v_i$، که v_i وزن در (۲۱.۳.۵) و w_i وزن در (۱۰.۳.۵) است. از (۲۳.۳.۵) با (۲۴.۳.۵) و $f(x) = h_i(x)$ استفاده می‌کنیم. نتیجه بلافاصله به دست می‌آید، زیرا اکنون $h_i(x)$ با استفاده از $\{x_i\}$ ساخته شده است.

(ج) فرمول خطا. با استنتاج بعضی از ویژگیهای سودمند دیگرِ وزنهای $\{w_i\}$ در (۱۰.۳.۵) آغاز می‌کنیم. با استفاده از تعریف (۲.۶.۳)ی $h_i(x)$

$$\begin{aligned} w_i &= \int_a^b w(x)h_i(x)dx = \int_a^b w(x)[۱ - ۲l_i'(x_i)(x - x_i)][l_i(x)]^۲ dx \\ &= \int_a^b w(x)[l_i(x)]^۲ dx - ۲l_i'(x_i) \int_a^b w(x)(x - x_i)[l_i(x)]^۲ dx \end{aligned}$$

با توجّه به (۱۹.۳.۵)، آخرین انتگرال صفر است، زیرا $\tilde{h}_i(x) = (x - x_i)[l_i(x)]^۲$. بنابراین

$$w_i = \int_a^b w(x)[l_i(x)]^۲ dx > ۰ \qquad i = ۱, ۲ \dots n \qquad (۲۵.۳.۵)$$

و همهٔ وزنها برای همهٔ مقادیر n مثبت‌اند.

برای ساختن w_i؛ با گذاشتن $f(x) = l_i(x)$ در (۲۰.۳.۵)، کار را شروع و توجه می‌کنیم که $E_n(f) = ۰$، زیرا درجهٔ (l_i) برابر $n-۱$ است، با استفاده از $l_i(x_j) = \delta_{ij}$ خواهیم داشت

$$w_i = \int_a^b w(x)l_i(x)dx \qquad i = ۱, \dots, n \qquad (۲۶.۳.۵)$$

برای ساده‌کردن بیشتر این فرمول، اتحاد کریستوفل ـ داربو (قضیهٔ۶.۴) را می‌توان به‌کار برد، و با یک رشته عملیات فرمول (۱۱.۳.۵) را به‌دست آورد. برای شرح جزئیات آیزکسون و کلر (۱۹۶۶ صص ۳۳۳ ـ ۳۳۴) را ببینید.

برای خطای انتگرالگیری اگر $f(x)$، $2n$ بار در $[a,b]$ پیوسته مشتقپذیر باشد، آنگاه

$$\begin{aligned} E_n(f) &= \int_a^b w(x)\mathscr{E}_n(x)dx \\ &= \int_a^b w(x)[\psi_n(x)]^2 f[x_1, x_1, \ldots, x_n, x_n, x]dx \\ &= f[x_1, x_1, \ldots, x_n, x_n, \xi]\int_a^b w(x)[\psi_n(x)]^2 dx \qquad \xi \in [a,b] \end{aligned}$$

که در آخرین مرحله از قضیهٔ مقدار میانگین انتگرال استفاده شده است. با استفاده از (۱۴.۳.۵) در آخرین انتگرال و گذاشتن مشتق به‌جای تفاضلات منقسم خواهیم داشت

$$E_n(f) = \frac{f^{(2n)}(\eta)}{(2n)!}\int_a^b w(x)\frac{[\varphi_n(x)]^2}{A_n^2}dx$$

که فرمول خطای (۱۰.۳.۵) را به‌دست می‌دهد. ■

انتگرالگیری گاوس ـ لژاندر برای $w(x) \equiv 1$ فرمول گاوس روی $[-1,1]$ چنین است

$$\int_{-1}^{1} f(x)dx \doteq \sum_{j=1}^{n} w_j f(x_j) \qquad (۲۷.۳.۵)$$

که گره‌ها صفرهای $P_n(x)$، چند جمله‌یی درجهٔ n لژاندر، در بازهٔ $[-1,1]$ هستند. وزنها عبارت‌اند از

$$w_i = \frac{-2}{(n+1)P'_n(x_i)P_{n+1}(x_i)} \qquad i = 1,2,\ldots n \qquad (۲۸.۳.۵)$$

و

$$E_n(f) = \frac{2^{2n+1}(n!)^4}{(2n+1)[(2n)!]^2}\cdot\frac{f^{(2n)}(\eta)}{(2n)!} = e_n\frac{f^{(2n)}(\eta)}{(2n)!} \qquad (۲۹.۳.۵)$$

برای $-1 < \eta < 1$.

برای انتگرالها در بازه‌های متناهی دیگر و تابع وزن $w(x) \equiv 1$، تبدیل متغیرهای خطی زیر را به‌کار می‌بریم:

$$\int_a^b f(t)dt = \left(\frac{b-a}{2}\right)\int_{-1}^{1} f\left(\frac{a+b+x(b-a)}{2}\right)dx \qquad (۳۰.۳.۵)$$

که انتگرال را به بازهٔ استاندهٔ $[-1,1]$ بدل می‌کند.

برای راحتی، جدول ۱۰.۵ را که گرهها و وزنها را برای فرمول (۲۷.۳.۵) برای مقادیر کوچک

1. Stroud

جدول ۱۰.۵ گرههای لژاندار-گاوس و وزنها

n	x_i	ω_i
۲	±۰٫۵۷۷۳۵۰۲۶۹۲	۱٫۰
۳	±۰٫۷۷۴۵۹۶۶۶۹۲	۰٫۵۵۵۵۵۵۵۵۵۶
	۰٫۰	۰٫۸۸۸۸۸۸۸۸۸۹
۴	±۰٫۸۶۱۱۳۶۳۱۱۶	۰٫۳۴۷۸۵۴۶۴۵۱
	±۰٫۳۳۹۹۸۱۰۴۳۶	۰٫۶۵۲۱۴۵۱۵۴۹
۵	±۰٫۹۰۶۱۷۹۸۴۵۹	۰٫۲۳۶۹۲۶۸۸۵۱
	±۰٫۵۳۸۴۶۹۳۱۰۱	۰٫۴۷۸۶۲۸۶۷۰۵
	۰٫۰	۰٫۵۶۸۸۸۸۸۸۸۹
۶	±۰٫۹۳۲۴۶۹۵۱۴۲	۰٫۱۷۱۳۲۴۴۹۲۴
	±۰٫۶۶۱۲۰۹۳۸۶۵	۰٫۳۶۰۷۶۱۵۷۳۰
	±۰٫۲۳۸۶۱۹۱۸۶۱	۰٫۴۶۷۹۱۳۹۳۴۶
۷	±۰٫۹۴۹۱۰۷۹۱۲۳	۰٫۱۲۹۴۸۴۹۶۶۲
	±۰٫۷۴۱۵۳۱۱۸۵۶	۰٫۲۷۹۷۰۵۳۹۱۵
	±۰٫۴۰۵۸۴۵۱۵۱۴	۰٫۳۸۱۸۳۰۰۵۰۵
	۰٫۰	۰٫۴۱۷۹۵۹۱۸۳۷
۸	±۰٫۹۶۰۲۸۹۸۵۶۵	۰٫۱۰۱۲۲۸۵۳۶۳
	±۰٫۷۹۶۶۶۴۷۷۴	۰٫۲۲۲۳۸۱۰۳۴۵
	±۰٫۵۲۵۵۳۲۴۰۹۹	۰٫۳۱۳۷۰۶۶۴۵۹
	±۰٫۱۸۳۴۳۴۶۴۲۵	۰٫۳۶۲۶۸۳۷۸۳۴

n می‌دهد، داده‌ایم. برای مقادیر بزرگ n جدول بسیار کامل در استرود و سکرست[1] (۱۹۶۶) را ببینید، که تا $n = ۵۱۲$ پیش رفته است.

مثال انتگرال (۱۱.۱.۵) را محاسبه کنید،

$$I = \int_{۰}^{\pi} e^x \cos(x) dx \doteq -۱۲٫۰۷۰۳۴۶۳۱۶۴$$

که قبلاً در بخش ۱.۵ به‌عنوان یک مثال برای قاعدهٔ ذوزنقه‌یی (جدول ۱.۵) و قاعدهٔ سیمپسون (جدول ۳.۵) به‌کار برده شده‌است. نتایجی که در جدول ۱۱.۵ داده شده‌اند مزیّت محسوس انتگرالگیری گاوسی را نشان می‌دهند.

1. Secrest

جدول ۱۱.۵ انتگرالگیری گاوسی برای (۱۱.۱.۵)

n	I_n	$I - I_n$
۲	−۱۲٫۳۳۶۲۱۰۴۶۵۷۰	۲٫۶۶E − ۱
۳	−۱۲٫۱۲۷۴۲۰۴۵۰۱۷	۵٫۷۱E − ۲
۴	−۱۲٫۰۷۰۱۸۹۴۹۰۲۹	−۱٫۵۷E − ۴
۵	−۱۲٫۰۷۰۳۲۸۵۳۵۸۹	−۱٫۷۸E − ۵
۶	−۱۲٫۰۷۰۳۴۶۳۳۱۱۰	۱٫۴۷E − ۸
۷	−۱۲٫۰۷۰۳۴۶۳۱۷۵۳	۱٫۱۴E − ۹
۸	−۱۲٫۰۷۰۳۴۶۳۱۶۳۹	−۴٫۲۵E − ۱۳

یک قضیّهٔ کلی خطا. در تلاش برای توضیح همگرایی بسیار خوب انتگرالگیری گاوسی، قضیّهٔ مفیدی را می‌آوریم. در زیر بخش بعد، خطا در انتگرالگیری گاوس ـ لژاندر را با تفصیل بیشتر بررسی خواهیم کرد.

قضیهٔ ۴.۵ فرض کنید $[a, b]$ متناهی باشد. در این صورت خطا در انتگرالگیری گاوسی،

$$E_n(f) = \int_a^b w(x)f(x)dx - \sum_{j=1}^{n} w_j f(x_j)$$

در رابطهٔ زیر صدق می‌کند

$$|E_n(f)| \leq 2\left[\int_a^b w(x)dx\right]\rho_{2n-1}(f) \qquad n \geq 1 \quad (۳۱.۳.۵)$$

که $\rho_{2n-1}(f)$ خطای مینیماکس (۱.۲.۴) است.

برهان برای هر چندجمله‌یی $p(x)$ از درجهٔ نابزرگتر از $2n-1$، $E_n(p) = 0$. همچنین تابع خطا به ازای هر $F, G \in C[a, b]$ در رابطهٔ زیر صدق می‌کند

$$E_n(F + G) = E_n(F) + E_n(G)$$

گیریم $p(x) = q^*_{2n-1}(x)$، تقریب مینیماکس از درجهٔ نابزرگتر از $2n-1$ برای $f(x)$ بر $[a, b]$

باشد. در این صورت

$$E_n(f) = E_n(f) - E_n(q^*_{2n-1}) = E_n(f - q^*_{2n-1})$$
$$= \int_a^b w(x)[f(x) - q^*_{2n-1}(x)]dx - \sum_{j=1}^{n} w_j[f(x_j) - q^*_{2n-1}(x_j)]$$
$$| E_n(f) | \leq \| f - q^*_{2n-1} \|_\infty \left[\int_a^b w(x)dx + \sum_{j=1}^{n} | w_j | \right]$$

با توجه به (۲۵.۳.۵)، همهٔ w_j ها بزرگتر از صفرند. همچنین چون $p(x) \equiv 1$ از درجهٔ صفر است، پس

$$\sum_{j=1}^{n} w_j = \int_a^b w(x)dx$$

و برهان (۳۱.۳.۵) کامل می‌شود. ■

از نتایج بخش ۶.۴ و ۷.۴، سرعت همگرایی به صفر $\rho_m(f)$ با هموارای انتگرالده، افزایش می‌یابد. از (۲۱.۳.۵)، همین نتیجه برای انتگرالگیری گاوسی درست است. در مقابل، قاعدهٔ ذوزنقه‌یی مرکب، بدون توجه به هموارای $f(x)$، معمولاً سریعتر از مرتبهٔ h^2 همگرا نخواهد بود [به‌ویژه، اگر $f'(b) - f'(a) \neq 0$]. انتگرالگیری گاوسی، برخلاف بسیاری از قواعد انتگرالگیری مرکب از هموارای بیشتر انتگرالده بهره می‌گیرد.

مثال قاعدهٔ انتگرالگیری گاوس ـ لژاندر را برای انتگرالگیری زیر در نظر می‌گیریم

$$I = \int_0^1 e^{-x^2}dx \doteq 0.۷۴۶۸۲۴۱۳۲۸۱۲ \qquad (۳۲.۳.۵)$$

جدول ۱۲.۵ شامل کرانهای خطا بر اساس فرمول (۳۱.۳.۵)،

$$| E_n(f) | \leq 2\rho_{2n-1}(f) \qquad (۳۳.۳.۵)$$

همراه با خطای واقعی است. کران خطا تقریباً به اندازهٔ خطای واقعی است.

بحث انتگرالگیری گاوس ـ لژاندر نخست سعی می‌کنیم عبارت خطای (۲۹.۳.۵) را بیشتر قابل فهم کنیم. ابتدا تعریف می‌کنیم

$$M_m = \operatorname*{Max}_{-1 \leq x \leq 1} \frac{| f^{(m)}(x) |}{m!} \qquad m \geq 0 \qquad (۳۴.۳.۵)$$

جدول ۱۲.۵ انتگرالگیری گاوسی (۳۲.۳.۵)

n	$E_n(f)$	(۳۲.۳.۵)
۱	$-۳٫۲۰\text{E}-۲$	$۱٫۰۶\text{E}-۱$
۲	$۲٫۲۹\text{E}-۴$	$۱٫۳۳\text{E}-۳$
۳	$۹٫۵۵\text{E}-۶$	$۳٫۲۴\text{E}-۵$
۴	$-۳٫۳۵\text{E}-۷$	$۹٫۲۴\text{E}-۷$
۵	$۶٫۰۵\text{E}-۹$	$۱٫۶۱\text{E}-۸$

برای ردهٔ بزرگی از توابع نامتناهی مشتقپذیر f بر $[-۱, ۱]$، داریم $\text{Sup}_{m\geq ۰} M_m < \infty$. برای مثال، اگر ناحیهٔ R از صفحهٔ مختلط با

$$R = \{z : | z - x | \leq ۱ \text{ به‌ازای یک } ۱ - \leq x \leq ۱\}$$

تعریف شده باشد، و $f(z)$ در R تحلیلی باشد، این امر درست است. در بسیاری از توابع، هرگاه $m \to \infty$ آنگاه $M_m \to ۰$، برای مثال $f(x) = e^x$ و $\cos x$.

از ترکیب (۲۹.۳.۵) و (۳۴.۳.۵) به‌دست می‌آوریم

$$| E_n(f) | \leq e_n M_{۲n} \qquad n \geq ۱ \qquad (۳۵.۳.۵)$$

و اندازهٔ e_n در بررسی سرعت همگرایی اساسی است.

برای آشنایی بیشتر با عامل e_n، می‌توان آن را با استفاده از فرمول استرلینگ

$$n! \doteq e^{-n} n^n \sqrt{۲\pi n}$$

که به مفهوم خطای نسبی وقتی $n \to \infty$، درست است براورد نمود. در این صورت چنین به‌دست می‌آوریم

$$e_n \doteq \frac{\pi}{۴n} \qquad \text{وقتی} \quad n \to \infty \qquad (۳۶.۳.۵)$$

این یک براورد کاملاً خوبی است. برای مثال، $e_۵ = ۰٫۰۰۲۹۳$ و (۳۶.۳.۵) براورد $۰٫۰۰۳۰۷$ را می‌دهد. چنانچه این برآورد با (۳۵.۳.۵) ترکیب شود، نتیجه می‌دهد:

$$| E_n(f) | \leq \frac{\pi}{۴n} \cdot M_{۲n} \qquad (۳۷.۳.۵)$$

که وقتی $n \to \infty$، به مفهوم مجانبی یک کران صحیح است. این رابطه نشان می‌دهد که $E_n(f) \to ۰$ یعنی بایک نرخ نمایی به‌صورت تابعی از n کاهش می‌یابد. این نرخ را با نرخ چندجمله‌ییهای $۱/(n^۲)$ و $۱/(n^۴)$ به ترتیب برای قاعده‌های ذوزنقه‌یی و سیمپسون مقایسه نمایید.

برای مطالعهٔ انتگرالده‌هایی که بینهایت مشتقپذیر نیستند، می‌توانیم شکل خطای هسته‌یی پئانو را درست مانند بخش ۱.۵ برای قاعده‌های سیمپسون و ذوزنقه‌یی به کار بریم. اگر $f(x)$ در $[-۱, ۱]$، rبار مشتقپذیر و $f^{(r)}(x)$ در $[-۱, ۱]$ انتگرالپذیر باشد، آنگاه برای $K_{n,r}(t)$، یک هستهٔ مناسب پئانو، داریم

$$E_n(f) = \int_{-۱}^{۱} K_{n,r}(t) f^{(r)}(t) dt \qquad n > \frac{r}{۲} \qquad (۳۸.۳.۵)$$

شیوهٔ ساختن $K_{n,r}(t)$ درست مانند ساختن هسته‌های پئانو (۲۱.۱.۵) و (۲۵.۱.۵) در بخش (۱.۵) است. با توجهٔ به (۳۸.۳.۵) داریم

$$| E_n(f) | \leq e_{n,r} M_r$$

$$e_{n,r} = r! \int_{-۱}^{۱} | K_{n,r}(t) | dt \qquad (۳۹.۳.۵)$$

مقادیر $e_{n,r}$ که در جدول ۱۳.۵ داده شده‌اند از استرود ـ سکرست (۱۹۶۶، صص ۱۵۲ ـ ۱۵۳) گرفته شده‌است. جدول ۱۳.۵ نشان می‌دهد که اگر f دوبار پیوسته مشتقپذیرباشد، انتگرالگیری گاوسی حداقل به همان سرعت قاعدهٔ ذوزنقه‌یی (۵.۱.۵) همگراست. با استفاده از (۳۹.۳.۵) و جدول، می‌توانیم یک کران تجربی بسازیم

$$| E_n(f) | \leq \frac{۰٫۴۲}{n^۲} \left[\underset{-۱ \leq x \leq ۱}{\text{Max}} | f''(x) | \right] \qquad (۴۰.۳.۵)$$

جدول ۱۳.۵ مقادیر ثابت خطای $e_{n,r}$ برای (۳۹.۳.۵)

n	$e_{n,۲}$	نسبت	$e_{n,۴}$	نسبت
۲	۰٫۱۶۲		۰٫۱۷۸	
۴	۰٫۴۳۷E − ۱	۳٫۷	۰٫۶۴۷E − ۲	۲۷٫۵
۸	۰٫۱۱۸E − ۱	۳٫۷	۰٫۴۱۷E − ۳	۱۵٫۵
۱۶	۰٫۳۱۱E − ۲	۳٫۸	۰٫۲۷۹E − ۴	۱۴٫۹
۳۲	۰٫۸۰۰E − ۳	۳٫۹	۰٫۱۸۳E − ۵	۱۵٫۳
۶۴	۰٫۲۰۳E − ۳	۳٫۹		
۱۲۸	۰٫۵۱۱E − ۴	۴٫۰		

از فرمول متناظر با (۷.۱.۵) برای قاعدهٔ ذوزنقه‌یی در $[-۱, ۱]$ چنین نتیجه می‌شود

$$|\text{خطای ذوزنقه‌یی}| \leq \frac{۰٫۶۷}{n^۲}\left[\underset{-۱\leq x\leq ۱}{\text{Max}} |f''(x)|\right]$$

که اندکی بزرگتر از (۴۰.۳.۵) است. در محاسبات واقعی، انتگرالگیری گاوسی ظاهراً همیشه بر قاعدهٔ ذوزنقه‌یی برتری دارد، بجز مواردی که انتگرالده‌ها دوره‌یی باشند و بازهٔ انتگرالگیری مضرب صحیحی از دورهٔ انتگرالده باشد، مانند جدول ۷.۵. بحث مشابهی را با استفاده از جدول ۱۳.۵ با $e_{n,۴}$ برای انتگرالده‌های $f(x)$ که چهار بار مشتقپذیر باشند می‌توان انجام داد. (مسألهٔ ۲۰ را ببینید).

مثال سه مثال دیگر می‌آوریم که به خوشرفتاری مثالهای جدولهای ۱۱.۵ و ۱۲.۵ نیستند. مثالهای زیر را در نظر می‌گیریم

$$I^{(۱)} = \int_۰^۱ \sqrt{x}\, dx = \frac{۲}{۳} \qquad I^{(۲)} = \int_۰^۵ \frac{dx}{۱+(x-\pi)^۲} \doteq ۲٫۳۳۹۷۶۶۲۸۳۶۷$$

$$I^{(۳)} = \int_۰^{۲\pi} e^{-x}\sin(۵۰x)dx \doteq ۰٫۰۱۹۹۵۴۶۶۹۲۷۸ \qquad (۴۱.۳.۵)$$

مقادیر جدول ۱۴.۵ نشان می‌دهند که انتگرالگیری گاوسی، علی‌رغم رفتار بد انتگرالده، هنوز خیلی کاراست.

نتایج $I^{(۱)}$ را با نتایج جدول ۶.۵ برای قاعده‌های ذوزنقه‌یی و سیمپسون مقایسه کنید. انتگرالگیری گاوسی با خطایی متناسب با $۱/n^۳$ همگراست و حال آنکه در جدول ۶.۵، خطا با نرخ متناسب با $۱/n^{۱٫۵}$ همگراست. انتگرال زیر را که در آن $\alpha > -۱$ و غیر صحیح، $f(x)$ هموار و $f(۰) \neq ۰$

جدول ۱۴.۵ مثالهای انتگرالگیری گاوسی (۴۱.۳.۵)

n	$I^{(۱)} - I_n^{(۱)}$	نسبت	$I^{(۲)} - I_n^{(۲)}$	$I^{(۳)} - I_n^{(۳)}$
۲	$-۷٫۲۲\text{E}-۳$		$۳٫۵۰\text{E}-۱$	$۳٫۴۸\text{E}-۱$
۴	$-۱٫۱۶\text{E}-۳$	۶٫۲	$-۹٫۱۹\text{E}-۲$	$-۱٫۰۴\text{E}-۱$
۸	$-۱٫۶۹\text{E}-۴$	۶٫۹	$-۴٫۰۳\text{E}-۳$	$-۱٫۸۰\text{E}-۲$
۱۶	$-۲٫۳۰\text{E}-۵$	۷٫۴	$-۶٫۲۴\text{E}-۷$	$-۳٫۳۴\text{E}-۱$
۳۲	$-۳٫۰۰\text{E}-۶$	۷٫۶	$-۲٫۹۸\text{E}-۱۱$	$۱٫۱۶\text{E}-۱$
۶۴	$-۳٫۸۴\text{E}-۷$	۷٫۸	–	$۱٫۵۳\text{E}-۱$
۱۲۸	$-۴٫۸۴\text{E}-۸$	۷٫۹	–	$۶٫۶۹\text{E}-۱۵$

درنظر می‌گیریم

$$I(f) = \int_{0}^{1} x^{\alpha} f(x)dx \qquad (۴۲.۳.۵)$$

دونالدسن[1] و الیوت[2] (۱۹۷۲) نشان داده‌اند که خطا در انتگرالگیری گاوس ـ لژاندر در (۴۲.۳.۵) دارای براورد مجانبی

$$E_n(f) \doteq \frac{c(f,\alpha)}{n^{2(1+\alpha)}} \qquad (۴۳.۳.۵)$$

است. این نتیجه با $I_n^{(1)}$ در جدول ۱۴.۵ با $\alpha = 1/2$ تطابق دارد. نتایج مهم دیگری از انتگرالگیری گاوس ـ لژاندر، در مقالهٔ دونالدسن و الیوت داده شده است.

همگرایی اولیهٔ $I_n^{(2)}$ به $I^{(2)}$ خیلی آهسته است ولی وقتی n افزایش می‌یابد، سرعت همگرایی به طور چشمگیری افزایش می‌یابد. برای $n \geq 64$، در محدودهٔ حساب ماشین، $I_n^{(2)} = I^{2}$. این نتایج را همچنین با نتایج جدول ۵.۵ برای قاعده‌های ذوزنقه‌یی و سیمپسون مقایسه کنید.

تقریبهای جدول ۱۴.۵ برای $I^{(3)}$ بسیارکم است، چون نوسان انتگرالده زیاد است. انتگرالده ۱۰۱ صفر در بازهٔ انتگرالگیری دارد. برای پیداکردن یک مقدار دقیق $I_n^{(3)}$، درجهٔ چندجمله‌یی تقریب زن در انتگرالگیری زیربنائیِ گاوسی باید بسیار بزرگ باشد. $I_n^{(3)}$، با $n = 128$، یک تقریب بسیار دقیق برای $I^{(3)}$ است.

توضیحات کلی انتگرالگیریهای گاوسی تعدادی نقاط قوت و ضعف دارند.

۱. به علّت شکل گرهها و وزنها و در نتیجه نیاز به استفاده از جدول، بسیاری از افراد یک فرمول ساده‌تری مانند قاعدهٔ سیمپسون، را ترجیح می‌دهند. وقتی که برای انتگرالگیری ازکامپیوتر استفاده می‌شود این امر مسأله‌ای نیست. برنامه‌هایی باید نوشته شوند که شامل این وزنها و نقاط گرهی برای مقادیر استانهٔ n باشند، مثلاً $n = 2, 4, 6, 8, \ldots, 512$ [که از استرود و سکرست (۱۹۶۶) گرفته شده‌اند.]. به علاوه تعداد زیادی برنامه‌های سریع برای محاسبهٔ نقاط گرهها و وزنها، برای بعضی از توابع وزن متداول، وجود دارند. در بین الگوریتمهای شناخته شده، بهتر از همه الگوریتمی است که در گلوب[3] و ولش[4] (۱۹۶۹) آمده است.

۲. براورد خطا مشکل است، بنابراین معمولاً

$$I - I_n \doteq I_m - I_n \qquad (۴۴.۳.۵)$$

1. Donaldson 2. Elliot 3. Golub 4. Welsch

را برای یک $m > n$ انتخاب می‌کنند. مثلاً $m = n + ۲$ را برای انتگرالده‌های خوشرفتار، و در غیر این صورت $m = ۲n$. این نتیجه‌ای بهتر از آنچه لازم است به‌دست می‌دهد، حتی با افزایش تعداد ارزشیابیهای تابعی، انتگرالگیری گاوسی باز از خیلی از روشهای دیگر سریعتر است.

۳. گره‌ها برای هر فرمول I_n از گرههای فرمولهای قبلی I_m متمایزند و این امر موجب یک عدم کارایی می‌شود. اگر I_n براساس براوردِ خطایی مانند (۴۴.۳.۵)، دقیق نباشد باید مقدار جدیدی برای I_n حساب کنیم. این باعث اتلاف انرژی خواهد بود، چون از هیچ یک از مقادیر قبلی انتگرالده نمی‌توان در محاسبهٔ مجدد I_n، استفاده کرد. این موضوع خیلی کاملتر در آخرین قسمت این بخش توضیح داده شده است و به روشهای جدیدی منجر می‌شود که این نقطه ضعف را ندارند. معهذا، در بسیاری از حالات، عدم کارایی انتگرالگیری گاوسی، به علت نرخ سریع همگرایی معمولاً مهم نخواهد بود.

۴. اگر قرار باشد که ردهٔ وسیعی از انتگرالهای با ماهیّت مشابه محاسبه شوند،مانند زیر عمل می‌کنیم.

تعدادی انتگرال نماینده شامل یک انتگرال که احتمالاً انتگرالده آن بدترین رفتار را دارد انتخاب می‌کنیم. n را مقداری تعیین می‌کنیم که $I_n(f)$ در بین مجموعهٔ نماینده‌ها دارای دقت کافی باشد. سپس این مقدار n را ثابت نگه می‌داریم و $I_n(f)$ را به‌عنوان انتگرال عددی تمام اعضای ردهٔ اصلی انتگرالها به‌کار می‌بریم.

۵. انتگرالگیری گاوسی برای خیلی از انتگرالده‌های نزدیک به تکین، به‌طور خیلی مؤثر عمل می‌کند، همان گونه که در (۴۳.۳.۵) برای (۴۲.۳.۵) نشان داده شده است. ولی تمام نقاط با رفتار تکین باید در نقاط انتهایی بازهٔ انتگرالگیری قرار گیرند. انتگرالگیری گاوسی برای انتگرالی چون

$$\int_{۰}^{۱} \sqrt{| x - ۰٫۷ |}dx$$

که دارای یک نقطهٔ تکین در بازهٔ انتگرالگیری است، بسیار ضعیف است. (بسیاری از روشهای انتگرالگیری عددی دیگر نیز برای این انتگرال ضعیف عمل می‌کنند). این انتگرال باید به شکل زیر شکسته و محاسبه شود

$$\int_{۰}^{۰٫۷} \sqrt{۰٫۷ - x}\, dx + \int_{۰٫۷}^{۱} \sqrt{x - ۰٫۷}\, dx$$

تعمیمهایی که از نقاط گرهی مجدداً استفاده می‌کنند فرض کنید که فرمول انتگرالگیری زیر را داریم:

$$I_n(f) = \sum_{k=۱}^{n} w_k f(x_k) \doteq \int_a^b w(x) f(x) dx \qquad (۴۵.۳.۵)$$

می‌خواهیم یک فرمول انتگرالگیری جدیدی بسازیم که در آن از n نقطهٔ x_1، ...، x_n و m نقطهٔ جدید x_{n+1}، ...، x_{n+m} استفاده شود:

$$I_{n+m}(f) = \sum_{k=1}^{n+m} v_k f(x_k) \doteq \int_a^b w(x)f(x)dx \qquad (۴۶.۳.۵)$$

این $n+2m$ پارامتر نامعین، یعنی گره‌های $x_{n+1}, \ldots, x_{n+m}$ و وزنهای u_1، ...، u_{n+m} باید طوری انتخاب شوند که درجهٔ دقت (۴۵.۳.۵) در حد امکان زیاد باشد. ما دنبال یک فرمول با درجهٔ دقت $n+2m-1$ هستیم. اینکه آیا چنین فرمولی را می‌توان با گرههای جدید x_{n+1}، ...، x_{n+m} که در $[a,b]$ قرار دارند، تعیین کرد معلوم نیست.

برای حالتی که (۴۵.۳.۵) یک فرمول گاوسی است، کرونرود تعمیهای (۴۶.۳.۵) را برای $m = n+1$ مطالعه نموده است. این نوع فرمولهای زوج، راه کم هزینه‌تری (در مقایسه با استفاده از یک قاعدهٔ گاوسی با $2n+1$ نقطهٔ گرهی) برای تولید برآورد خطا برای قاعدهٔ گاوسی به‌دست می‌دهد و درجهٔ دقت آنقدر بالاست که نوع دقت متناظر با قاعده‌های گاوسی را ایجاد می‌کند.

شکل دیگری از موضوع اخیر در پاترسن[1] (۱۹۶۸) ارائه شده‌است. برای $w(x) \equiv 1$، او با یک قاعدهٔ گاوس ـ لژاندر $I_{n,0}(f)$ شروع می‌نماید. سپس با تکرار فرمولهای (۴۶.۳.۵) دنباله‌ای از فورمولها را از عناصر قبلی دنباله با $m = n+1$ به وجود می‌آورد. یکی از مقاله‌های پاترسن (۱۹۷۳) شامل الگوریتمی است بر پایهٔ دنبالهٔ قواعد I_2، I_7، I_{15}، I_{31}، I_{63}، I_{127}، I_{255}؛ فرمول I_2 قاعدهٔ گاوس سه نقطه‌یی است. دنبالهٔ دیگری از این نوع $\{I_{10}, I_{21}, I_{43}, I_{87}\}$ توسط پیسنز[2] و همکاران (۱۹۸۳، صص ۲۷،۲۶،۱۹) داده شده است، که I_{10} قاعدهٔ گاوس ده نقطه‌یی است. تمام فرمولهای این چنینی پاترسن تاکنون، همه دارای نقاط گرهی داخل بازهٔ انتگرالگیری با وزنهای مثبت بوده‌اند.

درجهٔ دقت قاعده‌های پاترسن با تعداد نقاط گرهی افزایش می‌یابد. برای دنبالهٔ $I_3, I_7, \ldots, I_{255}$، که قبلاً اشاره شد، درجهٔ دقت به ترتیب چنین‌اند ۳۸۳، ۱۹۱، ۹۵، ۴۷، ۲۳، ۱۱، ۵ $= d$. چون وزنها مثبت‌اند، اثبات قضیهٔ ۴.۵ را می‌توان تکرار کرد تا نشان داد قاعده‌های پاترسن سریعاً همگرا هستند.

بحث بیشتر قاعده‌های پاترسن و کرونرود، به انضمام برنامه‌ها، در (صفحات ۱۷.۱۵) پیسنز و همکاران (۱۹۸۳) داده شده است. این کتابها، در ضمن، مراجع بسیاری از نوشته‌هایی را که در این زمینه منتشر شده‌اند، داده‌اند.

---

1. Patterson 2. Piessens et al

مثال گیریم (۴۵.۳.۵) قاعدهٔ سه‌نقطه‌یی گاوسی بر $[-۱, ۱]$ باشد:

$$I_۳(f) = \frac{۸}{۹}f(۰) + \frac{۵}{۹}[f(-\sqrt{۰٫۶}) + f(\sqrt{۰٫۶})] \qquad (۴۷.۳.۵)$$

قاعدهٔ کرون‌رود برای آن چنین است

$$\begin{aligned} I_۷(f) =& \alpha_۰ f(۰) + \alpha_۱[f(-\sqrt{۰٫۶}) + f(\sqrt{۰٫۶})] \\ &+ \alpha_۲[f(-\beta_۱) + f(\beta_۱)] + \alpha_۳[f(-\beta_۲) + f(\beta_۲)] \qquad (۴۸.۳.۵) \end{aligned}$$

که $\beta_۱^۲$ و $\beta_۲^۲$ به ترتیب ریشه‌های کوچکتر و بزرگتر معادلهٔ زیر هستند.

$$x^۲ - \frac{۱۰}{۹}x + \frac{۱۵۵}{۸۹۱} = ۰$$

وزنهای $\alpha_۳, \alpha_۲, \alpha_۱, \alpha_۰$ از انتگرالگیری چندجمله‌یی $p_۷(x)$ لاگرانژ بر $[-۱, ۱]$ که $f(x)$ را در نقاط $\{۰, \pm\sqrt{۰٫۶}, \pm\beta_۱, \pm\beta_۲\}$ درونیابی می‌نماید، به‌دست آمده‌اند. مقادیر تقریبی عبارتند از

$$\alpha_۰ = ۰٫۴۵۰۹۱۶۵۳۸۶۵۸ \quad \alpha_۱ = ۰٫۲۶۸۴۸۸۰۸۹۸۶۸$$
$$\alpha_۲ = ۰٫۴۰۱۳۹۷۴۱۴۷۷۶ \quad \alpha_۳ = ۰٫۱۰۴۶۵۶۲۲۶۰۲۶$$

۴.۵ فرمولهای خطای مجانبی و کاربردهای آنها

تعریف فرمول خطای مجانبی (۱۰.۱.۵) را برای یک فرمول انتگرالگیری عددی به یاد می‌آوریم: $\tilde{E}_n(f)$ یک فرمول خطای مجانبی برای $E_n(f) = I(f) - I_n(f)$ است اگر

$$\lim_{n\to\infty} \frac{\tilde{E}_n(f)}{E_n(f)} = ۱ \qquad (۱.۴.۵)$$

یا هم‌ارز با آن

$$\lim_{n\to\infty} \frac{E_n(f) - \tilde{E}_n(f)}{E_n(f)} = ۰$$

(۹.۱.۵) و (۱۸.۱.۵) از بخش ۱.۵ مثالهایی در این مورد هستند.

با به‌دست آوردن فرمول خطای مجانبی، شکل یا ساختار خطا را به‌دست می‌آوریم. با این اطلاعات، می‌توانیم یا خطا در $I_n(f)$ را مثل در جدولهای ۱.۵ و ۳.۵ براورد نماییم، یا یک فرمول جدید و دقیقتری مانند قاعدهٔ ذوزنقه‌یی اصلاح‌شده در (۱۲.۱.۵) پیدا کنیم. در این بخش، این هر دو حالت بعداً نشان داده شده و به روش انتگرالگیری همگرای سریع رامبرگ ختم شده‌است. مطلب را با بسط بیشتر فرمولهای خطای مجانبی آغاز می‌کنیم.

چندجمله‌ییهای برنولی برای استفاده در قضیهٔ بعد، چندجمله‌ییهای برنولی $B_n(x)$، $n \geq 0$ را معرفی می‌کنیم. این چندجمله‌ییها به‌طور ضمنی با تابع مولد زیر تعریف شده‌اند

$$\frac{t(e^{xt}-1)}{e^t-1} = \sum_{j=1}^{\infty} B_j(x)\frac{t^j}{j!} \qquad (۲.۴.۵)$$

چند تای اول این چند جمله‌ییها به قرار زیرند:

$$B_0(x) = 1, \quad B_1(x) = x, \quad B_2(x) = x^2 - x,$$
$$B_3(x) = x^3 - \frac{3x^2}{2} + \frac{x}{2}, \quad B_4(x) = x^2(1-x)^2 \qquad (۳.۴.۵)$$

در این چندجمله‌ییها

$$B_k(0) = 0 \qquad k \geq 1 \qquad (۴.۴.۵)$$

برای محاسبهٔ این چندجمله‌ییها روابط بازگشتی که به سادگی قابل محاسبه‌اند وجود دارند (مسألهٔ ۲۳ را ببینید).

اعداد برنولی نیز، که به طور ضمنی با رابطهٔ

$$\frac{t}{e^t-1} = \sum_{j=0}^{\infty} B_j \cdot \frac{t^j}{j!} \qquad (۵.۴.۵)$$

تعریف می‌شوند، مورد توجه ما هستند. چند تای اوّل این اعداد چنین‌اند:

$$B_0 = 1 \quad B_1 = -\frac{1}{2} \quad B_2 = \frac{1}{6} \quad B_4 = \frac{-1}{30}$$
$$B_6 = \frac{1}{42} \quad B_8 = \frac{-1}{30} \qquad (۶.۴.۵)$$

و برای تمام اعداد صحیح فرد $j \geq 3$، $B_j = 0$. برای پیدا کردن یک رابطه بین چندجمله‌ییهای $B_j(x)$ برنولی، از (۲.۴.۵) نسبت به x در $[0, 1]$ انتگرال می‌گیریم. در این صورت

$$1 - \frac{t}{e^t-1} = \sum_{1}^{\infty} \frac{t^j}{j!} \int_0^1 B_j(x)dx$$

و بنابراین

$$B_j = -\int_0^1 B_j(x)dx \qquad j \geq 1 \qquad (۷.۴.۵)$$

ما همچنین به یک تعریف توسیع دوره‌یی از $B_j(x)$ نیاز داریم

$$\overline{B}_j(x) = \begin{cases} B_j(x) & ۰ \le x \le ۱ \\ \overline{B}_j(x-۱) & x \ge ۱ \end{cases} \qquad (۸.۴.۵)$$

فرمول اویلر ـ مک لورن قضیهٔ زیر یک فرمول خیلی مفصّلی برای مقدار مجانبی خطا در قاعدهٔ ذوزنقه‌یی به‌دست می‌دهد. این قضیه اساس بسیاری از تحلیلهای مجانبی خطا، در این بخش، است. ارتباط آن با بعضی از فرمولهای دیگر انتگرالگیری، بعداً در این بخش ظاهر خواهد شد.

قضیهٔ ۵.۵ (فرمول اویلر ـ مک لورن) گیریم $m \ge ۰$، $n \ge ۱$، و تعریف می‌کنیم $h = (b-a)/n$، $x_j = a + jh$ برای $j = ۰, ۱, \ldots n$. همچنین فرض می‌کنیم $f(x)$، $۲m+۲$بار پیوسته مشتقپذیر بر $[a,b]$ برای مقداری از $m \ge ۰$ باشد. در این صورت برای خطا در قاعدهٔ ذوزنقه‌یی،

$$\begin{aligned} E_n(f) &= \int_a^b f(x)dx - h\sum_{j=۰}^{n}{}'' f(x_j) \\ &= -\sum_{i=۱}^{m} \frac{B_{۲i}}{(۲i)!} h^{۲i}[f^{(۲i-۱)}(b) - f^{(۲i-۱)}(a)] \\ &+ \frac{h^{۲m+۲}}{(۲m+۲)!}\int_a^b \overline{B}_{۲m+۲}\left(\frac{x-a}{h}\right) f^{(۲m+۲)}(x)dx \qquad (۹.۴.۵) \end{aligned}$$

توجه: پریم دوگانه روی نماد مجموعیابی بدین معناست که اولین و آخرین جمله قبل از جمع شدن باید نصف شوند.

برهان یک برهان کامل در رلستن[1] (۱۹۶۵، صص ۱۳۱ ـ ۱۳۳) داده شده و یک مطالعهٔ کلیتر در (بخش ۲) لاینس[2] و پیوری[3] (۱۹۷۳) آمده است. برهان رلستن کوتاه و دقیق است و از ویژگیهای خاص چندجمله‌ییهای برنولی کمال استفاده را می‌نماید. ما برهان ساده‌تر ولی با کلیت کمتری را با نشان دادن اینکه اثبات بر پایهٔ انتگرالگیری جزء به جزء و کمی عملیات جبری استادانه مبتنی است ارائه می‌کنیم.

برای اثبات (۹.۴.۵) در حالت کلی $n > ۱$، ابتدا اثبات را برای $n = ۱$ انجام می‌دهیم. پس

1. Ralston 2. Lyness 3. Puri

توجه خود را بر رابطهٔ زیر متمرکز می‌نماییم

$$E_1(f) = \int_0^h f(x)dx - \frac{h}{2}[f(0) + f(h)]$$
$$= \frac{1}{2}\int_0^h f''(x)x(x-h)dx \qquad (10.4.5)$$

فرمول اخیر از (21.1.5) به‌دست می‌آید. چون فرمول مجانبی زیر را می‌دانیم

$$E_1(f) \doteq -\frac{h^2}{12}[f'(h) - f'(0)]$$

کوشش می‌کنیم با عملیاتی روی (10.4.5) این نتیجه را به‌دست آوریم. می‌نویسیم

$$E_1(f) = \int_0^h f''(x)\left[-\frac{h^2}{12}\right]dx + \int_0^h f''(x)\left[\frac{x(x-h)}{2} + \frac{h^2}{12}\right]dx$$

پس

$$E_1(f) = -\frac{h^2}{12}[f'(h) - f'(0)] + \int_0^h f''(x)\left[\frac{x^2}{2} - \frac{xh}{2} + \frac{h^2}{12}\right]dx$$

با استفاده از انتگرالگیری جزء به جزء

$$E_1(f) = -\frac{h^2}{12}[f'(h) - f'(0)] + \left[f''(x)\left(\frac{x^3}{6} - \frac{x^2h}{4} + \frac{h^2x}{12}\right)\right]_0^h$$
$$- \int_0^h f^{(3)}(x)\left[\frac{x^3}{6} - \frac{x^2h}{4} + \frac{h^2x}{12}\right]dx$$

محاسبه کمیت داخل کروشه‌ها، در $x = 0$ و $x = h$، صفر به ما می‌دهد. با انتگرالگیری جزء به جزء مجدد، قسمت خارج انتگرال باز هم صفر می‌شود. در نتیجه

$$E_1(f) = -\frac{h^2}{12}[f'(h) - f'(0)] + \frac{1}{24}\int_0^h f^{(4)}(x)x^2(x-h)^2dx \qquad (11.4.5)$$

که همان (9.4.5) با $m = 1$ است. برای به‌دست آوردن حالت $m = 2$، ابتدا توجه می‌کنیم که

$$\frac{1}{24}\int_0^h x^2(x-h)^2dx = \frac{h^5}{720}$$

سپس، مانند قبل می‌نویسیم

$$\frac{1}{24}\int_0^h f^{(4)}(x)x^2(x-h)^2dx = \int_0^h f^{(4)}(x)\left[\frac{h^4}{720}\right]dx$$
$$+\int_0^h f^{(4)}(x)\left[\frac{x^2(x-h)^2}{24}-\frac{h^4}{720}\right]dx$$
$$=\frac{h^4}{720}[f^{(3)}(h)-f^{(3)}(0)]$$
$$+\int_0^h f^{(4)}(x)\left[\frac{x^4-2x^3h+x^2h^2}{24}-\frac{h^4}{720}\right]dx$$

اگر دوبار انتگرالگیری جزء به جزء را انجام دهیم، فرمول (۹.۴.۵) را برای حالت $m=2$ به‌دست می‌آوریم. این عمل را می‌توان به‌طور نامتناهی ادامه داد. در برهانی که رلستن ارائه داده از انتگرالگیری جزءبه‌جزء و بهره‌گیری از روابط خاص چندجمله‌ییهای برنولی استفاده شده است (مسألهٔ ۲۳ را ببینید).

برای اثبات حالت کلی $n>1$، می‌نویسیم

$$E_n(f)=\sum_{j=1}^{n}\left\{\int_{x_{j-1}}^{x_j} f(x)dx-\frac{h}{2}[f(x_{j-1})+f(x_j)]\right\}$$

برای حالت $m=1$، با به‌کار بردن (۱۱.۴.۵)،

$$E_n(f)=\sum_{j=1}^{n}\left\{-\frac{h^2}{12}[f'(x_j)-f'(x_{j-1})]\right\}$$
$$+\sum_{j=1}^{n}\frac{1}{24}\int_{x_{j-1}}^{x_j} f^{(4)}(x)(x-x_{j-1})^2(x-x_j)^2dx$$
$$=-\frac{h^2}{12}[f'(b)-f'(a)]+\frac{h^4}{24}\int_a^b f^{(4)}(x)\overline{B}_4\left(\frac{x-a}{h}\right)dx \quad (۱۲.۴.۵)$$

اثبات برای $m>1$، اساساً مشابه همین است. ■

جملهٔ خطا در (۹.۴.۵) بااستفاده از قضیهٔ مقدار میانگین انتگرال ساده می‌شود. می‌توان نشان داد که

$$B_{2j}(x)>0 \qquad 0<x<1 \qquad (۱۳.۴.۵)$$

و در نتیجه جملهٔ خطا در رابطهٔ زیر صدق می‌کند

$$\int_a^b \overline{B}_{2m+2}(\frac{x-a}{h})f^{(2m+2)}(x)dx = f^{(2m+2)}(\xi)\int_a^b \overline{B}_{2m+2}(\frac{b-a}{h})dx$$
$$= nf^{(2m+2)}(\xi)\int_a^{a+h} B_{2m+2}(\frac{x-a}{n})dx$$
$$= nhf^{(2m+2)}(\xi)\int_0^1 B_{2m+2}(u)du$$
$$= -(b-a)B_{2m+2}f^{(2m+2)}(\xi)$$

برای مقداری چون $a \le \xi \le b$

بنابراین (۹.۴.۵) چنین می‌شود

$$E_n(f) = -\sum_{i=1}^{m}\frac{B_{2i}}{(2i)!}h^{2i}[f^{(2i-1)}(b) - f^{(2i-1)}(a)]$$
$$-\frac{h^{2m+2}(b-a)B_{2m+2}}{(2m+2)!}f^{(2m+2)}(\xi) \qquad (۱۴.۴.۵)$$

برای مقداری چون $a \le \xi \le b$.

به‌عنوان یک فرع مهم (۹.۴.۵)، می‌توانیم نشان دهیم که قاعدهٔ ذوزنقه‌یی وقتی برای توابع دوره‌یی به‌کار برده شود، بسیار خوب عمل می‌کند.

فرع ۱ گیریم $f(x)$ برای $a \le x \le b$ به‌طور نامتناهی مشتقپذیر باشد و فرض می‌کنیم که تمام مشتقات مرتبهٔ فرد آن دوره‌یی و $b-a$ مضرب صحیحی از این دوره باشد. آنگاه مرتبهٔ همگرایی قاعدهٔ ذوزنقه‌یی $I_n(f)$ که برای

$$I(f) = \int_a^b f(x)dx$$

به‌کارگرفته شده است از هر توانی از h بزرگتر است. ■

برهان مستقیماً از فرضها دربارهٔ $f(x)$،

$$f^{(2j-1)}(b) = f^{(2j-1)}(a) \qquad j \ge 1 \qquad (۱۵.۴.۵)$$

نتیجتاً برای هر $m \ge 0$ با $h = (b-a)/n$، (۱۴.۴.۵) نتیجه می‌دهد که

$$I(f) - I_n(f) = -\frac{h^{2m+2}}{(2m+2)!}(b-a)B_{2m+2}f^{(2m+2)}(\xi), \quad a \le \xi \le b \quad (۱۶.۴.۵)$$

بنابراین وقتی $n \to \infty$ (و $h \to 0$) نرخ همگرایی متناسب می‌شود با h^{2m+2}. ولی چون m دلخواه بود، پس نتیجهٔ مطلوب به‌دست آمده است. ■

این نتیجه برای $f(x) = \exp(\cos(x))$ در جدول ۷.۵ نشان داده شده است. قاعدهٔ ذوزنقه‌یی اغلب بهترین قاعدهٔ انتگرالگیری برای انتگرالده‌های هموار دوره‌یی از نوعی است که در فرع قبلی مشخص شدند. برای مقایسهٔ فرمول گاوس ـ لژاندر و قاعدهٔ ذوزنقه‌یی برای خانوادهٔ یک پارامتری از توابع دوره‌یی، با رفتارهای گوناگون، دونالدسن و الیوت (۱۹۷۲، ص ۵۹۲) را ببینید. اینان توصیه می‌کنند که حتی برای انتگرالده‌های خیلی قله‌یی، قاعدهٔ ذوزنقه‌یی برتر است. در تحلیل قبلی به‌نظر می‌آمد که انتگرالگیری گاوسی برای انتگرالده‌های قله‌یی بهتر است. این نتیجه‌گیری تحلیل قبلی را اصلاح می‌کند.

فرمول مجموعیابی اویلر ـ مک لورن یکی از کاربردهای مهم (۹.۴.۵) یا (۱۴.۴.۵)، اگر چه انتگرالگیری عددی را شامل نمی‌شود، در مجموعیابی سری‌هاست.

فرع ۲ (فرمول مجموعیابی اویلر ـ مک لورن) فرض کنید $f(x)$، در $0 \le x < \infty$، برای مقداری از $m \le 0$ ، $2m+2$ بار پیوسته مشتقپذیر باشد. در این صورت برای تمام مقادیر $n \ge 1$،

$$\sum_{j=0}^{n} f(j) = \int_0^n f(x)dx + \frac{1}{2}[f(0) + f(n)] + \sum_{i=1}^{m} \frac{B_{2i}}{(2i)!}[f^{(2i-1)}(n) - f^{(2i-1)}(0)] - \frac{1}{(2m+2)!}\int_0^n \overline{B}_{2m+2}(x) f^{(2m+2)}(x)dx \qquad (۱۷.۴.۵)$$

■

برهان فقط $a = 0$ و $b = n$ را در (۹.۴.۵) بگذارید و جملات را به شکل مناسبی مرتب کنید. ■

مثال در یک مثال بعدی به مجموع

$$S = \sum_{1}^{\infty} \frac{1}{n^{3/2}} \qquad (۱۸.۴.۵)$$

نیاز داریم. اگر انتخاب واضح $f(x) = (x+1)^{-3/2}$ را در (۱۷.۴.۵) به‌کار بریم، نتایج مأیوس

کننده‌اند. ولی اگر $n \to \infty$، به‌دست می‌آوریم

$$S = \int_{۰}^{\infty} \frac{dx}{(x+۱)^{۳/۲}} + \frac{۱}{۲} - \sum_{۱}^{m} \frac{B_{۲i}}{(۲i)!} f^{(۲i-۱)}(۰)$$
$$- \frac{۱}{(۲m+۲)!} \int_{۰}^{\infty} \overline{B}_{۲m+۲}(x) f^{(۲m+۲)}(x) dx$$

و جملهٔ خطا برای هیچ انتخابی از m، کوچک نمی‌شود. ولی اگر سری را به دو قسمت تقسیم کنیم، می‌توانیم با آن به گونه‌ای خیلی دقیق برخورد کنیم.

گیریم $f(x) = (x+۱۰)^{-۳/۲}$، پس با $m = ۱$،

$$\sum_{۱۰}^{\infty} \frac{۱}{n^{۳/۲}} = \sum_{۰}^{\infty} f(j) = \int_{۰}^{\infty} \frac{dx}{(x+۱۰)^{۳/۲}} + \frac{۱}{۲(۱۰)^{۳/۲}} - \frac{(۱/۶)}{۲} \cdot \frac{(-۳/۲)}{(۱۰)^{۵/۲}} + E$$
$$E = -\frac{۱}{۲۴} \int_{۰}^{\infty} \overline{B}_{۴}(x) f^{(۴)}(x) dx$$

چون $\overline{B}_{۴}(x) \geq ۰$، $f^{(۴)}(x) > ۰$، داریم $E < ۰$. همچنین

$$۰ < -E < \frac{۱}{۲۴} \int_{۰}^{\infty} \left(\frac{۱}{۱۶}\right) f^{(۴)}(x) dx = \frac{۳۵}{(۱۰۲۴)(۱۰)^{۹/۲}} \doteq ۱٫۰۸ \times ۱۰^{-۶}$$

بنابراین

$$\sum_{۱۰}^{\infty} \frac{۱}{n^{۳/۲}} = ۰٫۶۴۸۶۶۲۲۰۵ + E$$

از مجموعیابی به‌طور مستقیم $\sum_{۱}^{۹} (۱/n^{۳/۲}) \doteq ۱٫۹۶۳۷۱۳۷۱۷$، به‌دست می‌آوریم

$$\sum_{۱}^{\infty} \frac{۱}{n^{۳/۲}} = ۲٫۶۱۲۳۷۵۹ + E \qquad ۰ < -E < ۱٫۰۸ \times ۱۰^{-۶} \qquad (۱۹.۴.۵)$$

برای اطلاعات بیشتر در مورد تکنیکهای مجموعیابی، به رلستن (۱۹۶۵، صص ۱۳۴ـ۱۳۸) مراجعه کنید. برای اینکه به اهمیت روش مجموعیابی بالا پی ببرید، برای به‌دست آوردن دقت مشابه برای S لازم است تعداد $۳٫۴۳ \times ۱۰^{۱۲}$ جمله به S علاوه کنید.

تعمیم فرمول اویلر ـ مک لورن برای انتگرالهایی که در آنها انتگرالده در بعضی نقاط مشتقپذیر نباشد، باز هم می‌توان یک بسط خطای مجانبی پیدا کرد. برای قاعدهٔ ذوزنقه‌یی و قاعده‌های دیگر انتگرالگیری عددی که برای انتگرالده‌های با نقاط تکین جبری و [یا] لگاریتمی به‌کار رفته‌اند، مقالهٔ

لاینس و نینهام[1] (۱۹۷۶) را ببینید. در پاراگرافهای بعد نتایج آنها را برای انتگرال

$$I = \int_{0}^{1} x^{\alpha} f(x)dx \qquad \alpha > 0 \qquad (۲۰.۴.۵)$$

با $f \in C^{m+1}[0,1]$ و بااستفاده از قاعدهٔ انتگرالگیری عددی ذوزنقه‌یی مشخص می‌کنیم.
فرض کنید α عدد صحیح نباشد،

$$E_n(f) = \sum_{j=0}^{m-1} \frac{c_j}{n^{\alpha+j+1}} + \sum_{j=1}^{m-1} \frac{d_j}{n^{j+1}} + O\left(\frac{1}{n^{m+1}}\right) \qquad (۲۱.۴.۵)$$

جملهٔ $O(1/n^{m+1})$ کمیّتی است با مقداری متناسب با $1/(n^{m+1})$ یا شاید کوچکتر از آن. ثابتها در زیر داده شده‌اند،

$$c_j = \frac{2\Gamma(\alpha+j+1)\sin[(\pi/2)(\alpha+j)]\zeta(\alpha+j+1)}{(2\pi)^{\alpha+j+1} j!} f^{(j)}(0)$$

$$d_j = 0 \qquad j \text{ زوج}$$

$$d_j = (-1)^{(j-1)/2} \cdot \frac{2\zeta(j+1)}{(2\pi)^{j+1}} g^{(j)}(1) \qquad j \text{ فرد}$$

با $g(x) = x^{\alpha} f(x)$، $\Gamma(x)$ تابع گاما و $\zeta(p)$ تابع زتا،

$$\zeta(p) = \sum_{j=1}^{\infty} \frac{1}{j^p} \qquad p > 1 \qquad (۲۲.۴.۵)$$

برای $0 < \alpha < 1$ با $m = 1$، برآورد خطای مجانبی را به‌دست می‌آوریم

$$E_n(f) = \frac{2\Gamma(\alpha+1)\sin[(\pi/2)\alpha]\zeta(\alpha+1)f(0)}{(2\pi)^{\alpha+1} n^{\alpha+1}} + O\left(\frac{1}{n^2}\right) \qquad f \in C^2[0,1] \qquad (۲۳.۴.۵)$$

برای مثال، با $I = \int_0^1 \sqrt{x} f(x)dx$ و استفاده از (۱۹.۴.۵) در محاسبهٔ $\zeta(3/2)$،

$$E_n(f) = \frac{c}{n\sqrt{n}} f(0) + O\left(\frac{1}{n^2}\right) \qquad c = \frac{\zeta(\frac{3}{2})}{4\pi} \doteq 0.208 \qquad (۲۴.۴.۵)$$

این مطلب به‌طور عددی در مثال جدول ۶.۵، در بخش ۱.۵، تأیید شده است.

1. Lyness and Ninham

برای تکینی لگاریتمی در نقاط انتهایی، ازکارهای لاینس و نینهام (۱۹۷۶) یک فرمول خطای مجانبی به شکل زیر نتیجه می‌شود

$$E_n(f) \doteq \frac{c.\log n}{n^p} \qquad (۲۵.۴.۵)$$

برای یک مقدار $p > ۰$ و یک مقدار ثابت c. برای منظورهای عددی، این مقدار اصولاً $O(۱/n^p)$ است. برای توجیه آن، حد زیر را با استفاده از قاعدهٔ هوپیتال برای $p > q$ محاسبه می‌کنیم:

$$\lim_{n\to\infty} \frac{\log(n)/n^p}{۱/n^q} = \lim_{n\to\infty} \frac{\log(n)}{n^{p-q}} = ۰$$

این بدان معناست که $\log(n)/n^p$ برای هر $q < p$ سریعتر از $۱/n^q$ کاهش می‌یابد. و روشن است که با سرعتی کمی کمتر از $۱/n^p$ کاهش می‌یابد،

برای محاسبات عملی، (۲۵.۴.۵) اصولاً $O(۱/n^p)$ است. برای مثال، حد خطاهای پی‌درپی زیر را محاسبه می‌کنیم.

$$\lim_{n\to\infty} \frac{I - I_n}{I - I_{۲n}} = \lim_{n\to\infty} \frac{c.\log(n)/n^p}{c.\log(۲n)/۲^p n^p} = \lim_{n\to\infty} ۲^p . \frac{\log(n)}{\log(۲n)}$$
$$= \lim_{n\to\infty} ۲^p . \frac{۱}{۱ + (\log ۲/\log n)} = ۲^P$$

این همان نسبت حدّی است که پیدا می‌شود وقتی که خطا درست $O(۱/n^p)$ است.

برونیابی ایتکین[1] با انگیزه‌ای از مطالب قبل، فرض می‌کنیم که فرمول انتگرالگیری دارای یک فرمول خطای مجانبی

$$I - I_n \doteq \frac{c}{n^p} \qquad p > ۰ \qquad (۲۶.۴.۵)$$

باشد. این رابطه همیشه برقرار نیست. برای مثال، انتگرالگیری گاوسی معمولاً در (۲۶.۴.۵) صدق نمی‌کند و قاعدهٔ ذوزنقه‌یی وقتی برای انتگرالده‌های دوره‌یی به‌کار رفته باشد در آن صدق نخواهد کرد. معذالک، بسیاری از قاعده‌های انتگرالگیری عددی، برای تعداد وسیعی از انتگرالده‌ها در آن صدق می‌کنند. با استفاده از این شکل مفروض برای خطا، سعی می‌کنیم که خطا را براورد نماییم. مشابه این کار، کار برونیابی ایتکین در بخش ۶.۲ است.

1. Aitkin

ابتدا p را براورد می‌نماییم. با استفاده از (۲۶.۴.۵)،

$$\frac{I_{۲n}-I_n}{I_{۴n}-I_{۲n}}=\frac{(I-I_n)-(I-I_{۲n})}{(I-I_{۲n})-(I-I_{۴n})}\doteq\frac{(c/n^p)-(c/۲^p n^p)}{(c/۲^p n^p)-(c/۴^p n^p)}=۲^p$$

$$R_{۴n}\equiv\frac{I_{۲n}-I_n}{I_{۴n}-I_{۲n}}\doteq۲^p \qquad (۲۷.۴.۵)$$

این روش یک راه ساده برای محاسبهٔ p به‌دست می‌دهد.

مثال استفاده از قاعدهٔ سیمپسون را برای $\int_۰^۱ x\sqrt{x}dx=$ ۰٫۴ درنظر می‌گیریم. در جدول ۱۵.۵، ستون R_n باید به $۲^{۲٫۵}\doteq$۵٫۶۶ میل کند، که یک نتیجهٔ نظری از لاینس و نینهام (۱۹۶۷) برای مرتبهٔ همگرایی است. روشن است که نتایج عددی نظریه را تأیید می‌کند.

برای براورد انتگرال I با دقت اضافی، گیریم، I_n، $I_{۲n}$ و $I_{۴n}$ محاسبه شده‌اند. با استفاده از (۲۶.۴.۵)

$$\frac{I-I_n}{I-I_{۲n}}\doteq۲^p\doteq\frac{I-I_{۲n}}{I-I_{۴n}}$$

و بنابراین

$$(I-I_n)(I-I_{۴n})\doteq(I-I_{۲n})^۲$$

جدول ۱۵.۵ خطاهای انتگرالگیری سیمپسون برای $\int_۰^۱ x\sqrt{x}dx$

n	I_n	$I-I_n$	$I_n-I_{n/۲}$	R_n
۲	۰٫۴۰۲۳۶۸۹۲۷۰۶۲	−۲٫۳۶۹ − ۳		
۴	۰٫۴۰۰۴۳۱۹۱۶۰۴۵	−۴٫۳۱۹ − ۴	−۱٫۹۳۷ − ۳	
۸	۰٫۴۰۰۰۷۷۲۴۹۴۴۷	−۷٫۷۲۵ − ۵	−۳٫۵۴۷ − ۴	۵٫۴۶
۱۶	۰٫۴۰۰۰۱۳۷۱۳۴۶۹	−۱٫۳۷۱ − ۵	−۶٫۳۵۴ − ۵	۵٫۵۸
۳۲	۰٫۴۰۰۰۰۲۴۲۷۸۴۶	−۲٫۴۲۸ − ۶	−۱٫۱۲۹ − ۵	۵٫۶۳
۶۴	۰٫۴۰۰۰۰۰۴۲۹۴۱۳	−۴٫۲۹۴ − ۷	−۱٫۹۹۸ − ۶	۵٫۶۵
۱۲۸	۰٫۴۰۰۰۰۰۰۷۵۹۲۴	−۷٫۵۹۲ − ۸	−۳٫۵۳۵ − ۷	۵٫۶۵
۲۵۶	۰٫۴۰۰۰۰۰۰۱۳۴۲۳	−۱٫۳۴۲ − ۸	−۶٫۲۵۰ − ۸	۵٫۶۶
۵۱۲	۰٫۴۰۰۰۰۰۰۰۲۳۷۳	−۲٫۳۷۳ − ۹	−۱٫۱۰۵ − ۸	۵٫۶۶

جدول ۱۶.۵ جدول تفاضلی برای انتگرالگیری سیمپسون

m	I_m	$\Delta I_m = I_{۲m} - I_m$	$\Delta^۲ I_m$
۱۶	۰٫۴۰۰۰۱۳۷۱۳۴۶۹		
		−۱٫۱۲۸۵۶۲۳E − ۵	
۳۲	۰٫۴۰۰۰۰۲۴۲۷۸۴۶		۹٫۲۸۷۱۹E − ۶
		−۱٫۹۹۸۴۳۳E − ۶	
۶۴	۰٫۴۰۰۰۰۰۴۲۹۴۱۳		

حال آن را نسبت به I حل می‌کنیم و عملیات جبری انجام می‌دهیم تا شکل مطلوبی به‌دست آید

$$I \doteq \tilde{I}_{۴n} \equiv I_{۴n} - \frac{(I_{۴n} - I_{۲n})^۲}{(I_{۴n} - I_{۲n}) - (I_{۲n} - I_n)} \qquad (۲۸.۴.۵)$$

مثال با استفاده از مسألهٔ قبل برای $f(x) = x\sqrt{x}$ و جدول ۱۵.۵، جدول تفاضلی را در جدول ۱۶.۵ به‌دست می‌آوریم. در این صورت

$$I \doteq \tilde{I}_{۶۴} = ۰٫۳۹۹۹۹۹۹۹۹۳۸۷$$

$$I - \tilde{I}_{۶۴} = ۶٫۱۳ \times ۱۰^{-۱۰} \qquad I - I_{۶۴} = -۴٫۲۹ \times ۱۰^{-۷}$$

بنابراین $\tilde{I}_{۶۴}$ یک بهبود قابل توجه در $I_{۶۴}$ است. ضمناً توجه کنید که $I_{۶۴} - \tilde{I}_{۶۴}$ یک تقریب عالی برای $I - I_{۶۴}$ است.

جمع‌بندی کنیم؛ یک قاعدهٔ انتگرالگیری عددی که در (۲۶.۴.۵) صدق می‌کند و سه مقدار I_n، $I_{۲n}$ و $I_{۴n}$ داده شده‌اند، $\tilde{I}_{۴n}$، برونیابی ایتکین (۲۸.۴.۵) را به‌دست می‌آوریم. این معمولاً یک بهبود کلی در $I_{۴n}$ به‌عنوان تقریب I خواهد بود؛ و بر این اساس

$$I - I_{۴n} \doteq \tilde{I}_{۴n} - I_{۴n} \qquad (۲۹.۴.۵)$$

با قاعده سیمپسون، یا هر فرمول بستهٔ مرکب دیگر نیوتن ـ کوتس، هزینهٔ محاسبهٔ I_n، $I_{۲n}$ و $I_{۴n}$ بیشتر از محاسبهٔ $I_{۴n}$ به تنهایی، یعنی $۴n + ۱$ محاسبهٔ تابع، نخواهد بود. و هرگاه الگوریتم به‌صورت صحیحی طراحی شده باشد، نیازی به ذخیرهٔ موقت برای تعداد زیادی مقدار تابع $f(x_j)$ نیست. به این دلیل هرگز نباید قاعدهٔ سیمپسون را فقط برای یک مقدار اندیس n به‌کار برد. بدون هیچگونه وقت اضافی، و بایک الگوریتم کمی پیچیده‌تر، می‌توان یک برونیابی ایتکین و یک براورد خطا پیدا کرد.

برونیابی ریچاردسن[1] اگر فرض کنیم انتگرالده $f(x)$ در انتگرال $I(f)$ به‌قدرکافی هموار است، آنگاه می‌توانیم عبارت خطای (۹.۴.۵) را به شکل زیر بنویسیم

$$I - I_n = \frac{d_2^{(0)}}{n^2} + \frac{d_4^{(0)}}{n^4} + \cdots + \frac{d_{2m}^{(0)}}{n^{2m}} + F_{n,m} \qquad (۳۰.۴.۵)$$

که I_n معرّف قاعدهٔ ذوزنقه‌یی است و

$$F_{n,m} = \frac{(b-a)^{2m+2}}{(2m+2)!n^{2m+2}} \int_a^b \overline{B}_{2m+2}\left(\frac{x-a}{h}\right) f^{(2m+2)}(x)dx$$

$$d_{2j}^{(0)} = -\frac{B_{2j}}{(2j)!}(b-a)^{2j}[f^{(2j-1)}(b) - f^{(2j-1)}(a)] \qquad (۳۱.۴.۵)$$

اگر چه سریهایی که با آنها کار می‌کنیم همیشه متناهی‌اند و یک جملهٔ خطا دارند، معمولاً به جملهٔ خطا توجه نمی‌کنیم

وقتی که n زوج است

$$I - I_{n/2} = \frac{4d_2^{(0)}}{n^2} + \frac{16d_4^{(0)}}{n^4} + \frac{64d_6^{(0)}}{n^6} + \ldots \qquad (۳۲.۴.۵)$$

(۳۰.۴.۵) را در ۴ ضرب و (۳۲.۴.۵) را از آن کم می‌کنیم:

$$4(I - I_n) - (I - I_{n/2}) = \frac{-12d_4^{(0)}}{n^4} - \frac{60d_6^{(0)}}{n^6} - \ldots$$

$$I = \frac{4I_n - I_{n/2}}{3} - \frac{4d_4^{(0)}}{n^4} - \frac{20d_6^{(0)}}{n^6} - \ldots$$

تعریف می‌کنیم

$$I_n^{(1)} = \frac{1}{3}[4I_n^{(0)} - I_{n/2}^{(0)}] \qquad n \text{زوج} \qquad n \geq 2 \qquad (۳۳.۴.۵)$$

و $I_m^{(0)} \equiv I_m$. ما $\{I_n^{(1)}\}$ را برونیاب ریچاردسن برای $\{I_n^{(0)}\}$ می‌نامیم.

دنبالهٔ

$$I_2^{(1)}, I_4^{(1)}, I_6^{(1)}, \ldots$$

یک قاعدهٔ انتگرالگیری عددی جدید است. برای خطا،

$$I - I_n^{(1)} = \frac{d_4^{(1)}}{n^4} + \frac{d_6^{(1)}}{n^6} + \cdots \qquad (۳۴.۴.۵)$$

$$d_4^{(1)} = -4d_4^{(0)}, d_6^{(1)} = -20d_6^{(0)}, \ldots \qquad (۳۵.۴.۵)$$

1. Richardson

برای آنکه فرمول صریحی برای $I_n^{(1)}$ پیدا کنیم، می‌گیریم $h=(b-a)/n$ و $x_j=a+jh$، $j=0,1,\dots,n$. سپس با استفاده از (۳۳.۴.۵) و تعریف قاعدهٔ ذوزنقه‌یی

$$I_n^{(1)}=\frac{4h}{3}\left[\frac{1}{2}f_0+f_1+\cdots+f_{n-1}+\frac{1}{2}f_n\right]$$
$$-\frac{2h}{3}\left[\frac{1}{2}f_0+f_2+f_4+\cdots+f_{n-2}+\frac{1}{2}f_n\right]$$

$$I_n^{(1)}=\frac{h}{3}[f_0+4f_1+2f_2+4f_3+\cdots+2f_{n-2}+4f_{n-1}+f_n] \qquad (۳۶.۴.۵)$$

که قاعدهٔ سیمپسون با n تقسیم جزیی است. برای خطا، با استفاده از (۳۵.۴.۵) و (۳۱.۴.۵)،

$$I-I_n^{(1)}=-\frac{h^4}{180}[f^{(3)}(b)-f^{(3)}(a)]+$$
$$\frac{h^6}{1512}[f^{(5)}(b)-f^{(5)}(a)]+\cdots \qquad (۳۷.۴.۵)$$

معنی این عبارت این است که کار با فرمول اویلر ـ مک لورن با عملیات جبری ساده به قاعدهٔ سیمپسون تبدیل می‌شود. ما مثال عددی ذکر نمی‌کنیم، زیرا طبق (۳۶.۴.۵)، $I_n^{(1)}$ درست، همان قاعدهٔ سادهٔ سیمپسون است.

استدلال فوق را که به $I_n^{(1)}$ منتهی شد، می‌توان ادامه داد و فرمولهای جدیدی به‌دست آورد. مانند گذشته، اگر n مضربی از ۴ باشد، آنگاه

$$I-I_{n/2}^{(1)}=\frac{16d_4^{(1)}}{n^4}+\frac{64d_6^{(1)}}{n^6}+\cdots$$
$$16(I-I_n^{(1)})-(I-I_{n/2}^{(1)})=\frac{-48d_6^{(1)}}{n^6}+\cdots$$
$$I=\frac{16I_n^{(1)}-I_{n/2}^{(1)}}{15}-\frac{48d_6^{(1)}}{15n^6}+\cdots \qquad (۳۸.۴.۵)$$

در این صورت

$$I-I_n^{(2)}=\frac{d_6^{(2)}}{n^6}+\frac{d_8^{(2)}}{n^8}+\cdots \qquad (۳۹.۴.۵)$$

با

$$I_n^{(2)}=\frac{16I_n^{(1)}-I_{n/2}^{(1)}}{15} \qquad n\geq 4 \qquad (۴۰.۴.۵)$$

و n بر ۴ بخشپذیر است. $\{I_n^{(2)}\}$ را برونیاب ریچاردسن $\{I_n^{(1)}\}$ می‌نامیم. اگر وزنهای واقعی انتگرالگیری $I_n^{(2)}$ را پیدا کنیم، مشابه با (۳۶.۴.۵)، خواهیم دید که $I_n^{(2)}$ همان قاعدهٔ مرکب بول[1] است.

1. Bool

با به‌کار بردن فرمولهای قبلی، می‌توانیم، با استفاده از (۳۹.۴.۵)، براوردهای سودمندی برای خطا به‌دست آوریم.

$$I - I_n^{(1)} = \frac{16 I_n^{(1)} - I_{n/2}^{(1)}}{15} - I_n^{(1)} + \frac{d_6^{(2)}}{n^6} + \cdots$$
$$= \frac{I_n^{(1)} - I_{n/2}^{(1)}}{15} + \frac{d_6^{(2)}}{n^6} + \cdots$$

با استفاده از $h = (b-a)/n$،

$$I - I_n^{(1)} = \frac{1}{15}[I_n^{(1)} - I_{n/2}^{(1)}] + O(h^6) \qquad (41.4.5)$$

و بنابراین

$$I - I_n^{(1)} \doteq \frac{1}{15}[I_n^{(1)} - I_{n/2}^{(1)}] \qquad (42.4.5)$$

چون هر دو عبارت $O(h^4)$ هستند و جملهٔ باقیمانده $O(h^6)$ است. این را براورد **خطای ریچاردسن** برای قاعدهٔ سیمپسون می‌نامند.

این روند برونیابی را می‌توان به طریق استقرایی ادامه داد. تعریف می‌کنیم

$$I_n^{(k)} = \frac{4^k I_n^{(k-1)} - I_{n/2}^{(k-1)}}{4^k - 1} \qquad n \geq 2^k \qquad (43.4.5)$$

با n مضربی از 2^k، $k \geq 1$. می‌توان نشان داد که خطا شکل زیر را دارد

$$I - I_n^{(k)} = \frac{d_{2k+2}^{(k)}}{n^{2k+2}} + \cdots$$
$$= A_k (b-a) h^{2k+2} f^{(2k+2)}(\zeta_n) \qquad a < \zeta_n < b \qquad (44.4.5)$$

که A_k ثابتی مستقل از f و h است و

$$d_{2k+2}^{(k)} = A_k (b-a)^{2k+2} [f^{(2k+1)}(b) - f^{(2k+1)}(a)]$$

بالاخره، می‌توان نشان داد که برای هر $f \in C[a,b]$،

$$\lim_{n \to \infty} I_n^{(k)}(f) = I(f) \qquad (45.4.5)$$

قاعده‌های $I_n^{(k)}(f)$ برای $k \geq 2$، هیچ رابطهٔ مستقیمی با قاعده‌های مرکب نیوتن ـ کوتس ندارند. برای جزئیات کامل باؤئر[1] و دیگران (۱۹۶۳) را ببینید.

1. Bauer

$$
\begin{array}{ccccc}
I_1^{(0)} & & & & \\
I_2^{(0)} & I_2^{(1)} & & & \\
I_4^{(0)} & I_4^{(1)} & I_4^{(2)} & & \\
I_8^{(0)} & I_8^{(1)} & I_8^{(2)} & I_8^{(3)} & \\
I_{16}^{(0)} & I_{16}^{(1)} & I_{16}^{(2)} & I_{16}^{(3)} & I_{16}^{(4)} \\
\vdots & \vdots & \vdots & \vdots & \vdots
\end{array}
$$

شکل ۴.۵ جدول انتگرالگیری رامبرگ

انتگرالگیری رامبرگ[1] تعریف می‌کنیم:

$$J_k(f) = I_{2^k}^{(k)} \qquad k = 0, 1, 2, \ldots \qquad (۴۶.۴.۵)$$

این قاعده، قاعدهٔ انتگرالگیری رامبرگ است. نمودار شکل ۴.۵ برای برونیابهای ریچاردسن در مورد قاعدهٔ ذوزنقه‌یی را درنظر می‌گیریم، که شماره تقسیمهای جزیی توانی از ۲ است. ستون اول قاعدهٔ ذوزنقه‌یی را نشان می‌دهد، ستون دوم‌قاعدهٔ سیمپسون را، و غیره. به موجب (۴۵.۴.۵)، هر ستون به $I(f)$ همگراست. قاعدهٔ رامبرگ استفاده از قطر است. چون هر ستون از ستون قبلی سریعتر همگرا می‌شود، با فرض اینکه $f(x)$ به‌طور نامتناهی مشتقپذیر است، می‌توان انتظار داشت که $I_k(f)$ برای هر k سریعتر از $\{I_n^{(k)}\}$ همگرا شود. معمولاً چنین است و نتیجتاً از اواخر دههٔ ۱۹۵۰، این روش بسیار عمومیّت یافته است. در مقایسه با روشهای انتگرالگیری گاوسی، انتگرالگیری رامبرگ این مزیّت را دارد که از طولهایی به فاصله‌های مساوی استفاده می‌کند. برای تحلیل کاملتر انتگرالگیری رامبرگ باؤئر و دیگران (۱۹۶۳) را ببینید.

مثال با استفاده از انتگرالگیری رامبرگ، انتگرال زیر رامحاسبه کنید

$$I = \int_0^{\pi} e^x \cos(x) dx = -\frac{1}{2}(e^{\pi} + 1)$$

از این انتگرال به‌عنوان یک مثال قبلاً در جدولهای ۱.۵، ۳.۵ و ۱۱.۵ به ترتیب برای قاعده‌های ذوزنقه‌یی، سیمپسون، و گاوس ـ لژاندر استفاده شد. نتایج رامبرگ در جدول ۱۷.۵ داده شده‌اند. این اعداد نشان می‌دهند که انتگرالگیری رامبرگ نسبت به قاعدهٔ سیمپسون برتری دارد، ولی انتگرالگیری گاوسی باز همگرایی سریعتری دارد.

1. Romberg

جدول ۱۷.۵ مثال انتگرالگیری رامبرگ

k	گره‌ها	$J_k(f)$	خطا
۰	۲	$-۳۴٫۷۷۸۵۱۸۶۶۰۲۶$	$۲٫۲۷\mathrm{E}+۱$
۱	۳	$-۱۱٫۵۹۲۸۳۹۵۵۳۴۲$	$-۴٫۷۸\mathrm{E}-۱$
۲	۵	$-۱۲٫۰۱۱۰۸۴۳۱۷۵۴$	$-۵٫۹۳\mathrm{E}-۲$
۳	۹	$-۱۲٫۰۷۰۴۲۰۴۱۲۸۷$	$۷٫۴۱\mathrm{E}-۵$
۴	۱۷	$-۱۲٫۰۷۰۳۴۷۲۰۸۷۳$	$۸٫۹۲\mathrm{E}-۷$
۵	۳۳	$-۱۲٫۰۷۰۳۴۶۳۱۶۳۲$	$-۶٫۸۲\mathrm{E}-۱۱$
۶	۶۵	$-۱۲٫۰۷۰۳۴۶۳۱۶۳۹$	$<۵٫۰۰\mathrm{E}-۱۲$

برای محاسبهٔ $J_k(f)$ برای یک مقدار خاص k، وقتی $J_{k-۱}(f),\ldots,J_۱(f)$ قبلاً محاسبه شده باشند، سطر

$$I_n^{(۰)}, I_n^{(۱)}, \ldots, I_n^{(k-۱)} \qquad n = ۲^{k-۱} \qquad (۴۷.۴.۵)$$

بایستی در یک حافظهٔ موقتی نگهداری شود. سپس $I_{۲k}^{(۰)}$، از $I_{۲k-۱}^{(۰)}$ و $۲^{k-۱}$ مقدار جدید تابع، محاسبه خواهد شد. با استفاده از (۳۳.۴.۵)، (۴۰.۴.۵) و (۴۳.۴.۵) سطر بعدی که در جدول جملهٔ $J_k(f)$است، محاسبه خواهد شد. $J_k(f)$ با $J_{k-۱}(f)$ باید مقایسه شود تا ملاحظه شود آیا دقت کافی وجود دارد تا $J_k(f)$ به‌عنوان یک تقریب دقیق برای $I(f)$ پذیرفته شود، یا نه.

ما این شیوه را به یک طریق صوری در الگوریتم زیر می‌آوریم. این الگوریتم به دلایل آموزشی گنجانده شده است و نبایستی به‌عنوان یک برنامهٔ جدی تلقی شود، مگر آنکه بعضی اصلاحات در آن صورت گیرد. برای مثال، آزمون خطا خیلی مقدماتی و قدیمی است و یک بررسی مطمئنی برای انتگرالهای عددی متناظر با مقادیر کوچک k، وقتی مقادیر تابع به قدر کافی نمونه‌گیری نشده‌است. بایستی ضمیمه شود،

الگوریتم Romberg (f,a,b,ϵ,int)

۱. تبصره: انتگرالگیری رامبرگ را برای محاسبهٔ int به‌عنوان براوردی از انتگرال

$$I = \int_a^b f(x)dx$$

به‌کار برید. عمل را متوقّف کنید وقتی که $|I - \mathrm{int}| \leq \epsilon$.

۲. مقادیر اولیه:

$$k := ۰, n := ۱,$$

$$T_۰ := R_۰ := \alpha_۰ := (b-a)[f(a)+f(b)]/۲$$

۳. حلقهٔ اصلی را شروع کنید: $n := ۲n\ ,\ k := k+۱\ ,\ h = (b-a)/n$

۴. $\text{sum} := \sum_{j=۱}^{n/۲} f(a+(۲j-۱)h)$

۵. $T_k := h.\text{sum} + \frac{۱}{۲}T_{k-۱}$

۶. $\beta_j := \alpha_j \qquad j = ۰, ۱, \ldots, k-۱$

۷. $\alpha_۰ := T_k \qquad m := ۱$

۸. تا مرحلهٔ ۱۰ برای $j = ۱, ۲, \ldots k$ عمل را انجام دهید.

۹. $m := ۴m$

۱۰. $\alpha_j := \frac{m.\alpha_{j-۱} - \beta_{j-۱}}{m-۱}$

۱۱. $R_k := \alpha_k$

۱۲. اگر $|R_k - R_{k-۱}| > \epsilon$، به مرحلهٔ ۳ بروید.

۱۳. چون $|R_k - R_{k-۱}| \leq \epsilon$، $\text{int} = R_{k-۱}$ را بپذیرید و الگوریتم را خاتمه دهید.

شکلهای متفاوتی از انتگرالگیری رامبرگ موجود است. برای مثال، راههای دیگر افزایش تعداد گرهها مطالعه شده است. برای یک بررسی خیلی کامل از نوشته‌هایی دربارهٔ انتگرالگیری رامبرگ، دیویس[1] و رابینوویتس[2] (۱۹۸۴، صص ۴۳۴ـ۴۴۶) را ببینید. یک برنامهٔ فورترن برای یک انتگرالگیری رامبرگ نیز در آنجا داده شده است.

۵.۵ انتگرالگیری عددی خودکار

در برنامهٔ انتگرالگیری عددی خودکار، مقدار تقریبی انتگرال با دقتی که کاربر مشخص می‌کند محاسبه می‌شود. کاربر نیازی به تعیین روش یا تعداد نقاط گرهی ندارد. برنامه‌های انتگرالگیری خودکار بسیار خوبی وجود دارند که بسیاری افراد از آنها استفاده می‌نمایند. با چنین برنامه‌ای، وقت کاربر برای برنامه‌نویسی صرف نخواهد شد، و نیازی به آشنایی با نظریهٔ انتگرالگیری عددی نخواهد داشت. معذالک، تقریباً همیشه ممکن است برنامهٔ خودکار را بهتر نمود، گرچه این امر به

1. Davis 2. Rabinowitz

شناخت کافی از انتگرالگیری عددی که برای مسأله به‌کار رفته، نیاز دارد. وقتی فقط به دفعات کمی انتگرالگیری عددی انجام می‌شود، انتگرالگیری خودکار اغلب راه خوبی برای صرفه‌جویی در وقت است. ولی برای مسائلی که مستلزم انتگرالگیری‌های زیاد است، شاید بهتر باشد که با صرف وقت روند انتگرالگیری عددی کم‌خرجتری پیدا کنیم.

یک برنامهٔ انتگرالگیری عددی خودکار، مانند یک «جعبهٔ سیاه» عمل می‌کند، بدون اینکه کاربر قادر باشد مراحل میانی محاسبات را ببیند. به همین علّت، مهمترین ویژگی لزوم چنین برنامه‌ای این است که قابل اعتماد باشد: مقدار تقریبی انتگرال که توسط برنامه به‌دست می‌آید باید دقت ادعا شده را داشته باشد. از جنبهٔ نظری، همان‌گونه که در پاراگراف بعد توضیح داده می‌شود، چنین الگوریتمی وجود ندارد. ولی برای آن نوع انتگرالده‌هایی که معمولاً در عمل در نظر گرفته می‌شوند، برنامه‌هایی هستند که تا درجهٔ بالایی قابل اعتمادند. این قابلیّت اعتماد می‌تواند بهبود یابد اگر کاربر شرح برنامه را بخواند و به محدودیتها و فرضهای برنامه توجّه کند.

برای پی‌بردن به عدم امکان برنامهٔ انتگرالگیریِ خودکارِ کاملاً اعتمادپذیر از لحاظ نظری، باید توجه داشت که این برنامه مقادیر انتگرالده $f(x)$ را در تعداد متناهی نقطه چون، $x_1, \ldots, x_n$ محاسبه می‌کند. در این صورت تعداد نامتناهی تابع پیوستهٔ $\hat{f}(x)$ موجودند که برای آنها

$$\hat{f}(x_i) = f(x_i) \qquad i = 1, \ldots n$$

و

$$\int_a^b \hat{f}(x)\, dx \neq \int_a^b f(x)dx$$

در واقع تعداد نامتناهی از این گونه توابع $\hat{f}(x)$ وجود دارند که به‌طور نامتناهی مشتقپذیرند. در مسائل عملی، نامحتمل می‌نماید که یک برنامهٔ انتگرالگیری عددی خوش‌ساخت، اعتمادپذیر نباشد، ولی این شدنی است. اعتمادپذیری برنامهٔ انتگرالگیری خودکار را می‌توان با سخت‌ترکردن آزمونهای خطا بالا برد، ولی این نیز باعث از دست رفتن کارایی برنامه می‌شود. معمولاً بین اعتمادپذیری و کارایی یک ارتباطی وجود دارد. برای بحث بیشتر دربارهٔ مسائل اعتمادپذیری و کارایی برنامه‌های انتگرالگیری خودکار لاینس[1] و کاگانوف[2] (۱۹۷۶) را ببینید.

انتگرالگیری تطبیقی برنامه‌های خودکار را می‌توان به دو دسته تقسیم کرد. (۱) آنهایی که در آنها از یک قاعدهٔ فراگیر مانند قاعدهٔ گاوس یا قاعدهٔ ذوزنقه‌یی با فاصله‌گذاری مساوی استفاده می‌کنند، و (۲) آنهایی که در آنها یک خط مشی **تطبیقی** به‌کار می‌برند، که در آن، قاعدهٔ انتگرالگیری جای

1. Lyness 2. Kaganove

نقاط گرهی و حتی تعریف آن را تغییر می‌دهد تا تغییر رفتار موضعی انتگرالده را منعکس نماید. در خط مشی فراگیر همان نوع براورد خطا که در بخش قبل بحث شد به‌کار برده می‌شود. اکنون مفهوم و طرز کار خط مشی تطبیقی را مورد بحث قرار می‌دهیم.

هموار و مشتقپذیری بسیاری از انتگرالده‌ها در نقاط مختلف بازهٔ انتگرالگیری $[a,b]$ تغییر می‌کنند. برای مثال، در

$$I=\int_{۰}^{۱}\sqrt{x}\,dx$$

انتگرالده در $x=۰$ دارای شیب بی‌نهایت است ولی در نقاط نزدیک به ۱ خوشرفتار است. در بیشتر روشهای عددی یک شبکهٔ یکنواخت از نقاط گرهی را به‌کار می‌برند، یعنی، چگالی نقاط گرهی را در سرتاسر بازهٔ انتگرالگیری تقریباً مساوی می‌گیرند. این روش شامل فرمولهای مرکب نیوتن ـ کوتس، انتگرالگیری گاوسی و انتگرالگیری رامبرگ می‌شود. هرگاه انتگرالده در نقطه‌ای چون α از بازهٔ $[a,b]$ بدرفتار باشد، نقاط گرهی بسیاری باید نزدیک α جا داده شوند تا این رفتار را تعدیل کنند. ولی این امر موجب می‌شود که نقاط گرهی بیشتر از حد نیاز، در سایر قسمتهای $[a,b]$ به‌کار روند. در انتگرالگیری تطبیقی سعی می‌شود تا نقاط گرهی را مطابق با رفتار انتگرالده جای دهند، به‌گونه‌ای که چگالی نقاط گرهی در نزدیکی نقاط بدرفتار انتگرالده بیشتر باشد.

اکنون مفهوم اساسی انتگرالگیری تطبیقی را با استفاده از شکل ساده شدهٔ **قاعدهٔ تطبیقی سیمپسون** شرح می‌دهیم. برای آنکه ببینیم چرا فاصله‌گذاری متغیر لازم است، قاعدهٔ سیمپسون را با چنین فاصله‌گذاری نقاط گرهی درنظر می‌گیریم:

$$I(f)=\sum_{j=۱}^{n/۲}\int_{x_{۲j-۲}}^{x_{۲j}}f(x)dx \doteq I_n(f)$$

$$=\sum_{j=۱}^{n/۲}\left(\frac{x_{۲j}-x_{۲j-۲}}{۶}\right)(f_{۲j-۲}+۴f_{۲j-۱}+f_{۲j}) \qquad (۱.۵.۵)$$

با $x_{۲j-۱}=(x_{۲j-۲}+x_{۲j})/۲$. با استفاده از (۱۵.۱.۵)

$$I(f)-I_n(f)=-\frac{۱}{۲۸۸۰}\sum_{j=۱}^{n/۲}(x_{۲j}-x_{۲j-۲})^۵f^{(۴)}(\xi_j) \qquad (۲.۵.۵)$$

با $x_{۲j-۲}\le\xi_j\le x_{۲j}$. واضح است که می‌خواهید $x_{۲j}-x_{۲j-۲}$ را مطابق با اندازهٔ $f^{(۴)}(\xi)$ انتخاب کنید که معمولاً نامعلوم است. اگر قدر مطلق $f^{(۴)}(x)$ بسیار متغیر باشد، نمی‌خواهید نقاط گرهی متساوی‌الفاصله باشند.

نمادگذاری، زیر را وارد می‌کنیم

$$I_{\alpha,\beta} = \int_{\alpha}^{\beta} f(x)dx$$

$$I^{(1)}_{\alpha,\beta} = \frac{h}{3}\left[f(\alpha)+4f\left(\frac{\alpha+\beta}{2}\right)+f(\beta)\right] \quad h=\frac{\beta-\alpha}{2} \qquad (3.5.5)$$

$$I^{(2)}_{\alpha,\beta} = I^{(1)}_{\alpha,\gamma} + I^{(1)}_{\gamma,\beta} \quad \gamma=\frac{\beta+\alpha}{2}$$

برای توضیح محاسبهٔ

$$I = \int_{a}^{b} f(x)dx$$

با الگوریتم تطبیقی، از یک تعریف بازگشتی استفاده می‌کنیم. فرض می‌کنیم $\varepsilon > 0$ داده شده است و می‌خواهیم یک انتگرال تقریبی $\tilde{I}$ پیدا کنیم که

$$| I - \tilde{I} | \leq \varepsilon \qquad (4.5.5)$$

با $\alpha = a$ و $\beta = b$ آغاز و $I^{(1)}_{\alpha,\beta}$ و $I^{(2)}_{\alpha,\beta}$ را حساب می‌کنیم. اگر

$$| I^{(2)}_{\alpha,\beta} - I^{(1)}_{\alpha,\beta} | \leq \varepsilon \qquad (5.5.5)$$

آنگاه $I^{(2)}_{\alpha,\beta}$ را به‌عنوان تقریب تطبیقی انتگرال $I_{\alpha,\beta}$ می‌پذیریم. در غیر این صورت می‌نویسیم $\varepsilon := \varepsilon/2$ و انتگرال تطبیقی برای $I_{\alpha,\beta}$ را برابر مجموع انتگرالهای تطبیقی برای $I_{\alpha,\gamma}$ و $I_{\gamma,\beta}$ قرار می‌دهیم که $\gamma = \alpha + \beta/2$، و هر کدام از انتگرالها باید با یک تحمل خطای ε محاسبه شوند.

در یک کاربرد عملی مانند کاربرد در یک برنامهٔ رایانه‌یی، محدودیتهای خیلی بیشتری به‌عنوان وسایل ایمنی، ضمیمه می‌شوند و برآورد خطا معمولاً بسیار پیچیده‌تراست. و در همهٔ محاسبه‌های تابع، دقّت می‌شود که اطمینان حاصل شود که انتگرالده در یک نقطه دوبار محاسبه نشود. این امر نیاز به یک شیوهٔ انباشتن ماهرانهٔ مقادیر $f(x)$ دارد که باید موقتاً نگهداری شوند چون برای محاسبات بعدی مورد احتیاج‌اند. تغییرات بسیار جزئی زیادی می‌توان به‌کار برد تا اجرای برنامه بهبود یابد، ولی معمولاً در مرحله اول مقدار زیادی تجربه و بررسی عملی لازم است. بدین دلیل و دلایل دیگر، توصیه شده است از شیوه‌های تطبیقی آزمایش شده استفاده شود. [مثلاً، دبور (۱۹۷۱) پیسنز و همکاران (۱۹۸۳)را ببینید.]. این موضوع در آخر بخش بیشتر توضیح داده شده است.

جدول ۱۸.۵ مثال تطبیقی (۶.۵.۵) سیمپسون

$[\alpha,\beta]$	$I^{(2)}$	$I-I^{(2)}$	$I-I^{(1)}$	$\mid I^{(2)}-I^{(1)} \mid$	ε
[۰٫۰, ۰٫۰۶۲۵]	۰٫۰۱۰۲۵۸	۱٫۶E − ۴	۴٫۵E − ۴	۲٫۹E − ۴	۰٫۰۰۰۳۱۲۵
[۰٫۰۶۲۵, ۰٫۰۱۲۵]	۰٫۰۱۹۰۴۶	۱٫۲E − ۷	۱٫۱E − ۶	۱٫۰E − ۶	۰٫۰۰۰۳۱۲۵
[۰٫۱۲۵, ۰٫۲۵]	۰٫۰۵۳۸۷۱	۴٫۵E − ۷	۳٫۶E − ۶	۴٫۰E − ۶	۰٫۰۰۰۶۲۵
[۰٫۲۵, ۰٫۵]	۰٫۱۵۲۳۶۸	۹٫۳E − ۷	۱٫۱E − ۵	۱٫۰E − ۵	۰٫۰۰۱۲۵
[۰٫۵, ۱٫۰]	۰٫۴۳۰۹۶۲	۲٫۴E − ۶	۳٫۰E − ۵	۲٫۸E − ۵	۰٫۰۰۲۵

مثال استفاده از شیوهٔ تطبیقی سادهٔ قبلی سیمپسون را برای محاسبهٔ انتگرال

$$I=\int_{۰}^{۱}\sqrt{x}\,dx \qquad (۶.۵.۵)$$

با $\varepsilon = ۰٫۰۰۵$ بر بازهٔ [۰ , ۱] درنظر می‌گیریم. بازه‌های نهایی $[\alpha,\beta]$ و انتگرالهای $I^{(2)}_{\alpha,\beta}$ در جدول ۱۸.۵ داده شده‌اند. ستونی که با ε مشخص شده، تحمل خطا را که در آزمون (۵.۵.۵) برای برآورد خطا در $I^{(1)}_{\alpha,\beta}$ به‌کار رفته نشان می‌دهد، خطایی که برای $I^{(1)}_{\alpha,\beta}$ برآورده شده بر بازهٔ [۰ , ۰٫۰۶۲۵] نادقیق ولی برای سایر زیربازه‌ها دقیق است. مقداری که برای برآورد $I_{\alpha,\beta}$ به‌کار برده شده عملاً برابر $I^{(2)}_{\alpha,\beta}$ و در همهٔ زیربازه‌ها به قدرکافی دقیق است. انتگرال کل که از مجموع تمام $I^{(1)}_{\alpha,\beta}$ها به‌دست می‌آید، عبارت است از

$$\tilde{I}=۰٫۶۶۶۵۰۵ \qquad I-\tilde{I}=۱٫۶E-۴$$

و کران محاسبه شده چنین است

$$\mid I-\tilde{I}\mid \leq ۳٫۳E-۴$$

که از مجموع مقادیر ستونی که با $\mid I^{(2)}-I^{(1)}\mid$ نشان داده‌شده به‌دست آمده است. توجه دارید که خطا بیشتر در اولین زیرفاصله متمرکز است، همان‌گونه که از رفتار انتگرالده در نزدیکی $x=۰$ می‌توانستیم پیش‌بینی نماییم. برای یک مثال که آزمون (۵.۵.۵)برای آن مناسب نیست، مسألهٔ ۳۲ را ببینید.

چند برنامهٔ انتگرالگیری خودکار یکی از معروفترین برنامه‌های انتگرالگیری خودکار، برنامهٔ تطبیقی CADRE[1] است که در دبور (۱۹۷۱) آمده و شامل وسیلهٔ شناسایی تکینیهای جبری در نقاط انتهایی بازهٔ انتگرالگیری است. فرمولهای خطای مجانبی لاینس و نیهام (۱۹۶۷) که برای یک حالت خاص در (۲۱.۴.۵) داده‌شده، به‌کار گرفته شده است تا روش انتگرالگیری همگرای

1. Cautious Adaptive Romberg Extrapolation

سریعتری را باز هم بر پایهٔ برونیابی مکرّر ریچاردسون، فراهم نماید. به تجربه دریافته‌اند که برنامهٔ CADRE هم کاملاً اعتمادپذیر و هم کارآست.

یک بسته نرم‌افزار که اخیراً ساخته شده QUADPACK است، که بعضی از برنامه‌های آن برای منظورهای کلی و بعضی دیگر برای رده‌های خاصی از انتگرالها تدوین شده است. این بسته نرم‌افزار نتیجهٔ یک همکاری دسته‌جمعی است و شرح کامل آن در پیسنز و همکاران (۱۹۸۳) داده شده است. این بسته به خوبی آزمایش شده و به‌نظر می‌آید که یک مجموعهٔ عالی از برنامه‌ها باشد.

بحث قبلی را با محاسبهٔ تقریبهای عددی انتگرالهای زیر روشن می‌سازیم:

$$I_۱ = \int_۰^۱ \frac{۴dx}{۱ + ۲۵۶(x - ۰٫۳۷۵)^۲} = \frac{۱}{۴}[\tan^{-۱}(۱۰) + \tan^{-۱}(۶)]$$

$$I_۲ = \int_۰^۱ x^۲\sqrt{x}\, dx = \frac{۲}{۷}$$

$$I_۳ = \int_۰^۱ \sqrt{x}\, dx = \frac{۲}{۳}$$

$$I_۴ = \int_۰^۱ \log(x)dx = -۱$$

$$I_۵ = \int_۰^۱ \log \mid x - ۰٫۷ \mid dx = ۰٫۳\log(۰٫۳) + ۰٫۷\log(۰٫۷) - ۱$$

$$I_۶ = \int_۰^{۲\pi} e^{-x}\sin(۵۰x)dx = \frac{۵۰}{۲۵۰۱}(۱ - e^{-۲\pi})$$

$$I_۷ = \int_{-۹}^{۱۰۰۰۰} \frac{dx}{\sqrt{\mid x \mid}} = ۲۰۶$$

از QUADPACK، برنامهٔ DQAGP را انتخاب می‌کنیم. این برنامه نیز برای تشخیص تکینیهای جبری در دو سر بازه، و ترمیم حضور آنها، راههایی دارد. برای اجرای بهتر، برنامه به‌کاربر اجازه می‌دهد که نقاط داخل بازهٔ انتگرالگیری را که انتگرالده در آن نقاط تکین است، مشخص نماید.

ما هر دو برنامهٔ CADREو DQAGP را برای محاسبهٔ انتگرالهای فوق، با تحمل خطای $۱۰^{-۲}$، $۱۰^{-۵}$، $۱۰^{-۸}$ به‌کار برده‌ایم. نتایج در جداول ۱۹.۵ و ۲۰.۵ داده شده‌اند. برای مقایسهٔ منصفانه‌تر CADRE و DQAGP، CADRE1 را برای دو انتگرال در هر یک از انتگرالهای $I_۵$ و $I_۷$ به‌کار برده‌ایم تا تکینیها در نقاط انتهایی بازه‌ها قرارگیرند. برای مثال، CADRE را برای هر یک از انتگرالهای موجود در

$$I_۷ = \int_{-۹}^{۰} \frac{dx}{\sqrt{-x}} + \int_{۰}^{۱۰۰۰۰} \frac{dx}{\sqrt{x}} \qquad (۷.۵.۵)$$

جدول ۱۹.۵ مثالهای انتگرالگیری برای CADRE

	خطای مطلوب					
	10^{-2}		10^{-5}		10^{-8}	
انتگرال	خطا	N	خطا	N	خطا	N
I_1	$A = 2{,}49\text{E}-6$	76	$A = 1{,}40\text{E}-7$	73	$A = 4{,}60\text{E}-11$	225
	$P = 5{,}30\text{E}-4$		$P = 4{,}45\text{E}-6$		$P = 2{,}48\text{E}-9$	
I_2	$A = 1{,}18\text{E}-5$	9	$A = 3{,}96\text{E}-7$	17	$P = 2{,}73\text{E}-10$	129
	$P = 3{,}27\text{E}-3$		$P = 3{,}56\text{E}-6$		$P = 2{,}81\text{E}-9$	
I_3	$A = 1{,}03\text{E}-4$	17	$A = 3{,}23\text{E}-8$	33	$A = 1{,}98\text{E}-9$	65
	$P = 2{,}98\text{E}-3$		$P = 4{,}43\text{E}-8$		$P = 2{,}86\text{E}-9$	
I_4	$A = 6{,}57\text{E}-5$	105	$A = 6{,}45\text{E}-8$	209	$A = 4{,}80\text{E}-9$	281
	$P = 4{,}98\text{E}-3$		$P = 9{,}22\text{E}-6$		$P = 1{,}55\text{E}-8$	
I_5	$A = 2{,}77\text{E}-5$	226	$A = 7{,}41\text{E}-8$	418	$A = 5{,}89\text{E}-9$	562
	$P = 3{,}02\text{E}-3$		$P = 1{,}00\text{E}-5$		$P = 1{,}11\text{E}-8$	
I_6	$A = 8{,}49\text{E}-6$	955	$A = 2{,}37\text{E}-8$	1171	$A = 4{,}30\text{E}-11$	2577
	$P = 8{,}48\text{E}-3$		$P = 1{,}67\text{E}-5$		$P = 2{,}07\text{E}-8$	
I_7	$A = 4{,}54\text{E}-4$	98	$A = 7{,}72\text{E}-7$	418	$A = **$	1506
	$P = 1{,}30\text{E}-3$		$P = 8{,}02\text{E}-6$		$P = **$	

به‌کار برده‌ایم. در جدولها، P معرّف کران خطای پیش‌بینی شده در برنامه، و A خطای مطلق واقعی در جواب محاسبه شده است. ستون N تعداد محاسبات انتگرالده را می‌دهد. در تمام نقاطی که انتگرالده نامعین بوده، به‌طور دلخواه مقدار تابع صفر گرفته شده است. مثالها با درجهٔ دقّت مضاعف در یک Prime 850 با واحد گردکردن $10^{-14} \times 1{,}4 \doteq 2^{-46}$ محاسبه شده‌اند.

در جدول ۱۹.۵، CADRE برای I_7 با تحمل خطای 10^{-8} نتیجه‌ای نداده است، با اینکه از (۷.۵.۵) استفاده شده است. در غیر از این حالت نتیجه کاملاً خوب بوده است. وقتی از تجزیه به دو انتگرال (۷.۵.۵) استفاده نشود و CADRE فقط یک بار برای بازهٔ $[-9{,}10000]$ به‌کار رود، هر سه تحمل خطا مردود است.

در جدول ۲۰.۵ خطای پیش‌بینی شده، در بعضی حالات از خطای واقعی کوچکتر است. به‌نظر می‌آید این اشکال به علت کار در حد نهایی درجهٔ دقّت حساب ماشین باشد، و در تمام حالات، خطای نهایی در حدود خواسته شده خوب بوده است.

در مقایسه، برنامه‌های DQAGP و CADRE هر دو اعتمادپذیر و کارا هستند. همچنین، هر دو برنامه، به‌طور نسبتاً ضعیفی برای انتگرال خیلی نوسانی I_6 عمل می‌کنند. این امر نشان

جدول ۲۰.۵ مثالهای انتگرالگیری برای DQAGP

	خطای مطلوب					
	10^{-2}		10^{-5}		10^{-8}	
انتگرال	خطا	N	خطا	N	خطا	N
I_1	$A = 2{,}88\text{E}-9$	105	$A = 5{,}40\text{E}-13$	147	$A = 5{,}40\text{E}-13$	147
	$P = 2{,}96\text{E}-3$		$P = 5{,}21\text{E}-10$		$P = 5{,}21\text{E}-10$	
I_2	$A = 1{,}17\text{E}-11$	21	$A = 1{,}17\text{E}-11$	21	$A = 1{,}17\text{E}-11$	21
	$P = 7{,}46\text{E}-9$		$P = 7{,}46\text{E}-9$		$P = 7{,}46\text{E}-9$	
I_3	$A = 4{,}79\text{E}-6$	21	$A = 4{,}62\text{E}-13$	189	$A = 4{,}62\text{E}-13$	189
	$P = 4{,}95\text{E}-3$		$P = 4{,}77\text{E}-14$		$P = 4{,}77\text{E}-14$	
I_4	$A = 5{,}97\text{E}-13$	231	$A = 5{,}97\text{E}-13$	231	$A = 5{,}97\text{E}-13$	231
	$P = 7{,}15\text{E}-14$		$P = 7{,}15\text{E}-14$		$P = 7{,}15\text{E}-14$	
I_5	$A = 8{,}67\text{E}-13$	462	$A = 8{,}67\text{E}-13$	462	$A = 8{,}67\text{E}-13$	462
	$P = 1{,}15\text{E}-13$		$P = 1{,}15\text{E}-13$		$P = 1{,}15\text{E}-13$	
I_6	$A = 1{,}00\text{E}-3$	525	$A = 6{,}33\text{E}-14$	861	$A = 5{,}33\text{E}-14$	1239
	$P = 4{,}36\text{E}-3$		$P = 8{,}13\text{E}-6$		$P = 7{,}12\text{E}-9$	
I_7	$A = 1{,}67\text{E}-10$	462	$A = 1{,}67\text{E}-10$	462	$A = 1{,}67\text{E}-10$	462
	$P = 1{,}16\text{E}-10$		$P = 1{,}16\text{E}-10$		$P = 1{,}16\text{E}-10$	

می‌دهد که I_6 بایستی با برنامه‌ای محاسبه شود که برای انتگرالهای نوسانی (مانند DQAWO در QUADPACK برای محاسبهٔ ضرایب فوریه) نوشته شده‌اند. از جدولها چنین برمی‌آید که، DAQGP تا اندازه‌ای در برخورد با انتگرالهای مشکل تواناتر است در حالی که برای سایر انتگرالها، در مقایسه با CADRE، کارایی یکسان دارد. مثالهای مفصّلتر برای CADRE در رابینسون[1] (۱۹۷۹) داده شده است.

در محاسبات پرحجم، خطرِ با استفاده کردن از برنامه‌های انتگرالگیری خود کار زیاد است، و ممکن است موجب نتایج اشتباه و عدم کارایی شود. برای توضیح در استفاده از این نوع برنامه‌ها در محاسبات وسیع و توصیه‌ها برای اینکه چه وقت آنها را نتخاب کنید، لاینس (۱۹۸۳) را ببینند. آنچه در زیر می‌آید از نکاتی است که او نتیجه گرفته است.

قاعدهٔ انتگرالگیری خودکار(AQR) یک بخش عملی و مؤثر در نرم‌افزار عددی است. مزیّت اصلی آن برای کاربر راحتی آن است. کاربر می‌تواند آن را از قفسهٔ کتابخانه برداشته، به جریان اندازد و به‌کار آن اطمینان داشته باشد. به دلیل این سهولت، معمولاً یک هزینهٔ جزیی در زمان CPU وجود دارد، این

1. Robinson

هزینهٔ اضافی دارای ضریبی در حدود ۳ است. **روال قاعدهٔ محاسبهٔ انتگرالگیری** (REQR) [قاعده انتگرالگیری غیرخودکار] این هزینهٔ اضافی را ندارد، ولی کدگذاری و کنترل REQR ممکن است چند ساعت از وقت کاربر را بگیرد. بنابراین اگر زمان پیش‌بینی‌شدهٔ CPU زیاد نباشد، بسیاری از کاربرها با طیب خاطر هزینهٔ اضافی را می‌پردازند تا در وقت صرفه‌جویی کنند و گرفتار دردسر نشوند.

مع‌ذالک، بعضی مسائل مشخص ـ معمولاً با مقیاس بزرگ ـ وجود دارند که AQR برای آنها طراحی نشده است و استفادهٔ حساب نشده از آن به زمان CPUی اضافی می‌انجامد که ضریب آن ۱۰۰ یا بیشتر خواهد بود. این مسائل در شرایطی پیش می‌آیند که در آنها تعداد زیادی انتگرالگیریهای مجزا وجود دارند و نتایج این انتگرالگیریها بعداً در بعضی از فرآیندهای عددی دیگر، به‌عنوان ورودی مورد استفاده قرار می‌گیرند. برای شناختن چنین وضعیتی، لازم است که فرایند عددی بعدی امتحان شود تا ببینند که آیا به تابع ورودی هموار نیاز دارد .... برای بعضی از این مسائل یک REQR کاملاً مناسب است، و حال آنکه یک AQR ممکن است به یک ناکامی عددی بینجامد.

۶.۵ انتگرالهای تکین

در این بخش محاسبهٔ تقریبی انتگرالهایی را مورد بحث قرار می‌دهیم که برای آنها روشهای بحث شده در بخشهای ۱.۵ تا ۴.۵، خوب عمل نمی‌کنند: این روشها شامل قواعد مرکب نیوتن ـ کوتس (مثلاً، قاعدهٔ ذوزنقه‌یی)، تربیع گاوس ـ لژاندر و انتگرالگیری رامبرگ هستند. انتگرالهایی که در اینجا بحث می‌شوند، اگر با قواعدی که اخیراً نامبرده شد حساب شوند، به دلایل مختلف، به همگرایی عددی ضعیف می‌انجامند. در اینجا در موارد زیر بحث خواهیم کرد انتگرالهایی که (۱) انتگرالده آنها در بازهٔ انتگرالگیری (a, b) دارای تکینی باشند، (۲) انتگرالهایی که بازهٔ انتگرالگیری نامتناهی دارند. روشهای انتگرالگیری تطبیقی را می‌توان برای این انتگرالها به‌کارگرفت، ولی اگر ماهیّت رفتار تکین آنها به دقّت بررسی و ترمیم شوند معمولاً ممکن است بتوان یک تقریب با همگرایی سریعتر پیدا کرد.

تبدیل متغیر انتگرالگیری اهمیت این روش را با چند مثال نشان خواهیم داد. برای

$$I = \int_a^b \frac{f(x)}{\sqrt{x}} dx \qquad (۱.۶.۵)$$

که $f(x)$ تابعی چندین بار پیوسته مشتقپذیر است، می‌گیریم $x = u^2$ و $0 \le u \le \sqrt{b}$. در این صورت

$$I = 2 \int_0^{\sqrt{b}} f(u^2) du$$

این انتگرال دارای انتگرالده هموار است و روشهای استانده را می‌توان برای آن به‌کار برد.
همچنین با تبدیل متغیّر $u = \sqrt{۱ - x}$،

$$\int_{۰}^{۱} \sin(x)\sqrt{۱ - x^{۲}}dx = ۲\int_{۰}^{۱} u^{۲}\sqrt{۲ - u^{۲}}\sin(۱ - u^{۲})du$$

انتگرالده سمت راست تعداد نامتناهی مشتق پیوسته در $[۰, ۱]$ دارد، حال آنکه مشتق انتگرالده نخست، در $x = ۱$ تکین بود.

برای یک بازهٔ انتگرالگیری نامتناهی، روش تبدیل متغیر باز هم مفید خواهد بود. فرض کنید

$$I = \int_{۱}^{\infty} \frac{f(x)}{x^{p}}dx \qquad p > ۱ \qquad (۲.۶.۵)$$

و $\lim_{x\to\infty} f(x)$ وجود دارد. همچنین فرض کنید $f(x)$ در بازهٔ $[۱, \infty)$ هموار است. با استفاده از تبدیل متغیر

$$x = \frac{۱}{u^{\alpha}} \qquad dx = \frac{-\alpha}{u^{۱+\alpha}} \qquad \text{برای} \quad \alpha > ۰$$

داریم

$$I = \alpha\int_{۰}^{۱} u^{p\alpha} f\left(\frac{۱}{u^{\alpha}}\right)\frac{du}{u^{۱+\alpha}} = \alpha\int_{۰}^{۱} u^{(p-۱)\alpha-۱} f\left(\frac{۱}{u^{\alpha}}\right)du \qquad (۳.۶.۵)$$

با انتخاب α به گونه‌ای که نمای $(p-۱)\alpha - ۱$ بزرگ شود، همواری انتگرالده جدید را در $u = ۰$ ماکسیمم می‌نماییم. برای مثال با

$$I = \int_{۱}^{\infty} \frac{f(x)}{x\sqrt{x}}dx$$

تبدیل متغیر $x = ۱/u^{۴}$ به انتگرال زیر می‌انجامد

$$I = ۴\int_{۰}^{۱} uf\left(\frac{۱}{u^{۴}}\right)du \qquad (۴.۶.۵)$$

اگر رفتار $f(x)$ را در $x = \infty$

$$f(x) = c_{۰} + \frac{c_{۱}}{x} + \frac{c_{۲}}{x^{۲}} + \cdots$$

فرض کنیم، آنگاه

$$uf(\frac{۱}{u^{۴}}) = c_{۰}u + c_{۱}u^{۵} + c_{۲}u^{۹} + \cdots$$

و (۴.۶.۵) یک انتگرالده هموار در $u = ۰$ خواهد داشت.

در برخورد با تکینی‌های انتگرالده در نقاط انتهایی انتگرال

$$I = \int_a^b f(x)dx \qquad (۵.۶.۵)$$

آیری[1] و همکاران (۱۹۷۰) یک فکر جالب دارند: تعریف می‌کنیم

$$\psi(t) = \exp\left(\frac{-c}{۱ - t^۲}\right) \qquad (۶.۶.۵)$$

$$\varphi(t) = a + \frac{b-a}{\gamma}\int_{-۱}^{t} \psi(u)du \qquad -۱ \le t \le ۱ \qquad (۷.۶.۵)$$

که در آن c یک ثابت مثبت است و

$$\gamma = \int_{-۱}^{۱} \psi(u)du$$

وقتی t از -۱ تا ۱ تغییر می‌کند، $\varphi(t)$ از a تا b تغییر می‌کند. با استفاده از $x = \varphi(t)$ به‌عنوان تبدیل متغیر در (۵.۶.۵)، به‌دست می‌آوریم

$$I = \int_{-۱}^{۱} f(\varphi(t))\varphi'(t)dt \qquad (۸.۶.۵)$$

تابع $\varphi'(t) = ((b-a)/\gamma)\psi(t)$ در $[-۱, ۱]$ بینهایت بار مشتقپذیر است و تابع وتمام مشتقات آن در $t = \pm ۱$ صفرند. در (۸.۶.۵) انتگرالده و تمام مشتقات آن در $t = \pm ۱$، برای تقریباً تمام توابع قابل توجه صفر می‌شوند. با استفاده از فرمول خطای (۹.۴.۵) برای قاعدهٔ ذوزنقه‌یی در $[-۱, ۱]$، می‌توان دید که قاعدهٔ ذوزنقه‌یی وقتی برای (۸.۶.۵) به‌کار برده شود، با سرعت زیاد همگراست. ما این روش را روش IMT می‌نامیم.

این روش در دِدونکر و پیسنز[2] (۱۹۷۶) و در مقایسه‌های کلی رابینسون (۱۹۷۹)، عرضه شده است و به‌عنوان یک راه بسیار اعتمادپذیر و کارا در مواجهه با انتگرالهای (۴.۶.۵) که در نقاط انتهایی تکینی دارند، شناخته شده است. دِدونکر و پیسنز (۱۹۷۶) همچنین انتگرالها را روی $[۰, \infty)$ نیز بررسی کرده‌اند، بدین طریق که ابتدا تبدیل متغیر $x = (۱ + u)/(۱ - u)$، $-۱ \le u < ۱$ و سپس تبدیل متغیر $u = \varphi(t)$ را به‌کار برده‌اند.

مثال روش قبلی (۵.۶.۵)ـ(۸.۶.۵) را با قاعدهٔ ذوزنقه‌یی برای محاسبهٔ انتگرال زیر به‌کار برید

$$I = \int_۰^۱ \sqrt{-\ln x}\, dx = \frac{\sqrt{\pi}}{۲} \doteq ۰٫۸۸۶۲۲۶۹ \qquad (۹.۶.۵)$$

---

1. Iri 2. de Doncker and Piessens

جدول ۲۱.۵ مثال برای روش IMT

گره‌ها	خطا
۲	$-۶٫۵۴\mathrm{E}-۲$
۴	$۵٫۸۲\mathrm{E}-۳$
۸	$-۱٫۳۰\mathrm{E}-۴$
۱۶	$۷٫۴۲\mathrm{E}-۶$
۳۲	$۱٫۱۷\mathrm{E}-۸$
۶۴	$۱٫۱۸\mathrm{E}-۱۲$

توجه داشته باشید که انتگرالده در هر دو نقطهٔ انتهایی دارای رفتار تکین است، اگرچه در این دو حالت رفتار متفاوت است. ثابت c در (۶.۶.۵) برابر ۴ است و محاسبهٔ (۷.۶.۵) از دِدوکر و رابینسون اخذ شده است. نتایج در جدول ۲۱.۵ نشان داده شده‌اند. ستونی که با گره‌ها مشخص شده تعداد نقاط گرهی را در بازهٔ $[۰, ۱]$ مشخص می‌نماید.

انتگرالگیری گاوسی در بخش ۳.۵ یک نظریه کلی فرمولهای انتگرالگیری گاوسی

$$\int_a^b w(x)f(x)dx \doteq \sum_{j=۱}^{n} w_{j,n} f(x_{j,n}) \qquad n \geq ۱$$

را که درجهٔ دقّت $۲n-۱$ دارند، توضیح دادیم. پیدا کردن گره‌ها و وزنها و شکل خطا در قضیهٔ ۳.۵ داده شده است. برای کار در این بخش ملاحظه می‌کنیم که: (۱) بازهٔ (a, b) ممکن است نامتناهی باشد، و (۲) $w(x)$ ممکن است در (a, b) تکینیهایی داشته باشد، به شرطی که منفی نباشد و در فرضهای (۸.۳.۴) و (۹.۳.۴) از بخش ۳.۴ صدق کند. برای همگرایی سریع، همچنین پیش‌بینی می‌کنیم که لازم است $f(x)$ تابعی هموار باشد، همان گونه که در انتگرالگیری گاوس ـ لژاندر در بخش ۳.۵ نشان داده شد.

وزنها و گره‌ها برای تعداد زیادی از توابع وزن $w(x)$ و بازه‌های (a, b) معلوم‌اند. جداول استرود و سکرست (۱۹۶۶) شامل انتگرالهای

$$\int_۰^\infty x^\alpha e^{-x} f(x)dx \qquad \int_{-\infty}^{\infty} e^{-x^۲} f(x)dx \qquad \int_۰^۱ \ln\left(\frac{۱}{x}\right) f(x)dx \quad (۱۰.۶.۵)$$

و بعضی انتگرالهای دیگر است. ثابت α بزرگتر از -۱ است. کتابهای دیگری هستند که جداولی برای انتگرالهای غیر از انتگرالهای (۱۰.۶.۵) دارند. همچنین مقالهٔ گلوب[1] و وِلش[2] (۱۹۶۹)

1. Golub 2. Welsch

شیوه‌ای برای پیدا کردن گره‌ها و وزنهای (۱۰.۶.۵) برمبنای حلّ مسألهٔ ویژه مقدار ماتریسی، به‌دست می‌دهد. یک برنامه داده شده است که بیشتر شامل انتگرالهای وزین متداولتری است که در انتگرالگیری گاوسی به‌کار برده می‌شوند. برای یک بحث اضافی از انتگرالگیری گاوس، با ارجاع به نوشتارها (از جمله جداول و برنامه‌ها)، دیویس و رابینویتس (۱۹۸۴، صص ۹۵ ـ ۱۳۲، ۲۲۲ ـ ۲۲۹) را ببینید.

مثال استفاده از انتگرالگیری گاوسی را برای محاسبهٔ انتگرال زیر نشان می‌دهیم:

$$I = \int_{0}^{\infty} g(x)dx$$

انتگرالگیری گاوس ـ لاگر[1] را که در آن $w(x) = e^{-x}$ اختیار شده به‌کار می‌بریم. سپس I را به‌صورت زیر می‌نویسیم

$$I = \int_{0}^{\infty} e^{-x}[e^{x}g(x)]dx = \int_{0}^{\infty} e^{-x}f(x)dx \qquad (۱۱.۶.۵)$$

نتایج را برای سه انتگرال زیر می‌دهیم

$$I^{(1)} = \int_{0}^{\infty} \frac{xdx}{e^{x}-1} = \frac{\pi^{2}}{6}$$
$$I^{(2)} = \int_{0}^{\infty} \frac{xdx}{(1+x^{2})^{5}} = \frac{1}{8}$$
$$I^{(3)} = \int_{0}^{\infty} \frac{dx}{1+x^{2}} = \frac{\pi}{2}$$

انتگرالگیری گاوس ـ لاگر برای انتگرالده‌هایی که وقتی $x \to \infty$، به‌طور نمایی کاهش می‌یابند، از همه بهتر است. برای انتگرالده‌هایی که وقتی $x \to \infty$، $O(\frac{1}{x^{p}})$، $p > 1$، هستند، نرخ همگرایی وقتی $p \to 1$ بسیار ضعیف می‌شود. این توضیحات در جدول ۲۲.۵ نمایش داده شده است. برای یک بحث صوری در همگرایی انتگرالگیری گاوس ـ لاگر، دیویس و رابینوویتس (۱۹۸۴، ص ۲۲۷) را ببینید.

یک مورد به‌ویژه سادهٔ انتگرالگیری گاوسی، انتگرال تکین زیر است

$$I(f) = \int_{-1}^{1} \frac{f(x)dx}{\sqrt{1-x^{2}}} \qquad (۱۲.۶.۵)$$

1. Gauss-Laguerre

جدول ۲۲.۵ مثالهایی از انتگرالگیری گاوس ـ لاگر

گره‌ها	$I^{(1)} - I_n^{(1)}$	$I^{(2)} - I_n^{(2)}$	$I^{(3)} - I_n^{(3)}$
۲	۱٫۰۱E − ۴	−۸٫۰۵E − ۲	۷٫۷۵E − ۲
۴	۱٫۲۸E − ۵	−۴٫۲۰E − ۲	۶٫۶۹E − ۲
۸	−۹٫۴۸E − ۸	۱٫۲۷E − ۲	۳٫۷۰E − ۲
۱۶	۳٫۱۶E − ۱۱	−۱٫۳۹E − ۳	۱٫۷۱E − ۲
۳۲	۷٫۱۱E − ۱۴	۳٫۰۵E − ۵	۸٫۳۱E − ۳
۶۴	—	۱٫۰۶E − ۷	۴٫۰۷E − ۳

با این تابع وزن، چند جمله‌ییهای متعامد چند جمله‌ییهای چبیشف $\{T_n(x),\ n \geq 0\}$ هستند. بنابراین گره‌های انتگرالگیری در (۱۰.۶.۵) چنین داده می‌شوند

$$x_{j,n} = \cos\left(\frac{2j-1}{2n}\pi\right) \qquad j = 1, \ldots, n \qquad (13.6.5)$$

و با توجّه به (۱۱.۳.۵)، وزنها عبارت‌اند از

$$w_{j,n} = \frac{\pi}{n} \qquad j = 1, \ldots, n$$

با استفاده از فرمول (۱۰.۳.۵) برای خطا، فرمول انتگرالگیری گاوسی برای (۱۲.۶.۵) با رابطهٔ زیر داده می‌شود

$$\int_{-1}^{1} \frac{f(x)dx}{\sqrt{1-x^2}} = \frac{\pi}{n}\sum_{j=1}^{n} f(x_{j,n}) + \frac{2\pi}{2^{2n}(2n)!} f^{(2n)}(\eta) \qquad (14.6.5)$$

برای $-1 < \eta < 1$

این فرمول بستگی به قاعدهٔ میانگاهی مرکب (۱۸.۲.۵) دارد. اگر تبدیل متغیر $x = \cos\theta$ در (۱۴.۶.۵) داده شود، چنین به دست خواهد آمد

$$\int_0^{\pi} f(\cos\theta)d\theta = \frac{\pi}{n}\sum_{j=1}^{n} f(\cos\theta_{j,n}) + E \qquad (15.6.5)$$

که $\theta_{j,n} = (2j-1)\pi/2n$. بنابراین انتگرالگیری گاوسی برای (۱۲.۶.۵) هم‌ارز است با قاعدهٔ میانگاهی مرکب که برای انتگرال سمت چپ (۱۵.۶.۵) به‌کار برده شده‌است. مانند قاعدهٔ ذوزنقه‌یی، قاعدهٔ میانگاهی دارای بسط خطایی است بسیار شبیه به بسط خطایی که در (۹.۴.۵)، با استفاده از فرمول اویلر ـ مک‌لورن داده شده است. فرع ۱ مربوط به قضیهٔ ۵.۵ نیز معتبر است، که نشان

می‌دهد قاعدهٔ میانگاهی مرکب برای توابع دوره‌یی بسیار دقیق است. این دقت زیاد در (۱۴.۶.۵) منعکس شده‌است. بنابراین انتگرالگیری گاوسی برای (۱۲.۶.۵) به فرمولی می‌انجامد که مناسب بودن آن از بسط خطای مجانبی برای قاعدهٔ میانگاهی مرکب وقتی که برای انتگرال سمت چپ (۱۵.۶.۵) به‌کار برده شود، به‌دست می‌آید.

برخورد تحلیلی با تکینی بازهٔ انتگرالگیری را به دو بخش تقسیم می‌کنیم، یک بخش شامل نقطهٔ تکینی است که بایستی به گونه‌ای تحلیلی به آن پرداخته شود. برای مثال، انتگرال زیر را درنظر می‌گیریم

$$I = \int_0^b f(x)\log(x)dx = \left[\int_0^\varepsilon + \int_\varepsilon^b\right] f(x)\log(x)dx \equiv I_1 + I_2 \quad (۱۶.۶.۵)$$

با فرض اینکه $f(x)$ در $[\varepsilon, b]$ هموار است، یک روش استانده‌ای برای محاسبهٔ I_2 به‌کار می‌بریم. برای I_1 فرض می‌کنیم $f(x)$ حول صفر دارای یک سری همگرای تیلر در $[0, \varepsilon]$ باشد. در این صورت

$$I_1 = \int_0^\varepsilon f(x)\log(x)dx = \int_0^\varepsilon \left(\sum_0^\infty a_j x^j\right)\log(x)dx$$
$$= \sum_0^\infty a_j \frac{\varepsilon^{j+1}}{j+1}\left[\log(\varepsilon) - \frac{1}{j+1}\right] \quad (۱۷.۶.۵)$$

برای مثال، در مورد

$$I = \int_0^{4\pi} \cos(x)\log(x)dx$$

تعریف می‌کنیم

$$I_1 = \int_0^{0.1} \cos(x)\log(x)dx \qquad \varepsilon = 0.1$$
$$= \varepsilon[\log(\varepsilon) - 1] - \frac{\varepsilon^3}{6}\left[\log(\varepsilon) - \frac{1}{3}\right] + \frac{\varepsilon^5}{600}\left[\log(\varepsilon) - \frac{1}{5}\right] - \cdots$$

که یک سری متناوب است و بنابراین روشن است که با سه جملهٔ اول یک تقریب دقیق برای I_1 به دست خواهد آمد. یک روش استانده می‌توان برای I_2 در $[0.1, 4\pi]$ به‌کار برد.

تکنیکهای مشابهی می‌توان برای بازه‌های نامتناهی انتگرالگیری $[a, \infty)$ به‌کار برد که انتگرال را در بازهٔ $[b, \infty)$ برای مقدار بزرگی از b حذف می‌کند. این روش در اینجا بحث نشده است.

انتگرالگیری حاصلضرب گیریم $I(f) = \int_a^b w(x)f(x)dx$ که در آن تابع وزن $w(x)$، نزدیک به تکین یا تکین، و $f(x)$ یک تابع هموار است. هدف اصلی این است که دنباله‌ای از توابع $f_n(x)$ تولید کنیم به‌طوری که

۱. $\| f - f_n \|_\infty = \underset{a \le x \le b}{\text{Max}} | f(x) - f_n(x) | \to ۰$ وقتی $n \to \infty$

۲. انتگرالهای

$$I_n(f) \equiv \int_a^b w(x) f_n(x) dx \qquad (۱۸.۶.۵)$$

به سادگی قابل محاسبه باشند.

این هدف تعمیم طرح (۲.۰.۵) است که در مقدمه آمده است. برای خطا،

$$\begin{aligned} | I(f) - I_n(f) | &\le \int_a^b | w(x) | | f(x) - f_n(x) | dx \\ &\le \| f - f_n \|_\infty \int_a^b | w(x) | dx \end{aligned} \qquad (۱۹.۶.۵)$$

بنابراین وقتی $I_n(f) \to I(f)$ ،$n \to \infty$، و نرخ همگرایی حداقل به همان نرخ همگرایی $f_n(x)$ به $f(x)$ در $[a, b]$ است.

در چارچوب فوق، برای تعریف روشهای انتگرالگیری حاصلضرب معمولاً $f_n(x)$ را از $f(x)$ با استفاده از درونیابی چندجمله‌یی تکه‌یی تعریف می‌کنند. برای نشان دادن طرح اصلی، با جبری ساده روش ذوزنقه‌یی حاصلضرب را برای محاسبهٔ انتگرال زیر تعریف می‌کنیم

$$I(f) = \int_۰^b f(x) \log(x) dx \qquad (۲۰.۶.۵)$$

گیریم $n \ge ۱$ ،$h = b/n$ و $x_j = jh$ برای $j = ۰, ۱ \dots n$. $f_n(x)$ را تابع خطی تکه‌یی که $f(x)$ را در گره‌های $x_۰, x_۱, \dots x_n$ درونیابی می‌نماید چنین تعریف می‌کنیم.

برای $x_{j-۱} \le x \le x_j$ ،$j = ۱, ۲, \cdots, n$ تعریف می‌کنیم

$$f_n(x) = \frac{۱}{h}[(x_j - x) f(x_{j-۱}) + (x - x_{j-۱}) f(x_j)] \qquad (۲۱.۶.۵)$$

با توجّه به (۱۰.۱.۳) می‌توان به سادگی نشان داد

$$\| f - f_n \|_\infty \le \frac{h^۲}{۸} \| f'' \|_\infty \qquad (۲۲.۶.۵)$$

مشروط بر آنکه $f(x)$ دوبار پیوسته مشتقپذیر در $b \le x \le ۰$ باشد. از (۱۹.۶.۵) کران خطا را به دست می‌آوریم

$$| I(f) - I_n(f) | \le \frac{h^۲}{۸} \| f'' \|_\infty \int_۰^b | \log(x) | dx \qquad (۲۳.۶.۵)$$

این روش تعریف $I_n(f)$ مشابه قاعدهٔ ذوزنقه‌یی (۵.۱.۵) است. قاعدهٔ (۵.۱.۵) را می‌توانستیم از انتگرالگیری تابع $f_n(x)$ قبلی نیز به‌دست آوریم. در این صورت تابع وزن $w(x) \equiv ۱$ می‌شد. به سادگی می‌توانیم مطالب فوق را با به‌کار بردن درونیابی چندجمله‌یی تکه‌یی از درجهٔ بالاتر تعمیم دهیم. استفاده از درونیابی مرتبهٔ دوم تکه‌یی برای تعریف $f_n(x)$، به یک فرمول $I_n(f)$ می‌انجامد که قاعدهٔ حاصلضرب سیمپسون خوانده می‌شود. و با استفاده از استدلالی که به (۲۳.۶.۵)) رسیدیم، می‌توان نشان داد که

$$| I(f) - I_n(f) | \leq \frac{\sqrt{۳}}{۲۷} h^۳ \| f^{(۳)} \|_\infty \int_۰^b | \log(x) | dx \qquad (۲۴.۶.۵)$$

فرمولهای مراتب بالاتر را می‌توان با استفاده از درونیابیها درجهٔ بالاتر به‌دست آورد.

برای محاسبهٔ $I_n(f)$ با استفاده از (۲۱.۶.۵) داریم

$$I_n(f) = \sum_{j=۱}^{n} \int_{x_{j-۱}}^{x_j} \log(x) \left[\frac{(x_j - x) f(x_{j-۱}) + (x - x_{j-۱}) f(x_j)}{h} \right] dx$$

$$= \sum_{k=۰}^{n} w_k f(x_k) \qquad (۲۵.۶.۵)$$

$$w_۰ = \frac{۱}{h} \int_{x_۰}^{x_۱} (x_۱ - x) \log(x) dx \qquad w_n = \frac{۱}{h} \int_{x_{n-۱}}^{x_n} (x - x_{n-۱}) \log(x) dx,$$

$$w_j = \frac{۱}{h} \int_{x_{j-۱}}^{x_j} (x - x_{j-۱}) \log(x) dx$$

$$+ \frac{۱}{h} \int_{x_j}^{x_{j+۱}} (x_{j+۱} - x) \log(x) dx \qquad j = ۱, \ldots, n-۱ \qquad (۲۶.۶.۵)$$

محاسبهٔ این وزنها را می‌توان خیلی ساده کرد. با تبدیل متغیر $x - x_{j-۱} = uh$، $۰ \leq u \leq ۱$، داریم

$$\frac{۱}{h} \int_{x_{j-۱}}^{x_j} (x - x_{j-۱}) \log(x) dx = h \int_۰^۱ u \log[(j - ۱ + u)h] du$$

$$= \frac{h}{۲} \log(h) + h \int_۰^۱ u \log(j - ۱ + u) du$$

و

$$\frac{۱}{h} \int_{x_{j-۱}}^{x_j} (x_j - x) \log(x) dx = h \int_۰^۱ (۱ - u) \log[(j - ۱ + u)h] du$$

$$= \frac{h}{۲} \log(h) + h \int_۰^۱ (۱ - u) \log(j - ۱ + u) du$$

تعریف می‌کنیم

جدول ۲۳.۵ وزنها برای قاعدهٔ ذوزنقه‌یی حاصلضرب

k	$\psi_1(k)$	$\psi_2(k)$
۰	−۰٫۲۵۰	−۰٫۷۵۰
۱	۰٫۲۵۰	۰٫۱۳۶۲۹۴۳۶۱۱
۲	۰٫۴۸۸۳۷۵۹۲۸۱	۰٫۴۲۱۱۶۶۵۷۶۸
۳	۰٫۶۴۸۵۷۷۸۵۴۵	۰٫۶۰۰۷۶۲۷۲۳۹
۴	۰٫۷۶۹۵۷۰۵۴۵۷	۰٫۷۳۲۴۴۱۵۷۲۰
۵	۰٫۸۶۶۸۶۰۲۷۴۷	۰٫۸۳۶۵۰۶۹۷۸۵
۶	۰٫۹۴۸۲۴۲۸۳۷۶	۰٫۹۲۲۵۷۱۳۹۰۴
۷	۱٫۰۱۸۲۰۱۶۵۲	۰٫۹۹۵۹۵۹۶۳۸۵

$$\psi_1(k) = \int_0^1 u \log(u+k)du$$
$$\psi_2(k) = \int_0^1 (1-u)\log(u+k)du \qquad (۲۷.۶.۵)$$

برای $k = 0, 1, 2 \ldots$ در این صورت

$$w_0 = \frac{h}{2}\log(h) + h\psi_2(0) \qquad w_n = \frac{h}{2}\log(h) + h\psi_1(n-1)$$
$$w_j = h\log(h) + h[\psi_i(j-1) + \psi_2(j)] \quad j = 1, 2 \ldots, n-1 \qquad (۲۸.۶.۵)$$

توابع $\psi_1(k)$ و $\psi_2(k)$ به h، b یا n بستگی ندارند. آنها را می‌توان محاسبه و در یک جدول نگهداری کرد و برای مقادیر متفاوت b و n به‌کار برد. برای مثال، یک جدول $\psi_1(k)$ و $\psi_2(k)$ به‌ازای $k = 0, 1, \ldots, 99$ را می‌توان برای هر مقدار $b > 0$ و هر $n \leq 100$ به‌کار برد. وقتی که جدول مقادیر $\psi_1(k)$ و $\psi_2(k)$ محاسبه شده باشد، هزینهٔ استفاده از قاعدهٔ ذوزنقه‌یی حاصلضرب بیش از هزینهٔ هیچ قاعدهٔ انتگرالگیری دیگری نخواهد بود.

انتگرالهای $\psi_1(k)$ و $\psi_2(k)$ در (۲۷.۶.۵) را می‌توان به‌طور صریح محاسبه کرد؛ بعضی از مقادیر آنها در جدول ۲۳.۵ داده شده است.

مثال $I = \int_0^1 (1/(x+2))\log(x)dx \doteq -۰٫۴۴۸۴۱۳۷$ را محاسبه کنید. مقادیر محاسبه شده در جدول ۲۴.۵ داده شده‌اند. آهنگ همگرایی محاسبه شده با مرتبهٔ همگرایی (۲۳.۶.۵) مطابقت دارد. انواع بسیاری درونیابی را می‌توان برای تعریف $f_n(x)$ به‌کار برد، ولی تاکنون در بیشتر کاربردها، چندجمله‌یی‌های درونیابی تکه‌یی با نقاط گرهی متساوی‌الفاصله به‌کار رفته‌اند. تابعهای وزن دیگری

جدول ۲۴.۵ مثال برای قاعدهٔ ذوزنقه‌یی حاصلضرب

n	I_n	$I - I_n$	نسبت
۱	−۰٫۴۵۸۳۳۳۳	۰٫۰۰۹۹۲	
۲	−۰٫۴۵۱۶۰۹۶	۰٫۰۰۳۲۰	۳٫۱۰
۴	−۰٫۴۴۹۳۰۱۱	۰٫۰۰۰۸۸۷	۲٫۶۱
۸	−۰٫۴۴۸۶۴۶۰	۰٫۰۰۰۲۳۲	۳٫۸۲

را نیز می‌توان به‌کار برد. برای مثال

$$w(x) = x^{\alpha} \qquad \alpha > -۱ \qquad x \geq ۰ \qquad (۲۹.۶.۵)$$

و باز هم تابعهای وزن را می‌توان به فرمولهای نسبتاً ساده‌تری شبیه (۲۸.۶.۵) بدل کرد. برای یک مقدار گنگ α، چون

$$w(x) = \frac{۱}{x^{\sqrt{۲}-۱}}$$

برای از بین بردن تکینی در انتگرال، دیگر تبدیل متغیر را نمی‌توان به‌کار برد. همچنین یکی دیگر از کاربردهای اصلی انتگرالگیری حاصلضرب در معادلات انتگرالی است که در آنها تابع هسته تکینی جبری و/یا لگاریتمی دارد. برای چنین معادلاتی حتی با تکینی ریشهٔ دوم دیگر تبدیل متغیر ممکن نیست. برای مثال، معادلهٔ

$$\lambda\varphi(x) - \int_a^b \frac{\varphi(y)dy}{|x-y|^{۱/۲}} = f(x) \qquad a \leq x \leq b$$

با مقادیر داده شدهٔ λ، a، b، f و تابع مجهول خواسته شدهٔ φ را درنظر می‌گیریم. انتگرالگیری حاصلضرب برای چنین معادلاتی به روندهای کارا می‌انجامد، به‌شرطی که $\varphi(y)$ یک تابع هموار باشد [ایتکین[1] ۱۹۷۶، ص ۱۰۶ را ببینید.]

برای توابع وزن پیچیده که در آنها وزنهای w_j را دیگر نمی‌توان محاسبه نمود، اغلب امکان دارد مسأله را به مسألهٔ دیگری برگرداند که انتگرالگیری عددی برای آن باز به‌سادگی قابل اجرا باشد. این مطلب را با یک مثال بررسی خواهیم کرد.

مثال انتگرال $I = \int_۰^{\pi} f(x)\log(\sin x)dx$ را درنظر می‌گیریم. انتگرالده در هر دو نقطهٔ $x = ۰$

1. Atkinson

و $x = \pi$ تکینی دارد. با استفاده از

$$\log(\sin x) = \log\left[\frac{\sin x}{x(\pi - x)}\right] + \log(x) + \log(x - \pi)$$

داریم

$$\begin{aligned} I &= \int_{\circ}^{\pi} f(x) \log\left[\frac{\sin x}{x(\pi - x)}\right] dx + \int_{\circ}^{\pi} f(x) \log(x) dx \\ &\quad + \int_{\circ}^{\pi} f(x) \log(\pi - x) dx \\ &\equiv I_{۱} + I_{۲} + I_{۳} \end{aligned} \qquad (۳۰.۶.۵)$$

انتگرال $I_{۱}$، انتگرالده بینهایت مشتقپذیر دارد و هر روش عددی استانده برای آن به خوبی اجرا می‌شود. انتگرال $I_{۲}$ قبلاً مورد بحث قرار گرفته است، با $w(x) = \log(x)$. برای $I_{۳}$، یک تبدیل متغیر به‌کار می‌بریم و می‌نویسیم

$$I_{۳} = \int_{\circ}^{\pi} f(x) \log(\pi - x) dx = \int_{\circ}^{\pi} f(\pi - z) \log(z)\, dz$$

از ترکیب با $I_{۲}$ داریم

$$I_{۲} + I_{۳} = \int_{\circ}^{\pi} \log(x)[f(x) + f(\pi - x)] dx$$

که برای آن همان عمل قبلی به‌کار می‌رود. با این دستکاریها حالتهای $w(x) = \log(x)$ و $w(x) = x^{\alpha}$ بسیار بیشتر از آن که در ابتدا تصور می‌شد، قابل استفاده است.

برای تحلیل خطای مجانبی انتگرالگیری حاصلضرب، اثر دِهوگ[۱] و وایس[۲] (۱۹۷۳) را ببینید، که در آن بعضی از تعمیمهای بسط اویلر ـ مک‌لورن داده شده است. با استفاده از نتایجی که به‌دست آورده‌اند، می‌توان نشان داد که خطا در قاعدهٔ حاصلضرب سیمپسون برابر $O(h^{۴}\log(h))$ است. بنابراین کران (۲۴.۶.۵) بر پایهٔ خطای درونیابی $f(x) - f_n(x)$، نرخ همگرایی صحیحی را پیش‌بینی نمی‌کند. این نتیجه مشابه با نتیجهٔ (۱۷.۱.۵) برای خطای قاعدهٔ سیمپسون است که در آنجا خطا کوچکتر از آن بود که با استفاده از درونیابی درجهٔ دوم می‌توانستیم نتیجه بگیریم.

۷.۵ مشتقگیری عددی

تقریبهای عددی برای مشتقات عمدتاً به دو منظور مورد استفاده واقع می‌شوند. نخست، علاقه‌مندیم مشتقات داده‌هایی را حساب کنیم که اغلب به‌طور تجربی به‌دست آمده‌اند. دوم، فرمولهای مشتقگیری

1. de Hoog 2. Weiss

عددی را برای پیدا کردن روشهای عددی در حل معادلات دیفرانسیل معمولی و معادلات دیفرانسیل جزیی به‌کار می‌بریم. ما این بخش را با به‌دست آوردن فرمولهایی که بیشترین کاربرد را برای مشتقگیری عددی دارند، آغاز می‌کنیم.

مسالهٔ مشتقگیری عددی از بعضی جهات مشکلتر از انتگرالگیری عددی است. وقتی از مقادیر تابع که به‌صورت تجربی معین شده‌اند، استفاده می‌کنیم، خطای این مقادیر، معمولاً به ناپایداری در مشتقگیری عددی تابع می‌انجامد. برعکس، انتگرالگیری عددی وقتی با چنین خطاهایی مواجه می‌شود پایدار است (مسألهٔ ۱۳ را ببینید).

فرمولهای کلاسیک یکی از راههای اساسی پیدا کردن تقریب عددی برای $f'(x)$، استفاده از مشتق چندجمله‌یی $p_n(x)$ است که $f(x)$ را در یک مجموعه نقاط گرهی معین درونیابی می‌کند. گیریم x_0، x_1، $\ldots$ x_n داده شده باشند و $p_n(x)$ تابع $f(x)$ را در این گره‌ها درونیابی کند. معمولاً $\{x_i\}$ها متساوی‌الفاصله‌اند. در این صورت از رابطهٔ زیر استفاده می‌کنیم

$$f'(x) \doteq p_n'(x) \tag{۱.۷.۵}$$

از روابط (۶.۱.۳)، (۴.۲.۳) و (۱۱.۲.۳) داریم:

$$\begin{aligned} p_n(x) &= \sum_{j=0}^{n} f(x_j) l_j(x) \\ l_j(x) &= \frac{\Psi_n(x)}{(x-x_j)\Psi_n'(x_j)} \\ &= \frac{(x-x_0)\ldots(x-x_{j-1})(x-x_{j+1})\ldots(x-x_n)}{(x_j-x_0)\ldots(x_j-x_{j-1})(x_j-x_{j+1})\ldots(x_j-x_n)} \\ \Psi_n(x) &= (x-x_0)\ldots(x-x_n) \end{aligned}$$

$$f(x) - p_n(x) = \Psi_n(x) f[x_0, \ldots, x_n, x] \tag{۲.۷.۵}$$

بنابراین

$$f'(x) \doteq p_n'(x) = \sum_{j=0}^{n} f(x_j) l_j'(x) \equiv D_n f(x) \tag{۳.۷.۵}$$

$$\begin{aligned} f'(x) - D_h f(x) =& \Psi_n'(x) f[x_0, \ldots, x_n, x] \\ &+ \Psi_n(x) f[x_0, \ldots, x_n, x, x] \end{aligned} \tag{۴.۷.۵}$$

که در مرحلهٔ آخر از رابطهٔ (۱۷.۲.۳) استفاده شده است. با استفاده از (۱۲.۲.۳)،

$$f'(x) - D_h f(x) = \Psi'_n(x)\frac{f^{(n+1)}(\xi_1)}{(n+1)!} + \Psi_n(x)\frac{f^{(n+2)}(\xi_2)}{(n+2)!} \qquad (۵.۷.۵)$$

که $\xi_1, \xi_2 \in \mathscr{H}\{x_0, \ldots, x_n, x\}$. فرمولهای مشتقگیری مراتب بالاتر و خطای آنها را می‌توان از مشتقگیری بیشتر از (۳.۷.۵) و (۴.۷.۵) به‌دست آورد.

عادیترین کاربردهای مطالب فوق در نقاط گرهی متساوی‌الفاصلهٔ $\{x_i\}$ است. در این حالت گیریم $x_i = x_0 + ih$، $i \geq 0$ و $h > 0$. به‌سادگی می‌توان نشان داد که

$$\Psi_n(x) = O(h^{n+1}) \qquad \Psi'_n(x) = O(h^n) \qquad (۶.۷.۵)$$

بنابراین

$$f'(x) - p'_n(x) = \begin{cases} O(h^n) & \Psi'_n(x) \neq 0 \\ O(h^{n+1}) & \Psi'_n(x) = 0 \end{cases} \qquad (۷.۷.۵)$$

اکنون مثالی برای هر حالت به دست می‌آوریم.

گیریم $n = 1$، پس $p_n(x)$ درست همان درونیابی خطی $(x_0, f(x_0))$ و $(x_1, f(x_1))$ خواهد بود. در این صورت رابطهٔ (۳.۷.۵) چنین به‌دست می‌دهد

$$f'(x_0) \doteq D_h f(x_0) \equiv \frac{1}{h}[f(x_0 + h) - f(x_0)] \qquad (۸.۷.۵)$$

از (۵.۷.۵)،

$$f'(x_0) - D_h f(x_0) = \frac{h}{2} f''(\xi_1) \qquad x_0 \leq \xi_1 \leq x_1 \qquad (۹.۷.۵)$$

زیرا $\Psi(x_0) = 0$.

برای بهبود این وضعیت با درونیابی خطی، انتخاب می‌کنیم $x = m \equiv (x_0 + x_1)/2$. در این صورت

$$f'(m) \doteq \frac{1}{h}[f(x_1) - f(x_0)]$$

این فرمول را با قراردادن $\delta = h/2$ بازنویسی می‌کنیم، تا به دست آوریم

$$f'(m) \doteq D_\delta f(m) = \frac{1}{2\delta}[f(m + \delta) - f(m - \delta)] \qquad (۱۰.۷.۵)$$

برای خطا، با استفاده از (۵.۷.۵) و $\Psi_{۲n}(m) = ۰$،

$$f'(m) - D_\delta f(m) = \frac{-\delta^۲}{۶} f^{(۳)}(\xi_۲) \qquad m - \delta \le \xi_۲ \le m + \delta \quad (۱۱.۷.۵)$$

به‌طورکلی، برای به‌دست آوردن مرتبهٔ بالاترِ حالت (۷.۷.۵)، می‌خواهیم نقاط گرهی $\{x_i\}$ را طوری انتخاب کنیم که $\Psi'_n(x) = ۰$. این درست خواهد بود اگر n فرد باشد و گره‌ها مانند (۱۰.۷.۵)، نسبت به x قرینه باشند.

برای پیدا کردن فرمولهایی از مراتب بالاتر وقتی که نقاط گرهی تماماً در یک طرف x باشند مقادیر بزرگتری از n را در (۳.۷.۵) به‌کار می‌بریم. برای مثال با $x = x_۰$ و $n = ۲$ داریم،

$$f'(x_۰) \doteq D_h f(x_۰) = \frac{۱}{۲h}[-۳f(x_۰) + ۴f(x_۱) - f(x_۲)] \qquad (۱۲.۷.۵)$$

$$f'(x_۰) - D_h f(x_۰) = \frac{h^۲}{۳} f^{(۳)}(\xi_۱) \qquad x_۰ \le \xi_۱ \le x_۲ \qquad (۱۳.۷.۵)$$

روش ضرایب نامعین یک روش دیگر برای پیدا کردن فرمولهایی برای انتگرالگیری عددی، مشتقگیری عددی و درونیابی، روش موسوم به ضرایب نامعین است. این روش اغلب هم‌ارز است با فرمولهای حاصل از فرمول درونیابی یک چندجمله‌یی، ولی گاهی به دست آوردن آن ساده‌تر است. ما این روش را با پیدا کردن فرمولی برای $f''(x)$ نشان خواهیم داد.

فرض می‌کنیم

$$f''(x) \doteq D_h^{(۲)} f(x) = Af(x+h) + Bf(x) + Cf(x-h) \qquad (۱۴.۷.۵)$$

که در آن A و B و C نامشخّص‌اند. به‌جای $f(x+h)$ و $f(x-h)$ بسطهای تیلر آنها را می‌گذاریم

$$f(x \pm h) = f(x) \pm hf'(x) + \frac{h^۲}{۲} f''(x) \pm \frac{h^۳}{۶} f^{(۳)}(x) + \frac{h^۴}{۲۴} f^{(۴)}(\xi_\pm)$$

که در آن $x - h \le \xi_- \le x \le \xi_+ \le x + h$. از قرار دادن این مقادیر در فرمول (۱۴.۷.۵) و مرتّب کردن برحسب توان h :

$$\begin{aligned} &Af(x+h) + Bf(x) + Cf(x-h) \\ &= (A+B+C)f(x) + h(A-C)f'(x) + \frac{h^۲}{۲}(A+C)f''(x) \\ &\quad + \frac{h^۳}{۶}(A-C)f^{(۳)}(x) + \frac{h^۴}{۲۴}[Af^{(۴)}(\xi_+) + Bf^{(۴)}(\xi_-)] \qquad (۱۵.۷.۵) \end{aligned}$$

برای آنکه این عبارت برابر $f''(x)$ شود، قرار می‌دهیم

$$A+B+C=۰ \qquad A-C=۰ \qquad A+C=\frac{۲}{h^۲}$$

جواب این دستگاه چنین است

$$A=C=\frac{۱}{h^۲} \qquad B=\frac{-۲}{h^۲} \qquad (۱۶.۷.۵)$$

که فرمول زیر را به دست می‌دهد

$$D_h^{(۲)}f(x)=\frac{f(x+h)-۲f(x)+f(x-h)}{h^۲} \qquad (۱۷.۷.۵)$$

برای خطا، (۱۶.۷.۵) را در (۱۵.۷.۵) قرار می‌دهیم و از (۱۷.۷.۵) استفاده می‌کنیم تا به‌دست آوریم

$$f''(x)-D_h^{(۲)}f(x)=\frac{h^۲}{۲۴}[f^{(۴)}(\xi_+)+f^{(۴)}(\xi_-)]$$

با استفاده از مسألهٔ ۱ فصل ۱ و با فرض اینکه $f(x)$ چهار بار پیوسته مشتقپذیر است

$$f''(x)-D_h^{(۲)}f(x)=-\frac{h^۲}{۱۲}f^{(۴)}(\xi) \qquad (۱۸.۷.۵)$$

به‌ازای $x-h\le\xi\le x+h$. فرمولهای (۱۷.۷.۵) و (۱۸.۷.۵) را می‌توانستیم از محاسبهٔ $p_۴''(x)$ درونیاب چندجمله‌یی درجهٔ دوم $f(x)$ در $x+h$، x، $x-h$ به‌دست آوریم ولی روش فوق احتمالاً ساده‌تر است.

نظر کلی در روش ضرایب نامعین آن است که ضرایب بسط تیلر را بر حسب h به‌گونه‌ای پیدا کنیم که مشتق (یا انتگرال) مورد نظر تا حد ممکن به مقدار حقیقی نزدیک باشد.

اثر خطا در مقادیر تابع فرمولهای قبلی برای پیدا کردن روشهای حل معادلات دیفرانسیل معمولی و معادلات دیفرانسیل جزیی مفیدند، ولی اگر برای مقادیر تابع که به‌صورت تجربی به‌دست آمده‌اند به‌کار روند ممکن است خطاهای جدی ایجاد نمایند. برای نشان‌دادن روشی که اثرات چنین خطاهایی را تحلیل نماید تقریب مشتق دوم (۱۷.۷.۵) را درنظر می‌گیریم.

مطلب را با بازنویسی (۱۷.۷.۵) به شکل زیر آغاز می‌کنیم

$$f''(x_۱)\doteq D_h^{(۲)}f(x_۱)=\frac{f(x_۲)-۲f(x_۱)+f(x_۰)}{h^۲}$$

که $x_j = x_0 + jh$. گیریم که مقادیر واقعی که به‌کار رفته‌اند $\tilde{f}_i$ باشند با

$$f(x_i) = \tilde{f}_i + \varepsilon_i \qquad i = 0, 1, 2 \qquad (۱۹.۷.۵)$$

مشتق عددی واقعی محاسبه شده چنین است

$$\tilde{D}_h^{(2)} f(x_1) = \frac{\tilde{f}_2 - 2\tilde{f}_1 + \tilde{f}_0}{h^2} \qquad (۲۰.۷.۵)$$

برای خطای آن (۱۹.۷.۵) را در (۲۰.۷.۵) می‌گذاریم و به‌دست می‌آوریم

$$f''(x_1) - \tilde{D}_h^{(2)} f(x_1) = f''(x_1) - \frac{f(x_2) - 2f(x_1) + f(x_0)}{h^2} + \frac{\varepsilon_2 - 2\varepsilon_1 + \varepsilon_0}{h^2}$$
$$= \frac{-h^2}{12} f^{(4)}(\xi) + \frac{\varepsilon_2 - 2\varepsilon_1 + \varepsilon_0}{h^2} \qquad (۲۱.۷.۵)$$

برای عبارت شامل $\{\varepsilon_i\}$، فرض کنید که این خطاها در بازهٔ $-E \le \varepsilon \le E$، تصادفی باشند. در این صورت

$$| f''(x_1) - \tilde{D}_h^{(2)} f(x_1) | \le \frac{h^2}{12} | f^{(4)}(\xi) | + \frac{4E}{h^2} \qquad (۲۲.۷.۵)$$

و این کران اخیر در بسیاری از مواد به‌دست می‌آید. یک مثال برای چنین خطاهایی، خطاهای گردکردن است که E کران اندازهٔ آنهاست.

کران خطا در (۲۲.۷.۵) در ابتدا با کاهش h کوچکتر می‌شود ولی برای h به اندازهٔ کافی نزدیک به صفر، خطا دوباره شروع به افزایش می‌کند. یک مقدار بهینه برای h وجود دارد که آن را h^* می‌نامیم که سمت راست (۲۲.۷.۵) را مینیمم می‌کند، و احتمالاً مقدار مشابهی نیز برای خطای واقعی $f''(x_1) - \tilde{D}_h^{(2)} f(x_1)$ وجود خواهد داشت.

مثال گیریم $f(x) = -\cos(x)$، با استفاده از تقریب عددی (۱۷.۷.۵)، $f''(0)$ را محاسبه می‌کنیم. در جدول ۲۵.۵، دوستون خطا داده‌ایم، خطاهای (۱) $D_h^{(2)} f(0)$ که دقیقاً محاسبه شده است؛ (۲) خطاهای $\tilde{D}_h^{(2)} f(0)$ که با حساب اعشاری هشت رقمی محاسبه وگرد شده‌است. در این حالت اخیر،

$$| f''(0) - \tilde{D}_h^{(2)} f(0) | \le \frac{h^2}{12} + \frac{2 \times 10^{-8}}{h^2} \qquad (۲۳.۷.۵)$$

کران (۲۳.۷.۵) در $h^* = 0.0022$ مینیمم می‌شود که با خطاهای $f''(0) - \tilde{D}_h^{(2)} f(0)$ که در جدول داده شده‌اند سازگار است. برای $D_h^{(2)} f(0)$ که دقیقاً محاسبه شده است، توجه کنید که وقتی h نصف می‌شود، خطاها با نسبت ۱/۴ کوچک می‌شوند که با فرمول خطای (۱۸.۷.۵) سازگار است.

جدول **۲۵.۵** مثال $D_h^{(۲)} f(۰)$ و $\tilde{D}_h^{(۲)}$

h	$f''(۰) - D_h^{(۲)} f(۰)$	نسبت	$f''(۰) - \tilde{D}_h^{(۲)} f(۰)$
۰٫۵	۲٫۰۷E − ۲		۲٫۰۷E − ۲
۰٫۲۵	۵٫۲۰E − ۳	۳٫۹۸	۵٫۲۰E − ۳
۰٫۱۲۵	۱٫۳۰E − ۳	۳٫۹۹	۱٫۳۰E − ۳
۰٫۰۶۲۵	۳٫۲۵E − ۴	۴٫۰۰	۳٫۲۵E − ۴
۰٫۰۳۱۲۵	۸٫۱۴E − ۵	۴٫۰۰	۸٫۴۵E − ۵
۰٫۰۱۵۶۲۵	۲٫۰۳E − ۵	۴٫۰۰	۲٫۵۶E − ۶
۰٫۰۰۷۸۱۲۵	۵٫۰۹E − ۶	۴٫۰۰	−۷٫۹۴E − ۵
۰٫۰۰۳۹۰۶۲۵	۱٫۲۷E − ۶	۴٫۰۰	−۷٫۹۴E − ۵
۰٫۰۰۱۹۵۳۱۲۵	۳٫۱۸E − ۷	۴٫۰۰	−۱٫۳۹E − ۳

بحث در آثار خواندنی

اگر چه مبحث انتگرالگیری عددی یکی از قدیمیترین مباحث آنالیز عددی است و نوشته‌های زیادی در مورد آنها موجود است، مقالات جدیدی به مقدار نسبتاً زیاد مرتباً منتشر می‌شوند. بسیاری از این مقاله‌ها روشهایی را برای رده‌هایی خاص از مسائل، مثل انتگرالهای نوسانی، مطرح می‌کنند، و عدّهٔ دیگر واکنشی به تغییرات در رایانه‌ها، مثل استفاده از معماریهای خط لولهٔ بُرداری هستند. بهترین بررسی انتگرالگیری عددی کار وسیع و مشروح دیویس و رابینوویتس (۱۹۸۴) است. این اثر شامل بررسی فراگیر از بیشتر روشهای انتگرالگیری، فهرست منابع جامع، مجموعه‌ای از برنامه‌های رایانه‌یی و فهرست برنامه‌های انتگرالگیری منتشر شده، است. همچنین شامل مقالهٔ آبراموویتس دربارهٔ «محاسبه عملی انتگرالها» است، که توصیه‌های بسیار عالی در روشهای تحلیلی انتگرالگیری به‌دست می‌دهد. کتابهای درسی مهم دیگر در مورد انتگرالگیری عددی عبارت‌اند از ، انگلس (۱۹۸۰)، کریلوف (۱۹۶۲)، استرود (۱۹۷۱). برای تاریخچهٔ روشهای انتگرالگیری عددی کلاسیک، گولدستاین (۱۹۷۷) را ببینید.

به علّت کمی جا، مجبور بوده‌ایم بعضی از مفاهیم مهم را حذف کنیم. مهمترین آنها عبارت‌اند از (۱) انتگرالگیری کلنشار–کرتیس[1] و (۲) انتگرالگیری چند متغیره. اولی مبتنی است بر انتگرالگیری بسط چبیشف انتگرالده: که به تجربه ثابت شده است که برای انواع بسیاری از انتگرالها روشی عالی است. این روش ابتکاری درکلنشار–کرتیس (۱۹۶۰) عرضه شده است. یک شرح معمولی از آن در (صفحات ۲۸ تا ۲۹) پیسنز و همکاران (۱۹۸۳) داده شده است. زمینهٔ انتگرالگیری

1. Clenshaw-Curits

چند متغیره یک زمینهٔ بسیار فعال پژوهشی است، و کتابهای درسی انگلس (۱۹۸۰) و استرود (۱۹۷۱) بهترین مدخل بر این زمینه‌اند. به لحاظ استفادهٔ گسترده از انتگرالگیری چند متغیره در روش عنصر متناهی برای حل معادلات دیفرانسیل جزئی، کتابهای روش عنصر متناهی اغلب شامل فرمولهای انتگرالگیری برای ناحیه‌های مثلثی و مستطیلی هستند.

انتگرالگیری عددی خودکار ـ وقتی احساس شد که اغلب انتگرالگیریهای عددی را می‌توان از این طریق محاسبه کرد ـ یک زمینهٔ بسیار فعال پژوهشی در دهه‌های ۱۹۶۰ و ۱۹۷۰ بود. اخیراً یک بازگشت به سوی استفادهٔ وسیع از انتگرالگیریهای غیر خودکار، به ویژه برای انتگرالهای دم‌دستی، صورت گرفته است. یک بحث عالی دربارهٔ مزایا و معایب نسبی انتگرالگیری خودکار در لاینس (۱۹۸۳) داده شده است. تواناترین و انعطاف‌پذیرترین برنامه‌های خودکار امروزی شاید آنهایی باشند که در QUADPACK داده شده‌اند، این برنامه‌ها را پیسنز و همکاران (۱۹۸۳) بحث و روشن کرده‌اند. نسخه‌هایی از QUADPACK در IMSL و NAG موجود است.

برای ریزرایانه‌ها و محاسبات دستی، قاعدهٔ سیمپسون، به دلیل سادگی، هنوز خیلی طرفدار دارد. مع‌هذا عنایت خاصی بایستی به روش گاوس به دلیل دقت بالاتر آن بشود. گره‌ها و وزنها در آبراموویتس و اشتگون (۱۹۶۴) و در استرود و سکرت (۱۹۶۶) اکنون در دسترس قرار داده شده‌اند و برنامه‌هایی برای محاسبهٔ آنها مهیا شده‌اند.

مشتقگیری عددی به تعبیر بخش ۶.۱ یک مسألهٔ بد حالت است. روندهای مشتقگیری عددی که در آنها این نکته درنظر گرفته شده است در ده پانزده سال گذشته توسعه یافته‌اند. به ویژه، اندرسون و بلوم‌فیلد (۱۹۷۴a) (۱۹۷۴b)، کالوم[1] (۱۹۷۲) وابا[2] (۱۹۸۰)، وُلترینگ[3] (۱۹۸۶) را ببینید.

مراجع

Abramowitz, M., and I. Stegun, Eds. (1964). *Handbook of Mathematical Functions*. National Bureau of Standards, U.S. Government Printing Office, Washington, D.C.

Anderssen, R., and P. Bloomfield (1974a). Numerical differentiation procedures for non-exact data, *Numer. Math.*, 22, 157–182.

Anderssen, R., and P. Bloomfield (1974b). A time series approach to numerical differentiation, *Technometrics* 16, 69–75.

Atkinson, K. (1976). *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind*. Society for Industrial and Applied Mathematics, Philadelphia.

1. Cullum (1971) 2. Wahba 3. Woltring

Atkinson, K. (1982). Numerical integration on the sphere, *J. Austr. Math. Soc. (Ser. B)* **23**, 332–347.

Bauer, F., H. Rutishauser, and E. Stiefel (1963). New aspects in numerical quadrature. In *Experimental Arithmetic, High Speed Computing, and Mathematics*, pp. 199–218. Amer. Math. Soc., Providence, R.I.

de Boor, C. (1971). CADRE: An algorithm for numerical quadrature. In *Mathematical Software*, pp. 201–209. Academic Press, New York.

Clenshaw, C., and A. Curtis (1960). A method for numerical integration on an automatic computer, *Numer. Math.* **2**, 197–205.

Cryer, C. (1982). *Numerical Functional Analysis*. Oxford Univ. Press (Clarendon), Oxford, England.

Cullum, J. (1971). Numerical differentiation and regularization, *SIAM J. Numer. Anal.* **8**, 254–265.

Davis, P. (1963). *Interpolation and Approximation*. Ginn (Blaisdell), Boston.

Davis, P., and P. Rabinowitz (1984). *Methods of Numerical Integration*, 2nd ed. Academic Press, New York.

Dixon, V. (1974). Numerical quadrature: A survey of the available algorithms. In *Software for Numerical Mathematics*, D. Evans, Ed., pp. 105–137. Academic Press, London.

Donaldson, J., and D. Elliott (1972). A unified approach to quadrature rules with asymptotic estimates of their remainders, *SIAM J. Numer. Anal.* **9**, 573–602.

de Doncker, E., and R. Piessens (1976). Algorithm 32: Automatic computation of integrals with singular integrand, over a finite or an infinite interval, *Computing* **17**, 265–279.

Engels, H. (1980). *Numerical Quadrature and Cubature*. Academic Press, New York.

Goldstine, H. (1977). *A History of Numerical Analysis*. Springer-Verlag, New York.

Golub, G., and J. Welsch (1969). Calculation of Gauss quadrature rules, *Math. Comput.* **23**, 221–230.

de Hoog, F., and R. Weiss (1973). Asymptotic expansions for product integration, *Math. Comput.* **27**, 295–306.

Iri, M., S. Moriguti, and Y. Takasawa (1970). On a numerical integration formula (in Japanese), J. Res. Inst. *Math. Sci.*, **91**, 82. Kyoto Univ., Kyoto, Japan.

Isaacson, E., and H. Keller (1966). *Analysis of Numerical Methods*. Wiley, New York.

Kronrod, A. (1965). *Nodes and Weights of Quadrature Formulas*, Consultants Bureau, New York.

Krylov, V. (1962). *Approximate Calculation of Integrals*. Macmillan, New York.

Lyness, J. (1983). When not to use an automatic quadrature routine, *SIAM Rev.* **25**, 63–88.

Lyness, J., and C. Moler (1967). Numerical differentiation of analytic functions, *SIAM J. Num. Anal.* **4**, 202–210.

Lyness, J., and J. Kaganove (1976). Comments on the nature of automatic quadratic routines, *ACM Trans. Math. Softw.* **2**, 65–81.

Lyness, J., and K. Puri (1973). The Euler-MacLaurin expansion for the simplex, *Math. Comput.* **27**, 273–293.

Lyness, J., and B. Ninham (1967). Numerical quadrature and asymptotic expansions, *Math. Comput.* **21**, 162–178.

Patterson, T. (1968). The optimum addition of points to quadrature formulae, *Math. Comput.* **22**, 847–856. (Includes a microfiche enclosure; a microfiche correction is in **23**, 1969.)

Patterson, T. (1973) Algorithm 468: Algorithm for numerical integration over a finite interval. *Commun. ACM* **16**, 694–699.

Piessens, R., E. deDoncker-Kapenga, C. Überhuber, and D. Kahaner (1983). *QUADPACK: A Subroutine Package for Automatic Integration*. Springer-Verlag, New York.

Ralston, A. (1965). *A First Course in Numerical Analysis*. McGraw-Hill, New York.

Robinson, I. (1979). A comparison of numerical integration programs, *J. Comput. Appl. Math.* **5**, 207–223.

Robinson, I., and E. de Doncker (1981). Automatic computation of improper integrals over a bounded or unbounded planar region, *Computing* **27**, 253–284.

Stenger, F. (1981). Numerical methods based on Whittaker cardinal or sinc functions, *SIAM Rev.* **23**, 165–224.

Stroud, A. (1971). *Approximate Calculation of Multiple Integrals*. Prentice-Hall, Englewood Cliffs, N.J.

Stroud, A., and D. Secrest (1966). *Gaussian Quadrature Formulas*. Prentice-Hall, Englewood Cliffs, N.J.

Wahba, G. (1980). Ill-posed problems: Numerical and statistical methods for mildly, moderately, and severely ill-posed problems with noisy data, Tech. Rep. #595, Statistics Dept., Univ. of Wisconsin, Madison. (Prepared for the Proc. Int. Symp. on Ill-posed Problems, Newark, Del., 1979.)

Woltring, H. (1986). A Fortran package for generalized, cross-validatory spline smoothing and differentiation, *Adv. Eng. Softw.* **8**, 104–113.

مسائل

۱. برنامه‌ای بنویسید که با استفاده از قاعدهٔ ذوزنقه‌یی و n تقسیم جزیی، $I = \int_a^b f(x)dx$ را محاسبه نماید. نتیجه را I_n بنامید. این برنامه را برای محاسبهٔ انتگرالهای زیر با $n = ۲, ۴, ۸, ۱۶, \ldots, ۵۱۲$ به‌کار برید

(الف) $\int_۰^۱ e^{-x^۲}dx$ (ب) $\int_۰^۱ x^{۵/۲}dx$ (ج) $\int_{-۴}^{۴} \frac{dx}{۱+x^۲}$

(د) $\int_۰^{۲\pi} \frac{dx}{۲+\cos(x)}$ (ه) $\int_۰^{\pi} e^x \cos(۴x)dx$

نرخ همگرایی I_n به I را به‌صورت تجربی با محاسبهٔ نسبتهای (۲۷.۴.۵) تحلیل کنید:

$$R_n = \frac{I_{۲n} - I_n}{I_{۴n} - I_{۲n}}$$

۲. مسألهٔ ۱ را با استفاده از قاعدهٔ سیمپسون تکرار کنید.

۳. قاعدهٔ ذوزنقه‌یی تصحیح شدهٔ (۱۲.۱.۵) را برای انتگرالهای مسألهٔ ۱ به‌کار برید. نتایج را با نتایج مسألهٔ ۲ که با قاعدهٔ سیمپسون به‌دست آمده است مقایسه کنید.

۴. به‌عنوان روش دیگر برای قاعدهٔ ذوزنقه‌یی تصحیح شدهٔ (۱۲.۱.۵)، چندجمله‌یی درجهٔ سوّم درونیابی ارمیت را برای $f(x)$ به‌کار برید تا

$$\int_a^b f(x)dx \doteq \left(\frac{b-a}{۲}\right)[f(a)+f(b)] - \frac{(b-a)^۲}{۱۲}[f'(b)-f'(a)]$$

را به‌دست آورید. فرمول خطا برای درونیابی ارمیت را به‌کار برید تا یک فرمول خطا برای تقریب پیشین پیدا شود. این نتایج را برای (۱۲.۱.۵)، با تقسیم $[a,b]$ به n جزء، تعمیم دهید.

۵. (الف) فرض کنید که در بازهٔ $[۰,b]$، $f(x)$ پیوسته و $f'(x)$ انتگرالپذیر باشد. نشان دهید که خطا در قاعدهٔ ذوزنقه‌یی برای محاسبهٔ $\int_۰^۱ f(x)dx$ به‌شکل زیر است

$$E_n(f) = \int_۰^۱ K(t)f'(t)dt$$

$$K(t) = \frac{t_{j-۱}+t_j}{۲} - t \qquad t_{j-۱} \le t \le t_j \qquad j = ۱, \ldots n$$

این نتایج را با (۲۲.۱.۵) که در آن $f'(x)$ پیوسته و $f''(x)$ انتگرالپذیر است، مقایسه کنید.

(ب) نتیجه را برای $f(x) = x^\alpha$ و $f(x) = x^\alpha \log(x)$، $۰ < \alpha < ۱$ به‌کار برید. این یک مرتبهٔ همگرایی به‌دست می‌دهد، اگرچه کوچکتر از مرتبهٔ واقعی است [مسألهٔ ۶ و (۲۳.۴.۵) را ببینید].

۶. با استفاده از برنامهٔ مسألهٔ ۱ نرخ همگرایی قاعدهٔ ذوزنقه‌یی را به‌طور تجربی برای $\int_۰^۱ x^\alpha \ln(x)dx$ در حالت $۰ \le \alpha \le ۱$ برای $\alpha = ۰٫۲۵, ۰٫۵, ۰٫۷۵, ۱٫۰$ به‌دست آورید.

۷. شکل مرکب قاعدهٔ بول را که در جدول ۸.۵ در درایهٔ $n = ۴$ داده شده‌است به‌دست آورید. فرمولهای خطایی مشابه فرمولهایی که در (۱۷.۱.۵) و (۱۸.۱.۵) برای قاعدهٔ سیمپسون داده شده‌اند پیدا کنید.

۸. مسألهٔ ۱ را با استفاده از قاعدهٔ بول که در مسألهٔ ۷ به‌دست آمده تکرار کنید.

۹. گیریم $p_۲(x)$ چندجمله‌یی درجهٔ دومی باشد که $f(x)$ را در $x = ۰, h, ۲h$، درونیابی می‌کند. با استفاده از این درونیاب یک فرمول انتگرالگیری عددی I_h برای $I = \int_۰^{۳h} f(x)dx$ پیدا کنید بسط یک سری تیلر $f(x)$ را به‌کار برید تا نشان دهید که

$$I - I_h = \frac{۳}{۸}h^۴ f^{(۳)}(۰) + O(h^۵)$$

۱۰. برای فرمول انتگرالگیری میانگاهی (۱۷.۲.۵)، فرمول خطای هسته‌یی پئانو زیر را به‌دست آورید

$$\int_۰^h f(x)dx - hf\left(\frac{h}{۲}\right) = \int_o^h K(t)f''(t)dt$$

$$K(t) = \begin{cases} \frac{۱}{۲}t^۲ & ۰ \le t \le \frac{h}{۲} \\ \frac{۱}{۲}(h-t)^۲ & \frac{h}{۲} \le t \le h \end{cases}$$

با استفاده از این فرمول جملهٔ خطا در (۱۷.۲.۵) را پیدا کنید.

۱۱. توابعی را که به طریق زیر تعریف شده‌اند درنظر می‌گیریم. گیریم $n > ۰$ و $h = (b-a)/n$، $t_j = a + jh$ برای $j = ۰, ۱, \ldots, n$. گیریم $f(x)$ در هر زیر بازهٔ $[t_{j-۱}, t_j]$ به‌ازای $j = ۱, \ldots, n$ خطی باشد. نشان دهید که مجموعهٔ همهٔ چنین fهایی، برای مقادیر $n \ge ۱$ در $C[a,b]$ چگال است.

۱۲. گیریم $w(x)$ یک تابع انتگرالپذیر بر بازهٔ $[a,b]$ باشد،

$$\int_a^b \mid w(x) \mid dx < \infty$$

و گیریم

$$\int_a^b w(x)f(x)dx \doteq \sum_{j=۱}^n w_{j,n} f(x_{j,n})$$

دنباله‌ای از قاعده‌های انتگرالگیری عددی باشد. قضیهٔ ۲.۵ را برای این حالت تعمیم دهید.

۱۳. فرض کنید قاعدهٔ انتگرالگیری عددی

$$\int_a^b f(x)dx \doteq \sum_{j=۱}^n w_{j,n} f(x_{j,n})$$

برای کلیهٔ توابع پیوسته همگرا باشد. اثر خطاها را در مقادیر تابع درنظر بگیرید. فرض کنید که از $\tilde{f}_i \doteq f(x_i)$ با

$$| f(x_i) - \tilde{f}_i | \leq \varepsilon \qquad ۱ \leq i \leq n$$

استفاده می‌کنیم. تأثیر این خطاها در مقادیر تابع، بر این انتگرالگیری عددی چه خواهد بود؟

۱۴. انتگرالگیری گاوس ـ لژاندر را برای انتگرالهای مسألهٔ ۱ به‌کار برید. نتایج را با نتایج قاعده‌های ذوزنقه‌یی و سیمپسون مقایسه کنید.

۱۵. انتگرالگیری گاوس ـ لژاندر را برای محاسبهٔ $\int_{-۴}^{۴} dx/(۱+x^۲)$ با فرمول $n = ۲, ۴, ۶, ۸$ نقطهٔ گرهی به‌کار برید. نتایج را با نتایج جدول ۹.۵ که از فرمول نیوتن ـ کوتس حاصل شده است مقایسه کنید.

۱۶. ثابت کنید که گره‌ها و وزنهای گاوس ـ لژاندر در $[-۱, ۱]$ به‌طور متقارن حول $x = ۰$ توزیع شده‌اند.

۱۷. فرمول: انتگرالگیری گاوسی را برای

$$I(f) = \int_{۰}^{۱} f(x) \log\left(\frac{۱}{x}\right) dx$$

که در آن تابع وزن $w(x) = \log(۱/x)$ است پیدا کنید. راهنمایی: مسألهٔ ۲۰ (الف) از فصل ۴ را ببینید. همچنین از مشابه (۷.۳.۵) برای محاسبهٔ وزنها استفاده کنید نه از طریق فرمول (۱۱.۳.۵).

۱۸. فرمولهای انتگرالگیری گاوسی یک نقطه‌یی و دو نقطه‌یی را برای

$$I = \int_{۰}^{۱} x f(x) dx \doteq \sum_{j=۱}^{n} w_j f(x_j)$$

با تابع وزن $w(x) = x$ به دست آورید.

۱۹. برای انتگرال $I = \int_{-۱}^{۱} \sqrt{۱-x^۲} f(x) dx$ با وزن $w(x) = \sqrt{۱-x^۲}$، فرمولهای صریح برای وزنها و گره‌های فرمول انتگرالگیری گاوسی به‌دست آورید. همچنین فرمول خطا را پیدا کنید. راهنمایی: مسألهٔ ۲۴ فصل ۴ را ببینید.

۲۰. با استفاده از ستون $e_{n,۴}$ جدول ۱۳.۵، فرمول خطای مرتبهٔ ۴ مشابه فرمول خطای مرتبهٔ ۲ (۴۰.۳.۵) را برای $e_{n,۲}$، پیدا کنید. آن را با فرمول خطای مرتبهٔ چهار (۱۷.۱.۵) سیمپسون مقایسه کنید.

۲۱. در فرمول (۴۸.۳.۵) کرونرود[1] وزنها را از حل دستگاه چهار معادلهٔ خطی، می‌توان محاسبه کرد. این دستگاه را پیدا و حل کنید تا مقادیر داده شدهٔ بعد از (۴۸.۳.۵) تأیید شود. راهنمایی: طریقی را که به (۷.۳.۵) انجامید به‌کار برید.

---

۱. Kronrod

۲۲. فرمول هفت نقطه‌یی گاوس ـ لژاندر را با فرمول هفت نقطه‌یی (۴۸.۳.۵) کرونرود مقایسه کنید. هر یک از آنها را برای انتگرالهای گوناگون به کار برید و خطاهای متناظر آنها را مقایسه کنید.

۲۳. (الف) رابطهٔ (۷.۴.۵) را برای چندجمله‌ییهای $B_j(x)$ برنولی و اعداد B_j برنولی پیدا کنید. نشان دهید که برای تمام اعداد صحیح فرد $j \geq ۳$، $B_j = ۰$.

(ب) اتحادهای زیر را به‌دست آورید

$$B'_j(x) = jB_{j-۱}(x) \qquad j \geq ۴ \text{ و زوج}$$

$$B'_j(x) = j[B_{j-۱}(x) + B_{j-۱}] \qquad j \geq ۳ \text{ و فرد}$$

این اتحادها را می‌توان برای یک اثبات کلی فرمول اویلر ـ مک لورن (۹.۴.۵) به‌کار گرفت.

۲۴. با استفاده از فرمول مجموعیابی اویلر ـ مک لورن (۱۷.۴.۵)، برآوردی برای $\zeta(\frac{۵}{۴})$ با دقتی تا سه رقم اعشار پیدا کنید. $\zeta(p)$، تابع زتا در (۲۲.۴.۵) تعریف شده است.

۲۵. فرمول خطای مجانبی برای قاعدهٔ ذوزنقه‌یی را که برای $\int_۰^۱ \sqrt[۳]{x} f(x)dx$ به‌کار رفته پیدا کنید. برآورد در مسألهٔ ۲۴ را به‌کار برید.

۲۶. جدول انتگرالهای تقریبی I_n در زیر را که با استفاده از قاعدهٔ سیمپسون به‌دست آمده‌اند درنظر می‌گیریم. مرتبهٔ همگرایی I_n به I را پیش‌بینی کنید.

n	I_n
۲	۰٫۲۸۴۵۱۷۷۹۶۸۶
۴	۰٫۲۸۵۵۹۲۵۴۵۷۶
۸	۰٫۲۸۵۷۰۲۴۸۷۴۸
۱۶	۰٫۲۸۵۷۱۳۱۷۷۳۱
۳۲	۰٫۲۸۵۷۱۴۱۸۳۶۳
۶۴	۰٫۲۸۵۷۱۴۲۷۶۴۳

یعنی اگر $I - I_n \doteq c/n^p$، p چیست؟ آیا به‌نظر می‌آید که این شکل معتبری برای خطای این داده‌ها باشد؟ مقداری برای c و خطای $I_{۶۴}$ پیش‌بینی کنید. n به چه بزرگی باید انتخاب شود اگر بخواهیم I_n خطایی کوچکتر از $۱۰^{-۱۱}$ داشته باشد؟

۲۷. فرض کنید که خطا در یک فرمول انتگرالگیری بسط مجانبی زیر را داشته باشد

$$I - I_n = \frac{C_۱}{n\sqrt{n}} + \frac{C_۲}{n^۲} + \frac{C_۳}{n^۲\sqrt{n}} + \frac{C_۴}{n^۳} + \cdots$$

فرآیند برونیابی ریچاردسن از بخش ۵.۴ را برای پیدا کردن فرمول‌هایی برای C_1 و C_2 تعمیم دهید. فرض کنیم سه مقدار I_n، I_{2n}، I_{4n} محاسبه شده باشند. این مقادیر را برای محاسبهٔ C_1 و C_2 و برآوردی از I باخطایی از مرتبهٔ $1/n^2\sqrt{n}$ به‌کار برید.

۲۸. برای قاعدهٔ ذوزنقه‌یی (که با $I_n^{(T)}$ نمایش داده شده) در محاسبهٔ $I = \int_a^b f(x)dx$، فرمول خطای مجانبی را به شکل زیر داریم

$$I - I_n^{(T)} = -\frac{h^2}{12}[f'(b) - f'(a)] + O(h^4)$$

و برای فرمول میانگاهی $I_n^{(M)}$ داریم

$$I - I_n^{(M)} = \frac{h^2}{24}[f'(b) - f'(a)] + O(h^4)$$

به شرط آنکه f به اندازهٔ کافی در $[a, b]$ مشتقپذیر باشد. با استفاده از این نتایج، از ترکیب $I_n^{(T)}$ و $I_n^{(M)}$ یک فرمول انتگرالگیری عددی جدید $\tilde{I}_n$ با درجهٔ همگرایی بالاتر به‌دست آورید. وزنهای فرمول جدید $\tilde{I}_n$ را بنویسید.

۲۹. یک فرمول خطای مجانبی برای قاعدهٔ سیمپسون، قابل مقایسه با فرمول اویلر ـ مک لورن (۹.۴.۵) برای قاعدهٔ ذوزنقه‌یی، به‌دست آورید. از (۹.۴.۵)، (۳۳.۴.۵) و (۳۶.۴.۵) به همان‌گونه که در (۳۷.۴.۵) عمل کردیم، استفاده کنید.

۳۰. نشان دهید که فرمول (۴۰.۴.۵) برای $I_n^{(2)}$ قاعدهٔ مرکب بول[1] است. مسألهٔ ۷ را ببینید.

۳۱. الگوریتم رامبرگ بخش ۴.۵ را اجرا کنید و سپس آن را برای انتگرالهای مسألهٔ ۱ به‌کار برید. نتایج را با نتایج قاعده‌های سیمپسون و ذوزنقه‌یی مقایسه کنید.

۳۲. انتگرال $I = \int_0^1 x^\alpha dx$ را با استفاده از قاعدهٔ تطبیقی سیمپسون که به دنبال (۳.۵.۵) شرح داده شده بود محاسبه کنید. برای اینکه ببینید آزمون (۵.۵.۵) ممکن است به شکست انجامد حالت $-1 < \alpha < 0$ را درنظر بگیرید که انتگرالده به‌طور دلخواه در $x = 0$ برابر صفر گرفته شده است. نشان دهید که برای مقادیر به اندازهٔ کافی کوچک ε، آزمون (۵.۵.۵) در زیر بازهٔ $[0, h]$ هرگز صدق نمی‌کند. توجه داشته باشید که برای ε مشخص در $[0, 1]$، تحمل خطا برای $[0, h]$ برابر εh خواهد بود.

۳۳. قاعدهٔ سیمپسون با فاصله‌گذاری برابر را برای انتگرالگیری $\int_0^1 \log(x)dx$ به‌کار برید. به‌ازای $x = 0$، انتگرالده را برابر صفر بگیرید. نتایج را با آنچه در جداول (۱۹.۵) و (۲۰.۵) برای I_4 داده شده مقایسه نمایید.

---

1. Bool

۳۴. از یک برنامهٔ انتگرالگیری تطبیقی [مثل DQAGP از پیسنز و همکاران (۱۹۸۳)]برای محاسبهٔ انتگرالهای مسألهٔ ۱ استفاده کنید. نتایج را با نتایج مسائل ۱، ۲، ۱۴، ۳۱ مقایسه نمایید.

۳۵. رفتار تکین انتگرالده در

$$I = \int_0^1 f(x)\log(x)dx$$

را با استفاده از تبدیل متغیر $x = t^r$، $r > 0$، کاهش دهید. هموارى انتگرالده نتیجه را تحلیل کنید. همچنین رفتار تجربی قاعده‌های ذوزنقه‌یی و سیمپسون را برای مقادیر گوناگون r بررسی نمایید.

۳۶. روش IMT را، که به دنبال (۵.۶.۵) توضیح داده شده‌است، برای محاسبهٔ $I = \int_0^\infty f(x)dx$، با تبدیل متغیری از $(0, \infty]$ به یک بازهٔ متناهی، به‌کار برید. از همین روش برای محاسبهٔ انتگرالهایی که به دنبال (۱۱.۶.۵) آمده‌اند استفاده کنید.

۳۷. انتگرالگیری گاوس ـ لاگر با $n = 2, 4, 6, 8$ نقطهٔ گرهی را برای محاسبهٔ انتگرالهای زیر به‌کار برید. از فرمول (۱۱.۶.۵) استفاده کنید تا انتگرالها به شکل مناسب درآیند.

(الف) $\int_0^\infty e^{-x^2}dx = \frac{\sqrt{\pi}}{2}$ (ب) $\int_0^\infty \frac{xdx}{(1+x^2)^2} = \frac{1}{2}$

(ج) $\int_0^\infty \frac{\sin(x)}{x}dx = \frac{\pi}{2}$

۳۸. انتگرالهای تکین زیر را در حدود تحمل خطای مشخص شده محاسبه کنید. از برنامهٔ خودکار استفاده نکنید.

(الف) $\int_0^{4\pi} \cos(u)\log(u)du \qquad \varepsilon = 10^{-8}$

(ب) $\int_0^{2/\pi} x^2 \sin\left(\frac{1}{x}\right)dx \qquad \varepsilon = 10^{-3}$

(ج) $\int_0^1 \frac{\cos(x)}{\sqrt[3]{x^2}}dx \qquad \varepsilon = 10^{-8}$

(د) $\int_0^1 \frac{\sqrt{1-x^4}}{x^\alpha}dx \qquad \varepsilon = 10^{-5}, \quad \alpha = 1 - \frac{1}{\pi}$

۳۹. برای محاسبهٔ انتگرالهای مسألهٔ ۳۸ از یک برنامهٔ تطبیقی (مثلاً DQAGP) استفاده کنید.

۴۰. (الف) قاعدهٔ انتگرالگیری ذوزنقه‌یی حاصلضرب را برای محاسبهٔ

$$I(f) = \int_0^b \frac{f(x)}{x^\alpha}dx \qquad 0 < \alpha \le 1$$

بسط دهید. وزنها را به شکل مناسبی، مشابه با (۲۸.۶.۵) برای قاعدهٔ ذوزنقه‌یی حاصلضرب وقتی $w(x) = \log(x)$، انتخاب کنید. همچنین یک فرمول خطا مشابه با (۲۴.۶.۵) به‌دست آورید.

(ب) با استفاده از نتایج قسمت (الف) برای محاسبهٔ انتگرالهای زیر یک برنامهٔ ساده بنویسید

$$\int_{0}^{1} \frac{e^x}{x^{1/\pi}} dx \quad \textbf{(i)} \qquad\qquad \int_{0}^{1} \frac{dx}{\sin(x^{2/\pi})} \quad \textbf{(ii)}$$

راهنمائی: برای قسمت (ii) ابتدا $u = x^{2/\pi}$ بگیرید.

۴۱. برای نشان دادن بدحالتی مشتقگیری از یک تابع $y(t)$ در یک بازهٔ $[0, 1]$، محاسبهٔ $x(t) = y'(t)$ و $x_n(t) = y'_n(t)$ را، با

$$y_n(t) = y(t) + \frac{1}{n} t^n \qquad n \geq 1$$

درنظر می‌گیریم. تعریف بدحالت و خوش حالت را از بخش ۶.۱ به یاد آورید. با استفاده از ساخت بالا نشان دهید که محاسبهٔ $x(t)$ نسبت به تغییرات y ناپایدار است. برای اندازه‌گیری تغییرات در x و y از نرم ماکسیمم (۸.۱.۱) و (۴.۱.۴) استفاده کنید.

۴۲. با استفاده از رایانه یا ماشین حساب خود، مثال مشتقگیری عددی (۲۳.۷.۵) و جدول ۲۵.۵ را تکرار کنید.

۴۳. نتایج خطا برای $D_\delta f(m)$ در (۱۰.۷.۵) را مربوط به اثرات خطاهای گرد کردن، شبیه به نتایج (۲۱.۷.۵)ـ(۲۲.۷.۵) برای $\tilde{D}_h^{(2)} f(x)$، به‌دست آورید. آن را برای $f(x) = e^x$ در $x = 0$ به‌کار برید و نتایج را با نتایج واقعی محاسبات دستی خود مقایسه نمایید.

۶

روشهای عددی برای معادلات دیفرانسیل معمولی

معادلات دیفرانسیل یکی از مهمترین ابزارهای ریاضی هستند که در مسائل مدلسازی در علوم فیزیکی به‌کار برده می‌شوند. در این فصل ما روشهای عددی حل مسائل معادلات دیفرانسیل معمولی را به‌دست آورده و تحلیل می‌نماییم. شکل اصلی مسأله‌ای که ما مطالعه می‌کنیم، مسألهٔ مقدار اولیه است:

$$y' = f(x,y) \qquad y(x_0) = Y_0 \qquad (۱۰.۶)$$

تابع $f(x,y)$ باید به‌ازای همهٔ مقادیر (x,y) در حوزه‌ای چون D از صفحهٔ xy، پیوسته باشد و (x_0, Y_0) نقطه‌ای در D است. نتایجی که برای (۱۰.۶)به‌دست می‌آید، هم برای دستگاه معادلات دیفرانسیل و هم برای معادلات از مراتب بالاتر، به‌آسانی تعمیم داده می‌شوند، به شرطی که نمادهای برداری و ماتریسی مناسب به‌کارگرفته شوند. این تعمیمها در دو بخش آینده بحث و نشان داده شده‌اند.

گوییم $Y(x)$ یک جواب (۱۰.۶) در $[a,b]$ است اگر برای همهٔ مقادیر $a \le x \le b$،

۱. $(x, Y(x)) \in D$

۲. $Y(x_0) = Y_0$

۳. $Y'(x)$ موجود و $Y'(x) = f(x, Y(x))$.

در سرتاسر این فصل، نماد $Y(x)$ معرّف جواب درست هر مسألهٔ معادلهٔ دیفرانسیلی است که در نظر گرفته می‌شود.

مثال ۱. معادلهٔ کلی دیفرانسیل مرتبهٔ اول خطی به شکل زیر است

$$y' = a_0(x)y + g(x) \qquad a \le x \le b$$

که در آن ضرایب $a_0(x)$ و $g(x)$ بنا به فرض در $[a, b]$ پیوسته‌اند. حوزهٔ D برای این مسأله چنین است

$$D = \{(x, y) \mid a \le x \le b, \quad -\infty < y < \infty\}$$

جواب دقیق این معادله را در هرکتاب درسی مقدماتی معادلات دیفرانسیل [مثلاً بویس[1] و دیپریما[2] (۱۹۸۶)] می‌توان یافت. به عنوان یک حالت خاص، معادلهٔ زیر را در نظر می‌گیریم

$$y' = \lambda y + g(x) \qquad 0 \le x < \infty \qquad (۲.۰.۶)$$

با $g(x)$ پیوسته در $[0, \infty)$. جوابی که در $Y(0) = Y_0$ صدق می‌کند با رابطهٔ زیر داده می‌شود

$$Y(x) = Y_0 e^{\lambda x} + \int_0^x e^{\lambda(x-t)} g(t)dt \qquad 0 \le x < \infty \qquad (۳.۰.۶)$$

این رابطه بعداً برای نشان دادن نتایج نظری گوناگون به‌کار خواهد رفت.

۲. معادلهٔ $y' = -y^2$ غیرخطی است. $Y(x) \equiv 0$ یکی از جوابهای آن است، بقیهٔ جوابها به شکل

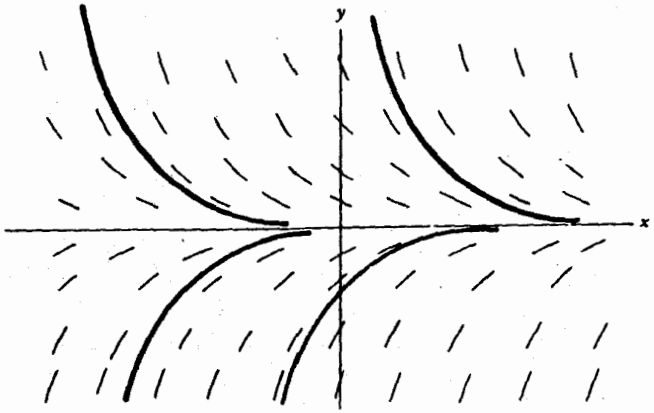
$$Y(x) = \frac{1}{x + c}$$

هستند، ثابت c اختیاری است. توجه کنید که $|Y(-c)| = \infty$. بنابراین همواری کلّی $f(x, y) = -y^2$، همواری جوابها را تضمین نمی‌کند.

برای آنکه شناختی هندسی از جوابهای یک معادلهٔ دیفرانسیل مرتبهٔ اوّل به‌دست آوریم، می‌توانیم به **میدان سوهایی** که این معادله در صفحهٔ xy القا می‌کند نظر افکنیم. اگر $Y(x)$ جوابی باشد که از (x_0, Y_0) بگذرد، آنگاه شیب $Y(x)$ در (x_0, Y_0) برابر است با $Y'(x_0) = f(x_0, Y_0)$. در حوزهٔ Dی $f(x, y)$ یک مجموعه نقاط (x, y) را گرفته یک پاره‌خط کوتاه با شیب $f(x, y)$ از هریک از نقاط (x, y) رسم می‌کنیم.

مثال معادله $y' = -y$ را در نظر بگیرید. میدان سوها در شکل ۱.۶ داده شده است، و چندین

1. Boyce 2. Diprima



شکل ۱.۶ میدان سوها برای $y' = -y$

جواب نمونه در آن رسم شده است. از نمودار روشن است که تمام جوابها در رابطهٔ زیر صدق می‌کنند

$$\lim_{x \to \infty} Y(x) = 0$$

برای آنکه ترسیم میدان سوهای $y' = f(x, y)$ را ساده‌تر کنیم، دنبال آن خمهایی از صفحهٔ xy می‌گردیم که در امتداد آنها $f(x, y)$ ثابت باشد. رابطهٔ

$$f(x, y) = c$$

را به‌ازای مقادیر مختلف c حل می‌کنیم. هر جواب $Y(x)$ معادلهٔ $y' = f(x, y)$ که منحنی $f(x, y) = c$ را قطع کند، در نقطهٔ برخورد در رابطهٔ $Y'(x) = c$ صدق می‌نماید. خمهای $f(x, y) = c$ را **خمهای تراز** این معادلهٔ دیفرانسیل می‌نامند.

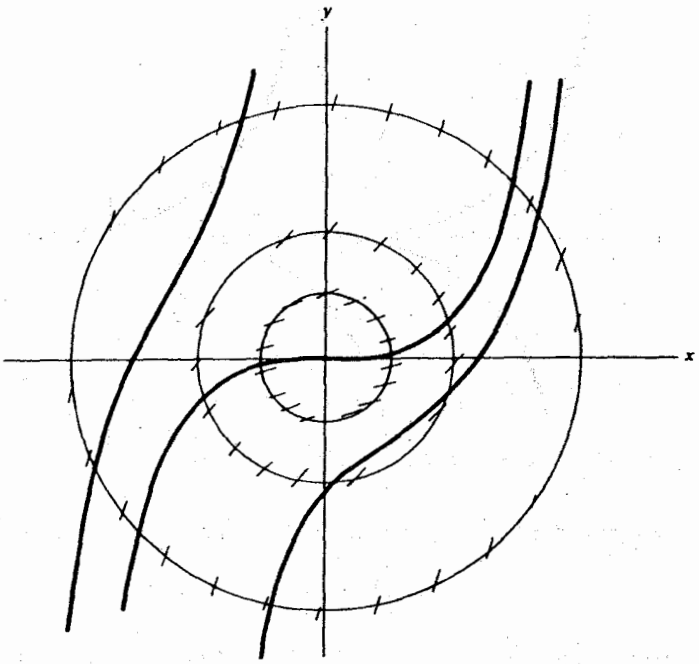
مثال نمایش رفتار کیفی جوابهای

$$y' = x^2 + y^2$$

را رسم کنید. خمهای تراز معادله با

$$x^2 + y^2 = c \qquad c > 0$$

داده می‌شوند که دایره‌هایی به مرکز $(0, 0)$ و به شعاع $\sqrt{c}$ هستند. میدان سوها با بعضی جوابهای نمونه در شکل ۲.۶ داده شده است. سه دایره با شعاعهای ۴ و ۱ و $\frac{1}{4}$ $= c$ رسم شده است. این ترسیم یک اطلاع کیفی دربارهٔ جواب $Y(x)$ در اختیار ما می‌گذارد و گاهی اوقات می‌تواند مفید باشد.



شکل ۲.۶ میدان سوها برای $y' = x^2 + y^2$

۱.۶ وجود، یکتایی و نظریهٔ پایداری

از مثالهای مقدمه باید مشهود باشد که در بیشتر حالات، مسألهٔ مقدار اولیه (۱.۰.۶) یک جواب یکتا دارد. پیش از شروع به آنالیز عددی (۱.۰.۶) بعضی قضایای نظری آن را ارائه می‌دهیم. شرایطی داده شده است تا یکتایی جواب را تضمین نماید، و ما پایداری جواب را هنگامی که دادهٔ Y_0 و مشتق $f(x,y)$ به اندازهٔ کوچکی تغییر نمایند، بررسی خواهیم کرد. این قضایا برای بهتر فهمیدن روشهای عددی که بعداً ارائه خواهند شد لازم‌اند. از این پس فرض می‌شود که حوزهٔ D در شرط فنی فرعی زیر صدق می‌نماید: اگر دو نقطهٔ (x, y_1) و (x, y_2) هر دو به D تعلق داشته باشند، آنگاه پاره‌خط قائمی که این دو نقطه را به هم وصل می‌کند نیز به D تعلق دارد.

قضیهٔ ۱.۶ گیریم $f(x,y)$ یک تابع پیوسته از x و y به‌ازای جمیع مقادیر (x,y) در D، و (x_0, Y_0) یک نقطهٔ داخلی در D باشد. فرض می‌کنیم $f(x,y)$ در شرط **لیپشیتس** زیر

$$| f(x,y_1) - f(x,y_2) | \leq K | y_1 - y_2 | \qquad (x,y_1),(x,y_2) \in D \text{ به ازای همهٔ مقادیر}$$
(۱.۱.۶)

برای مقداری از $K \geq 0$، صدق می‌کند. در این صورت در یک بازهٔ $I = [x_0 - \alpha, x_0 + \alpha]$ که مناسب انتخاب شده باشد، (۱.۰.۶) جواب یکتای $Y(x)$ دارد.

برهان برهان این قضیه را می‌توانید در اغلب کتابهای نظریهٔ معادلات دیفرانسیل معمولی [مثلاً بویس و دیپریما (۱۹۸۶، ص ۹۵)] ببینید. به همین جهت ما آن را حذف می‌کنیم. ■

نشان می‌دهیم که شرط لیپشیتس برقرار است اگر $\partial f(x,y)/\partial y$ در D موجود و کراندار باشد. می‌گیریم

$$K = \underset{(x,y)\in D}{\text{Max}} \left|\frac{\partial f(x,y)}{\partial y}\right| \qquad (۲.۱.۶)$$

سپس با استفاده از قضیهٔ مقدار میانگین

$$f(x,y_1) - f(x,y_2) = \frac{\partial f(x,\xi)}{\partial y}(y_1 - y_2)$$

برای مقداری از ξ بین y_1 و y_2. با استفاده از (۲.۱.۶) نتیجهٔ (۱.۱.۶) فوراً به‌دست می‌آید.

مثال دو مثال زیر این قضیه را روشن می‌سازند

۱. معادلهٔ $y' = 1 + \sin(xy)$ را با

$$D = \{(x,y) \mid 0 \leq x \leq 1, -\infty < y < \infty\}$$

در نظر می‌گیریم. برای محاسبهٔ K، ثابت لیپشیتس، (۲.۱.۶) را به‌کار می‌بریم. پس

$$\frac{\partial f(x,y)}{\partial y} = x \cdot \cos(xy) \qquad K = 1$$

بنابراین برای هر (x_0, Y_0)، با $0 < x_0 < 1$، یک جواب $Y(x)$ مربوط به مسألهٔ مقدار اولیه در بازه‌ای چون $[x_0 - \alpha, x_0 + \alpha] \subset [0, 1]$ وجود دارد.

۲. مسألهٔ

$$y' = \frac{2x}{a^2}y^2 \qquad y(0) = 1$$

را برای هر ثابت $a > 0$ در نظر می‌گیریم. جواب چنین است

$$Y(x) = \frac{a^2}{a^2 - x^2} \qquad -a < x < a$$

اگر a کوچک باشد، جواب فقط در بازهٔ کوچکی وجود دارد. برای تعیین ثابت لیپشیتس باید

$$\frac{\partial f(x,y)}{\partial y} = \frac{4xy}{a^2}$$

را حساب کنیم. برای آنکه یک ثابت متناهی لیپشیتس روی D داشته باشیم، ناحیهٔ D باید نسبت به x و y کراندار باشد، مثلاً $-c \le x \le c$ و $-b \le y \le b$. در این صورت، قضیهٔ قبلی بیان می‌کند که یک جواب $Y(x)$ در بازه‌ای چون $-\alpha \le x \le \alpha$ با $\alpha \le c$، وجود دارد.

مبادا چنین پنداشت که می‌توانیم مشتق جزیی در نقطهٔ آغازی $(x_0, Y_0) = (0, 1)$ را حساب کرده اطلاعات کافی برای براورد ثابت لیپشیتس K به‌دست آوریم، توجه داشته باشید که برای هر مقدار $a > 0$

$$\frac{\partial f(0,1)}{\partial y} = 0$$

پایداری جواب پایداری جواب $Y(x)$ هنگامی بررسی می‌شود که مسألهٔ مقدار اولیه به اندازهٔ کوچکی تغییر یابد. این مطلب به بحث پایداری بخش ۶.۱ فصل ۱ مرتبط است. ما مسألهٔ اختلال‌یافتهٔ زیر را، با همان مفروضات $f(x;y)$ در قضیهٔ ۱.۶ در نظر می‌گیریم

$$\begin{aligned} y' &= f(x,y) + \delta(x) \\ y(x_0) &= Y_0 + \varepsilon \end{aligned} \qquad (۳.۱.۶)$$

وانگهی فرض می‌کنیم به‌ازای جمیع مقادیر x هایی که به‌ازای یک y، $(x,y) \in D$ باشد، $\delta(x)$ پیوسته است. در این صورت می‌توان نشان داد که مسألهٔ (۳.۱.۶) یک جواب یکتا دارد که با $Y(x;\delta,\varepsilon)$ نشان داده می‌شود.

قضیهٔ ۲.۶ همان فرضهای قضیهٔ ۱.۶ را می‌پذیریم. در این صورت مسألهٔ (۲.۱.۶) در بازه‌ای چون $[x_0 - \alpha, x_0 + \alpha]$، به‌ازای مقداری از $\alpha > 0$، یک جواب یکتای $Y(x;\delta,\varepsilon)$ یکنواخت دارد برای تمام اختلالهای ε و $\delta(x)$ که $\delta(x)$ برای مقادیر به اندازه کافی کوچک ε_0 در

$$|\varepsilon| \le \varepsilon_0 \qquad \|\delta\|_\infty \le \varepsilon_0$$

صدق می‌کند. به‌علاوه، اگر $Y(x)$ جواب مسألهٔ اختلال نیافته باشد، آنگاه

$$\max_{|x-x_0|\le\alpha} |Y(x)-Y(x;\delta,\varepsilon)| \le k[|\varepsilon|+\alpha\|\delta\|_\infty] \qquad (۴.۱.۶)$$

که در آن، با استفاده از ثابت لیپشیتس K از (۱.۱.۶)، $k=1/(1-\alpha K)$.

برهان به‌دست آوردن (۴.۱.۶) بسیار شبیه به اثبات قضیهٔ ۱.۶ است و می‌توان آن را در بسیاری از کتابهای درسی کارشناسی ارشد در معادلات دیفرانسیل معمولی یافت. ■

با استفاده از این قضیه، می‌توانیم بگوییم که مسألهٔ مقدار اولیهٔ (۱.۰.۶)، به تعبیر بخش ۶.۱ فصل ۱، خوش‌حالت یا پایدار است. اگر تغییرات کوچکی در معادلهٔ دیفرانسیل یا در مقدار اولیه داده شود جواب نیز به اندازهٔ کوچکی تغییر می‌کند. جواب Y به داده‌های مسأله، یعنی تابع f و مقدار اولیهٔ Y_0، بستگی پیوسته دارد.

در بخش ۶.۱ اشاره شد که مسأله می‌تواند پایدار ولی نسبت به محاسبات عددی بد ـ وضع باشد. این وضعیت برای معادلات دیفرانسیل مصداق دارد، اگرچه در عمل زیاد پیش نمی‌آید. برای اینکه بهتر بفهمیم که چه موقع ممکن است چنین اتفاقی بیفتد، ما اختلال Y را که از اختلالهای مسأله نتیجه می‌شود برآورد می‌کنیم. برای ساده‌کردن بحث، فقط اختلال ε در مقدار اولیهٔ Y_0 را در نظر می‌گیریم. اختلالات $\delta(x)$ در معادله تقریباً به همان‌گونه که در (۴.۱.۶) نشان داده شده است در جواب نهایی مسأله وارد می‌شوند.

مانند (۳.۱.۶)، اختلالی در مقدار آغازی Y_0 ایجاد می‌کنیم، گیریم $Y(x;\varepsilon)$ معرّف جواب اختلال یافته باشد. پس

$$\begin{aligned} &Y'(x;\varepsilon)=f(x,Y(x;\varepsilon)) \qquad x_0-\alpha\le x\le x_0+\alpha \\ &Y(x_0;\varepsilon)=Y_0+\varepsilon \end{aligned} \qquad (۵.۱.۶)$$

معادلات (۱.۰.۶) برای $Y(x)$ را از معادلات فوق کم می‌کنیم و قرار می‌دهیم $Z(x)=Y(x;\varepsilon)-Y(x)$. در این صورت

$$\begin{aligned} Z'(x;\varepsilon) &= f(x,Y(x;\varepsilon))-f(x,Y(x)) \\ &\doteq \frac{\partial f(x,Y(x))}{\partial y} Z(x;\varepsilon) \end{aligned} \qquad (۶.۱.۶)$$

و $Z(x_\circ;\varepsilon)=\varepsilon$. تقریب (۶.۱.۶) وقتی معتبر است که $Y(x;\varepsilon)$ به اندازهٔ کافی به $Y(x)$ نزدیک باشد، که برای مقادیر کوچک ε و بازهٔ کوچک $[x_\circ-\alpha,x_\circ+\alpha]$ چنین است.

به سادگی می‌توانیم معادلهٔ دیفرانسیل تقریبی (۶.۱.۶) را حل کنیم و به‌دست آوریم

$$Z(x;\varepsilon)\doteq\varepsilon\cdot\exp\left[\int_{x_\circ}^{x}\frac{\partial f(t,Y(t))}{\partial y}dt\right] \qquad (۷.۱.۶)$$

اگر مشتق جزئی در رابطهٔ

$$\frac{\partial f(t,Y(t))}{\partial y}\leq\circ \qquad |\,x_\circ-t\,|\leq\alpha \qquad (۸.۱.۶)$$

صدق کند آنگاه، هرگاه x افزایش یابد احتمال دارد $Z(x,\varepsilon)$ به وسیلهٔ ε کراندار بماند. در این حالت گوییم مسألهٔ مقدار اولیه خوش ـ وضع است.

به‌عنوان یک مثال از رفتار معکوس، مسألهٔ زیر را در نظر می‌گیریم

$$y'=\lambda y+g(x) \qquad y(\circ)=Y_\circ \qquad (۹.۱.۶)$$

با $\lambda>\circ$. پس $\partial f/\partial y=\lambda$ و می‌توانیم دقیقاً حساب کنیم

$$Z(x;\varepsilon)=\varepsilon e^{\lambda x}$$

پس تغییر در $Y(x)$ با افزایش x بسیار بزرگ می‌گردد.

مثال معادلهٔ

$$y'=۱۰۰y-۱۰۱e^{-x} \qquad y(\circ)=۱ \qquad (۱۰.۱.۶)$$

دارای جواب $Y(x)=e^{-x}$ است. مسألهٔ اختلال‌یافتهٔ

$$y'=۱۰۰y-۱۰۱e^{-x} \qquad y(\circ)\equiv۱+\varepsilon$$

دارای جواب

$$Y(x;\varepsilon)=e^{-x}+\varepsilon e^{۱۰۰x}$$

است که به سرعت از جواب واقعی دور می‌شود. گوییم (۱۰.۱.۶) یک مسألهٔ بد ـ وضع است.

برای اینکه یک مسأله خوش ـ وضع باشد، نیاز داریم که انتگرال

$$\int_{x_\circ}^{x}\frac{\partial f(t,Y(t))}{\partial t}dt$$

با افزایش x، کران بالای صفر یا یک عدد مثبت کوچک داشته باشد. پس اختلال $Z(x;\varepsilon)$ با مضرب ثابتی از ε کراندار می‌شود، که مقدار ثابت خیلی بزرگ نخواهد بود.

در حالتی که (۸.۱.۶) برقرار ولی قدرمطلق مشتق نسبی بزرگ باشد، با افزایش x، اختلال $Z(x;\varepsilon)$ به سرعت به صفر می‌گراید. چنین معادلاتی خوش‌وضع به‌حساب می‌آیند، ولی آنها نیز می‌توانند برای بیشتر روشهای عددی این فصل مشکل‌آفرین باشند. این معادلات را **معادلات دیفرانسیل سرسخت** (stiff) خوانند و ما در بخش ۹.۶ به آنها باز می‌گردیم.

دستگاه معادلات دیفرانسیل مطالب این فصل به دستگاه m معادلهٔ مرتبهٔ اول زیر تعمیم می‌یابد.

$$\begin{aligned} y_1' &= f_1(x, y_1, \dots, y_m) \qquad y_1(x_0) = Y_{1,0} \\ &\vdots \\ y_m' &= f_m(x, y_1, \dots, y_m) \qquad y_m(x_0) = Y_{m,0} \end{aligned} \tag{۱۱.۱.۶}$$

این دستگاه اغلب با استفاده از نمادهای برداری به شکل یک معادلهٔ مرتبهٔ اول نوشته می‌شود. قرار می‌دهیم

$$\mathbf{y}(x) = \begin{bmatrix} y_1(x) \\ \vdots \\ y_m(x) \end{bmatrix} \qquad \mathbf{f}(x, \mathbf{y}) = \begin{bmatrix} f_1(x, \mathbf{y}) \\ \vdots \\ f_m(x, \mathbf{y}) \end{bmatrix} \qquad \mathbf{Y}_0 = \begin{bmatrix} Y_{1,0} \\ \vdots \\ Y_{m,0} \end{bmatrix} \tag{۱۲.۱.۶}$$

در این صورت دستگاه (۱۱.۱.۶) می‌تواند به شکل زیر نوشته شود

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}) \qquad \mathbf{y}(x_0) = \mathbf{Y}_0 \tag{۱۳.۱.۶}$$

با استفاده از نرمهای برداری به جای قدرمطلق، در واقع تمام نتایج قبلی برای این مسألهٔ مقدار اولیهٔ برداری تعمیم می‌یابد.

برای معادلات از مراتب بالاتر، مسألهٔ مقدار اولیه

$$\begin{aligned} y^{(m)} &= f(x, y, y', \dots, y^{(m-1)}) \\ y(x_0) &= Y_0, \dots, y^{(m-1)}(x_0) = Y_0^{(m-1)} \end{aligned} \tag{۱۴.۱.۶}$$

را با معرفی توابع مجهول جدید زیر، می‌توان به یک دستگاه معادلهٔ مرتبهٔ اول برگرداند

$$y_1 = y y_2 = y', \dots, y_m = y^{(m-1)}$$

این توابع در دستگاه زیر صدق می‌کنند

$$\begin{aligned} y_1' &= y_2 & y_1(x_0) &= Y_0 \\ y_2' &= y_3 & y_2(x_0) &= Y_0' \\ &\vdots & &\vdots \\ y_{m-1}' &= y_m & & \\ y_m' &= f(x, y_1, \dots y_m) & y_m(x_0) &= Y_0^{(m-1)} \end{aligned} \qquad (۱۵.۱.۶)$$

مثال معادلهٔ خطی مرتبهٔ دوّم زیر

$$y'' = a_1(x)y' + a_0(x)y + g(x) \qquad y(x_0) = \alpha \quad y'(x_0) = \beta$$

به دستگاه زیر تبدیل می‌شود

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ a_0(x) & a_1(x) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} 0 \\ g(x) \end{bmatrix} \quad \begin{bmatrix} y_1(x_0) \\ y_2(x_0) \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

در شکل برداری که در آن $A(x)$ ماتریس ضرایب است

$$\mathbf{y}' = A(x)\mathbf{y} + \mathbf{G}(x) \qquad \mathbf{y}(x_0) = Y_0 = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

که یک دستگاه معادلات دیفرانسیل خطّی مرتبهٔ **اول** است.

روشهای ویژه‌ای برای معادلات مرتبهٔ mام وجود دارند، ولی این روشها فقط برای $m = 2$ که در کاربردهای مکانیک نیوتونی از قانون دوّم مکانیک نیوتن ظاهر می‌شوند، به اندازهٔ قابل‌توجهی بسط داده شده‌اند. بیشتر معادلات مراتب بالا پس از تبدیل به یک دستگاه معادلات مرتبهٔ اول هم‌ارز، همان گونه که در بالا توضیح داده شد، حل می‌شوند.

۲.۶ روش اویلر

متداولترین روشهای عددی برای حل (۱.۰.۶)، **روشهای تفاضلات متناهی** نامیده می‌شوند. مقادیر تقریبی برای جواب در یک مجموعه از نقاط گرهی

$$x_0 < x_1 < x_2 < \dots < x_n < \dots \qquad (۱.۲.۶)$$

به‌دست آمده‌اند و مقدار تقریبی در هر x_n با استفاده از مقادیری که در مراحل قبلی محاسبه شده، به‌دست آمده است. ما بحث را با یک روش ساده ولی از جنبهٔ محاسبه غیرکارا که به لئونهارت اویلر

نسبت داده می‌شود آغاز می‌کنیم. تحلیل آن بسیاری از جنبه‌های آنالیز روشهای تفاضلات متناهی کاراتر را در بر دارد، بی‌آنکه پیچیدگی اضافی آنها را داشته باشد. ابتدا چندین شکل پیداکردن روش اویلر را بیان و با تحلیل کامل همگرایی و پایداری موضوع را دنبال می‌کنیم. یک فرمول مجانبی خطا به‌دست آورده این بخش را با تعمیم نتایج قبلی برای یک دستگاه معادلات به پایان می‌رسانیم.

مانند قبل، $Y(x)$ نمایانگر جواب درست (۱.۰.۶) است:

$$Y'(x) = f(x, Y(x)) \qquad Y(x_0) = Y_0 \qquad (۲.۲.۶)$$

جواب تقریبی را با $y(x)$، و مقادیر $y(x_0), y(x_1), \ldots, y(x_n), \ldots$ را اغلب با $y_0, y_1, \ldots, y_n, \ldots$ نشان می‌دهیم. اندازهٔ گامهای مساوی $h > 0$ را برای تعریف نقاط گره‌یی به‌کار می‌بریم.

$$x_j = x_0 + jh \qquad j = 0, 1, \ldots$$

وقتی جوابهای عددی را به‌ازای مقادیر مختلف h مقایسه می‌کنیم، نماد $y_h(x)$ را نیز برای اشاره به $y(x)$ با طول گام h به‌کار می‌بریم. مسألهٔ (۱.۰.۶) را در یک بازهٔ متناهی ثابت که همیشه با نماد $[x_0, b]$ نشان داده می‌شود، حل خواهیم کرد. نماد $N(h)$ نشان دهندهٔ بزرگترین اندیس N است که برای آن

$$x_N \leq b \qquad x_{N+1} > b$$

در بخشهای بعدی، تغییر اندازهٔ گام را در هر نقطهٔ x_n مورد بحث قرار می‌دهیم تا خطا را کنترل کنیم.

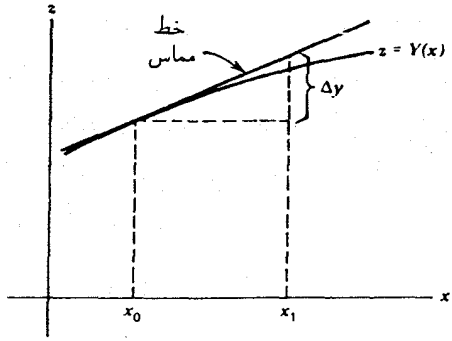
پیداکردن روش اویلر روش اویلر چنین تعریف شده است

$$y_{n+1} = y_n + hf(x_n, y_n) \qquad n = 0, 1, 2, \ldots \qquad (۳.۲.۶)$$

با $y_0 \doteq Y_0$. این روش از چهار دیدگاه نشان داده شده است.

۱. از دیدگاه هندسی. نمودار جواب $Y(x)$ در شکل ۳.۶ را در نظر می‌گیریم. خط مماس بر نمودار $Y(x)$ در x_0 را رسم می‌کنیم و این خط را به‌عنوان یک تقریب خم برای $x_0 \leq x \leq x_1$ به‌کار می‌بریم. پس

$$\frac{\Delta y}{h} = Y'(x_0) = f(x_0, Y_0)$$
$$Y(x_1) - Y(x_0) \doteq \Delta y = hY'(x_0)$$
$$Y(x_1) \doteq Y(x_0) + hf(x_0, Y(x_0))$$



شکل ۳.۶ تعبیر هندسی روش اویلر

با تکرار این استدلال در $[x_1, x_2]$، $[x_2, x_3]$، ... فرمول کلی (۳.۲.۶) را به‌دست می‌آوریم.

۲. سری تیلر. $Y(x_{n+1})$ را حول x_n بسط می‌دهیم.

$$Y(x_{n+1}) = Y(x_n) + hY'(x_n) + \frac{h^2}{2}Y''(\xi_n) \qquad x_n \le \xi_n \le x_{n+1} \qquad (۴.۲.۶)$$

با حذف عبارت خطا، روش (۳.۲.۶) ی اویلر را به‌دست می‌آوریم. جملهٔ

$$T_n = \frac{h^2}{2}Y''(\xi_n) \qquad (۵.۲.۶)$$

خطای برشی یا **خطای گسسته‌سازی** در x_{n+1} خوانده می‌شود. ما خطای برشی را در این کتاب به‌کار می‌بریم.

۳. مشتقگیری عددی. از تعریف مشتق

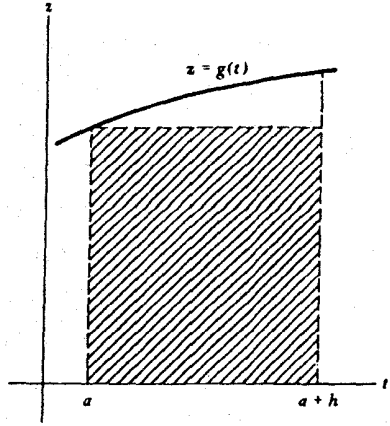
$$\frac{Y(x_{n+1}) - Y(x_n)}{h} \doteq Y'(x_n) = f(x_n, Y(x_n))$$
$$Y(x_{n+1}) \doteq Y(x_n) + hf(x_n, Y(x_n))$$

۴. انتگرالگیری عددی. از $Y'(t) = f(t, Y(t))$ در $[x_n, x_{n+1}]$ انتگرال می‌گیریم:

$$Y(x_{n+1}) = Y(x_n) + \int_{x_n}^{x_{n+1}} f(t, Y(t))dt \qquad (۶.۲.۶)$$

روش سادهٔ انتگرالگیری عددی را در نظر می‌گیریم

$$\int_a^{a+h} g(t)dt \doteq hg(a) \qquad (۷.۲.۶)$$



شکل ۴.۶ نمایش (۷.۲.۶)

که قاعدهٔ **مستطیلی دست چپ** خوانده می‌شود. قسمت هاشورزدهٔ شکل ۴.۶ انتگرال عددی را نشان می‌دهد. با استفاده از این انتگرال در (۶.۲.۶)، مانند گذشته خواهیم داشت:

$$Y(x_{n+1}) \doteq Y(x_n) + hf(x_n, Y(x_n))$$

از سه روش تحلیلی (۲) تا (۴)، هم (۲) و هم (۴) ساده‌ترین حالات یک مجموعهٔ روشهای دقیق افزایشی هستند. روش (۲) به **روشهای تک گامی**، به‌ویژه فرمولهای رونگه ـ کوتا می‌انجامد. روش (۴) ما را به **روشهای چند گامی**، به‌ویژه روشهای مصحّح ـ پیشگو می‌رساند. شاید شگفت‌آور باشد که روش (۳) اغلب به روشهای موفق دیگری نمی‌رسد. اولین مثال برای (۳)، روش میانگاهی در بخش ۴.۶ است که به مسائل ناپایداری عددی می‌انجامد. در مقابل، در بخش ۹.۶، مشتقگیری عددی به‌کار رفته است که یک رده از روشهایی را برای حلّ معادلات دیفرانسیل سرسخت به‌دست می‌دهد. قسمت عمدهٔ این فصل به روشهای چند مرحله‌یی اختصاص داده شده است، تا اندازه‌ای بدین علّت که آنها معمولاً کاراترین رده از روشها هستند، و تا اندازه‌ای به این دلیل که تحلیل آنها پیچیده‌تر از روشهای رونگه ـ کوتاست. این روش اخیر را در بخش ۱۰.۶ ادامه خواهیم داد.

قبل از تحلیل روش اویلر، چند مثال عددی می‌آوریم. این مثالها همچنین برای توضیح قسمتی از نظریه‌ای که ارائه شده به‌کار گرفته می‌شوند.

مثال ۱. معادلهٔ $y' = y$، $y(0) = 1$ را درنظر می‌گیریم. جواب درست آن $Y(x) = e^x$ است. نتایج عددی در جدول ۱.۶ برای چند مقدار h داده شده‌اند. جوابهای $y_h(x_n)$ به جای آنکه در

جدول ۱.۶ روش اویلر برای مثال (۱)

	x	$y_h(x)$	$Y(x)$	$Y(x)-y_h(x)$
$h=$ ۰٫۲	۰٫۴۰	۱٫۴۴۰۰۰	۱٫۴۹۱۸۲	۰٫۰۵۱۸۲
	۰٫۸۰	۲٫۰۷۳۶۰	۲٫۲۲۵۵۴	۰٫۱۵۱۹۴
	۱٫۲۰	۲٫۹۸۵۹۸	۳٫۳۲۰۱۲	۰٫۳۳۴۱۳
	۱٫۶۰	۴٫۲۹۹۸۲	۴٫۹۵۳۰۳	۰٫۶۵۳۲۱
	۲٫۰۰	۶٫۱۹۱۷۴	۷٫۳۸۹۰۶	۱٫۱۹۷۳۲
$h=$ ۰٫۱	۰٫۴۰	۱٫۴۶۴۱۰	۱٫۴۹۱۸۲	۰٫۰۲۷۷۲
	۰٫۸۰	۲٫۱۴۳۵۹	۲٫۲۲۵۵۴	۰٫۰۸۱۹۵
	۱٫۲۰	۳٫۱۳۸۴۳	۳٫۳۲۰۱۲	۰٫۱۸۱۶۹
	۱٫۶۰	۴٫۵۹۴۹۷	۴٫۹۵۳۰۳	۰٫۳۵۸۰۶
	۲٫۰۰	۶٫۷۲۷۵۰	۷٫۳۸۹۰۶	۰٫۶۶۱۵۶
$h=$ ۰٫۰۵	۰٫۴۰	۱٫۴۷۷۴۶	۱٫۴۹۱۸۲	۰٫۰۱۴۳۷
	۰٫۸۰	۲٫۱۸۲۸۷	۲٫۲۲۵۵۴	۰٫۰۴۲۶۷
	۱٫۲۰	۳٫۲۲۵۱۰	۳٫۳۲۰۱۲	۰٫۰۹۵۰۲
	۱٫۶۰	۴٫۷۶۴۹۴	۴٫۹۵۳۰۳	۰٫۱۸۸۰۹
	۲٫۰۰	۷٫۰۳۹۹۹	۷٫۳۸۹۰۶	۰٫۳۴۹۰۷

تمام نقاطی که حساب شده، داده شوند، فقط در چند نقطه داده شده‌اند. توجه کنید که در هر نقطه، وقتی h نصف می‌شود، خطا نیز تقریباً نصف می‌شود.

۲. معادله

$$y' = \frac{۱}{۱+x^۲} - ۲y^۲ \qquad y(۰) = ۰ \qquad (۸.۲.۶)$$

را در نظر می‌گیریم. جواب درست مسأله

$$Y(x) = \frac{x}{۱+x^۲}$$

است. نتایج در جدول ۲.۶ داده شده‌اند. بازهم به رفتار خطا وقتی h کاهش می‌یابد توجه کنید.

تحلیل همگرایی در هر مرحله از روش اویلر، یک خطای برشی اضافی (۵.۲.۶) وارد می‌شود. ما اثر تجمّعی این خطاها را تحلیل می‌کنیم. خطای $Y(x)-y(x)$را **خطای کلی** نامند، و ستونهای آخر جدولهای ۱.۶ و ۲.۶ مثالهایی از خطای کلی‌اند.

جدول ۲.۶ روش اویلر برای مثال (۲)

	x	$y_h(x)$	$Y(x)$	$Y(x)-y_h(x)$
$h=$ ۰٫۲	۰٫۰۰	۰٫۰	۰٫۰	۰٫۰
	۰٫۴۰	۰٫۳۷۶۳۱	۰٫۳۴۴۸۳	−۰٫۰۳۱۴۸
	۰٫۸۰	۰٫۵۴۲۲۸	۰٫۴۸۷۸۰	−۰٫۰۵۴۴۸
	۱٫۲۰	۰٫۵۲۷۰۹	۰٫۴۹۱۸۰	−۰٫۰۳۵۲۹
	۱٫۶۰	۰٫۴۶۶۳۲	۰٫۴۴۹۴۴	−۰٫۰۱۶۸۹
	۲٫۰۰	۰٫۴۰۶۸۲	۰٫۴۰۰۰۰	−۰٫۰۰۶۸۲
$h=$ ۰٫۱	۰٫۴۰	۰٫۳۶۰۸۵	۰٫۳۴۴۸۳	−۰٫۰۱۶۰۳
	۰٫۸۰	۰٫۵۱۳۷۱	۰٫۴۸۷۸۰	−۰٫۰۲۵۹۰
	۱٫۲۰	۰٫۵۰۹۶۱	۰٫۴۹۱۸۰	−۰٫۰۱۷۸۱
	۱٫۶۰	۰٫۴۵۸۷۲	۰٫۴۴۹۴۴	−۰٫۰۰۹۲۸
	۲٫۰۰	۰٫۴۰۴۱۹	۰٫۴۰۰۰۰	−۰٫۰۰۴۱۹
$h=$ ۰٫۰۵	۰٫۴۰	۰٫۳۵۲۸۷	۰٫۳۴۴۸۳	−۰٫۰۰۸۰۴
	۰٫۸۰	۰٫۵۰۰۴۹	۰٫۴۸۷۸۰	−۰٫۰۱۲۶۸
	۱٫۲۰	۰٫۵۰۰۷۳	۰٫۴۹۱۸۰	−۰٫۰۰۸۹۲
	۱٫۶۰	۰٫۴۵۴۲۵	۰٫۴۴۹۴۴	−۰٫۰۰۴۸۱
	۲٫۰۰	۰٫۴۰۲۲۷	۰٫۴۰۰۰۰	−۰٫۰۰۲۲۷

برای آنکه یک برداشت شهودی از رفتار خطای کلی روش اویلر پیدا کنیم، مسألهٔ خیلی سادهٔ زیر را در نظر می‌گیریم

$$y' = ۲x \qquad y(۰) = ۰ \qquad (۹.۲.۶)$$

$Y(x) = x^۲$ جواب آن است. روش اویلر برای این مسأله چنین است

$$y_{n+۱} = y_n + ۲hx_n \qquad y_۰ = ۰$$

سپس با استقراء می‌توان تحقیق کرد که

$$y_n = x_{n-۱}x_n \qquad n \geq ۱$$

برای خطا

$$Y(x_n) - y_n = x_n^۲ - x_n x_{n-۱} = hx_n \qquad (۱۰.۲.۶)$$

پس خطای کلی در هر نقطهٔ ثابت x متناسب با h است. و این نکته با رفتار مثالها در جدولهای ۱.۶ و ۲.۶ توافق دارد، که در آنها هرگاه h نصف شود خطا در حدود نصف تقلیل می‌یابد.

برای تحلیل کامل خطا، مطلب را با لم زیر آغاز می‌کنیم، که در تحلیل روشهای تفاضلات متناهی بسیار سودمند است.

لم ۱ به‌ازای هر مقدار حقیقی x

$$1 + x \leq e^x$$

و به‌ازای هر $x \geq -1$

$$0 \leq (1+x)^m \leq e^{mx} \qquad (11.2.6)$$

برهان با استفاده از قضیهٔ تیلر

$$e^x = 1 + x + \frac{x^2}{2} e^{\xi}$$

با ξ بین 0 و x. چون باقیمانده همیشه مثبت است، قسمت اول ثابت می‌شود. به دنبال آن فرمول (۱۱.۲.۶) به سادگی نتیجه می‌شود. ■

در بقیهٔ این فصل، فرض می‌شود که تابع $f(x,y)$ در شرط قویتر لیپشیتس زیر صدق می‌کند:

$$|f(x,y_1) - f(x,y_2)| \leq \mathbf{k} |y_1 - y_2| \qquad -\infty < y_1, y_2 < \infty \qquad x_0 \leq x \leq b \qquad (12.2.6)$$

برای مقداری از $K \geq 0$. اگرچه این شرط از شرط لازم قویتر است، ولی اثبات را آسان می‌کند. و اگر تابع $f(x,y)$ که در شرط ضعیفتر (۱.۱.۶) صدق می‌کند، و یک جواب $Y(x)$ در مسألهٔ مقدار اولیهٔ (۱.۰.۶) داده شده باشد، تابع f را می‌توان به گونه‌ای تغییر داد که در شرط (۱۲.۲.۶) صدق نماید، بی‌آنکه جواب $Y(x)$ تغییر یابد یا ویژگی اساسی مسألهٔ (۱.۰.۶) و جواب عددی آن عوض شود. [برای جزئیات شمپاین[1] و گوردون[2] (۱۹۷۵) (ص ۲۴) را ببینید].

قضیهٔ ۳.۶ فرض می‌کنیم که $Y(x)$، جواب (۱.۰.۶)، دارای مشتق مرتبهٔ دوّم کراندار در $[x_0, b]$ باشد. در این صورت جواب $\{y_h(x_n) \mid x_0 \leq x_n \leq b\}$ که از روش (۳.۲.۶)ی اویلر به‌دست

1. Shampine 2. Gordon

آمده است در رابطهٔ زیر صدق می‌کند

$$\underset{x_0 \le x_n \le b}{\text{Max}} \mid Y(x_n) - y_h(x_n) \mid \le e^{(b-x_0)K} \mid e_0 \mid + \left[\frac{e^{(b-x_0)K} - 1}{K}\right] \tau(h) \qquad (۱۳.۲.۶)$$

که

$$\tau(h) = \frac{h}{2} \parallel Y'' \parallel_\infty \qquad (۱۴.۲.۶)$$

و $e_0 = Y_0 - y_h(x_0)$.

به‌علاوه اگر $h \to 0$ وقتی

$$\mid Y_0 - y_h(x_0) \mid \le c_1 h \qquad (۱۵.۲.۶)$$

برای مقداری از $c_1 \ge 0$ (مثلاً $Y_0 = y_0$ به ازای جمیع مقادیر h) برقرار باشد، آنگاه یک ثابت $B \ge 0$ وجود دارد که برای آن

$$\underset{x_0 \le x_n \le b}{\text{Max}} \mid Y(x_n) - y_h(x_n) \mid \le Bh \qquad (۱۶.۲.۶)$$

برهان گیریم $e_n = Y(x_n) - y(x_n)$، $n \ge 0$، و تعریف $N(h)$ در ابتدای بخش را به یاد می‌آوریم. براساس خطای برشی در (۵.۲.۶) تعریف می‌کنیم

$$\tau_n = \frac{h}{2} Y''(\xi_n) \qquad 0 \le n \le N(h) - 1$$

با استفاده از (۱۴.۲.۶) به‌سادگی نتیجه می‌شود

$$\underset{0 \le n \le N-1}{\text{Max}} \mid \tau_n \mid \le \tau(h)$$

از روابط (۴.۲.۶)، (۲.۲.۶) و (۳.۲.۶) داریم

$$Y_{n+1} = Y_n + hf(x_n, Y_n) + h\tau_n \qquad (۱۷.۲.۶)$$

$$y_{n+1} = y_n + hf(x_n, y_n) \qquad 0 \le n \le N(h) - 1 \qquad (۱۸.۲.۶)$$

که از نماد معمولی $Y_n \equiv Y(x_n)$ استفاده کرده‌ایم. از کم‌کردن (۱۸.۲.۶) از (۱۷.۲.۶)،

$$e_{n+1} = e_n + h[f(x_n, Y_n) - f(x_n, y_n)] + h\tau_n \qquad (۱۹.۲.۶)$$

به‌دست می‌آید. و کران آن را با استفاده از (۱۲.۲.۶) حساب می‌کنیم

$$| e_{n+1} | \leq | e_n | + hK | Y_n - y_n | + h | \tau_n |$$

$$| e_{n+1} | \leq (1+hK) | e_n | + h\tau(h) \qquad 0 \leq n \leq N(h)-1 \qquad (۲۰.۲.۶)$$

رابطهٔ (۲۰.۲.۶) را می‌توانیم به‌صورت بازگشتی به‌کاربرده به‌دست آوریم،

$$| e_n | \leq (1+hK)^n | e_0 | + \{1 + (1+hK) + \ldots + (1+hK)^{n-1}\} h\tau(h)$$

با استفاده از فرمول مجموع سری هندسی متناهیِ

$$1 + r + r^2 + \ldots + r^{n-1} = \frac{r^n - 1}{r - 1} \qquad r \neq 1 \qquad (۲۱.۲.۶)$$

خواهیم داشت

$$| e_n | \leq (1+hK)^n | e_0 | + [\frac{(1+hK)^n - 1}{K}] \tau(h) \qquad (۲۲.۲.۶)$$

با استفاده از لم ۱،

$$(1+hK)^n \leq e^{nhK} = e^{(x_n - x_0)K} \leq e^{(b-x_0)K}$$

و این رابطه با (۲۲.۲.۶) قسمت اصلی قضیه، (۱۳.۲.۶)، قضیه را به‌دست خواهد داد.

باقی قضیه، (۱۶.۲.۶)، یک نتیجهٔ نمایانِ (۱۳.۲.۶) است که ثابت B با

$$B = c_1 e^{(b-x_0)K} + \left[\frac{e^{(b-x_0)K} - 1}{K}\right] \cdot \frac{\| Y'' \|_\infty}{2}$$

داده شده است. و این اثبات را کامل می‌نماید. ■

نتیجه (۱۶.۲.۶) ایجاب می‌کند که هرگاه طول گام h نصف شود، خطا دست‌کم با ضریب یک‌دوّم کاهش یابد. این مطلب با مثالهای جدولهای ۱.۶ و ۲.۶ تائید شده است. بعداً در این بخش نشان داده شده است که (۱۶.۲.۶) دقیقاً نرخ درست همگرایی را به‌دست می‌دهد (مسألهٔ ۷ را نیز ببینید).

کران (۱۶.۲.۶) سرعت درست همگرایی در روش اویلر را به‌دست می‌دهد، ولی ضریب ثابت B در اغلب معادلات خیلی بیش از اندازه بزرگ است. مثلاً، با مثال قبلی (۸.۲.۶)، فرمول

(۱۳.۲.۶) پیشگویی می‌کند که خطا با b بزرگ می‌شود. ولی به‌روشنی در جدول ۲.۶ مشاهده می‌شود، وقتی x افزایش می‌یابد، خطا کاهش پیدا می‌کند. ما اصلاح زیر را برای (۱۳.۲.۶) ارائه می‌کنیم که در خیلی حالات مانند (۸.۲.۶) به‌کار خواهد رفت.

فرع همان فرضهای قضیهٔ ۳.۶ را در نظر می‌گیریم؛ به‌علاوه فرض می‌کنیم

$$\frac{\partial f(x,y)}{\partial y} \leq 0 \qquad (۲۳.۲.۶)$$

برای $x_0 \leq x \leq b$ و $-\infty < y < \infty$. در این صورت به‌ازای جمیع مقادیر h به اندازهٔ کافی کوچک

$$| Y(x_n) - y_h(x_n) | \leq | e_0 | + \frac{h}{2}(x_n - x_0) \underset{x_0 \leq x_n \leq b}{\text{Max}} | Y''(x) | \qquad (۲۴.۲.۶)$$

به‌ازای $x_0 \leq x_n \leq b$. ■

برهان قضیهٔ مقدار میانگین را برای معادلهٔ خطای (۱۹.۲.۶) به‌کار می‌بریم:

$$e_{n+1} = \left[1 + h\frac{\partial f(x_n, \zeta_n)}{\partial y}\right] e_n + \frac{h^2}{2} Y''(\xi_n) \qquad (۲۵.۲.۶)$$

با ζ_n بین $y_h(x_n)$ و $Y(x_n)$. از همگرایی $y_h(x)$ به $Y(x)$ در $[x_0, b]$، می‌فهمیم که مشتقهای جزئی $\partial f(x_n, \zeta_n)/\partial y$ به $\partial f(x, Y(x))/\partial y$ میل می‌کنند، و بنابراین باید از نظر قدرمطلق در $[x_0, b]$ کراندار باشند. $h_0 > 0$ را طوری می‌گیریم که برای تمام مقادیر $h \leq h_0$،

$$1 + h\frac{\partial f(x_n, \zeta_n)}{\partial y} \geq -1 \qquad x_0 \leq x_n \leq b \qquad (۲۶.۲.۶)$$

از (۲۳.۲.۶) نتیجه می‌گیریم که سمت چپ نیز به‌ازای جمیع مقادیر h از بالا با مقدار ۱ کراندار شده است. این نتایج را در (۲۵.۲.۶) به‌کار می‌بریم تا به‌دست آید

$$| e_{n+1} | \leq | e_n | + \frac{h^2}{2} | Y''(\xi_n) | \qquad (۲۷.۲.۶)$$

با استقرا می‌توانیم نشان دهیم

$$| e_n | \leq | e_0 | + \frac{h^2}{2} [| Y''(\xi_0) | + \ldots + | Y''(\xi_{n-1}) |]$$

که به‌سادگی به (۲۴.۲.۶) منجر می‌شود. ■

نتیجهٔ (۲۴.۲.۶) خیلی بهتر از کران قبلی (۱۳.۲.۶) است، به جای $\exp(K(b-x_0))$ مقدار $b-x_0$ (کران x_n-x_0) گذاشته شده است، که با افزایش b با سرعت کمتری بزرگ می‌شود. این قضیه مستقیماً برای مثال قبلی (۸.۲.۶) کاربرد ندارد، ولی با یک بررسی دقیق برهان در این حالت، نشان داده می‌شود که برهان هنوز معتبر است.

تحلیل پایداری تحلیل پایداریِ مسألهٔ مقدار اولیه را که در قضیهٔ ۲.۶ داده شده بود به یاد آورید. برای درنظرگرفتن طرح مشابهی برای روش اویلر، روش عددی زیر را با $z_0 = y_0 + \varepsilon$ در نظر می‌گیریم

$$z_{n+1} = z_n + h[f(x_n, z_n) + \delta(x_n)] \qquad 0 \le n \le N(h)-1 \qquad (۲۸.۲.۶)$$

مقایسهٔ این دو فرمول مانند مقایسهٔ (۵.۱.۶) با (۱.۰.۶) است، که پایداری مسألهٔ مقدار اولیه را نشان می‌دهد. ما دو جواب عددی $\{z_n\}$ و $\{y_n\}$ را وقتی $h \to 0$، مقایسه می‌کنیم.
گیریم $e_n = z_n - y_n$، $n \ge 0$. پس $e_0 = \varepsilon$، و با کم‌کردن (۳.۲.۶) از (۲۸.۲.۶)،

$$e_{n+1} = e_n + h[f(x_n, z_n) - f(x_n, y_n)] + h\delta(x_n)$$

این فرمول درست شکل (۱۹.۲.۶) را دارد. اگر شیوه‌ای مشابه آنچه را که به دنبال (۱۹.۲.۶) آمده است به‌کار بریم، خواهیم داشت

$$\underset{0 \le n \le N(h)}{\text{Max}} \mid z_n - y_n \mid \le e^{(b-x_0)K} \mid \varepsilon \mid + \left[\frac{e^{(b-x_0)K}-1}{K}\right] \parallel \delta \parallel_\infty$$

نتیجتاً ثابتهایی مانند k_1 و k مستقل از h، وجود دارند به‌طوری که

$$\underset{0 \le n \le N(h)}{\text{Max}} \mid z_n - y_n \mid \le k_1 \mid \varepsilon \mid + k_2 \parallel \delta \parallel_\infty \qquad (۲۹.۲.۶)$$

این نتیجه مشابه نتیجهٔ (۴.۱.۶) برای مسألهٔ اصلی (۱.۰.۶) است. این رابطه می‌گوید که روش اویلر یک روش عددی پایدار برای حل مسألهٔ مقدار اولیه (۱.۰.۶) است. تأکید می‌کنیم که تمام روشهای عددی برای مسائل مقدار اولیه چنین شکل پایداری دارند و از پایداری مسألهٔ اصلی (۱.۰.۶) پیروی می‌کنند. به‌علاوه، به پایداری به شکلهای دیگر نیز نیاز داریم، که بعداً معرفی می‌شوند. در آینده، می‌گیریم $\delta(x) \equiv 0$ و فقط اثر اختلال در مقدار اولیهٔ Y_0 را در نظر می‌گیریم. این امر موجب سادگی تحلیل می‌شود و نتایج به همان اندازه سودمند خواهند بود.

تحلیل خطای گردکردن در هر مرحله از روش اویلر خطایی وارد می‌کنیم، خطایی که از خطاهای گردکردن اعمال انجام‌شده حاصل شده است. این عدد، که با نماد ρ_n نشان داده می‌شود، **خطای گردکردن موضعی** خوانده می‌شود. اگر نتیجهٔ مقادیر عددی را $\tilde{y}_n$ بنامیم، خواهیم داشت

$$\tilde{y_{n+1}} = \tilde{y}_n + hf(x_n, \tilde{y}_n) + \rho_n \qquad n = 0, 1, \ldots, N(h) - 1 \qquad (۳۰.۲.۶)$$

مقادیر $\tilde{y}_n$ اعدادی با ارقام اعشاری متناهی هستند که عملاً در رایانه به‌دست آمده‌اند، و y_n مقداری است که اگر حساب دقیق به‌کار می‌رفت به‌دست می‌آمد. گیریم $\rho(h)$ کرانی برای خطاهای گردکردن باشد،

$$\rho(h) = \underset{0 \le n \le N(h)-1}{\text{Max}} \mid \rho_n \mid \qquad (۳۱.۲.۶)$$

در یک وضعیت عملی، با استفاده از رایانه‌یی که طول کلمهٔ ثابت داشته باشد، کران $\rho(h)$ وقتی $h \to 0$، کاهش نمی‌یابد. بلکه تقریباً ثابت می‌ماند و $\rho(h)/ \parallel Y \parallel_\infty$ با واحد خطای گردکردن u در رایانه، متناسب خواهد بود، یعنی u کوچکترین عدد رایانه‌یی است که برای آن $1 + u > 1$ [فرمول (۱۲.۲.۱) فصل ۱ را ببینید].

برای آنکه اثر خطاهای گردکردن در (۳۰.۲.۶) را ببینیم، این رابطه را از جواب درست در (۴.۲.۶) کم می‌کنیم تا حاصل شود:

$$\tilde{e}_{n+1} = \tilde{e}_n + h[f(x_n, Y_n) - f(x_n, \tilde{y}_n)] + h\tau_n - \rho_n$$

که در آن $\tilde{e}_n = Y(x_n) - \tilde{y}_n$. مانند اثبات قضیهٔ ۳.۶ عمل می‌کنیم ولی $\tau_n - \rho_n/h$ را با τ_n که در برهان قبلی بود یکی می‌گیریم. در این صورت به‌دست می‌آوریم

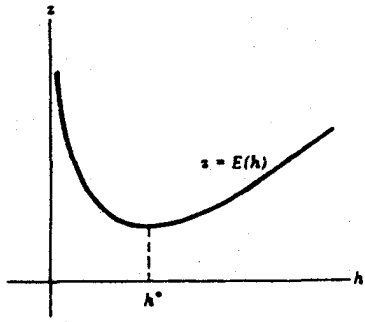
$$\mid \tilde{e}_n \mid \le e^{(b-x_0)k} \mid Y_0 - \tilde{y}_0 \mid + \left[\frac{e^{(b-x_0)K} - 1}{K}\right] \left[\tau(h) + \frac{\rho(h)}{h}\right] \qquad (۳۲.۲.۶)$$

برای بررسی بیشتر این کران، گیریم $u \doteq \rho(h)/ \parallel Y \parallel_\infty$، همان‌گونه که قبلاً بحث شد. پس

$$\mid \tilde{e}_n \mid \le c \left[\frac{h}{2} \parallel Y'' \parallel_\infty + \frac{u \parallel Y \parallel_\infty}{h}\right] \equiv E(h) \qquad (۳۳.۲.۶)$$

رفتار کیفی $E(h)$ در نمودار شکل ۵.۶ نشان داده شده است. $E(h)$ در h^* مینیمم می‌شود، و هر کاهش دیگر، باعث افزایش خطای کران $E(h)$ خواهد شد.

این $E(h)$، بدترین حالت ممکن را برای اثر خطاهای گردکردن می‌دهد. در عمل، هم اندازه و هم علامت خطاهای گردکردن تغییر می‌کنند. برآیند حذف باعث می‌شود که $\tilde{e}_n$ با سرعتی کمتر از



شکل ۵.۶ خم خطا برای (۳۳.۲.۶)

سرعتی که بر اثر جملهٔ $1/h$ که در $E(h)$ و (۳۲.۲.۶) حاصل می‌شود، افزایش یابد. ولی هنوز هم یک مقدار بهینهٔ h^* وجود دارد و برای h کمتر از آن، خطا باز افزایش می‌یابد. برای یک تحلیل کاملتر هنریچی (۱۹۶۲، صص ۳۵ـ۵۹) را ببینید.

معمولاً خطای گردکردن مسألهٔ مهمی نیست. ولی، اگر دقت مطلوب به‌بهترین دقتی که به‌علت محدودیت طول کلمه در رایانه می‌توان دست یافت، نزدیک باشد، به اثرات گردکردن باید توجه بیشتری بشود. در یک رایانه بزرگ IBM (جدول ۱.۱ را ببینید) با حساب دقّت مضاعف، اگر $h \geq$ ۰٫۰۰۱، آنگاه ماکسیمم u/h برابر $۲٫۲ \times ۱۰^{-۱۳}$ خواهد شد که u واحد گردکردن است. بنابراین خطای گردکردن معمولاً مسألهٔ قابل‌توجهی به‌وجود نخواهد آورد مگر آنکه تحمّلهای خطای بسیار کوچکی خواسته شده باشد. ولی در دقت ساده، با همین محدودیت روی h، ماکسیمم u/hبرابر $۰٫۵ \times ۱۰^{-۴}$ می‌شود، و با تحمل خطایی به این اندازه، (که یک مقدار غیرمنطقی نیست)، خطای گردکردن عامل مؤثرتری خواهد بود.

مثال مسألهٔ

$$y' = -y + ۲\cos(x) \qquad y(۰) = ۱$$

را که جواب درست آن $Y(x) = \sin(x) + \cos(x)$ است حل می‌کنیم. این مسأله را با روش اویلر و استفاده از سه حساب مختلف حل می‌کنیم: (۱) حساب با ممیز شناور با چهار رقم اعشار و قطع‌کردن؛ (۲) حساب ممیز شناور با چهار رقم اعشار و گردکردن؛ و (۳) حساب دقیق، یا حساب با دقت بسیار بالا. در دو حالت اوّل، واحد خطاهای گردکردن به‌ترتیب $u =$ ۰٫۰۰۱ و $u =$ ۰٫۰۰۰۵ هستند. کران (۳۲.۲.۶) را برای حالتهای (۱) و (۲) به‌کار می‌بریم، حال آنکه

جدول ۳.۶ مثال اثرات خطای گردکردن در روش اویلر

h	x	اعشاری قطع‌شده	اعشاری گردشده	حساب دقیق
۰٫۰۴	۱	$-۱٫۰۰\mathrm{E}-۲$	$-۱٫۷۰\mathrm{E}-۲$	$-۱٫۷۰\mathrm{E}-۲$
	۲	$۱٫۱۷\mathrm{E}-۲$	$۱٫۸۳\mathrm{E}-۲$	$-۱٫۸۳\mathrm{E}-۲$
	۳	$-۱٫۲۰\mathrm{E}-۳$	$-۲٫۸۰\mathrm{E}-۳$	$-۲٫۷۸\mathrm{E}-۳$
	۴	$۱٫۰۰\mathrm{E}-۲$	$۱٫۶۰\mathrm{E}-۲$	$۱٫۵۳\mathrm{E}-۲$
	۵	$۱٫۱۳\mathrm{E}-۲$	$۱٫۹۶\mathrm{E}-۲$	$۱٫۹۴\mathrm{E}-۲$
۰٫۰۲	۱	$۷٫۰۰\mathrm{E}-۳$	$-۹٫۰۰\mathrm{E}-۳$	$-۸٫۴۶\mathrm{E}-۳$
	۲	$۴٫۰۰\mathrm{E}-۳$	$-۹٫۱۰\mathrm{E}-۳$	$-۹٫۱۳\mathrm{E}-۳$
	۳	$۲٫۳۰\mathrm{E}-۳$	$-۱٫۴۰\mathrm{E}-۳$	$-۱٫۴۰\mathrm{E}-۳$
	۴	$-۶٫۰۰\mathrm{E}-۳$	$۸٫۰۰\mathrm{E}-۳$	$۷٫۶۲\mathrm{E}-۳$
	۵	$-۶٫۰۰\mathrm{E}-۳$	$۸٫۵۰\mathrm{E}-۳$	$۹٫۶۳\mathrm{E}-۳$
۰٫۰۱	۱	$۲٫۸۰\mathrm{E}-۲$	$-۳٫۰۰\mathrm{E}-۳$	$-۴٫۲۲\mathrm{E}-۳$
	۲	$۲٫۲۸\mathrm{E}-۲$	$-۴٫۳۰\mathrm{E}-۳$	$-۴٫۵۶\mathrm{E}-۳$
	۳	$۷٫۴۰\mathrm{E}-۳$	$-۴٫۰۰\mathrm{E}-۴$	$-۷٫۰۳\mathrm{E}-۴$
	۴	$-۲٫۳۰\mathrm{E}-۲$	$۳٫۰۰\mathrm{E}-۳$	$۳٫۸۰\mathrm{E}-۳$
	۵	$-۲٫۴۱\mathrm{E}-۲$	$۴٫۶۰\mathrm{E}-۳$	$۴٫۸۱\mathrm{E}-۳$

حالت ۳ در کران نظری (۲۴.۲.۶) صدق می‌نماید. خطاها برای هر سه شکل روش اویلر در جدول ۳.۶ داده شده‌اند. خطا برای جوابهای حاصله با استفاده از حساب اعشاری، بر پایهٔ جواب درست $Y(x)$ که تا چهار رقم گرد شده است محاسبه شده‌اند.

برای حالت حساب اعشاری قطع‌شده، با $h = ۰٫۰۲$، تأثیر روی خطاها آغاز شده است. با $h = ۰٫۰۱$، خطای قطع‌کردن، اثر قابل‌ملاحظه‌ای بر خطای کلی دارد. برعکس، خطاها، با استفاده از حساب گردشده به‌تدریج کاهش می‌یابند، اگرچه $h = ۰٫۰۱$ قدری متأثر شده است. در این مسأله مانند بسیاری دیگر، استفاده از حساب گردشده بسیار برتر از حساب قطع شده است.

تحلیل خطای مجانبی یک برآورد خطای مجانبی در روش اویلر، با چشم‌پوشی از اثرات خطای گردکردن به‌دست آمده است. قبل از هر چیز، برای آسان‌کردن عملیات جبری در این تحلیل یک نمادگذاری لازم است. اگر $B(x,h)$ تابعی باشد که برای $x_۰ \le x \le b$ و به‌ازای جمیع مقادیر به اندازهٔ کافی کوچک h تعریف شده باشد، آنگاه نماد

$$B(x,h) = O(h^p)$$

به‌ازای مقداری از $p > 0$، بدین معناست که به ازای جمیع مقادیر به اندازهٔ کافی کوچک h، مقدار ثابتی مانند c وجود دارد به طوری که

$$| B(x,h) | \leq ch^p \qquad x_0 \leq x \leq b \qquad (۳۴.۲.۶)$$

اگر B فقط به h بستگی داشته باشد، همین نوع کران به‌دست می‌آید.

قضیهٔ ۴.۶ فرض می‌کنیم که $Y(x)$ جواب مسألهٔ مقدار اولیهٔ (۱.۰.۶) و سه بار مشتقپذیر باشد. فرض می‌کنیم

$$f_y(x,y) \equiv \frac{\partial f(x,y)}{\partial y} \qquad f_{yy}(x,y) \equiv \frac{\partial^2 f(x,y)}{\partial y^2}$$

به‌ازای $x_0 \leq x \leq b$ و $-\infty < y < \infty$، پیوسته و کراندار باشند. گیریم مقدار اولیهٔ $y_h(x_0)$ در رابطهٔ زیر صدق کند

$$Y_0 - y_h(x_0) = \delta_0 h + O(h^2) \qquad (۳۵.۲.۶)$$

معمولاً این خطا صفر است و بنابراین $\delta_0 = 0$.

در این صورت خطا در روش (۳.۲.۶) اویلر در رابطهٔ زیر صدق می‌کند

$$Y(x_n) - y_h(x_n) = D(x_n)h + D(h^2) \qquad (۳۶.۲.۶)$$

که در آن $D(x)$ جواب مسألهٔ مقدار اولیهٔ خطی زیر است:

$$D'(x) = f_y(x, Y(x))D(x) + \frac{1}{2}Y''(x) \qquad D(x_0) = \delta_0 \quad (۳۷.۲.۶)$$

برهان قضیهٔ تیلر را به‌کار می‌بریم،

$$Y(x_{n+1}) = Y(x_n) + hY'(x_n) + \frac{h^2}{2}Y''(x_n) + \frac{h^3}{6}Y^{(3)}(\xi_n)$$

به‌ازای یک $x_n \leq \xi_n \leq x_{n+1}$. رابطهٔ (۳.۲.۶) را از رابطهٔ فوق کم کرده از (۲.۲.۶) استفاده می‌کنیم تا به‌دست آوریم

$$\begin{aligned} e_{n+1} =& e_n + h[f(x_n, Y_n) - f(x_n, y_n)] \\ &+ \frac{h^2}{2}Y''(x_n) + \frac{h^3}{6}Y^{(3)}(\xi_n) \end{aligned} \qquad (۳۸.۲.۶)$$

$f(x_n, y_n)$ را به‌عنوان تابعی از y_n در نظر گرفته قضیهٔ تیلر را برای آن به‌کار می‌بریم،

$$f(x_n, y_n) = f(x_n, Y_n) + (y_n - Y_n) f_y(x_n, Y_n) + \frac{1}{2}(y_n - Y_n)^2 f_{yy}(x_n, \zeta_n)$$

به‌ازای مقداری از ζ_n بین y_n و Y_n. اگر از این رابطه در (۳۸.۲.۶) استفاده کنیم

$$e_{n+1} = [1 + h f_y(x_n, Y_n)] e_n + \frac{h^2}{2} Y''(x_n) + B_n$$

$$B_n = \frac{h^3}{6} Y^{(3)}(\xi_n) - \frac{1}{2} h f_{yy}(x_n, \zeta_n) e_n^2 \qquad (۳۹.۲.۶)$$

با استفاده از (۱۶.۲.۶)

$$B_n = O(h^3) \qquad (۴۰.۲.۶)$$

چون B_n نسبت به سایر جملات (۳۹.۲.۶) کوچک است، قسمت عمدهٔ خطا را با حذف B_n به‌دست می‌آوریم. گیریم g_n، معرّف قسمت عمدهٔ خطا، به‌طور ضمنی با

$$g_{n+1} = [1 + h f_y(x_n, Y_n)] g_n + \frac{h^2}{2} Y''(x_n) \qquad (۴۱.۲.۶)$$

تعریف شده است، که

$$g_0 = \delta_0 h \qquad (۴۲.۲.۶)$$

قسمت عمدهٔ خطا در مقدار اولیهٔ (۳۵.۲.۶) است. چون انتظار داریم که $g_n \doteq e_n$، و چون $e_n = O(h)$، دنبالهٔ $\{\delta_n\}$ را که به‌طور ضمنی این گونه معرفی می‌کنیم

$$g_n = h\delta_n \qquad (۴۳.۲.۶)$$

در (۴۱.۲.۶) می‌گذاریم و h را حذف کرده دوباره مرتب می‌نماییم، به‌دست می‌آید

$$\delta_{n+1} = \delta_n + h[f_y(x_n, Y_n)\delta_n + \frac{1}{2} Y''(x_n)] \qquad x_0 \le x_n \le b \qquad (۴۴.۲.۶)$$

مقدار اولیهٔ δ_0 در (۳۵.۲.۶) مستقل از h تعریف شده است.

معادلهٔ (۴۴.۲.۶) روش اویلر برای حل مسألهٔ مقدار اولیهٔ (۳۷.۲.۶) است؛ بنابراین به موجب قضیه ۳.۶،

$$D(x_n) - \delta_n = O(h) \qquad x_0 \le x_n \le b \qquad (۴۵.۲.۶)$$

این رابطه را با (۴۳.۲.۶) ترکیب می‌کنیم،

$$g_n = D(x_n)h + O(h^۲) \qquad (۴۶.۲.۶)$$

برای کامل‌کردن اثبات، باید نشان دهیم که g_nدرواقع قسمت اصلی خطای e_n است. نماد زیر را معرفی می‌کنیم

$$k_n = e_n - g_n \qquad (۴۷.۲.۶)$$

در این صورت $k_۰ = e_۰ - g_۰ = O(h^۲)$، که از (۳۵.۲.۶) و (۴۲.۲.۶) نتیجه می‌شود. رابطهٔ (۴۱.۲.۶) را از (۳۹.۲.۶) کم می‌کنیم و (۴۰.۲.۶) را به‌کار می‌بریم

$$k_{n+۱} = [۱ + hf_y(x_n, Y_n)]k_n + B_n$$
$$| k_{n+۱} | \leq (۱ + hK) | k_n | + O(h^۳)$$

این رابطه به شکل (۲۰.۲.۶) در برهان قضیهٔ ۳.۶ است که به‌جای جملهٔ $h\tau(h)$ جملهٔ $O(h^۳)$ گذاشته شده است. با استفاده از همان محاسبه

$$| k_n | = O(h^۲) \qquad (۴۸.۲.۶)$$

از ترکیب (۴۶.۲.۶) ـ (۴۸.۲.۶) داریم

$$e_n = g_n + k_n = [hD(x_n) + O(h^۲)] + O(h^۲)$$

که رابطهٔ (۳۶.۲.۶) را ثابت می‌کند. ■

تابع $D(x)$ به‌ندرت صریحاً به‌دست می‌آید، ولی شکل خطا در (۳۶.۲.۶) اطلاعات کیفی مفیدی در اختیار می‌گذارد. این اغلب به‌عنوان پایهٔ شیوه‌های برونیابی به‌کار می‌رود، که بعضی از آنها در بخشهای بعد مورد بحث قرار گرفته‌اند.

مثال مسأله

$$y' = -y \qquad y(۰) = ۱$$

را با جواب $Y(x) = e^{-x}$ در نظر می‌گیریم. معادله برای $D(x)$ چنین است

$$D'(x) = -D(x) + \frac{۱}{۲}e^{-x} \qquad D(۰) = ۰$$

و جواب این معادله

جدول ۴.۶ مثال (۴۹.۲.۶) از قضیهٔ ۴.۶

x_n	$Y_n - y_n$	$hD(x_n)$
۰ر۴	۰ر۰۰۶۸۹	۰ر۰۰۶۷۰
۰ر۸	۰ر۰۰۹۲۰	۰ر۰۰۸۹۹
۱ر۲	۰ر۰۰۹۲۱	۰ر۰۰۹۰۴
۱ر۶	۰ر۰۰۸۱۹	۰ر۰۰۸۰۸
۲ر۰	۰ر۰۰۶۸۲	۰ر۰۰۶۷۷

$$D(x) = \frac{۱}{۲}xe^{-x}$$

است. این رابطه فرمول مجانبی زیر را برای خطا در روش اویلر به‌دست می‌دهد

$$Y(x_n) - y_h(x_n) \doteq \frac{h}{۲}x_n e^{-x_n} \qquad (۴۹.۲.۶)$$

جدول ۴.۶ شامل خطاهای واقعی و خطاهایی است که با (۴۹.۲.۶) به‌ازای $h = ۰ر۰۵$ پیش‌بینی شده است. توجه نمایید که با افزایش x، خطا کاهش می‌یابد، درست مثل خطای جواب $Y(x)$. ولی خطای نسبی به‌طور خطی نسبت به x، افزایش می‌یابد،

$$\frac{Y(x_n) - y_h(x_n)}{Y(x_n)} \doteq \frac{h}{۲}x_n$$

همچنین، براورد (۴۹.۲.۶) به مراتب بهتر از کرانی است که توسط (۱۳.۲.۶) در قضیهٔ ۳.۶ داده شده است. آن کران چنین است

$$| Y(x_n) - y_h(x_n) | \leq \frac{h}{۲}(e^{x_n} - ۱)$$

و به‌طور نمایی با x افزایش می‌یابد.

دستگاه معادلات برای ساده‌کردن بیان، فقط دستگاه مرتبهٔ دوّم زیر را در نظر می‌گیریم:

$$\begin{aligned} y_۱' &= f_۱(x, y_۱, y_۲) \quad y_۱(x_۰) = Y_{۱,۰} \\ y_۲' &= f_۲(x, y_۱, y_۲) \quad y_۲(x_۰) = Y_{۲,۰} \end{aligned} \qquad (۵۰.۲.۶)$$

تعمیم آن به دستگاههای مراتب بالاتر ساده است. روش اویلر برای حل (۵۰.۲.۶) چنین است

$$\begin{aligned} y_{۱,n+۱} &= y_{۱,n} + hf_۱(x_n, y_{۱,n}, y_{۲,n}) \\ y_{۲,n+۱} &= y_{۲,n} + hf_۲(x_n, y_{۱,n}, y_{۲,n}) \end{aligned} \qquad (۵۱.۲.۶)$$

که یک تعمیم روشن از (۳.۲.۶) است.

تمام نتایج پیشین این بخش، به‌صورت (۵۱.۲.۶) تعمیم پیدا می‌کنند، و شکل تعمیم بسیار روشن خواهد بود اگر نماد گذاری برداری (۱۳.۱.۶) را که در بخش قبل بیان شد برای (۵۰.۲.۶) و (۵۱.۲.۶) به‌کار بریم. می‌نویسیم

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}) \qquad \mathbf{y}(x_0) = \mathbf{Y}_0$$

به‌جای قدر مطلق، نُرم (۱۶.۱.۱) از فصل ۱ را به‌کار می‌بریم

$$\| \mathbf{y} \| = \max_i | y_i |$$

برای تعمیم شرط (۱۲.۲.۶) لیپشیتس، قضیهٔ تیلر (قضیهٔ ۱.۵) برای توابع چندمتغیره را به‌کار می‌بریم تا به‌دست آوریم

$$\| \mathbf{f}(x, \mathbf{z}) - \mathbf{f}(x, \mathbf{y}) \| \leq K \| \mathbf{z} - \mathbf{y} \| \qquad (۵۲.۲.۶)$$

$$K = \max_i \sum_j \max_{\substack{x_0 \leq x \leq b \\ -\infty < w_1, w_2 < \infty}} \left| \frac{\partial f_i(x, w_1, w_2)}{\partial w_j} \right| \qquad (۵۳.۲.۶)$$

ماتریس ژاکوبی زیر نقش $\partial f(x,y)/\partial y$ را دارد

$$\mathbf{f}_y(x, \mathbf{y}) = \begin{bmatrix} \dfrac{\partial f_1}{\partial y_1} & \dfrac{\partial f_1}{\partial y_2} \\ \dfrac{\partial f_2}{\partial y_1} & \dfrac{\partial f_2}{\partial y_2} \end{bmatrix} \qquad (۵۴.۲.۶)$$

به‌عنوان مثال، فرمول خطای مجانبی (۳۶.۲.۶) چنین می‌شود

$$\mathbf{Y}(x_n) - y_h(x_n) = h\mathbf{D}(x_n) + \mathbf{R}_n \qquad \| \mathbf{R}_n \| = O(h^2) \qquad (۵۵.۲.۶)$$

که $\mathbf{D}(x)$ جواب دستگاه خطی زیر است

$$\mathbf{D}'(x) = \mathbf{f}_{\mathbf{y}}(x, \mathbf{Y}(x))\mathbf{D}(x) + \frac{1}{2}\mathbf{Y}''(x) \qquad \mathbf{D}(x_0) = \delta_0 \qquad (۵۶.۲.۶)$$

که از ماتریس قبلی $\mathbf{f}_{\mathbf{y}}(x, \mathbf{y})$ استفاده شده است.

مثال مسألهٔ آونگ زیر را حل کنید

$$\theta''(t) = -\sin(\theta(t)) \qquad \theta(0) = \frac{\pi}{2} \qquad \theta'(0) = 0$$

جدول ۵.۶ روش اویلر برای مثال (۵۷.۲.۶)

h	x_n	$y_{۱,n}$	$Y_۱(x_n)$	خطاها	$y_{۲,n}$	$Y_۲(x_n)$	خطاها
۰٫۲	۰٫۲	۱٫۵۷۰۸	۱٫۵۵۰۸	−۰٫۰۲۰۰	−۰٫۲۰۰۰۰	−۰٫۱۹۹۹۹۹	۰٫۰۰۰۰۰۱
	۰٫۶	۱٫۴۵۰۸	۱٫۳۹۱۰	−۰٫۰۵۹۸	−۰٫۵۹۹۸۴	−۰٫۵۹۸۰۶	۰٫۰۰۱۷۸
	۱٫۰	۱٫۱۷۱۱	۱٫۰۷۴۹	−۰٫۰۹۶۲	−۰٫۹۹۲۶۷	−۰٫۹۷۵۵۰	۰٫۰۱۷۱۷
۰٫۱	۰٫۲	۱٫۵۶۰۸	۱٫۵۵۰۸	−۰٫۰۱۰۰	−۰٫۲۰۰۰۰	−۰٫۱۹۹۹۹۹	۰٫۰۰۰۰۰۱
	۰٫۶	۱٫۴۲۰۸	۱٫۳۹۱۰	−۰٫۰۲۹۸	−۰٫۵۹۹۲۷	−۰٫۵۹۸۰۶	۰٫۰۰۱۲۱
	۱٫۰	۱٫۱۲۲۳	۱٫۰۷۴۹	−۰٫۰۴۷۴	−۰٫۹۸۵۶۸	−۰٫۹۷۵۵۰	۰٫۰۱۰۱۸

با انتخاب $y_۱ = \theta$ و $y_۲ = \theta'$ و گذاشتن x به‌جای t، معادلهٔ فوق به دستگاه زیر تبدیل می‌شود

$$\begin{aligned} y_۱' &= y_۲ & y_۱(۰) &= \frac{\pi}{۲} \\ y_۲' &= -\sin(y_۱) & y_۲(۰) &= ۰ \end{aligned} \qquad (۵۷.۲.۶)$$

نتایج عددی در جدول ۵.۶ داده شده‌اند. توجه کنید وقتی h نصف می‌شود خطا تقریباً نصف می‌شود.

۳.۶ روشهای چندگامی

این بخش شامل مقدمه‌ای بر روشهای چندگامی است. بعضی روشهای خاص با جزئیات بیشتر در بخشهای بعدی بررسی شده‌اند. یک نظریهٔ کاملتر در بخش ۸.۶ داده شده است.

مانند قبل، گیریم $h > ۰$ و گره‌ها را با $x_n = x_۰ + nh$، $n \geq ۰$، تعریف می‌کنیم. شکل کلی روشهای چندمرحله‌یی که در نظر گرفته خواهند شد به قرار زیر است

$$y_{n+۱} = \sum_{j=۰}^{p} a_j y_{n-j} + h \sum_{j=-۱}^{p} b_j f(x_{n-j}, y_{n-j}) \qquad n \geq p \qquad (۱.۳.۶)$$

ضرایب $a_۰$، ... ، a_p، $b_{-۱}$، $b_۰$، ... ، b_p ثابت‌اند و $p \geq ۰$. اگر $a_p \neq ۰$ یا $b_p \neq ۰$ برقرار باشد روش را $p+۱$ مرحله‌یی خوانند، زیرا $p+۱$ مقدار قبلی، در محاسبهٔ $y_{n+۱}$ به‌کار گرفته شده‌اند. مقادیر $y_۱$، ... ، y_p باید به وسایل دیگری به‌دست آیند؛ این مطلب در بخشهای دیگر بحث شده است. روش اویلر یک مثال از روش یک گامی با شرطهای $p = ۰$ و

$$a_۰ = ۱ \qquad b_۰ = ۱ \qquad b_{-۱} = ۰$$

است.

اگر $b_{-1} = 0$، آنگاه y_{n+1} فقط در سمت چپ معادلهٔ (۱.۳.۶) ظاهر می‌شود. چنین فرمولهایی را **روشهای صریح** نامند. اگر $b_{-1} \neq 0$، آنگاه y_{n+1} در دو طرف معادلهٔ (۱.۳.۶) وجود دارد و فرمول را **روش ضمنی** خوانند. وجود جواب y_{n+1} را برای h به اندازهٔ کافی کوچک، می‌توان با استفاده از نظریهٔ نقطهٔ ثابت بخش ۵.۲ نشان داد. روشهای ضمنی معمولاً با روشهای بارستی حل می‌شوند که برای روش ذوزنقه‌یی در بخش ۵.۶ بحث شده است.

مثال ۱. روش میانگاهی با رابطهٔ

$$y_{n+1} = y_{n-1} + 2hf(x_n, y_n) \qquad n \geq 1 \qquad (۲.۳.۶)$$

تعریف شده است که یک روش دوگامی صریح است. این روش در بخش ۴.۶ بررسی شده است.

۲. روش ذوزنقه‌یی با رابطهٔ زیر تعریف شده است،

$$y_{n+1} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1})] \qquad n \geq 0 \qquad (۳.۳.۶)$$

این روش یک روش یک‌گامی ضمنی است و در بخشهای ۵.۶ و ۶.۶ مورد بحث قرارگرفته است.

برای هر تابع مشتقپذیر $Y(x)$، خطای برشی برای انتگرالگیری $Y'(x)$ را با رابطهٔ زیر تعریف می‌کنیم

$$T_n(Y) = Y(x_{n+1}) - \left[\sum_{j=0}^{p} a_j Y(x_{n-j}) + h \sum_{j=-1}^{p} b_j Y'(x_{n-j})\right] \qquad n \geq p \qquad (۴.۳.۶)$$

تابع $\tau_n(Y)$ را با رابطهٔ زیر تعریف می‌کنیم

$$\tau_n(Y) = \frac{1}{h} T_n(Y) \qquad (۵.۳.۶)$$

برای اثبات همگراییِ جواب تقریبیِ $\{y_n \mid x_0 \leq x_n \leq b\}$ از رابطهٔ (۱.۳.۶) به جواب $Y(x)$ از مسألهٔ مقدار اولیهٔ (۱.۰.۶) لازم است که داشته باشیم

$$\tau(h) \equiv \max_{x_p \leq x_n \leq b} | \tau_n(Y) | \to 0 \qquad \text{وقتی } h \to 0 \qquad (۶.۳.۶)$$

این رابطه اغلب **شرط سازگاری** برای روش (۱.۳.۶) خوانده می‌شود. سرعت همگرایی جواب $\{y_n\}$ به جواب درست $Y(x)$ وابسته به سرعت همگرایی در (۶.۳.۶) است، و بنابراین نیاز داریم

شرایطی را بدانیم که تحت آنها برای انتخاب مطلوبی از $m \geq ۱$ داشته باشیم

$$\tau(h) = O(h^m) \qquad (۷.۳.۶)$$

ما اکنون استلزامهای (۶.۳.۶) و (۷.۳.۶) برای ضرایب در (۱.۳.۶) را تحقیق می‌کنیم. نتیجهٔ همگرایی برای (۱.۳.۶) بعداً به صورت قضیهٔ ۶.۶ داده شده است.

قضیهٔ ۵.۶ گیریم عدد صحیح $m \geq ۱$ داده شده باشد. برای آنکه (۶.۳.۶) به‌ازای همهٔ توابع پیوسته مشتقپذیر $Y(x)$ برقرار باشد، یعنی برای اینکه روش (۱.۳.۶) سازگار باشد، لازم و کافی است که

$$\sum_{j=۰}^{p} a_j = ۱ \qquad -\sum_{j=۰}^{p} j a_j + \sum_{j=-۱}^{p} b_j = ۱ \qquad (۸.۳.۶)$$

و برای آنکه (۷.۳.۶) به‌ازای همهٔ توابع $m+۱$ بار پیوسته مشتقپذیر $Y(x)$ معتبر باشد، لازم و کافی است که (۸.۳.۶) برقرار بوده و

$$\sum_{j=۰}^{p} (-j)^i a_j + i \sum_{j=-۱}^{p} (-j)^{i-۱} b_j = ۱ \qquad i = ۲, \ldots, m \qquad (۹.۳.۶)$$

برهان توجه نمایید که به ازای جمیع مقادیر ثابت α و β و همهٔ توابع مشتقپذیر Y و W، داریم

$$T_n(\alpha Y + \beta W) = \alpha T_n(Y) + \beta T_n(W) \qquad (۱۰.۳.۶)$$

برای بررسی نتایج (۶.۳.۶) و (۷.۳.۶)، با استفاده از قضیهٔ ۴.۱ تیلر، $Y(x)$ را حول x_n بسط می‌دهیم تا به‌دست آوریم

$$Y(x) = \sum_{i=۰}^{m} \frac{۱}{i!}(x - x_n)^i Y^{(i)}(x_n) + R_{m+۱}(x) \qquad (۱۱.۳.۶)$$

با فرض آنکه $Y(x)$، $m+۱$ بار پیوسته مشتقپذیر است. با گذاردن در (۴.۳.۶) و استفاده از (۱۰.۳.۶) داریم

$$T_n(Y) = \sum_{i=۰}^{m} \frac{۱}{i!} Y^{(i)}(x_n) T_n((x - x_n)^i) + T_n(R_{m+۱})$$

لازم است که $T_n((x - x_n)^i)$ را به ازای $i \geq ۰$ محاسبه کنیم.

برای $i = 0$

$$T_n(1) = c_0 \equiv 1 - \sum_{j=0}^{p} a_j \qquad (۱۲.۳.۶)$$

برای $i \geq 1$

$$\begin{aligned} T_n((x-x_n)^i) &= (x_{n+1}-x_n)^i - \left[\sum_{j=0}^{p} a_j (x_{n-j}-x_n)^i + h \sum_{j=-1}^{p} b_j i (x_{n-j}-x_n)^{i-1}\right] \\ &= c_i h^i \end{aligned} \qquad (۱۳.۳.۶)$$

$$c_i = 1 - \left[\sum_{j=0}^{p} (-j)^i a_j + i \sum_{j=-1}^{p} (-j)^{i-1} b_j\right] \qquad i \geq 1$$

و از آنجا نتیجه می‌گیریم

$$T_n(Y) = \sum_{i=0}^{m} \frac{c_i}{i!} h^i Y^{(i)}(x_n) + T_n(R_{m+1}) \qquad (۱۴.۳.۶)$$

اگر باقیماندهٔ $R_{m+1}(x)$ را به شکل زیر بنویسیم

$$R_{m+1}(x) = \frac{1}{(m+1)!}(x-x_n)^{m+1} Y^{(m+1)}(x_n) + \dots$$

آنگاه

$$T_n(R_{m+1}) = \frac{c_{m+1}}{(m+1)!} h^{m+1} Y^{(m+1)}(x_n) + O(h^{m+2}) \qquad (۱۵.۳.۶)$$

با فرض اینکه Y، $m+2$ بار مشتقپذیر است.

برای پیداکردن شرط سازگاری (۶.۳.۶)، لازم است که $\tau(h) = O(h)$، و این ایجاب می‌نماید $T_n(Y) = O(h^2)$. با استفاده از (۱۴.۳.۶) با $m = 1$، باید داشته باشیم $c_0 = 0$ ، c_1 که مجموعهٔ معادلات (۸.۳.۶) را خواهد داد. در بعضی کتابها، این معادلات را شرایط سازگاری نامیده‌اند. برای به‌دست آوردن (۷.۳.۶) برای مقداری از $m \geq 1$، باید داشته باشیم $T_n(Y) = O(h^{m+1})$. که درستی آن از (۱۴.۳.۶) و (۱۳.۳.۶) به‌دست می‌آید اگر و فقط اگر $c_i = 0$، $i = 0, 1, \dots, m$. این، شرایط (۹.۳.۶) را ثابت و برهان را کامل می‌کند. ■

بزرگترین مقدار m که برای آن (۷.۳.۶) برقرار باشد **مرتبه** یا **مرتبهٔ همگرایی** روش (۱.۳.۶) خوانده می‌شود. در بخش ۷.۶ پیداکردن روشهایی از هر مرتبهٔ دلخواه را، بررسی خواهیم نمود.

اکنون یک قضیهٔ همگرایی برای جواب (۱.۳.۶) خواهیم آورد. گرچه، این قضیه همهٔ روشهای چندگامی را که همگرا هستند شامل نمی‌شود، بیشتر روشهای سودمند رایج را در برمی‌گیرد. علاوه بر آن، برهان آن بسیار ساده‌تر از برهان قضیهٔ کلی ۸.۶ از بخش ۸.۶ است.

قضیهٔ ۶.۶ حل مسألهٔ مقدار اولیهٔ زیر را با استفاده از روش چندگامی (۱.۳.۶) درنظر می‌گیریم

$$y' = f(x,y) \qquad y(x_0) = Y_0 \qquad x_0 \le x \le b$$

گیریم خطاهای اولیه در رابطهٔ زیر صدق می‌کنند

$$\eta(h) \equiv \operatorname*{Max}_{0 \le i \le p} | Y(x_i) - y_h(x_i) | \to 0 \quad h \to 0 \quad \text{وقتی} \qquad (۱۶.۳.۶)$$

فرض کنید که این روش سازگار است، یعنی در (۶.۳.۶) صدق می‌کند. و بالاخره، فرض کنید ضرایب a_j همگی نامنفی هستند،

$$a_j \ge 0 \qquad j = 0, 1, \ldots, p \qquad (۱۷.۳.۶)$$

در این صورت روش (۱.۳.۶) همگراست و برای مقادیر مناسبی از c_1 و c_2،

$$\operatorname*{Max}_{x_0 \le x_n \le b} | Y(x_n) - y_h(x_n) | \le c_1 \eta(h) + c_2 \tau(h) \qquad (۱۸.۳.۶)$$

اگر روش (۱.۳.۶) از مرتبهٔ m باشد و اگر خطاهای اولیه در $\eta(h) = O(h^m)$ صدق کنند، آنگاه سرعت همگرایی این روش برابر $O(h^m)$ است.

برهان رابطهٔ (۴.۳.۶) را دوباره می‌نویسیم و از $Y'(x) = f(x, Y(x))$ استفاده می‌کنیم تا به‌دست آوریم

$$Y(x_{n+1}) = \sum_{j=0}^{p} a_j Y(x_{n-j}) + h \sum_{j=-1}^{p} b_j f(x_{n-j}, Y(x_{n-j})) + h\tau_n(Y)$$

با کم‌کردن (۱.۳.۶) از آن و استفاده از نماد $e_i = Y(x_i) - y_i$،

$$e_{n+1} = \sum_{j=0}^{p} a_j e_{n-j} + h \sum_{j=-1}^{p} b_j \left[f(x_{n-j}, Y_{n-j}) - f(x_{n-j}, y_{n-j}) \right] + h\tau_n(Y)$$

شرط لیپشیتس و فرض (۱۷.۳.۶) را به‌کار می‌بریم تا به‌دست آوریم

$$| e_{n+1} | \leq \sum_{j=0}^{p} a_j | e_{n-j} | + hK \sum_{j=-1}^{p} | b_j | | e_{n-j} | + h\tau(h)$$

تابع خطای کرانی زیر را وارد می‌کنیم

$$f_n = \max_{0 \leq i \leq n} | e_i | \qquad n = 0, 1, \ldots, N(h)$$

با به‌کاربردن این تابع،

$$| e_{n+1} | \leq \sum_{j=0}^{p} a_j f_n + hK \sum_{j=-1}^{p} | b_j | f_{n+1} + h\tau(h)$$

و با استفاده از (۸.۳.۶)

$$| e_{n+1} | \leq f_n + hcf_{n+1} + h\tau(h) \qquad c = K \sum_{j=-1}^{p} | b_j |$$

روشن است که طرف راست یک کران برای f_n است، و بنابراین

$$f_{n+1} \leq f_n + hcf_{n+1} + h\tau(h)$$

برای $hc \leq 1/2$ که وقتی $h \to 0$، باید درست باشد،

$$\begin{aligned} f_{n+1} &\leq \frac{f_n}{1 - hc} + \frac{h}{1 - hc}\tau(h) \\ &\leq (1 + 2hc) f_n + 2h\tau(h) \end{aligned}$$

با توجه به اینکه $f_p = \eta(h)$، مانند (۲۰.۲.۶) که در برهان قضیهٔ ۳.۶ آمده است، عمل می‌کنیم. پس

$$f_n \leq e^{2c(b-x_0)}\eta(h) + \left[\frac{e^{2c(b-x_0)} - 1}{c}\right]\tau(h) \qquad x_0 \leq x_n \leq b \quad (۱۹.۳.۶)$$

برهان را کامل می‌کند. ■

نتیجه‌های این قضیه را می‌توان با مفروضات ضعیفتری اثبات نمود؛ به‌ویژه، به جای ((۱۷.۳.۶) می‌توان شرط بسیار ضعیفتری گذاشت. این نتایج در بخش ۸.۶ داده شده‌اند. برای پیداکردن یک نرخ همگرایی $O(h^m)$ برای روش (۱.۳.۶)، لازم است که هر مرحله دارای خطای

$$T_n(Y) = O(h^{m+1})$$

باشد. ولی مقادیر اولیه y_0، ...، y_p فقط لازم است با دقت $O(h^m)$ محاسبه شوند، زیرا $\eta(h) = O(h^m)$ برای (۱۸.۳.۶) کافی است. مثالهایی برای نشان‌دادن استفاده از روش مرتبهٔ پایینتر برای تولید مقادیر اولیهٔ y_0، ...، y_n در بخشهای آینده داده شده است.

نتیجهٔ (۱۹.۳.۶) را می‌توان برای حالات خاص بهبود بخشید، ولی سرعت همگرایی به‌همان حالت خواهد ماند. مثالهایی از این قضیه در بخشهای بعد آمده است. همانند تحلیل در روش اویلر، یک تحلیل کامل پایداری می‌توان ارائه داد، از جمله نتیجه‌ای را به شکل (۲۹.۲.۶). این برهان تغییریافتهٔ سادهٔ برهان قضیهٔ ۶.۶ است. همچنین یک تحلیل خطای مجانبی می‌توان ارائه داد؛ مثالهایی در دو بخش بعد داده شده‌اند.

۴.۶ روش میانگاهی

ما روش میانگاهی را تعریف و تحلیل می‌نماییم، و آن را برای نشان‌دادن بعضی مفاهیم که با روش اویلر ممکن نبود، به‌کار می‌بریم. همانند روش اویلر می‌توانیم روش میانگاهی را به چندین راه پیدا کنیم، و در اینجا از مشتقگیری عددی استفاده می‌کنیم. با توجّه به (۱۱.۷.۵) فصل ۵، داریم

$$g'(a) = \frac{g(a+h) - g(a-h)}{2h} - \frac{h^2}{6} g^{(3)}(\xi)$$

برای مقداری از $a - h \leq \xi \leq a + h$. با به‌کار بردن این رابطه در

$$Y'(x_n) = f(x_n, Y(x_n))$$

داریم

$$\frac{Y(x_{n+1}) - Y(x_{n-1})}{2h} - \frac{h^2}{6} Y^{(3)}(\xi_n) = f(x_n, Y(x_n))$$

با $x_{n-1} \leq \xi_n \leq x_{n+1}$. از حل این رابطه برحسب $Y(x_{n+1})$،

$$Y(x_{n+1}) = Y(x_{n-1}) + 2hf(x_n, Y(x_n)) + \frac{1}{3} h^3 Y^{(3)}(\xi_n) \quad (۱.۴.۶)$$

روش **میانگاهی** با حذف آخرین جمله به‌دست می‌آید:

$$y_{n+1} = y_{n-1} + 2hf(x_n, y_n) \qquad n \geq 1 \qquad (۲.۴.۶)$$

این یک روش صریح دوگامی است و مرتبهٔ همگرایی آن دو است. مقدار y_1 باید با روش دیگری محاسبه شود.

روش میانگاهی را می‌توانستیم از به‌کار بردن قاعدهٔ انتگرالگیری عددی میانگاهی (۱۷.۲.۵) در انتگرال زیر که صورت انتگرالی معادلهٔ دیفرانسیل (۱.۰.۶) است به‌دست آوریم:

$$Y(x_{n+1}) = Y(x_{n-1}) + \int_{x_{n-1}}^{x_{n+1}} f(t, Y(t))dt \qquad (۳.۴.۶)$$

ما از ذکر جزئیات صرف‌نظر می‌کنیم. این راه در بخش ۷.۶، برای پیداکردن روشهای چندگامی دیگر به‌کار رفته است.

برای تحلیل همگرایی (۲.۴.۶)، از قضیهٔ ۶.۶ استفاده می‌کنیم. برای خطای برشی، به راحتی از (۱.۴.۶) به‌دست می‌آوریم که

$$\tau_n(Y) = \frac{1}{3}h^2 Y^{(3)}(\xi_n) \qquad x_{n-1} \le \xi_n \le x_n \qquad (۴.۴.۶)$$

یک شکل برهان بهتر قضیهٔ ۶.۶ برای روش میانگاهی، نتیجه می‌دهد که

$$\underset{x_0 \le x_n \le b}{\text{Max}} \mid Y(x_n) - y_h(x_n) \mid \le e^{2K(b-x_0)}\eta(h) + \left[\frac{e^{2K(b-x_0)} - 1}{2K}\right]\left[\frac{1}{3}h^2 \| Y^{(3)} \|_\infty\right] \qquad (۵.۴.۶)$$

$$\eta(h) = \text{Max}\{| Y_0 - y_0 |, | Y(x_1) - y_h(x_1) |\}$$

با فرض آنکه برای جمیع مقادیر h، $y_0 = Y_0$، نیاز داریم که $Y(x_1) - y_h(x_1) = O(h^2)$، تا مرتبهٔ کلی همگرایی $O(h^2)$ را در (۵.۴.۶) داشته باشیم. با توجه به (۴.۲.۶)، یکی از روشهای تک‌گامی اویلر این ویژگی مورد نظر را داراست؛

$$y_1 = y_0 + hf(x_0, y_0) \qquad y_0 = Y_0 \qquad (۶.۴.۶)$$

$$Y(x_1) - y_1 = \frac{h^2}{2}Y''(\zeta) \qquad x_0 \le \zeta \le x_1 \qquad (۷.۴.۶)$$

با این مقدار اولیه $y_h(x_1) = y_1$، خطای نتیجهٔ (۵.۴.۶) ایجاب می‌کند که

$$\underset{x_0 \le x_n \le b}{\text{Max}} \mid Y(x_n) - y_h(x_n) \mid = O(h^2) \qquad (۸.۴.۶)$$

می‌توانیم یک تحلیل پایداری کامل، مشابه با آنچه برای روش اویلر داده شد، برای روش میانگاهی بیاوریم. اگر برای سادگی فرض کنیم که $\eta(h) = O(h^3)$، آنگاه فرمول خطای مجانبی

زیر را خواهیم داشت

$$Y(x_n) - y_h(x_n) = D(x_n)h^2 + O(h^3) \qquad x_0 \le x_n \le b$$

$$D' = f_y(x, Y(x))D + \frac{1}{6}Y^{(3)}(x) \qquad D(x_0) = 0 \qquad (۹.۴.۶)$$

تفاوت اثبات این نتایج با آنچه در روش اویلر دیدیم کم است و ما از ذکر اثبات صرف‌نظر می‌کنیم.

پایداری ضعیف همان گونه که قبلاً اشاره شد، روش میانگاهی دارای همان نوع پایداری است که برای روش اویلر در (۲۸.۲.۶) و (۲۹.۲.۶) نشان داده شد. ولی به هرحال، این برای منظورهای عملی کافی نیست. نشان خواهیم داد که روش میانگاهی برای مفهوم دیگری از پایداری که باید تعریف شود، رضایتبخش نیست.

راه حل عددی مسألهٔ زیر را در نظر می‌گیریم

$$y' = \lambda y \qquad y(0) = 1 \qquad (۱۰.۴.۶)$$

که دارای جواب $Y(x) = e^{\lambda x}$ است. این مسأله به‌عنوان یک مسألهٔ نمونه برای حالت کلّیتر مسألهٔ (۱.۰.۶) به‌کار خواهد رفت، موضوعی که در بخش ۸.۶ توضیح خواهیم داد. در اینجا کافی است توجه کنیم که اگر یک روش عددی برای مسألهٔ ساده‌ای چون (۱۰.۴.۶)، بد عمل کند، آنگاه غیرمحتمل است که چنین روشی برای معادلات دیفرانسیل پیچیده‌تر، خوب عمل کند.

روش میانگاهی برای معادلهٔ (۱۰.۴.۶) چنین است

$$y_{n+1} = y_{n-1} + 2h\lambda y_n \qquad n \ge 1 \qquad (۱۱.۴.۶)$$

ما جواب درست این معادله را پیدا و با جواب $Y(x) = e^{\lambda x}$ مقایسه می‌کنیم. معادلهٔ (۱۱.۴.۶) یک مثال از **معادلهٔ تفاضلی خطی** مرتبهٔ دو است. یک نظریهٔ عمومی برای معادلات تفاضلی خطی مرتبهٔ pام وجود دارد که با نظریهٔ معادلات دیفرانسیل خطی مرتبهٔ p مشابه است. بیشتر روشهای حل معادلات دیفرانسیل، مشابهی برای حل معادلات تفاضلی دارند و این خود یک راهنما برای حل معادلهٔ (۱۱.۴.۶) خواهد بود. ما کار را با جستجوی جوابهای مستقل خطی برای معادلهٔ تفاضلی آغاز می‌کنیم. سپس این جوابها را برای یافتن جواب کلّی ترکیب خواهیم کرد. برای نظریهٔ کلی معادلات تفاضلی خطی هنریچی (۱۹۶۲، صص ۲۱۰ـ۲۱۵) را ببینید.

در قیاس با جوابهای نمایی برای معادلات دیفرانسیل خطی، جوابی برای معادلهٔ (۱۱.۴.۶) به شکل زیر به‌ازای مقدار مجهول r فرض می‌کنیم

$$y_n = r^n \qquad n \ge 0 \qquad (۱۲.۴.۶)$$

و آن را در معادلهٔ (۱۱.۴.۶) قرار می‌دهیم تا شرایط لازم برای r را پیدا کنیم

$$r^{n+1} = r^{n-1} + 2h\lambda r^n$$

با حذف r^{n-1} داریم،

$$r^2 = 1 + 2h\lambda r \qquad (۱۳.۴.۶)$$

این استدلال برگشت‌پذیر است. اگر r در معادلهٔ درجهٔ دوم (۱۳.۴.۶) صدق کند، آنگاه (۱۲.۴.۶) در (۱۱.۴.۶) صدق خواهد کرد.

معادلهٔ (۱۳.۴.۶) را **معادلهٔ مشخصهٔ** روش میانگاهی نامند. ریشه‌های آن چنین‌اند

$$r_0 = h\lambda + \sqrt{1 + h^2\lambda^2} \qquad r_1 = h\lambda - \sqrt{1 + h^2\lambda^2} \qquad (۱۴.۴.۶)$$

جواب عمومی (۱۱.۴.۶) عبارتست از

$$y_n = \beta_0 r_0^n + \beta_1 r_1^n \qquad n \geq 0 \qquad (۱۵.۴.۶)$$

ضرایب β_0 و β_1 به نحوی انتخاب می‌شوند که مقادیر y_0 و y_1 که در ابتدا داده شده بودند با مقادیری که با (۱۵.۴.۶) محاسبه شده‌اند مطابقت نمایند:

$$\beta_0 + \beta_1 = y_0$$

$$\beta_0 r_0 + \beta_1 r_1 = y_1$$

جواب عمومی این دستگاه چنین است

$$\beta_0 = \frac{y_1 - r_1 y_0}{r_0 - r_1} \qquad \beta_1 = \frac{y_0 r_0 - y_1}{r_0 - r_1}$$

برای آنکه درک مستقیمی از این فرمولها به‌دست آوریم مقادیر اولیهٔ دقیق را درنظر می‌گیریم

$$y_0 = 1 \qquad y_1 = e^{\lambda h}$$

سپس با استفاده از قضیهٔ تیلر،

$$\beta_0 = \frac{e^{\lambda h} - r_1}{2\sqrt{1 + h^2\lambda^2}} = 1 + O(h^2\lambda^2)$$

$$\beta_1 = \frac{r_0 - e^{\lambda h}}{2\sqrt{1 + h^2\lambda^2}} = O(h^3\lambda^3) \qquad (۱۶.۴.۶)$$

از این مقادیر، وقتی $h \to 0$ داریم $\beta_0 \to 1$ و $\beta_1 \to 0$. بنابراین در فرمول (۱۵.۴.۶)، جملهٔ $\beta_0 r_0^n$ باید متناظر جواب درست $e^{\lambda x_n}$ باشد، زیرا وقتی $h \to 0$، جمله $\beta_1 r_1^n \to 0$. در واقع

$$r_0^n = e^{\lambda x_n}[1 + O(h)] \qquad (۱۷.۴.۶)$$

که اثبات آن به عهده خواننده واگذار می‌شود.

برای آنکه دشواری حل عددی $y' = \lambda y$ را در استفاده از (۱۵.۴.۶) ببینیم، اندازه‌های نسبی r_0 و r_1 را به دقّت بررسی می‌کنیم. ما فقط حالتی را که λ حقیقی است در نظر می‌گیریم. برای $-\infty < \lambda < 0$ و همهٔ مقادیر h،

$$r_0 > | r_1 | > 0$$

بنابراین مقدار r_1^n با سرعتی کمتر از r_0^n افزایش می‌یابد، و جملهٔ درست در جواب عمومی (۱۵.۴.۶) یعنی $\beta_0 r_0^n$ غالب می‌شود.

ولی، برای $-\infty < \lambda < 0$، خواهیم داشت

$$0 < r_0 < 1 \qquad r_1 < -1 \qquad h > 0$$

و در نتیجه، $\beta_1 r_1^n$ با افزایش n و مقدار ثابت h، سرانجام بر $\beta_0 r_0^n$ غالب خواهد شد، بدون توجه به اینکه h در ابتدا چقدر کوچک انتخاب شده‌باشد. وقتی $n \to \infty$، جملهٔ $\beta_0 r_0^n \to 0$ ؛ حال آنکه $\beta_1 r_1^n$ از نظر قدرمطلق افزایش می‌یابد و علامت آن با افزایش n یک در میان عوض می‌شود.

جملهٔ $\beta_1 r_1^n$، **جواب مزاحم** روش عددی (۱۱.۴.۶) نامیده می‌شود، زیرا متناظر با هیچ یک از جوابهای معادلهٔ دیفرانسیل اصلی $y' = \lambda y$ نیست. معادلهٔ دیفرانسیل اصلی یک خانواده جواب تک‌پارامتری دارد که به مقدار اولیهٔ Y_0 بستگی دارد، ولی جواب تقریبی (۱۱.۴.۶) یک خانواده جواب دو پارامتری (۱۵.۴.۶) دارد که وابسته به مقادیر y_0 و y_1 است. جواب جدید $\beta_1 r_1^n$ یک آفرینش روش عددی است؛ برای مسألهٔ (۱۰.۴.۶) با $\lambda < 0$، این موجب می‌شود که جواب عددی وقتی $x_n \to \infty$، از جواب درست مسأله، دور شود. به علّت این رفتار گوییم روش میانگاهی فقط ضعیف ـ پایدار است.

پس از آنکه قدری از نظریه را که لازم داریم ارائه کردیم، مجدداً در بخش ۸.۶ به بحث فوق باز می‌گردیم. عملی‌بودن قابلیّت کاربرد مسألهٔ نمونهٔ (۱۰.۴.۶) را با در نظر گرفتن علامت $\partial f(x, Y(x))/\partial y$ تعمیم می‌دهیم. اگر منفی باشد، آنگاه ناپایداری ضعیف روش میانگاهی در حل مسألهٔ مقدار اولیهٔ متناظر، معمولاً ظاهر می‌شود. این موضوع در مثال دوم در زیر نشان داده شده است.

جدول ۶.۶ مثال (۱) ناپایداری روش میانگاهی

x_n	y_n	$Y(x_n)$	خطاها
۰٫۲۵	۰٫۷۵۰۰	۰٫۷۷۸۸	۰٫۰۲۸۸
۰٫۵۰	۰٫۶۲۵۰	۰٫۶۰۶۵	−۰٫۰۱۸۵
۰٫۷۵	۰٫۴۳۷۵	۰٫۴۷۲۴	۰٫۰۳۴۹
۱٫۰۰	۰٫۴۰۶۳	۰٫۳۶۷۹	−۰٫۰۳۸۴
۱٫۲۵	۰٫۲۳۴۴	۰٫۲۸۶۵	۰٫۰۵۲۱
۱٫۵۰	۰٫۲۸۹۱	۰٫۲۲۳۱	−۰٫۰۶۵۹
۱٫۷۵	۰٫۰۸۹۸	۰٫۱۷۳۸	۰٫۰۸۳۹
۲٫۰۰	۰٫۲۴۴۱	۰٫۱۳۵۳	−۰٫۱۰۸۸
۲٫۲۵	−۰٫۰۳۲۲	۰٫۱۰۵۴	۰٫۱۳۷۶

مثال ۱. مسألهٔ نمونهٔ (۱۰.۴.۶) را با $\lambda = -1$ در نظر می‌گیریم. نتایج عددی در جدول ۶.۶ به‌ازای $h = ۰٫۲۵$ داده شده است. مقدار y_1 با استفاده از روش اویلر، مانند (۶.۴.۶)، به‌دست آمده است. از مقادیر جدول مشاهده می‌شود که قدرمطلق جواب مزاحم افزایش می‌یابد. برای $x_n = ۲٫۲۵$، جواب عددی y_n منفی می‌شود، و علامت آن با هر مرحله متوالیاً تغییر می‌کند.

۲. مسألهٔ زیر را در نظر می‌گیریم

$$y' = x - y^۲ \qquad y(۰) = ۰$$

جواب $Y(x)$ به ازای $x \geq ۰$ اکیداً صعودی است؛ به ازای مقدار بزرگ x، $Y(x) \doteq \sqrt{x}$. گرچه این جواب صعودی است

$$\frac{\partial f(x,y)}{\partial y} = -۲y < ۰ \qquad \text{به‌ازای } y > ۰$$

بنابراین انتظار داریم که روش میانگاهی نوعی ناپایداری نشان دهد. این مطلب توسط نتایج داده‌شده در جدول ۷.۶ با طول گام $h = ۰٫۲۵$، تأیید شده است. جواب عددی در اطراف $x_n = ۲٫۲۵$ رو به کاهش می‌گذارد و در $x_n = ۳٫۲۵$، جواب y_n منفی می‌شود.

۵.۶ روش ذوزنقه‌یی

روش ذوزنقه‌یی را برای معرفی روشهای ضمنی و مفاهیم وابسته به آنها، به‌کار می‌بریم. به‌علاوه، روش ذوزنقه‌یی، به‌دلیل ویژگی پایداری خاصی که دارد، خود سودمند است. برای معرفی قاعدهٔ ذوزنقه‌یی از انتگرالگیری عددی استفاده می‌کنیم.

جدول ۷.۶ مثال (۲) ناپایداری میانگاهی

x_n	y_n	$Y(x_n)$	خطاها
۰٫۲۵	۰٫۰	۰٫۰۳۱۲	۰٫۰۳۱۲
۰٫۵۰	۰٫۱۲۵۰	۰٫۱۲۳۵	−۰٫۰۱۵
۰٫۷۵	۰٫۲۴۲۲	۰٫۲۷۰۰	۰٫۰۲۷۸
۱٫۰۰	۰٫۴۷۰۷	۰٫۴۵۵۵	−۰٫۰۱۵۱
۱٫۲۵	۰٫۶۳۱۴	۰٫۶۵۸۵	۰٫۰۲۷۱
۱٫۵۰	۰٫۸۹۶۳	۰٫۸۵۷۴	−۰٫۰۳۸۹
۱٫۷۵	۰٫۹۷۹۷	۱٫۰۳۷۶	۰٫۰۵۷۹
۲٫۰۰	۱٫۲۹۱۴	۱٫۱۹۳۶	−۰٫۰۹۷۸
۲٫۲۵	۱٫۱۴۵۹	۱٫۳۲۶۴	۰٫۱۸۱
۲٫۵۰	۱٫۷۵۹۹	۱٫۴۴۰۵	−۰٫۳۱۹
۲٫۷۵	۰٫۸۴۷۲	۱٫۵۴۰۴	۰٫۶۹۳
۳٫۰۰	۲٫۷۷۶۰	۱٫۶۳۰۲	−۱٫۱۵
۳٫۲۵	−۱٫۵۰۵۸	۱٫۷۱۲۵	۳٫۲۲

از معادلهٔ دیفرانسیل $Y'(t) = f(t, Y(t))$ روی $[x_n, x_{n+1}]$ انتگرال می‌گیریم تا به‌دست آوریم

$$Y(x_{n+1}) = Y(x_n) + \int_{x_n}^{x_{n+1}} f(t, Y(t))dt$$

قاعدهٔ ذوزنقه‌یی سادهٔ (۱۲.۱.۵) و (۱۴.۱.۵) را به‌کار می‌بریم تا به‌دست آوریم

$$Y(x_{n+1}) = Y(x_n) + \frac{h}{2}[f(x_n, Y(x_n)) + f(x_{n+1}, Y(x_{n+1}))] - \frac{h^3}{12}Y^{(3)}(\xi_n) \qquad (۱.۵.۶)$$

به‌ازای یک مقدار $x_n \leq \xi_n \leq x_{n+1}$. با حذف جملهٔ باقی‌مانده، روش ذوزنقه‌یی

$$y_{n+1} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1})] \qquad n \geq 0 \qquad (۲.۵.۶)$$

را به‌دست می‌آوریم. این یک روش تک‌گامی با مرتبهٔ همگرایی $O(h^2)$ است. و نیز یک مثال ساده از روش ضمنی است، زیرا y_{n+1} در هردو طرف (۲.۵.۶) ظاهر شده است. یک مثال عددی در آخر این بخش داده شده است.

راه‌حلّ بارَستی فرمول (۲.۵.۶) یک معادلهٔ غیرخطی با ریشهٔ y_{n+1} است، و هر یک از روشهای کلی فصل ۲ را می‌توان برای حل آن به‌کار برد. بارست خطی ساده (بخش ۵.۲ را ببینید) مناسبترین و معمولاً کافی است. گیریم $y_{n+1}^{(0)}$ یک حدس اولیهٔ خوب برای جواب y_{n+1} باشد، و تعریف می‌کنیم

$$y_{n+1}^{(j+1)} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(j)}] \qquad j = 0, 1, \ldots \qquad (۳.۵.۶)$$

حدس اولیه معمولاً با استفاده از یک روش روشن به‌دست می‌آید.

برای تحلیل بارست و تعیین شرایطی که تحت آنها، بارست همگرا می‌شود، رابطهٔ (۳.۵.۶) را از (۲.۵.۶) کم می‌کنیم تا به‌دست آوریم

$$y_{n+1} - y_{n+1}^{(j+1)} = \frac{h}{2}[f(x_{n+1}, y_{n+1}) - f(x_{n+1}, y_{n+1}^{(j)})] \qquad (۴.۵.۶)$$

شرط لیپشیتس (۱۲.۲.۶) را به‌کار می‌بریم تا کران آن را با مقدار زیر به‌دست آوریم

$$| y_{n+1} - y_{n+1}^{(j+1)} | \leq \frac{hK}{2} | y_{n+1} - y_{n+1}^{(j)} | \qquad j \geq 0 \qquad (۵.۵.۶)$$

اگر

$$\frac{hK}{2} < 1 \qquad (۶.۵.۶)$$

آنگاه بارَستهای $y_{n+1}^{(j)}$ وقتی $j \to \infty$ به y_{n+1} همگرا می‌شوند. یک برآورد دقیقتر از نرخ همگرایی با کاربستن قضیهٔ مقدار میانگین برای (۴.۵.۶) به‌دست می‌آید:

$$y_{n+1} - y_{n+1}^{(j+1)} \doteq \frac{h}{2} f_y(x_{n+1}, y_{n+1}) \left[y_{n+1} - y_{n+1}^{(j)} \right] \qquad (۷.۵.۶)$$

اغلب در عمل طول گام h و حدس اوّلیهٔ $y_{n+1}^{(0)}$ به گونه‌ای انتخاب می‌شوند که مسلم شود فقط به محاسبهٔ یک بارست نیاز است، و در آن صورت $y_{n+1} \doteq y_{n+1}^{(1)}$ انتخاب می‌شود.

محاسبه y_{n+1} از y_n شامل یک خطای برشی $O(h^3)$ است [(۱.۵.۶) را ببینید]. برای نگهداری این مرتبهٔ دقّت، بارست نهایی $y_{n+1}^{(i)}$ که برای نشان‌دادن y_{n+1} انتخاب شده است، باید در رابطهٔ $| y_{n+1} - y_{n+1}^{(i)} | = O(h^3)$ صدق کند. واگر بخواهیم خطای بارست اهمیت کمتری داشته باشد (همانطورکه در بخش بعد عمل خواهیم کرد)، باید $y_{n+1}^{(i)}$ چنان انتخاب شود که در رابطهٔ زیر صدق نماید.

$$| y_{n+1} - y_{n+1}^{(i)} | = O(h^4) \qquad (۸.۵.۶)$$

برای تحلیل خطا در انتخاب یک حدس اوّلیهٔ y_{n+1}، باید مفهوم **حل موضعی** را معرفی نماییم. این مفهوم برای روشن ساختن اینکه دقیقاً چه جوابی توسط اغلب برنامه‌های رایانه‌یی برای حل معادلات

دیفرانسیل به‌دست می‌آید نیز مهم خواهد بود. گیریم $u_n(x)$ معرّف جواب معادلهٔ دیفرانسیلی باشد که از نقطهٔ (x_n, y_n) عبور می‌کند:

$$u_n'(x) = f(x, u_n(x)) \qquad u_n(x_n) = y_n \qquad (۹.۵.۶)$$

در مرحلهٔ x_n، با دانستن y_n، کمیت $u_{n+1}(x)$ است که سعی در محاسبهٔ آن داریم، نه $Y(x_{n+1})$.
با به‌کار بردن روشی که به (۱.۵.۶) منجر شد به‌ازای مقداری چون $x_n \le \xi_n \le x_{n+1}$، داریم:

$$u_n(x_{n+1}) = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, u_n(x_{n+1}))] - \frac{h^3}{12}u_n^{(3)}(\xi_n) \quad (۱۰.۵.۶)$$

گیریم $\tilde{e}_{n+1} = u_n(x_{n+1}) - y_{n+1}$، که آن را **خطای موضعی** در محاسبهٔ y_{n+1} از y_n می‌نامیم.
رابطهٔ (۲.۵.۶) را از رابطهٔ بالا کم می‌کنیم تا به‌دست آوریم

$$\begin{aligned}\tilde{e}_{n+1} &= \frac{h}{2}[f(x_{n+1}, u_n(x_{n+1})) - f(x_{n+1}, y_{n+1})] - \frac{h^3}{12}u_n^{(3)}(\xi_n) \\ &= \frac{h}{2}f_y(x_{n+1}, y_{n+1})\tilde{e}_{n+1} + O(h\tilde{e}_{n+1}^2) - \frac{h^3}{12}u_n^{(3)}(x_n) + O(h^4)\end{aligned}$$

که در آن دوبار قضیهٔ مقدار میانگین را به‌کار برده‌ایم. می‌توان نشان داد که به‌ازای جمیع مقادیر به اندازهٔ کافی کوچک h،

$$\tilde{e}_{n+1} = O(h^3)$$

به‌طور دقیقتر

$$\tilde{e}_{n+1} = \left[1 - \frac{h}{2}f_y(x_{n+1}, y_{n+1})\right]^{-1} \cdot \left[-\frac{h^3}{12}u_n^{(3)}(x_n) + O(h^4)\right]$$

$$u_n(x_{n+1}) - y_{n+1} = -\frac{h^3}{12}u_n^{(3)}(x_n) + O(h^4) \qquad (۱۱.۵.۶)$$

این رابطه نشان می‌دهد که خطای موضعی، در اصل همان خطای برشی است.
اگر روش اویلر برای محاسبهٔ $y_{n+1}^{(0)}$ به‌کار رفته باشد،

$$y_{n+1}^{(0)} = y_n + hf(x_n, y_n) \qquad (۱۲.۵.۶)$$

در آن صورت $u_n(x_{n+1})$ را می‌توان بسط داده نشان داد که

$$u_n(x_{n+1}) - y_{n+1}^{(0)} = \frac{h^2}{2}u''_n(\xi_n) \qquad x_n \le \xi_n \le x_{n+1} \qquad (۱۳.۵.۶)$$

از ترکیب با رابطهٔ (۱۱.۵.۶)،

$$y_{n+1} - y_{n+1}^{(0)} = O(h^2) \qquad (۱۴.۵.۶)$$

برای برقراری (۸.۵.۶) ازکران (۵.۵.۶) نتیجه می‌شود که باید دو بارست محاسبه و سپس از $y_{n+1}^{(2)}$ برای معرّفی y_{n+1} استفاده کنیم.

با استفاده از روش میانگاهی می‌توانیم حدس اولیهٔ دقیقتری به‌دست آوریم

$$y_{n+1}^{(0)} = y_{n-1} + 2hf(x_n, y_n) \qquad (۱۵.۵.۶)$$

برای برآورد خطا، از همان راهی که به (۱.۴.۶) منجر شد استفاده می‌کنیم تا به‌دست آوریم

$$u_n(x_{n+1}) = u_n(x_{n-1}) + 2hf(x_n, u_n(x_n)) + \frac{h^3}{3}u_n^{(3)}(\eta_n)$$

برای مقداری از $x_{n-1} \le \eta_n \le x_{n+1}$. (۱۱.۵.۶) را از آن کم می‌کنیم،

$$u_n(x_{n+1}) - y_{n+1}^{(0)} = u_n(x_{n-1}) - y_{n-1} + \frac{h^3}{3}u_n^{(3)}(\eta_n)$$

کمیّت $u_n(x_{n-1}) - y_{n-1}$ را می‌توان به طریقی مشابه راهی که در (۱۱.۵.۶) به‌کار برده شد محاسبه نمود و نتیجهٔ مشابهی به‌دست آورد:

$$u_n(x_{n-1}) - y_{n-1} = \frac{h^3}{12}u_n^{(3)}(x_n) + O(h^4) \qquad (۱۶.۵.۶)$$

پس

$$u_n(x_{n+1}) - y_{n+1}^{(0)} = \frac{5h^3}{12}u_n^{(3)}(x_n) + O(h^4)$$

و از ترکیب این رابطه با (۱۱.۵.۶)

$$y_{n+1} - y_{n+1}^{(0)} = \frac{h^3}{2}u_n^{(3)}(x_n) + O(h^4) \qquad (۱۷.۵.۶)$$

با حدس اولیهٔ (۱۵.۵.۶)، یک بارست از (۳.۵.۶) برای برقراری (۸.۵.۶) بر مبنای کران در (۵.۵.۶) کافی خواهد بود.

فرمولهای (۱۲.۵.۶) و (۱۵.۵.۶) را **فرمولهای پیشگو** نامند، و فرمول بارستی ذوزنقه‌یی (۳.۵.۶) **فرمول تصحیح‌کننده** خوانده می‌شود. این دو فرمول روی هم **روش پیشگو ـ تصحیح‌کننده** را تشکیل می‌دهند و پایهٔ روشی هستند که برای کنترل اندازهٔ خطای موضعی می‌توان به‌کار برد. این مطلب در بخش آینده نشان داده شده است.

همگرایی و نتایج پایداری همگرایی روش ذوزنقه‌یی با قضیهٔ ۶.۶ تضمین شده است. فرض کنیم $hk \leq ۱$،

$$\underset{x_۰ \leq x_n \leq b}{\text{Max}} \mid Y(x_n) - y_h(x_n) \mid \leq e^{۲K(b-x_۰)} \mid e_۰ \mid + \left[\frac{e^{۲K(b-x_۰)} - ۱}{K}\right]\left[\frac{h^۲}{۱۲} \parallel Y^{(۳)} \parallel_\infty\right] \quad (۱۸.۵.۶)$$

پیداکردن فرمول خطای مجانبی مشابه پیداکردن آن در روش اویلر است. با فرض $e_۰ = \delta h^۲ + O(h^۳)$، می‌توانیم نشان دهیم

$$\begin{aligned} Y(x_n) - y_h(x_n) &= D(x_n)h^۲ + O(h^۳) \\ D'(x) &= f_y(x, Y(x))D(x) - \frac{۱}{۱۲}Y^{(۴)}(x) \qquad D(x_۰) = \delta_۰ \end{aligned} \quad (۱۹.۵.۶)$$

برای روش ذوزنقه‌یی نیز می‌توان یک فرمول پایداری داد، به شکل استانده، مانند آنچه در (۲۸.۲.۶) و (۲۹.۲.۶) برای روش اویلر داده شده است. اثبات را به عهدهٔ خواننده واگذار می‌کنیم.

همان‌طور که در روش میانگاهی عمل کردیم، می‌توانیم اثر به‌کار بردن روش ذوزنقه‌یی را در مورد معادلهٔ نمونهٔ

$$y' = \lambda y \qquad y(۰) = ۱ \quad (۲۰.۵.۶)$$

که جواب آن $Y(x) = e^{\lambda x}$ است، بررسی کنیم. برای آنکه انگیزش انجام چنین کاری را بیشتر نشان دهیم، روش ذوزنقه‌یی را که برای مسألهٔ خطی زیر به‌کار رفته در نظر می‌گیریم

$$y' = \lambda y + g(x) \qquad y(۰) = Y_۰ \quad (۲۱.۵.۶)$$

یعنی

$$y_{n+۱} = y_n + \frac{h}{۲}[\lambda y_n + g(x_n) + \lambda y_{n+۱} + g(x_{n+۱})] \qquad n \geq ۰ \quad (۲۲.۵.۶)$$

با $y_۰ = Y_۰$. در این صورت روش عددی اختلال یافته را در نظر می‌گیریم

$$z_{n+۱} = z_n + \frac{h}{۲}[\lambda z_n + g(x_n) + \lambda z_{n+۱} + g(x_{n+۱})] \qquad n \geq ۰$$

با $z_۰ = Y_۰ + \varepsilon$. برای تحلیل اثر اختلال در مقدار اولیه گیریم $w_n = z_n - y_n$. با کم‌کردن داریم

$$w_{n+۱} = w_n + \frac{h}{۲}[\lambda w_n + \lambda w_{n+۱}] \qquad n \geq ۰ \qquad w_۰ = \varepsilon \quad (۲۳.۵.۶)$$

این همان روش ذوزنقه‌یی است که برای مسألهٔ نمونه به‌کار رفته است، جز اینکه مقدار اولیه به جای یک، ε است. جواب عددی (۲۳.۵.۶) در واقع ε برابر آن جواب عددی است که برای (۲۰.۵.۶) به‌دست آمده است. بنابراین رفتار جواب عددی مسألهٔ نمونه، رفتار پایداری روش ذوزنقه‌یی را که برای (۲۱.۵.۶) به‌کار رفته‌بود به‌دست می‌دهد.

مسائل نمونهٔ جالب برای ما آنهایی هستند که در آنها λ حقیقی و منفی یا λ مختلط با قسمت حقیقی منفی است. دلیل چنین انتخابی این است که، در چنین صورتی مسألهٔ معادلهٔ دیفرانسیل (۲۱.۵.۶) خوش‌ـوضع است، همان گونه که در (۸.۱.۶) ذکر شده است، و جالبترین حالاتی که مستثنا شده‌اند $\lambda = ۰$ و λ انگاری محض هستند.

با به‌کار بردن قاعدهٔ ذوزنقه‌یی برای (۲۰.۵.۶)،

$$y_{n+۱} = y_n + \frac{h\lambda}{۲}[y_n + y_{n+۱}] \qquad y_۰ = ۱$$

پس

$$y_{n+۱} = \left[\frac{۱ + (h\lambda/۲)}{۱ - (h\lambda/۲)}\right] y_n \qquad n \geq ۰$$

به طریق استقراء

$$y_n = \left[\frac{۱ + (h\lambda/۲)}{۱ - (h\lambda/۲)}\right]^n \qquad n \geq ۰ \qquad (۲۴.۵.۶)$$

به شرطی که $h\lambda \neq ۲$. برای حالت $\lambda < ۰$ حقیقی، می‌نویسیم

$$r = \frac{۱ + (h\lambda/۲)}{۱ - (h\lambda/۲)} = ۱ + \frac{h\lambda}{۱ - (h\lambda/۲)} = -۱ + \frac{۲}{۱ - (h\lambda/۲)}$$

این رابطه نشان می‌دهد که به‌ازای جمیع مقادیر $h > ۰$، داریم $-۱ < r < ۱$. بنابراین

$$\mathrm{Lim}_{n\to\infty} y_n = ۰ \qquad (۲۵.۵.۶)$$

برای کراندار‌بودن $\{y_n\}$ هیچ محدودیتی برای h وجود ندارد، و بنابراین پایداری روش عددی (۲۲.۵.۶) برای همهٔ مقادیر $h > ۰$ و همهٔ مقادیر $\lambda < ۰$، تضمین می‌شود. این یک حکم قویتری است از آنچه که ممکن است برای بیشتر روشهای عددی گفته شود، معمولاً h باید به اندازهٔ کافی کوچک باشد تا پایداری تضمین شود. در برخی کاربردها، مثل معادلات دیفرانسیل سرسخت، این یک نکتهٔ مهمّی است. این ویژگی که (۲۵.۵.۶) به‌ازای جمیع مقادیر $h > ۰$ و همهٔ مقادیر مختلط λ با $\mathrm{Real}(\lambda) < ۰$ برقرار است، **A–پایداری** (یا پایداری مطلق‌ـم.) خوانده می‌شود. ما آن را در بخش ۸.۶ و مسألهٔ ۳۷ بیشتر باز خواهیم کرد.

برآورد خطای ریچاردسن این برآورد خطا در بخش ۴.۵ معرفی، و هم برای پیشگویی خطا [مانند براورد خطا در (۴۲.۴.۵)] و هم برای به‌دست‌آوردن یک روش انتگرالگیری عددی همگرای سریعتر [مانند برونیابی در (۴۰.۴.۵)] به‌کار گرفته شد. این برآورد را می‌توان به هر دو طریق فوق برای حل معادلات دیفرانسیل به‌کار برد، اگرچه ما بیشتر آن را برای پیشگویی خطا به‌کار می‌بریم.

گیریم $y_h(x)$ و $y_{۲h}(x)$ معرّف جوابهای عددی $y' = f(x,y)$ در $[x_۰, b]$ باشند، که با استفاده از روش ذوزنقه‌یی (۲.۵.۶) به‌دست آمده‌اند. در این صورت با استفاده از (۱۹.۵.۶)،

$$Y(x_n) - y_h(x_n) = D(x_n)h^۲ + O(h^۳)$$
$$Y(x_n) - y_{۲h}(x_n) = ۴D(x_n)h^۲ + O(h^۳)$$

از ضرب اولین معادله در ۴ وکم‌کردن آن از دومین معادله، و حل نسبت به $Y(x_n)$ به‌دست می‌آوریم:

$$Y(x_n) = \frac{۱}{۳}[۴y_h(x_n) - y_{۲h}(x_n)] + O(h^۳) \qquad (۲۶.۵.۶)$$

سمت راست فرمول یک مرتبهٔ همگرایی بالاتری نسبت به روش ذوزنقه‌یی دارد ولی باید توجه کرد که برای آن نیاز به محاسبهٔ $y_h(x_n)$ و $y_{۲h}(x_n)$ در تمام نقاط گرهی x_n در $[x_۰, b]$ است.

کاربرد بیشتر فرمول (۲۶.۵.۶) در پیشگویی خطای فراگیر در $y_h(x)$ است. با استفاده از (۲۶.۵.۶)،

$$Y(x_n) - y_h(x_n) = \frac{۱}{۳}[y_h(x_n) - y_{۲h}(x_n)] + O(h^۳)$$

با توجه به رابطهٔ (۱۹.۵.۶)، طرف چپ معادلهٔ فوق $O(h^۲)$ است و بنابراین اولین جملهٔ سمت راست نیز باید $O(h^۲)$ باشد. بنابراین

$$Y(x_n) - y_h(x_n) \doteq \frac{۱}{۳}[y_h(x_n) - y_{۲h}(x_n)] \qquad (۲۷.۵.۶)$$

یک برآورد مجانبی خطاست. این یک شیوهٔ عملی در برآورد خطای فراگیر است، اگرچه روشی که ما برای پیداکردن آن به‌کار بردیم برای گره‌های با طول گام متغیّر، مجاز نخواهد بود.

مثال مسألهٔ زیر را در نظر می‌گیریم

$$y' = -y^۲ \qquad y(۰) = ۱$$

جدول ۸.۶ روش ذوزنقه‌یی و برآورد خطای ریچاردسن

x	$y_{۲h}(x)$	$Y(x)-y_{۲h}(x)$	$y_h(x)$	$Y(x)-y_h(x)$	$\frac{۱}{۳}[y_h(x)-y_{۲h}x)]$
۱٫۰	۰٫۴۸۳۱۴۴	۰٫۰۱۶۸۵۶	۰٫۴۹۶۰۲۱	۰٫۰۰۳۹۷۹	۰٫۰۰۴۲۹۲
۲٫۰	۰٫۳۲۳۶۱۰	۰٫۰۰۹۷۲۳	۰٫۳۳۰۹۹۱	۰٫۰۰۲۳۴۲	۰٫۰۰۲۴۶۰
۳٫۰	۰٫۲۴۳۸۹۰	۰٫۰۰۶۱۱۰	۰٫۲۴۸۵۲۱	۰٫۰۰۱۴۷۹	۰٫۰۰۱۵۴۳
۴٫۰	۰٫۱۹۴۸۳۸	۰٫۰۰۴۱۶۲	۰٫۱۹۸۹۹۱	۰٫۰۰۱۰۰۹	۰٫۰۰۱۰۵۱
۵٫۰	۰٫۱۶۳۶۵۸	۰٫۰۰۳۰۰۸	۰٫۱۶۵۹۳۷	۰٫۰۰۰۷۳۰	۰٫۰۰۰۷۵۹

که دارای جواب $Y(x)=۱/(۱+x)$ است. نتایج جدول ۸.۶ برای طول گامهای $h=$ ۰٫۲۵ و $۲h=$ ۰٫۵۰ هستند. ستون آخر برآورد خطای (۲۷.۵.۶) است که یک برآوردگر دقیق از خطای واقعی $Y(x)-y_h(x)$ است.

۶.۶ الگوریتمی پیشگو-تصحیح‌کننده از مرتبهٔ پایین

در این بخش، الگوریتم نسبتاً ساده‌ای برای حل مسألهٔ مقدار اولیهٔ (۱.۰.۶) آورده شده است. در این الگوریتم از روش ذوزنقه‌یی (۲.۵.۶) استفاده می‌شود و اندازهٔ خطای موضعی با تغییر طول گام h کنترل می‌شود. این روش به علّت مرتبهٔ پایین همگرایی‌اش مفید نیست، ولی برخی از مفاهیم و تکنیکهای موجود در ساختن الگوریتمی پیشگو-تصحیح‌کننده، با طول گام متغیر را نشان می‌دهد.

هر مرحله از x_n به $x_{n+۱}$ عبارت از ساختن $y_{n+۱}$ از y_n و $y_{n-۱}$ است و $y_{n+۱}$ یک جواب تقریبی (۲.۵.۶) بر پایهٔ استفاده از بارستی از (۳.۵.۶) خواهد بود. یک مرحلهٔ عادی دارای طول گام $h=x_n-x_{n-۱}=x_{n+۱}-x_n$ است، پیشگوی میانگاهی (۱۵.۵.۶) به‌کارگرفته شده و خطای موضعی از تفاضل فرمولهای پیشگو و تصحیح‌کننده، تخمین زده شده است. وقتی تغییر طول گام انجام شده باشد، پیشگوی (۱۲.۵.۶) ی اویلر به‌کارگرفته شده است.

استفاده‌کننده از این الگوریتم باید علاوه بر تعیین پارامترهایی که مسألهٔ معادلهٔ دیفرانسیل (۱.۰.۶) را تعریف می‌کنند، چندین پارامتر را نیز معین نماید. طول گام h تغییر می‌کند و کاربر باید h_{Min} و h_{Max} را که طول گام را محدود می‌نمایند معین کند. کاربر همچنین باید یک مقدار اولیه برای h معین کند و این مقدار باید طوری باشد که به‌ازای آن $\frac{۱}{۲}hf_y(x_۰,y_۰)$ از لحاظ قدرمطلق، به اندازهٔ کافی کوچکتر از یک، مثلاً ۰٫۱ باشد. این کمّیت سرعت همگرایی بارستی را در (۳.۵.۶) مشخص می‌کند که بعداً در همین بخش به دنبال یک مثال عددی مورد بحث قرار گرفته است. یک تحمل خطای ε باید داده شود، و طول گام h باید طوری انتخاب شده باشد که خطای برشی

موضعی در هر مرحله در رابطه زیر، صدق کند

$$\frac{۱}{۴}\varepsilon h \leq |\text{ trunc }| \leq \varepsilon h \qquad (۱.۶.۶)$$

این فرمول را کنترل‌کنندهٔ **خطا در طول گام واحد** گویند. اهمیت آن در نزدیک به پایان همین بخش مورد بحث قرار گرفته است.

نمادگذاری بخش قبل ادامه دارد. تابع $u_n(x)$ جواب $y' = f(x,y)$ است که از نقطهٔ (x_n, y_n) می‌گذرد. خطای موضعی که باید برآورد و کنترل شود، فرمول (۱۱.۵.۶) است که خطای پیداکردن $u_n(x_{n+۱})$ با استفاده از روش ذوزنقه‌یی است:

$$u_n(x_{n+۱}) - y_{n+۱} = -\frac{h^۳}{۱۲}u_n^{(۳)}(x_n) + O(h^۴) \qquad h = x_{n+۱} - x_n \qquad (۲.۶.۶)$$

اگر y_n به اندازهٔ کافی به $Y(x_n)$ نزدیک باشد، آنگاه این عبارت یک تقریب خوب برای خطای برشی بسیار وابسته در (۱.۵.۶) است:

$$\frac{-h^۳}{۱۲}Y^{(۳)}(\xi_n)$$

و (۲.۶.۶) تنها کمیّتی است که اطلاعات لازم برای کنترل آن را در دست داریم.

انتخاب طول گام اولیه مسأله عبارت است از پیداکردن مقدار اولیهٔ h و نقطهٔ گرهی $x_۱ = x_۰ + h$ که برای آن $|\ y_۱ - Y(x_۱)\ |$ در کرانهای (۱.۶.۶) صدق کند. با مقدار اولیهٔ h که توسط کاربر داده می‌شود، مقدار $y_h(x_۱)$ با استفاده از پیشگوی (۱۲.۵.۶) اویلر و دوبار بارستن در (۳.۵.۶) به‌دست می‌آید. با همین روند، مقادیر $y_{h/۲}(x_۰ + h/۲)$ و $y_{h/۲}(x_۱)$ نیز محاسبه می‌شوند. روند برونیابی ریچاردسن برای پیشگویی خطا در $y_h(x)$ به‌کار گرفته می‌شود،

$$Y(x_۱) - y_h(x_۱) \doteq \frac{۴}{۳}[y_{h/۲}(x_۱) - y_h(x_۱)] \qquad (۳.۶.۶)$$

اگر این خطا در کرانهای (۱.۶.۶) صدق نماید، مقدار h پذیرفته می‌شود و مرحلهٔ ذوزنقه‌یی معمولی با استفاده از پیشگوی میانگاهی (۱۱.۵.۶) آغاز می‌شود. ولی اگر (۳.۶.۶) در (۱.۶.۶) صدق نکند، آنگاه یک مقدار جدید برای h انتخاب می‌شود.

با استفاده از مقادیر

$$f_۰ = f(x_۰, y_۰) \qquad f_۱ = f\left(x_۰ + \frac{h}{۲}, y_{h/۲}\left(x_۰ + \frac{h}{۲}\right)\right) \qquad f_۲ = f(x_۱, y_{h/۲}(x_۱))$$

تقریب زیر را پیدا می‌کنیم

$$Y^{(۳)}(x_۰) \doteq D_۳y \equiv \frac{(f_۲ - ۲f_۱ + f_۰)}{(h^۲/۴)} \qquad (۴.۶.۶)$$

این تقریب، تقریبی است که در آن از تفاضل منقسم مرتبهٔ دوم $Y' = f(x, y)$ استفاده می‌شود؛ برای مثال لم ۲ بخش ۴.۳ را به‌کار می‌بریم. برای هر طول گام کوچک h، خطای برشی در $x_۰ + h$ به خوبی با فرمول زیر تقریب زده می‌شود:

$$-\frac{h^۳}{۱۲}Y^{(۳)}(\xi_۰) \doteq -\frac{h^۳}{۱۲}D_۳y$$

طول گام جدید به گونه‌ای انتخاب می‌شود که

$$\left| \frac{h^۳}{۱۲}D_۳y \right| = \frac{۱}{۲}\varepsilon h$$

$$h = \sqrt{\frac{۶\varepsilon}{|\, D_۳y \,|}} \qquad (۵.۶.۶)$$

این باید خطای برشی اولیه را تقریباً در وسط ناحیهٔ (۱.۶.۶) به‌عنوان محک خطا درگام واحد، قرار دهد. با این مقدار جدید h، آزمون (۳.۶.۶)، به‌عنوان کنترل درستی عمل، مجدداً تکرار می‌شود.

با انتخاب h به گونه‌ای که وقتی خطای برشی نصف یا دو برابر می‌شود، درکران (۱.۶.۶) صدق کند، عدم نیاز به تغییر طول گام را برای چندین گام تامین کرده‌ایم، به شرطی که مشتق $Y^{(۳)}(x)$ به‌سرعت تغییر ننماید. تغییر طول گام پرهزینه‌تر از یک گام معمولی است، و ما می‌خواهیم نیاز به چنین تغییری را مینیمم نماییم.

گام پیشگو ـ تصحیح‌کنندهٔ معمولی طول گام h در $x_{n+۱} - x_n = x_n - x_{n-۱} = h$ صدق می‌کند. برای پیداکردن مقدار $y_{n+۱}$، قاعدهٔ پیشگوی میانگاهی (۱۵.۵.۶) را به‌کار می‌بریم و در (۳.۵.۶) یک بارست اجرا می‌کنیم. خطای موضعی (۱۱.۵.۶) با استفاده از (۱۷.۵.۶) برآورد می‌شود:

$$-\frac{۱}{۶}(y_{n+۱} - y^{(۰)}_{n+۱}) = -\frac{h^۳}{۱۲}u^{(۳)}_n(x_n) + O(h^۴)$$

$$= [u_n(x_{n+۱}) - y_{n+۱}] + O(h^۴) \qquad (۶.۶.۶)$$

بنابراین خطای موضعی را با استفاده از رابطهٔ

$$\text{trunc} \equiv -\frac{۱}{۶}(y_{n+۱} - y^{(۰)}_{n+۱}) \qquad (۷.۶.۶)$$

اندازه می‌گیریم. اگر خطای برشی در (۱.۶.۶) صدق کند، h تغییر داده نمی‌شود و محاسبات با این روند گام منظم ادامه خواهد یافت. ولی وقتی (۱.۶.۶) برقرار نباشد، مقادیر y_{n+1} و x_{n+1} کنار گذاشته می‌شوند و طول گام جدیدی بر پایهٔ مقدار خطای برشی، انتخاب می‌شود.

تغییر طول گام با استفاده از (۶.۶.۶)،

$$-\frac{u_n^{(3)}(x_n)}{12} \doteq \frac{\text{trunc}}{h_0^3}$$

که h_0 طول گام به‌کار برده‌شده در محاسبهٔ trunc را نشان می‌دهد. برای یک طول گام دلخواه h، خطای موضعی در به‌دست آوردن y_{n+1} با عبارت زیر برآورد می‌شود؛

$$u_n(x_n+h)-y_h(x_{n+1}) \doteq -\frac{h^3}{12}u_n^{(3)}(x_n) \doteq \left[\frac{h}{h_0}\right]^3 \text{trunc}$$

h را به گونه‌ای انتخاب می‌کنیم که

$$\left[\frac{h}{h_0}\right]^3 \mid \text{trunc} \mid = \frac{1}{2}\varepsilon h$$

$$h = \sqrt{\frac{\varepsilon h_0^3}{2 \mid \text{trunc} \mid}} \qquad (۸.۶.۶)$$

y_{n+1} را با استفاده از پیشگوی اویلر و دو مرتبه بارستن در (۳.۵.۶) محاسبه می‌کنیم و سپس به گام پیشگو- تصحیح‌کنندهٔ عادی باز می‌گردیم. برای احتراز از تغییرات سریع در h که می‌تواند به خطاهای قابل‌ملاحظه‌ای منجر شود، مقدار جدید h هرگز نباید بیش از دو برابر مقدار قبلی افزایش یابد. اگر مقدار جدید h از $h_{\min}$ کمتر باشد، آنگاه محاسبه متوقف می‌شود. ولی اگر h بزرگتر از $h_{\max}$ باشد، فقط $h = h_{\max}$ را انتخاب کرده محاسبه را ادامه می‌دهیم. این امر ممکن است مشکلاتی در برداشته باشد، که به دنبال مثال عددی مورد بحث قرار گرفته است.

الگوریتم $\text{Detrap}(f, x_0, y_0, x_{\text{end}}, \epsilon, h, h_{\text{Min}}, h_{\text{Max}}, \text{ier})$

۱. تبصره: مسأله‌ای که باید حل شود $Y'=f(x,y)$ و $Y(x_0)=y_0$ به‌ازای $x_0 \le x \le x_{\text{end}}$ است، با روشی که قبلاً در این بخش گفته شد. مقادیر جواب تقریبی در هر نقطهٔ گرهی درج می‌شود. در مورد پارامتر خطای ε و پارامترهای طول گام قبلاً در این بخش بحث شد. متغیر ier یک شاخص خطا و یک خروجی در پایان الگوریتم است: $\text{ier} = 0$ یعنی یک بازگشت

معمولی؛ ier = ۱ یعنی در بعضی نقاط گرهی $h = h_{\text{Max}}$؛ و ier = ۲، یعنی چون $h < h_{\text{Min}}$، انتگرالگیری خاتمه یافته است زیرا $h \geq h_{\text{Min}}$ یک شرط لازم است.

۲. آغاز کنید: ier := ۰ و loop := ۱،

۳. تبصره: یک مقدار آغازی برای h انتخاب کنید

۴. با استفاده از روش (۲.۵.۶) مقادیر $y_h(x_۰ + h)$، $y_{h/۲}(x_۰ + h/۲)$، $y_{h/۲}(x_۰ + h)$ را محاسبه کنید. در هر مورد، از پیشگوی (۱۲.۵.۶) اویلر استفاده کرده آن را با دو بارست (۳.۵.۶) دنبال کنید

۵. برای خطا در $y_h(x_۰ + h)$ از

$$\text{trunc} := \frac{۴}{۳}[y_{h/۲}(x_۰ + h) - y_h(x_۰ + h)]$$

استفاده کنید.

۶. اگر $|\text{trunc}| \leq \varepsilon h$ یا اگر $|\text{trunc}| \leq \frac{۱}{۴}\varepsilon h$ یا loop= ۲ آنگاه $x_۱ := x_۰ + h$، $y_۱ := y_h(x_۰ + h)$، $x_۱$ و $y_۱$ را درج کنید و به مرحلهٔ ۱۰ بروید.

۷. از (۴.۶.۶)، $D_۳y \doteq y^{(۳)}(x_۰)$ را حساب کنید. اگر $D_۳y \neq ۰$، آنگاه

$$h := \left[\frac{۶\varepsilon}{|D_۳y|}\right]^{۱/۲}$$

اگر $D_۳y = ۰$، آنگاه $h := h_{\text{Max}}$ و loop := ۲.

۸. اگر $h < h_{\text{Min}}$، آنگاه ier := ۲ و خارج شوید. اگر $h > h_{\text{Max}}$ آنگاه $h := h_{\text{Max}}$، ier := ۱، loop := ۲.

۹. به مرحلهٔ ۴ بروید.

۱۰. تبصره: این بخش الگوریتم شامل مرحلهٔ پیشگو ـ تصحیح‌کنندهٔ عادی با کنترل خطاست.

۱۱. بگیرید $x_۲ := x_۱ + h$ و $y_۲^{(۰)} := y_۱ + ۲hf(x_۱, y_۲)$. $y_۲$ را با یک بارستِ (۵.۶.۳) به‌دست آورید.

۱۲. $\text{trunc} := -\frac{۱}{۶}(y_۲ - y_۲^{(۰)})$.

۱۳. اگر $|\text{trunc}| > \epsilon h$ یا $|\text{trunc}| < \frac{۱}{۴}\epsilon h$ به مرحلهٔ ۱۶ بروید.

۱۴. $x_۲$ و $y_۲$ را درج کنید.

۱۵. $x_۰ := x_۱$، $x_۱ := x_۲$، $y_۰ := y_۱$ و $y_۱ := y_۲$. اگر $x_۱ < x_{\text{end}}$، به مرحلهٔ ۱۱ بروید. وگرنه خارج شوید.

۱۶. تبصره: طول گام را تغییر دهید.

۱۷. $x_1 := x_0$، $y_0 := y_1$، $h_0 := h$، و با استفاده از (۸.۶.۶)، h را محاسبه کنید.

۱۸. $h := \text{Min}\{h, 2h_0\}$.

۱۹. اگر $h < h_{\text{Min}}$، آنگاه ier := ۲ و خارج شوید. اگر $h > h_{\text{Max}}$، آنگاه ier := ۱ و $h := h_{\text{Max}}$.

۲۰. $y_1^{(0)} := y_0 + hf(x_0, y_0)$ و y_1 را با انجام دو بارست (۳.۵.۶) محاسبه کنید.

۲۱. x_1 و y_1 را درج کنید.

۲۲. اگر $x_1 < x_{\text{end}}$، به مرحلهٔ ۱۰ برگردید، وگرنه خارج شوید.

در مثال زیر یک اجرای Detrap به‌کار رفته که در آن trunc نیز درج شده است. جزیی به برنامه‌نویسی اضافه شده است تا خطای برشی y_1 در مرحلهٔ ۲۰ پیشگویی شود.

مثال مسألهٔ زیر را در نظر می‌گیریم

$$y' = \frac{1}{1+x^2} - 2y^2 \qquad y(0) = 0 \qquad (9.6.6)$$

جواب درست آن چنین است:

$$Y(x) = \frac{x}{1+x^2}$$

این یک مسألهٔ جالب برای آزمون Detrap است و به خوبی اجرا می‌شود. این معادله در بازهٔ $[0, 10]$ با $h_{\text{Min}} = 0.001$ و $h_{\text{Max}} = 1.0$، $h = 0.1$، $\varepsilon = 0.0005$ حل شده است. جدول ۹.۶ شامل بعضی از نتایج است، از جمله خطای فراگیر درست، و خطای موضعی درست که با True le نشان داده شده است. این خطای آخری با استفاده از یک روش عددی دقیقتر محاسبه شده است. به علّت کمبود جا، فقط قسمتهایی از خروجی را در این جدول آورده‌ایم.

چندین نکته را با این مثال نشان خواهیم داد. نخست، مرحلهٔ ۱۸ برای احتراز از طول گامهایی بسیار بزرگ لازم است. برای مسألهٔ (۹.۶.۶) داریم

$$Y^{(3)}(x) = \frac{-6(x^4 - 6x^2 + 1)}{(1+x^2)^4}$$

که در $x \doteq \pm 0.414$ و $x \doteq \pm 2.414$ صفر می‌شود. بنابراین، بر پایهٔ رابطهٔ (۲.۶.۶) و نزدیک‌بودن $u_n(x)$ به $Y(x)$، خطای موضعی در حل مسألهٔ (۹.۶.۶) در نزدیکی این نقاط بسیار کوچک

جدول ۹.۶ مثال از الگوریتم Detrap

x_n	h	y_n	$Y(x_n)-y_n$	trunc	True le
۰٫۰۲۲۷	۰٫۰۲۲۷	۰٫۰۲۲۶۸۹	۵٫۸۴E − ۶	۵٫۸۴E − ۶	۵٫۸۴E − ۶
۰٫۰۴۵۴	۰٫۰۲۲۷	۰٫۰۴۵۳۰۸	۱٫۱۷E − ۵	۵٫۸۳E − ۶	۵٫۸۴E − ۶
۰٫۰۶۸۱	۰٫۰۲۲۷	۰٫۰۶۷۷۸۷	۱٫۷۴E − ۵	۵٫۷۶E − ۶	۵٫۷۵E − ۶
۰٫۰۹۰۸	۰٫۰۲۲۷	۰٫۰۹۰۰۶۰	۲٫۲۸E − ۵	۵٫۶۲E − ۶	۵٫۶۱E − ۶
۰٫۲۷۲۵	۰٫۰۲۲۷	۰٫۲۵۳۵۹۴	۵٫۱۶E − ۵	۲٫۹۶E − ۶	۲٫۸۵E − ۶
۰٫۳۰۶۵	۰٫۰۳۴۰	۰٫۲۸۰۱۲۵	۵٫۶۶E − ۵	۶٫۷۴E − ۶	۶٫۷۹E − ۶
۰٫۳۴۰۵	۰٫۰۳۴۰	۰٫۳۰۵۰۸۴	۶٫۰۱E − ۵	۶٫۲۱E − ۶	۵٫۷۳E − ۶
۰٫۳۷۴۶	۰٫۰۳۴۰	۰٫۳۲۸۴۱۱	۶٫۱۱E − ۵	۴٫۲۸E − ۶	۳٫۵۴E − ۶
۰٫۴۴۰۸	۰٫۰۶۶۲	۰٫۳۶۹۰۱۹	۵٫۰۵E − ۵	−۶٫۵۶E − ۶	−۵٫۲۰E − ۶
۰٫۵۰۷۰	۰٫۰۶۶۲	۰٫۴۰۳۲۹۷	۲٫۴۴E − ۵	۱٫۰۴E − ۵	−۲٫۱۲E − ۵
۰٫۵۷۳۲	۰٫۰۶۶۲	۰٫۴۳۱۴۶۹	−۲٫۰۳E − ۵	−۲٫۹۲E − ۵	−۴٫۲۱E − ۵
۰٫۶۱۳۸	۰٫۰۴۰۶	۰٫۴۴۵۸۷۹	−۲٫۹۹E − ۵	−۱٫۱۲E − ۵	−۱٫۱۰E − ۵
۱٫۹۵۹۵	۰٫۱۳۵	۰٫۴۰۴۹۸۲	−۱٫۰۲E − ۴	−۱٫۶۴E − ۵	−۱٫۶۷E − ۵
۲٫۰۹۴۲	۰٫۱۳۵	۰٫۳۸۸۹۴۴	−۱٫۰۳E − ۴	−۱٫۷۹E − ۵	−۲٫۱۱E − ۵
۲٫۳۱۷۲	۰٫۲۲۳	۰٫۳۶۳۸۶۴	−۶٫۵۷E − ۵	۱٫۲۷E − ۵	۸٫۱۵E − ۶
۲٫۷۶۳۲	۰٫۴۴۶	۰٫۳۱۹۶۴۹	۳٫۴۴E − ۴	۴٫۴۱E − ۴	۳٫۷۸E − ۴
۳٫۰۶۶۴	۰٫۳۰۳	۰٫۲۹۴۴۴۷	۳٫۲۱E − ۴	۹٫۳۹E − ۵	۸٫۴۱E − ۵
۷٫۶۹۵۹	۰٫۶۷۲	۰٫۱۲۷۳۹۶	۳٫۸۷E − ۴	۸٫۷۷E − ۵	۱٫۱۲E − ۴
۸٫۶۹۵۹	۱٫۰۰۰	۰٫۱۱۳۱۰۰	۳٫۹۶E − ۴	۱٫۷۳E − ۴	۱٫۵۷E − ۴
۹٫۶۹۵۹	۱٫۰۰۰	۰٫۱۰۱۶۲۵	۴٫۲۷E − ۴	۱٫۱۸E − ۴	۱٫۶۸E − ۴
۱۰٫۶۹۵۹	۱٫۰۰۰	۰٫۰۹۲۲۷۳	۴٫۱۱E − ۴	۹٫۴۵E − ۵	۱٫۲۱E − ۴

می‌شود. این امر موجب پیشگویی یک h بسیار بزرگ، در (۸.۶.۶) می‌شود، مقداری که برای نقاط بعدی x_n بسیار بزرگ است. در $x_n = ۲٫۷۶۳۲$، گام ۱۸ الگوریتم، برای احتراز از مقدار بزرگ گمراه‌کنندهٔ h لازم بود. همان‌گونه که می‌توان مشاهده کرد، خطای موضعی در $x_n = ۲٫۷۶۳۲$ به لحاظ مقدار بزرگتر h، بسیار افزایش می‌یابد. کمی بعد از آن، طول گام h کوچک می‌شود تا اندازهٔ خطای موضعی را کاهش دهد.

در تمام نتیجه‌گیریهای این بخش و بخش بعد، برآوردهایی انجام شده‌اند که اگر h به اندازهٔ کافی کوچک می‌بود بسیار دقیق می‌بودند. در بیشتر حالات کمیّت قاطع، در واقع $hf_y(x_n, y_n)$ است،

همانند (۷.۵.۶)، هنگامی که نرخ همگرایی بارستِ (۳.۵.۶) تحلیل می‌شد. در حالت بارست ذوزنقه‌یی (۳.۵.۶)، این نرخ همگرایی چنین است

$$\text{Rate} \doteq \frac{۱}{۲} h f_y(x_n, y_n) \qquad (۱۰.۶.۶)$$

و برای مسألهٔ (۹.۶.۶) این نرخ چنین است:

$$\text{Rate} \doteq -۲hy_n$$

اگر این کمیت نزدیک ۱ باشد چندین بارست برای به‌دست آوردن یک مقدار دقیق $y_{n+۱}$ لازم خواهد بود. مطابق جدول، اندازهٔ این نرخ، وقتی h افزایش می‌یابد، تقریباً بزرگ می‌شود. در $x = ۲٫۳۱۷۲$ این نرخ حدود ۰٫۱۶۲ است. این نرخ به اندازهٔ کافی کوچک به‌نظر می‌آید، ولی خطای موضعی بیشتر از قبل نادقیق است، و این شاید بر اثر بارست نادقیق به‌دست آمده در (۳.۵.۶) باشد. می‌توان الگوریتم را پیچیده‌تر نمود تا مسائل با یک h بسیار بزرگ شناسایی شود، ولی انتخاب یک h_{Max} با اندازه‌ای معقول، نیز مفید خواهد بود.

خطای کلی با دادن یک کران خطا، مشابه کران (۱۸.۵.۶) برای یک طول گام ثابت، مسأله را آغاز می‌کنیم. می‌نویسیم

$$Y(x_{n+۱}) - y_{n+۱} = [Y(x_{n+۱}) - u_n(x_{n+۱})] + [u_n(x_{n+۱}) - y_{n+۱}] \qquad (۱۱.۶.۶)$$

برای آخرین جمله، فرض می‌کنیم که محک خطا در طول گام واحد (۱.۶.۶) برقرار باشد:

$$| u_n(x_{n+۱}) - y_{n+۱} | \leq \varepsilon(x_{n+۱} - x_n) \qquad (۱۲.۶.۶)$$

برای جمله دیگر در (۱۱.۶.۶)، شکلهای معادلات انتگرالی

$$\begin{aligned} Y(x) &= Y(x_n) + \int_{x_n}^{x} f(t, Y(t))dt \\ u_n(x) &= y_n + \int_{x_n}^{x} f(t, u_n(x))dt \qquad x \geq x_n \end{aligned} \qquad (۱۳.۶.۶)$$

را معرفی می‌کنیم. این دو رابطه را از هم کم می‌کنیم و کرانها را با استفاده از شرط لیپشیتس می‌گیریم

$$| Y(x) - u_n(x) | \leq | e_n | + K(x - x_n) \max_{x_n \leq t \leq x} | Y(t) - u_n(t) | \qquad x \geq x_n$$

که در آن $e_n = Y(x_n) - y_n$. با استفاده از این رابطه، به‌دست می‌آوریم

$$| Y(x_{n+1}) - u_n(x_{n+1}) | \leq \frac{1}{1 - K(x_{n+1} - x_n)} | e_n | \qquad (۱۴.۶.۶)$$

می‌نویسیم

$$H = \underset{x_0 \leq x_n \leq b}{\text{Max}} (x_{n+1} - x_n) = \text{Max } h$$

و فرض می‌کنیم

$$HK < 1$$

از ترکیب روابط (۱۱.۶.۶)، (۱۲.۶.۶) و (۱۴.۶.۶)، به‌دست می‌آوریم

$$| e_{n+1} | \leq \frac{1}{1 - HK} | e_n | + \varepsilon H \qquad x_0 \leq x_n \leq b$$

این رابطه، مانند آنچه در قضیهٔ ۳.۶، دیدیم به‌سادگی حل می‌شود و برای یک مقدار مناسب $c > 0$ به‌دست می‌آوریم

$$| Y(x_n) - y_n | \leq e^{c(b-x_0)K} | Y(x_0) - y_0 | + \left[\frac{e^{c(b-x_0)} - 1}{c}\right] \varepsilon \qquad (۱۵.۶.۶)$$

این یک نتیجهٔ اساسی خطا در استفاده از طول گام متغیر، و یک توجیه جزئی از محک خطا در (۱.۶.۶) است.

در بعضی از مواقع، می‌توانیم یک کران واقع بینانه‌تری پیدا کنیم. برای سادگی فرض می‌کنیم برای جمیع مقادیر (x, y)، $f_y(x, y) \leq 0$. روابط (۱۳.۶.۶) را از هم کم می‌کنیم

$$\begin{aligned} Y(x) - u_n(x) &= e_n + \int_{x_n}^{x} [f(t, Y(t)) - f(t, u_n(t))] dt \\ &= e_n + \int_{x_n}^{x} \frac{\partial f(t, \zeta(t))}{\partial y} [Y(t) - u_n(t)] dt \end{aligned}$$

در مرحلهٔ آخر از قضیهٔ مقدار میانگین ۲.۱ استفاده می‌شود، و می‌توان نشان داد که $\partial f(t, \zeta(t))/\partial y$ یک تابع پیوسته از t است. این نشان می‌دهد که $v(x) \equiv Y(x) - u_n(x)$ یک جواب مسألهٔ خطی زیر است

$$v'(x) = \frac{\partial f(x, \zeta(x))}{\partial y} v(x) \qquad v(x_n) = e_n$$

جواب این مسألهٔ خطی همراه با فرض $f_y(x, y) \leq 0$ به‌دست می‌دهد

$$| Y(x_{n+1}) - u_n(x_{n+1}) | \leq | e_n |$$

شرط $f_y(x,y) \leq 0$ مرتبط است با مسائل مقدار اولیهٔ خوش ـ وضع که قبلاً در (۸.۱.۶) به آن اشاره شد. از ترکیب با (۱۱.۶.۶) و (۱۲.۶.۶) به‌دست می‌آوریم

$$| Y(x_{n+1}) - y_{n+1} | \leq | Y(x_n) - y_n | + \varepsilon(x_{n+1} - x_n)$$

این نامعادله را حل می‌کنیم، یک کران واقع‌بینانه‌تر زیر به‌دست می‌آید

$$| Y(x_n) - y_n | \leq | Y(x_0) - y_0 | + \varepsilon(x_n - x_0) \qquad (۱۶.۶.۶)$$

این رابطه تا اندازه‌ای رفتار خوب مثال در جدول ۶ـ۹ را توضیح می‌دهد؛ نتایج نظری بهتری هم وجود دارد. ولی نتایج (۱۵.۶.۶) و (۱۶.۶.۶)، برای توجیه استفاده از آزمون (۱.۶.۶) که خطا را در هر گام کنترل می‌کند، کافی خواهند بود. در دستگاه معادلات $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$، به جای شرط $f_y(x,y) \leq 0$ این شرط که، ویژه مقدارهای ماتریس ژاکوبی $\mathbf{f}_y(x, \mathbf{Y}(x))$ دارای اجزای حقیقی صفر یا منفی باشند، گذاشته می‌شود.

الگوریتم *Detrap*، را به چند طریق می‌توان بهتر کرد. ولی این ساختمان، یک الگوریتم پیشگو ـ تصحیح‌کننده با طول گام متغیر را نشان می‌دهد. خروجیها در یک مجموعهٔ نامناسب از نقاط گرهی x_n درج می‌شوند ولی با یک الگوریتم سادهٔ درونیابی می‌توان این مشکل را حل کرد. پیشگوها را می‌توان بهتر کرد ولی آن نیز الگوریتم را پیچیده‌تر می‌کند. در بخش بعد به بحث در الگوریتمهای پیشگو ـ تصحیح‌کنندهٔ عملیِ فعلاً موجود که اغلب آنها از نظر مرتبه نیز متغیرند باز می‌گردیم.

۷.۶ پیداکردن روشهای چند گامی از مراتب بالاتر

فرمول عمومی روش $p+1$ گامی (۱.۳.۶) از بخش ۳.۶ برای حل مسألهٔ مقدار اولیهٔ (۱.۹.۶) را یادآوری می‌کنیم:

$$y_{n+1} = \sum_{j=0}^{p} a_j y_{n-j} + h \sum_{j=-1}^{p} b_j f(x_{n-j}, y_{n-j}) \qquad n \geq p \quad (۱.۷.۶)$$

یک نظریه برای این روشها در بخش ۳.۶ داده شده است. بعضی روشهای مراتب بالاتر خاص را اکنون به‌دست می‌آوریم. دو راه اساسی برای به‌دست آوردن چنین فرمولهایی وجود دارند: (۱) روش ضرایب نامعین و (۲) انتگرالگیری عددی. روشهای انتگرالگیری عددی امروزه از همه رایجترند ولی دیدگاه روش ضرایب نامعین در تحلیل و مطالعهٔ روشهای عددی هنوز مهم است.

فرمولهای ضمنی را می‌توان از راه بارستی، کاملاً مشابه با (۲.۵.۶) و (۳.۵.۶) در روش ذوزنقه‌یی، حل کرد. در (۱.۷.۶) اگر $b_{-1} \neq 0$، این بارست به شکل زیر تعریف می‌شود:

$$y_{n+1}^{(l+1)} = \sum_{j=0}^{p} [a_j y_{n-j} + h b_j f(x_{n-j}, y_{n-j})] + h b_{-1} f(x_{n+1}, y_{n+1}^{(l)}) \qquad l \geq 0 \tag{۲.۷.۶}$$

این بارست همگراست اگر $hb_{-1}K < 1$، که K ثابت لیپشیتس برای $f(x,y)$ است که در (۱۲.۲.۶) آمده است. نرخ همگرایی خطی، مشابه با (۵.۵.۶) در روش ذوزنقه‌یی، با $hb_{-1}K$ کراندار می‌شود.

ما دنبال یک زوج فرمول، یک تصحیح‌کننده و یک پیشگو، هستیم. فرض کنید که فرمول تصحیح‌کننده از مرتبهٔ mباشد، یعنی در هر مرحله خطای برشی $O(h^{m+1})$ باشد. اغلب فقط یک بارست در (۲.۷.۶) محاسبه می‌شود، و این بدان معناست که پیشگو باید حداقل از درجهٔ $m-1$ باشد تا خطای برشی در $y_{n+1}^{(1)}$ نیز $O(h^{m+1})$ باشد. بحث خطای بارستی روش ذوزنقه‌یی را در بخش ۵.۶، بین (۳.۵.۶) و (۱۷.۵.۶)، ببینید که از پیشگوهای اویلر و میانگاهی استفاده می‌کند. مفاهیم اصلی بدون هیچ تغییر قابل‌ملاحظه‌ای به (۱.۷.۶) و (۲.۷.۶) انتقال می‌یابند.

روش ضرایب نامعین اگر قرار باشد فرمول (۱.۷.۶) از مرتبهٔ $m \geq 1$ باشد، آنگاه با توجّه به قضیهٔ ۵.۶ لازم و کافی است که

$$\sum_{j=0}^{p} a_j = 1$$

$$\sum_{j=0}^{p} a_j (-j)^i + i \sum_{j=-1}^{p} b_j (-j)^{i-1} = 1 \qquad i = 1, 2, \ldots, m \tag{۳.۷.۶}$$

برای یک روش صریح، شرط اضافی $b_{-1} = 0$ لازم است.

برای یک روش ضمنی کلی، $2p+3$ پارامتر $\{a_j, b_j\}$ وجود دارند که باید تعیین شوند و $m+1$ معادله وجود دارند. ممکن است چنین تصور شود که می‌توان تساوی $m+1 = 2p+3$ را نوشت، ولی این امر از نقطهٔ نظر پایداری و همگرایی (۱.۷.۶) بسیار نامعقول است. این نکته در بخش بعد و در مسائل نشان داده شده است. معمولاً، بهترین کار آن است که برای یک روش ضمنی $m \leq p+2$ انتخاب شود. برای یک روش صریح، ملاحظات پایداری اغلب زیاد مهم نیست زیرا این روش معمولاً یک پیشگو برای یک فرمول ضمنی است.

مثال همهٔ روشهای دوگامی مرتبه دوّم را پیدا کنید. فرمول (۱.۷.۶) چنین است

$$y_{n+1} = a_0 y_n + a_1 y_{n-1} + h[b_{-1} f(x_{n+1}, y_{n+1}) + b_0 f(x_n, y_n) + b_1 f(x_{n-1}, y_{n-1})] \quad n \geq 1 \quad (۴.۷.۶)$$

ضرایب باید در رابطهٔ (۳.۷.۶) با $m = 2$ صدق نمایند.

$$a_0 + a_1 = 1 \qquad -a_1 + b_{-1} + b_0 + b_1 = 1 \qquad a_1 + 2b_{-1} - 2b_1 = 1$$

از حل این دستگاه داریم:

$$a_1 = 1 - a_0 \qquad b_{-1} = 1 - \frac{1}{4}a_0 - \frac{1}{2}b_0 \qquad b_1 = 1 - \frac{3}{4}a_0 - \frac{1}{2}b_0 \qquad (۵.۷.۶)$$

که a_0 و b_0 متغیرند. روش میانگاهی حالت ویژه‌ای است که در آن $a_0 = 0$، $b_0 = 2$. ضرایب a_0 و b_0 را می‌توان طوری انتخاب کرد که پایداری بهبود یابد، یا خطای برشی کوچک شود، یا یک فرمول صریح به‌دست آید، یا ترکیبی از این موارد حاصل شود. شرایطی که پایداری و همگرایی را تضمین نمایند (غیر از $0 \leq a_0 \leq 1$ و استفاده از قضیهٔ ۶.۶) پیش از نظریهٔ کلی برای (۱.۷.۶) در بخش بعد، نمی‌توان داد.

با برقراری (۳.۷.۶)، قضیهٔ ۵.۶ ایجاب می‌کند که خطای برشی، برای همهٔ مقادیر $Y(x)$ که $m+1$ بار بر $[x_0, b]$ پیوسته مشتقپذیر است، در

$$T_n(Y) = O(h^{m+1})$$

صدق کند. این شرط برای بیشتر منظورهای نظری و عملی کافی است. ولی گاهی در ساختن الگوریتمهای پیشگو ـ تصحیح‌کننده، ترجیح داده می‌شود که خطا دارای شکل زیر باشد

$$T_n(Y) = d_m h^{m+1} Y^{(m+1)}(\xi_n) \qquad x_{n-p} \leq \xi_n \leq x_{n+1} \qquad (۶.۷.۶)$$

که d_m مقدار ثابتی است مستقل از n. مثالها عبارت‌اند از، روش اویلر، روش میانگاهی و روش ذوزنقه‌یی.

برای به‌دست آوردن فرمول (۶.۷.۶) با فرض برقراری (۳.۷.۶)، ابتدا خطای برشی $T_n(y)$ را به شکل یک انتگرال با استفاده از مفهوم هستهٔ پئانو که در آخرین قسمت بخش ۱.۵ داده شده،

بیان می‌کنیم. $Y(x)$ را حول x_{n-p} بسط می‌دهیم

$$Y(x)=\sum_{i=0}^{m}\frac{(x-x_{n-p})^i}{i!}Y^{(i)}(x_{n-p})+R_{m+1}(x)$$
$$R_{m+1}(x)=\frac{1}{m!}\int_{x_{n-p}}^{x_{n+1}}(x-t)_+^m Y^{(m+1)}(t)dt$$
$$(x-t)_t^m=\begin{cases}(x-t)^m & x\geq t\\ 0 & x<t\end{cases}$$

این بسط را در فرمول (۴.۳.۶) به‌جای $T_n(Y)$ قرار می‌دهیم. با استفاده از این فرض که این روش از مرتبهٔ m است خواهیم داشت

$$T_n(Y)=T_n(R_{m+1})$$
$$=R_{m+1}(x_{n+1})-\left[\sum_{j=0}^{p}a_jR_{m+1}(x_{n-j})+h\sum_{j=-1}^{p}b_jR'_{m+1}(x_{n-j})\right]$$
$$T_n(Y)=\int_{x_{n-p}}^{x_{n+1}}G(t-x_{n-p})Y^{(m+1)}(t)dt \qquad (۷.۷.۶)$$

با هسته پئانو

$$G(s)=\frac{1}{m!}\left\{(x_{p+1}-s)_+^m-\left[\sum_{j=0}^{p}a_j(x_{p-j}-s)_+^m+hm\sum_{j=-1}^{p}b_j(x_{p-j}-s)_+^{m-1}\right]\right\}$$
$$=T_p((x-s)_+^m) \qquad 0\leq s\leq x_{p+1} \qquad (۸.۷.۶)$$

تابع $G(s)$ اغلب تابع نفوذ نامیده می‌شود.

$m+1$ شرط (۱.۷.۶) برای ضرایب (۳.۷.۶) بدین معناست که $r=(2p+3)-(m+1)$ پارامتر آزاد وجود دارد [برای فرمول صریح $r=(2p+3)-(m+2)$]. بنابراین r پارامتر آزاد برای تعیین $G(s)$ در دست است. ما مقادیری برای این پارامترها انتخاب می‌کنیم که $G(s)$ در $[0,x_{p+1}]$ دارای یک علامت شود. و آنگاه با استفاده از قضیهٔ مقدار میانگین، خواهیم داشت

$$T_{n+1}(Y)=Y^{(m+1)}(\xi_n)\int_{x_{n-p}}^{x_{n+1}}G(t-x_{n-p})dt$$

با شرط $x_{n-p} \le \xi_n \le x_{n+1}$. با عملیات بیشتر، (۶.۷.۶) را با

$$d_m = \frac{1}{m!}\int_0^{p+1}\left[\nu^m - \sum_{j=0}^{p} a_j(\nu-j-1)_+^m - m\sum_{j=-1}^{p} b_j(\nu-j-1)_+^{m-1}\right]d\nu \qquad (۹.۷.۶)$$

به‌دست می‌آوریم. باز هم این رابطه وابسته به $G(s)$ است که به‌ازای $0 \le s \le x_{p+1}$ دارای یک علامت است.

مثال فرمول (۴.۷.۶) را در نظر گرفته فرض کنید که فرمول صریح است ($b_{-1} = 0$). در این صورت

$$y_{n+1} = a_0 y_n + a_1 y_{n-1} + h[b_0 f(x_n, y_n) + b_1 f(x_{n-1}, y_{n-1})] \qquad n \ge 1 \qquad (۱۰.۷.۶)$$

با

$$a_1 = 1 - a_0 \qquad b_0 = 2 - \frac{1}{2}a_0 \qquad b_1 = -\frac{1}{2}a_0$$

فرمول قبلی را به‌کار می‌بریم

$$\begin{aligned} G(s) &= \frac{1}{2}\Big[(x_2 - s)^2 - a_0(x_1 - s)_+^2 - a_1(x_0 - s)_+^2 \\ &\qquad - 2hb_0(x_1 - s)_+ - 2hb_1\,(x_0 - s)_+\Big] \\ &= \begin{cases} \frac{1}{2}s[s(1-a_0) + a_0 h] & 0 \le s \le h \\ \frac{1}{2}(x_2 - s)^2 & h \le s \le 2h \end{cases} \end{aligned}$$

شرط اینکه $G(s)$ در فاصلهٔ $[0, 2h]$ دارای یک علامت باشد برقرار است، اگر و تنها اگر $a_0 \ge 0$. پس

$$T_{n+1}(Y) = (\frac{1}{12}a_0 + \frac{1}{3})h^3 Y^{(3)}(\xi_n) \qquad x_{n-1} \le \xi_n \le x_{n+1} \quad (۱۱.۷.۶)$$

توجه کنید که خطای برشی مینیمم است وقتی که $a_0 = 0$؛ بنابراین در میان همهٔ روشهای صریح دوگامی مرتبه دوّم، روش میانگاهی برای داشتن خطای برشی کوچک، از همه بهتر است. برای بحث بیشتر از این مثال، مسألهٔ ۳۱ را ببینید.

روشهای مبتنی بر انتگرالگیری عددی موضوع اصلی مسألهٔ زیر است. معادلهٔ دیفرانسیل را با انتگرالگیری در بازه‌ای چون $[x_{n-r}, x_{n+1}]$ به‌صورت زیر به‌دست می‌آوریم

$$Y(x_{n+1}) = Y(x_{n-r}) + \int_{x_{n-r}}^{x_{n+1}} f(t, Y(t))dt \qquad (۱۲.۷.۶)$$

به‌ازای یک $r \geq 0$ و همهٔ مقادیر $n \geq r$. یک چندجمله‌یی $P(t)$ درست می‌کنیم که انتگرالده $Y'(t) = f(t, Y(t))$ را در یک مجموعه نقاط گرهی $\{x_i\}$ درونیابی کند و سپس از $P(t)$ در بازهٔ $[x_{n-r}, x_{n+1}]$ انتگرال می‌گیریم تا تقریبی برای (۱۲.۷.۶) به‌دست آید. هر سه روش قبلی، اویلر، میانگاهی، و ذوزنقه‌یی، را می‌توان از این طریق به‌دست آورد.

مثال قاعدهٔ انتگرال‌گیری سیمپسون، (۱۳.۱.۵) و (۱۵.۱.۵) را برای معادلهٔ زیر به‌کار می‌بریم

$$Y(x_{n+1}) = Y(x_{n-1}) + \int_{x_{n-1}}^{x_{n+1}} f(t, Y(t))dt$$

این کار چنین نتیجه می‌دهد

$$Y_{n+1} = Y_{n-1} + \frac{h}{3}[Y'_{n-1} + 4Y'_n + Y'_{n+1}] - \frac{h^5}{90}Y^{(5)}(\xi_n)$$

برای مقداری چون $x_{n-1} \leq \xi_n \leq x_{n+1}$. این تقریب بر پایهٔ انتگرالگیری چندجمله‌یی درجهٔ دوم که $Y' = f(t, Y(t))$ را در نقاط گرهی x_{n-1}، x_n و x_{n+1} درونیابی می‌نماید، قرار دارد. برای سادگی نمادگذاری $Y_j = Y(x_j)$ و $\mathbf{Y}'_j = Y'(x_j)$ را به‌کار می‌بریم. از حذف جملهٔ خطا، فرمول ضمنی مرتبهٔ چهار زیر به‌دست می‌آید

$$y_{n+1} = y_{n-1} + \frac{h}{3}[f(x_{n-1}, y_{n-1}) + 4f(x_n, y_n) + f(x_{n+1}, y_{n+1})] \quad n \geq 1 \quad (۱۳.۷.۶)$$

این فرمول یک فرمول مشهور و تصحیح‌کنندهٔ یک الگوریتم پیشگو ـ تصحیح‌کنندهٔ کلاسیک است که به **روش میلن**[1] معروف است. بنابر قضیهٔ ۶.۶ این روش همگراست و

$$\underset{x_0 \leq x_n \leq b}{\text{Max}} \mid Y(x_n) - y_n \mid = O(h^4)$$

ولی این روش، به همان شکلی که برای روش نقطهٔ میانی در بخش ۴.۶ بیان شد پایداری ضعیفی دارد.

1. Milne

روشهای آدامز این روشها، روشهای چندگامی هستند که بیشترین کاربرد را دارند، و برای تولید الگوریتمهای پیشگو ـ تصحیح‌کننده‌ای به‌کار می‌روند که در آنها خطا، با دو تغییر، یکی طول گام h و دیگری مرتبهٔ روش، کنترل می‌شود. این موضوع مفصلاً در این بخش بحث‌شده و یک مثال عددی هم داده شده است.

برای به‌دست آوردن این روشها، از صورت انتگرالی زیر استفاده می‌کنیم

$$Y_{n+1} = Y_n + \int_{x_n}^{x_{n+1}} f(t, Y(t))dt \qquad (۱۴.۷.۶)$$

چندجمله‌ییهایی که $Y'(t) = f(t, Y(t))$ را درونیابی می‌کنند ساخته می‌شوند و سپس از آنها بر بازهٔ $[x_n, x_{n+1}]$ انتگرالگیری می‌شود تا تقریبی برای Y_{n+1} به‌دست آید. مطلب را از فرمولهای صریح یا پیشگو آغاز می‌کنیم.

حالت ۱. روشهای آدامز ـ بشفورت[1] گیریم $P_p(t)$ یک چندجمله‌یی از درجهٔ نابزرگتر از P باشد که $Y'(t)$ را در x_{n-p}، ...، x_n درونیابی نماید. مناسبترین شکل برای $P_p(t)$ فرمول تفاضلی پسرو نیوتن، (۱۱.۳.۳) است که حول x_n بسط داده شده است:

$$\begin{aligned} P_p(t) =& Y'_n + \frac{(t-x_n)}{h} \nabla Y'_n + \frac{(t-x_n)(t-x_{n-1})}{2!h^2} \nabla^2 Y'_n + \dots \\ &+ \frac{(t-x_n)\dots(t-x_{n-p+1})}{p!h^p} \nabla^p Y'_n \end{aligned} \qquad (۱۵.۷.۶)$$

که در آن $\nabla Y'_n = Y'(x_n) - Y'(x_{n-1})$، و تفاضلات پسرو از مراتب بالاتر به‌طور مشابه تعریف می‌شوند [(۱۰.۳.۳) را ببینید]. خطای درونیابی با رابطهٔ زیر داده می‌شود

$$\begin{aligned} E_p(t) &= (t-x_{n-p})\dots(t-x_n)Y'[x_{n-p}, \dots, x_n, t] \\ &= \frac{(t-x_{n-p})\dots(t-x_n)}{(p+1)!} Y^{(p+2)}(\zeta_t) \qquad x_{n-p} \le \zeta_t \le x_{n+1} \end{aligned} \qquad (۱۶.۷.۶)$$

به شرطی که $x_{n-p} \le t \le x_{n+1}$، و $Y(t)$، $(p+2)$ بار پیوسته مشتقپذیر باشد. فرمولهای (۸.۲.۳) و (۱۱.۱.۳) از فصل ۳ را برای اثبات درستی (۱۶.۷.۶) ببینید.

انتگرال $P_p(t)$ چنین داده شده است:

$$\begin{aligned} \int_{x_n}^{x_{n+1}} P_p(t)dt &= hY'_n + \sum_{j=1}^{p} \frac{1}{j!h^j} \nabla^j Y'_n \int_{x_n}^{x_{n+1}} (t-x_n)\dots(t-x_{n+1-j})dt \\ &= h\sum_{j=0}^{p} \gamma_j \nabla^j Y'_n \end{aligned} \qquad (۱۷.۷.۶)$$

1. Bashforth

جدول ۱۰.۶ ضرایب آدامز ـ بشفورت

γ_0	γ_1	γ_2	γ_3	γ_4	γ_5
۱	$\frac{۱}{۲}$	$\frac{۵}{۱۲}$	$\frac{۳}{۸}$	$\frac{۲۵۱}{۷۲۰}$	$\frac{۹۵}{۲۸۸}$

ضرایب γ_j با تبدیل متغیر $s=(t-x_n)/h$ به‌صورت زیر به‌دست می‌آیند

$$\gamma_j=\frac{۱}{j!}\int_0^1 s(s+۱)\dots(s+j-۱)ds \qquad j\geq ۱ \qquad (۱۸.۷.۶)$$

که γ_0 برابر ۱ اختیار می‌شود.

جدول ۱۰.۶ شامل چند مقدار اول γ_j است. گی‌یر[1] (۱۹۷۱، صص ۱۰۴ـ۱۱۱) شامل اطلاعات بیشتری، از جمله یک تابع مولد برای ضرایب است.

خطای برشی در به‌کاربردن چندجمله‌یی درونیاب $P_p(t)$ در (۱۴.۷.۶) با رابطهٔ زیر داده می‌شود

$$T_n(Y)=\int_{x_n}^{x_{n+1}} E_p(t)dt=\int_{x_n}^{x_{n+1}}(t-x_n)\dots(t-x_{n-p})Y'[x_{n-p},\dots,x_n,t]dt$$

با فرض اینکه $Y(x)$ به‌ازای $x_n\leq x\leq x_{x+1}$، $p+2$ بار پیوسته مشتقپذیر است، از قضیهٔ ۳.۳، فصل ۳ نتیجه می‌شود که تفاضل منقسم در انتگرال اخیر یک تابع پیوسته از t است. چون چندجمله‌یی $(t-x_n)\dots(t-x_{n-p})$ در $[x_n,x_{n+1}]$ نامنفی است، می‌توان قضیهٔ مقدار میانگین در انتگرالها (قضیهٔ ۳.۱) را به‌کار برد و برای مقداری چون $x_{n-p}\leq\zeta\leq x_{n+1}$ به‌دست آورد

$$T_n(Y)=Y'[x_{n-p},\dots,x_n,\zeta]\int_{x_n}^{x_{n+1}}(t-x_n)\dots(t-x_{n-p})dt$$

رابطهٔ (۱۸.۷.۶) را برای محاسبهٔ انتگرال و (۱۲.۲.۳) را برای تبدیل تفاضل منقسم به‌کار می‌بریم. پس

$$T_n(Y)=\gamma_{p+1}h^{p+2}Y^{(p+2)}(\xi_n) \qquad x_{n-p}\leq\xi_n\leq x_{n+1} \qquad (۱۹.۷.۶)$$

یک شکل دیگری وجود دارد که برای برآورد خطای برشی بسیار مفید است. با به‌کاربردن قضیهٔ مقدار میانگین (قضیهٔ ۲.۱) و لم مشابه تفاضلات پسرو از لم ۲ از بخش ۴.۳ داریم:

$$T_n(Y)=h\gamma_{p+1}\nabla^{p+1}Y'_n+O(h^{p+2}) \qquad (۲۰.۷.۶)$$

قسمت اصلی این خطا، $h\gamma_{p+1}\nabla^{p+1}Y'_n$، آخرین جمله می‌شد اگر به‌جای $P_p(t)$ چندجمله‌یی درونیاب $P_{p+1}(t)$ به‌کار برده می‌شد. این شکل خطا یک ابزار اصلی در الگوریتمهایی است که مرتبهٔ روش را برای کنترل خطای برشی، تغییر می‌دهد.

1. Gear

جدول ۱۱.۶ فرمولهای آدامز ـ بشفورت

$p=0$	$Y_{n+1}=Y_n+hY'_n+\frac{1}{2}h^2Y''(\xi_n)$
$p=1$	$Y_{n+1}=Y_n+\frac{h}{2}[3Y'_n-Y'_{n-1}]+\frac{5}{12}h^3Y^{(3)}(\xi_n)$
$p=2$	$Y_{n+1}=Y_n+\frac{h}{12}[23Y'_n-16Y'_{n-1}+5Y'_{n-2}]+\frac{3}{8}h^4Y^{(4)}(\xi_n)$
$p=3$	$Y_{n+1}=Y_n+\frac{h}{24}[55Y'_n-59Y'_{n-1}+37Y'_{n-2}-9Y'_{n-3}]+\frac{251}{720}h^5Y^{(5)}(\xi_n)$

با استفاده از معادلات (۱۷.۷.۶) و (۱۹.۷.۶)، معادلهٔ (۱۴.۷.۶) به‌ازای یک $x_{n-p}\le\xi_n\le x_{n+1}$ چنین می‌شود

$$Y_{n+1}=Y_n+h\sum_{j=0}^{p}\gamma_j\,\nabla^j Y'_n+\gamma_{p+1}h^{p+2}Y^{(p+2)}(\xi_n) \qquad (۲۱.۷.۶)$$

روش عددی متناظر آن چنین است

$$y_{n+1}=y_n+h\sum_{j=0}^{p}\gamma_j\,\nabla^j y'_n \qquad n\ge p \qquad (۲۲.۷.۶)$$

در فرمول فوق، $y'_j\equiv f(x_j,y_j)$، و به‌عنوان یک مثال از تفاضلات پسرو داریم $\nabla y'_j=y'_j-y'_{j-1}$. جدول ۱۱.۶ شامل این فرمولها برای $p=0,1,2,3$ است. این معادلات به شکل معمولیتر (۱.۷.۶) نوشته شده‌اند، که در آنها بستگی به مقدار $f(x_{n-j},y_{n-j})$ به‌طور صریح نشان داده شده است. توجه کنید که حالت $p=0$ همان روش اویلر است.

فرمولهای آدامز ـ بشفورت در فرضهای قضیهٔ ۶.۶ صدق می‌نمایند، بنابراین روشهایی همگرا و پایدارند. گذشته از آن، ناپایداری از نوع موجود در روش میانگاهی (۲.۴.۶) و روش (۱۳.۷.۶) سیمپسون را ندارند. اثبات این مطلب در بخش بعد داده شده است و بحث بیشتر تا نزدیک به انتهای آن بخش، هنگام معرّفی مفهوم پایداری نسبی به تعویق انداخته شده است.

حالت ۲. روشهای آدامز ـ مولتن[1] بازهم فرمول انتگرالی (۱۴.۷.۶) را به‌کار می‌بریم ولی $Y'(t)=f(t,Y(t))$ را در $p+1$ نقطهٔ x_{n+1}، ...، x_{n-p+1}، به‌ازای $p\ge 0$، درونیابی می‌نماییم. نحوهٔ عمل دقیقاً مانند روشهای آدامز ـ بشفورت است، و ما فقط نتایج نهایی را می‌آوریم. معادلهٔ (۱۴.۷.۶) به معادلهٔ زیر تبدیل می‌شود

$$Y_{n+1}=Y_n+h\sum_{j=0}^{p}\delta_j\,\nabla^j Y'_{n+1}+\delta_{p+1}h^{p+2}Y^{(p+2)}(\zeta_n) \qquad (۲۳.۷.۶)$$

1. Moulton

جدول ۱۲.۶ ضرایب آدامز ـ مولتن

δ_0	δ_1	δ_2	δ_3	δ_4	δ_5
۱	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$

جدول ۱۳.۶ فرمولهای آدامز ـ مولتن

$p=0$	$Y_{n+1}=Y_n+hY'_{n+1}-\frac{1}{2}h^2Y''(\zeta_n)$
$p=1$	$Y_{n+1}=Y_n+\frac{h}{2}[Y'_{n+1}+Y'_n]-\frac{1}{12}h^3Y^{(3)}(\zeta_n)$
$p=2$	$Y_{n+1}=Y_n+\frac{h}{12}[5Y'_{n+1}+8Y'_n-Y'_{n-1}]-\frac{1}{24}h^4Y^{(4)}(\zeta_n)$
$p=3$	$Y_{n+1}=Y_n+\frac{h}{24}[9Y'_{n+1}+19Y'_n-5Y'_{n-1}+Y'_{n-2}]-\frac{19}{720}h^5Y^{(5)}(\zeta_n)$

با $x_{n-p+1}\le\zeta_n\le x_{n+1}$. ضرایب δ_j با رابطهٔ زیر تعریف شده‌اند

$$\delta_j=\frac{1}{j!}\int_0^1(s-1)s(s+1)\dots(s+j-2)ds \qquad j\ge 1 \quad (۲۴.۷.۶)$$

با $\delta_0=1$، و چند مقدار آنها در جدول ۱۲.۶ داده شده‌اند. خطای برشی را درست مانند (۲۰.۷.۶) می‌توان به شکل زیر نوشت

$$T_{n+1}(Y)=h\delta_{p+1}\nabla^{p+1}Y'_{n+1}+O(h^{p+3}) \qquad (۲۵.۷.۶)$$

روش عددی متناظر با (۲۳.۷.۶) چنین است

$$y_{n+1}=y_n+h\sum_{j=0}^{p}\delta_j\nabla^j y'_{n+1} \qquad n\ge p-1 \qquad (۲۶.۷.۶)$$

که در آن، مانند قبل، $y'_j\equiv f(x_j,y_j)$. جدول ۱۳.۶ شامل چند فرمول مرتبهٔ پایین برای $p=0,1,2,3$ است. توجه کنید که حالت $p=1$ همان روش ذوزنقه‌یی است.

فرمول (۲۶.۷.۶) یک روش ضمنی است، و بنابراین برای حل آن با روش بارستی یک پیشگو لازم است. مفاهیم اصلی در اینجا درست همان مفاهیم موجود در بخش ۵.۶ برای حل بارستی روش ذوزنقه‌یی است. اگر یک الگوریتم پیشگو ـ تصحیح‌کننده با مرتبهٔ ثابت خواسته شده باشد، و اگر فقط یک بارست باید حساب شود، آنگاه در یک فرمول آدامز ـ مولتن از مرتبهٔ $m\ge 2$ می‌توان یک پیشگو از مرتبهٔ m یا $m-1$ به‌کار برد. مزیّت به‌کارگیری پیشگوی مرتبهٔ $m-1$ آن است که پیشگو و تصحیح‌کننده هر دو از مقادیر مشتق در نقاط گرهی x_n، x_{n-1}، ...، x_{n-m+2}،

استفاده می‌کنند. برای مثال، فرمول مرتبهٔ دوم آدامز- مولتن با فرمول مرتبهٔ یک آدامز- بشفورت به‌عنوان پیشگو، درست همان روش ذوزنقه‌یی با پیشگوی اویلر است. این موضوع در بخش ۵.۶ مورد بحث واقع و نشان داده شد که مناسب است؛ در هر دو روش تنها از یک مقدار قبلی مشتق، $f(x_n, y_n)$ استفاده می‌شود.

یک مثال جالبتر، روش مرتبه چهار زیر است؛

$$y_{n+1}^{(0)} = y_n + \frac{h}{24}[23f(x_n, y_n) - 16f(x_{n-1}, y_{n-1}) + 5f(x_{n-2}, y_{n-2})]$$
$$y_{n+1}^{(j+1)} = y_n + \frac{h}{24}[9f(x_{n+1}, y_{n+1}^{(j)}) + 19f(x_n, y_n) - 5f(x_{n-1}, y_{n-1}) + f(x_{n-2}, y_{n-2})]$$
(۲۷.۷.۶)

معمولاً فقط یک بارست محاسبه می‌شود، گرچه این امر، شکل خطای برشی در (۲۳.۷.۶) را تغییر می‌دهد. گیریم $u_n(x)$ جواب $y' = f(x, y)$ گذرنده بر نقطهٔ (x_n, y_n) باشد. پس خطای برشی در به‌کاربردن تقریب $y_{n+1}^{(1)}$ عبارتست از،

$$u_n(x_{n+1}) - y_{n+1}^{(1)} = [u_n(x_{n+1}) - y_{n+1}] + [y_{n+1} - y_{n+1}^{(1)}]$$

با استفاده از (۲۳.۷.۶) و بسط خطای بارستی،

$$u_n(x_{n+1}) - y_{n+1}^{(1)} = \delta_4 h^5 u_n^{(5)}(x_n) + \frac{3h}{8} f_y(x_n, y_n)(y_{n+1} - y_{n+1}^{(0)}) + O(h^6)$$
(۲۸.۷.۶)

دو جملهٔ اول طرف راست، پس از علامت تساوی، از مرتبهٔ h^5 اند. اگر یا (۱) بارستهای بیشتری محاسبه شود، یا (۲) پیشگوی از مرتبهٔ بالاتری به‌کار گرفته شود، آنگاه قسمت اصلی خطای برشی فقط $\delta_4 h^5 u_n^{(5)}(x_n)$ خواهد شد. و این از نظر برآورد خطا، حالت مطلوبتری است.

برای روشهای از مرتبهٔ مشابه، خطای برشی فرمولهای آدامز- مولتن به مراتب کوچکتر از خطای آدامز- بشفورت است. برای مثال، خطای برشی فرمول مرتبهٔ چهار آدامز- مولتن ۰٫۰۷۶ خطای برشی فرمول آدامز- بشفورت مرتبهٔ چهار است. این است دلیل عمدهٔ استفاده از روشهای ضمنی، اگرچه، ملاحظات دیگری نیز مطرح است. همچنین توجه کنید که خطای برشی فرمول مرتبهٔ چهار آدامز- مولتن بیش از دو برابر خطای برشی روش سیمپسون است. دلیل استفاده از فرمول آدامز- مولتن این است که ویژگیهای پایداری آن خیلی بهتر از ویژگیهای پایداری روش سیمپسون است.

جدول ۱۴.۶ مثال عددی برای روش آدامز- مولتن

x	خطا برای $h = ۰٫۱۲۵$	خطا برای $h = ۰٫۰۶۲۵$	نسبت
۲٫۰	۲٫۰۷E − ۵	۱٫۲۱E − ۶	۱۷٫۱
۴٫۰	۲٫۲۱E − ۶	۱٫۲۰E − ۷	۱۸٫۳
۶٫۰	۳٫۷۴E − ۷	۲٫۰۰E − ۸	۱۸٫۷
۸٫۰	۱٫۰۰E − ۷	۵٫۲۴E − ۹	۱۹٫۱
۱۰٫۰	۳٫۵۸E − ۸	۱٫۸۳E − ۹	۱۹٫۶

مثال روش (۲۷.۷.۶) برای حل معادلهٔ

$$y' = \frac{۱}{۱+x^۲} - ۲y^۲ \qquad y(۰) = ۰ \qquad (۲۹.۷.۶)$$

که دارای جواب $Y(x) = x/(۱+x^۲)$ است، به‌کار رفته است. مقادیر اوّلیهٔ $y_۱$، $y_۲$ و $y_۳$ را برای ساده‌کردن مثال، مقادیر درست انتخاب کرده‌اند. مقادیر جوابها با دو مقدار h محاسبه شده‌اند، و خطاهای حاصل در چند نقطهٔ گرهی در جدول ۱۴.۶ داده شده‌اند. ستونی که با نسبت مشخص شده است نسبت خطا با $h = ۰٫۱۲۵$ به خطا در $h = ۰٫۰۶۲۵$ است. توجه نمایید که مقادیر نسبت نزدیک به ۱۶ هستند، که باید انتظار می‌داشتیم، زیرا روش (۲۷.۷.۶) از مرتبهٔ چهار است.

روشهای متغیّر- مرتبه در حال حاضر مشهورترین الگوریتم‌های پیشگو-اصلاحگر خطای برشی را هم با تغییر طول گام و هم با تغییر مرتبهٔ روش کنترل می‌نمایند، و در تمام این الگوریتم‌ها از فرمولهای خانوادهٔ آدامز استفاده می‌کنند. اولین برنامه‌های رایانه‌یی از این نوع، که به‌طور گسترده مورد استفاده قرار گرفته بود DIFSUB از گی‌یر (۱۹۷۱، صص ۱۵۸-۱۶۶) و برنامه‌های کرو[1] بودند. پس از آن، چنین برنامه‌های متغیر- مرتبهٔ آدامز نوشته شدند که، با GEAR از هایندمارش[2] (۱۹۷۴) و با DE/STEP از شمپاین و گوردن (۱۹۷۵) از متداولترین آنها هستند. برنامه GEAR باز هم بهبود یافت و به برنامه LSODE تبدیل شد که قسمتی از یک بسته برنامهٔ بزرگی است که ODEPACK خوانده می‌شود [هایندمارش (۱۹۸۳) را برای توضیح ببینید]. برنامه DE/STEP بهبود بیشتری یافته و به برنامهٔ DDEABM تبدیل شد و قسمتی از یک بستهٔ کلی دیگری به‌نام DEPAC است [برای توضیح به شمپاین و واتس[3] (۱۹۸۰) مراجعه کنید]. در همهٔ حالات، برنامه‌هایی که قبلاً گفتیم برآورد خطایی را به‌کار می‌برند که بر پایهٔ فرمولهای (۲۳.۷.۶) و (۲۵.۷.۶) استوار است، اگرچه این

1. Krogh 2. Hindmarsh 3. Watts

فرمولها [و فرمولهای (۲۲.۷.۶) و (۲۶.۷.۶)] برای طول گام متغیر، ممکن است نیاز به اصلاح داشته باشند. برنامه‌ها از بعضی جنبه‌های مهم ولی فنّی تغییر می‌کنند، از جمله تعداد بارستهای اصلاحگر آدامز ـ مولتن و شکلی که در آن اطلاعات گذشته دربارهٔ مشتق f پیش کشیده می‌شود و شکل درونیابی جواب برای نقطهٔ گرهی خروجی جاری. به علّت کمبود جا، و پیچیدگی موضوعات وارده، بیش از این در موضوع تفاوتهای این برنامه‌ها بحث نمی‌کنیم. برای بعضی ملاحظات دیگر در مورد این برنامه‌ها گوپتا[1] و همکاران (۱۹۸۵، صص ۱۶ـ۱۹) را ببینید.

با مجاز دانستن تغییر مرتبه، هیچ مشکلی برای پیدا کردن مقادیر اولیهٔ روشهای مرتبهٔ بالاتر آدامز نخواهد بود. برنامه‌ها با فرمول مرتبهٔ دوم ذوزنقه‌یی با پیشگوی اویلر آغاز می‌شوند، آنگاه با فراهم‌آمدن مقادیر اولیه افزایش می‌یابند. اگر جواب به سرعت تغییر یابد، آنگاه برنامه معمولاً فرمولی از مرتبهٔ پایین انتخاب می‌کند، حال آنکه برای یک جواب با تغییرات آهسته‌تر و هموارتر، مرتبه معمولاً بالاتر است.

در برنامهٔ DEی شمپاین و گوردن (۱۹۸۵، صص ۱۸۶ـ۲۰۹)، خطای برشی در x_{n+1} که $trunc$ خوانده شده، باید در رابطهٔ زیر صدق کند

$$| \text{trunc}_j | \leq \text{ABSERR} + \text{RELERR} * | y_{n,j} | \qquad (۳۰.۷.۶)$$

این رابطه باید برای هر مؤلفهٔ خطای برشی و هر $y_{n,j}$، مؤلفهٔ متناظر جواب y_n دستگاه معادلات دیفرانسیل داده شده برقرار باشد. مقادیر ABSERR و RELERR را کاربر می‌دهد. مقدار trunc با فرض یکنواخت‌بودن فاصله‌ها، تقریباً با

$$\text{trunc} \doteq h\delta_{p+1} \nabla^{p+1} y'_{n+1}$$

داده می‌شود. این عبارت، خطای برشی فرمول p مرحله‌یی زیر است

$$y^{(p)}_{n+1} = y_n + h \sum_{j=0}^{p} \delta_j \nabla^j y'_{n+1}$$

وقتی trunc در آزمون (۳۰.۷.۶) صدق کند، مقدار y_{n+1} چنین می‌شود

$$y_{n+1} \equiv y^{(p+1)}_{n+1} = y^{(p)}_{n+1} + \text{trunc} \qquad (۳۱.۷.۶)$$

بنابراین خطای برشی واقعی $O(h^{P+2})$ است، و چنانچه با (۳۰.۷.۶) ترکیب شود، می‌توان نشان داد که خطای برشی y_{n+1} در یک محک خطا در گام واحد صدق می‌نماید، که مشابه خطای

1. Gupta

برشی است که با (۱.۶.۶) برای الگوریتم Detrap در بخش ۶.۶ بیان شده است. برای یک بحث مفصل شمپاین و گوردن (۱۹۷۵، ص ۱۰۰) را ببینید.

برنامهٔ DE (و جانشین آن DDEABM) در کنترل خطا، از جمله انتخاب مرتبه و طول گام، خیلی پیچیده است. نمی‌توان از آن به‌قدر کافی در این جای محدود، در این کتاب بحث کرد ولی بهترین مرجع، کتاب شمپاین و گوردن (۱۹۷۵) است که به الگوریتمهای متغیر ـ مرتبهٔ آدامز اختصاص یافته است. برنامه‌های DE و DDEABM از لحاظ کنترل خطا و راحتی کاربر خوب طراحی شده‌اند. هرکدام به شکل قابل‌حمل هم نوشته شده است، و به‌طورکلی برنامه‌ای است که برای حل معادلات دیفرانسیل خیلی توصیه شده است.

مثال مسألهٔ زیر را در نظر می‌گیریم

$$y' = \frac{y}{۴}\left(۱ - \frac{y}{۲۰}\right) \qquad y(۰) = ۱ \qquad (۳۲.۷.۶)$$

که دارای جواب زیر است

$$Y(x) = \frac{۲۰}{(۱ + ۱۹e^{-(x/۴)})}$$

برای حل این مسأله DDEABM با مقادیر خروجی در $x = ۲, ۴, ۶, \ldots, ۲۰$ به‌کار برده شده است. سه مقدار ABSERR به‌کار برده شده و در هر حالت RELERR=0. خطاهای کلی واقعی در جدول ۱۵.۶ داده شده‌اند. ستونی که با NFE مشخص شده تعداد محاسبات $f(x, y)$ را که از $x_۰$ شروع می‌شود و برای به‌دست آوردن $y_h(x)$ لازم است، نشان می‌دهد.

خطای کلّی (فراگیر) برنامه‌های رایانه‌یی خودکار که قبلاً توضیح داده شدند خطای موضعی یا خطای برشی را کنترل می‌کنند. ولی خطای کلی جواب را کنترل نمی‌کنند. معمولاً خطای برشی

جدول ۱۵.۶ مثال برنامهٔ خودکار DDEABM

	ABSERR = $۱۰^{-۳}$		ABSERR = $۱۰^{-۶}$		ABSERR = $۱۰^{-۹}$	
x	خطا	NFE	خطا	NFE	خطا	NFE
۴٫۰	−۳٫۲۶E − ۵	۱۵	۱٫۲۴E − ۶	۲۸	۲٫۸۶E − ۱۰	۵۲
۸٫۰	۶٫۰۰E − ۴	۲۱	۳٫۸۶E − ۶	۴۲	−۱٫۹۸E − ۹	۷۶
۱۲٫۰	۱٫۷۰E − ۳	۲۵	۴٫۹۳E − ۶	۵۴	−۲٫۴۱E − ۹	۱۰۲
۱۶٫۰	۹٫۱۳E − ۴	۳۱	۳٫۷۳E − ۶	۶۴	−۱٫۸۶E − ۹	۱۲۴
۲۰٫۰	۹٫۱۶E − ۴	۳۷	۱٫۷۹E − ۶	۷۴	−۹٫۵۸E − ۱۰	۱۳۸

در این برنامه‌ها آنقدر کوچک نگهداشته می‌شوند که خطای کلی در یک محدوده، قابل‌قبول است، گرچه این امر تضمین نمی‌شود. دلایل برای این خطای کلّی کوچک خیلی شبیه دلایلی است که در بخش ۶.۶ تشریح شده است؛ به‌ویژه (۱۵.۶.۶) و (۱۶.۶.۶) را به یاد آورید.

خطای کلی را می‌توان کنترل کرد، ما مثالی در ذیل می‌آوریم. ولی حتی با یک برآورد خطای کلی، نمی‌توانیم آن را برای بسیاری از معادلات کنترل نماییم. علّت آن این است که خطای کلی ترکیب شده است از اثرات همهٔ خطاهای برشی گذشته، و کاهش طول گام در مرحلهٔ جاری خطاهای مراحل گذشته را تغییر نخواهد داد. معمولاً، اگر خطای کلی بسیار بزرگ باشد، معادله باید دوباره با طول گام کوچکتری حل شود.

برای نمایش خطای کلی چندین روش ارائه شده است. یکی از آنها، برونیابی ریچاردسن، در بخش ۱۰.۶ برای روش رونگه ـ کوتا توضیح داده شده است. در زیر روش دیگری برای روش بخش ۶.۶ ارائه خواهیم داد. برای یک بازنگری کلی ازموضوع به اسکیل[1] (۱۹۸۶) مراجعه کنید.

برای روش ذوزنقه‌یی، جواب درست $Y(x)$ در

$$Y(x_{n+1})=Y(x_n)+\frac{h}{2}[f(x_n,Y(x_n))+f(x_{n+1},Y(x_{n+1}))]-\frac{h^3}{12}Y^{(3)}(\xi_n)$$

صدق می‌کند که $h=x_{n+1}-x_n$ و $x_n\le\xi_n\le x_{n+1}$. اگر قاعدهٔ ذوزنقه‌یی

$$y_{n+1}=y_n+\frac{h}{2}[f(x_n,y_n)+f(x_{n+1},y_{n+1})]$$

را از آن کم کنیم خواهیم داشت

$$e_{n+1}=e_n+\frac{h}{2}\{[f(x_n,y_n+e_n)-f(x_n,y_n)]$$
$$+[f(x_{n+1},y_{n+1}+e_{n+1})-f(x_{n+1},y_{n+1})]\}-\frac{h^3}{12}Y^{(3)}(\xi_n)\quad(۳۳.۷.۶)$$

با فرض $e_n=Y(x_n)-y_n$، $n\ge 0$. این معادله، معادلهٔ خطا برای روش ذوزنقه‌یی است و ما سعی می‌کنیم آن را به‌طور تقریبی برای محاسبهٔ e_{n+1} حل کنیم.

به الگوریتم Detrap از بخش ۶.۶ بازمی‌گردیم، به‌جای عبارت برشی در (۳۳.۷.۶) متغیر Trunc را که در Detrap محاسبه شده می‌گذاریم. سپس (۳۳.۷.۶) را نسبت به $\hat{e}_{n+1}$ که تقریبی برای خطای کلی درست e_{n+1} است، حل می‌کنیم. می‌توانیم این معادله را نسبت به $\hat{e}_{n+1}$ با روشهای

1. Skeel

ریشه‌یابی گوناگون حل کنیم، ما بارست‌های نقطه ـ ثابت ساده را به‌ازای $j \geq 0$ به‌کار می‌بریم

$$\hat{e}_{n+1}^{(j+1)} = \hat{e}_n + \frac{h}{2}\{[f(x_n, y_n + \hat{e}_n) - f(x_n, y_n)] + [f(x_{n+1}, y_{n+1} + \hat{e}_{n+1}^{(j)}) - f(x_{n+1}, y_{n+1})]\} + \text{trunc} \quad (۳۴.۷.۶)$$

$\hat{e}_{n+1}^{(0)} = \hat{e}_n$ را به‌کار می‌بریم و چون این فقط به‌منظورهای توضیحی است، در (۳۴.۷.۶) چندین بارست را انجام می‌دهیم. این فکر ساده خیلی به روشهای تصحیح تفاضلی اسکیل (۱۹۸۶) وابسته است.

مثال محاسبات جدول ۹.۶ برای Detrap در حل معادلهٔ (۹.۶.۶) را تکرار می‌کنیم. ما همان پارامترها را برای Detrap به‌کار می‌بریم. نتایج در جدول ۱۶.۶ برای همان مقادیر x_n در جدول ۹.۶ داده شده‌اند. این نتایج نشان می‌دهند که e_n و $\hat{e}_n$ تقریباً همیشه به اندازهٔ معقولی، البته از نظر قدر مطلق، بهم نزدیک‌اند. تقریب $e_n \doteq \hat{e}_n$ در نزدیکی $x = 0.5$ ضعیف است و این به‌دلیل ضعف برآورد خطای برشی در Detrap است. حتی این نتایج ضعیف هم در این مسأله از بین می‌روند و برای مقادیر بزرگتر x_n، تقریب $e_n \doteq \hat{e}_n$ هنوز سودمند است.

۸.۶ نظریهٔ همگرایی و پایداری برای روشهای چندگامی

در این بخش یک نظریهٔ کامل همگرایی و پایداری برای روش چند گامی زیر ارائه شده است

$$y_{n+1} = \sum_{j=0}^{p} a_j y_{n-j} + h \sum_{j=-1}^{p} b_j f(x_{n-j}, y_{n-j}) \qquad x_{p+1} \leq x_{n+1} \leq b \quad (۱.۸.۶)$$

در این بررسی، عملیات بخش ۳.۶ تعمیم داده خواهد شد، و ابزار ریاضی لازم به‌وجود خواهد آمد تا بررسی شود که آیا ضعیف ـ پایداریِ روش (۱.۸.۶)، فقط به‌دلیل ناپایداری از نوع مرتبط با روش میانگاهی است یا به دلیل دیگر.

مطلب را با چند تعریف آغاز می‌کنیم. مفهوم پایداری که با روش اویلر معرفی شد [(۲۸.۲.۶) و (۲۰.۲.۶) را ببینید]، اکنون تعمیم می‌یابد. گیریم $\{y_n \mid 0 \leq n \leq N(h)\}$ جواب (۱.۸.۶)، برای معادلهٔ دیفرانسیلی چون $y' = f(x, y)$، برای همهٔ مقادیر به اندازه کافی کوچک h، $h \leq h_0$ باشد. یادآور می‌شویم که $N(h)$ بزرگترین زیرنمایهٔ N است که برای آن $x_N \leq b$. به‌ازای هر مقدار $h \leq h_0$ اختلالی در مقادیر اولیهٔ $y_0, \ldots, y_p$ به مقادیر جدید z_0، ...، z_p به‌وجود می‌آوریم

جدول ۱۶.۶ محاسبه خطای کلی برای Detrap

x_n	h	e_n	$\hat{e}_n$	trunc
۰٫۰۲۲۷	۰٫۰۲۲۷	۵٫۸۴E − ۶	۵٫۸۳E − ۶	۵٫۸۴E − ۶
۰٫۰۴۵۴	۰٫۰۲۲۷	۱٫۱۷E − ۵	۱٫۱۶E − ۵	۵٫۸۳E − ۶
۰٫۰۶۸۱	۰٫۰۲۲۷	۱٫۷۴E − ۵	۱٫۷۳E − ۵	۵٫۷۶E − ۶
۰٫۰۹۰۸	۰٫۰۲۲۷	۲٫۲۸E − ۵	۲٫۲۸E − ۵	۵٫۶۲E − ۶
۰٫۲۷۲۵	۰٫۰۲۲۷	۵٫۱۶E − ۵	۵٫۱۹E − ۵	۲٫۹۶E − ۶
۰٫۳۰۶۵	۰٫۰۳۴۰	۵٫۶۶E − ۵	۵٫۶۶E − ۵	۶٫۷۴E − ۶
۰٫۳۴۰۵	۰٫۰۳۴۰	۶٫۰۱E − ۵	۶٫۰۵E − ۵	۶٫۲۱E − ۶
۰٫۳۷۴۶	۰٫۰۳۴۰	۶٫۱۱E − ۵	۶٫۲۱E − ۵	۴٫۲۸E − ۶
۰٫۴۴۰۸	۰٫۰۶۶۲	۵٫۰۵E − ۵	۵٫۰۴E − ۵	−۶٫۵۶E − ۶
۰٫۵۰۷۰	۰٫۰۶۶۲	۲٫۴۴E − ۵	۳٫۵۷E − ۵	−۱٫۰۴E − ۵
۰٫۵۷۳۲	۰٫۰۶۶۲	−۲٫۰۳E − ۵	۴٫۳۴E − ۶	−۲٫۹۲E − ۵
۰٫۶۱۳۸	۰٫۰۴۰۶	−۲٫۹۹E − ۵	−۶٫۷۶E − ۶	−۱٫۱۲E − ۵
۱٫۹۵۹۵	۰٫۱۳۵	−۱٫۰۲E − ۴	−۸٫۷۶E − ۵	−۱٫۶۴E − ۵
۲٫۰۹۴۲	۰٫۱۳۵	−۱٫۰۳E − ۴	−۸٫۶۸E − ۵	−۱٫۷۹E − ۵
۲٫۳۱۷۲	۰٫۲۲۳	−۶٫۵۷E − ۵	−۵٫۰۸E − ۵	۱٫۲۷E − ۵
۲٫۷۶۳۲	۰٫۴۴۶	۳٫۴۴E − ۴	۳٫۱۷E − ۴	۴٫۴۱E − ۴
۳٫۰۶۶۴	۰٫۳۰۳	۳٫۲۱E − ۴	۲٫۹۶E − ۴	۹٫۳۹E − ۵
۷٫۶۹۵۹	۰٫۶۷۲	۳٫۸۷E − ۴	۲٫۶۹E − ۴	۸٫۷۷E − ۵
۸٫۶۹۵۹	۱٫۰۰۰	۳٫۹۶E − ۴	۳٫۰۵E − ۴	۱٫۷۳E − ۴
۹٫۶۹۵۹	۱٫۰۰۰	۴٫۲۷E − ۴	۲٫۹۴E − ۴	۱٫۱۸E − ۴
۱۰٫۶۹۵۹	۱٫۰۰۰	۴٫۱۱E − ۴	۲٫۷۷E − ۴	۹٫۴۵E − ۵

به‌طوری که شرط

$$\underset{0\le n\le p}{\text{Max}} \mid y_n - z_n \mid \le \varepsilon \qquad 0 < h \le h_0 \qquad (۲.۸.۶)$$

برقرار باشد. توجه کنید که مقادیر آغازی احتمالاً به h بستگی دارند. گوییم یک خانواده جواب $\{y_n \mid 0 \le n \le N(h)\}$ پایدار است اگر مقدار ثابتی چون c مستقل از $h < h_0$ وجود داشته باشد به‌طوری که به‌ازای جمیع مقادیر به اندازه کافی کوچک ε داشته باشیم

$$\underset{0\le n\le N(h)}{\text{Max}} \mid y_n - z_n \mid \le c\varepsilon \qquad 0 < h \le h_0 \qquad (۳.۸.۶)$$

همهٔ مسائل معادلهٔ دیفرانسیل

$$y' = f(x,y) \qquad y(x_0) = Y_0 \qquad (۴.۸.۶)$$

را در نظر می‌گیریم که در آنها $f(x,y)$ مشتق پیوسته باشد و در شرط (۱۲.۲.۶) لیپشیتس صدق کند، و فرض می‌کنیم که جوابهای تقریبی $\{y_n\}$ همگی پایدارند. در این صورت می‌گوییم (۱.۸.۶) یک **روش عددی پایدار** است.

به منظور تعریف همگرایی برای یک مسألهٔ داده‌شدهٔ (۴.۸.۶)، فرض می‌کنیم که مقادیر اولیهٔ y_0، ...، y_p در رابطهٔ زیر صدق می‌کنند

$$\eta(h) \equiv \max_{0 \le n \le p} | Y(x_n) - y_n | \to 0 \qquad h \to 0 \text{ وقتی} \qquad (۵.۸.۶)$$

در این صورت گفته می‌شود جواب $\{y_n\}$ به $Y(x)$ همگراست اگر

$$\max_{x_0 \le x_n \le b} | Y(x_n) - y_n | \to 0 \qquad h \to 0 \text{ وقتی} \qquad (۶.۸.۶)$$

اگر (۱.۸.۶) برای تمام مسائل (۴.۸.۶) همگرا باشد، آنگاه آن را یک **روش عددی همگرا** نامند.

تعریف سازگاری را که در بخش ۳.۶ داده‌شده به یاد می‌آوریم. روش (۱.۸.۶) **سازگار** است اگر برای تمام توابع پیوسته مشتقپذیر $Y(x)$ در $[x_0, b]$،

$$\frac{1}{h} \max_{x_p \le x_n \le b} | T_n(Y) | \to 0 \qquad h \to 0 \text{ وقتی}$$

یا هم‌ارز با آن از قضیهٔ ۵.۶، ضرایب $\{a_j\}$ و $\{b_j\}$ باید در روابط زیر صدق کنند

$$\sum_{j=0}^{p} a_j = 1 \qquad -\sum_{j=0}^{p} j a_j + \sum_{j=-1}^{p} b_j = 1 \qquad (۷.۸.۶)$$

می‌توان نشان داد که همگرایی، سازگاری را ایجاب می‌کند؛ در نتیجه، ما فقط روشهایی را در نظر می‌گیریم که در (۷.۸.۶) صدق می‌کنند. به‌عنوان یک مثال از برهان لزوم (۷.۸.۶)، فرض همگرایی (۱.۸.۶) برای مسألهٔ

$$y' \equiv 0 \qquad y(0) = 1$$

اولین شرط (۷.۸.۶) را ایجاب خواهد کرد. کافی است بگیریم $y_p = 1 = \ldots = y_0$ و به نتایج همگرایی y_{p+1} به $Y(x) \equiv 1$ توجه کنیم.

پایداری نسبی $\Longrightarrow$ شرط قوی ریشه

$\Downarrow$ $\Downarrow$

پایداری $\Longleftrightarrow$ شرط ریشه $\Longleftrightarrow$ همگرایی

شکل ۶.۶ شرح اجمالی نظریهٔ روشهای چندگامی سازگار

همگرایی و پایداری (۱.۸.۶) به ریشه‌های چندجمله‌یی زیر وابسته‌اند

$$\rho(r) = r^{p+1} - \sum_{j=0}^{p} a_j r^{p-j} \qquad (۸.۸.۶)$$

ملاحظه می‌کنید که با توجّه به شرط سازگاری (۷.۸.۶) داریم $\rho(1) = 0$. گیریم r_0، ... ، r_p معرّف ریشه‌های $\rho(r)$ باشند، که هر ریشه مطابق چندگانی خود تکرار شده است، و گیریم $r_0 = 1$.
روش (۱.۸.۶) در **شرط ریشه** صدق می‌نماید اگر

.۱

$$| r_j | \leq 1 \qquad j = 0, 1, \cdots, p \qquad (۹.۸.۶)$$

.۲

$$| r_j | = 1 \Rightarrow \rho'(r_j) \neq 0 \qquad (۱۰.۸.۶)$$

شرط اول ایجاب می‌کند که تمام ریشه‌های $\rho(r)$ در دایرهٔ واحد $\{z : | z | \leq 1\}$ واقع در صفحهٔ مختلط باشند. شرط (۱۰.۸.۶) بیان می‌کند که تمام ریشه‌های واقع بر مرز دایره باید ریشه‌های سادهٔ $\rho(r)$ باشند.

نتایج عمدهٔ این بخش در شکل ۶.۶ نشان داده شده‌اند، گرچه بعضی از آنها اثبات نشده‌اند. **شرط قوی ریشه** و مفهوم **پایداری نسبی** بعداً در این بخش معرفی خواهد شد.

نظریهٔ پایداری تمام روشهای عددی که در بخشهای قبل داده شده‌اند پایدار بودند. اکنون یک مثال از روش ناپایدار می‌آوریم. این امر به‌منظور ایجاد انگیزه برای مطالعهٔ نظریهٔ کلی پایداری صورت گرفته است.

مثال فرمول کلّی (۱۰.۷.۶) را برای یک روش صریح مرتبهٔ دوّم دوگامی به‌یاد آورید و a_0 را ۳

انتخاب کنید، در این صورت روش

$$y_{n+1} = 3y_n - 2y_{n-1} + \frac{h}{2}[f(x_n, y_n) - 3f(x_{n-1}, y_{n-1})] \qquad n \geq 1 \quad (11.8.6)$$

را با خطای برشی

$$T_n(Y) = \frac{7}{12}h^3 Y^{(3)}(\xi_n) \qquad x_{n-1} \leq \xi_n \leq x_{n+1}$$

را به‌دست می‌آوریم. حل مسألهٔ $y(0) = 0$، $y' \equiv 0$، را در نظر می‌گیریم که $Y(x) \equiv 0$ جواب آن است. اگر از $y_1 = 0$ استفاده کنیم، واضح می‌شود که $y_n = 0$، $n \geq 0$، جواب عددی $y_0 = 0$ آن است. اختلالی در داده‌های آغازی به اندازهٔ $z_0 = \varepsilon/2$، $z_1 = \varepsilon$ به‌ازای یک $\varepsilon \neq 0$ ایجاد می‌نماییم. در این صورت می‌توان نشان داد که جواب عددی متناظر، عبارت است از

$$z_n = \varepsilon \times 2^{n-1} \qquad n \geq 0 \qquad (12.8.6)$$

استدلالی که منجر به پیداشدن این جواب شده است بعداً در شرایط کلیتری داده خواهد شد. برای ملاحظهٔ اثر این اختلال در جواب اصلی،

$$\max_{x_0 \leq x_n \leq b} | y_n - z_n | = \max_{0 \leq x_n \leq b} | \varepsilon | 2^{n-1} = | \varepsilon | 2^{N(h)-1}$$

از آنجا که وقتی $h \to 0$، $N(h) \to \infty$، انحراف $\{z_n\}$ از $\{y_n\}$ وقتی $h \to 0$، به‌طور افزایشی زیاد می‌شود. روش (۱۱.۸.۶) ناپایدار است، و هرگز نباید مورد استفاده قرار گیرد. همچنین باید توجه کرد که شرط ریشه نیز نقض شده است، زیرا $\rho(r) = r^2 - 3r + 2$ دارای ریشه‌های $r_0 = 1$ و $r_1 = 2$ است.

برای بررسی پایداری (۱.۸.۶)، فقط معادلهٔ خاص

$$y' = \lambda y \qquad y(0) = 1 \qquad (13.8.6)$$

را در نظر می‌گیریم، که $Y(x) = e^{\lambda x}$ جواب آن است. نتایج به‌دست آمده به مطالعهٔ پایداری مسألهٔ معادله دیفرانسیل کلی انتقال می‌یابد. یک دلیل شهودی، به‌سادگی به‌دست می‌آید. $Y'(x) = f(x, y(x))$ را حول (x_0, y_0) بسط می‌دهیم تا به‌دست آوریم

$$\begin{aligned} Y'(x) &\doteq f(x_0, Y_0) + f_x(x_0, Y_0)(x - x_0) + f_y(x_0, Y_0)(Y(x) - Y_0) \\ &= \lambda(Y(x) - Y_0) + g(x) \end{aligned} \qquad (14.8.6)$$

که در آن $\lambda = f_y(x_\circ, Y_\circ)$ و $g(x) = f(x_\circ, Y_\circ) + f_x(x_\circ, Y_\circ)(x - x_\circ)$. اگر $(x - x_\circ)$ به‌اندازهٔ کافی کوچک باشد، این تقریب یک تقریب معتبر است. عبارت $V(x) = Y(x) - Y_\circ$ را وارد می‌کنیم

$$V'(x) \doteq \lambda V(x) + g(x) \qquad (۱۵.۸.۶)$$

جملهٔ ناهمگن $g(x)$ در به‌دست آوردن پایداری عددی همه‌جا حذف می‌شود، زیرا ما با تفاضلهای جوابهای معادله سروکار داریم. حذف $g(x)$ در (۱۵.۸.۶)، معادلهٔ نمونهٔ (۱۳.۸.۶) را به‌دست می‌دهد. برای توضیح بیشتر، به نتایج پایداری (۵.۱.۶) تا (۱۰.۱.۶) و به نتایج پایداری ذوزنقه‌یی در (۲۰.۵.۶) تا (۲۳.۵.۶) رجوع کنید.

در حالتی که $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$ یک دستگاه m معادلهٔ دیفرانسیل مانند (۱۳.۱.۶) را نشان می‌دهد، مشتق جزیی $\mathbf{f}_\mathbf{y}(x, \mathbf{y})$، مانند (۵۴.۲.۶)، یک ماتریس ژاکوبی می‌شود:

$$[\mathbf{f}_\mathbf{y}(x, \mathbf{y})]_{ij} = \frac{\partial f_i}{\partial y_j} \qquad ۱ \le i, j \le m$$

پس مسألهٔ نمونه، به یک دستگاه m معادلهٔ دیفرانسیل خطی

$$\mathbf{y}' = \Lambda \mathbf{y} + \mathbf{g}(x) \qquad (۱۶.۸.۶)$$

بدل می‌شود که $\Lambda = \mathbf{f}_y(x_\circ, \mathbf{Y}_\circ)$. می‌توان نشان داد که در بیشتر حالات، این دستگاه به دستگاه هم‌ارز زیر بدل می‌شود

$$z_i' = \lambda_i z_i + \gamma_i(x) \qquad ۱ \le i \le m \qquad (۱۷.۸.۶)$$

که $\lambda_۱، \ldots، \lambda_m$ ویژه مقدارهای ماتریس Λ هستند (مسألهٔ ۲۴ را ببینید). با (۱۷.۸.۶) ما به معادلهٔ نمونهٔ سادهٔ (۱۳.۸.۶) بازگشته‌ایم، به شرطی که λ عدد مختلط باشد تا بتواند تمام ویژه مقدارهای ممکنِ Λ را شامل شود.

با به‌کار بردن (۱.۸.۶) برای معادلهٔ نمونهٔ (۱۳.۸.۶)، به‌دست می‌آوریم

$$y_{n+۱} = \sum_{j=\circ}^{p} a_j y_{n-j} + h\lambda \sum_{j=-۱}^{p} b_j y_{n-j} \qquad (۱۸.۸.۶)$$

$$(۱ - h\lambda b_{-۱}) y_{n+۱} - \sum_{j=\circ}^{p} (a_j + h\lambda b_j) y_{n-j} = \circ \qquad n \ge p \quad (۱۹.۸.۶)$$

این معادله یک **معادلهٔ تفاضلی خطی همگن** از مرتبهٔ $p+1$ است و نظریهٔ حلپذیری آن کاملاً مشابه نظریهٔ معادلات دیفرانسیل خطی همگن مرتبهٔ $p+1$ است. به‌عنوان یک مرجع کلی، هنریچی (۱۹۶۲، صص ۲۱۰ـ۲۱۵)، اَیزکسون و کلر (۱۹۶۶، صص ۴۰۵ تا ۴۱۷) را ببینید.

برای پیدا کردن جواب عمومی، سعی می‌کنیم ابتدا به دنبال جوابهای خاصّی به شکل

$$y_n = r^n \qquad n \geq 0$$

بگردیم. اگر بتوانیم $p+1$ جواب مستقل خطی پیدا کنیم، آنگاه هر ترکیب خطی یک جواب عمومی (۱۹.۸.۶) خواهد بود.

با گذاردن $y_n = r^n$ در (۱۹.۸.۶) و حذف r^{n-p}، به‌دست می‌آوریم

$$(1 - h\lambda b_{-1})r^{p+1} - \sum_{j=0}^{p}(a_j + h\lambda b_j)r^{p-j} = 0 \qquad (۲۰.۸.۶)$$

این معادله، **معادلهٔ مشخصه** نامیده می‌شود، و سمت چپ آن **چندجمله‌یی مشخصه** است. ریشه‌ها را **ریشه‌های مشخصه** نامند. تعریف می‌کنیم

$$\sigma(r) = b_{-1}r^{p+1} + \sum_{j=0}^{p} b_j r^{p-j}$$

و تعریف (۸.۸.۶) را برای $\rho(r)$ را به یاد می‌آوریم. معادلهٔ (۲۰.۸.۶) به معادلهٔ زیر تبدیل می‌شود

$$\rho(r) - h\lambda\sigma(r) = 0 \qquad (۲۱.۸.۶)$$

ریشه‌های مشخصه را با

$$r_0(h\lambda), r_1(h\lambda), \ldots, r_p(h\lambda)$$

نشان می‌دهیم که می‌توان نشان داد به‌طور پیوسته به مقدار $h\lambda$ بستگی دارند. وقتی $h\lambda = 0$، معادلهٔ (۲۱.۸.۶) به‌صورت $\rho(r) = 0$ در می‌آید، برای r_j، ریشه‌های قبلی $\rho(r) = 0$، داریم $r_j(0) = r_j$، $j = 0, 1, \ldots, p$. چون $r_0 = 1$ یک ریشهٔ $\rho(r)$ است، گیریم که $r_0(h\lambda)$ ریشهٔ (۲۱.۸.۶) باشد که برای آن $r_0(0) = 1$. به دلایلی که بعداً روشن خواهد شد، ریشهٔ $r_0(h\lambda)$، **ریشهٔ اصلی** خوانده می‌شود. اگر ریشه‌های $r_j(h\lambda)$ همگی متمایز باشند، جواب عمومی (۱۹.۸.۶) چنین است

$$y_n = \sum_{j=0}^{p} \gamma_j [r_j(h\lambda)]^n \qquad n \geq 0 \qquad (۲۲.۸.۶)$$

ولی اگر $r_j(h\lambda)$ یک ریشه با چندگانی $\nu > 1$ باشد، آنگاه ν جواب مستقل خطی (۱۹.۸.۶) به‌صورت زیرند؛

$$\{[r_j(h\lambda)]^n\}, \{n[r_j(h\lambda)]^n\}, \ldots, \left\{n^{\nu-1}[r_j(h\lambda)]^n\right\}$$

با این جوابها همراه با جوابهایی که از سایر ریشه‌ها به‌دست می‌آیند، می‌توان جواب عمومی (۱۹.۸.۶)، قابل‌مقایسه با (۲۲.۸.۶) را تولید کرد.

قضیهٔ ۷.۶ شرط سازگاری (۷.۸.۶) مفروض است. روش چندگامی (۱.۸.۶) پایدار است اگر و تنها اگر شرط ریشهٔ (۹.۸.۶) و (۱۰.۸.۶) برقرار باشد.

برهان ۱. اثبات را با نشان دادن لزوم شرط ریشه برای پایداری آغاز می‌کنیم. بدین منظور، خلاف آن را با فرض اینکه برای مقداری از j

$$|r_j(0)| > 1$$

در نظر می‌گیریم. معادلهٔ دیفرانسیل $y' \equiv 0$، $y(0) = 0$، با جواب $Y(x) \equiv 0$ را در نظر می‌گیریم. پس (۱.۸.۶) چنین می‌شود

$$y_{n+1} = \sum_{j=0}^{p} a_j y_{n-j} \qquad n \geq p \qquad (۲۳.۸.۶)$$

اگر $y_0 = y_1 = \ldots = y_p = 0$، روشن است که جواب عددی برای جمیع مقادیر $n \geq 0$، $y_n = 0$ خواهد شد. برای مقادیر اختلال یافتهٔ اولیه مقادیر زیر را می‌گیریم

$$z_0 = \varepsilon, z_1 = \varepsilon r_j(0), \ldots, z_p = \varepsilon r_j(0)^p \qquad (۲۴.۸.۶)$$

برای این مقادیر اولیه

$$\underset{0 \leq n \leq p}{\text{Max}} \mid y_n - z_n \mid = \varepsilon \mid r_j(0) \mid^p$$

که یک کران یکنواخت برای همهٔ مقادیر کوچک h است، زیرا سمت راست مستقل از h است. این کران نیز وقتی $\varepsilon \to 0$، به صفر میل می‌کند.

جواب (۲۴.۸.۶) با شرایط اولیهٔ (۲۴.۸.۶) به‌صورت زیر است

$$z_n = \varepsilon[r_j(0)]^n \qquad n \geq 0$$

برای انحراف از $\{y_n\}$ داریم

$$\underset{x_0 \le x_n \le b}{\text{Max}} \mid y_n - z_n \mid = \varepsilon \mid r_j(0) \mid^{N(h)}$$

وقتی $h \to 0$، $N(h) \to \infty$ وکران نامتناهی می‌شود. و این ثابت می‌کند که وقتی $|r_j(0)| > 1$، این روش ناپایدار است. اگر شرط ریشه با فرض نادرست‌بودن (۱۰.۸.۶) نقض شود، برهان مشابهی می‌توان ارائه نمود. این برهان به مسألهٔ ۲۹ واگذار شده است.

۲. فرض می‌کنیم که شرط ریشه برقرار باشد. برهان پایداری را به معادله‌ای به صورت (۱۳.۸.۶) محدود می‌کنیم. یک برهان برای معادلهٔ کلی $y' = f(x, y)$ می‌توان داد که با کمی تغییر برهان زیر است. برهان کلی شامل حل معادلات تفاضلی خطی ناهمگن است [برای یک بحث کلی اَیزکسون و کلر (۱۹۶۶، صص ۴۰۵ـ۴۱۷) را ببینید.] برای آنکه برهان بازهم ساده‌تر شود، فرض می‌کنیم که ریشه‌های $r_j(0)$، $j = 0, 1, \ldots, p$، همگی متمایز باشند. این امر عیناً برای $r_j(h\lambda)$ درست است، مشروط بر آنکه مقدار h به اندازه کافی کوچک، مثلاً $0 \le h \le h_0$، نگهداشته شود.

گیریم $\{y_n\}$ و $\{z_n\}$ دو جواب (۱۹.۸.۶) در $[x_0, b]$ باشند، و فرض می‌کنیم،

$$\underset{0 \le n \le p}{\text{Max}} \mid y_n - z_n \mid \le \varepsilon \qquad 0 < h \le h_0 \qquad (۲۵.۸.۶)$$

خطای $e_n = y_n - z_n$ را وارد می‌کنیم. با کم‌کردن و به‌کاربردن (۱۹.۸.۶) برای هر جواب،

$$(1 - h\lambda b_{-1})e_{n+1} - \sum_{j=0}^{p}(a_j + h\lambda b_j)e_{n-j} = 0 \qquad (۲۶.۸.۶)$$

برای $x_{p+1} \le x_{n+1} \le b$. جواب عمومی چنین است

$$e_n = \sum_{j=0}^{p} \gamma_j [r_j(h\lambda)]^n \qquad n \ge 0 \qquad (۲۷.۸.۶)$$

ضرایب $\gamma_0(h), \ldots, \gamma_p(h)$ باید طوری انتخاب شوند که

$$\begin{aligned} \gamma_0 + \gamma_1 + \ldots + \gamma_p &= e_0 \\ \gamma_0 r_0(h\lambda) + \ldots + \gamma_p r_p(h\lambda) &= e_1 \\ &\vdots \\ \gamma_0 [r_0(h\lambda)]^p \ldots + \gamma_p [r_p(h\lambda)]^p &= e_p \end{aligned}$$

در این صورت جواب (۲۷.۸.۶) با اختلالات اولیهٔ داده‌شدهٔ e_0 ، ...، e_p مطابقت دارد و در معادلهٔ تفاضلی (۲۶.۸.۶) صدق می‌نماید. با استفاده ازکران (۲۵.۸.۶) و نظریهٔ دستگاه معادلات

خطی، به‌سادگی می‌توان نشان داد که برای مقدار ثابتی چون $c_1 > 0$،

$$\max_{0 \le i \le p} |\gamma_i| \le c_1 \varepsilon \qquad 0 < h \le h_0 \tag{۲۸.۸.۶}$$

ما برهان را نمی‌آوریم گرچه با مفاهیمی که در فصل ۷ و ۸ معرفی شده‌اند می‌توان به‌سادگی آن را انجام داد. برای کراندارنمودن e_n در $[x_0, b]$، باید هر جملهٔ $[r_j(h\lambda)]^n$ را کراندار کنیم. برای این کار بسط

$$r_j(u) = r_j(0) + u r_j'(\zeta) \tag{۲۹.۸.۶}$$

را به‌ازای مقداری از ζ بین 0 و u در نظر می‌گیریم. برای محاسبهٔ $r_j'(u)$، از معادلهٔ مشخصهٔ زیر مشتق می‌گیریم

$$\rho(r_j(u)) - u\sigma(r_j(u)) = 0$$

پس

$$r_j'(u) = \frac{\sigma(r_j(u))}{\rho'(r_j(u)) - u\sigma'(r_j(u))} \tag{۳۰.۸.۶}$$

با فرض اینکه $r_j(0)$ یک ریشهٔ سادهٔ $\rho(r) = 0$، $0 \le j \le p$، نتیجه می‌شود که $\rho'(r_j(0)) \ne 0$، و با توجه به پیوستگی، به‌ازای جمیع مقادیر به اندازهٔ کافی کوچک u، $\rho'(r_j(u)) \ne 0$. مخرج (۳۰.۸.۶) ناصفر است و می‌توانیم $r_j'(u)$ را کراندار کنیم

$$|r_j'(u)| \le c_2 \qquad \text{جمیع مقادیر } |u| \le u_0$$

برای مقداری از $u_0 > 0$.

با استفاده از این نتیجه از (۲۹.۸.۶) و از شرط ریشهٔ (۹.۸.۶)، به‌ازای جمیع مقادیر $0 < h \le h_0$ داریم

$$\begin{aligned} |r_j(h\lambda)| &\le |r_j(0)| + c_2 |h\lambda| \le 1 + c_2 |h\lambda| \\ |[r_j(h\lambda)]^n| &\le [1 + c_2 |h\lambda|]^n \le e^{c_2 n |h\lambda|} \le e^{c_2 (b - x_0)|\lambda|} \end{aligned} \tag{۳۱.۸.۶}$$

از ترکیب با (۲۷.۸.۶) و (۲۸.۸.۶) به‌ازای یک مقدار مناسب ثابت c_2، داریم

$$\max_{x_0 \le x_n \le b} |e_n| \le c_3 |\varepsilon| e^{c_2 (b - x_0)|\lambda|} \qquad 0 < h \le h_0$$

برای مقدار مناسب ثابت c_3. ■

نظریهٔ همگرایی قضیهٔ زیر تعمیم قضیهٔ ۶.۶ از بخش ۳.۶، با شرایط لازم و کافی است که برای همگرایی روشهای چند گامی داده شده است.

قضیهٔ ۸.۶ شرط سازگاری (۷.۸.۶) مفروض است. در این صورت روش چندگامی (۱.۸.۶) همگراست اگر و تنها اگر شرط ریشهٔ (۹.۸.۶)، (۱۰.۸.۶) برقرار باشد.

برهان ۱. با نشان‌دادن لزوم شرط ریشه برای همگرایی آغاز و باز هم از مسألهٔ $y' \equiv 0$ و $y(0) = 0$ با جواب $Y(x) \equiv 0$ استفاده می‌کنیم. روش چندگامی (۱.۸.۶) چنین می‌شود

$$y_{n+1} = \sum_{j=0}^{p} a_j y_{n-j} \qquad n \geq p \qquad (۳۲.۸.۶)$$

با انتخاب y_0، ...، y_p به نحوی که در شرط زیر صدق کنند

$$\eta(h) \equiv \operatorname*{Max}_{0 \leq n \leq p} \mid y_n \mid \to 0 \qquad \text{وقتی } h \to 0 \qquad (۳۳.۸.۶)$$

فرض می‌کنیم که شرط ریشه نقض شده است. نشان می‌دهیم که (۳۲.۸.۶) به $Y(x) \equiv 0$ همگرا نیست.

فرض می‌کنیم که یک $\mid r_j(0) \mid > 1$. یک جواب قابل‌قبول (۳۲.۸.۶) چنین است

$$y_n = h[r_j(0)]^n \qquad x_0 \leq x_n \leq b \qquad (۳۴.۸.۶)$$

شرط (۳۳.۸.۶) برقرار است زیرا

$$\eta(h) = h \mid r_j(0) \mid^p \to 0 \qquad \text{وقتی } h \to 0$$

ولی جواب $\{y_n\}$ همگرا نیست. نخست،

$$\operatorname*{Max}_{x_0 \leq x_n \leq b} \mid Y(x_n) - y_n \mid = h \mid r_j(0) \mid^{N(h)}$$

آن مقادیر از h را که $h = b/N(h)$ در نظر می‌گیریم. می‌توان از قاعدهٔ هوپیتال استفاده نموده نشان داد که

$$\operatorname*{Lim}_{N \to \infty} \frac{b}{N} \mid r_j(0) \mid^N = \infty$$

که نشان می‌دهد (۳۲.۸.۶) به جواب $Y(x) \equiv 0$ همگرا نیست.

فرض می‌کنیم از شرط ریشه، (۹.۸.۶) برقرار باشد، ولی یک $r_j(\circ)$ ریشهٔ چندگانهٔ $\rho(r)$ باشد و $|r_j(\circ)| = ۱$. آنگاه اثباتی مشابه آنچه گفته شد کماکان برقرار است ولی باید از جواب زیر استفاده کنیم

$$y_n = hn[r_j(\circ)]^n \qquad \circ \le n \le N(h)$$

و این برهان لزوم شرط ریشه را کامل می‌کند.

۲. فرض کنید که شرط ریشه برقرار است. مانند قضیهٔ قبل، اثبات همگرایی برای یک معادلهٔ دیفرانسیل دلخواه خیلی مشکل است. می‌توانید به آیزکسون و کلر (۱۹۶۶، صص ۴۰۵ـ۴۱۷) مراجعه کنید. برهان ارائه شده به معادله‌ای به‌صورت (۱۳.۸.۶) محدود شده است، و باز هم برای آسان‌بودن اثبات فرض می‌کنیم که ریشه‌های $r_j(\circ)$ متمایزند.

روش چندگامی (۱.۸.۶) برای معادلهٔ نمونهٔ $y' = \lambda y$ و $y(\circ) = ۱$ به (۱۸.۸.۶) تبدیل می‌شود. نشان می‌دهیم که جملهٔ $\gamma_\circ[r_\circ(h\lambda)]^n$ در جواب عمومی

$$y_n = \sum_{j=\circ}^{p} \gamma_j[r_j(h\lambda)]^n \qquad (۳۵.۸.۶)$$

به جواب $Y(x) = e^{\lambda x}$ در $[\circ, b]$، همگراست. جملات دیگر $\gamma_j[r_j(h\lambda)]^n$، $j = ۱, ۲, \ldots, p$، **جوابهای مزاحم** هستند و می‌توان نشان داد که وقتی $h \to \circ$ این جوابها به صفر می‌گرایند (مسألهٔ ۳۰ را ببینید).

با استفاده از قضیهٔ تیلر، $r_\circ(h\lambda)$ را بسط می‌دهیم

$$r_\circ(h\lambda) = r_\circ(\circ) + h\lambda r'(\circ) + O(h^۲)$$

از (۳۰.۸.۶) داریم

$$r'_\circ(\circ) = \frac{\sigma(۱)}{\rho'(۱)}$$

که با استفاده از شرط سازگاری (۷.۸.۶)، این تساوی به $r'_\circ(\circ) = ۱$ منجر می‌شود. پس روی هر بازهٔ متناهی $\circ \le x_n \le b$،

$$r_\circ(h\lambda) = ۱ + h\lambda + O(h^۲) = e^{\lambda h} + O(h^۲)$$
$$[r_\circ(h\lambda)]^n = e^{\lambda n h}[۱ + O(h^۲)]^n = e^{\lambda x_n}[۱ + O(h)] \qquad (۳۶.۸.۶)$$

بنابراین

$$\underset{0\le x_n\le b}{\text{Max}} \mid [r_0(h\lambda)]^n - e^{\lambda x_n} \mid \to 0 \qquad h \to 0 \text{ وقتی که} \qquad (۳۷.۸.۶)$$

اکنون باید نشان دهیم که وقتی $h \to 0$، $\gamma_0 \to 1$.

ضرایب $\gamma_0(h)$، ...، $\gamma_p(h)$ در دستگاه خطی زیر صدق می‌نمایند

$$\begin{aligned} \gamma_0 + \gamma_1 + \ldots + \gamma_p &= y_0 \\ \gamma_0[r_0(h\lambda)] + \ldots + \gamma_p[r_p(h\lambda)] &= y_1 \\ &\vdots \\ \gamma_0[r_0(h\lambda)]^p + \ldots + r_p[r_p(h\lambda)]^p &= y_p \end{aligned} \qquad (۳۸.۸.۶)$$

مقادیر آغازی $y_0, \ldots, y_p$ بستگی به h دارند، و بنا به فرض باید در رابطهٔ زیر صدق کنند

$$\eta(h) \equiv \underset{0\le n\le p}{\text{Max}} \mid e^{\lambda x_n} - y_n \mid \to 0 \qquad h \to 0 \text{ وقتی}$$

ولی این امر ایجاب می‌کند که

$$\underset{h\to 0}{\text{Lim}} y_n = 1 \qquad 0 \le n \le p \qquad (۳۹.۸.۶)$$

ضریب γ_0 را می‌توان با استفاده از قاعدهٔ کرامر در حل (۳۸.۸.۶) به‌دست آورد:

$$\gamma_0 = \frac{\begin{vmatrix} y_0 & 1 & \ldots & 1 \\ y_1 & r_1 & \ldots & r_p \\ \vdots & & & \\ y_p & r_1^p & \ldots & r_p^p \end{vmatrix}}{\begin{vmatrix} 1 & 1 & \ldots & 1 \\ r_0 & r_1 & \ldots & r_p \\ \vdots & & & \\ r_0^p & r_1^p & \ldots & r_p^p \end{vmatrix}} \qquad (۴۰.۸.۶)$$

مخرج کسر برای $r_p(0), \ldots, r_1(0), r_0(0) = 1$ به دترمینان واندرموند میل می‌کند که ناصفر است زیرا ریشه‌ها متمایزند (مسألهٔ ۱، فصل ۳ را ببینید). با استفاده از (۳۹.۸.۶)، صورت کسر نیز وقتی $h \to 0$ به همان مقدار میل خواهد کرد. بنابراین وقتی $h \to 0$ آنگاه $\gamma_0 \to 1$. با استفاده

از این ویژگی همراه با (۳۷.۸.۶) و مسألهٔ ۳۰، جواب $\{y_n\}$ به سمت $Y(x) = e^{\lambda x}$ در $[\circ, b]$ میل خواهد کرد. ■

آنچه ذیلاً می‌آید نتیجه مشهودی است و نتیجهٔ نمایان قضایای ۷.۶ و ۸.۶ است.

فرع گیریم (۱.۸.۶) یک روش چندگامی سازگار باشد. در این صورت این روش همگراست اگر و تنها اگر پایدار باشد. ■

پایداری نسبی و پایداری ضعیف بازهم معادلهٔ نمونهٔ (۱.۸.۶) و جواب عددی آن (۳۲.۸.۶) را در نظر می‌گیریم. قضیهٔ قبل می‌گوید که جوابهای انگلی $\gamma_j[r_j(h\lambda)]^n$ وقتی $h \to \circ$، به صفر می‌گرایند. ولی برای یک h ثابت و x_n صعودی، می‌خواهیم این جوابها نسبت به قسمت اصلی جواب $\gamma_\circ[r_\circ(h\lambda)]^n$ کوچک بمانند. این امر وقتی میسر است که ریشه‌های مشخصه، به‌ازای جمیع مقادیر به اندازهٔ کافی کوچک h در رابطهٔ زیر صدق کنند

$$| r_j(h\lambda) | \leq r_\circ(h\lambda) \qquad j = ۱, ۲, \ldots, p \qquad (۴۱.۸.۶)$$

این نابرابری ما را به تعریف پایداری نسبی زیر هدایت می‌کند.

گوییم روش (۱.۸.۶) **نسبی ـ پایدار** است اگر ریشه‌های مشخصهٔ $r_j(h\lambda)$، به‌ازای مقادیر به اندازهٔ کافی کوچک $| h\lambda |$ در رابطهٔ (۴۱.۸.۶) صدق کنند. و می‌گویند روش در **شرط قوی ریشه** صدق می‌کند اگر

$$| r_j(\circ) | < ۱ \qquad j = ۱, ۲, \ldots, p \qquad (۴۲.۸.۶)$$

این شرط، به‌سادگی کنترل می‌شود، و پایداری نسبی را ایجاب می‌کند. فقط کافی است از پیوستگی ریشه‌های $r_j(h\lambda)$ نسبت به $h\lambda$ استفاده کنیم تا ببینیم که شرط (۴۲.۸.۶) شرط (۴۱.۸.۶) را ایجاب می‌کند. پایداری نسبی شرط قوی ریشه را تضمین نمی‌کند، اگرچه برای بیشتر روشها، آنها هم‌ارزند [مسألهٔ ۳۶ (ب) را ببینید]. اگر یک روش چند گامی پایدار باشد ولی نسبی ـ پایدار نباشد آن را **ضعیف ـ پایدار** خوانند.

مثال ۱. برای روش میانگاهی

$$r_\circ(h\lambda) = ۱ + h\lambda + O(h^۲) \qquad r_۱(h\lambda) = -۱ + h\lambda + O(h^۲) \qquad (۴۳.۸.۶)$$

بنابر (۴۱.۸.۶) وقتی $\lambda < ۰$، ضعیف ـ پایدار است که با آنچه قبلاً در بخش ۴.۶ بیان شد مطابقت می‌کند.

۲. روشهای آدامز ـ بشفورت و آدامز ـ مولتن، (۲۲.۷.۶) و (۲۶.۷.۶)، وقتی $h = ۰$، دارای یک چندجمله‌یی مشخصه هستند،

$$\rho(r) = r^{p+۱} - r^p \qquad (۴۴.۸.۶)$$

ریشه‌ها عبارتند از $r_۰ = ۱$ و $r_j = ۰$، $j = ۱, ۲, \ldots, p$ بنابراین شرط قوی ریشه برقرار است و روشهای آدامز نسبی ـ پایدارند.

ناحیه‌های پایداری در بحثهای قبلی پایداری، لازم بود مقادیر h به اندازهٔ کافی کوچک باشند تا بتوان عملیات را ادامه داد. اصلاً اشاره نشده بود که h چه اندازه باید کوچک باشد. روشن است که اگر لازم شود h بسیار کوچک باشد، آنگاه این روش برای بسیاری از مسائل غیرعملی خواهد بود. بنابراین لازم است مقادیر مجاز h را بررسی کنیم. چون پایداری به ریشه‌های مشخصه بستگی دارد، و چون خود این ریشه‌ها نیز به $h\lambda$ وابسته‌اند، ما علاقه‌مندیم مقادیری از $h\lambda$ را پیدا کنیم که برای آنها روش چندگامی (۱.۸.۶) به تعبیری پایدار باشد. برای اینکه بتوانیم در همهٔ وضعیتهایی که در حل دستگاه معادلات دیفرانسیل پیش می‌آیند، بحث کنیم لازم است، همان‌گونه که به دنبال (۱۷.۸.۶) اشاره شد، λ بتواند مقادیر مختلط اختیار کند.

برای دیدن انگیزهٔ بحث اخیر، پایداری روش اویلر را در نظر می‌گیریم. روش اویلر را برای معادلهٔ

$$y' = \lambda y + g(x) \qquad y(۰) = Y_۰ \qquad (۴۵.۸.۶)$$

به‌کار می‌بریم، نتیجهٔ زیر به‌دست می‌آید

$$y_{n+۱} = y_n + h[\lambda y_n + g(x_n)] \qquad n \geq ۰ \qquad y_۰ = Y_۰ \quad (۴۶.۸.۶)$$

سپس مسألهٔ اختلال یافته را در نظر می‌گیریم

$$z_{n+۱} = z_n + h[\lambda z_n + g(x_n)] \qquad n \geq ۰ \qquad z_۰ = Y_۰ + \varepsilon \quad (۴۷.۸.۶)$$

برای معادلهٔ اصلی (۴۵.۸.۶)، این اختلال $Y_۰$ به جوابهای $Y(x)$ و $Z(x)$ می‌انجامد که در رابطهٔ زیر صدق می‌کنند

$$Y(x) - Z(x) = \varepsilon e^{\lambda x} \qquad x \geq ۰$$

در این مسألهٔ اصلی، معمولاً به حالتی توجه داریم که $\text{Real}(\lambda) \leq ۰$، زیرا در این صورت $| Y(x) - Z(x) |$ وقتی $x \to ۰$، کراندار باقی می‌ماند. توجه خود را به حالت $\text{Real}(\lambda) < ۰$ معطوف می‌داریم. در این حالت وقتی $x \to \infty$، $Y(x) - Z(x) \to ۰$. برای چنین λای می‌خواهیم مقادیری از h را پیدا کنیم که جوابهای عددی (۴۶.۸.۶) و (۴۷.۸.۶) رفتار وابسته به $Y(x)$ و $Z(x)$ را حفظ کنند.

گیریم $e_n = z_n - y_n$. با کم‌کردن (۴۶.۸.۶) از (۴۷.۸.۶)،

$$e_{n+۱} = e_n + h\lambda e_n = (۱ + h\lambda)e_n \qquad e_۰ = \varepsilon$$

و با استقراء

$$e_n = (۱ + h\lambda)^n \varepsilon \qquad (۴۸.۸.۶)$$

پس وقتی $x_n \to \infty$، $e_n \to ۰$ اگر و تنها اگر

$$| ۱ + h\lambda | < ۱ \qquad (۴۹.۸.۶)$$

این رابطه یک مجموعه از مقادیر مختلط $h\lambda$ به‌دست می‌دهد که از نقاط درون دایرهٔ به شعاع یک به مرکز ۱ − در صفحهٔ مختلط تشکیل شده است. اگر $h\lambda$ به این مجموعه تعلق داشته باشد، آنگاه وقتی $x_n \to \infty$، $y_n - z_n \to ۰$، ولی نه در غیر این صورت.

برای آنکه ببینیم این بحث برای همگرایی نیز اهمیت دارد، توجه می‌کنیم که معادلهٔ دیفرانسیل اصلی را می‌توان به‌عنوان یک حالت اختلال‌یافتهٔ معادلهٔ تقریب‌زنندهٔ (۴۶.۸.۶) منظور کرد. از به‌کاربردن (۱۷.۲.۶) در (۴۵.۸.۶)،

$$Y_{n+۱} = Y_n + h[\lambda Y_n + g(x_n)] + \frac{h^۲}{۲} Y''(\xi_n) \qquad (۵۰.۸.۶)$$

در اینجا در هر مرحله یک اختلال معادلهٔ (۴۶.۸.۶) را داریم، نه فقط یک اختلال در نقطهٔ اولیهٔ $x_۰ = ۰$. با این حال، می‌توان نشان داد که تحلیل پایداری قبلی را برای این اختلال (۴۶.۸.۶) می‌توان به‌کار برد. فرمول خطای (۴۸.۸.۶) باید به نحو مناسبی اصلاح شود، ولی باز هم فرمول به کران (۴۹.۸.۶) زیاد وابسته خواهد بود. (مسألهٔ ۴۰ را ببینید).

مثال روش اویلر را برای مسألهٔ

$$y' = \lambda y + (۱ - \lambda)\cos(x) - (۱ + \lambda)\sin(x) \qquad y(۰) = ۱ \quad (۵۱.۸.۶)$$

جدول ۱۷.۶ روش اویلر برای (۵۱.۸.۶)

λ	x	خطا: $h = ۰٫۵$	خطا: $h = ۰٫۱$	خطا: $h = ۰٫۰۱$
-۱	۱	$-۲٫۴۶\mathrm{E}-۱$	$-۴٫۳۲\mathrm{E}-۲$	$-۴٫۲۲\mathrm{E}-۳$
	۲	$-۲٫۵۵\mathrm{E}-۱$	$-۴٫۶۴\mathrm{E}-۲$	$-۴٫۵۵\mathrm{E}-۳$
	۳	$-۲٫۶۶\mathrm{E}-۲$	$-۶٫۷۸\mathrm{E}-۳$	$-۷٫۲۲\mathrm{E}-۴$
	۴	$۲٫۲۷\mathrm{E}-۱$	$۳٫۹۱\mathrm{E}-۲$	$۳٫۷۸\mathrm{E}-۳$
	۵	$۲٫۷۲\mathrm{E}-۱$	$۴٫۹۱\mathrm{E}-۲$	$۴٫۸۱\mathrm{E}-۳$
-۱۰	۱	$۳٫۹۸\mathrm{E}-۱$	$-۶٫۹۹\mathrm{E}-۳$	$-۶٫۹۹\mathrm{E}-۴$
	۲	$۶٫۹۰\mathrm{E}+۰$	$-۲٫۹۰\mathrm{E}-۳$	$-۳٫۰۸\mathrm{E}-۴$
	۳	$۱٫۱۱\mathrm{E}+۲$	$۳٫۸۶\mathrm{E}-۳$	$۳٫۶۴\mathrm{E}-۴$
	۴	$۱٫۷۷\mathrm{E}+۳$	$۷٫۰۷\mathrm{E}-۳$	$۷٫۰۴\mathrm{E}-۴$
	۵	$۲٫۸۳\mathrm{E}+۴$	$۳٫۷۸\mathrm{E}-۳$	$۳٫۹۷\mathrm{E}-۴$
-۵۰	۱	$۳٫۲۶\mathrm{E}+۰$	$۱٫۰۶\mathrm{E}+۳$	$-۱٫۳۹\mathrm{E}-۴$
	۲	$۱٫۸۸\mathrm{E}+۳$	$۱٫۱۱\mathrm{E}+۹$	$-۵٫۱۶\mathrm{E}-۵$
	۳	$۱٫۰۸\mathrm{E}+۶$	$۱٫۱۷\mathrm{E}+۱۵$	$۸٫۲۵\mathrm{E}-۵$
	۴	$۶٫۲۴\mathrm{E}+۸$	$۱٫۲۳\mathrm{E}+۲۱$	$۱٫۴۱\mathrm{E}-۴$
	۵	$۳٫۵۹\mathrm{E}+۱۱$	$۱٫۲۸\mathrm{E}+۲۷$	$۷٫۰۰\mathrm{E}-۵$

که $Y(x) = \sin(x) + \cos(x)$ جواب صحیح آن است، به‌کار برید. ما نتایج را برای چندین مقدار λ و h داده‌ایم. برای $\lambda = -۱, -۱۰, -۵۰$، کران (۴۹.۸.۶) برای h به ترتیب کرانهایی به شکل زیر به‌دست می‌دهد.

$$۰ < h < ۲ \qquad ۰ < h < \frac{۱}{۵} = ۰٫۲ \qquad ۰ < h < \frac{۱}{۲۵} = ۰٫۰۴$$

همان‌گونه که از جدول ۱۷.۶ می‌توان دید، استفاده از مقادیر بزرگتر h نتایج عددی نامطلوبی به بار می‌آورد.

آنچه در بالا با روش اویلر به‌دست آوردیم انگیزهٔ روش کلی، برای پیداکردن مجموعهٔ همهٔ $h\lambda$ هایی است که برای آنها روش (۱.۸.۶) پایدار است. چون ما فقط حالتهای $\mathrm{Real}(\lambda) < ۰$ را درنظر می‌گیریم، $\{y_n\}$، جواب عددی (۱.۸.۶) در حل معادلهٔ نمونهٔ $y' = \lambda y$ را می‌خواهیم که با $x_n \to \infty$ برای همهٔ مقادیر اولیهٔ $y_۰$، ...، y_p به صفر میل کند. مجموعهٔ همهٔ $h\lambda$ هایی که برای آنها این موضوع درست باشد **ناحیهٔ پایداری مطلق** روش (۱.۸.۶) خوانده می‌شود. هرچه

این ناحیه وسیعتر باشد، محدودیت h برای داشتن جواب عددی پایدار، کمتر است.

وقتی (۱.۸.۶) برای معادلهٔ نمونه به‌کار رود، معادلهٔ قبلی (۱۸.۸.۶) را به‌دست می‌آوریم، و به شرطی که ریشه‌های مشخصهٔ $r_0(h\lambda)$، ...، $r_p(h\lambda)$ متمایز باشند جواب آن با (۲۲.۸.۶)، یعنی

$$y_n = \sum_{j=0}^{p} \gamma_j [r_j(h\lambda)]^n \qquad n \geq 0$$

داده می‌شود. برای آنکه وقتی $n \to \infty$، این جواب به‌ازای همهٔ مقادیر اولیهٔ y_0، ...، y_p به صفر میل کند، لازم و کافی است که داشته باشیم

$$| r_j(h\lambda) | < 1 \qquad j = 0, 1, \ldots, p \qquad (۵۲.۸.۶)$$

مجموعهٔ همهٔ مقادیر $h\lambda$ که در این مجموعهٔ نامساویها صدق کنند نیز ناحیهٔ پایداری مطلق خوانده می‌شود. این ناحیه در مجموعه‌ای که در پاراگراف قبلی تعریف شده قرار دارد، و معمولاً با آن مجموعه برابر است. در پیداکردن ناحیهٔ پایداری مطلق، فقط با (۵۲.۸.۶) کار می‌کنیم.

مثال روش مرتبهٔ دو آدامز ـ بشفورت

$$y_{n+1} = y_n + \frac{h}{2}[3y'_n - y'_{n-1}] \qquad n \geq 1 \qquad (۵۳.۸.۶)$$

را درنظر می‌گیریم. معادله مشخصه چنین است

$$r^2 - (1 + \frac{3}{2}h\lambda)r + \frac{1}{2}h\lambda = 0$$

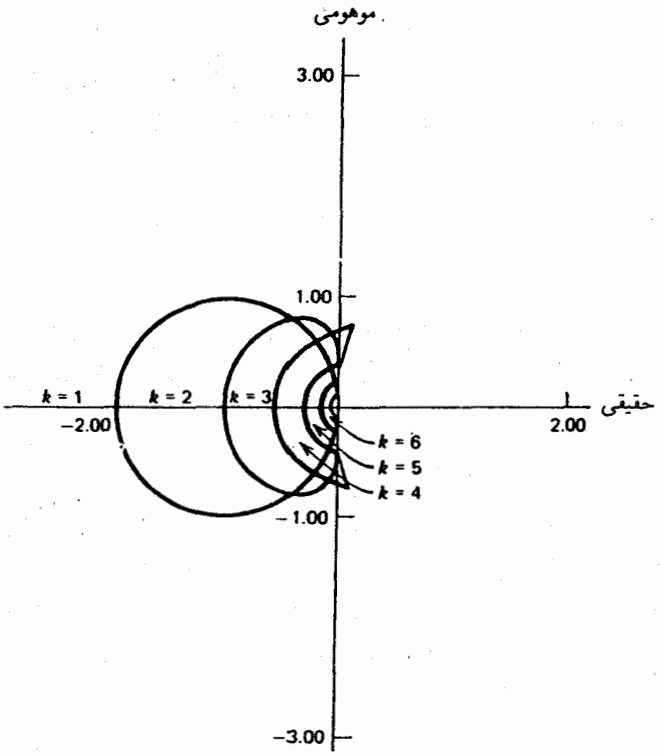
و ریشه‌ها عبارت‌اند از

$$r_0 = \frac{1}{2}\left\{1 + \frac{3}{2}h\lambda + \sqrt{1 + h\lambda + \frac{9}{4}h^2\lambda^2}\right\}$$
$$r_1 = \frac{1}{2}\left\{1 + \frac{3}{2}h\lambda - \sqrt{1 + h\lambda + \frac{9}{4}h^2\lambda^2}\right\}$$

ناحیهٔ پایداری مطلق، مجموعهٔ $h\lambda$ هاست که برای آنها

$$| r_0(h\lambda) | < 1 \qquad | r_1(h\lambda) | < 1$$

به‌ازای مقادیر حقیقی λ، مقادیر قابل قبول $h\lambda$، عبارت‌اند از $-1 < h\lambda < 0$.

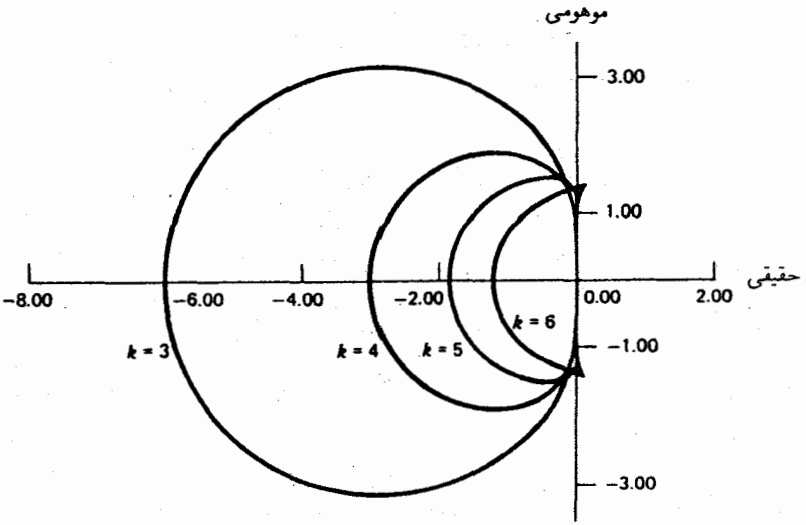


شکل ۷.۶ نواحی پایداری برای روشهای آدامز ـ بشفورت. روش مرتبهٔ k در داخل ناحیه‌ای که در سمت چپ مرکز مشخص شده پایدار است [با اجازه، از صفحهٔ ۱۳۱ گی‌یر (۱۹۷۱) گرفته شده است.]

کرانهای نواحی پایداری مطلق روشهای آدامز ـ بشفورت و آدامز ـ مولتن به ترتیب در شکلهای ۷.۶ و ۸.۶ داده شده‌اند. در مورد فرمولهای آدامز ـ مولتن با یک بارست از پیشگوی آدامز ـ بشفورت، نواحی پایداری مطلق در شمپاین و گوردون (۱۹۷۵، صص ۱۳۵ـ۱۴۰) داده شده‌اند.

از این نمودارها، دیده می‌شود که وقتی مرتبهٔ روش افزایش یابد، ناحیهٔ همگرایی مطلق کوچکتر می‌شود. و برای فرمولهای هم‌مرتبه، در دو فرمول آدامز ـ مولتن و آدامز ـ بشفورت، ناحیهٔ همگرایی مطلق در اولی به مراتب بزرگتر است. اندازهٔ این ناحیه‌ها، معمولاً از نقطهٔ نظر عملی به خوبی قابل‌قبول‌اند. برای مثال، مقادیر حقیقی $h\lambda$ در ناحیهٔ پایداری مطلق برای فرمول مرتبهٔ چهار آدامز ـ مولتن با $-۳ < h\lambda < ۰$ داده شده است. در اغلب حالات، این محدودیّت زیادی برای h نخواهد بود.

خانوادهٔ فرمولهای آدامز، برای ایجاد الگوریتمهای متغیر ـ مرتبه، بسیار مناسب‌اند، و ناحیه‌های پایداری کاملاً قابل‌قبول‌اند. این روشها برای مسائلی که λ منفی و قدرمطلق آن بزرگ باشد، مشکلاتی



شکل ۸.۶ نواحی پایداری برای روشهای آدامز ـ بشفورت. روش مرتبهٔ k در داخل ناحیه‌ای که مشخص شده پایدار است. [از صفحهٔ ۱۳۱ گی‌یر (۱۹۷۱) با اجازه برداشته شده است.]

دارند، و این نوع مسائل با روشهای دیگر بهتر حل می‌شوند، که در بخش بعد ملاحظه خواهیم نمود.

روشهای ویژه‌ای وجود دارند که برای آنها ناحیهٔ پایداری مطلق شامل تمام مقادیر مختلط $h\lambda$ با $\text{Real}\ (\lambda) < 0$ است. این روشها Aـ**پایدار** خوانده می‌شوند، و با آنها برای داشتن پایداری از نوعی که ملاحظه کرده‌ایم، هیچ محدودیتی برای h وجود نخواهد داشت. قاعدهٔ ذوزنقه‌یی یک مثال از چنین روش است. [(۲۴.۵.۶) ـ (۲۵.۵.۶) را ببینید].

مثال روش پسرو اویلر را در نظر می‌گیریم:

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) \qquad n \geq 0 \qquad (۵۴.۸.۶)$$

با به‌کاربردن این روش در معادلهٔ نمونهٔ $y' = \lambda y$ و حل آن، داریم

$$y_n = \left[\frac{1}{1 - h\lambda}\right]^n y_0 \qquad n \geq 0 \qquad (۵۵.۸.۶)$$

بنابراین وقتی $x_n \to \infty$، $y_n \to 0$ اگر و تنها اگر

$$\frac{1}{|1 - h\lambda|} < 1$$

جدول ۱۸.۶ مثال قاعدهٔ ذوزنقه: $h = ۰٫۵$

x	خطا: $\lambda = -۱$	خطا: $\lambda = -۱۰$	خطا: $\lambda = -۵۰$
۲	$-۱٫۱۳\mathrm{E}-۲$	$-۲٫۷۸\mathrm{E}-۳$	$-۷٫۹۱\mathrm{E}-۴$
۴	$-۱٫۴۳\mathrm{E}-۲$	$-۸٫۹۱\mathrm{E}-۵$	$-۸٫۹۱\mathrm{E}-۵$
۶	$۲٫۰۲\mathrm{E}-۲$	$۲٫۷۷\mathrm{E}-۳$	$۴٫۷۲\mathrm{E}-۴$
۸	$-۲٫۸۶\mathrm{E}-۳$	$-۲٫۲۲\mathrm{E}-۳$	$-۵٫۱۱\mathrm{E}-۴$
۱۰	$-۱٫۷۹\mathrm{E}-۲$	$-۹٫۲۳\mathrm{E}-۴$	$-۱٫۵۶\mathrm{E}-۴$

این رابطه به‌ازای جمیع مقادیر $h\lambda$ با $\mathrm{Real}(\lambda) < ۰$ درست است و روش پسرو اویلر یک روش A ـ پایدار است.

مثال روش ذوزنقه‌یی را برای مسألهٔ (۵۱.۸.۶)، که قبلاً با روش اویلر حل شده به‌کار می‌بریم. از طول گام $h = ۰٫۵$ را برای $\lambda = -۱, -۱۰, -۵۰$ استفاده می‌کنیم. نتایج در جدول ۱۸.۶ داده شده‌اند. این نتایج نشان می‌دهند که وقتی $|\lambda|$ افزایش می‌یابد $\mathrm{Real}(\lambda) < ۰$، قاعدهٔ ذوزنقه‌یی ناپایدار نمی‌شود.

داشتن روشهای A ـ پایدار از مراتب بزرگتر از ۲ مفید خواهد بود. ولی یک قضیه از دال‌کوییست[1] (۱۹۶۳) نشان می‌دهد که چنین روشهایی وجود ندارند. ما در بخش آینده بعضی روشهای مراتب بالاتر را که بیشترین ویژگیهای لازم پایداری را دارا هستند بررسی خواهیم کرد.

۹.۶ معادلات دیفرانسیل سرسخت و روش خطوط

حل عددی معادلات دیفرانسیل سرسخت، در طی ده پانزده سال اخیر بسیار مورد مطالعه قرار گرفته است. این‌گونه معادلات (از جمله دستگاههای معادلات دیفرانسیل) درکاربردهای روزافزونی در موضوعات بسیار دور از هم مثل جنبش‌شناسی شیمیایی و حل عددی معادلات با مشتقات جزیی ظاهر شده‌اند. در این بخش ما بعضی از نکات اصلی این موضوع را مطرح می‌نماییم و رابطهٔ آن را با حل عددی یک معادلهٔ سادهٔ گرما نشان می‌دهیم.

تعاریف زیادی از مفهوم معادلهٔ دیفرانسیل سرسخت وجود دارد. مهمترین جنبهٔ مشترک این تعاریف آن است که چنین معادلاتی وقتی با روشهای معمولی (مثلاً، روشهای آدامز، بخش ۷.۶) حل شوند، طول گام h اجباراً باید بسیار کوچک گرفته شود تا پایداری تأمین گردد ـ بسیار کوچکتر از آنچه ممکن است در مورد خطای برشی، لازم به‌نظر آید. یک گواه این، حل معادلهٔ (۵۱.۸.۶)

1. Dahlquist

به روش اویلر است، که در جدول ۱۷.۶ داده شده است. در آن حالت، مجهول $Y(x)$ با λ تغییر نمی‌کرد و بنابراین خطای برشی مستقل از λ بود. ولی خطای واقعی قویاً تحت تأثیر قدرمطلق λ قرار گرفته بود، زیرا لازم بود که λ در شرط پایداری $|1 + h\lambda| < 1$ صدق نماید تا همگرایی به‌دست آید. وقتی $|\lambda|$ افزایش می‌یافت، اندازه h باید مطابق با آن کوچک می‌شد. این رفتار وقتی از روشهای عددی استانده برای معادلات دیفرانسیل سرسخت استفاده می‌شود امری عادی است، با این تفاوت عمده که مقادیر واقعی $|\lambda|$ بسیار بزرگترند، برای مثال $\lambda = -10^6$.

اکنون به متداولترین ردهٔ این معادلات دیفرانسیل می‌نگریم، بر این پایه که بررسی خود را به خطی‌نمودن دستگاه $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$، که در (۱۴.۸.۶) تا (۱۷.۸.۶) شرح داده شده معطوف می‌داریم:

$$\mathbf{y}' = \Lambda\mathbf{y} + \mathbf{g}(x) \qquad (۱.۹.۶)$$

با $\Lambda = \mathbf{f}_y(x_0, \mathbf{Y}_0)$، ماتریس ژاکوبی $\mathbf{f}$. معادلهٔ دیفرانسیل $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$ را **سرسخت** گوییم اگر λ_j، یکی از ویژه مقدارهای Λ، یا به‌طورکلیتر $\mathbf{f}_y(x, \mathbf{y})$، دارای قسمت حقیقی منفی با اندازهٔ بسیار بزرگ باشد. ما روشهای عددی برای معادلات دشوار را با درنظرگرفتن اثر آنها روی معادلهٔ نمونهٔ

$$y' = \lambda y + g(x) \qquad (۲.۹.۶)$$

مطالعه می‌کنیم که $\text{Real}(\lambda)$ منفی و از لحاظ اندازه بسیار بزرگ است. این روش، محدودیتهای خاص خود را دارد که بعضی از آنها را بعداً نشان خواهیم داد، ولی به ما امکاناتی می‌دهد که روشهای غیر رضایتبخش راکنار بگذاریم، و همچنین ما را به فکر بعضی روشهای رضایتبخش ممکن می‌رساند.

مفهوم ناحیهٔ پایداری مطلق، که در بخش اخیر وارد شد، ابزار اولیه در مطالعهٔ پایداری یک روش عددی در حل معادلات دیفرانسیل سرسخت است. ما روشهایی را جستجو می‌کنیم که ناحیهٔ پایداری آنها شامل تمام محور حقیقی منفی، و همچنین تا آنجا که ممکن است، شامل ناحیهٔ چپ صفحهٔ مختلط باشد. برای پیداکردن این گونه روشها چند راه وجود دارد، ولی ما فقط یکی از آنها را، که **فرمولهای مشتقگیری پسرو**[1] $(BDFs)$ را به‌دست می‌دهد، مورد بحث قرار می‌دهیم.

گیریم $P_p(x)$ یک چندجمله‌یی از درجهٔ نابزرگتر از p باشد که $Y(x)$ را در نقاط x_n، x_{n+1}، ... ، x_{n-p+1} برای مقداری از $p \geq 1$، درونیابی نماید:

$$P_p(x) = \sum_{j=-1}^{p-1} Y(x_{n-j}) l_{j,n}(x) \qquad (۳.۹.۶)$$

1. backward differentiation formulas

جدول ۱۹.۶ ضرایب روش BDF (۶.۹.۶)

p	β	α_0	α_1	α_2	α_3	α_4	α_5
۱	۱	۱					
۲	$\frac{۲}{۳}$	$\frac{۴}{۳}$	$-\frac{۱}{۳}$				
۳	$\frac{۶}{۱۱}$	$\frac{۱۸}{۱۱}$	$-\frac{۹}{۱۱}$	$\frac{۲}{۱۱}$			
۴	$\frac{۱۲}{۲۵}$	$\frac{۴۸}{۲۵}$	$-\frac{۳۶}{۲۵}$	$\frac{۱۶}{۲۵}$	$-\frac{۳}{۲۵}$		
۵	$\frac{۶۰}{۱۳۷}$	$\frac{۳۰۰}{۱۳۷}$	$-\frac{۳۰۰}{۱۳۷}$	$\frac{۲۰۰}{۱۳۷}$	$-\frac{۷۵}{۱۳۷}$	$\frac{۱۲}{۱۳۷}$	
۶	$\frac{۶۰}{۱۴۷}$	$\frac{۳۶۰}{۱۴۷}$	$-\frac{۴۵۰}{۱۴۷}$	$\frac{۴۰۰}{۱۴۷}$	$-\frac{۲۲۵}{۱۴۷}$	$\frac{۷۲}{۱۴۷}$	$-\frac{۱۰}{۱۴۷}$

که در آن $\{l_{j,n}(x)\}$ توابع پایهٔ درونیابی لاگرانژ برای نقاط گرهی x_{n+1}، ...، x_{n-p+1} هستند [(۵.۱.۳) را ببینید.] از رابطهٔ زیر استفاده می‌کنیم

$$P_p'(x_{n+1}) \doteq Y'(x_{n+1}) = f(x_{n+1}, Y(x_{n+1})) \qquad (۴.۹.۶)$$

از ترکیب با (۳.۹.۶) و حل آن نسبت به $Y(x_{n+1})$، داریم

$$Y(x_{n+1}) \doteq \sum_{j=0}^{p-1} \alpha_j Y(x_{n-j}) + h\beta f(x_{n+1}, Y(x_{n+1})) \qquad (۵.۹.۶)$$

روش p گامی فرمول مشتقگیری پسرو با رابطهٔ زیر داده می‌شود

$$y_{n+1} = \sum_{j=0}^{p-1} \alpha_j y_{n-j} + h\beta f(x_{n+1}, y_{n+1}) \qquad (۶.۹.۶)$$

ضرایب در حالتهای $p = 1, \ldots, 6$ در جدول ۱۹.۶ داده شده‌اند. حالت $p = 1$ همان روش پسرو اویلر است، که به‌دنبال (۵۴.۸.۶) در بخش اخیر بحث شد. خطای برشی برای (۶.۹.۶) را از فرمول خطا برای مشتقگیری عددی، که در (۵.۷.۵) داده شده بود، می‌توان به‌دست آورد:

$$T_n(Y) = -\frac{\beta}{p+1} h^{p+1} Y^{(p+1)}(\xi_n) \qquad (۷.۹.۶)$$

برای مقداری از $x_{n-p+1} \le \xi_n \le x_{n+1}$.

ناحیه‌های پایداری مطلق برای فرمولهای جدول ۱۹.۶ درگی‌پر (۱۹۷۱، صص ۲۱۵ـ۲۱۶) داده شده‌اند. برای ساختن این نواحی، باید تمام مقادیر $h\lambda$ را که برای آنها

$$| r_j(h\lambda) | < 1 \qquad j = 0, \ldots, p \qquad (8.9.6)$$

پیدا کنیم، که ریشه‌های مشخصهٔ $r_j(h\lambda)$ جوابهای معادلهٔ

$$r^p = \sum_{j=0}^{p-1} \alpha_j r^{p-1-j} + h\lambda\beta r^p \qquad (9.9.6)$$

هستند. می‌توان نشان داد که برای $p = 1$ و $p = 2$، فرمولهای مشتقگیری پسرو A ـ پایدارند، و برای $3 \leq p \leq 6$، ناحیهٔ پایداری مطلق با افزایش p، کوچک می‌شود، گرچه در هر حالت تمام محور حقیقی منفی را در بر می‌گیرد. برای $p \geq 7$، ناحیه‌های پایداری مطلق، برای حل مسائل سرسخت، قابل‌قبول نیستند. برای بحث بیشتر در مورد این ناحیه‌های پایداری، گی‌پر (۱۹۷۱، فصل ۱۱) و لامبرت (۱۹۷۳، فصل ۸) را ببینید.

باز هم با روشهای BDF و با سایر روشهایی که فقط بر پایهٔ ناحیهٔ پایداری مطلق انتخاب شده‌اند، اشکالاتی وجود دارند. ابتدا، با معادلهٔ نمونهٔ $y' = \lambda y$ آغاز می‌کنیم، اگر $\text{Real}(\lambda)$ منفی و قدرمطلق آن بزرگ باشد، آنگاه جواب $Y(x)$ بسیار سریع به صفر می‌گراید، و چنانچه $\text{Real}(\lambda)$ افزایش یابد، $Y(x)$ سریعتر به صفر همگرا می‌شود، ما می‌خواهیم که همین رفتار برای جواب عددی معادلهٔ نمونهٔ $\{y_n\}$ برقرار باشد. ولی با قاعدهٔ ذوزنقه‌یی A ـ پایدار، جواب [از (۲۴.۵.۶)] چنین است

$$y_n = \left[\frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}}\right]^n y_0 \qquad n \geq 0$$

اگر $| \text{Real}(\lambda) |$ بزرگ باشد، آنگاه کسر داخل کروشه نزدیک -1 می‌شود، و y_n خیلی آهسته به صفر کاهش می‌یابد. با استفاده از نوع استدلالی که برای روش اویلر در (۴۵.۸.۶) ـ (۵۰.۸.۶) به‌کار رفت، اثر اختلالات، برای مقادیر بزرگ λ، به‌سرعت به صفر کاهش نمی‌یابد. بنابراین ممکن است روش ذوزنقه‌یی برای مسائل سرسخت، یک انتخاب کاملاً رضایتبخش نباشد. در مقایسه، روش پسرو A ـ پایدار اویلر رفتار مطلوب را داراست. با توجّه به (۵۵.۸.۶)، جواب مسألهٔ نمونه چنین است

$$y_n = \left[\frac{1}{1 - h\lambda}\right]^n y_0 \qquad n \geq 0$$

وقتی $|\lambda|$ افزایش می‌یابد، دنبالهٔ $\{y_n\}$ سریعتر به صفر می‌گراید. بنابراین جواب پسرو اویلر رفتار جواب درست معادلهٔ نمونه را بهتر منعکس می‌نماید.

اشکال دوم در مورد ناحیه‌های پایداری این است که این ناحیه‌ها بر مبنای مقدار ثابت λ و مسائل خطی پایه‌گذاری شده‌اند. خطی‌کردن (۱.۹.۶) اغلب معتبر است ولی نه همیشه. برای مثال، مسألهٔ خطی مرتبه دوّم زیر را که در آن یکی از ضرایب ثابت نیست در نظر می‌گیریم:

$$y'' + ay' + (1 + b \cdot \cos(2\pi x))y = g(x) \qquad x \geq 0 \qquad (10.9.6)$$

آن را به دستگاه هم‌ارز زیر برمی‌گردانیم

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= -(1 + b\cos(2\pi x))y_1 - ay_2 + g(x) \end{aligned} \qquad (11.9.6)$$

فرض می‌کنیم که $a > 0$ و $|b| < 1$. ویژه مقدارهای معادلهٔ همگن $[g(x) \equiv 0]$ عبارت‌اند از

$$\lambda = \frac{-a \pm \sqrt{a^2 - 4[1 + b.\cos(2\pi x)]}}{2} \qquad (12.9.6)$$

این ویژه مقدارها، اعداد حقیقی منفی یا اعداد مختلط با قسمت حقیقی منفی هستند. براساس نظریهٔ پایداری برای حالت ضرایب ثابت (یا Λ ثابت)، فرض خواهیم کرد که اثر تمام اختلالات در داده‌های اولیه وقتی $x \to \infty$، از بین می‌روند. ولی درواقع، قسمت همگن (۱۰.۹.۶) جوابهای بیکران خواهد داشت. بنابراین اختلالاتی در مقادیر اولیه وجود دارند که موجب جوابهای اختلال‌یافتهٔ بیکران در (۱۰.۹.۶) می‌شوند. این موضوع، اعتبار استفاده از معادلهٔ نمونهٔ $y' = \lambda y + g(x)$ را زیر سؤال می‌برد. کاربرد آن روشهایی را می‌طلبد که ممکن است بخواهیم بیشتر دربارهٔ آنها مطالعه کنیم، ولی این روش بخودی خود، برای دربرگرفتن انواع فراوان مسائل خطی و غیرخطی کافی نخواهد بود. مثال (۱۰.۹.۶) از اِیکِن[1]، (۱۹۸۵، ص ۲۶۹) گرفته شده است.

حل روش تفاضلات متناهی ما این مسأله را با در نظرگرفتن روش پسرو اویلر نشان می‌دهیم:

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) \qquad n \geq 0 \qquad (13.9.6)$$

اگر فرمول معمولی بارستی

$$y_{n+1}^{(j+1)} = y_n + hf(x_{n+1}, y_{n+1}^{(j)}) \qquad j \geq 0 \qquad (14.9.6)$$

به‌کار رود، آنگاه

1. Aiken

$$y_{n+1} - y_{n+1}^{(j+1)} \doteq h\frac{\partial f(x_{n+1}, y_{n+1})}{\partial y}[y_{n+1} - y_{n+1}^{(j)}]$$

برای همگرایی نیاز داریم که داشته باشیم

$$| h\frac{\partial f(x_{n+1}, y_{n+1})}{\partial y} | < 1 \qquad (۱۵.۹.۶)$$

ولی در معادله‌های سرسخت، لازمهٔ این رابطه بسیار کوچک‌بودن h است، چیزی که ما می‌خواهیم از آن پرهیز نماییم. بنابراین روش ریشه‌یابی دیگری باید برای پیداکردن y_{n+1} در (۱۳.۹.۶) به‌کار بریم.

متداولترین روشها برای حل (۱۳.۹.۶) روشهایی هستند که بر پایهٔ روش نیوتن استوارند. برای یک معادلهٔ دیفرانسیل، روش نیوتن برای پیداکردن y_{n+1} چنین است

$$y_{n+1}^{(j+1)} = y_{n+1}^{(j)} - [1 - hf_y(x_{n+1}, y_{n+1}^{(j)})]^{-1}[y_{n+1}^{(j)} - y_n - hf(x_{n+1}, y_{n+1}^{(j)})] \qquad (۱۶.۹.۶)$$

برای $j \geq 0$. یک حدس اولیهٔ خام $y_{n+1}^{(0)} = y_n$ است، اگرچه معمولاً این جواب را می‌توان بهتر کرد. برای دستگاه معادلات دیفرانسیل، روش نیوتن خیلی گران می‌شود. برای کاهش هزینه، ماتریس

$$I - h\mathbf{f}_y(x_{n+1}, \mathbf{z}) \qquad \text{برای مقداری از } \mathbf{z} \doteq \mathbf{y}_n \qquad (۱۷.۹.۶)$$

را برای جمیع مقادیر j و برای چندین مقدار پی‌درپی n به‌کار می‌بریم. بنابراین روش نیوتن [بخش ۱۱.۲ را ببینید] برای حل دستگاه نمونهٔ (۱۳.۹.۶) با رابطهٔ زیر تقریب زده می‌شود

$$\begin{aligned}[I - h\mathbf{f}_y(x_{n+1}, \mathbf{z})]\boldsymbol{\delta}^{(j)} &= \mathbf{y}_{n+1}^{(j)} - \mathbf{y}_n - h\mathbf{f}(x_{n+1}, \mathbf{y}_{n+1}^{(j)}) \\ \mathbf{y}_{n+1}^{(j+1)} &= \mathbf{y}_{n+1}^{(j)} + \boldsymbol{\delta}^{(j)}\end{aligned} \qquad (۱۸.۹.۶)$$

برای $j \geq 0$. این مسأله به حل چند دستگاه خطی که یک ماتریس ضرایب دارند، می‌انجامد. این کار را در مقایسه با وقتی که ماتریس اصلاح شده است، بسیار آسانتر می‌توان انجام داد (مطالب بخش ۱.۸ را ببینید). ماتریس رابطهٔ (۱۷.۹.۶) باید متناوباً روزآمد شود، ولی در مقایسه با روش دقیق نیوتن، صرفه‌جویی هنوز هم خیلی مهم است. برای بحث بیشتر این موضوع اِیکن (۱۹۸۵) و گوپتا و همکاران (۱۹۸۵، صص ۲۲ـ۲۵) را ببینید. برای یک بازنگری برنامه‌های رایانه‌یی در حل معادلات دیفرانسیل سرسخت، اِیکن (۱۹۸۵، فصل ۴) را ببینید.

روش خطوط معادلهٔ دیفرانسیل جزیی سهموی زیر را در نظر می‌گیریم

$$U_t = U_{xx} + G(x,t) \qquad ۰ < x < ۱ \qquad t > ۰ \tag{۱۹.۹.۶}$$

$$U(۰,t) = d_۰(t) \qquad U(۱,t) = d_۱(t) \qquad t \geq ۰ \tag{۲۰.۹.۶}$$

$$U(x,۰) = f(x) \qquad ۰ \leq x \leq ۱ \tag{۲۱.۹.۶}$$

نمادهای U_t و U_{xx} به ترتیب مشتقات جزئی نسبت به t و x را نشان می‌دهند. تابع مجهول $U(x,t)$ به زمان t و متغیر مکانی x بستگی دارد. شرایط (۲۰.۹.۶)، شرایط مرزی خوانده می‌شوند، و (۲۱.۹.۶) شرط اولیه نامیده می‌شود. جواب U را می‌توان دمای یک میلهٔ عایق‌دار به طول ۱ تعبیر نمود، که $U(x,t)$ دما در مکان x و زمان t است؛ بنابراین (۱۹.۹.۶) اغلب **معادلهٔ گرما** نامیده می‌شود. توابع G، $d_۰$، $d_۱$ و f داده شده و هموار فرض می‌شوند. برای مطالعهٔ مبانی نظری (۱۹.۹.۶) ـ (۲۱.۹.۶)، ویدر[1] (۱۹۷۵) یا هر کتاب مقدماتی معمولی در معادلات دیفرانسیل جزئی را ببینید. ما در اینجا **روش خطوط** را برای حلّ U می‌آوریم، روشی عددی که در ده پانزده سال اخیر بسیار متداول شده است. این روش هم به حل یک دستگاه معادلات دیفرانسیل معمولی سرسخت می‌انجامد.

گیریم $m > ۰$ یک عدد صحیح باشد، $۱/m$ را با δ نشان می‌دهیم و تعریف می‌کنیم،

$$x_j = j\delta \qquad j = ۰, ۱, \ldots, m$$

معادلهٔ (۱۹.۹.۶) را با تقریب‌زدن مشتق مکانی، گسسته می‌سازیم. فرمولهای (۱۷.۷.۵) و (۱۸.۷.۵) را برای تقریب‌زدن مشتقات مرتبهٔ دوّم به یاد می‌آوریم. با استفاده از این فرمولها،

$$U_{xx}(x_j,t) = \frac{U(x_{j+۱},t) - ۲U(x_j,t) + U(x_{j-۱},t)}{\delta^۲} - \frac{\delta^۲}{۱۲}\frac{\partial^۴ U(\xi_j,t)}{\partial x^۴}$$

برای $j = ۱, ۲, \ldots, m-۱$. با گذاردن در (۱۹.۹.۶) خواهیم داشت:

$$U_t(x_j,t) = \frac{U(x_{j+۱},t) - ۲U(x_{j,t}) + U(x_{j-۱},t)}{\delta^۲} + G(x_j,t) - \frac{\delta^۲}{۱۲}\cdot\frac{\partial^۴ U(\xi_j,t)}{\partial x^۴} \qquad ۱ \leq j \leq m-۱ \tag{۲۲.۹.۶}$$

معادلهٔ (۱۹.۹.۶) را باید در هر نقطهٔ گرهی داخلی x_j تقریب بزنیم. مجهول ξ_j به $[x_{j-۱}, x_{j+۱}]$ تعلق دارد: $\xi_j \in [x_{j-۱}, x_{j+۱}]$.

1. Widder

جملهٔ آخر در (۲۲.۹.۶)، خطای برشی در مشتقگیری عددی، را حذف می‌کنیم. با تحمیل تساوی در معادلهٔ تقریبی حاصل‌شده، به‌دست می‌آوریم

$$u_j'(t) = \frac{۱}{\delta^۲}[u_{j+۱}(t) - ۲u_j(t) + u_{j-۱}(t)] + G(x_j, t) \qquad (۲۳.۹.۶)$$

برای $j = ۱, ۲, \ldots, m-۱$. توابع $u_j(t)$ تقریبهای $U(x_j, t)$، $۱ \le j \le m-۱$، درنظر گرفته شده‌اند. این فرمول تقریب (۱۹.۹.۶) با **روش خطوط** است، و یک دستگاه $m-۱$ معادلهٔ دیفرانسیل معمولی است. توجه می‌کنیم که $u_۰(t)$ و $u_m(t)$ که به‌ازای $j = ۱$ و $j = m-۱$ در (۲۳.۹.۶) موردنیازند، با استفاده از (۲۰.۹.۶) به‌دست آمده‌اند:

$$u_۰(t) = d_۰(t) \qquad u_m(t) = d_۱(t) \qquad (۲۴.۹.۶)$$

شرط اولیه برای (۲۳.۹.۶) به توسط (۲۱.۹.۶) داده شده است:

$$u_j(۰) = f(x_j) \qquad ۱ \le j \le m-۱ \qquad (۲۵.۹.۶)$$

نامگذاری **روش خطوط** از حل $U(x,t)$ در طول خطوط (x_j, t)، $t \ge ۰$، $۱ \le j \le m-۱$، در صفحهٔ (x,t) گرفته شده است.

تحت مفروضات هموارای مناسب در توابع $d_۰$، $d_۱$، G و f، می‌توان نشان داد که

$$\underset{\substack{۰\le j\le m\\ ۰\le t\le T}}{\text{Max}} \mid U(x_j,t) - u_j(t) \mid \le C_T\delta^۲ \qquad (۲۶.۹.۶)$$

بنابراین برای تکمیل فرایند حل، فقط لازم است دستگاه (۲۳.۹.۶) را حل کنیم.

راحت‌تر است که (۲۳.۹.۶) را به شکل ماتریسی بنویسیم. نمادهای زیر را وارد می‌کنیم

$$\mathbf{u}(t) = [u_۱(t), \ldots, u_{m-۱}(t)]^T \qquad \mathbf{u}_۰ = [f(x_۱), \ldots, f_{m-۱}(x)]^T$$

$$\mathbf{g}(t) = \left[\frac{۱}{\delta^۲}d_۰(t) + G(x_۱,t), G(x_۲,t), \ldots, G(x_{m-۲},t), \frac{۱}{\delta^۲}d_۱(t) + G(x_{m-۱},t)\right]^T$$

$$\Lambda = \frac{۱}{\delta^۲}\begin{bmatrix} -۲ & ۱ & ۰ & ۰ & \cdots & ۰ \\ ۱ & -۲ & ۱ & ۰ & & \vdots \\ ۰ & ۱ & -۲ & ۱ & & \\ \vdots & & & \ddots & & ۰ \\ & & & & ۱ & -۲ & ۱ \\ ۰ & ۰ & \cdots & & ۰ & ۱ & -۲ \end{bmatrix} \qquad (۲۷.۹.۶)$$

ماتریس Λ از مرتبهٔ $m-1$ است. در تعریفهای $\mathbf{u}$ و $\mathbf{g}$، اندیس فوقانی T ترانهادهٔ ماتریس را نشان می‌دهد، لذا $\mathbf{u}$ و $\mathbf{g}$ بردارهای ستونی به طول $m-1$ هستند. با استفاده از این ماتریسها، معادلهٔ (۲۳.۹.۶) ـ (۲۵.۹.۶) را می‌توان دوباره به شکل زیر نوشت

$$\mathbf{u}'(t) = \Lambda \mathbf{u}(t) + g(t) \qquad \mathbf{u}(0) = \mathbf{u}_0 \qquad (۲۸.۹.۶)$$

اگر از روش اویلر استفاده کنیم، روش عددی زیر را خواهیم داشت:

$$\mathbf{V}_{n+1} = \mathbf{V}_n + h[\Lambda \mathbf{V}_n + \mathbf{g}(t_n)] \qquad \mathbf{V}_0 = \mathbf{u}_0 \qquad (۲۹.۹.۶)$$

که در آن $t_n = nh$ و $\mathbf{V}_n \doteq \mathbf{u}(t_n)$. این روش یک روش عددی معروفی برای معادلهٔ گرماست، که **روش صریح ساده** نامیده می‌شود. پایداری (۲۹.۹.۶) و چند روش دیگر برای حل (۲۸.۹.۶) را بررسی می‌نماییم.

معادلهٔ (۲۸.۹.۶) به شکل معادلهٔ نمونهٔ (۱.۹.۶) است، و بنابراین به‌ویژه مقدارهای λ برای امتحان سرسختی دستگاه نیاز داریم. می‌توان نشان داد که این ویژه مقدارها همگی حقیقی‌اند و با رابطهٔ زیر داده می‌شوند:

$$\lambda_j = -\frac{4}{\delta^2} \sin^2\left(\frac{j\pi}{2m}\right) \qquad 1 \le j \le m-1 \qquad (۳۰.۹.۶)$$

اثبات این را به مسألهٔ ۶ فصل ۷ احاله می‌نماییم. اگر مستقیماً این فرمول را امتحان کنیم، داریم

$$\lambda_{m-1} \le \lambda_j \le \lambda_1 \qquad (۳۱.۹.۶)$$

$$\lambda_{m-1} = -\frac{4}{\delta^2} \sin^2\left(\frac{(m-1)\pi}{2m}\right) \doteq \frac{-4}{\delta^2}$$

$$\lambda_1 = -\frac{4}{\delta^2} \sin^2\left(\frac{\pi}{2m}\right) \doteq -\pi^2$$

این تقریبها برای مقادیر بزرگ m هم معتبرند. می‌توان دید که اگر δ کوچک باشد، (۲۸.۹.۶) یک دستگاه سرسخت خواهد بود.

با استفاده از (۳۱.۹.۶) و (۴۹.۸.۶) برای تحلیل پایداری (۲۹.۹.۶)، داریم

$$|1 + h\lambda_j| < 1 \qquad j = 1, \ldots, m-1$$

با به‌کار بردن (۳۰.۹.۶)، این رابطه ما را به حکم هم‌ارز زیر می‌رساند

$$0 < \frac{4h}{\delta^2} \sin^2\left(\frac{j\pi}{2m}\right) < 2 \qquad 1 \le j \le m-1$$

این رابطه وقتی برقرار می‌شود که $۲ \leq ۴\frac{h}{\delta^۲}$ یا

$$h \leq \frac{۱}{۲}\delta^۲ \qquad (۳۲.۹.۶)$$

اگر δ کوچک، مثلاً $\delta = ۰٫۰۱$، باشد، آنگاه مرحلهٔ زمانی h باید خیلی کوچک انتخاب شود تا پایداری به‌دست آید.

برخلاف این محدودیت (۳۲.۹.۶) با روش اویلر، روش پسرو اویلر هیچ محدودیتی ندارد زیرا A ـ پایدار است. این روش چنین می‌شود

$$\mathbf{V}_{n+۱} = \mathbf{V}_n + h[\Lambda \mathbf{V}_{n+۱} + g(t_{n+۱})] \qquad \mathbf{V}_۰ = \mathbf{u}_۰ \quad (۳۳.۹.۶)$$

برای حل این دستگاه خطی برحسب $\mathbf{V}_{n+۱}$ داریم،

$$(\mathbf{I} - h\Lambda)\mathbf{V}_{n+۱} = \mathbf{V}_n + h\mathbf{g}(t_{n+۱}) \qquad (۳۴.۹.۶)$$

این یک دستگاه سه قطری معادلات خطی است (بخش ۳.۸ را ببینید). این دستگاه را می‌توان خیلی سریع حل کرد، با تقریباً $۵m$ عمل حساب در هر مرحلهٔ زمانی، صرفنظر از بهای محاسبهٔ طرف راستِ (۳۴.۸.۶). هزینهٔ حل روش (۲۹.۹.۶) اویلر تقریباً به همین اندازه است، و بنابراین حل (۳۴.۹.۶)، زیاد وقت‌گیر نیست.

مثال مسألهٔ معادلهٔ دیفرانسیل جزئی (۱۹.۹.۶) ـ (۲۱.۹.۶) را با توابع G، $d_۰$، $d_۱$ و f که از جواب معلوم مسأله

$$U = e^{-۰٫۱t}\sin(\pi x) \qquad ۰ \leq x \leq ۱ \quad t \geq ۰ \qquad (۳۵.۹.۶)$$

به‌دست می‌آیند، حل کنید. نتایج روش (۲۹.۹.۶) اویلر در جدول ۲۰.۶ و نتایج روش (۳۳.۹.۶) پسرو اویلر در جدول ۲۱.۶ داده شده‌اند.

برای روش اویلر، m را برابر ۴، ۸، ۱۶ می‌گیریم و برای حفظ پایداری، h را با توجّه به (۳۲.۹.۶) مساوی $\delta^۲/۲$ اختیار می‌کنیم. باید توجه کرد که این انتخاب به ترتیب به طول مرحله‌های زمانی $h \doteq ۰٫۰۳۱$، $h \doteq ۰٫۰۰۷۸$ و $h \doteq ۰٫۰۰۲۰$ می‌انجامد. از (۲۶.۹.۶) و فرمول خطا برای روش اویلر، انتظار داریم که خطا متناسب با $\delta^۲$ باشد، زیرا $h = \delta^۲/۲$. این امر ایجاب می‌کند که وقتی m دو برابر می‌شود، خطا بر ۴ تقسیم شود، و نتایج جدول ۲۰.۶، این امر را تأیید می‌کند. در این جدول، ستون خطا معرّف خطای ماکسیمم در نقاط گرهی (x_j, t)، $۰ \leq j \leq n$، برای زمان داده‌شدهٔ t است.

جدول ۲۰.۶ روش خطوط: روش اویلر

t	خطا $m = ۴$	نسبت	خطا $m = ۸$	نسبت	خطا $m = ۱۶$
۱٫۰	۳٫۸۹E − ۲	۴٫۰۹	۹٫۵۲E − ۳	۴٫۰۲	۲٫۳۷E − ۳
۲٫۰	۳٫۱۹E − ۲	۴٫۰۹	۷٫۷۹E − ۳	۴٫۰۲	۱٫۹۴E − ۳
۳٫۰	۲٫۶۱E − ۲	۴٫۰۹	۶٫۳۸E − ۳	۴٫۰۱	۱٫۵۹E − ۳
۴٫۰	۲٫۱۴E − ۲	۴٫۱۰	۵٫۲۲E − ۳	۴٫۰۲	۱٫۳۰E − ۳
۵٫۰	۱٫۷۵E − ۲	۴٫۰۹	۴٫۲۸E − ۳	۴٫۰۴	۱٫۰۶E − ۳

جدول ۲۱.۶ روش خطوط: روش پسرو اویلر

t	خطا $m = ۴$	خطا $m = ۸$	خطا $m = ۱۶$
۱٫۰	۴٫۴۵E − ۲	۱٫۱۰E − ۲	۲٫۸۶E − ۳
۲٫۰	۳٫۶۵E − ۲	۹٫۰۱E − ۳	۲٫۳۴E − ۳
۳٫۰	۲٫۹۹E − ۲	۷٫۳۷E − ۳	۱٫۹۲E − ۳
۴٫۰	۲٫۴۵E − ۲	۶٫۰۴E − ۳	۱٫۵۷E − ۳
۵٫۰	۲٫۰۰E − ۲	۴٫۹۴E − ۳	۱٫۲۹E − ۳

برای حل (۲۸.۹.۶) با روش پسرو اویلر، دیگر هیچ نیازی به ارتباط بین مرحلهٔ مکانی δ و مرحلهٔ زمانی h، نخواهد بود. با بررسی فرمول خطای (۲۶.۹.۶) در روش خطوط و فرمول خطای برشی (۷.۹.۶) ($p = ۱$ می‌گیریم) برای روش پسرو اویلر، ملاحظه می‌کنیم که خطا در حل مسألهٔ (۱۹.۹.۶) ـ (۲۱.۹.۶) متناسب با $h + \delta^۲$ است. برای تابع مجهول U ی (۳۴.۹.۶)، یک تغییر آهسته با تغییر t وجود دارد. بنابراین برای خطای برشی مربوط به انتگرالگیری زمانی، باید بتوانیم مرحلهٔ زمانی نسبتاً بزرگی برای h در مقایسه با مرحلهٔ مکانی δ، انتخاب نماییم تا دو منبع خطا از نظر اندازه نسبتاً مساوی باشند. در جدول ۲۱.۶، از $h = ۰٫۱$ و $m = ۴, ۸, ۱۶$ استفاده کرده‌ایم. توجه نمایید که این مرحلهٔ زمانی بسیار بزرگتر از آن است که در جدول ۲۰.۶ برای روش اویلر به‌کار برده شده است و بنابراین روش پسرو اویلر، برای این مسألهٔ خاص بسیار کاراتر است.

برای بحث بیشتر دربارهٔ روش خطوط، ایکن (۱۹۸۵، صص ۱۲۴ـ۱۴۸) را ببینید. برای بعضی برنامه‌های روش خطوط در حل دستگاه معادلات دیفرانسیل جزئی سهموی غیرخطی، با یک و دو متغیر مکانی، سینکاوک و مَدسِن[1] (۱۹۷۵) و ملگارت[2] و سینکاوک (۱۹۸۱) را ببینید.)

1. Sincovec & Madsen 2. Malgaard

۱۰.۶ روشهای تک گامی و روشهای رونگه ـ کوتا

روشهای تک‌گامی برای حل $y' = f(x, y)$ فقط نیاز به دانستن جواب عددی y_n دارند تا مقدار بعدی y_{n+1} را محاسبه نمایند. این امر برای روشهای تک‌گامی مزایای روشنی نسبت به روشهای $p-$ گامی که از چندین مقدار قبلی $\{y_n, \ldots, y_{n-p+1}\}$ استفاده می‌کنند به‌وجود می‌آورد، زیرا مقادیر اولیهٔ $\{y_1, \ldots, y_{p-1}\}$ در روشهای چندگامی باید با روش عددی دیگری محاسبه شوند.

یکی از معروفترین روشهای تک‌گامی، روشهای رونگه ـ کوتاست. برنامه‌نویسی آنها نسبتاً آسان و کنترل خطای برشی آنها در مقایسه با روشهای چند گامی، بسیار ساده است. برای روشهای چندگامی با مرتبهٔ ثابت، که غالباً در گذشته به‌کار می‌رفت، روشهای رونگه ـ کوتا وسیلهٔ عادی برای محاسبهٔ مقادیر اولیهٔ موردنیاز این روشهای چند گامی، بودند. بزرگترین عیب روشهای رونگه ـ کوتا آن است که، در مقایسه با روشهای چندگامی، برای رسیدن به یک دقت مفروض، به محاسبهٔ خیلی بیشتر $f(x, y)$ نیاز دارد. بعداً بعضی نتایج مقایسهٔ برنامه‌های متغیر ـ مرتبهٔ آدامز و برنامه‌های ثابت ـ مرتبهٔ رونگه ـ کوتا را یادآوری می‌نماییم.

ساده‌ترین روش تک‌گامی بر پایهٔ استفاده از سری تیلر مبتنی است. فرض کنید $Y(x)$، $(r+1)$ بار پیوسته مشتقپذیر باشد که $Y(x)$ جواب مسألهٔ مقدار اولیهٔ

$$y' = f(x, y) \qquad y(x_0) = Y_0 \qquad (۱.۱۰.۶)$$

است. با استفاده از قضیهٔ تیلر، $Y(x_1)$ را حول x_0 بسط می‌دهیم:

$$Y(x_1) = Y(x_0) + hY'(x_0) + \ldots + \frac{h^r}{r!}Y^{(r)}(x_0) + \frac{h^{r+1}}{(r+1)!}Y^{(r+1)}(\xi) \quad (۲.۱۰.۶)$$

به‌ازای مقداری چون $x_0 \leq \xi \leq x_1$. با حذف جملهٔ باقیمانده، یک تقریب برای $Y(x_1)$ خواهیم داشت به شرطی که بتوانیم $Y''(x_0)$، ...، $Y^{(r)}(x_0)$ را حساب کنیم. از $Y'(x) = f(x, Y(x))$ مشتق می‌گیریم تا به‌دست آوریم

$$Y''(x) = f_x(x, Y(x)) + f_y(x, Y(x))Y'(x),$$
$$Y''(x) = f_x + f_y f$$

به همین ترتیب ادامه می‌دهیم تا مشتقات مراتب بالاتر $Y(x)$ را به‌دست آوریم.

جدول ۲۲.۶ مثال روش (۳.۱۰.۶) سری تیلر

		$h = ۰٫۰۶۲۵$	$h = ۰٫۱۲۵$	
x	$y_h(x)$	$Y(x) - y_h(x)$	$Y(x) - y_h(x)$	نسبت
۲٫۰	۰٫۳۳۳۶۴۹	−۳٫۲E − ۴	−۱٫۴E − ۳	۴٫۴
۴٫۰	۰٫۲۰۰۱۳۵	−۱٫۴E − ۴	−۵٫۹E − ۴	۴٫۳
۶٫۰	۰٫۱۴۲۹۳۱	−۷٫۴E − ۵	−۳٫۲E − ۴	۴٫۳
۸٫۰	۰٫۱۱۱۱۵۷	−۴٫۶E − ۵	−۲٫۰E − ۴	۴٫۳
۱۰٫۰	۰٫۰۹۰۹۴۱	−۳٫۱E − ۵	−۱٫۴E − ۴	۴٫۳

مثال مسألهٔ زیر را با جواب $Y(x) = ۱/(۱ + x)$ در نظر می‌گیریم،

$$y' = -y^۲ \qquad y(۰) = ۱$$

پس $Y'' = -۲YY' = ۲Y^۳$ و (۲.۱۰.۶) با $r = ۱$ به‌دست می‌دهد

$$Y(x_۱) = Y_۰ - hY_۰^۲ + h^۲Y_۰^۳ + \frac{h^۳}{۶}Y^{(۳)}(\xi_۰) \qquad x_۰ \le \xi_۰ \le x_۱$$

جملهٔ باقیمانده را حذف کرده‌ایم تا تقریبی برای $Y(x_۱)$ به‌دست آوریم. به همین ترتیب، از این مقدار می‌توان برای محاسبهٔ تقریب $Y(x_۲)$ و غیره نیز استفاده کرد. روش عددی چنین است:

$$y_{n+۱} = y_n - hy_n^۲ + h^۲y_n^۳ \qquad n \ge ۰ \qquad (۳.۱۰.۶)$$

جدول ۲۲.۶ شامل خطاهای این جواب عددی در یک مجموعهٔ انتخابی از نقاط گرهی است. برای فاصلهٔ نقاط در هر شبکه، $h = ۰٫۱۲۵$ و $h = ۰٫۰۶۲۵$ به‌کار برده شده است و نسبت خطاهای به‌دست آمده نیز داده شده است. توجه نمایید وقتی h نصف می‌شود، نسبت تقریباً ۴ می‌شود. این را می‌توان به‌طور نظری توجیه کرد زیرا می‌توان نشان داد که نرخ همگرایی $O(h^۲)$ است، برهان مشابه همان برهانی است که در قضیهٔ ۳.۶ داده شده است یا در قضیهٔ ۹.۶ بعداً در این بخش داده خواهد شد.

روش سری تیلر می‌تواند نتایج عالی به‌دست دهد. ولی استفاده از آن، به‌دلیل مشتقگیری تحلیلی $f(x, y)$، پردردسر است. محاسبهٔ مشتقات ممکن است خیلی مشکل و محاسبهٔ آنها به‌ویژه در دستگاه معادلات وقتگیر باشد. این مشتقگیریها را می‌توان با استفاده از زبان عملیات نمادی با رایانه انجام داد، سپس به سادگی روش سری تیلر را تولید کرد. مع‌هذا، احتمالاً باز هم محاسبهٔ مشتقات وقتگیر است، و به نظر می‌آید که روشهایی که بر پایهٔ محاسبهٔ فقط $f(x, y)$ هستند کاراتر باشند. برای پیروی از روش سری تیلر، وقتی فقط $f(x, y)$ محاسبه می‌شود، به فرمولهای رونگه ـ کوتا برمی‌گردیم.

روشهای رونگه ـ کوتا روشهای رونگه ـ کوتا دقیقاً به بسط سری تیلر $Y(x)$ در (۲.۱۰.۶) وابسته‌اند ولی مشتقگیری f در استفاده از این روشها هرگز لازم نیست. برای آسانی نمادگذاری، روش رونگه ـ کوتا را مختصراً با RK نمایش می‌دهیم. تمام روشهای RK بدین شکل نوشته می‌شوند:

$$y_{n+1} = y_n + hF(x_n, y_n, h; f) \qquad n \geq 0 \qquad (۴.۱۰.۶)$$

با مثالهایی از تابع F آغاز می‌کنیم، و بعداً فرضهای مربوط به آن را مورد بحث قرار می‌دهیم. ولی در اینجا، باید متوجه باشیم برای تمام مقادیر کوچک h می‌خواهیم داشته باشیم

$$F(x, Y(x), h; f) \doteq Y'(x) = f(x, Y(x)) \qquad (۵.۱۰.۶)$$

خطای برشی برای (۴.۱۰.۶) را با

$$T_n(Y) = Y(x_{n+1}) - Y(x_n) - hF(x_n, Y(x_n), h; f) \qquad n \geq 0 \qquad (۶.۱۰.۶)$$

تعریف و $\tau_n(Y)$ را به طور ضمنی با رابطهٔ زیر تعریف می‌کنیم

$$T_n(Y) = h\tau_n(Y)$$

رابطهٔ (۶.۱۰.۶) را دوباره مرتّب می‌کنیم و به‌دست می‌آوریم

$$Y(x_{n+1}) = Y(x_n) + hF(x_n, Y(x_n), h; f) + h\tau_n(Y) \qquad n \geq 0 \qquad (۷.۱۰.۶)$$

در قضیهٔ ۹.۶، این رابطه با (۴.۱۰.۶) مقایسه می‌شود تا همگرایی $\{y_n\}$ به Y ثابت شود.

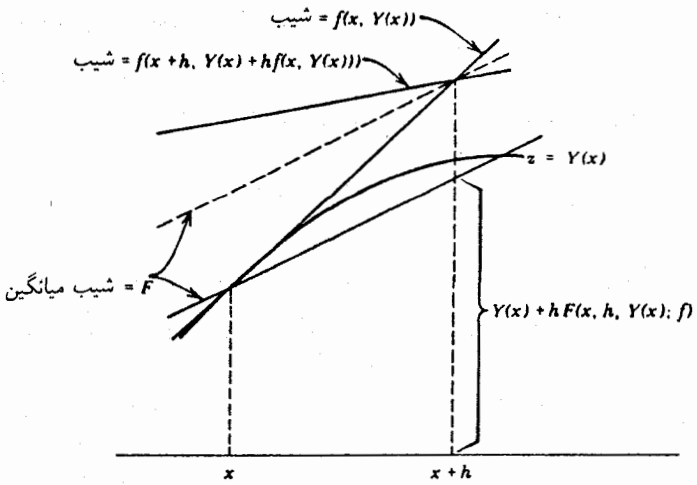
مثال ۱. روش ذوزنقه‌یی را در نظر می‌گیریم و آن را با یک بارست با استفاده از روش اویلر به‌عنوان پیشگو حل می‌کنیم.

$$y_{n+1} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_n + hf(x_n, y_n))] \qquad n \geq 0 \qquad (۸.۱۰.۶)$$

در نمادگذاری (۴.۱۰.۶)،

$$F(x, y, h; f) = \frac{1}{2}[f(x, y) + f(x + h, y + hf(x, y))]$$

همان‌طورکه در شکل ۹.۶ می‌توان دید، F یک **شیب میانگین** $Y(x)$ در $[x, x + h]$ است.



شکل ۹.۶ نمایش روش (۸.۱۰.۶) رونگه ـ کوتا

۲. روش زیر نیز بر پایهٔ به‌دست آوردن یک شیب میانگین برای جواب در $[x_n, x_{n+1}]$ است:

$$y_{n+1} = y_n + hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hf(x_n, y_n)) \qquad n \geq 0 \quad (۹.۱۰.۶)$$

در این حالت

$$F(x, y, h; f) = f(x + \frac{1}{2}h, y + \frac{1}{2}hf(x, y))$$

پیداکردن فرمولی برای خطای برشی به پیدا کردن این روشها مربوط است، و این مطلب برای روشهای RK از مراتب بالاتر نیز درست است. پیداکردن روشهای RK با به‌دست آوردن یک خانواده از فرمولهای مرتبهٔ دوم شامل (۸.۱۰.۶) و (۹.۱۰.۶) روشن می‌شود. فرض می‌کنیم که F شکل کلی زیر را دارد

$$F(x, y, h; f) = \gamma_1 f(x, y) + \gamma_2 f(x + \alpha h, y + \beta h f(x, y)) \quad (۱۰.۱۰.۶)$$

که در آن چهار ثابت γ_1، γ_2، α و β باید معین شوند.

از قضیهٔ تیلر (قضیهٔ ۵.۱) برای توابع دو متغیره استفاده کرده عبارت دوّم طرف راست (۱۰.۱۰.۶) را تا جملات شامل مشتقات مرتبهٔ دوّم بسط می‌دهیم. این بسط به ما خواهد داد

$$F(x, y, h; f) = \gamma_1 f(x, y) + \gamma_2 \{f(x, y) + h[\alpha f_x + \beta f f_y] + h^2[\frac{1}{2}\alpha^2 f_{xx} + \alpha\beta f_{xy} f + \frac{1}{2}\beta^2 f^2 f_{yy}]\} + O(h^3) \quad (۱۱.۱۰.۶)$$

همچنین به بعضی از مشتقات $Y'(x) = f(x, Y(x))$ نیاز داریم، یعنی به

$$Y'' = f_x + f_y f$$
$$Y^{(3)} = f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_y f_x + f_y^2 f \qquad (۱۲.۱۰.۶)$$

برای خطای برشی،

$$T_n(Y) = Y(x_{n+1}) - Y(x_n) - hF(x_n, Y(x_n), h; f)$$
$$= hY_n' + \frac{h^2}{2}Y_n'' + \frac{h^3}{6}Y_n^{(3)} + O(h^4) - hF(x_n, Y_n, h; f)$$

اگر مقادیر (۱۱.۱۰.۶) و (۱۲.۱۰.۶) را در این عبارت قرار دهیم و برحسب توانهای h مرتّب کنیم به‌دست می‌آوریم

$$T_n(Y) = h[1 - \gamma_1 - \gamma_2]f + h^2[(\frac{1}{2} - \gamma_2\alpha)f_x + (\frac{1}{2} - \gamma_2\beta)f_y f]$$
$$+ h^3[(\frac{1}{6} - \frac{1}{2}\gamma_2\alpha^2)f_{xx} + (\frac{1}{3} - \gamma_2\alpha\beta)f_{xy}f + (\frac{1}{6} - \frac{1}{2}\gamma_2\beta^2)f_{yy}f^2$$
$$+ \frac{1}{6}f_y f_x + \frac{1}{6}f_y^2 f] + O(h^4) \qquad (۱۳.۱۰.۶)$$

تمام مشتقات در (x_n, Y_n) محاسبه شده‌اند.

مایلیم که خطای برشی در حد ممکن، با سرعت بیشتری به صفر همگرا شود. اگر f مجاز باشد به طور دلخواه تغییر نماید، ضریب h^3 نمی‌تواند در حالت کلّی صفر شود. لازمهٔ اینکه ضرایب h و h^2 صفر شوند چنین خواهد شد

$$\gamma_1 + \gamma_2 = 1 \qquad \gamma_2\alpha = \frac{1}{2} \qquad \gamma_2\beta = \frac{1}{2} \qquad (۱۴.۱۰.۶)$$

و از آنجا نتیجه می‌شود

$$T_n(Y) = O(h^3)$$

دستگاه (۱۴.۱۰.۶) نامعین است و جواب عمومی آن

$$\gamma_1 = 1 - \gamma_2 \qquad \alpha = \beta = \frac{1}{2\gamma_2} \qquad (۱۵.۱۰.۶)$$

است، که γ_2 دلخواه است. هر دو حالت (۸.۱۰.۶) (با $\gamma_2 = 1/2$) و (۹.۱۰.۶) (با $\gamma_2 = 1$) حالتهای خاص این جواب‌اند.

با گذاردن این مقادیر در (۱۳.۱۰.۶)، می‌توانیم جملهٔ اصلی خطای برشی را که به γ_2 بستگی دارد پیدا کنیم. در بعضی حالات، در حالی‌که f می‌تواند به طور دلخواه تغییر کند، مقدار γ_2 طوری انتخاب می‌شود که ضریب h^3 در حد ممکن کوچک باشد. برای مثال، اگر (۱۳.۱۰.۶) را به شکل

$$T_n(Y) = c(f,\gamma_2)h^3 + O(h^4) \qquad (۱۶.۱۰.۶)$$

بنویسیم، نابرابری کوشی ـ شوارتس [(۸.۱.۷) را در فصل ۷ ببینید] را می‌توان به‌کار برده و نشان داد

$$| c(f,\gamma_2) | \leq c_1(f)c_2(\gamma_2) \qquad (۱۷.۱۰.۶)$$

که در آن

$$c_1(f) = [f_{xx}^2 + f_{xy}^2 f^2 + f_{yy}^2 f^4 + f_y^2 f_x^2 + f_y^4 f^2]^{1/2}$$

$$c_2(\gamma_2) = \left[(\frac{1}{6} - \frac{1}{2}\gamma_2\alpha^2)^2 + (\frac{1}{3} - \gamma_2\alpha\beta)^2 + (\frac{1}{6} - \frac{1}{2}\gamma_2\beta^2)^2 + \frac{1}{18}\right]^{1/2}$$

که α و β در (۱۵.۱۰.۶) داده شده‌اند. مینیمم مقدار $c_2(\gamma_2)$ به‌ازای $\gamma_2 = ۰٫۷۵$ حاصل می‌شود و $c_2(۰٫۷۵) = ۱/\sqrt{۱۸}$. روش عددی مرتبهٔ دوّم به‌دست آمده چنین خواهد شد

$$y_{n+1} = y_n + \frac{h}{4}\left[f(x_n,y_n) + 3f(x_n + \frac{2}{3}h, y_n + \frac{2}{3}hf(x_n,y_n))\right] \qquad n \geq 0 \qquad (۱۸.۱۰.۶)$$

این روش بهینه است به این معنی که $c_2(\gamma_2)$، ضریب جملهٔ $c_1(f)h^3$ در خطای برشی را مینیمم می‌نماید. برای بحث مفصل از راه تحلیل خطای برشی در روشهای RK، شمپاین (۱۹۸۶) را ببینید.

فرمولهای مراتب بالاتر را می‌توان ساخت و به روشی مشابه تحلیل نمود، اگرچه محاسبات جبری بسیار پیچیده خواهند شد. فرض کنید فرمولی برای $F(x,y,h;f)$ به شکل زیر داده شده است

$$F(x,y,h;f) = \sum_{j=1}^{p} \gamma_j V_j \qquad (۱۹.۱۰.۶)$$

$$V_1 = f(x,y)$$

$$V_j = f\left(x + \alpha_j h, y + h\sum_{i=1}^{j-1}\beta_{ji}V_i\right) \qquad j = 2,\ldots,p \qquad (۲۰.۱۰.۶)$$

جدول ۲۳.۶ مرتبهٔ ماکسیمم روشهای رونگه ـ کوتا

تعداد ارزیابی‌های تابع	۱	۲	۳	۴	۵	۶	۷	۸
ماکسیمم مرتبهٔ روش	۱	۲	۳	۴	۴	۵	۶	۶

این ضرایب را می‌توان به‌گونه‌ای انتخاب نمود که جملات پیشرو در خطای برشی را صفر نمایند، همان‌گونه که برای (۱۰.۱۰.۶) و (۱۴.۱۰.۶) انجام شد. واضح است که بین تعداد محاسبات $f(x,y)$ که آن را p می‌نامیم، و مرتبهٔ ماکسیمم ممکن که می‌توان برای خطای برشی به‌دست آورد، ارتباطی وجود دارد. این ارتباط در جدول ۲۳.۶ داده شده است که بخشی از آن منسوب به بوچر[1] (۱۹۶۵) است.

تا تقریباً ۱۹۷۰، متداولترین روش RK، احتمالاً فرمول کلاسیک اصلی بوده، که یک تعمیم روش سیمپسون است. این روش چنین است

$$y_{n+1} = y_n + \frac{h}{6}[V_1 + 2V_2 + 2V_3 + V_4] \qquad (21.10.6)$$

$$V_1 = f(x_n, y_n) \qquad V_2 = f\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hV_1\right)$$

$$V_3 = f\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hV_2\right) \qquad V_4 = f(x_n + h, y_n + hV_3)$$

می‌توان ثابت کرد که این فرمول از مرتبهٔ چهار با $T_n(Y) = O(h^5)$ است. اگر $f(x,y)$ به y بستگی نداشته باشد، این فرمول به قاعدهٔ انتگرالگیری سیمپسون بدل می‌شود.

مثال مسألهٔ

$$y' = \frac{1}{1+x^2} - 2y^2 \qquad y(0) = 0 \qquad (22.10.6)$$

با جواب $Y = x/(1+x^2)$ را در نظر می‌گیریم. روش (۲۱.۱۰.۶) با طول گام ثابت به‌کار برده‌شده و نتایج در جدول ۲۴.۶ داده شده‌اند. طول گامها $h = 0.25$ و $h = 0.5$ بوده‌اند. ستون نسبت، نسبت خطا در نقاط گرهی مربوطه را، وقتی h نصف‌شده، نشان می‌دهد. ستون آخر مثالی است از فرمول (۲۴.۱۰.۶) از مطالب حاصل از برونیابی ریچاردسن. چون برای روش (۲۱.۱۰.۶) تساوی $T_n(Y) = O(h^5)$ برقرار است، قضیهٔ ۹.۶ ایجاب می‌کند که نرخ همگرایی $y_h(x)$ به $Y(x)$، $O(h^4)$ باشد. مقدار نظری نسبت، ۱۶ است و وقتی h باز هم کاهش یابد، این نسبت به ۱۶ نزدیکتر می‌شود.

1. Butcher

جدول ۲۴.۶ مثالی برای روش (۲۱.۱۰.۶) رونگه ـ کوتا

x	$y_h(x)$	$Y(x)-y_h(x)$	$Y(x)-y_{2h}(x)$	نسبت	$\frac{1}{15}[y_h(x)-y_{2h}(x)]$
۲٫۰	۰٫۳۹۹۹۵۶۹۹	۴٫۳E − ۵	۱٫۰E − ۳	۲۴	۶٫۷E − ۵
۴٫۰	۰٫۲۳۵۲۹۱۵۹	۲٫۵E − ۶	۷٫۰E − ۵	۲۸	۴٫۵E − ۶
۶٫۰	۰٫۱۶۲۱۶۱۷۹	۳٫۷E − ۷	۱٫۲E − ۵	۳۲	۷٫۷E − ۷
۸٫۰	۰٫۱۲۳۰۷۶۸۳	۹٫۲E − ۸	۳٫۴E − ۶	۳۶	۲٫۲E − ۷
۱۰٫۰	۰٫۰۹۹۰۰۹۸۷	۳٫۱E − ۸	۱٫۳E − ۶	۴۱	۸٫۲E − ۸

روشهای رونگه ـ کوتا، همانند روشهای چندگامی، دارای فرمولهای خطاهای مجانبی هستند. برای (۲۱.۱۰.۶)،

$$Y(x)-y_h(x)=D(x)h^4+O(h^5) \qquad (۲۳.۱۰.۶)$$

که $D(x)$ در مسألهٔ مقدار اولیهٔ معینی صدق می‌کند. برهان این قضیّه، تعمیم برهان قضیهٔ ۹.۶ و شبیه است به اشتقاق قضیهٔ ۴.۶ برای روش اویلر، در بخش ۲.۶. نتیجهٔ (۲۳.۱۰.۶) را می‌توان برای به‌دست آوردن یک برآورد خطا به‌کاربرد درست همان‌گونه که در روش ذوزونقه‌یی در فرمول (۲۶.۵.۶) از بخش ۵.۶ عمل شده بود. برای یک طول گام $2h$،

$$Y(x)-y_{2h}(x)=16D(x)h^4+O(h^5)$$

اگر مانند عمل قبلی برای (۲۷.۵.۶) عمل کنیم به‌دست می‌آوریم

$$Y(x)-y_h(x)=\frac{1}{15}[y_h(x)-y_{2h}(x)]+O(h^5) \qquad (۲۴.۱۰.۶)$$

و اولین جملهٔ سمت راست برآورد برای خطای سمت چپ است. این برآورد در آخرین ستون جدول ۲۴.۶ نشان داده شده است.

تحلیل همگرایی برای به‌دست آوردن همگرایی کلّی (۴.۱۰.۶)، لازم است $\tau_n(Y)\to 0$ وقتی $h\to 0$. چون

$$\tau_n(Y)=\frac{Y(x_{n+1})-Y(x_n)}{h}-F(x_n,Y(x_n),h;f)$$

می‌خواهیم

$$F(x,Y(x),h;f)\to Y'(x)=f(x,Y(x))$$

وقتی $h\to 0$. به عبارت دقیق‌تر، تعریف می‌کنیم

$$\delta(h) = \underset{\substack{x_0 \le x \le b \\ -\infty < y < \infty}}{\text{Max}} \mid f(x,y) - F(x,y,h;f) \mid \qquad (۲۵.۱۰.۶)$$

و فرض می‌کنیم

$$\delta(h) \to ۰ \quad \text{وقتی} \quad h \to ۰ \qquad (۲۶.۱۰.۶)$$

این شرط را گاهی اوقات **شرط سازگاری** روش (۴.۱۰.۶) رونگه ـ کوتا خوانند.

ما همچنین به یک شرط لیپشیتس برای F نیاز داریم: به‌ازای تمام مقادیر $x_0 \le x \le b$ و $-\infty < y, z < \infty$، و برای جمیع مقادیر کوچک $h > ۰$،

$$\mid F(x,y,h;f) - F(x,z,h;f) \mid \le L \mid y - z \mid \qquad (۲۷.۱۰.۶)$$

این شرط معمولاً با استفاده از شرط (۱۲.۲.۶) لیپشیتس برای $f(x,y)$، ثابت می‌شود. برای مثال، با روش (۹.۱۰.۶)،

$$\begin{aligned} &\mid F(x,y,h;f) - f(x,z,h;f) \mid \\ &\quad = \left| f\left(x + \frac{h}{۲}, y + \frac{h}{۲} f(x,y)\right) - f\left(x + \frac{h}{۲}, z + \frac{h}{۲} f(x,z)\right) \right| \\ &\quad \le K \left| y - z + \frac{h}{۲}[f(x,y) - f(x,z)] \right| \\ &\quad \le K \left(۱ + \frac{h}{۲} K\right) \mid y - z \mid \end{aligned}$$

L را برابر $K(۱ + \frac{۱}{۲}K)$ برای $h \le ۱$ انتخاب می‌کنیم.

قضیهٔ ۹.۶ فرض می‌کنیم روش (۴.۱۰.۶) رونگه ـ کوتا در شرط (۲۷.۱۰.۶) لیپشیتس صدق می‌نماید. در این صورت برای مسألهٔ مقدارآغازی (۱.۱۰.۶)، جواب $\{y_n\}$ در رابطهٔ زیر صدق می‌کند

$$\underset{x_0 \le x_n \le b}{\text{Max}} \mid Y(x_n) - y_n \mid \le e^{(b-x_0)L} \mid Y_0 - y_0 \mid + \left[\frac{e^{(b-x_0)L} - ۱}{L}\right] \tau(h) \qquad (۲۸.۱۰.۶)$$

که در آن

$$\tau(h) \equiv \underset{x_0 \le x_n \le b}{\text{Max}} \mid \tau_n(Y) \mid \qquad (۲۹.۱۰.۶)$$

اگر شرط سازگاری (۲۶.۱۰.۶) نیز برقرار باشد، جواب عددی $\{y_n\}$ به $Y(x)$ همگراست.

برهان رابطهٔ (۴.۱۰.۶) را از (۷.۱۰.۶) کم می‌کنیم تا به‌دست آید

$$e_{n+1} = e_n + h[F(x_n, Y_n, h; f) - F(x_n, y_n, h; f)] + h\tau_n(Y) \qquad (۳۰.۱۰.۶)$$

که در آن $e_n = Y(x_n) - y_n$. شرط لیپشیتس (۲۷.۱۰.۶) را به‌کار می‌بریم و از (۲۹.۱۰.۶) استفاده می‌کنیم تا رابطهٔ زیر پیدا شود

$$| e_{n+1} | \leq (1 + hL) | e_n | + h\tau_n(h) \qquad x_0 \leq x_N \leq b$$

با برهانی همانند برهان همگرایی برای روش اویلر، از رابطهٔ فوق به‌آسانی به نتیجهٔ (۲۸.۱۰.۶) می‌رسیم. [(۲۰.۲.۶) و (۲۲.۲.۶) را ببینید].

در بیشتر حالات، با محاسبهٔ مستقیم معلوم می‌شود که وقتی $h \to 0$، $\tau(h) \to 0$، و در چنین حالتی، همگرایی $\{y_n\}$ به $Y(x)$ فوری ثابت می‌شود. ولی آنچه که لازم است بدانیم برقرار بودن (۲۶.۱۰.۶) است. برای ملاحظهٔ این امر، می‌نویسیم

$$\begin{aligned} h\tau_n(Y) &= Y(x_{n+1}) - Y(x_n) - hF(x_n, Y(x_n), h; f) \\ &= hY'(x_n) + \frac{h^2}{2}Y''(\xi_n) - hF(x_n, Y(x_n), h; f) \\ h \mid \tau_n(Y) \mid &\leq h\delta(h) + \frac{h^2}{2} \parallel Y'' \parallel_\infty \\ \tau(h) &\leq \delta(h) + \frac{1}{2}h \parallel Y'' \parallel_\infty \end{aligned}$$

بنابراین وقتی $h \to 0$، $\tau(h) \to 0$ که برهان را کامل می‌کند. ■

فرع بعدی یک نتیجهٔ بلافصل (۲۸.۱۰.۶) است.

فرع اگر روش (۴.۱۰.۶) رونگه ـ کوتا دارای خطای برشی $T_n(Y) = O(h^{m+1})$ باشد، آنگاه نرخ همگرایی $\{y_n\}$ به $Y(x)$ برابر $O(h^m)$ است. ■

به‌دست آوردن یک فرمول خطای مجانبی برای روش (۴.۱۰.۶) رونگه ـ کوتا دشوار نیست مشروط بر آنکه یک چنین فرمولی برای خطای برشی معلوم باشد. فرض می‌کنیم

$$T_n(Y) = \varphi(x_n)h^{m+1} + O(h^{m+2}) \qquad (۳۱.۱۰.۶)$$

که $\varphi(x)$ از روی $Y(x)$ و $f(x,Y(x))$ معلوم شده است. به‌عنوان یک مثال، نتیجهٔ (۱۳.۱۰.۶) را در به‌دست آوردن این بسط برای روشهای مرتبهٔ دوم RK ببینید. شکلهای قویتر (۲۶.۱۰.۶) و (۲۷.۱۰.۶) نیز لازم‌اند. فرض می‌کنیم

$$F(x,y,h;f)-F(x,z,h;f)=\frac{\partial F(x,y,h;f)}{\partial y}(y-z)+O((y-z)^{۲}) \quad (۳۲.۱۰.۶)$$

و همچنین فرض می‌کنیم

$$\delta_{۱}(h)\equiv \underset{\substack{x_{۰}\le x\le b\\ -\infty<y<\infty}}{\mathrm{Max}} \left|\frac{\partial f(x,y)}{\partial y}-\frac{\partial F(x,y,h;f)}{\partial y}\right| \to ۰ \qquad h\to ۰ \quad \text{وقتی} \quad (۳۳.۱۰.۶)$$

در عمل، هر دو این نتایج به‌سادگی اثبات می‌شوند. با این مفروضات، می‌توانیم فرمول زیر را به‌دست آوریم

$$Y(x)-y_h(x)=D(x)h^m+O(h^{m+۱}) \qquad (۳۴.۱۰.۶)$$

که $D(x)$ در مسألهٔ مقدار آغازی زیر صدق می‌کند

$$D'=f_y(x,Y(x))D(x)+\varphi(x) \qquad D(x_{۰})=۰ \quad (۳۵.۱۰.۶)$$

قضایای پایداری را، مشابه با قضایایی که برای روشهای چندگامی بیان شد، می‌توان برای روشهای رونگه ـ کوتا بیان کرد. نوع اساسی پایداری، که تقریباً در ابتدای بخش ۸.۶ تعریف شده، به‌سادگی ثابت می‌شود. برهان، یک تغییر سادهٔ برهان پایداری روش (۲۹.۲.۶) اویلر است، و ما آن را به مسألهٔ ۴۹ واگذار می‌کنیم. یک تفاوت اساسی با نظریهٔ چندگامی این است که با روشهای RK هیچ جواب انگلی ایجاد نمی‌شود؛ بنابراین مفهوم پایداری نسبی برای روشهای RK کاربرد ندارد. ناحیه‌های پایداری مطلق را مانند نظریهٔ چندگامی می‌توان مطالعه کرد، ولی ما آن را در اینجا حذف و در ضمن مسائل مطرح می‌کنیم.

برآورد خطای برشی برای کنترل اندازهٔ خطای برشی، ابتدا باید بتوانیم آن را برآورد کنیم. گیریم $u_n(x)$ معرّف جواب $y'=f(x,y)$ گذرنده بر (x_n,y_n) باشد [(۹.۵.۶) را ببینید]. می‌خواهیم خطا در $y_h(x_n+۲h)$ را نسبت به $u_n(x_n+۲h)$ برآورد نماییم.

با به‌کار بردن (۳۰.۱۰.۶) برای خطای $y_h(x)$ در مقایسه با $u_n(x)$، و استفاده از فرمول مجانبی (۳۱.۱۰.۶)

$$\begin{aligned} u_n(x_{j+1}) - y_h(x_{j+1}) =& u_n(x_j) - y_h(x_j) \\ &+ h\{F(x_j, u_n(x_j), h; f) - F(x_j, y_h(x_j), h; f)\} \\ &+ \varphi_n(x_j)h^{m+1} + O(h^{m+2}) \qquad j \geq n \end{aligned}$$

با توجّه به این رابطه به سادگی می‌توان ثابت کرد که

$$u_n(x_n + h) - y_h(x_n + h) = \varphi_n(x_n)h^{m+1} + O(h^{m+2})$$
$$u_n(x_n + 2h) - y_h(x_n + 2h) = 2\varphi_n(x_n)h^{m+1} + O(h^{m+2})$$

اگر همین شیوه را برای $y_{2h}(x_j)$ به‌کار بریم،

$$u_n(x_n + 2h) - y_{2h}(x_n + 2h) = 2^{m+1}\varphi_n(x_n)h^{m+1} + O(h^{m+2})$$

از دو معادلهٔ اخیر،

$$\begin{aligned} &u_n(x_n + 2h) - y_h(x_n + 2h) \\ &= \frac{1}{2^m - 1}[y_h(x_n + 2h) - y_{2h}(x_n + 2h)] + O(h^{m+2}) \end{aligned} \qquad (۳۶.۱۰.۶)$$

و اولین جملهٔ سمت راست یک برآورد مجانبی برای خطای سمت چپ است.

محاسبهٔ $y_h(x_n + 2h)$ از روی $y_h(x_n) \equiv y_n$ را یک مرحلهٔ الگوریتم در نظر می‌گیریم. فرض می‌کنیم که کاربَر یک تحمل خطای ε انتخاب کرده است و

$$\begin{aligned} \text{trunc} &\equiv u_n(x_n + 2h) - y_h(x_n + 2h) \\ &\doteq \frac{1}{2^m - 1}[y_h(x_n + 2h) - y_{2h}(x_n + 2h)] \end{aligned} \qquad (۳۷.۱۰.۶)$$

باید در رابطهٔ زیر صدق کند

$$0.5\varepsilon h \leq |\text{ trunc }| \leq 2\varepsilon h \qquad (۳۸.۱۰.۶)$$

این فرمول، خطادرگام واحد را کنترل می‌کند و برطبق این رابطه خطا باید نه خیلی بزرگ باشد نه خیلی کوچک. مفهوم خطا در طول گام واحد را در (۱.۶.۶) بخش ۶.۶ به یاد آورید.

اگر (۳۸.۱۰.۶) برقرار باشد، محاسبات با همان h ادامه می‌یابد. ولی اگر برقرار نباشد، یک مقدار جدید $\hat{h}$ باید انتخاب شود. گیریم $\hat{h}$ چنین انتخاب شود

$$۲ \mid \varphi_n(x_n) \mid \hat{h}^{m+۱} \doteq \varepsilon h$$

که در آن $\varphi_n(x_n)$ از رابطهٔ زیر تعیین می‌شود

$$\varphi_n(x_n) \doteq \frac{\text{trunc}}{۲h^{m+۱}}$$

با این مقدار جدید $\hat{h}$، خطای برشی جدید باید نزدیک میانگاهی (۳۸.۱۰.۶) باشد. الگوریتم به این صورت با چندین روش به‌کار گرفته شده است. برای مثال گی‌یر (۱۹۷۱، صص ۸۳ـ۸۴) را برای الگوریتم مشابهی برای یک روش مرتبهٔ چهار RK ببینید.

در بسیاری از برنامه‌ها، برآورد خطای (۳۷.۱۰.۶) به‌مقادیر جاری محاسبات $y_h(x_n + ۲h)$ اضافه می‌شود و یک نتیجهٔ دقیقتری به‌دست می‌دهد. این روش را **برونیابی موضعی** نامند. در چنین صورتی، محک خطا در گام واحد جایگزین محک (۳۸.۱۰.۶) **خطا در گام** می‌شود:

$$۰٫۵\varepsilon \leq \mid \text{trunc} \mid \leq ۲\varepsilon \qquad (۳۹.۱۰.۶)$$

در چنین مواردی، می‌توان نشان داد که خطای موضعی مقادیر برونیابی‌شدهٔ $y_n(x_n + ۲h)$ در یک محک اصلاح‌شدهٔ خطا درگام واحد صدق می‌نماید. [شمپاین و گوردون (۱۹۷۵، ص ۱۰۰) را ببینید]. در اجرای این روش، به‌نظر می‌رسد برنامه‌هایی که از برونیابی موضعی و محک خطا در گام استفاده می‌کنند کاراتر از برنامه‌هایی هستند که از (۳۸.۱۰.۶) استفاده می‌کنند و از برونیابی موضعی استفاده نمی‌کنند.

برای بهتر پی‌بردن به هزینهٔ روشهای RK با برآورد خطایی که قبلاً داده شده‌است، فقط روشهای مرتبهٔ چهار RK با چهار محاسبهٔ $f(x, y)$ در هرگام RK را در نظر می‌گیریم. در رفتن از x_n به $x_n + ۲h$ برای به‌دست آوردن $y_h(x_n + ۲h)$ به هشت محاسبه نیاز است و سه محاسبهٔ اضافی برای به‌دست آوردن $y_{۲h}(x_n + ۲h)$ لازم است. بنابراین یک گام الگوریتمهای متغیّرگام به یازده محاسبهٔ f نیاز دارد. اگرچه در مقایسه با روش چندگامی، روش متغیرگام RK نسبتاً گران است، ولی بسیار پایدار و قابل‌اعتماد است، و برنامه‌نویسی رایانه‌یی آن تا حدی آسان است.

روشهای رونگه ـ کوتا ـ فِلبرک[1] روشهای رونگه ـ کوتا ـ فِلبرک روشهای RK هستند که در آنها خطای برشی از مقایسهٔ جواب محاسبه‌شدهٔ $y_{n+۱}$ با نتیجهٔ متناظری که از یک فرمول مرتبهٔ بالاتر

1. Fehlberg

RK به‌دست آمده، محاسبه می‌شود. رایجترین این روشها از فلبرک است [مثلاً، فلبرک (۱۹۷۰) را ببینید]؛ در حال حاضر این روشها رایجترین روشهای RK هستند. برای روشن‌شدن مطلب، ما فقط رایجترین زوج فرمولهای رونگه ـ کوتا ـ فلبرک (RKF) از مرتبهٔ ۴ و ۵ را درنظر می‌گیریم. این فرمولها همزمان محاسبه می‌شوند و اختلاف آنها را برآورد خطای برشی روش مرتبهٔ چهار می‌گیرند.

از جدول ۲۳.۶ دیده می‌شود که روش مرتبهٔ پنج RK به شش محاسبهٔ f در هرگام نیاز دارد. از این‌رو فلبرک در فرمول مرتبهٔ چهار، به‌جای چهار محاسبهٔ معمول پنج محاسبهٔ f را انتخاب کرد. این آزادی انتخاب ۴ یا ۵ محاسبهٔ f در فرمول مرتبهٔ چهار امکان خطای برشی کوچکتر را به‌وجود آورده، و این موضوع بعداً نشان داده شده است.

مانند قبل، تعریف می‌کنیم

$$V_1 = f(x_n, y_n),$$

$$V_j = f\left(x_n + a_j h, y_n + h\sum_{i=1}^{j-1} \beta_{ji} V_i\right) \qquad j = 2, \ldots, 6 \qquad (40.10.6)$$

فرمولهای مراتب چهار و پنج به‌ترتیب چنین‌اند

$$y_{n+1} = y_n + h\sum_{i=1}^{5} \gamma_j V_j \qquad (41.10.6)$$

$$\hat{y}_{n+1} = y_n + h\sum_{i=1}^{6} \hat{\gamma}_j V_j \qquad (42.10.6)$$

خطای برشی در y_{n+1} تقریباً برابر است با

$$\text{trunc} = \hat{y}_{n+1} - y_{n+1} = h\sum_{j=1}^{6} c_j V_j \qquad (43.10.6)$$

ضرایب در جدولهای ۲۵.۶ و ۲۶.۶ داده شده‌اند.

برای مقایسهٔ روش سنّتی RK (۲۱.۱۰.۶) و روش قبلی RKF، خطاهای برشی در حل مسألهٔ سادهٔ

$$y' = x^4 \qquad y(0) = 0$$

را در نظر می‌گیریم. خطاهای برشی برای (۴۱.۱۰.۶) و (۲۱.۱۰.۶) به‌ترتیب چنین‌اند

$$\text{RKF}(41.10.6) : T_{n+1}(Y) \doteq 0.00048h^5$$

$$\text{RK}(21.10.6) : T_{n+1}(Y) \doteq -0.00083h^5 \qquad n \geq 0$$

جدول ۲۵.۶ ضرایب α_j و β_{ji} برای روش RKF

		β_{ji}				
j	α_j	$i=1$	۲	۳	۴	۵
۲	$\frac{1}{4}$	$\frac{1}{4}$				
۳	$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$			
۴	$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$		
۵	۱	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$	
۶	$\frac{1}{2}$	$-\frac{8}{27}$	۲	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$

جدول ۲۶.۶ ضرایب γ_j، $\hat{\gamma}_j$، c_j برای روش RKF

j	۱	۲	۳	۴	۵	۶
γ_j	$\frac{25}{216}$	۰	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	
$\hat{\gamma}_j$	$\frac{16}{135}$	۰	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$
c_j	$\frac{1}{360}$	۰	$-\frac{128}{4275}$	$-\frac{2197}{75240}$	$\frac{1}{50}$	$\frac{2}{55}$

این نتایج نشان می‌دهند که روش RKF معمولاً دارای خطای برشی کوچکتری است، گرچه در عمل، تفاوت به این بزرگی نخواهد بود. باید توجه کرد که در روش سنتی (۲۱.۱۰.۶) با طول گام h و برآورد خطای (۳۷.۱۰.۶) به یازده محاسبهٔ f برای رفتن از $y_h(x_n)$ به $y_h(x_n+2h)$ نیاز است. و روش RKF (۴۱.۱۰.۶) برای رفتن از $y_h(x_n)$ به $y_h(x_n+2h)$ دوازده محاسبهٔ f را لازم دارد. در نتیجه تلاش محاسباتی در رفتن از x_n به x_n+2h قابل‌مقایسه‌اند، و درست است که خطای آنها را برای یک مقدار h، مقایسه نماییم.

مثال روش (۴۱.۱۰.۶) RKF را برای حل مسألهٔ (۲۲.۱۰.۶) به‌کار می‌بریم. این مسأله قبلاً یک مثال برای روش کلاسیک (۲۱.۱۰.۶) RK بود. مانند قبل، $h=0.25$؛ و نتایج در جدول ۲۷.۶ داده شده‌اند. مقدار نظری نسبت بازهم ۱۶ است و واضح است که هنوز به آن مقدار نرسیده است. وقتی h کاهش یابد، این نسبت بیشتر به ۱۶ نزدیک می‌شود. استفاده

جدول ۲۷.۶ مثالی از روش RKF

x	$y_h(x)$	$Y(x)-y_h(x)$	$Y(x)-y_{۲h}(x)$	نسبت	$\frac{۱}{۱۵}[y_h(x)-y_{۲h}(x)]$
۲٫۰	۰٫۴۰۰۰۰۸۸۱	−۸٫۸E − ۶	−۵٫۰E − ۴	۵۷	−۳٫۳E − ۵
۴٫۰	۰٫۲۳۵۲۹۴۶۹	−۵٫۸E − ۷	−۴٫۰E − ۵	۶۹	−۲٫۶E − ۶
۶٫۰	۰٫۱۶۲۱۶۲۲۶	−۹٫۵E − ۸	−۷٫۹E − ۶	۸۳	−۵٫۲E − ۷
۸٫۰	۰٫۱۲۳۰۷۶۹۵	−۲٫۶E − ۸	−۲٫۵E − ۶	۹۵	−۱٫۶E − ۷
۱۰٫۰	۰٫۰۹۹۰۰۹۹۱	−۹٫۴E − ۹	−۱٫۰E − ۶	۱۰۶	−۶٫۶E − ۸

از فرمول درونیابی (۲۴.۱۰.۶) ریچاردسن در آخرین ستون داده شده است، و به‌روشنی، خطا بیش از اندازهٔ واقعی برآورد شده است. معهذا، باز هم یک برآورد خطای مفید است، به‌این دلیل که تصوری از اندازهٔ خطای کلی به‌دست می‌دهد.

روش (۴۱.۱۰.۶) و (۴۲.۱۰.۶) معمولاً با برونیابی موضعی به‌کار برده می‌شود، همان‌گونه که بعداً نشان داده شده است. این روش خیلی مورد مطالعه قرار گرفته است تا معلوم شود آیا بهبودهایی امکان‌پذیر هست یا نه. اخیراً، شمپاین (۱۹۸۶) تحلیلی ارائه داده است که بعضی فرمولهای بهبودیافتهٔ RKF را که بر پایهٔ چندین محک برای مقایسهٔ فرمولهای رونگه ـ کوتا استوار است، پیشنهاد می‌نماید. تاکنون، این فرمولها در برنامه‌های رایانه‌یی با کیفیت بالا به‌کار برده نشده‌اند، گرچه انتظار می‌رود که مورد استفاده قرار گیرند.

برنامه‌های خودکار رونگه ـ کوتا ـ فِلبرک با استفاده از (۴۳.۱۰.۶) می‌توان یک برنامهٔ RKF با طول گام متغیر نوشت که خطای برشی در فرمول مرتبهٔ چهار (۴۱.۱۰.۶) را برآورد و کنترل کند. یک چنین روشی را ل. شمپاین و ه. واتس نوشته‌اند و در شمپاین و واتس (۱۹۷۶a) توضیح داده شده است. ویژگیهای کلی آن به‌صورت زیر است. برنامه RKF45 نامگذاری شده و کاربر برنامه باید دو پارامتر خطای ABSERR و RELERR را مشخص کند. خطای برشی $y_{n+۱}$ در (۴۳.۱۰.۶) باید در محک خطا در گام

$$|\ \mathrm{trunc}_j\ | \le \mathrm{ABSERR} + \mathrm{RELERR}* |\ y_{n,j}\ | \qquad (۴۴.۱۰.۶)$$

برای هر مؤلفهٔ y_n، مانند $y_{n,j}$ صدق کند، که y_n جواب محاسبه‌شدهٔ دستگاه معادلات دیفرانسیلی است که باید حل شود. ولی بعد، نتیجه نهایی محاسبه، برای استفاده در محاسبات بعدی، به جای فرمول مرتبهٔ چهارم $y_{n+۱}$ فرمول مرتبهٔ پنجم $\hat{y}_{n+۱}$ از (۴۲.۱۰.۶) به‌کار می‌رود. بنابراین به جای جملات y_n در طرف راست (۴۱.۱۰.۶) و (۴۲.۱۰.۶) باید $\hat{y}_n$ گذاشته شود. بنابراین RKF45

در واقع یک روش مرتبهٔ پنج است، که طول گام h را، با کنترل‌کردن خطای برشی در فرمول مرتبهٔ چهار (۴۱.۱۰.۶)، انتخاب می‌نماید. مانند قبل، این عمل برونیابی موضعی نامیده شده، و می‌توان نشان داد که کران (۴۴.۱۰.۶) برای y_{n+1}، ایجاب می‌کند که $\hat{y}_{n+1}$ در یک کران اصلاح‌شدهٔ خطا در طول گام واحد در خطای برشی صدق نماید. استدلال شبیه استدلالی است که در شمپاین و گوردن (۱۹۷۵، ص ۱۰۰) برای روش آدامز متغیر-مرتبهٔ داده شده است.

آزمونهای شمپاین و همکاران (۱۹۷۹) نشان می‌دهند که RKF45 یک برنامهٔ برتر RK است، برنامه‌ای که برای گنجاندن در کتابخانهٔ برنامه‌های حل معادلات دیفرانسیل معمولی انتخابی بسیار عالی است. این برنامه خیلی متداول شده و در چندین کتاب درسی [مثلاً فورسایت و همکاران (۱۹۷۷)] آمده است.

به‌طور کلّی، مقایسه‌هایی که در انرایت[1] و هال[2] (۱۹۷۶) داده‌شده، نشان می‌دهند که روشهای RKF برتر از روشهای دیگر RK هستند. مقایسه با روشهای چندگامی مشکلتر است. روشهای چندگامی، کمتر به محاسبهٔ مشتق $f(x,y)$ نیاز دارند، ولی هزینه‌های بالاسری درگام برای روشهای چندگامی خیلی بیشتر از روشهای RK است. یک راه داوری برای آنکه ببینیم کدام روش بهتر است به‌کار برده شود، این است که هزینهٔ محاسبهٔ $f(x,y)$ را با هزینهٔ بالاسری در روشهای چندگامی، مقایسه کنیم. ملاحظات دیگری هم وجود دارند، برای مثال، اندازهٔ دستگاه معادلات دیفرانسیلی است که باید حل شود. یک بحث کلی از این عوامل و اثرات آنها در انرایت و هال (۱۹۷۶) و شمپاین و همکاران (۱۹۷۶) داده شده است.

مثال مسألهٔ

$$y' = \frac{y}{۴}\left(۱ - \frac{y}{۲۰}\right) \qquad y(۰) = ۱$$

را، که جواب آن

$$y(x) = \frac{۲۰}{(۱ + ۱۹e^{-x/۴})}$$

است، حل کنید.

مسأله با RKF45 حل شده و جوابها در نقاط $x = ۲, ۴, ۶, \ldots, ۲۰$ داده شده‌اند. سه مقدار ABSERR به‌کار برده شده و در تمام حالات $\text{RELERR} = ۱۰^{-۱۲}$. خطاهای کلی نتیجه در جدول ۲۸.۶ داده شده است. ستونی که با NFE مشخص شده تعداد محاسبه‌های $f(x,y)$ را برای به‌دست آوردن جواب داده‌شده مشخص می‌کند. نتیجه را با آنچه که در جدول (۱۵.۶) که

1. Enright 2. Hull

جدول ۲۸.۶ مثالی از برنامهٔ رونگه ـ کوتا ـ فلبرگ RKF45

	ABSERR $= ۱۰^{-۳}$		ABSERR $= ۱۰^{-۶}$		ABSERR $= ۱۰^{-۹}$	
x	خطا	NFE	خطا	NFE	خطا	NFE
۴٫۰	−۱٫۴E − ۵	۱۹	−۲٫۲E − ۷	۴۳	−۵٫۵E − ۱۰	۱۲۱
۸٫۰	−۳٫۴E − ۵	۳۱	−۵٫۶E − ۷	۷۹	−۱٫۳E − ۹	۲۲۹
۱۲٫۰	−۳٫۷E − ۵	۴۳	−۵٫۵E − ۷	۱۰۳	−۸٫۶E − ۱۰	۳۱۲
۱۶٫۰	−۳٫۳E − ۵	۵۵	−۵٫۴E − ۷	۱۲۷	−۱٫۲E − ۹	۳۹۵
۲۰٫۰	−۱٫۸E − ۶	۶۷	−۱٫۶E − ۷	۱۶۳	−۴٫۳E − ۱۰	۵۰۳

از برنامهٔ چندگامی متغیر-مرتبهٔ DDEABM به‌دست آمده بود مقایسه نمایید.

صورتی از RKF45در دست است که خطای کلی در جواب محاسبه‌شده را، با استفاده از برآورد خطای ریچاردسن مانند (۲۴.۱۰.۶)، برآورد می‌نماید. برای بحث در مورد این برنامه که GERK خوانده می‌شود، شمپاین و واتس (۱۹۷۶b) را ببینید، که شامل یک بحث در مورد مسألهٔ عمومی برآورد خطای کلی نیز هست.

روشهای ضمنی رونگه ـ کوتا ما فقط روشهای صریح RK را بررسی کرده‌ایم، زیرا اینها، روشهای اساسی در کاربردهای کنونی هستند. ولی فرمولهای ضمنی RK هم وجود دارند. با تعمیم فرمولهای صریح (۱۹.۱۰.۶) و (۲۰.۱۰.۶)، فرمولهای زیر را در نظر می‌گیریم

$$y_{n+۱} = y_n + hF(x_n, y_n, h; f) \qquad (۴۵.۱۰.۶)$$

$$F(x, y, h; f) = \sum_{j=۱}^{p} \gamma_j V_j$$

$$V_j = f\left(x + \alpha_j h, y + h\sum_{i=۱}^{p} \beta_{ji} V_i\right) \qquad j = ۱, \ldots, p \qquad (۴۶.۱۰.۶)$$

ضرایب $\{\alpha_j, \beta_{ji}, \gamma_j\}$ روش مورد نظر را مشخص می‌نمایند. در یک روش صریح برای $i \geq j$، داریم $\beta_{ji} = ۰$.

در سالهای اخیر، روشهای ضمنی به‌طور دامنه‌داری مورد مطالعه قرار گرفته‌اند، زیرا بعضی از آنها دارای ویژگیهای پایداری مساعدی برای حل معادلات دیفرانسیل سرسخت هستند. برای هر مرتبه‌ای از همگرایی، روشهای A ـ پایداری از همان مرتبه وجود دارند. از این لحاظ، روشهای ضمنی RK بر روشهای چندگامی برتری دارند، زیرا در روشهای چندگامی هیچ روش A ـ پایدار بیش از مرتبهٔ دو وجود ندارد. گسترده‌ترین برنامه‌های حل معادلات دیفرانسیل سرسخت، در حال

حاضر، بر پایهٔ فرمولهای مشتقگیری پسرو قرار دارند که در بخش ۹.۶ توضیح داده شده است. ولی کار بسیار زیادی برای ایجاد برنامه‌های مشابه بر پایهٔ روشهای ضمنی RK وجود دارد. برای یک مطالعهٔ مقدماتی در این زمینه، اِیکن (۱۹۸۵، بخش ۱.۳) را ببینید.

۱۱.۶ مسائل مقدار مرزی

تا اینجا، ما فقط روشهای عددی برای حل **مسائل مقدار اولیه** در معادلات دیفرانسیل را مطالعه کردیم. برای چنین مسائلی، شرایط جواب معادلهٔ دیفرانسیل در یک نقطه معین‌شده است، که **نقطهٔ اولیه** خوانده می‌شود. اکنون مسائلی را بررسی می‌کنیم که در آنها، شرایط در بیش از یک نقطه بر جواب مجهول تحمیل شده است. چنین مسائلی در معادلات دیفرانسیل **مسائل مقدار مرزی** یا BVP خوانده می‌شوند.

یک مسألهٔ نمونه‌یی که مطالعه می‌کنیم BVPی دونقطه‌یی است:

$$y'' = f(x, y, y') \qquad a < x < b$$
$$A \begin{bmatrix} u(a) \\ u'(a) \end{bmatrix} + B \begin{bmatrix} u(b) \\ u'(b) \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \qquad (۱.۱۱.۶)$$

جملات A و B معرّف ماتریسهای 2×2 هستند و γ_1 و γ_2 ثابتهای داده شده‌اند. نظریهٔ BVP از این نوع بسیار پیچیده‌تر از نظریهٔ مسألهٔ مقدار اولیه است (قضیه‌های ۱.۶ و ۲.۶ در بخش ۱.۶ را ببینید). برای نمایاندن مشکلات ممکن، مثالهای زیر را می‌آوریم.

مثال ۱. BVPی خطی دونقطه‌یی زیر را در نظر می‌گیریم

$$y'' = -\lambda y \qquad 0 < x < 1$$
$$y(0) = y(1) = 0 \qquad (۲.۱۱.۶)$$

اگر λ یکی از اعداد

$$\pi^2, 4\pi^2, 9\pi^2, \ldots, n^2\pi^2, \ldots \qquad (۳.۱۱.۶)$$

نباشد، این BVP دارای جواب یکتای $Y(x) \equiv 0$ است. در غیر این صورت، تعداد نامتناهی جواب

$$Y(x) = C \sin(\sqrt{\lambda} x) \qquad (۴.۱۱.۶)$$

وجود دارد، که λ از (۶-۱۱-۳) انتخاب می‌شود و C ثابت اختیاری است.

۲. همچنین مسألهٔ دونقطه‌یی زیر را در نظر می‌گیریم

$$y'' = -\lambda y + g(x) \qquad ۰ < x < ۱$$
$$y(۰) = y(۱) = ۰ \qquad (۵.۱۱.۶)$$

اگر λ از (۳.۱۱.۶) انتخاب نشده باشد و اگر $g(x)$ در $a \le x \le b$ پیوسته باشد، آنگاه این مسأله یک جواب یکتای $Y(x)$ دارد که در $[۰, ۱]$ دو بار پیوسته مشتقپذیر است. به عکس اگر $\lambda = \pi^۲$، آنگاه مسألهٔ (۵.۱۱.۶) دارای جواب است اگر و تنها اگر $g(x)$ در رابطهٔ زیر صدق کند

$$\int_۰^۱ g(x)\sin(\pi x)dx = ۰$$

در حالتی که این رابطه برقرار باشد، جواب با رابطهٔ

$$Y(x) = C\sin(\pi x) + \frac{۱}{\pi}\int_۰^x g(t)\sin(\pi(x-t))dt \qquad (۶.۱۱.۶)$$

داده می‌شود که C یک ثابت اختیاری است. نتیجهٔ مشابهی برای λ ی دیگری که از (۳.۱۱.۶) انتخاب شده باشد برقرار است.

چند قضیّه از نظریهٔ BVPی دو نقطه‌یی اکنون داده می‌شود، تا به فهم بیشتر مطالب و عرضهٔ روشهای عددی برای حل آنها، ما را یاری نمایند. ما کار را با مسألهٔ دونقطه‌یی برای معادله دیفرانسیل خطی مرتبهٔ دوم آغاز می‌کنیم:

$$y'' = p(x)y' + q(x)y + g(x) \qquad a < x < b$$
$$A\begin{bmatrix} u(a) \\ u'(a) \end{bmatrix} + B\begin{bmatrix} u(b) \\ u'(b) \end{bmatrix} = \begin{bmatrix} \gamma_۱ \\ \gamma_۲ \end{bmatrix} \qquad (۷.۱۱.۶)$$

مسألهٔ همگن حالتی است که در آن $g(x) \equiv ۰$ و $\gamma_۱, \gamma_۲ = ۰$.

قضیهٔ ۱۰.۶ مسألهٔ ناهمگن (۷.۱۱.۶) دارای جواب یکتای $Y(x)$ بر $[a, b]$، برای هر مجموعه از داده‌های $\{g(x), \gamma_۱, \gamma_۲\}$ است اگر و تنها اگر مسألهٔ همگن مربوطهٔ آن فقط جواب بدیهی $Y(x) \equiv ۰$ داشته باشد.

برهان استکگولد[1] (۱۹۷۹، ص ۱۹۷) را ببینید. مثالهای قبل مثالهایی از این قضیه‌اند. ■

1. Stakgold

برای شرایطی که در آنها، مسألهٔ همگن (۷.۱۱.۶)، فقط جواب صفر داشته باشد، مسألهٔ خطی خاصتر زیر را در نظر می‌گیریم:

$$\begin{aligned} &y'' = p(x)y' + q(x)y + g(x) \qquad a < x < b \\ &a_0 y(a) - a_1 y'(a) = \gamma_1 \qquad b_0 y(b) + b_1 y'(b) = \gamma_2 \end{aligned} \tag{۸.۱۱.۶}$$

در این مسأله می‌گوییم که شرایط مرزی جدا شده‌اند. فرض می‌کنیم که شرایط زیر برقرار باشند

$$\begin{aligned} &q(x) > 0 \qquad a \le x \le b \\ &a_0 a_1 \ge 0 \qquad b_0 b_1 \ge 0 \\ &|a_0| + |a_1| \neq 0 \qquad |b_0| + |b_1| \neq 0 \qquad |a_0| + |b_0| \neq 0 \end{aligned} \tag{۹.۱۱.۶}$$

در این صورت مسألهٔ همگن (۸.۱۱.۶) فقط دارای جواب صفر است، قضیهٔ ۱۰.۶ را می‌توان به‌کار برد، و مسألهٔ ناهمگن برای هر مجموعه از داده‌های $\{g(x), \gamma_1, \gamma_2\}$ دارای جواب یکتاست. برای اثبات این قضیّه کلر (۱۹۶۸، ص ۱۱) را ببینید. مثال (۵.۱۱.۶) توضیحی برای این قضیّه است. این مثال همچنین نشان می‌دهد که شرایط (۹.۱۱.۶) لازم نیستند. مسألهٔ (۵.۱۱.۶) برای بیشتر انتخابهای منفی $q(x) = \lambda$ یکتا حلپذیر است.

نظریهٔ مسألهٔ غیرخطی (۱.۱۱.۶) به مراتب پیچیده‌تر از نظریهٔ مسألهٔ خطی (۷.۱۱.۶) است. ما یک مقدّمهٔ مختصری از این نظریه برای مسألهٔ محدودتر زیر می‌آوریم.

$$\begin{aligned} &y'' = f(x, y, y') \qquad a < x < b \\ &a_0 y(a) - a_1 y'(a) = \gamma_1 \qquad b_0 y_1(b) + b_1 y'(b) = \gamma_2 \end{aligned} \tag{۱۰.۱۱.۶}$$

فرض بر این است که تابع f در همهٔ نقاط (x, u_i, v)، (x, u, v_i) در ناحیهٔ

$$R = \{(x, u, v) \mid a \le x \le b, -\infty < u, v < \infty\}$$

در شرط لیپشیتس زیر صدق می‌کند

$$\begin{aligned} &|f(x, u_1, v) - f(x, u_2, v)| \le K |u_1 - u_2| \\ &|f(x, u, v_1) - f(x, u, v_2)| \le K |v_1 - v_2| \end{aligned} \tag{۱۱.۱۱.۶}$$

این شرط بسیار قویتر از شرط مورد نیاز است، ولی بیان قضیهٔ زیر و تحلیل روشهای عددی را که بعداً داده خواهد شد آسان می‌کند.

قضیهٔ ۱۱.۶ برای مسألهٔ (۱۰.۱۱.۶)، فرض می‌کنیم $f(x,u,v)$ در ناحیهٔ R پیوسته باشد و در شرط (۱۱.۱۱.۶) لیپشیتس صدق کند. به‌علاوه فرض می‌کنیم که در R، f در روابط زیر صدق می‌کند:

$$\frac{\partial f(x,u,v)}{\partial u} > 0 \qquad |\frac{\partial f(x,u,v)}{\partial v}| \leq M \qquad (۱۲.۱۱.۶)$$

برای مقداری از ثابت $M > 0$. برای شرایط مرزی (۱۰.۱۱.۶)، فرض می‌کنیم

$$a_0 a_1 \geq 0 \qquad b_0 b_1 \geq 0$$
$$|a_0|+|a_1| \neq 0 \qquad |b_0|+|b_1| \neq 0 \qquad |a_0|+|b_0| \neq 0 \quad (۱۳.۱۱.۶)$$

در این صورت BVP (۱۰.۱۱.۶) دارای جواب یکتاست.

برهان کلر (۱۹۶۸، ص ۹) را ببینید. قضیهٔ یکتایی قبلی برای مسألهٔ خطی (۸.۱۱.۶) یک حالت خاص این قضیه است. ■

BVPهای غیرخطی ممکن است نایکتا حلپذیر باشند و فقط تعدادی متناهی جواب، داشته باشند. این وضعیّت در مسائل خطی پیش نمی‌آید، همان‌طور که در (۲.۱۱.۶) و (۶.۱۱.۶) نشان داده شد نایکتایی زیادی است و همیشه به معنای بی‌نهایت جواب است. یک مثال نایکتا حلپذیری، مسألهٔ مرتبهٔ دوم زیر است:

$$\frac{d}{dx}\left[I(x)\frac{dy}{dx}\right] + \lambda \sin(y) = 0 \qquad 0 < x < 1$$
$$y'(0) = y'(1) = 0 \qquad (۱۴.۱۱.۶)$$

که در مطالعهٔ کمانش ستون پیش می‌آید. عامل λ متناسب با باری است که بر ستون گذاشته شده است؛ وقتی λ از حد معینی تجاوز نماید، مسألهٔ (۱۴.۱۱.۶) یک جواب غیر از جواب صفر دارد. برای تفصیل بیشتر این مسأله، کلر و آنتمان[1] (۱۹۶۹، ص ۴۳) را ببینید.

همانند مطالب قبلی دربارهٔ مسائل مقدار اولیه [(۱۱.۱.۶) تا (۱۵.۱.۶) بخش ۱.۶ را ببینید]، تمام مسائل مقدار مرزی از مراتب بالا را می‌توان به‌صورت مسائلی برای یک دستگاه معادلات مرتبهٔ اول نوشت. این شکل کلی برای یک BVP دونقطه‌یی در یک دستگاه معادلات مرتبهٔ اول چنین است:

1. Antman

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}) \qquad a < x < b$$
$$A\mathbf{y}(a) + B\mathbf{y}(b) = \gamma \qquad (۱۵.۱۱.۶)$$

که یک دستگاه n معادلهٔ مرتبهٔ اول را نشان می‌دهد. کمیتهای $\mathbf{y}(x)$، $\mathbf{f}(x, \mathbf{y})$ و γ بردارهایی با n مؤلفه و A و B ماتریسهایی از مرتبهٔ $n \times n$ هستند. یک نظریه برای این‌گونه BVPها وجود دارد، مشابه با نظریه‌ای که برای مسألهٔ دونقطه‌یی (۱.۱۱.۶) ذکر کردیم، ولی به دلایل کمبود جا، از ذکر آن صرفنظر می‌کنیم.

در بقیهٔ این بخش، روشهای عددی اصلی برای حل BVPی دونقطه‌یی (۱.۱۱.۶) را به اختصار توضیح می‌دهیم. این روشها، برای دستگاه معادلات مرتبهٔ اول از قبیل (۱۵.۱۱.۶) تعمیم می‌یابند، ولی به لحاظ کمبود جا، آن قضایا را ذکر نمی‌کنیم. در بیشتر مطالب ارائه‌شده از کلر (۱۹۶۸) پیروی کرده‌ایم، و نظریهٔ دستگاههای مرتبهٔ اول را در آنجا داده‌ایم. برخلاف وضعیت مسائل مقدار اولیه، اغلب بهتر است که با BVPی مرتبهٔ بالاتر مستقیماً برخورد نماییم تا اینکه مسأله را به شکل یک دستگاه مرتبهٔ اول به‌طور عددی حل کنیم. ارائهٔ روشهای عددی مسألهٔ دونقطه‌ای، کمتر پیچیده است، و بنابراین ما بحث دربارهٔ مسألهٔ دو نقطه را به جای دستگاه (۱۵.۱۱.۶)، انتخاب نموده‌ایم.

روشهای پرتابی یکی از راههای متداول حل BVPی دونقطه‌یی، تبدیل آن است به مسأله‌ای که در آن از برنامه‌ای برای حل مسائل مقدار اولیه بتوان استفاده کرد. اکنون یک چنین روشی را برای BVPی (۱۰.۱۱.۶) به‌دست می‌دهیم.

مسألهٔ مقدار اولیهٔ زیر را در نظر می‌گیریم

$$y'' = f(x, y, y') \qquad a < x < b$$
$$y(a) = a_1 s - c_1\gamma_1 \qquad y'(a) = a_0 s - c_0\gamma_1 \qquad (۱۶.۱۱.۶)$$

که به پارامتر s بستگی دارد، و c_0 و c_1 ثابتهای اختیاری هستند که در شرط زیر صدق می‌کنند

$$a_1 c_0 - a_0 c_1 = 1$$

جواب (۱۶.۱۱.۶) را با $Y(x; s)$ نشان می‌دهیم. در این صورت به‌سادگی دیده می‌شود که برای، همهٔ مقادیر s که Y وجود داشته باشد

$$a_0 Y(a; s) - a_1 Y'(a; s) = \gamma_1$$

چون Y یک جواب (۱۶.۱۱.۶) است، تنها چیزی که لازم است تا این Y جواب (۱۰.۱۱.۶) باشد این است که در شرط مرزی در b که باقیمانده صدق کند. این بدان معناست که $Y(x;s)$ باید در شرط زیر صدق کند

$$\varphi(s) \equiv b_0 Y(b;s) + b_1 Y'(b;s) - \gamma_2 = 0 \qquad (۱۷.۱۱.۶)$$

این معادله یک معادلهٔ غیرخطی برحسب s است. اگر s^* یک ریشهٔ $\varphi(s)$ باشد، آنگاه $Y(x;s)$ در BVPی (۱۰.۱۱.۶) صدق می‌کند. می‌توان نشان داد که با مفروضات مناسبی روی f و شرایط مرزی آن، معادلهٔ (۱۷.۱۱.۶) دارای یک جواب یکتای s^* خواهد بود [کلر (۱۹۶۸، ص ۹) را ببینید]. می‌توانیم یک روش ریشه‌یابی برای معادلهٔ غیرخطی به‌کار بریم و s^* را پیدا کنیم. این طریق پیداکردن جواب برای BVP **روش پرتابی** خوانده می‌شود. این نام از پرتاب‌شناسی گرفته شده که در آنجا سعی می‌کنند شرایط اولیهٔ لازم در $x = a$ را معلوم نمایند تا مقدار معینی در $x = b$ به‌دست آید.

هریک از روشهای ریشه‌یابی فصل ۲ را می‌توان برای حل $\varphi(s) = 0$ به‌کار برد. هر محاسبهٔ $\varphi(s)$ شامل حل (۱۶.۱۱.۶) بر $[a,b]$ است، و درنتیجه می‌خواهیم تعداد چنین محاسباتی را مینیمم سازیم. به‌عنوان یک مثال خاص از یک روش همگرای سریع و مهم، به روش نیوتن می‌نگریم:

$$s_{m+1} = s_m - \frac{\varphi(s_m)}{\varphi'(s_m)} \qquad m = 0, 1, \ldots \qquad (۱۸.۱۱.۶)$$

برای محاسبهٔ $\varphi'(s)$، از شرط (۱۷.۱۱.۶) مشتق می‌گیریم تا به‌دست آوریم

$$\varphi'(s) = b_0 \xi_s(b) + b_1 \xi_s'(b) \qquad (۱۹.۱۱.۶)$$

که

$$\xi_s(x) = \frac{\partial Y(x;s)}{\partial s} \qquad (۲۰.۱۱.۶)$$

برای پیداکردن $\xi_s(x)$ از معادلهٔ زیر نسبت به s مشتق می‌گیریم

$$Y''(x;s) = f(x, Y(x;s), Y'(x;s))$$

در این صورت ξ_s در مسألهٔ مقدار اولیهٔ زیر صدق می‌کند

$$\begin{aligned} \xi_s''(x) =& f_2(x, Y(x;s), Y'(x;s))\xi_s(x) \\ &+ f_3(x, Y(x;s), Y'(x,s))\xi_s'(x) \qquad (۲۱.۱۱.۶) \\ \xi_s(a) = a_1 \qquad & \xi_s'(a) = a_0 \end{aligned}$$

توابع f_2 و f_3 معرّف مشتقات جزیی $f(x,u,v)$ به ترتیب نسبت به u و v هستند. مقادیر اولیه همان مقادیری هستند که در (۱۶.۱۱.۶) و از تعریف ξ_s به‌دست آمده‌اند.

در عمل مسائل (۱۶.۱۱.۶) و (۲۱.۱۱.۶) را به یک دستگاه چهار معادلهٔ مرتبهٔ اول با مجهولات Y، Y'، ξ_s و ξ_s' برمی‌گردانیم. این دستگاه به‌طور عددی مثلاً با یک روش مرتبهٔ p و طول گام h حل می‌شود. با نمادگذاری مشابهی برای سایر مجهولات، فرض می‌کنیم $y_h(x;s)$ تقریب $Y(x;s)$ را نشان دهد. از قضایای پیشین برای حل مسائل مقدار اولیه، می‌توان نشان داد که این جوابهای تقریبی از لحاظ خطا از مرتبهٔ $O(h^p)$ خواهند بود. با فرضهای مناسب در مسألهٔ اصلی (۱۰.۱۱.۶)، می‌توان نشان داد که ریشهٔ به‌دست آمدهٔ s_h^* نیز دارای خطای $O(h^p)$ خواهد بود، و همچنین است برای جواب تقریبی $y_h(x;s_h^*)$ وقتی با جواب $Y(x;s^*)$ مسألهٔ مقدار مرزی مقایسه شود. برای جزئیات این تحلیل کلر (۱۹۶۸، صص ۴۷ـ۵۴) را ببینید.

مثال BVPی دو نقطه‌یی زیر را در نظر می‌گیریم

$$\begin{aligned} y'' &= -y + \frac{2(y')^2}{y} \qquad -1 < x < 1 \\ y(-1) &= y(1) = (e+e^{-1})^{-1} \doteq 0.324027137 \end{aligned} \tag{۲۲.۱۱.۶}$$

جواب درست آن $Y(x) = (e^x + e^{-x})^{-1}$ است. مسألهٔ مقدار اولیه (۱۵.۱۱.۶) برای روش پرتابی چنین است

$$\begin{aligned} y'' &= -y + \frac{(2y')^2}{y} \qquad -1 < x < 1 \\ y(-1) &= (e+e^{-1})^{-1} \qquad y'(-1) = s \end{aligned} \tag{۲۳.۱۱.۶}$$

مسألهٔ مربوط به (۲۱.۱۱.۶) برای $\xi_s(x)$ عبارت است از:

$$\begin{aligned} \xi_s'' &= \left[-1 - 2\left(\frac{y'}{y}\right)^2\right]\xi_s + 4\frac{y'}{y}\xi_s' \\ \xi_s(-1) &= 0 \qquad \xi_s'(-1) = 1 \end{aligned} \tag{۲۴.۱۱.۶}$$

در معادلهٔ ξ_s'' از $Y(x;s)$، جواب (۲۳.۱۱.۶)، استفاده می‌شود. تابع $\varphi(s)$ برای محاسبهٔ s^* با رابطهٔ زیر داده می‌شود

$$\varphi(s) \equiv Y(1;s) - (e+e^{-1})^{-1}$$

جدول ۲۹.۶ روش پرتابی برای حل (۲۲.۱۱.۶)

$n=\frac{۲}{h}$	$s^*-s_h^*$	نسبت	E_h	نسبت
۴	۴٫۰۱E − ۳		۲٫۸۳E − ۲	
۸	۱٫۵۲E − ۳	۲٫۶۴	۷٫۳۰E − ۳	۳٫۸۸
۱۶	۴٫۶۴E − ۴	۳٫۲۸	۱٫۸۲E − ۳	۴٫۰۱
۳۲	۱٫۲۷E − ۴	۳٫۶۴	۴٫۵۴E − ۴	۴٫۰۱
۶۴	۳٫۳۴E − ۵	۳٫۸۲	۱٫۱۴E − ۴	۴٫۰۰

برای استفاده از تعریف روش نیوتن، داریم

$$\varphi'(s)=\xi_s(۱)$$

با استفاده از Y، جواب درست (۲۲.۱۱.۶) و شرط $y'(-۱)=s$ در (۲۳.۱۱.۶)، به‌سادگی s^*، ریشهٔ مطلوب $\varphi(s)$ به‌دست می‌آید

$$s^*=Y'(-۱)=\frac{e-e^{-۱}}{(e+e^{-۱})^۲}\doteq ۰٫۲۴۵۷۷۷۱۷۴$$

برای حل مسألهٔ مقدار اولیهٔ (۲۳.۱۱.۶)ـ(۲۴.۱۱.۶)، از روش مرتبهٔ دوم (۹.۱۰.۶) رونگه‌ـکوتا با طول گام $h=۲/n$ استفاده نموده‌ایم. نتایج برای چندین مقدار n در جدول ۲۹.۶ داده شده است. جواب (۲۴.۱۱.۶) با $y_h(x;s)$ نشان داده شده است و ریشهٔ حاصل

$$\varphi_h(s)\equiv y_h(۱;s)-(e+e^{-۱})^{-۱}=۰$$

با s_h^* نشان داده شده است. برای خطا در $y_h(x;s_h^*)$، گیریم

$$E_h=\underset{۰\le i\le n}{\text{Max}}\mid Y(x_i)-y_h(x_i;s_h^*)\mid$$

که $\{x_i\}$ها نقاط گرهی‌اند که از آنها در حل مسألهٔ مقدار اولیه استفاده شده است. ستونهایی که با نسبت مشخص شده‌اند ضرایبی را مشخص می‌کنند که با آنها، وقتی n دو برابر (یا h نصف) می‌شود، خطا کاهش می‌یابد. به‌طور نظری این ضرایب باید به ۴ نزدیک شوند زیرا روش رونگه‌ـکوتا دارای خطایی از مرتبهٔ $O(h^۴)$ است. و همان‌گونه که انتظار می‌رود، در تجربه این ضرایب به ۴٫۰ نزدیک می‌شوند. برای بارست (۱۸.۱۱.۶)ی نیوتن، در هر حالت $s_۰=۰٫۲$ به‌کار برده شده است. بارست وقتی پایان یافته است که آزمون

$$\mid s_{m+۱}-s_m\mid\le ۱۰^{-۱۰}$$

برآورده شده است. با این گزینشها، روش نیوتن در هر حالت به ۶ بارست نیاز داشته است جز حالت $n = 4$ (که هفت بارست لازم بوده است). ولی اگر $s_0 = 0$ به‌کار برده می‌شد، برای حالت $n = 4$، ۲۵ بارست لازم می‌شد، که اهمیت یک گزینش خوب حدس اولیهٔ s_0 را نشان می‌دهد.

مسائل چندی در روش پرتابی پیش می‌آید. اول، یک حدس کلی s_0 برای بارست نیوتن وجود ندارد، و با یک گزینش ضعیف، این بارست ممکن است واگرا شود. بدین دلیل، روش اصلاح‌شدهٔ (۱۱.۱۱.۲) ـ (۱۲.۱۱.۲) نیوتن در بخش ۱۱.۲ ممکن است لازم باشد تا همگرایی به‌وجود آید. مسألهٔ دوّم اینکه انتخاب $y_h(x; s)$ ممکن است نسبت به h، s و سایر ویژگیهای مسألهٔ مقدار مرزی، خیلی حساس باشد. برای مثال، اگر خطی‌سازی مسألهٔ مقدار آغازی (۱۶.۱۱.۶) ویژه مقدارهای مثبت بزرگی داشته باشد، انتخاب $Y(x; s)$ احتمالاً نسبت به تغییرات s حساس می‌شود. برای بحث در مورد تمام این مسائل، کلر (۱۹۶۸، فصل ۲)، اشتور[1] و بورلیرش[2] (۱۹۸۰، بخش ۳۰۷)، و فاکس[3] (۱۹۸۰، صص ۱۸۰ـ۱۸۶) را ببینید. بررسی بعضی از این مسائل BVPهای خطی آسانتر است، مانند (۸.۱۱.۶)، آن‌گونه که در کلر (۱۹۶۸، فصل ۲) صورت گرفته است.

روشهای تفاضل متناهی BVP دونقطه‌یی

$$\begin{aligned} y'' &= f(x; y, y') \qquad a < x < b \\ y(a) &= \gamma_1 \qquad y(b) = \gamma_2 \end{aligned} \tag{۲۵.۱۱.۶}$$

را در نظر می‌گیریم که جواب درست آن با $Y(x)$ نشان داده شده است. گیریم $n > 1$، $h = (b-a)/n$، $x_j = a + jh$ به‌ازای $j = 0, 1, \ldots, n$. در هر نقطهٔ گرهی داخلی x_i، $0 < i < n$، $Y''(x_i)$ و $Y'(x_i)$ را تقریب می‌زنیم:

$$\begin{aligned} Y''(x_i) &= \frac{Y_{i+1} - 2Y_i + Y_{i-1}}{h^2} - \frac{h^2}{12} Y^{(4)}(\xi_i) \\ Y'(x_i) &= \frac{Y_{i+1} - Y_{i-1}}{2h} - \frac{h^2}{6} Y^{(3)}(\eta_i) \end{aligned} \tag{۲۶.۱۱.۶}$$

به‌ازای یک مقدار $x_{i-1} \le \xi_i$ و $\eta_i \le x_{i+1}$، $i = 1, 2, \ldots, n-1$. با حذف آخرین جمله‌های خطا، و استفاده از این تقریبها در معادلهٔ دیفرانسیل، دستگاه غیرخطی تقریبی زیر را خواهیم داشت:

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = f\left(x_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h}\right) \qquad i = 1, \ldots, n-1 \tag{۲۷.۱۱.۶}$$

1. Stoer 2. Bulirsch 3. Fox

این دستگاه، یک دستگاه $n-1$ معادلهٔ غیرخطی برحسب $n-1$ مجهول $y_1, y_2, \ldots, y_{n-1}$ است. مقادیر $y_0 = \gamma_1$ و $y_n = \gamma_2$ با توجه به شرایط مرزی معلوم‌اند.

تحلیل خطا در $\{y_i\}$ در مقایسه با $\{Y(x_i)\}$ خیلی دشوارتر از آن است که در اینجا بتوان بیان کرد، زیرا نیاز به تحلیل حلپذیری دستگاه معادلات غیرخطی دارد. اساساً، اگر $Y(x)$ چهار بار مشتقپذیر باشد، اگر مسألهٔ (۲۵.۱۱.۶) در ناحیه‌ای پیرامون نمودار $Y(x)$ بر $[a,b]$، به‌طور یکتا حلپذیر باشد، و اگر $f(x,u,v)$ به اندازهٔ کافی مشتقپذیر باشد، آنگاه یک جواب (۲۷.۱۱.۶) وجود دارد و در رابطهٔ زیر صدق می‌کند

$$\max_{0 \le i \le n} | Y(x_i) - y_i | = O(h^2) \qquad (۲۸.۱۱.۶)$$

برای تحلیلی از این موضوع به کلر (۱۹۷۶، بخش ۲.۳) یا کلر (۱۹۶۸، بخش ۲.۳) مراجعه نمایید. وانگهی با فرضهای بیشتر روی f و همواری Y، می‌توان نشان داد که

$$Y(x_i) - y_i = \tau(x_i)h^2 + O(h^4) \qquad (۲۹.۱۱.۶)$$

که $\tau(x)$ مستقل از h است. این را می‌توان برای توجیه برونیابی ریچاردسن در به‌دست‌آوردن نتایجی که سریعتر همگرا باشند، به‌کار برد. [در سایر روشها، بهبود همگرایی بر پایهٔ تصحیح خطا در تقریبهای تفاضلی مرکزی (۲۷.۱۱.۶) قرار دارند.]

دستگاه (۲۷.۱۱.۶) به راههای گوناگونی حل می‌شود، که بعضی از آنها اصلاح سادهٔ روشهایی است که در بخش ۱۰.۲ و ۱۱.۲ داده شدند. در شکل ماتریسی، داریم

$$\frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 & & \cdots & 0 \\ 1 & -2 & 1 & & & \vdots \\ \vdots & & \ddots & & & \\ & & & 1 & -2 & 1 \\ 0 & \cdots & & 0 & 1 & -2 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{bmatrix}$$

$$= \begin{bmatrix} f\left(x_1, y_1, \dfrac{y_2 - y_0}{2h}\right) \\ f\left(x_2, y_2, \dfrac{y_3 - y_1}{2h}\right) \\ \vdots \\ f\left(x_{n-1}, y_{n-1}, \dfrac{y_n - y_{n-2}}{2h}\right) \end{bmatrix} - \begin{bmatrix} \dfrac{\gamma_1}{h^2} \\ 0 \\ \vdots \\ \dfrac{\gamma_2}{h^2} \end{bmatrix}$$

که آن را به شکل زیر نشان می‌دهیم.

$$\frac{1}{h^2}A\mathbf{y} = \hat{\mathbf{f}}(\mathbf{y}) + \mathbf{g} \qquad (۳۰.۱۱.۶)$$

ماتریس A ناتکین است [قضیهٔ ۲.۸ فصل ۸ را ببینید]؛ و دستگاه خطّی $A\mathbf{z} = \mathbf{b}$ به‌سادگی حلپذیر است، همان‌طور که قبل از قضیهٔ ۲.۸ فصل ۸ آمده است. روش نیوتن (بخش ۱۱.۲ را ببینید) برای حل (۳۰.۱۱.۶) به‌صورت زیر داده می‌شود

$$\mathbf{y}^{(m+1)} = \mathbf{y}^{(m)} - \left[\frac{1}{h^2}A - F(\mathbf{y}^{(m)})\right]^{-1}\left[\frac{1}{h^2}A\mathbf{y}^{(m)} - \hat{\mathbf{f}}(\mathbf{y}^{(m)}) - \mathbf{g}\right] \qquad (۳۱.۱۱.۶)$$

$$F(\mathbf{y}) = \left[\frac{\partial \hat{f}_i}{\partial y_j}\right] \qquad 1 \le i, j \le n-1$$

ماتریس ژاکوبی به علّت شکل خاص $\hat{\mathbf{f}}(\mathbf{y})$ به‌طور قابل‌ملاحظه‌ای ساده می‌شود:

$$[F(\mathbf{y})]_{ij} = \frac{\partial f\left(x_i, y_i, \dfrac{y_{i+1} - y_{i-1}}{2h}\right)}{\partial y_i}$$

این عبارت صفر است مگر برای $j = i+1$ یا $j = i-1$:

$$[F(y)]_{ii} = f_2\left(x_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h}\right) \qquad 1 \le i \le n-1$$

$$[F(y)]_{i,i-1} = \frac{-1}{2h}f_3\left(x_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h}\right) \qquad 2 \le i \le n-1$$

$$[F(y)]_{i,i+1} = \frac{1}{2h}f_3\left(x_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h}\right) \qquad 1 \le i \le n-2$$

که $f_2(x,u,v)$ و $f_3(x,u,v)$ به ترتیب مشتقات جزیی نسبت به u و v را نشان می‌دهند. بنابراین ماتریسی که باید در (۳۱.۱۱.۶) معکوس شود به شکل خاصی است که **سه‌قطری** خوانده می‌شود. گیریم

$$B_m = \frac{1}{h^2}A - F(\mathbf{y}^{(m)}) \qquad (۳۲.۱۱.۶)$$

می‌توانیم (۳۱.۱۱.۶) را دوباره به شکل زیر بنویسیم

$$\begin{aligned} \mathbf{y}^{(m+1)} &= \mathbf{y}^{(m)} - \boldsymbol{\delta}^{(m)} \\ B_m\boldsymbol{\delta}^{(m)} &= \tfrac{1}{h^2}A\mathbf{y}^{(m)} - \mathbf{f}(\mathbf{y}^{(m)}) - \mathbf{g} \end{aligned} \qquad (۳۳.۱۱.۶)$$

جدول ۳۰.۶ روش تفاضل متناهی برای حل (۲۲.۱۱.۶)

$n = \frac{۲}{h}$	E_h	نسبت
۴	۲٫۶۳E − ۲	
۸	۵٫۸۷E − ۳	۴٫۴۸
۱۶	۱٫۴۳E − ۳	۴٫۱۱
۳۲	۳٫۵۵E − ۴	۴٫۰۳
۶۴	۸٫۸۶E − ۵	۴٫۰۱

همان‌گونه که در بخش ۲.۸ فصل ۸ نشان داده شده است، این دستگاه خطی به‌سادگی و به سرعت حلپذیر است. می‌توان نشان داد که تعداد ضربها و تقسیمها در حدود $۵n$ است، که در حل یک دستگاه $n-۱$ معادلهٔ خطی، زیاد نیست. صرفه‌جویی بیشتر را می‌توان، با تغییرندادن B_m یا تغییر آن فقط پس از چند بارست (۳۳.۱۱.۶)، به وجود آورد. برای یک بازنگری وسیع و بحث در حل دستگاههای غیرخطی که در ارتباط با حل BVPها پیش می‌آیند به دویفل‌هارت[1] (۱۹۷۹) مراجعه نمایید.

مثال روش تفاضل متناهی (۲۷.۱۱.۶) را قبلاً برای حل BVPی (۲۲.۱۱.۶)، که پیش از این برای مثال روش پرتابی استفاده شده بود، به‌کار برده‌ایم. نتایج برای دوبرابرکردنهای متوالی $n = ۲/h$ در جدول ۳۰.۶ داده شده است. دستگاه غیرخطی (۲۷.۱۱.۶) همان‌گونه که در (۳۳.۱۱.۶) توضیح داده شد، با روش نیوتن، حل شده است. حدس آغازی چنین بوده است:

$$y_h^{(۰)}(x_i) = (e + e^{-۱})^{-۱} \qquad i = ۰, ۱, \ldots, n$$

که بر پایهٔ وصل‌کردن مقادیر مرزی به یکدیگر با یک خط مستقیم، حاصل شده است. کمیّت

$$d_h = \underset{۰ \le i \le n}{\text{Max}} \mid y_i^{(m+۱)} - y_i^{(m)} \mid$$

برای هر بارست محاسبه شده و هرگاه شرط زیر برقرار شده، بارست خاتمه یافته است. در همهٔ حالات،

$$d_h \le ۱۰^{-۱۰}$$

تعداد بارستهای محاسبه‌شده ۵ یا ۶ بوده است. برای خطا، گیریم

$$E_h = \underset{۰ \le i \le n}{\text{Max}} \mid Y(x_i) - y_h(x_i) \mid$$

1. Deuflhard

که در آن y_h جواب (۲۷.۱۱.۶) است که از روش نیوتن به‌دست آمده است. به موجب (۲۸.۱۱.۶) و (۲۹.۱۱.۶) باید انتظار داشته باشیم که مقادیر E_h، وقتی h نصف می‌شود، با ضریبی حدود ۴ کاهش یابد و این چیزی است که در جدول مشاهده می‌نماییم.

روشهای مراتب بالاتر را می‌توان به چند راه به‌دست آورد. (۱) تقریبات از مرتبهٔ بالاتر را برای مشتقات به‌کاربرد تا (۲۶.۱۱.۶) بهبود یابد؛ (۲) برونیابی ریچاردسن را بر پایهٔ (۲۹.۱۱.۶) به‌کاربرد و مانندانتگرالگیری رامبرگ، برونیابی را تکرارنمود تا روشهای از مراتب به دلخواه بالاتر به‌دست آیند، (۳) خطاهای برشی در (۲۶.۱۱.۶) با تفاضلات مراتب بالاتر، با استفاده از مقادیر محاسبه‌شدهٔ y_h، تقریب‌زدن با استفاده از این مقادیر برای اصلاحات (۲۷.۱۱.۶)، و به‌دست‌آوردن یک تقریب دقیقتر برای معادلهٔ دیفرانسیل در (۲۵.۱۱.۶)، که به جواب دقیقتر می‌انجامد. تمام این فنون به‌کار گرفته شده‌اند، و بعضی به‌صورت برنامه‌های رایانه‌یی کاملاً پیچیده در آمده‌اند. برای بحث بیشتر و برای مثالهایی از برنامه‌های رایانه‌یی، فاکس (۱۹۸۰، ص ۱۹۱)، جین[1] (۱۹۸۴، فصل ۴)، و پریرا[2] (۱۹۷۹) را ببینید.

روشها و مسائل دیگر چند روش دیگر در حل مسائل مقدار مرزی به‌کار رفته‌اند. بهترین آنها احتمالاً **روش هم‌مکانی** است. برای بحث دربارهٔ روشهای هم‌مکانی ردّین[3] (۱۹۷۹)، دویفل‌هارت (۱۹۷۹) و اَشر و راسل (۱۹۸۵) را ببینید. برای یک برنامهٔ مهم رایانه‌یی هم‌مکانی اَشر و همکاران (۱۹۸۱b، ۱۹۸۱a) را ببینید.

یک روش دیگر برای حل مسألهٔ مقدار مرزی، این است که یک معادلهٔ انتگرالی صورت هم‌ارز آن را حل کنیم. دستاوردهای بسیارکمتری برای چنین روشهای عددی وجود دارند، ولی در بعضی موارد این روشها می‌توانند خیلی موثر باشند. برای آشنایی با این روش کلر (۱۹۶۸، فصل ۴) را ببینید.

همچنین انواع دیگری از مسائل مقدار مرزی وجود دارند، که بعضی شامل نوعی رفتارهای تکین هستند، که در اینجا بحث نکرده‌ایم. برای همهٔ این مطالب به مقالات گزارشی اَشر و راسل (۱۹۸۵)، عزیز[4] (۱۹۸۵)، چایلدز و همکاران (۱۹۷۹) و گلادول و سه‌یرز (۱۹۸۰) را ببینید؛ همچنین (فصل ۴) کلر (۱۹۷۶) را برای مسائل تکین ملاحظه نمایید. برای بحث نرم‌افزارها، چایلدز و همکاران (۱۹۷۹)، گلادول[5] و سه‌یرز[6] (۱۹۸۰) و انرایت[7] را ببینید.

---

1. Jain 2. Pereyra 3. Reddien 4. Aziz
5. Gladwell 6. Sayers 7. Enright

گفتگو دربارهٔ منابع

معادلات دیفرانسیل معمولی و جزئی شکل اصلی مدل ریاضی در علوم و مهندسی هستند، و در نتیجه، حل عددی معادلات دیفرانسیل یک زمینهٔ وسیع تحقیق است. دو کتاب درسی کلاسیک که معلومات در این زمینه را قبل از کاربرد فراگیر رایانه‌های رقمی، نشان می‌دهند کولاتس[1] (۱۹۶۶) و میلن[2] (۱۹۵۳) هستند. بعضی از کتابهای عمومی مهم که از ۱۹۶۰ به بعد در حل عددی معادلات دیفرانسیل معمولی نوشته شده‌اند عبارت‌اند از هنریچی (۱۹۶۲)، گی‌یر (۱۹۷۱)، لاپیدوس[3] و زاینفلت[4] (۱۹۷۱)، لامبرت[5] (۱۹۷۳)، اشتتر[6] (۱۹۷۳)، هال و وات (۱۹۷۶)، شمپاین و گوردن (۱۹۷۵)، واندرهوون[7] (۱۹۷۷)، اورتگا و پول (۱۹۸۱) و بوچر[8] (۱۹۸۷). یک ارزیابی مفید در گوپتا و همکاران (۱۹۸۵) داده شده است.

نظریهٔ جدید همگرایی و پایداری روشهای چندگامی، که در بخش ۸.۶ معرفی شدند از دال‌کوییست (۱۹۵۶) آغاز شده است. یک بیان تاریخی در دال‌کوییست (۱۹۸۵) داده شده است. کتاب درسی هنریچی (۱۹۶۲) به‌صورت یک گزارش کلاسیک این نظریه، شامل بسط و کاربرد آن درآمده است. کتاب گی‌یر (۱۹۷۱) یک گزارش جدیدتر از همهٔ این روشهاست، به‌ویژه روشهای متغیّر ـ مرتبه. اشتتر (۱۹۷۳) یک تحلیل مجرّد کامل و عمومی برای نظریهٔ عددی حل مسائل مقدار اولیه می‌دهد. یک گزارش کامل از روشهای رونگه ـ کوتا، بسط و تحلیل خطای آن روشها، تا سال ۱۹۷۰ توسط لاپیدوس و زاینفلت (۱۹۷۱) داده شده است. هال و وات (۱۹۷۶) یک بررسی از تمام جنبه‌های حل معادلات دیفرانسیل معمولی، شامل بسیاری از مباحث خاص کرده‌اند، که در ده سال اخیر مورد توجه بیشتر قرار گرفته‌اند.

اولین کاربرد قابل ملاحظهٔ مفهوم روش متغیّر ـ مرتبه از آنِ گی‌یر (۱۹۷۱) و کرو[9] (۱۹۶۹) است. چنین روشهایی نسبت به روش چندگامیِ ثابت مرتبه از نظر کارایی برتری دارند، و هیچ روش اضافی برای شروع انتگرالگیری یا برای تغییر طول گام نمی‌طلبند. یک بیان بسیار خوب از روش متغیر ـ مرتبهٔ آدامز در شمپاین و گوردن (۱۹۷۵)، شامل برنامهٔ عالی DE/STEP داده شده است. برنامه‌های مهم دیگر پیشین که بر پایهٔ فرمولهای خانوادهٔ آدامز، نهاده شده‌اند، برنامه‌هایی بوده‌اند در کرو (۱۹۶۹)، در DIFSUB از گی‌یر (۱۹۶۹) و در GEAR از هایندمارش[10] (۱۹۷۴) است. برنامهٔ اخیر گی‌یر توسعهٔ بیشتری یافته و به یک بستهٔ بزرگ چندتابعی تبدیل شده است که ODEPACK نامیده می‌شود که در هایندمارش (۱۹۸۳) توضیح داده شده است. انواع مختلف

1. Collatz 2. Milne 3. Lapidus 4. Seinfeld
5. Lambert 6. Stetter 7. Vander Houwen 8. Butcher
9. Krogh 10. Hindmarsh

این برنامه‌ها و حل‌کننده‌های دیگر معادلات دیفرانسیل درکتابخانه‌های IMSL و NAG موجودند.

روشهای رونگه ـ کوتا یک زمینهٔ فعال پژوهشی نظری و توسعهٔ برنامه‌یی هستند و یک توسعهٔ بسیارکلی در بوچر (۱۹۸۷) داده شده است. روشهای جدید برای مسائل غیرسرسخت در حال پدیدآمدن هستند؛ برای مثال شمپاین (۱۹۸۶) و شمپاین و باکا[1] (۱۹۸۶) را ببینید. همچنین علاقهٔ زیادی به روشهای ضمنی رونگه ـ کوتا برای حل معادلات دیفرانسیل سرسخت وجود دارد. برای مروری به این روشها ایکن (۱۹۸۵، صص ۷۰ـ۹۲) (۱۹۸۶) را ببینید. یک رقیب بسیار مهم برای برنامهٔ RKF45، برنامه DVERK است که در هال و همکاران (۱۹۷۶) توضیح داده شده است. این برنامه براساس طرحی است از نوع فلبرگ[2] با یک جفت فرمولهای مرتبهٔ ۵ و ۶.

دستهٔ سوم از روشها در این کتاب نادیده گرفته شده است، روشهایی که بر پایهٔ روش برونیابی استوارند. تحقیقات فعلی در این زمینه توسط گرو[3] (۱۹۶۸) و بولیرش و اشتور (۱۹۶۶) آغاز شده است. هدف اصلی، انجام برونیابی مکرر از یک روش ساده است تا رفته‌رفته روشهایی از مراتب بالاتر و بالاتر به‌دست آید. درواقع، این عمل یک راه‌دیگری برای ایجاد روشهای متغیر ـ مرتبه به‌دست خواهد داد. این روشها در آزمونهای انرایت و هال (۱۹۷۶)، شمپاین و همکاران (۱۹۷۶)، نسبتاً خوب اجرا شده‌اند، ولی داوری دربارهٔ آنها این بوده که از جنبهٔ نظری و عملی به اندازهٔ روشهای چندگامی و رونگه ـ کوتا پیشرفته نیستند. برای یک بازنگری تازه در این زمینه، دویفل‌هارت (۱۹۸۵) را ببینید. همچنین شمپاین و باکا (۱۹۸۶) را ببینید که در آن روشهای برونیابی به‌عنوان یک نمونه از روشهای متغیّر - مرتبهٔ رونگه ـ کوتا مورد بحث قرار گرفته است.

برآورد خطای کلی زمینه‌ای است که در آن مقالات نسبتاً کمی انتشار یافته‌اند. برای یک بررسی کلی اسکیل[4] (۱۹۸۶) را ببینید. تا آنجا که اطلاع داریم، تنها برنامهٔ رایانه‌یی متداول، برنامهٔ GERK از شمپاین و واتس (۱۹۷۶a) است. کارهای بیشتری در این زمینه لازم است. بسیاری ازکاربرهای بسته‌های خودکار تحت تاثیر این گمان اشتباه قرار دارند که برنامهٔ خودکاری که آنها به‌کار می‌برند خطای کلی را کنترل می‌نماید، ولی این کنترل خطای کلی از لحاظ عملی امکان ندارد. در بسیاری از موارد داشتن تصوّری از اندازهٔ واقعی خطای کلی موجود در جواب عددی قاعدتاً مهم است.

مسائل مقدار مرزی برای معادلات دیفرانسیل معمولی مبحث مهم دیگری است، ولی هم نظریهٔ عمومی آنها و هم تحلیل عددی آنها بسیار پیچیده‌تر از مسألهٔ مقدار اولیه‌اند. کتابهای درسی مهم کلر (۱۹۷۸) و (۱۹۶۹) هستند، وگزارشهای چایلد و همکاران (۱۹۷۹) چندین مقالهٔ مهم برای تهیهٔ برنامه‌های رایانه‌یی ارائه می‌نماید. برای مقالات دیگر مجموعهٔ عزیز (۱۹۷۵)، هال[5] و وات (۱۹۷۶) و اَشر[6] و راسل[7] (۱۹۸۵) را ببینید. این زمینه در مقایسه با مسألهٔ مقدار اولیه برای معادلات غیرسرسخت،

1. Baca 2. Fehlberg 3. Gragg 4. Skeel
5. Hall 6. Ascher 7. Russell

هنوز نسبتاً جدید است. مطالعهٔ برنامه‌های رایانه‌یی در چندین جهت ادامه دارد و بعضی برنامه‌های نسبتاً خوب در سالهای اخیر تهیه شده‌اند. کارهای زیادتری در برنامه‌نویسی با استفاده از روش پرتابی انجام گرفته است ولی برنامه‌های عملی برای روشهای هم‌مکانی و تفاضل متناهی در حال تهیه هستند. برای بحثی از این برنامه‌ها انرایت (۱۹۸۵) و گلادول و سِه‌یِرز (۱۹۸۰، صص ۲۷۳ـ۳۰۳) را ببینید. چند برنامهٔ مقدار مرزی در مجموعهٔ انتشارات IMSL و NAG داده شده‌اند.

معادلات دیفرانسیل سرسخت یکی از زمینه‌های خاصی است که در ده سال اخیر اهمیت خیلی بیشتری یافته‌اند. بهترین بررسی کلی در این زمینه در ایکن (۱۹۸۵) داده شده است. در این کتاب مثالهایی از چگونگی پیدایش این مسائل، نظریهٔ روشهای عددی برای حل مسائل سرسخت، و یک بررسی از برنامه‌های رایانه‌یی موجود برای حل این مسائل داده شده است. بسیاری از کتابهای درسی دیگری که در فهرست منابع آمده‌اند، همچنین به مسألهٔ معادلات دیفرانسیل سرسخت اشاره دارند. ما همچنین مطالعهٔ کتاب شمپاین و گی‌یِر (۱۹۷۹) را توصیه می‌کنیم.

معادلات با جواب خیلی نوسانی، در برخی از کاربردها ظاهر می‌شوند. برای بحث در مورد این مسائل، ایکن (۱۹۸۵، صص ۱۱۱ـ۱۲۳) را ببینید. روش خطوط، در حل معادلات دیفرانسیل با مشتقات جزیی وابسته به زمان، یک شیوهٔ کلاسیک است که در سالهای اخیر خیلی متداول گشته است. این روش در ایکن (۱۹۸۵، صص ۱۲۴ـ۱۲۸)، سینکاوک و مَدْسِن (۱۹۷۵) و ملخارت و سینکاوک (۱۹۸۱) موردبحث قرار گرفته است.

باز یک زمینهٔ دیگر قابل توجه، حل دستگاههای آمیختهٔ معادلات دیفرانسیل و جبری (DAE) است. این نام به دستگاههایی داده می‌شود که در آنها، n مجهول، $m < n$ معادلهٔ دیفرانسیل، و $n - m$ معادلهٔ جبری، شامل همان n تابع مجهول، وجود دارند. چنین مسائلی در زمینه‌های کاربردی بسیاری ظاهر می‌شوند. یک چنین زمینهٔ بسیار مورد توجه در سالهای اخیر طرح کمکی رایانه‌یی (CAD) است. برای مطالعهٔ مقالاتی که در مورد این مسائل و مسائل دیگر DAE کاربرد دارند به راینبولت[1] (۱۹۸۴) و (۱۹۸۶) مراجعه نمایید.

به‌دلیل ایجاد تعدادی برنامه‌های خودکار برای حل معادلات دیفرانسیل، مطالعات تجربی چندی انجام گرفته است تا اجرای این برنامه‌ها را برآورد کنند و آنها را با هم مقایسه نمایند. بعضی از مقایسه‌های مهم در انرایت و هال (۱۹۷۶)، انرایت و همکاران (۱۹۷۵) و شمپاین و همکاران (۱۹۷۶) داده شده‌اند. از کار آنها روشن می‌شود که برنامه‌ها و همچنین روشها باید مقایسه شوند. کاربردهای مختلف یک برنامه برای یک روش، ممکن است تفاوتهای عمده‌ای در اجرا داشته باشد. هیچ نتایج مشابهی در مقایسهٔ برنامه‌های مقدار مرزی در دست نیست.

1. Rheinboldt

مراجع

Aiken, R., Ed. (1985). *Stiff Computation*. Oxford Univ. Press, Oxford, England.

Ascher, U. (1986). Collocation for two-point boundary value problems revisited, *SIAM J. Numer. Anal.* **23**, 596–609.

Ascher, U., J. Christiansen, and R. Russell (1981a). Collocation software for boundary-value ODEs, *ACM Trans. Math. Softw.* **7**, 209–222.

Ascher, U., J. Christiansen, and R. Russell (1981b). COLSYS: collocation software for boundary-value ODEs, *ACM Trans. Math. Softw.* **7**, 223–229.

Ascher, U., and R. Russell, Eds. (1985). *Numerical Boundary Value ODEs*. Birkhäuser, Basel.

Aziz, A. K., Ed. (1975). *Numerical Solutions of Boundary Value Problems for Ordinary Differential Equations*. Academic Press, New York.

Boyce, W., and R. Diprima (1986). *Elementary Differential Equations and Boundary Value Problems*, 4th ed. Wiley, New York.

Bulirsch, R., and J. Stoer (1966). Numerical treatment of ordinary differential equations by extrapolation, *Numer. Math.* **8**, 1–13.

Butcher, J. (1965). On the attainable order of Runge–Kutta methods, *Math. Comput.* **19**, 408–417.

Butcher, J. (1987). *The Numerical Analysis of Ordinary Differential Equations*. Wiley, New York.

Childs, B., M. Scott, J. Daniel, E. Denman, and P. Nelson, Eds. (1979). *Codes for Boundary-Value Problems in Ordinary Differential Equations*. Lecture Notes in Computer Science 76, Springer-Verlag, New York.

Coddington, E., and N. Levinson (1955). *Theory of Ordinary Differential Equations*. McGraw-Hill, New York.

Collatz, L. (1966). *The Numerical Treatment of Differential Equations*, 3rd ed. Springer-Verlag, New York.

Dahlquist, G. (1956). Numerical integration of ordinary differential equations, *Math. Scandinavica* **4**, 33–50.

Dahlquist, G. (1963). A special stability property for linear multistep methods, *BIT* **3**, 27–43.

Dahlquist, G. (1985). 33 years of numerical instability, part 1, *BIT* **25**, 188–204.

Deuflhard, P. (1979). Nonlinear equation solvers in boundary value problem codes. In B. Childs, M. Scott, J. Daniel, E. Denman, and P. Nelson (Eds.), *Codes for Boundary-Value Problems in Ordinary Differential Equations*, pp. 40–66. Lecture Notes in Computer Science 76, Springer-Verlag, New York.

Deuflhard, P. (1985). Recent progress in extrapolation methods for ordinary differential equations, *SIAM Rev.* **27**, 505–536.

Enright, W. (1985). Improving the performance of numerical methods for two-point boundary value problems. In U. Ascher and R. Russell (Eds.), *Numerical Boundary Value ODEs*, pp. 107–120. Birkhäuser, Basel.

Enright, W., and T. Hull (1976). Test results on initial value methods for non-stiff ordinary differential equations, *SIAM J. Numer. Anal.* **13**, 944–961.

Enright, W., T. Hull, and B. Lindberg (1975). Comparing numerical methods for stiff systems of O.D.E.'s, *BIT* **15**, 10–48.

Fehlberg, E. (1970). Klassische Runge–Kutta–Formeln vierter und niedrigerer Ordnumg mit Schrittweiten–Kontrolle und ihre Anwendung auf Wärmeleitungsprobleme, *Computing* **6**, 61–71.

Forsythe, G., M. Malcolm, and C. Moler (1977). *Computer Methods for Mathematical Computations*. Prentice-Hall, Englewood Cliffs, N.J.

Fox, L. (1980). Numerical methods for boundary-value problems. In I. Gladwell and D. Sayers (Eds.), *Computational Techniques for Ordinary Differential Equations*, pp. 175–217. Academic Press, New York.

Gear, C. W. (1971). *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice-Hall, Englewood Cliffs, N.J.

Gladwell, I., and D. Sayers, Eds. (1980). *Computational Techniques for Ordinary Differential Equations*. Academic Press, New York.

Gragg, W. (1965). On extrapolation algorithms for ordinary initial value problems, *SIAM J. Numer. Anal.* **2**, 384–403.

Gupta, G., R. Sacks-Davis, and P. Tischer (1985). A review of recent developments in solving ODEs, *Comput. Surv.* **17**, 5–47.

Hall, G., and J. Watt, Eds. (1976). *Modern Numerical Methods for Ordinary Differential Equations*. Oxford Univ. Press, Oxford, England.

Henrici, P. (1962). *Discrete Variable Methods in Ordinary Differential Equations*. Wiley, New York.

Hindmarsh, A. (1974). GEAR: Ordinary differential equation solver. Lawrence Livermore Rep. UCID-30001, Rev. 3, Livermore, Calif.

Hindmarsh, A. (1983). ODEPACK: A systematized collection of ODE solvers. In *Numerical Methods for Scientific Computation*, R. Stepleman, Ed. North-Holland, Amsterdam.

Hull, T., W. Enright, and K. Jackson (1976). User's guide for DVERK: A subroutine for solving non-stiff ODEs. Dept. Computer Sci. Tech. Rep. 100, Univ. of Toronto, Toronto, Ont., Canada.

Isaacson, E., and H. Keller (1966). *Analysis of Numerical Methods*. Wiley, New York.

Jain, M. K. (1984). *Numerical Solution of Differential Equations*, 2nd ed. Halstead Press, New York.

Keller, H. (1968). *Numerical Methods for Two-Point Boundary Value Problems*. Ginn (Blaisdell), Boston.

Keller, H. (1976). *Numerical Solution of Two-Point Boundary Value Problems*. Regional Conference Series in Applied Mathematics 24, Society for Industrial and Applied Mathematics, Philadelphia.

Keller, J., and S. Antman, Eds. (1969). *Bifurcation Theory and Nonlinear Eigenvalue Problems*. Benjamin, New York.

Krogh, F. (1969), VODQ/SVDQ/DVDQ—Variable order integrators for the numerical solution of ordinary differential equations. Section 314 Subroutine Writeup, Jet Propulsion Lab., Pasadena, Calif.

Lambert, J. (1973). *Computational Methods in Ordinary Differential Equations*. Wiley, New York.

Lapidus, L., and W. Schiesser, Eds. (1976). *Numerical Methods for Differential Equations: Recent Developments in Algorithms, Software*. New Applications, Academic Press, New York.

Lapidus, L., and J. Seinfeld (1971). *Numerical Solution of Ordinary Differential Equations*. Academic Press, New York.

Lentini, M., M. Osborne, and R. Russell (1985). The close relationship between methods for solving two-point boundary value problems, *SIAM J. Numer. Anal.* **22**, 280–309.

Melgaard, D., and R. Sincovec (1981). Algorithm 565: PDETWO/PSETM/GEARB: Solution of systems of two-dimensional nonlinear partial differential equations, *ACM Trans. Math. Softw.* **7**, 126–135.

Miller, R. K. and A. Michel (1982). *Ordinary Differential Equations*. Academic Press, New York.

Milne, W. (1953). *Numerical Solution of Differential Equations*. Wiley, New York.

Ortega, J., and W. Poole (1981). *An Introduction to Numerical Methods for Differential Equations*, Pitman, New York.

Pereyra, V. (1979). PASVA3: An adaptive finite difference Fortran program for first order nonlinear, ordinary differential equation problems. In B. Childs, M. Scott, J. Daniel, E. Denman, and P. Nelson (Eds.), *Codes for Boundary-Value Problems in Ordinary Differential Equations*, pp. 67–88. Lecture Notes in Computer Science 76, Springer-Verlag, New York.

Reddien, G. (1979). Projection methods. In B. Childs, M. Scott, J. Daniel, E. Denman, and P. Nelson (Eds.), *Codes for Boundary-Value Problems in Ordinary Differential Equations*, pp. 206–227. Lecture Notes in Computer Science 76, Springer-Verlag, New York.

Rheinboldt, W. (1984). Differential-algebraic systems as differential equations on manifolds, *Math. Comput.* **43**, 473–482.

Rheinboldt, W. (1986). *Numerical Analysis of Parametrized Nonlinear Equations*. Wiley, New York.

Shampine, L. (1985). Local error estimation by doubling, *Computing* **34**, 179–190.

Shampine, L. (1986). Some practical Runge–Kutta formulas, *Math. Comput.* **46**, 135–150.

Shampine, L., and L. Baca (1986). Fixed versus variable order Runge–Kutta, *ACM Trans. Math. Softw.* **12**, 1–23.

Shampine, L., and C. W. Gear (1979). A user's view of solving stiff ordinary differential equations, *SIAM Rev.* **21**, 1–17.

Shampine, L., and M. Gordon (1975). *Computer Solution of Ordinary Differential Equations*. Freeman, San Francisco.

Shampine, L., and H. Watts (1976a). Global error estimation for ordinary differential equations, *ACM Trans. Math. Softw.* **2**, 172–186.

Shampine, L., and H. Watts (1976b). Practical solution of ordinary differential equations by Runge–Kutta methods. Sandia Labs. Tech. Rep. SAND 76-0585, Albuquerque, N.Mex.

Shampine, L., and H. Watts (1980). DEPAC—Design of a user oriented package of ODE solvers. Sandia National Labs. Rep. SAND79-2374, Albuquerque, N.Mex.

Shampine, L., H. Watts, and S. Davenport (1976). Solving nonstiff ordinary differential equations—The state of the art, *SIAM Rev.* **18**, 376–411.

Sincovec, R., and N. Madsen (1975). Software for nonlinear partial differential equations, *ACM Trans. Math. Softw.* **1**, 232–260.

Skeel, R. (1986). Thirteen ways to estimate global error, *Numer. Math.* **48**, 1–20.

Stakgold, I. (1979). *Green's Functions and Boundary Value Problems*. Wiley, New York.

Stetter, H. (1973). *Analysis of Discretization Methods for Ordinary Differential Equations*. Springer-Verlag, New York.

Stoer, J., and R. Bulirsch (1980). *Introduction to Numerical Analysis*. Springer-Verlag, New York.

Van der Houwen, P. J. (1977). *Construction of Integration Formulas for Initial Value Problems*. North-Holland, Amsterdam.

Widder, D. (1975). *The Heat Equation*. Academic Press, New York.

مسائل

۱. میدان سودار $y' = x - y^۲$ را رسم کنید، سپس یک نمونه از منحنیهای جواب را در آن رسم، و سعی کنید که رفتار جواب $Y(x)$ را وقتی $x \to \infty$، حدس بزنید.

۲. ثابتهای لیپشیتس را برای توابع زیر مانند (۲.۱.۶) معین کنید.

(الف) $f(x,y) = ۲y/x \quad x \geq ۱$

(ب) $f(x,y) = \tan^{-۱}(y)$

(ج) $f(x,y) = (x^۳ - ۲)^{۲۷}/(۱۷x^۲ + ۴)$

(د) $f(x,y) = x - y^۲$، $| y | \leq ۱۰$

۳. مسائل زیر را به دستگاه معادلات مرتبهٔ اول تبدیل کنید.

(الف) $y'' - ۳y' + ۲y = ۰$، $y(۰) = ۱$، $y'(۰) = ۱$

(ب) $(۱ - y^۲)y' + y = ۰$، $y'' - ۰٫۱(۱ - y^۲)y' + y = ۰$، $y(۰) = ۱$، $y'(۰) = ۰$ (معادلهٔ واندرپول)

(ج) $x'' = -x/r^۳$، $y'' = -y/r^۳$، $r = \sqrt{x^۲ + y^۲}$ (معادلهٔ مداری)، $x(۰) = ۴$، $x'(۰) = ۰$، $y(۰) = ۰$، $y'(۰) = ۲$

۴. گیریم $Y(x)$، جواب مسألهٔ مقدار اولیهٔ (۱.۰.۶)، در صورت وجود، باشد. با انتگرالگیری نشان دهید که Y در رابطهٔ زیر صدق می‌کند

$$Y(x) = Y_0 + \int_{x_0}^{x} f(t, Y(t))dt$$

به عکس، نشان دهید که اگر این معادله یک جواب پیوسته بر بازهٔ $x_0 \le x \le b$ داشته باشد، مسألهٔ مقدار اولیهٔ (۱.۰.۶) نیز همان جواب را دارد.

۵. معادلهٔ انتگرالی مسألهٔ ۴، حداقل به‌طور نظری، با استفاده از بارست

$$Y_{m+1}(x) = Y_0 + \int_{x_0}^{x} f(t, Y_m(t))dt \qquad x_0 \le x \le b$$

برای $m \ge 0$ با $Y_0(x) \equiv Y_0$ حل شده است. این بارست را **بارست پیکار** (Picard) گویند، و با مفروضات مناسبی، می‌توان نشان داد که بارستهای $\{Y_m(x)\}$ به‌طور یکنواخت به $Y(x)$ همگرا هستند. با محاسبهٔ بارستهای Y_1، Y_2، Y_3 ی پیکار این حکم را برای مسائل زیر روشن کنید و آنها را با جواب درست Y مقایسه کنید.

(الف) $y' = -y$، $y(0) = 1$

(ب) $y' = -xy$، $y(1) = 2$

(ج) $y' = y + 2\cos(x)$، $y(0) = 1$

۶. یک برنامهٔ رایانه‌یی بنویسید که معادلهٔ $y' = f(x, y)$، $y(x_0) = y_0$ را با روش اویلر حل کند. آن را به گونه‌ای بنویسید که برای تابع دلخواه f، طول گام h، و بازهٔ $[x_0, b]$ بتوان به‌کار برد. با استفاده از این برنامه معادلهٔ $y' = x^2 - y$، $y(0) = 1$ را برای $0 \le x \le 4$، با طول گامهای $h = 0.25, 0.125, 0.0625$ به توالی حل کنید. برای هر مقدار h، جواب درست، جواب تقریبی، خطا، خطای نسبی را در نقاط گرهی $x = 0, 0.25, 0.50, 0.75, \ldots, 4.00$ را چاپ کنید. جواب درست چنین است: $Y(x) = x^2 - 2x + 2 - e^{-x}$. خروجی برنامهٔ خود را تحلیل کنید و اظهارنظرهای کتبی برای آن بنویسید. تحلیل خروجی به اندازهٔ به‌دست آوردن آن مهم است.

۷. برای مسألهٔ $y' = y$، $y(0) = 1$، جواب صریح $\{y_n\}$ را برای تقریب اویلر این معادله بنویسید. با استفاده از آن نشان دهید که برای $x_n = 1$، وقتی $h \to 0$، داریم $Y(1) - y_n \doteq (h/2)e$.

۸. حل مسألهٔ زیر را با روش اویلر در نظر می‌گیریم

$$y' = -y^2 \qquad x \ge 0 \qquad y(0) = 1$$

کران (۱۳.۲.۶) را با به‌کاربردن $K = ۲$ به‌عنوان ثابت لیپشیتس، با برآورد مجانبی حاصل از (۳۶.۲.۶) مقایسه کنید. جواب درست $Y(x) = ۱/(۱ + x)$ است.

۹. نشان دهید که روش اویلر در تقریب‌زدن جواب $Y(x) = (\frac{۲}{۳}x)^{۳/۲}$، $x \geq ۰$، در مسألهٔ $y' = y^{۱/۳}$، $y(۰) = ۰$ به شکست می‌انجامد. توضیح دهید چرا.

۱۰. برای معادلات مسألهٔ ۳ معادلات تفاضلی تقریب زننده را که با استفاده از روش اویلر به‌دست می‌آیند بنویسید.

۱۱. مثال خطای گردکردن به روش اویلر را با نتایجی که در جدول ۳.۶ نشان داده شده است به یاد آورید. سعی کنید رفتار مشابهی را در رایانهٔ خود با کاهش تدریجی h ایجاد کنید تا اینکه خطا روبه افزایش گذارد.

۱۲. مسألهٔ زیر را به یک دستگاه معادلات مرتبهٔ اول تبدیل کنید.

$$y^{(۳)} + ۴y'' + ۵y' + ۲y = -۴\sin(x) - ۲\cos(x)$$
$$y(۰) = ۱ \quad y'(۰) = ۰ \quad y''(۰) = -۱$$

با استفاده از روش اویلر این دستگاه را حل و از راه تجربی، خطا را مطالعه کنید. جواب درست $Y = \cos x$ است.

۱۳. (الف) شرط (۵۲.۲.۶) لیپشیتس را برای یک دستگاه دو معادلهٔ دیفرانسیل پیدا کنید.
(ب) ثابت کنید که روش (۵۱.۲.۶) به جواب (۵۰.۲.۶) همگراست.

۱۴. روش دوگامی زیر را در نظر می‌گیریم

$$y_{n+۱} = \frac{۱}{۲}(y_n + y_{n-۱}) + \frac{h}{۴}[۴y'_{n+۱} - y'_n + ۳y'_{n-۱}] \qquad n \geq ۱$$

که در آن $y'_n \equiv f(x_n, y_n)$. نشان دهید که این یک روش مرتبهٔ دو است و جملهٔ پیشرو در خطای برشی را که به‌صورت (۱۵.۳.۶) نوشته شده پیدا کنید.

۱۵. فرض می‌کنیم که روش چندگامی (۱.۳.۶) سازگار است و همانند قضیهٔ ۶.۶ در $a_j \geq ۰$، $j = ۰, ۱, \ldots, p$ صدق می‌کند. پایداری (۱.۳.۶) را مشابه با (۲۸.۲.۶)، برای روش اویلر ثابت کنید، ولی δ را متحّد با صفر بگیرید.

۱۶. برنامه‌ای برای حلّ $y' = f(x, y)$، $y(x_۰) = y_۰$ با استفاده از قاعدهٔ میانگاهی (۲.۴.۶) بنویسید. از طول گام ثابت h استفاده کنید. برای مقدار اولیهٔ $y_۱$، روش اویلر را به‌کار برید:

$$y_۱ = y_۰ + hf(x_۰, y_۰)$$

با این برنامه مسائل زیر را حل کنید:

(الف) $y' = -y^2$ ، $y(0) = 1$؛ $Y(x) = 1/(1+x)$.

(ب) $y' = \frac{y}{4}\left[1 - \frac{y}{20}\right]$ ، $y(0) = 1$؛ $Y(x) = \frac{20}{[1 + 19e^{-x/4}]}$.

(ج) $y' = -y + 2\cos(x)$ ، $y(0) = 1$؛ $Y(x) = \cos(x) + \sin(x)$.

(د) $y' = y - 2\sin(x)$ ، $y(0) = 1$ ، $Y(x) = \cos(x) + \sin(x)$

برای بازهٔ $[x_0, b] = [0, 5]$ با $h = 0.25$، 0.5 حل کنید. جواب عددی $y_n = y_h(x_n)$ را در هر نقطهٔ گرهی همراه با خطای درست چاپ کنید. دربارهٔ نتایج خود بحث کنید.

۱۷. برنامه‌ای برای حلّ $y' = f(x,y)$ ، $y(x_0) = y_0$ با استفاده از قاعدهٔ ذوزنقه‌یی (۲.۵.۶) با طول گام ثابت h بنویسید. در بارست (۳.۵.۶)، روش میانگاهی را به‌عنوان پیشگوی $y_{n+1}^{(0)}$ به‌کار برید. تعداد بارستهای J را یک متغیر ورودی بگیرید، $J \geq 1$. معادلات مسألهٔ ۱۶ را با $h = 0.5$ و $h = 0.25$ برای $[x_0, b] = [0, 10]$ حل کنید. ابتدا با $J = 1$ برای تمامی بازه حل کنید سپس فرایند را با $J = 2$ و $J = 3$ تکرارکنید. دربارهٔ نتایج بحث کنید. چگونه خطای کلی با J تغییر می‌کند؟

۱۸. فرمول خطای مجانبی (۱۹.۵.۶) را برای روش ذوزنقه‌یی به‌دست آورید.

۱۹. یک برنامه بنویسید که الگوریتم $Detrap$ از بخش ۶.۶ را اجرا نماید. با استفاده از این برنامه معادلات داده‌شده در مسألهٔ ۱۶ را با $\varepsilon = 0.001$ حل کنید.

۲۰. درونیاب درجهٔ دوّم را برای $Y'(x) = f(x, Y(x))$ در x_n، x_{n-1}، x_{n-2} به‌کار برید تا فرمول زیر به‌دست آید

$$Y(x_{n+1}) = Y_{n-3} + \frac{4h}{3}[2Y'_n - Y'_{n-1} + 2Y'_{n-2}] + \frac{28}{90}h^5 Y^{(5)}(\xi_n)$$

وقتی خطای برشی حذف شود روش

$$y_{n+1} = y_{n-3} + \frac{4h}{3}[2y'_n - y'_{n-1} + 2y'_{n-2}] \qquad n \geq 3$$

را به‌دست می‌آوریم. این فرمول پیشگو برای روش میلن، (۱۳.۷.۶)، است.

۲۱. نشان دهید که روش (۱۳.۷.۶) سیمپسون تنها ضعیف ـ پایدار است، به همان تعبیری که در روش میانگاهی در بخش ۴.۶ آمده است.

۲۲. برنامه‌ای برای حلّ $y' = f(x,y)$ ، $y(x_0) = y_0$ ، $x_0 \leq x \leq b$ با استفاده از فرمول مرتبهٔ چهار آدامز ـ مولتن با طول گام ثابت h بنویسید. از فرمول مرتبهٔ چهار آدامز ـ بشفورت به‌عنوان پیشگو استفاده کنید. مقادیر اولیهٔ y_1، y_2، y_3 را با استفاده از جواب درست مسأله تولید کنید.

معادلات مسألهٔ ۱۶ را با $h = ۰٫۵$ و $h = ۰٫۲۵$ حل کنید. جوابهای به‌دست‌آمده و خطاهای درست را چاپ کنید. برای مقایسه با روش (۲۷.۷.۶)، مثال جدول ۱۴.۶ را نیز حل کنید. دقت برونیابی ریچاردسن را برای براورد خطای

$$Y(x) - y_h(x) \doteq \frac{۱}{۱۵}[y_h(x) - y_{۲h}(x)]$$

که بر پایهٔ خطای کلی از مرتبهٔ چهار استوار است کنترل نمایید. دربارهٔ تمام نتایج خود بحث کنید.

۲۳. **(الف)** برای ضرایب γ_i و δ_i از فرمولهای آدامز ـ بشفورت و آدامز ـ مولتن نشان دهید که $\delta_i = \gamma_i - \gamma_{i-۱}$، $i \geq ۱$.

(ب) برای فرمول p گامی (۲۶.۷.۶) آدامز ـ مولتن، ثابت کنید

$$y_{n+۱} = y^{(۰)}_{n+۱} + h\gamma_p \nabla^{p+۱} y'_{n+۱} \qquad (۱)$$

که $y^{(۰)}_{n+۱}$ از فرمول $p+۱$ گامی (۲۲.۷.۶) آدامز ـ بشفورت

$$y^{(۰)}_{n+۱} = y_n + h\sum_{j=۰}^{p} \gamma_j \nabla^j y'_n \qquad (۲)$$

است. این فرمولها هر دو از مرتبهٔ $p+۱$ هستند. فرمول (۱) در به‌دست آوردن تصحیح‌کننده از پیشگو به‌کار می‌آید، و بر پایهٔ اجرای تفاضلات پسرو معینی از یک مرحله به مرحلهٔ بعدی مبتنی است. یک نتیجهٔ خیلی زیاد وابسته وجود دارد وقتی که یک پیشگوی p ـ گامی (مرتبهٔ p) برای حل یک تصحیح‌کنندهٔ p گامی (مرتبهٔ $p+۱$) به‌کار برده می‌شود. [شمپاین وگوردن (۱۹۷۵، ص ۵۱) را ببینید].

۲۴. معادلهٔ نمونهٔ (۱۶.۸.۶) را که در آن Λ یک ماتریس مربعی مرتبهٔ m است در نظر می‌گیریم. فرض کنید که $\Lambda = P^{-۱}DP$، که D یک ماتریس قطری با درایه‌های $\lambda_۱$، ...، λ_m است. تابع برداری مجهول جدید $z = P^{-۱}y(x)$ را وارد می‌کنیم. نشان دهید (۱۶.۸.۶) به شکلی که در (۱۷.۸.۶) داده‌شده تبدیل می‌شود، که نشان‌دهندهٔ تبدیل به یک معادلهٔ نمونهٔ یک‌بعدی است.

۲۵. با پیروی از مفاهیم بخشهای ۴.۶ و ۸.۶، یک شیوهٔ کلی برای حل معادلهٔ تفاضلی خطی زیر ارائه دهید

$$y_{n+۱} = a_۰ y_n + a_۱ y_{n-۱}$$

این روش را برای حل معادلات زیر به‌کار برید

(الف) $y_{n+1} = -\frac{1}{2}y_n + \frac{1}{2}y_{n-1}$

(ب) $y_{n+1} = y_n - \frac{1}{4}y_{n-1}$.

راهنمایی: فرمولی را که به دنبال (۲۲.۸.۶) آمده است ببینید.

۲۶. معادلهٔ تفاضلی خطی مرتبهٔ سهٔ

$$u_{n+1} = u_n + c(u_{n-1} - u_{n-2}) \qquad n \geq 2 \qquad 0 < c < 1$$

را که u_0، u_1، u_2 داده شده‌اند، حل کنید. دربارهٔ $\lim_{n\to\infty} u_n$ چه می‌توان گفت؟

۲۷. روش عددی زیر را در نظر می‌گیریم

$$y_{n+1} = 4y_n - 3y_{n-1} - 2hf(x_{n-1}, y_{n-1}) \qquad n \geq 1$$

مرتبهٔ آن را معین کنید. با یک مثال نشان دهید که این روش ناپایدار است.

۲۸. نشان دهید که روش دوگامی

$$y_{n+1} = 2y_{n-1} - y_n + h[\frac{5}{2}y'_n + \frac{1}{2}y'_{n-1}]$$

از مرتبهٔ ۲ و ناپایدار است. همچنین مستقیماً نشان دهید که در حل $y' = f(x, y)$ ممکن است همگرا نباشد.

۲۹. قسمت (۱) برهان قضیهٔ ۷.۶ را، که در آن شرط ریشه با فرض $|r_j| \leq 1$ و $\rho'(r_j) = 0$ برای مقداری از j، نقض شده است کامل کنید.

۳۰. برای قسمت (۲) از برهان قضیهٔ ۸.۶ نشان دهید که هرگاه $h \to 0$ و $1 \leq j \leq p$، آنگاه $\gamma_j[r_j(h\lambda)]^n \to 0$.

۳۱. (الف) مقادیر a_0 را در روش صریح مرتبهٔ دوّم (۱۰.۷.۶) که برای آن این روش پایدار است معین کنید.

(ب) اگر فقط خطای برشی (۱۱.۷.۶) در نظرگرفته شود، با رعایت محدودیت پایداری در قسمت (الف)، چگونه a_0 باید انتخاب شود؟

(ج) برای تامین یک ناحیهٔ بزرگ پایداری، با رعایت قسمت (الف)، a_0 چگونه باید انتخاب شود؟

۳۲. (الف) فرمول کلی را برای تمام روشهای دوگامی مرتبهٔ ۳ به‌دست آورید. این فرمولها یک خانواده از روشهای تک ـ پارامتری، مثلاً وابسته به ضریب a_1، خواهند بود.

(ب) برای پایداربودن روش قسمت (الف) چه محدودیتهایی برای a_1 باید وجود داشته باشد؟

(ج) اگر خطای برشی به شکل

$$T_{n+1}(Y) = \beta h^{4} Y^{(4)}(x_n) + O(h^5)$$

نوشته شود فرمولی برای β برحسب a_1 به‌دست آورید. (لازم نیست که هستهٔ پئانو یا تابع اثر این روش را بسازید). a_1 چگونه انتخاب شود تا با رعایت محدودیت پایداری قسمت (ب)، خطای برشی مینیمم شود؟

(د) ناحیهٔ پایداری مطلق برای روشهای قسمت (الف) را در نظر می‌گیریم. این ناحیه برای روش قسمت (ج) چه باشد تا ضریب β ی خطای برشی به حداقل برسد؟ مقدار دیگری از a_1 بدهید که روش پایدار به‌دست دهد و ناحیهٔ پایداری مطلق بزرگتری داشته باشد. دربارهٔ انتخاب تقریبی a_1 برای یک ناحیهٔ بهینه بحث کنید.

۳۳. (الف) تمام فرمولهای صریح مرتبهٔ چهار به شکل زیر را پیدا کنید

$$y_{n+1} = a_0 y_n + a_1 y_{n-1} + a_2 y_{n-2} + h[b_0 y'_n + b_1 y'_{n-1} + b_2 y'_{n-2}] \qquad n \geq 2$$

(ب) نشان دهید که هر چنین روشی ناپایدار است.

۳۴. یک روش چندگامی ضمنی مرتبهٔ چهار غیر از آنچه درکتاب داده شده است به‌دست آورید. آن را نسبی ـ پایدار سازید.

۳۵. برای چندجمله‌یی $\rho(r) = r^{p+1} - \sum_{0}^{p} a_j r^{p-j}$، فرض کنید $a_j \geq 0$ ، $0 \leq j \leq p$، و $\sum_{0}^{p} a_j = 1$. نشان دهید که ریشه‌های $\rho(r)$ در شرط ریشهٔ (۹.۸.۶) و (۱۰.۸.۶) صدق می‌نماید. این مستقیماً نشان می‌دهد که قضیهٔ ۶.۶ یک فرع قضیهٔ ۸.۶ است.

۳۶. (الف) روشهایی به شکل زیر را در نظر می‌گیریم

$$y_{n+1} = y_{n-q} + h \sum_{j=-1}^{p} b_j f(x_{n-j}, y_{n-j})$$

با $q \geq 1$. نشان دهید که این روشها در شرط قوی ریشه صدق نمی‌نمایند. درنتیجه، اغلب چنین روشهایی فقط ضعیف ـ پایدارند.

(ب) یک مثال با $q = 1$ پیدا کنید که نسبتاً ـ پایدار باشد.

۳۷. نشان دهید که ناحیهٔ پایداری مطلق روش ذوزنقه‌یی مجموعهٔ تمام کمیتهای مختلط $h\lambda$ با $\text{Real}(\lambda) < 0$ است.

۳۸. روش پسرو اویلر را برای حل مسألهٔ (۵۱.۸.۶) به‌کار برید. چون این معادله خطی است، معادلهٔ ضمنی برای y_{n+1} را می‌توان دقیقاً حل کرد. نتایج خود را با آنچه در جدول ۱۷.۶ برای روش اویلر داده‌شده مقایسه کنید.

۳۹. مسألهٔ ۳۸ را با استفاده از فرمول مرتبهٔ ۲ (۶.۹.۶) BDF تکرار کنید. برای پیداکردن y_1 جهت استفاده در (۶.۸.۶) از روش پسرو اویلر استفاده کنید.

۴۰. معادلهٔ نمونهٔ (۵۰.۸.۶) را که به‌عنوان یک اختلال از روش (۴۶.۸.۶) اویلر در نظرگرفته شده است به‌یاد آورید. برای حالت خاص ثابت $Y''(x) \equiv$، رفتار خطای $Y(x_n) - y_n$ را که به $h\lambda$ وابسته است تحلیل نمایید. نشان دهید باز هم برای خوشرفتاربودن خطا شرط (۴۹.۸.۶) موردنیاز است.

۴۱. فرمول خطای برشی (۷.۹.۶) را برای فرمولهای مشتقگیری پسرو به‌دست آورید.

۴۲. برای حل $y' = f(x,y)$ روش عددی زیر را در نظر می‌گیریم

$$y_{n+1} = y_n + \frac{h}{2}[y'_n + y'_{n+1}] + \frac{h^2}{12}[y''_n - y''_{n+1}] \qquad n \geq 0$$

در اینجا $y'_n = f(x_n, y_n)$،

$$y_n'' = \frac{\partial f(x_n, y_n)}{\partial x} + f(x_n, y_n)\frac{\partial f(x_n, y_n)}{\partial y}$$

که این فرمول براساس مشتق‌گیری از $Y'(x) = f(x, Y(x))$ به‌دست آمده است.

(الف) نشان دهید که این روش یک روش مرتبهٔ چهار است: $T_n(Y) = O(h^5)$.

(ب) نشان دهید که ناحیهٔ پایداری مطلق شامل تمام محور حقیقی منفی از صفحهٔ مختلط $h\lambda$ است.

۴۳. روش خطوط را که در (۲۳.۹.۶) ـ (۲۵.۹.۶) داده‌شده برای مسألهٔ زیر تعمیم دهید:

$$U_t = a(x,t)U_{xx} + G(x,t,U(x,t)) \qquad 0 < x < 1 \qquad t > 0$$

$$U(0,t) = d_0(t) \qquad U(1,t) = d_1(t) \qquad t \geq 0$$

$$U(x,0) = f(x) \qquad 0 \leq x \leq 1$$

برای آنکه مسأله خوش‌ـ تعریف باشد فرض می‌کنیم $a(x,t) > 0$، $0 \leq x \leq 1$ و $t \geq 0$.

۴۴. (الف) اگر یک حلاّل دستگاه جبری خطی سه قطری در اختیار داشته باشید، برنامه‌ای بنویسید که روش خطوط را برای مسألهٔ (۱۹.۹.۶) ـ (۲۱.۹.۶) اجرا کند. مثال کتاب با مجهول (۳۵.۹.۶) با روش پسرو اویلر حل شده است. اکنون روش خطوط را با استفاده از قاعدهٔ ذوزنقه‌یی اجرا کنید و نتایج خود را با آنچه در جدول ۲۰.۶ برای روش پسرو اویلر داده‌شده مقایسه نمایید.

(ب) با روش BDF مرتبهٔ دوم تکرار کنید.

۴۵. یک روش مرتبهٔ سه سری تیلر برای حل $y' = -y^2$ به‌دست آورید. نتایج عددی را با نتایج جدول ۲۲.۶ مقایسه نمایید.

۴۶. با استفاده از روش سری تیلر بخش ۱۰.۶، یک روش مرتبهٔ چهار برای حل $y' = x - y^2$، $y(0) = 0$، تولید کنید. طول گامهای ثابت $h = 0.5$، $h = 0.25$ و $h = 0.125$ را پشت‌سرهم

به‌کار برده مسأله را برای $0 \le x \le 10$ حل کنید. خطای کلی را با براورد خطای (۲۴.۱۰.۶) که بر پایهٔ برونیابی ریچاردسن استوار است براورد کنید.

۴۷. برنامه‌ای برای حل $y' = f(x,y)$، $y(x_0) = y_0$، با استفاده از روش کلاسیک (۲۱.۱۰.۶) رونگه ـ کوتا بنویسید و طول گام h را ثابت بگیرید.

(الف) با استفاده از این برنامه، معادلات مسألهٔ ۱۶ را حل کنید.

(ب) معادلهٔ $y' = x - y^2$، $y(0) = 0$، را برای $h = 0.5, 0.25, 0.125$ حل کنید. نتایج را با نتایج مسألهٔ ۴۶ مقایسه کنید.

۴۸. فرمول سه‌گامی رونگه ـ کوتا را که ذیلاً داده شده در نظر می‌گیریم

$$y_{n+1} = y_n + h[\gamma_1 V_1 + \gamma_2 V_2 + \gamma_3 V_3]$$
$$V_1 = f(x_n, y_n), \quad V_2 = f(x_n + \alpha_2 h, y_n + h\beta_{21}, V_1)$$
$$V_3 = f(x_n + \alpha_3 h, y_n + h(\beta_{31} V_1 + \beta_{32} V_2))$$

مجموعهٔ معادلاتی را که ضرایب $\{\gamma_j, \alpha_j, \beta_j\}$ باید در آنها صدق کنند تا این فرمول از مرتبهٔ ۳ باشد به‌دست آورید. یک جواب خاص این معادلات را پیدا کنید.

۴۹. ثابت کنید که اگر روش (۴.۱۰.۶) رونگه ـ کوتا در رابطهٔ (۲۷.۱۰.۶) صدق کند، روش پایدار است.

۵۰. روش کلاسیک (۲۱.۱۰.۶) رونگه ـ کوتا را برای مسألهٔ آزمونی (۵۱.۸.۶) برای مقادیر چندی از λ و h به‌کار برید. برای مثال، $\lambda = -1, -10, -50$ و $h = 0.5, 0.1, 0.01$ را مانند جدول ۱۷.۶ امتحان کنید.

۵۱. قسمت حقیقی ناحیهٔ پایداری مطلق را برای روش (الف) (۸.۱۰.۶)، (ب) (۹.۱۰.۶) (ج) (۲۱.۱۰.۶) رونگه ـ کوتا به‌دست آورید. ما به رفتار جواب عددی معادلهٔ دیفرانسیل $y' = \lambda y$ با $\mathrm{Real}(\lambda) < 0$ علاقه‌مندیم. به‌ویژه به مقادیری از $h\lambda$ علاقه‌مندیم که برای آنها جواب عددی با $x_n \to \infty$ به صفر میل می‌کند.

۵۲. (الف) با به‌کاربردن روش (۸.۱۰.۶) رونگه ـ کوتا، معادلهٔ

$$y' = -y + x^{0.1}[1.1 + x] \qquad y(0) = 0$$

را که جواب آن $Y(x) = x^{1.1}$ است حل کنید. این معادله را در $[0, 5]$ حل کنید و خطاها را در $x = 1, 2, 3, 4, 5$ چاپ کنید. طول گامهای $h = 0.1, 0.05, 0.025, 0.0125, 0.00625$

به‌کار برید. خطاهایی را معین کنید که بر اثر آنها وقتی h نصف می‌شود خطاها کاهش یابند. چگونه این عمل با نرخ نظری معمولی همگرایی $O(h^۲)$ مقایسه می‌شود؟ نتایج خود را توضیح دهید.
(ب) چه مشکلی پیش می‌آید وقتی بخواهیم معادلهٔ قسمت (الف) را با روش ناکوچکتر از مرتبهٔ ۲ تیلر حل کنیم؟ این نتیجه چه چیزی را دربارهٔ جواب به ما می‌گوید؟

۵۳. مسألهٔ مقدار مرزی (۱.۱۱.۶) را به یک مسألهٔ مقدار مرزی هم‌ارز برای یک دستگاه معادلات مرتبهٔ اول، مانند (۱۵.۱۱.۶) تبدیل کنید.

۵۴. (الف) مسألهٔ مقدار مرزی (۲۵.۱۱.۶) را در نظر می‌گیریم. برای تبدیل آن به یک مسألهٔ هم‌ارز با شرایط مرزی صفر، می‌نویسیم $y(x) = z(x) + w(x)$، که $w(x)$ خط راستی است که در شرایط مرزی زیر صدق می‌نماید: $w(a) = \gamma_۱$، $w(b) = \gamma_۲$. یک مسألهٔ جدید مقدار مرزی برای $z(x)$ به‌دست آورید.
(ب) این روند را برای مسألهٔ (۱۰.۱۱.۶) تعمیم دهید. یک مسألهٔ جدید با شرایط مرزی صفر به‌دست آورید. چه مفروضاتی، اگر فرضی لازم باشد، برای ضرایب $a_۰$، $a_۱$، $b_۰$، $b_۱$ لازم است؟

۵۵. با استفاده از روش پرتابی بخش ۱۱.۶، مسائل مقدار مرزی زیر را حل کنید. نرخ همگرایی را وقتی h تغییر می‌کند مطالعه کنید.
(الف) $y(۲) = \frac{۲}{۳}, y(۱) = \frac{۱}{۲}$؛ $y'' = \frac{-۲}{x}yy'$ $۱ < x < ۲$,
$Y(x) = x/(۱ + x)$ جواب درست است.

(ب) $y(\pi/۴) = ۱$، $y(۰) = ۰$، و $y'' = ۲yy'$ $۰ < x < \frac{\pi}{۴}$. $Y(x) = \tan(x)$ جواب درست است.

۵۶. برنامه‌های معادلهٔ دیفرانسیل را که مرکز محاسبات خودتان تهیه دیده بررسی نمایید. آنهایی را در نظر بگیرید که با تغییر طول گام و احتمالاً مرتبه، خطای برشی را به‌طور خودکار کنترل می‌نمایند. برنامه‌ها را برحسب اینکه چندگامی (ثابت ـ مرتبه یا متغیر ـ مرتبه)، رونگه ـ کوتا یا برونیابی رده‌بندی کنید. یکی از این برنامه‌ها را با DDEABM [از بخش ۷.۶ و شمپاین و گوردن (۱۹۷۵)] و RKF45 [از بخش ۹.۶ و شمپاین و واتس (۱۹۷۶b)] با حل‌کردن مسألهٔ

$$y' = \frac{y}{۴}\left[۱ - \frac{y}{۲۰}\right] \qquad y(۰) = ۱$$

با خطاهای مطلق مطلوب $۱۰^{-۳}$، $۱۰^{-۶}$ و $۱۰^{-۹}$ مقایسه نمایید. نتایج را با نتایج جدولهای ۱۵.۶ و ۲۸.۶ مقایسه نمایید.

۵۷. مسألهٔ

$$y' = \frac{۱}{t+۱} + c.\tan^{-۱}(y(t)) - \frac{۱}{۲} \qquad y(۰) = ۰$$

را که c یک ثابت داده شده است، در نظر می‌گیریم. چون $y'(۰) = \frac{۱}{۲}$، جواب $y(t)$ وقتی t افزایش پیدا کند بدون توجه به مقدار c در آغاز زیاد می‌شود. تا آنجا که می‌توانید، نشان دهید که مقداری برای c وجود دارد که آن را c^* می‌نامیم و برای آن (۱) اگر $c > c^*$، جواب $y(t)$ به بینهایت افزایش می‌یابد و (۲) اگر $c < c^*$، $y(t)$ ابتداً زیاد می‌شود و به حداکثر می‌رسد و بعد کاهش می‌یابد. c^* را با دقت ۰٫۰۰۰۰۵ تعیین و سپس جواب مربوط $y(t)$ را برای $۰ \le t \le ۵۰$ پیدا کنید.

۵۸. دستگاه زیر را در نظر بگیرید

$$x'(t) = Ax - Bxy \qquad y'(t) = Cxy - Dy$$

این دستگاه به مدل شکار و شکارچی لوتکا ـ ولترا معروف است، که $x(t)$ تعداد شکار و $y(t)$ تعداد شکارچی در زمان t را مشخص می‌کند،

(الف) گیریم $A = ۴$، $B = ۲$، $C = ۱$، $D = ۳$. مدل را حداقل تا سه رقم معنی‌دار برای $۰ \le t \le ۵$ حل کنید. مقادیر اولیه $x(۰) = ۳$ و $y(۰) = ۵$ هستند. x و y را به‌صورت توابعی از t رسم کنید و x را نسبت به y رسم نمایید.

(ب) همان مدل را با $x(۰) = ۳$ و به ترتیب ۲، ۱٫۵، ۱ $= y(۰)$ حل کنید. x را نسبت به y در هر حالت رسم کنید. چه چیزی مشاهده می‌کنید؟ چرا نقطهٔ $(۳, ۲)$ نقطهٔ تعادل خوانده می‌شود؟

۷

جبر خطی

حل دستگاه معادلات خطی و به‌دست‌آوردن ویژه مقدارها و ویژه‌بردارهای ماتریس دو مسألهٔ بسیار مهم‌اند که در موقعیّتهای گوناگون پیش می‌آیند. به‌عنوان مقدمهٔ بحث در این گونه مسائل در فصلهای آینده، بعضی از قضایای جبرخطی را ارائه می‌دهیم. بخش نخست شامل مروری بر مطالب فضاهای برداری، ماتریسها، و دستگاههای خطی است که در بسیاری از دروس جبر خطّی در دورهٔ کارشناسی تدریس می‌شوند. این قضایا فقط خلاصه شده‌اند، و هیچ‌گونه راه به‌دست‌آوردن آنها ذکر نشده است. در بقیّهٔ بخشها، از ویژه‌مقدارها، شکلهای متعارف ماتریسها، نرمهای بردار و ماتریس و قضایای اختلال در وارون ماتریسها بحث می‌شود. اگر لازم باشد می‌توان از خواندن این فصل صرف‌نظر کرد و هر زمان در فصلهای ۸ و ۹ لازم شود به آن مراجعه کرد. نمادگذاری، بخش ۱.۷ و نماد نرم در بخش ۳.۷ باید به اجمال مورد توجه قرار گیرند.

۱.۷ فضاهای برداری، ماتریسها و دستگاههای خطی

قطع نظر از جزئیات، **فضای برداری** V مجموعه‌ای است از اشیائی به نام **بردار** که برای آنها **جمع‌برداری** و **ضرب اسکالر** تعریف شده‌اند. هر فضای برداری V یک مجموعهٔ اسکالر وابسته دارد و در این کتاب این مجموعه ممکن است اعداد حقیقی $\mathbf{R}$ یا اعداد مختلط $\mathbf{C}$ باشد. عملیات برداری باید در قواعد

استانداً شرکتپذیری، تعویضپذیری و توزیعپذیری معینی صدق نمایند، که در اینجا فهرست نمی‌شوند. W، زیرمجموعهٔ فضای برداری V زیرفضای V خوانده می‌شود، اگر W خود یک فضای برداری باشد و از همان عملیات در V استفاده شود. برای یک بحث کامل در نظریهٔ فضاهای برداری، به هر یک ازکتابهای جبر خطی دورهٔ کارشناسی می‌توان مراجعه نمود [برای مثال، آنتون[1] (۱۹۸۴، فصل ۳)، هالموس[2] (۱۹۸۵، فصل ۱)، نوبل[3] (۱۹۶۹، فصلهای ۴ و ۱۴)، استرانگ[4] (۱۹۸۰، فصل ۲).

مثال ۱. $V = \mathbf{R}^n$، مجموعهٔ تمام nـ تاییهای $(x_1, \ldots, x_n)$، با درایه‌های حقیقی x_i، و $\mathbf{R}$ مجموعهٔ اسکالرهای وابسته به آن است.

۲. $V = \mathbf{C}^n$، مجموعه تمام nـ تاییها با درایه‌های مختلط و $\mathbf{C}$ مجموعهٔ اسکالرهاست.

۳. $V =$ مجموعهٔ چندجمله‌ییهای از درجهٔ نابزرگتر از n، برای یک n مفروض، یک فضای برداری است. اسکالرها می‌توانند $\mathbf{R}$ یا $\mathbf{C}$، بسته به کاربرد مطلوب، باشند.

۴. $V = C[a, b]$، مجموعهٔ همهٔ توابع پیوستهٔ حقیقی مقدار [یا مختلط مقدار] بر بازهٔ $[a, b]$ با مجموعهٔ اسکالر R [یا C]، یک فضای برداری است. مثال (۳) یک زیرفضای $C[a, b]$ است.

تعریف گیریم V یک فضای برداری باشد و $v_1, v_2, \ldots, v_m \in V$.

۱. گوییم v_1، v_2، ...، v_m **خطی ـ وابسته‌اند** اگر یک مجموعه از اسکالرهای α_1، ...، α_m که حداقل یکی از آنها مخالف صفر است، وجود داشته باشد به‌طوری که

$$\alpha_1 v_1 + \ldots + \alpha_m v_m = 0$$

چون حداقل یکی از اسکالرها، مثلاً α_i، مخالف صفر است می‌توانیم بنویسیم

$$v_i = -\frac{\alpha_1}{\alpha_i} v_1 - \cdots - \frac{\alpha_{i-1}}{\alpha_i} v_{i-1} - \frac{\alpha_{i+1}}{\alpha_i} v_{i+1} \cdots - \frac{\alpha_m}{\alpha_i} v_m$$

گوییم v_i یک ترکیب خطی از بردارهای v_1، ...، v_{i-1}، v_{i+1}، ...، v_m است. برای اینکه یک مجموعه از بردارها خطی ـ وابسته باشد، باید یکی از آنها یک ترکیب خطی از بقیه باشد.

۲. بردارهای v_1، ...، v_m را **خطی ـ مستقل** گوییم اگر وابسته نباشند. به عبارت دیگر تنها انتخاب

1. Anton 2. Halmos 3. Noble 4. Strang

اسکالرهای α_1، ...، α_m برای آنکه رابطهٔ

$$\alpha_1 v_1 + \cdots + \alpha_m v_m = 0$$

برقرار باشد، انتخاب نمایان $\alpha_1 = \ldots = \alpha_m = 0$ باشد. هیچ‌یک از v_i ها را نتوان به‌صورت ترکیبی از بقیه نوشت.

۳. $\{v_1, \ldots, v_m\}$ یک پایه برای V است اگر برای هر $v \in V$، یک انتخاب یکتا از اسکالرهای α_1، ...، α_m موجود باشد به‌طوری که

$$v = \alpha_1 v_1 + \ldots + \alpha_m v_m$$

توجه نمایید که از این رابطه نتیجه می‌شود که v_1، ...، v_m مستقل‌اند. اگر چنین پایهٔ متناهی وجود داشته باشد، گوییم V متناهی‌ـ بُعد است. در غیر این صورت، آن را نامتناهی‌ـ بُعد خوانیم.

قضیهٔ ۱.۷ اگر V یک فضای برداری با پایهٔ $\{v_1, \ldots, v_m\}$ باشد، هر پایهٔ V دقیقاً m بردار دارد. عدد m بُعد V خوانده می‌شود.

مثال ۱. $\{1, x, x^2, \ldots, x^n\}$ یک پایه برای فضای V از چندجمله‌ییهای از درجهٔ نابزرگتر از n است. بنابراین بُعد V $n+1$ است.

۲. R^n و C^n دارای پایهٔ $\{e_1, \ldots, e_n\}$ هستند، که در آن

$$e_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0) \qquad (۱.۱.۷)$$

که ۱ در موضع iام قرار دارد. بعد R^n و C^n برابر n است. این پایه را پایهٔ استانده برای R^n و C^n خوانند و بردارهای آنها را **بردارهای یکه** نامند.

۳. $C[a, b]$ نامتناهی‌ـ بُعد است.

ماتریسها و دستگاههای خطی ماتریسها آرایه‌های مستطیلی از اعداد حقیقی یا مختلط هستند، و در حالت کلی، ماتریس از مرتبهٔ $m \times n$ دارای شکل زیر است:

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \qquad (۲.۱.۷)$$

ماتریس مرتبهٔ n، نامگذاری مختصر ماتریس مربعی از مرتبهٔ $n \times n$ است. ماتریسها با حروف بزرگ و درایه‌های آنها معمولاً با حروف کوچک نشان داده می‌شوند، که معمولاً نظایر نام ماتریس‌اند، همان‌طور که هم‌اکنون نشان داده شد. تعاریف زیر عملیات معمولی روی ماتریسها را به‌دست می‌دهند.

تعریف ۱. گیریم A و B از مرتبهٔ $m \times n$ باشند. مجموع A و B ماتریس $C = A + B$ از مرتبهٔ $m \times n$ با درایه‌های زیر است

$$c_{ij} = a_{ij} + b_{ij}$$

۲. گیریم A از مرتبه $m \times n$ و α یک اسکالر باشد. در این صورت $C = \alpha A$، مضرب اسکالر A از مرتبهٔ $m \times n$ است و با

$$c_{ij} = \alpha a_{ij}$$

داده می‌شود.

۳. گیریم A از مرتبهٔ $m \times n$ و B از مرتبهٔ $n \times p$ باشد. در این صورت حاصلضرب $C = AB$ از مرتبهٔ $m \times p$ است که به شکل زیر داده می‌شود

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

۴. گیریم A از مرتبهٔ $m \times n$ باشد. **ترانهادهٔ** $C = A^T$ از مرتبهٔ $n \times m$ است و با رابطهٔ زیر داده می‌شود

$$c_{ij} = a_{ji}$$

ترانهادهٔ مزدوج $C = A^*$ نیز از مرتبهٔ $n \times m$ است و

$$c_{ij} = \overline{a}_{ji}$$

نماد $\overline{z}$ معرّف مزدوج مختلط عدد مختلط z است و z حقیقی است اگر و تنها اگر $\overline{z} = z$. ترانهادهٔ مزدوج A^* **ماتریس الحاقی** A نیز خوانده می‌شود.

ویژگیهای حسابی ماتریسها که در زیر آمده‌اند، بدون اشکال زیادی ثابت می‌شوند و اثبات آنها به عهدهٔ خواننده واگذار می‌شود

$$\begin{array}{lll} (A+B)+C=A+(B+C) & \text{(ب)} & A+B=B+A \quad \text{(الف)} \\ A(BC)=(AB)C & \text{(د)} & A(B+C)=AB+AC \quad \text{(ج)} \\ (AB)^T=B^TA^T & \text{(و)} & (A+B)^T=A^T+B^T \quad \text{(ه)} \end{array} \qquad (۳.۱.۷)$$

توجه به این امر مهم است که بدانیم در بسیاری از کاربردها برای برقراری ویژگیهای فوق لازم نیست ماتریسها مربعی باشند.

فضاهای برداری R^n و C^n معمولاً با مجموعهٔ بردارهای ستونی از مرتبهٔ $n \times ۱$ به ترتیب با درایه‌های حقیقی و مختلط تعریف می‌شوند. دستگاه خطی

$$\begin{aligned} a_{۱۱}x_۱+\ldots+a_{۱n}x_n &= b_۱ \\ &\vdots \\ a_{m۱}x_۱+\ldots+a_{mn}x_n &= b_m \end{aligned} \qquad (۴.۱.۷)$$

را می‌توان به صورت $Ax=b$ نوشت که A ماترس (۲.۱.۷) است، و

$$x=[x_۱,\ldots,x_n]^T \qquad b=[b_۱,\ldots,b_m]^T$$

بردار b یک بردار داده‌شده در R^m است و جواب x بردار مجهولی در R^n. استفاده از ضرب ماتریسی، دستگاه خطی (۴.۱.۷) را به شکل ساده‌تر و شهودیتر $Ax=b$ در می‌آورد.

ما اکنون چند تعریف اضافی دیگر برای ماتریسها، از جمله بعضی ماتریسهای خاص می‌دهیم.

تعریف ۱. **ماتریس صفر** از مرتبهٔ $m \times n$ همهٔ درایه‌هایش برابر صفرند. آن را با نماد $۰_{m\times n}$ یا به‌طور ساده‌تر با ۰ نشان می‌دهند. برای هر ماتریس A از مرتبه $m \times n$

$$A+۰=۰+A=A$$

۲. ماتریس همانی از مرتبهٔ n با $I=[\delta_{ij}]$ تعریف می‌شود،

$$\delta_{ij}=\begin{cases} ۱ & i=j \\ ۰ & i\neq j \end{cases} \qquad (۵.۱.۷)$$

برای جمیع مقادیر $1 \leq i, j \leq n$. برای هر ماتریس A از مرتبهٔ $m \times n$ و هر ماتریس B از مرتبهٔ $n \times p$

$$AI = A \qquad IB = B$$

نماد δ_{ij} معرّف **تابع دلتای کرونکر** است.

۳. گیریم A یک ماتریس مربعی از مرتبهٔ n باشد. اگر یک ماتریس مربعی B از مرتبهٔ n وجود داشته باشد به‌طوری که $AB = BA = I$، آنگاه A را **وارونپذیر** یا **وارون** B خوانند. می‌توان نشان داد که ماتریس B یکتاست و وارون A را با نماد A^{-1} نشان می‌دهیم.

۴. ماتریس A **متقارن** نامیده می‌شود اگر $A^T = A$، و **ارمیتی** خوانده می‌شود اگر $A = A^*$. اصطلاح متقارن معمولاً فقط برای ماتریسهای حقیقی به‌کار برده می‌شود. ماتریس A **پادمتقارن** نامیده می‌شود اگر $A^T = -A$. الزاماً همهٔ ماتریسهای متقارن، ارمیتی، یا پادمتقارن باید مربعی نیز باشند.

۵. گیریم A یک ماتریس $m \times n$ باشد. رتبهٔ سطری A تعداد سطرهای خطی مستقل A است که عناصری از R^n یا C^n تلقّی می‌شوند، و رتبهٔ ستونی تعداد ستونهای خطی مستقل است. می‌توان نشان داد (مسألهٔ ۴) که این دو عدد همیشه مساوی‌اند و این را **رتبهٔ** A خوانند.

برای تعریف و ویژگیهای دترمینان یک ماتریس مربعی A، می‌توان به هرکتاب درسی جبر خطی مراجعه نمود [برای مثال، آنتون (۱۹۸۴، فصل ۲)؛ نوبل (۱۹۶۹، فصل ۷)؛ استرانگ (۱۹۸۰، فصل ۴) را ببینید]. بسیاری از قضایای وارون ماتریسها و حلپذیری دستگاههای خطی را در قضیهٔ زیر خلاصه می‌کنیم.

قضیهٔ ۲.۷ گیریم A یک ماتریس مربعی با عناصری از R (یا C) باشد، و فضای برداری (یا $V = C^n$ باشد. در این صورت عبارات زیر هم‌ارزند:

۱. $Ax = b$ برای هر $b \in V$ یک جواب یکتای $x \in V$ دارد.

۲. $Ax = b$ برای هر $b \in V$ یک جواب $x \in V$ دارد.

۳. $Ax = 0$ ایجاب می‌کند $x = 0$.

۴. A^{-1} وجود دارد.

۵. دترمینان $(A) \neq 0$.

۶. رتبهٔ $(A) = n$.

برهانی در اینجا آورده نشده است، یک تمرین بسیار خوب، اثبات هم‌ارزی بعضی از این احکام است. مفاهیم استقلال خطی و پایه را همراه با قضیهٔ ۱.۷ به‌کار برید. همچنین از تجزیهٔ زیر استفاده کنید

$$Ax = x_1 A_{*1} + \ldots + x_n A_{*n} \qquad x \in \mathbf{R}^n \text{ یا } \mathbf{C}^n \qquad (۶.۱.۷)$$

که A_{*j} معرّف ستون j ام در A است. این عبارت می‌گوید که فضای همهٔ بردارهایی به شکل Ax توسط ستونهای A پدید می‌آیند، ولو اینکه خطی وابسته باشند.

فضاهای برداری با حاصلضرب داخلی یکی از دلایل مهم که مسائل را به صورت مسائل هم‌ارز در جبر خطی بیان می‌کنند عرضهٔ یک دیدگاه هندسی مسأله است. آنچه در این فرایند مهم است، حاصلضرب داخلی و تعامد است.

تعریف ۱. حاصلضرب داخلی دو بردار $x, y \in \mathbf{R}^n$ چنین تعریف می‌شود

$$(x, y) = \sum_{i=1}^{n} x_i y_i = x^T y = y^T x$$

و برای بردارهای $x, y \in \mathbf{C}^n$، حاصلضرب داخلی با رابطهٔ زیر تعریف می‌شود

$$(x, y) = \sum_{i=1}^{n} x_i \overline{y_i} = y^* x$$

۲. نرم اقلیدسی x در $\mathbf{C}^n$ یا $\mathbf{R}^n$ چنین تعریف می‌شود

$$\| x \|_2 = \sqrt{(x, x)} = \sqrt{| x_1 |^2 + \ldots + | x_n |^2} \qquad (۷.۱.۷)$$

اثبات قضایای زیر نسبتاً ساده است و به عهدهٔ خواننده واگذار شده است. گیریم V معرف $\mathbf{C}^n$ یا $\mathbf{R}^n$ باشد

۱. برای $x, y, z \in V$

$$(x + y, z) = (x, z) + (y, z) \qquad (x, y + z) = (x, y) + (x, z)$$

۲. برای هر $x, y \in V$

$$(\alpha x, y) = \alpha (x, y)$$

و برای $V = C^n$ و $\alpha \in C$

$$(x, \alpha y) = \overline{\alpha}(x, y)$$

۳. در C^n، $(x, y) = \overline{(y, x)}$؛ و در R^n، $(x, y) = (y, x)$.
۴. برای هر $x \in V$

$$(x, x) \geq 0$$

و $(x, x) = 0$، اگر و تنها اگر $x = 0$.
۵. برای هر $x, y \in V$،

$$|(x, y)|^2 \leq (x, x)(y, y) \qquad (۸.۱.۷)$$

که **نابرابری کوشی ـ شوارتس** خوانده می‌شود، و عیناً از راه اثبات (۳.۴.۴) فصل ۴ ثابت می‌شود. با استفاده از نرم اقلیدسی می‌توانیم بنویسیم

$$|(x, y)| \leq \| x \|_2 \| y \|_2 \qquad (۹.۱.۷)$$

۶. برای هر $x, y \in V$

$$\| x + y \|_2 \leq \| x \|_2 + \| y \|_2 \qquad (۱۰.۱.۷)$$

این رابطه **نابرابری مثلّثی** است. برای یک تعبیر هندسی، توضیحات قبلی در بخش ۱.۴ فصل ۴ را برای نرم $\| f \|_\infty$ روی $C[a, b]$ ببینید. برای اثبات (۱۰.۱.۷)، طرز به‌دست‌آوردن (۴.۴.۴) را در فصل چهار ملاحظه کنید.

۷. برای هر ماتریس مربعی از مرتبهٔ n و برای هر $x, y \in C^n$،

$$(Ax, y) = (x, A^* y) \qquad (۱۱.۱.۷)$$

قبلاً **حاصلضرب داخلی** برای معرفی طول اقلیدسی به‌کار رفته بود، ولی حداقل در فضاهایی که مجموعهٔ اسکالر R باشد، برای تعبیری از زاویه نیز به‌کار می‌رود.

تعریف ۱. برای x و y در $\mathbf{R}^n$، **زاویه بین** x و y با رابطهٔ زیر تعریف می‌شود

$$\mathscr{A}(x, y) = \cos^{-1}\left[\frac{(x, y)}{\| x \|_2 \| y \|_2}\right]$$

توجه نمایید که به موجب نابرابری (۹.۱.۷) کوشی ـ شوارتس، شناسه بین ۱− و ۱ است. این تعریف را می‌توان به‌طور ضمنی به شکل زیر نوشت

$$(x, y) = \| x \|_2 \| y \|_2 \cos(\measuredangle) \qquad (۱۲.۱.۷)$$

که یک فرمول آشنا از کاربرد حاصلضرب نقطه‌یی در R^2 و R^3 است.

۲. دو بردار x و y **متعامدند** اگر و تنها اگر $(x, y) = 0$. انگیزهٔ این اصطلاح از (۱۲.۱.۷) بوده است. اگر $\{x^{(1)}, \ldots, x^{(n)}\}$ یک پایه برای C^n یا R^n باشد و اگر $(x^{(i)}, x^{(j)}) = 0$ برای تمام مقادیر $i \neq j$، $n \leq i, j \leq n$، آنگاه $\{x^{(1)}, \ldots, x^{(n)}\}$ را پایهٔ **متعامد** خوانیم. اگر تمام بردارهای پایه دارای طول اقلیدسی ۱ باشند، پایه را **یکا متعامد** خوانند.

۳. ماتریس مربعی U را **یکانی** گویند اگر

$$U^* U = U U^* = I$$

اگر ماتریس U حقیقی باشد به جای یکانی آن را **متعامد** خوانند. سطرها [یا ستونها]ی یک ماتریس یکانی مرتبهٔ n، یک پایهٔ یکا متعامد برای $\mathbf{C}^n$ تشکیل می‌دهند، و همچنین است برای ماتریسهای متعامد و R^n.

مثال ۱. زاویهٔ بین بردارهای $y = (۳, ۲, ۱)$، $x = (۱, ۲, ۳)$ با رابطهٔ زیر داده می‌شود:

$$\mathcal{A} = \cos^{-1}\left[\frac{۱۰}{۱۴}\right] \doteq ۰٫۷۷۵ \text{ رادیان}$$

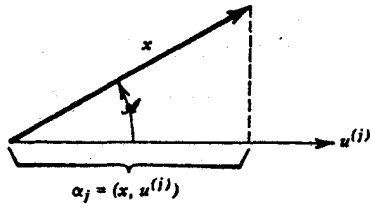
۲. ماتریسهای

$$U_1 = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \qquad U_2 = \begin{bmatrix} \frac{۱}{\sqrt{۲}} & \frac{i}{\sqrt{۲}} \\ \frac{i}{\sqrt{۲}} & \frac{۱}{\sqrt{۲}} \end{bmatrix}$$

یکانی‌اند، و اولی متعامد است.

برای فضای برداری R^n یا $V = C^n$، داشتن یک پایهٔ یکا متعامد مفید است، زیرا در این صورت به آسانی می‌توان هر بردار دلخواه را به مؤلفه‌های آن در امتداد بردارهای پایه تجزیه کرد. به بیان دقیقتر، گیریم $\{u^{(1)}, \ldots, u^{(n)}\}$ پایهٔ یکا متعامد V باشد و $x \in V$. با استفاده از این پایه

$$x = \alpha_1 u^{(1)} + \ldots + \alpha_n u^{(n)}$$



شکل ۱.۷ نمایش (۱۵.۱.۷)

ضرایب α_1، ...، α_n یکتا هستند. برای پیداکردن α_j، حاصلضرب داخلی x و $u^{(j)}$ را تشکیل می‌دهیم و سپس با استفاده از ویژگیهای یکا متعامدی پایه خواهیم داشت

$$(x, u^{(j)}) = \alpha_1(u^{(1)}, u^{(j)}) + \ldots + \alpha_n(u^{(n)}, u^{(j)})$$
$$= \alpha_j \qquad (۱۳.۱.۷)$$

بنابراین

$$x = \sum_{j=1}^{n} (x, u^{(j)}) u^{(j)} \qquad (۱۴.۱.۷)$$

شکل ۱.۷ یک تعبیر هندسی از این رابطه را نشان می‌دهد. با استفاده از (۱۳.۱.۷)،

$$\alpha_j = (x, u^{(j)}) = \| x \|_2 \| u^{(j)} \|_2 \cos(\measuredangle(x, u^{(j)}))$$
$$= \| x \|_2 \cos(\measuredangle(x, u^{(j)})) \qquad (۱۵.۱.۷)$$

بنابراین ضریب α_j طول تصویر قائم x روی محوری است که با $u^{(j)}$ مشخص می‌شود. فرمول (۱۴.۱.۷) تعمیم تجزیهٔ بردار x با استفاده از پایهٔ استانداردِ $\{e^{(1)}, \ldots, e^{(n)}\}$ است، که قبلاً تعریف شده بود.

مثال گیریم $V = \mathbf{R}^2$، و پایهٔ یکا متعامد

$$u^{(1)} = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) \qquad u^{(2)} = \left(-\frac{\sqrt{3}}{2}, \frac{1}{2}\right)$$

را در نظر می‌گیریم. در این صورت برای یک بردار داده‌شدهٔ $x = (x_1, x_2)$، می‌توان نشان داد که

$$x = \alpha_1 u^{(1)} + \alpha_2 u^{(2)}$$

$$\alpha_1 = (x_1, u^{(1)}) = \frac{x_1 + x_2\sqrt{3}}{2} \qquad \alpha_2 = (x, u^{(2)}) = \frac{x_2 - x_1\sqrt{3}}{2}$$

برای مثال

$$(۱, ۰) = \frac{۱}{۲}u^{(۱)} - \frac{\sqrt{۳}}{۲}u^{(۲)}$$

۲.۷ ویژه‌بردارها و شکلهای متعارف (کانونی) ماتریسها

عدد λ، مختلط یا حقیقی، یک **ویژه مقدار** ماتریس مربعی A است اگر یک بردار $x \in C^n$، $x \neq ۰$، وجود داشته باشد به‌طوری که

$$Ax = \lambda x \qquad (۱.۲.۷)$$

بردار x را **ویژه‌بردار** متناظر با ویژه مقدار λ نامند. با توجّه به قضیهٔ ۲.۷ و احکام (۳) و (۵)، λ یک ویژه مقدار A است اگر و تنها اگر

$$\det(A - \lambda I) = ۰ \qquad (۲.۲.۷)$$

این معادله را **معادلهٔ مشخصهٔ** A خوانند و برای تحلیل آن تابع زیر را معرفی می‌کنیم

$$f_A(\lambda) \equiv \det(A - \lambda I)$$

اگر A از مرتبهٔ n باشد، $f_A(\lambda)$ یک چندجمله‌یی دقیقاً از مرتبه n است که **چندجمله‌یی مشخصهٔ** A نامیده می‌شود. برای اثبات اینکه این یک چندجمله‌یی است این دترمینان را برحسب دترمینانهای جزئی متوالیاً بسط می‌دهیم، تا عبارت زیر به‌دست آید

$$\begin{aligned} f_A(\lambda) &= \det(A - \lambda I) \\ &= \det \begin{bmatrix} a_{۱۱} - \lambda & a_{۱۲} & \dots & a_{۱n} \\ a_{۲۱} & a_{۲۲} - \lambda & a_{۲۳} \dots & a_{۲n} \\ \vdots & & \ddots & \\ a_{n۱} & & \dots & a_{nn} - \lambda \end{bmatrix} \\ &= (a_{۱۱} - \lambda)(a_{۲۲} - \lambda) \dots (a_{nn} - \lambda) \\ &\quad + (\text{جملاتی از درجهٔ نابزرگتر از } n - ۲) \\ f_A(\lambda) &= (-۱)^n \lambda^n + (-۱)^{n-۱}(a_{۱۱} + \dots, + a_{nn})\lambda^{n-۱} \\ &\quad + (\text{جملاتی از درجهٔ نابزرگتر از } n - ۲) \end{aligned} \qquad (۳.۲.۷)$$

همچنین توجه نمایید که جملهٔ ثابت عبارت است از

$$f_A(0) = \det(A) \qquad (۴.۲.۷)$$

با توجه به ضریب λ^{n-1}، تعریف می‌کنیم

$$\operatorname{trace}(A) = a_{11} + a_{22} + \ldots + a_{nn} \qquad (۵.۲.۷)$$

که اغلب یک کمیت مورد علاقه در مطالعهٔ A است.

چون $f_A(\lambda)$ از درجهٔ n است، دقیقاً n ویژه مقدار برای A موجود است، به شرطی که ریشه‌های چندگانه را برحسب درجهٔ چندگانی آنها به حساب آوریم. هر ماتریس حداقل یک زوج ویژه مقدار ـ ویژه بردار دارد و ماتریس A از مرتبهٔ $n \times n$، حداکثر n ویژه مقدار متمایز خواهد داشت.

مثال ۱. چندجمله‌یی مشخصهٔ

$$A = \begin{bmatrix} ۲ & ۱ & ۰ \\ ۱ & ۳ & ۱ \\ ۰ & ۱ & ۲ \end{bmatrix}$$

عبارت است از

$$f_A(\lambda) = -\lambda^۳ + ۷\lambda^۲ - ۱۴\lambda + ۸$$

ویژه‌مقدارها عبارت‌اند از، $\lambda_۱ = ۱$، $\lambda_۲ = ۲$، $\lambda_۳ = ۴$ و ویژه‌بردارهای متناظر آنها چنین‌اند

$$u^{(۱)} = \begin{bmatrix} ۱ \\ -۱ \\ ۱ \end{bmatrix} \qquad u^{(۲)} = \begin{bmatrix} ۱ \\ ۰ \\ -۱ \end{bmatrix} \qquad u^{(۳)} = \begin{bmatrix} ۱ \\ ۲ \\ ۱ \end{bmatrix}$$

توجه نمایید که این ویژه‌بردارها بر هم عمود و بنابراین خطی مستقل‌اند. چون بُعد $R^۳$ (و $C^۳$) برابر ۳ است، این ویژه‌بردارها یک پایهٔ متعامد برای $R^۳$ (و $C^۳$) تشکیل می‌دهند. این نکته توضیحی برای قضیهٔ ۴.۷ است که بعداً بیان خواهد شد.

۲. برای ماتریس

$$A = \begin{bmatrix} ۱ & ۰ & ۰ \\ ۰ & ۱ & ۰ \\ ۰ & ۰ & ۱ \end{bmatrix} \qquad f_A(\lambda) = (۱ - \lambda)^۳$$

و سه ویژه‌بردار خطی مستقل برای ویژه‌مقدار $\lambda = ۱$ وجود دارد، برای مثال

$$[۱, ۰, ۰]^T \quad [۰, ۱, ۰]^T \quad [۰, ۰, ۱]^T$$

تمام ویژه‌بردارهای دیگر ترکیبهای خطی از این سه بردارند.

۳. برای ماتریس

$$A = \begin{bmatrix} ۱ & ۱ & ۰ \\ ۰ & ۱ & ۱ \\ ۰ & ۰ & ۱ \end{bmatrix} \qquad f_A(\lambda) = (۱ - \lambda)^۳$$

ماتریس A فقط یک ویژه‌بردار خطی مستقل برای ویژه‌مقدار $\lambda = ۱$ یعنی

$$x = [۱, ۰, ۰]^T$$

و مضارب آن دارد.

چندگانگی جبری یک ویژه‌مقدار ماتریس A، چندگانگی ریشهٔ $f_A(\lambda)$ است و **چندگانگی هندسی** آن بیشترین عدد ویژه‌بردارهای خطی مستقل آن ماتریس متناظر با ویژه‌مقدار آن است. مجموع چندگانگیهای جبری ویژه‌بردارهای ماتریس A از مرتبهٔ $n \times n$، نسبت به اختلالات جزئی در A، ثابت، یعنی n، است. ولی مجموع چندگانگیهای هندسی ممکن است با اختلالات جزئی، تغییرات بزرگی داشته باشد و این امر موجب می‌شود که محاسبهٔ عددی ویژه‌بردارها اغلب یک مسألهٔ مشکل باشد. همچنین چندگانگیهای جبری و هندسی لازم نیست مساوی باشند، همان‌گونه که مثالهای قبل نشان می‌دهند.

تعریف گیریم A و B ماتریسهای مربعی هم مرتبه باشند. در این صورت A **متشابه** با B است اگر یک ماتریس ناتکین P وجود داشته باشد که

$$B = P^{-۱}AP \qquad (۶.۲.۷)$$

توجه نمایید که این رابطه یک رابطهٔ متقارن است زیرا

$$A = Q^{-۱}BQ \qquad Q = P^{-۱}$$

رابطهٔ (۶.۲.۷) را می‌توان چنین تعبیر کرد که بگوییم A و B نمایشهای یک تبدیل خطی T از V به V، [$V = R^n$ یا C^n] هستند ولی نسبت به پایه‌های مختلف V. ماتریس P **ماتریس تبدیل پایه** خوانده می‌شود، و این ماتریس دو نمایش مختلف یک بردار $x \in V$ را نسبت به دوپایه‌ای که مورد استفاده قرار گرفته به هم مربوط می‌کند. [(آنتون (۱۹۸۴، بخش ۵.۵) یا نوبل (۱۹۶۹، بخش ۵.۱۴) را برای جزئیات بیشتر ببینید].

ما اکنون چند ویژگی سادهٔ ماتریسهای متشابه و ویژه مقدارهای آنها را بیان می‌کنیم.

۱. اگر A و B متشابه باشند، آنگاه $f_A(\lambda) = f_B(\lambda)$. برای اثبات این حکم از (۶.۲.۷) استفاده می‌کنیم تا نشان دهیم

$$\begin{aligned} f_B(\lambda) &= \det(B - \lambda I) = \det[P^{-1}(A - \lambda I)P] \\ &= \det(P^{-1}) \det(A - \lambda I) \det(P) = f_A(\lambda) \end{aligned}$$

زیرا

$$\det(P) \det(P^{-1}) = \det(PP^{-1}) = \det(I) = 1$$

۲. ویژه‌مقدارهای ماتریسهای متشابه A و B دقیقاً یکی هستند و یک تناظر یک به یک بین این ویژه‌بردارها وجود دارد. اگر $Ax = \lambda x$، آنگاه با استفاده از

$$P^{-1}AP(P^{-1}x) = \lambda P^{-1}x$$

$$Bz = \lambda z \qquad z = P^{-1}x \tag{۷.۲.۷}$$

بدیهی است که $z \neq 0$، زیرا در غیر این صورت x صفر می‌شود. همچنین برای هر ویژه‌بردار داده‌شدهٔ z از B، این استدلال را می‌توان معکوس کرد تا ویژه‌بردار متناظر $x = Pz$ برای A تولید شود.

۳. چون $f_A(\lambda)$ تحت تبدیلهای تشابهی ناورداست، ضرایب $f_A(\lambda)$ نیز تحت چنین تبدیلهای تشابهی، ناوردا خواهند بود. به‌ویژه برای ماتریسهای متشابه A و B،

$$\operatorname{trace}(A) = \operatorname{trace}(B) \qquad \det(A) = \det(B) \tag{۸.۲.۷}$$

شکلهای متعارف اکنون به ذکر چندین شکل متعارف مهم ماتریسها می‌پردازیم. این شکلها بین ساختار یک ماتریس و ویژه‌مقدارها و ویژه‌بردارهای آن ارتباط برقرار می‌کنند، و در کاربردهای گوناگون عرضه‌های دیگر ریاضیات و علوم مورد استفاده قرار می‌گیرند.

قضیهٔ ۳.۷ (شکل نرمال شور) گیریم A دارای مرتبهٔ n با عناصری از $\mathbf{C}$ باشد. در این صورت یک ماتریس یکانی وجود دارد به‌طوری که

$$T \equiv U^* AU \qquad (۹.۲.۷)$$

یک ماتریس بالا مثلثی است.

چون T مثلّثی است و چون $U^* = U^{-۱}$،

$$f_A(\lambda) = f_T(\lambda) = (\lambda - t_{۱۱}) \dots (\lambda - t_{nn}) \qquad (۱۰.۲.۷)$$

و بنابراین ویژه‌مقدارهای A عناصر قطری T هستند.

برهان برهان به استقراء روی n، مرتبهٔ A، انجام می‌گیرد. قضیه برای $n = ۱$ با استفاده از $U = [۱]$ درست است. گیریم که قضیّه برای تمام ماتریسهای مرتبهٔ $n \leq k - ۱$ درست باشد، و ثابت می‌کنیم که قضیّه برای تمام ماتریسهای از مرتبهٔ $n = k$ درست است.

گیریم $\lambda_۱$ یک ویژه‌مقدار A باشد، و $u^{(۱)}$ ویژه‌بردار متناظر آن باشد با $\| u^{(۱)} \|_۲ = ۱$. یک پایهٔ یکا متعامد برای C^k می‌گیریم که $u^{(۱)}$ اولین بردار پایه باشد و آن را $\{u^{(۱)}, \dots, u^{(k)}\}$ می‌نامیم. ماتریس $P_۱$ را با

$$P_۱ = [u^{(۱)}, u^{(۲)}, \dots, u^{(k)}]$$

تعریف می‌کنیم که به شکل افرازی با ستونهای متعامد $u^{(k)}, \dots, u^{(۱)}$ نوشته شده است. پس $P_۱^* P_۱ = I$ و بنابراین $P_۱^{-۱} = P_۱^*$. تعریف می‌کنیم

$$B_۱ = P_۱^* A P_۱$$

ادعا می‌کنیم

$$B_۱ = \begin{bmatrix} \lambda_۱ & \alpha_۲ & \dots & \alpha_k \\ ۰ & & & \\ \vdots & & A_۲ & \\ ۰ & & & \end{bmatrix}$$

که در آن A_2 از مرتبهٔ $k-1$ است، α_2، ...، α_k اعدادی هستند. برای اثبات این ادعا از ضرب ماتریسهای افرازشده استفاده می‌نماییم:

$$AP_1 = A[u^{(1)}, \dots, u^{(k)}] = [Au^{(1)}, \dots, Au^{(k)}]$$
$$= [\lambda_1 u^{(1)}, v^{(2)}, \dots, v^{(k)}] \qquad v^{(j)} = Au^{(j)}$$
$$B_1 = P_1^* AP_1 = [\lambda_1 P_1^* u^{(1)}, P_1^* v^{(2)}, \dots, P_1^* v^{(k)}]$$

چون $P_1^* P_1 = I$، داریم $P_1^* u^{(1)} = e^{(1)} = [1, 0, \dots, 0]^T$. بنابراین

$$B_1 = [\lambda_1 e^{(1)}, w^{(2)}, \dots, w^{(k)}] \qquad w^{(j)} = P_1^* v^{(j)}$$

که دارای شکل مطلوب است.

با فرض استقرا، یک ماتریس یکانی $\hat{P}_2$ از مرتبهٔ $k-1$ موجود است به‌طوری که

$$\hat{T} = \hat{P}_2^* A_2 \hat{P}_2$$

یک ماتریس بالا مثلثی از مرتبهٔ $k-1$ است. تعریف می‌کنیم،

$$P_2 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \hat{P}_2 & \\ 0 & & & \end{bmatrix}$$

پس P_2 یکانی است و

$$P_2^* B_1 P_2 = \begin{bmatrix} \lambda_1 & \gamma_2 & & \gamma_k \\ 0 & & & \\ \vdots & & \hat{P}_2^* A_2 \hat{P}_2 & \\ 0 & & & \end{bmatrix} = \begin{bmatrix} \lambda_1 & \gamma_2 & \dots & \gamma_k \\ 0 & & & \\ \vdots & & \hat{T} & \\ 0 & & & \end{bmatrix} \equiv T$$

یک ماتریس بالامثلثی است. بنابراین

$$T = P_2^* B_1 P_2 = P_2^* P_1^* AP_1 P_2 = (P_1 P_2)^* A (P_1 P_2)$$
$$T = U^* AU \qquad U = P_1 P_2$$

و به‌سادگی معلوم می‌شود که U یکانی است. پس، استقرا و برهان کامل است. ■

مثال برای ماتریس

$$A = \begin{bmatrix} ۰ر۲ & ۰ر۶ & ۰ \\ ۱ر۶ & -۰ر۲ & ۰ \\ -۱ر۶ & ۱ر۶ & ۳ر۰ \end{bmatrix}$$

ماتریسهای قضیه و (۹.۲.۷) چنین‌اند

$$T = \begin{bmatrix} ۱ & ۰ & -۱ \\ ۰ & ۳ & ۲ \\ ۰ & ۰ & -۱ \end{bmatrix} \qquad U = \begin{bmatrix} ۰ر۶ & ۰ & -۰ر۸ \\ ۰ر۸ & ۰ & ۰ر۶ \\ ۰ & ۱ر۰ & ۰ \end{bmatrix}$$

این راه معمولی محاسبهٔ ویژه‌مقدارها نیست، و فقط برای توضیح قضیه باید درنظرگرفته شود. این قضیه فقط به‌عنوان یک ابزار نظری به‌کار می‌رود، نه یک ابزار محاسباتی.

با استفاده از (۸.۲.۷) و (۹.۲.۷)

$$\text{trace}(A) = \lambda_۱ + \lambda_۲ + \ldots + \lambda_n \qquad \det(A) = \lambda_۱\lambda_۲\ldots\lambda_n \qquad (۱۱.۲.۷)$$

که $\lambda_۱$، ...، λ_n ویژه‌مقدارهای A هستند که باید عناصر قطری T را تشکیل دهند. به‌عنوان یک کاربرد خیلی مهمتر، قضیه معروف زیر را داریم.

قضیهٔ ۴.۷ (قضیهٔ محورهای اصلی) گیریم A یک ماتریس ارمیتی از مرتبهٔ n باشد، یعنی $A^* = A$. پس A دارای n ویژه‌مقدار $\lambda_۱$، ...، λ_n است که الزاماً متمایز نیستند، و n ویژه‌بردار متناظر $u_۱$، ...، u_n دارند که یک پایهٔ یکا متعامد C^n را تشکیل می‌دهند. اگر A حقیقی باشد، ویژه‌بردارهای $u_۱$، ...، u_n را می‌توان حقیقی گرفت که یک پایهٔ یکا متعامد برای R^n تشکیل خواهند داد. بالاخره یک ماتریس یکانی U وجود دارد به‌طوری که

$$U^*AU = D \equiv \text{diag}[\lambda_۱, \ldots, \lambda_n] \qquad (۱۲.۲.۷)$$

یک ماتریس قطری با عناصر قطری $\lambda_۱$، ...، λ_n است. اگر A حقیقی هم باشد، U را می‌توان متعامد گرفت.

برهان با توجه به قضیهٔ ۳.۷، یک ماتریس یکانی U موجود است که

$$U^*AU = T$$

که T بالامثلثی است. ترانهادهٔ مزدوج دو طرف را تشکیل می‌دهیم تا به‌دست آوریم

$$T^* = (U^*AU)^* = U^*A^*(U^*)^* = U^*AU = T$$

چون T^* پایین مثلثی است، باید داشته باشیم

$$T = \operatorname{diag}[\lambda_1, \dots, \lambda_n]$$

همچنین، $T^* = T$ شامل مزدوجهای مختلط تمام عناصر T است، و بنابراین همهٔ عناصر قطری T باید حقیقی باشند.

U را به صورت زیر می‌نویسیم

$$U = [u^{(1)}, \dots, u^{(n)}]$$

پس $T = U^*AU$ ایجاب می‌کند که $AU = UT$،

$$A[u^{(1)}, \dots, u^{(n)}] = [u^{(1)}, \dots, u^{(n)}] \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

$$[Au^{(1)}, \dots, Au^{(n)}] = [\lambda_1 u^{(1)}_1, \dots, \lambda_n u^{(n)}]$$

و

$$Au^{(j)} = \lambda_j u^{(j)} \qquad j = 1, \dots, n \qquad (۱۳.۲.۷)$$

چون ستونهای U یکا متعامدند و چون بعد C^n برابر n است، پس این بردارها باید یک پایهٔ یکا متعامد برای C^n تشکیل دهند. ما از ذکر برهان نتایجی که از حقیقی‌بودن A ناشی می‌شوند صرف‌نظر می‌کنیم. پس برهان کامل است. ■

مثال از مثال ۱ در این بخش، ماتریس

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

دارای ویژه‌مقدارهای $\lambda_1 = 1$، $\lambda_2 = 2$ و $\lambda_3 = 4$ و ویژه‌بردارهای یکا متعامد متناظر زیر است

$$u^{(1)} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \qquad u^{(2)} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \qquad u^{(3)} = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

اینها یک پایهٔ یکا متعامد برای R^3 یا C^3 تشکیل می‌دهند.

یک شکل متعارف دوّمی وجود دارد که اخیراً در حل مسائل جبر خطی عددی، به‌ویژه در حل دستگاه معادلات خطی فرامعیّن، بسیار مهم شده‌است. این دستگاهها از برازش داده‌های تجربی با استفاده از شیوه‌های خطّی کمترین مربعات به‌دست می‌آیند. [گلوب و وان لون[1] (۱۹۸۳، فصل ۶) و لوسن[2] و هانسن[3] (۱۹۷۴) را ببینید].

قضیهٔ ۵.۷ (تجزیهٔ تکین ـ مقدار) گیریم A از مرتبهٔ $n \times m$ باشد. در این صورت ماتریسهای یکانی U و V از مرتبه‌های به ترتیب m و n وجود دارند به‌طوری که ماتریس

$$V^* A U = F \qquad (۱۴.۲.۷)$$

یک ماتریس مستطیلی «قطری» از مرتبهٔ $n \times m$ به شکل زیر است

$$F = \begin{bmatrix} \mu_1 & & & 0 \\ & \mu_2 & & \\ & & \ddots & \\ 0 & & & \mu_r \\ & & & & 0 \\ & & & & & \ddots \end{bmatrix} \qquad (۱۵.۲.۷)$$

اعداد $\mu_1, \ldots, \mu_r$ **مقادیر تکین** A نامیده می‌شوند. این مقادیر همگی حقیقی و مثبت‌اند و آنها را می‌توان به شکلی مرتب کرد که

$$\mu_1 \geq \mu_2 > \ldots \geq \mu_r > 0 \qquad (۱۶.۲.۷)$$

که r رتبهٔ ماتریس A است.

1. Van Loan 2. Lawson 3. Hanson

برهان ماتریس مربعی A^*A از مرتبهٔ m را در نظر می‌گیریم. این ماتریس یک ماتریس ارمیتی است، و در نتیجه قضیهٔ ۴.۷ را می‌توان برای آن به‌کار برد. ویژه‌مقدارهای A^*A همه حقیقی، به‌علاوه همگی نامنفی‌اند. برای پی‌بردن به آن، فرض می‌کنیم

$$A^*Ax = \lambda x \qquad x \neq ۰$$

پس

$$(x, A^*Ax) = (x, \lambda x) = \lambda \parallel x \parallel_۲^۲$$

$$(x, A^*Ax) = (Ax, Ax) = \parallel Ax \parallel_۲^۲$$

$$\lambda = \left(\frac{\parallel Ax \parallel_۲}{\parallel x \parallel_۲}\right)^۲ \geq ۰$$

از این قضیّه همچنین ثابت می‌شود که

$$A^*Ax = ۰, x \in C^n \text{ اگر و تنها اگر } Ax = ۰ \qquad (۱۷.۲.۷)$$

با توجه به قضیهٔ ۳.۷، یک ماتریس یکانی U از مرتبه $m \times m$ موجود است به‌طوری که

$$U^*A^*AU = \operatorname{diag}[\lambda_۱, \dots, \lambda_r, ۰, \dots, ۰] \qquad (۱۸.۲.۷)$$

که همهٔ λ_iها، $۱ \leq i \leq r$، مخالف صفر و همگی مثبت‌اند. چون A^*A دارای مرتبهٔ m است، اندیس نابزرگتر از m است. **مقادیر تکین** زیر را معرفی می‌کنیم

$$\mu_i = \sqrt{\lambda_i} \qquad i = ۱, \dots, r \qquad (۱۹.۲.۷)$$

U می‌تواند طوری انتخاب شود که ترتیب (۱۶.۲.۷) به‌دست آید. با استفاده از ماتریس قطری مرتبهٔ m

$$D = \operatorname{diag}[\mu_۱, \dots, \mu_r, ۰, \dots, ۰]$$

(۱۸.۲.۷) را می‌توانیم به شکل زیر بنویسیم

$$(AU)^*(AU) = D^۲ \qquad (۲۰.۲.۷)$$

گیریم $W = AU$. پس از (۳۰.۲.۷)، $W^*W = D^2$. اگر W را به شکل

$$W = [W^{(1)}, \dots, W^{(m)}] \qquad W^{(j)} \in C^n$$

بنویسیم داریم

$$(W^{(j)}, W^{(j)}) = \begin{cases} \mu_j^2 & 1 \le j \le r \\ 0 & j > r \end{cases} \qquad (۲۱.۲.۷)$$

و

$$(W^{(i)}, W^{(j)}) = 0 \qquad \text{اگر } i \neq j \qquad (۲۲.۲.۷)$$

از (۲۱.۲.۷) نتیجه می‌شود، $W^{(j)} = 0$ اگر $j > r$. و از (۲۲.۲.۷) نتیجه می‌شود که r ستون اول W عناصر متعامد در C^n، و بنابراین خطی مستقل‌اند. و این امر ایجاب می‌کند که $r \le n$.

تعریف می‌کنیم

$$V^{(j)} = \frac{1}{\mu_j} W^{(j)} \qquad j = 1, \dots, r \qquad (۲۳.۲.۷)$$

این یک مجموعهٔ یکا متعامد در C^n است. اگر $r < n$، آنگاه $V^{(r+1)}, \dots, V^{(m)}$ را طوری انتخاب می‌کنیم که $\{V^{(1)}, \dots, V^{(n)}\}$ یک پایهٔ یکا متعامد در C^n باشد. تعریف می‌کنیم

$$V = [V^{(1)}, \dots, V^{(n)}] \qquad (۲۴.۲.۷)$$

به‌سادگی معلوم می‌شود که V یک ماتریس یکانی از مرتبهٔ $n \times n$ است و مستقیماً می‌توان تحقیق کرد که $VF = W$، که F به‌صورت (۱۵.۲.۷) است. بنابراین

$$VF = AU$$

که (۱۴.۲.۷) را ثابت می‌کند. برهان r بودن رتبهٔ (A) و به‌دست آوردن سایر ویژگیهای تجزیهٔ مقدار تکین به مسألهٔ ۱۹ واگذار شده است. تجزیهٔ تکین ـ مقدار در فصل ۹، در حل کمترین مربعات دستگاههای خطی فرا معیّن به‌کار برده شده است. ■

برای عرضهٔ اساسیترین شکل متعارف، نماد زیر را وارد می‌کنیم. ماتریس $n \times n$ زیر را تعریف می‌کنیم

$$J_n(\lambda) = \begin{bmatrix} \lambda & ۱ & ۰ & \dots & ۰ \\ ۰ & \lambda & ۱ & & \\ \cdot & & \cdot & & \\ \cdot & & & \cdot & ۱ \\ \cdot & & & & \cdot \\ ۰ & \cdot & \cdot & \cdot & \lambda \end{bmatrix} \quad n \geq ۱ \qquad (۲۵.۲.۷)$$

که $J_n(\lambda)$ دارای ویژه مقدار λ با چندگانگی جبری n و چندگانگی هندسی ۱ است. این $J_n(\lambda)$ **بلوک ژوردان** نامیده می‌شود.

قضیهٔ ۶.۷ (شکل متعارف ژوردان) گیریم A از مرتبهٔ n باشد. آنگاه یک ماتریس ناتکین P وجود دارد که برای آن

$$P^{-۱}AP = \begin{bmatrix} J_{n_۱}(\lambda_۱) & & & ۰ \\ & J_{n_۲}(\lambda_۲) & & \\ & & \ddots & \\ ۰ & & & J_{n_r}(\lambda_r) \end{bmatrix} \qquad (۲۶.۲.۷)$$

ویژه‌مقدارهای $\lambda_۱$، ...، λ_rلازم نیست متمایز باشند. اگر A ارمیتی باشد، قضیهٔ ۴.۷ ایجاب می‌کند $n_۱ = n_۲ = \dots = n_r = ۱$، زیرا در این حالت مجموع چندگانگیهای هندسی باید برابر n، مرتبهٔ ماتریس A، باشد.

اغلب مناسبتر است که (۲۶.۲.۷) را به شکل

$$P^{-۱}AP = D + N$$
$$D = \operatorname{diag}[\lambda_۱, \dots, \lambda_r] \qquad (۲۷.۲.۷)$$

بنویسیم، که هر λ_i، n_i بار، بر قطر D ظاهر می‌شود. درایه‌های ماتریس N همه صفرند جز احتمالاً ۱ های واقع بر بالای قطر. این ماتریس یک ماتریس **پوچتوان** است و به عبارت دقیقتر، در رابطهٔ زیر صدق می‌کند

$$N^n = ۰ \qquad (۲۸.۲.۷)$$

قضیهٔ شکل متعارف ژوردان به‌سادگی قابل اثبات نیست، و برای مطالعهٔ بیشتر این موضوع پربار، خوانندگان را به کتابهای جبر خطی زیادی که نوشته‌شده ارجاع می‌دهیم [برای مثال فرانکلین (۱۹۶۸، فصل ۵)؛ هالموس (۱۹۵۸، بخش ۵۸)؛ یا، نوبل (۱۹۶۹، فصل ۱۱) را ببینید].

۳.۷ نرمهای برداری و ماتریسی

نرم اقلیدسی $\|x\|_2$ قبلاً معرّفی شده است و اغلب آن را برای اندازه‌گیری بزرگی بردار به‌کار می‌برند. ولی موارد زیادی هستند که اندازه‌گیری بزرگی بردار از راههای دیگری مناسبتر است. از این‌رو مفهوم کلی نرم بردار را تعریف می‌کنیم.

تعریف گیریم V یک فضای برداری باشد و $N(x)$ تابعی حقیقی ـ مقدار که در V تعریف شده است. در این صورت $N(x)$ یک نرم است اگر:

(N۱) برای همهٔ مقادیر $x \in V$ $N(x) \geq 0$، و $N(x) = 0$ اگر و تنها اگر $x = 0$.

(N۲) برای همهٔ مقادیر $x \in V$ و تمام اسکالرهای α، $N(\alpha x) = |\alpha| N(x)$

(N۳) برای همهٔ مقادیر $x, y \in V$، $N(x+y) \leq N(x) + N(y)$

$N(x) = \| x \|$ نماد معمولی است. نماد $N(x)$ به‌کار برده شده تا تأکید شود که نرم، تابعی است با حوزهٔ V و برد اعداد حقیقی نامنفی. فاصلهٔ x از y را با $\| x - y \|$ تعریف می‌کنیم. نتایج سادهٔ این تعریف، **نابرابری مثلثی** در شکل دیگر آن

$$\| x - z \| \leq \| x - y \| + \| y - z \|$$

و **معکوس نابرابری مثلثی** زیر است.

$$|\| x \| - \| y \|| \leq \| x - y \| \qquad x, y \in V \qquad (۱.۳.۷)$$

مثال ۱. pـ نرم را برای $1 \leq p < \infty$ چنین تعریف می‌کنیم

$$\| x \|_p = \left[\sum_{1}^{n} | x_j |^p \right]^{1/p} \qquad x \in \mathbf{C}^n \qquad (۲.۳.۷)$$

۲. نرم ماکسیمم چنین است

$$\| x \|_\infty = \max_{1 \leq j \leq n} | x_j | \qquad x \in C^n \qquad (۳.۳.۷)$$

انگیزهٔ استفاده از زیرنمایهٔ ∞ برای نرم، نتیجهٔ مسألهٔ ۲۳ است.

۳. برای فضای برداری $V = C[a,b]$، نرمهای تابعی $\| f \|_۲$ و $\| f \|_\infty$ به‌ترتیب در فصلهای ۴ و ۱ معرفی شده‌اند.

مثال بردار $x = (۱, ۰, -۲, ۱)$ را در نظر می‌گیریم. آنگاه

$$\| x \|_۱ = ۴ \qquad \| x \|_۲ = \sqrt{۶} \qquad \| x \|_\infty = ۲$$

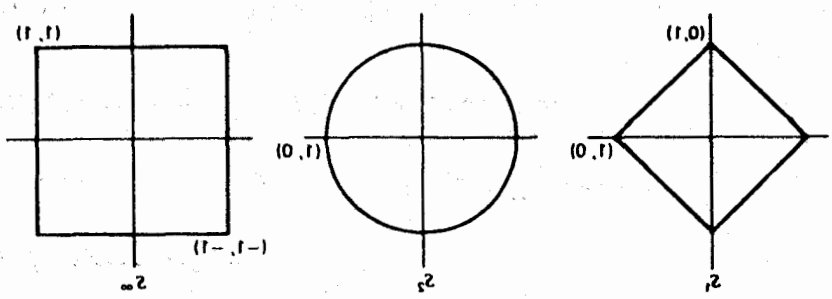
اثبات نُرم‌بودن $\| \cdot \|_p$ برای یک p ی کلی، بدیهی نیست. اثبات حالتهای $p = ۱$ و $p = \infty$ ساده است و $\| \cdot \|_۲$ در بخش ۱.۴ بررسی شده است. ولی برای $۱ < p < \infty$ و $p \neq ۲$ مشکل است نشان دهیم که $\| \cdot \|_p$ در نابرابری مثلّثی صدق می‌کند. این مطلب برای ما مسألهٔ مهمی نیست زیرا حالتهای عمدهٔ مورد نظر ما $p = ۱, ۲, \infty$ هستند. برای آنکه یک شهود هندسی از این نرمها به‌دست دهیم، **دایره‌های واحد**

$$S_p = \{x \in \mathbf{R}^۲ \mid \| x \|_p = ۱\} \qquad p = ۱, ۲, \infty \qquad (۴.۳.۷)$$

را در شکل ۲.۷ رسم کرده‌ایم.

اکنون چند قضیّه مربوط به نرمهای مختلف را ثابت می‌کنیم. با قضیّهٔ زیر در مورد پیوستگی $N(x) = \| x \|$، به‌صورت تابعی از x آغاز می‌کنیم.

لم گیریم $N(x)$ یک نرم در C^n (یا R^n) باشد. آنگاه $N(x)$ یک تابع پیوسته از $x_۱$، $x_۲$، ...، x_n، مؤلفه‌های بردار x است.



شکل ۲.۷ دایرهٔ واحد S_p با استفاده از نرم‌برداری $\| \cdot \|_p$

برهان می‌خواهیم نشان دهیم که

$$x_i \doteq y_i \qquad i = ۱, ۲, \ldots, n$$

ایجاب می‌کند که

$$N(x) \doteq N(y)$$

با استفاده از معکوس نابرابری مثلثی (۱.۳.۷)

$$| N(x) - N(y) | \leq N(x - y) \qquad x, y \in C^n$$

با توجّه به (۱.۱.۷)، تعریف پایهٔ استانده $\{e^{(۱)}, \ldots, e^{(n)}\}$ را برای C^n به خاطر می‌آوریم. پس

$$x - y = \sum_{j=۱}^{n} (x_j - y_j) e^{(j)}$$

$$N(x - y) \leq \sum_{j=۱}^{n} | x_j - y_j | N(e^{(j)}) \leq \| x - y \|_\infty \sum_{j=۱}^{n} N(e^{(j)})$$

$$| N(x) - N(y) | \leq c \| x - y \|_\infty \qquad c = \sum_{j=۱}^{n} N(e^{(j)}) \qquad (۵.۳.۷)$$

و این رابطه اثبات را کامل می‌کند. ■

باید توجه داشت که این رابطه همچنین ثابت می‌کند که برای هر نرم برداری N در C^n یک $c > ۰$ وجود دارد به‌طوری که

$$N(x) \leq c \| x \|_\infty \qquad \text{برای هر } x \in C^n \qquad (۶.۳.۷)$$

تنها کافی است که در (۵.۳.۷)، y را مساوی صفر قرار دهیم. قضیهٔ زیر عکس این قضیّه است.

قضیهٔ ۷.۷ (هم‌ارزی نرمها) گیریم N و M نرمهایی در C^n یا R^n باشند. در این صورت ثابتی مانند $c_۱ > ۰$ و $c_۲$ وجود دارند به‌طوری که

$$c_۱ M(x) \leq N(x) \leq c_۲ M(x) \qquad \text{برای هر } x \in V \qquad (۷.۳.۷)$$

برهان کافی است حالتی را در نظر بگیریم که N دلخواه باشد و $M(x) = \| x \|_\infty$. ترکیب دو چنین حکمی قضیهٔ کلی را به‌دست خواهد داد. بنابراین می‌خواهیم نشان دهیم که ثابتهایی چون c_1 و c_2 وجود دارند به‌طوری که

$$c_1 \| x \|_\infty \leq N(x) \leq c_2 \| x \|_\infty \qquad (۸.۳.۷)$$

یا هم‌ارز با آن

$$c_1 \leq N(z) \leq c_2 \qquad \text{برای هر } z \in S \qquad (۹.۳.۷)$$

که در آن S مجموعهٔ تمام نقاط z در C^n است به‌طوری که $\| z \|_\infty = 1$. نامساوی سمت راست (۹.۳.۷) بلافاصله از (۶.۳.۷) نتیجه می‌شود.

توجه کنید که S یک مجموعهٔ بسته و کراندار در C^n است و N یک تابع پیوسته در S. یک قضیهٔ استانده در حساب دیفرانسیل و انتگرال پیشرفته داریم که N به ماکسیمم و مینیمم خود در S در نقاطی از S می‌رسد، یعنی ثابتهای c_1، c_2 و نقاط z_1، z_2 در S وجود دارند به‌طوری که

$$\text{برای هر } z \in S \qquad c_1 = N(z_1) \leq N(z) \leq N(z_2) = c_2$$

واضح است که $c_2 \geq 0$، c_1. و اگر $c_1 = 0$، آنگاه $N(z_1) = 0$. ولی در این صورت $z_1 = 0$، که با ساختمان S که در آن $\| z \|_\infty = 1$، متناقض است. این مطلب (۹.۳.۷) را ثابت می‌کند و برهان قضیه کامل می‌شود. **توجه:** این قضیه برای فضاهای نامتناهی ـ بعد تعمیم نمی‌یابد. ■

بسیاری از روشهای عددی برای مسائلی که متضمّن دستگاههای خطی هستند، دنباله‌ای از بردارهای $\{x^{(m)} \mid m \geq 0\}$ تولید می‌کنند و ما می‌خواهیم راجع به همگرایی این دنباله به یک بردار x صحبت کنیم.

تعریف دنبالهٔ بردارهای $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}, \ldots\}$ در C^n یا R^n، به بردار x **همگرا** خوانده می‌شود اگر و تنها اگر

$$\| x - x^{(m)} \| \to 0 \qquad \text{وقتی} \qquad m \to \infty$$

توجه نمایید که انتخاب نرم تصریح نشده است. برای فضاهای متناهی‌بعد، مهم نیست که چه نرمی به‌کار برده شده است. گیریم M و N دو نرم بر C^n باشند. در این صورت از (۷.۳.۷)،

$$c_1 M(x - x^{(m)}) \leq N(x - x^{(m)}) \leq c_2 M(x - x^{(m)}) \qquad m \geq 0$$

و $M(x-x^{(m)})$ به صفر همگرا می‌شود اگر و تنها اگر $N(x-x^{(m)})$ به صفر همگرا شود. بنابراین $x^{(m)}$ با نرم M به x همگرا می‌شود اگر و تنها اگر با نرم N همگرا باشد. این یک نتیجهٔ مهمّی است و برای فضاهای نامتناهی بعد درست نیست.

نرمهای ماتریسی مجموعهٔ تمام ماتریسهای $n \times n$ با درایه‌های مختلط را می‌توان هم‌ارز با فضای برداری C^{n^2} در نظر گرفت که یک عمل ضربی خاص به فضای برداری اضافه شده است. بنابراین نرم ماتریسی باید در سه شرط معمول $N1$ تا $N3$ صدق کند. به علاوه، می‌خواهیم دو شرط دیگر هم داشته باشد.

تعریف نرم ماتریسی در $N1-N3$ و شرایط زیر صدق می‌کند

$$\| AB \| \leq \| A \| \| B \| \quad \text{(N4)}.$$

(N5) معمولاً فضای برداری که باآن کار می‌کنیم، $\mathbf{R}^n$ یا $V = C^n$، یک نرم برداری دارد، که آن را $\| x \|_v$، $x \in V$ می‌نامیم. می‌خواهیم که نرمهای ماتریسی و برداری سازگار باشند:

$$\| Ax \|_v \leq \| A \| \| x \|_v \qquad \text{برای هر } A \text{ و هر } x \in V$$

مثال گیریم A ماتریس $n \times n$ باشد و $\| \cdot \|_v = \| \cdot \|_2$. در این صورت برای $x \in C^n$، با استفاده از نامساوی (۸.۱.۷) کوشی ـ شوارتس:

$$\| Ax \|_2 = \left[\sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right|^2 \right]^{1/2} \leq \left[\sum_{i=1}^{n} \left\{ \sum_{j=1}^{n} | a_{ij} |^2 \right\} \left\{ \sum_{j=1}^{n} | x_j |^2 \right\} \right]^{1/2}$$

پس

$$\| Ax \|_2 \leq F(A) \| x \|_2 \qquad F(A) = \left[\sum_{i,j=1}^{n} | a_{ij} |^2 \right]^{1/2} \qquad (۱۰.۳.۷)$$

$F(A)$ را **نرم فروبنیوس** A خوانند. ویژگی $N5$ با استفاده از (۱۰.۳.۷) مستقیماً ثابت می‌شود. ویژگیهای $N1$ تا $N3$ برقرارند زیرا $F(A)$ نرم اقلیدسی در C^{n^2} است. اثبات $N4$ باقی می‌ماند. با استفاده از نابرابری کوشی ـ شوارتس

$$F(AB) = \left[\sum_{i,j=1}^{n} \left|\sum_{k=1}^{n} a_{ik}b_{kj}\right|^2\right]^{1/2}$$
$$\leq \left[\sum_{i,j=1}^{n} \left\{\sum_{k=1}^{n} |a_{ik}|^2\right\}\left\{\sum_{k=1}^{n} |b_{kj}|^2\right\}\right]^{1/2}$$
$$= F(A)F(B)$$

پس $F(A)$ یک نرم‌ماتریسی، سازگار با نرم اقلیدسی است.

معمولاً وقتی یک فضای برداری با یک نرم $\|\cdot\|_v$ داده شده است نرم ماتریسی وابسته به آن با رابطهٔ زیر تعریف می‌شود

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|_v}{\|x\|_v} \qquad (۱۱.۳.۷)$$

این نرم را اغلب نرم **عملگر** می‌خوانند و با تعریفی که دارد در N_5 صدق می‌کند:

$$\|Ax\|_v \leq \|A\|\|x\|_v \qquad x \in \mathbf{C}^n \qquad (۱۲.۳.۷)$$

برای یک ماتریس A، نرم عملگر که بر اثر نرم برداری $\|x\|_p$ القا شده با نماد $\|A\|_p$ نمایش داده می‌شود. مهمترین موارد در جدول ۱.۷ داده شده‌اند، و طرز پیداکردن آنها بعداً داده شده است. تعریف زیر را برای تعریف $\|A\|_2$ نیاز داریم

تعریف گیریم A یک ماتریس دلخواه باشد. **طیف** A مجموعهٔ همهٔ ویژه‌مقدارهای A است که با $\sigma(A)$ نمایش داده می‌شود. **شعاع طیفی** ماکسیمم اندازهٔ این ویژه‌مقدارهاست، و با نماد زیر بیان می‌شود

$$r_\sigma(A) = \max_{\lambda\in\sigma(A)} |\lambda| \qquad (۱۳.۳.۷)$$

جدول ۱.۷ نرمهای برداری و نرمهای عملگر ماتریسی متناظر

نرم ماتریسی	نرم برداری
$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} \|a_{ij}\|$	$\|x\|_1 = \sum_{i=1}^{n} \|x_i\|$
$\|A\|_2 = \sqrt{r_\sigma(A^*A)}$	$\|x\|_2 = \left[\sum_{j=1}^{n} \|x_i\|^2\right]^{1/2}$
$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} \|a_{ij}\|$	$\|x\|_\infty = \max_{1\leq i\leq n} \|x_i\|$

برای آنکه نشان دهیم (۱۱.۳.۷) در حالت کلّی یک نرم است، ابتدا نشان می‌دهیم که متناهی است. از قضیه ۷.۷، ثابتهای $c_1, c_2 > 0$ با

$$c_1 \| x \|_2 \leq \| x \|_v \leq c_2 \| x \|_2 \qquad x \in \mathbf{C}^n$$

وجود دارند. بنابراین

$$\frac{\| Ax \|_v}{\| x \|_v} \leq \frac{c_2 \| Ax \|_2}{c_1 \| x \|_2} \leq \frac{c_2}{c_1} F(A)$$

که ثابت می‌کند $\| A \|$ متناهی است.

در اینجا جالب است که به معنای هندسی $\| A \|$ توجه نماییم

$$\| A \| = \operatorname{Sup}_{x \neq 0} \frac{\| Ax \|_v}{\| x \|_v} = \operatorname{Sup}_{x \neq 0} \left\| A \left(\frac{x}{\| x \|_v} \right) \right\|_v = \operatorname{Sup}_{\|z\|_v = 1} \| Az \|_v$$

با توجه به اینکه اگر $\| z \|_v \leq 1$ انتخاب شود این سوپرمم تغییر نخواهد کرد،

$$\| A \| = \operatorname{Sup}_{\|z\|_v \leq 1} \| Az \| \qquad (۱۴.۳.۷)$$

گیریم

$$B = \{ z \in C^n \mid \| z \| \leq 1 \}$$

گوی واحد نسبت به $\| \cdot \|_v$ باشد. پس

$$\| A \| = \operatorname{Sup}_{z \in B} \| Az \|_v = \operatorname{Sup}_{w \in A(B)} \| w \|_v$$

که $A(B)$ نگاشت B است، وقتی A بر آن اثر کرده باشد. بنابراین $\| A \|$ تأثیر A را بر گوی واحد اندازه‌گیری می‌نماید، و اگر $\| A \| > 1$، $\| A \|$ کشیدگی ماکسیمم گوی B را وقتی A بر آن عمل کرده است نشان می‌دهد.

برهان در زیر ثابت می‌کنیم که نرم عملگر $\|A\|$ یک نرم ماتریسی است.

۱. واضح است که $\| A \| \geq 0$، و اگر $A = 0$، آنگاه $\| A \| = 0$. برعکس، اگر $\| A \| = 0$، آنگاه $\| Ax \|_v = 0$ برای جمیع مقادیر x. بنابراین برای تمام مقادیر x، $Ax = 0$، که ایجاب می‌کند $A = 0$.

۲. گیریم α یک اسکالر باشد. آنگاه

$$\| \alpha A \| = \sup_{\|x\|_v \leq 1} \| \alpha Ax \|_v = \sup_{\|x\|_v \leq 1} | \alpha | \| Ax \|_v$$
$$= | \alpha | \sup_{\|x\|_v < 1} \| Ax \|_v = | \alpha | \| A \|$$

۳. برای هر $x \in \mathbf{C}^n$

$$\| (A+B)x \|_v = \| Ax + Bx \|_v \leq \| Ax \|_v + \| Bx \|_v$$

زیرا $\| \cdot \|_v$ یک نرم است. با استفاده از ویژگی (۱۲.۳.۷)

$$\| (A+B)x \|_v \leq \| A \| \| x \|_v + \| B \| \| x \|_v$$
$$\frac{\| (A+B)x \|_v}{\| x \|_v} \leq \| A \| + \| B \|$$

که ایجاب می‌کند

$$\| A+B \| \leq \| A \| + \| B \|$$

۴. برای هر $x \in \mathbf{C}^n$، (۱۲.۳.۷) را به‌کار می‌بریم تا به‌دست آوریم

$$\| (AB)x \|_v = \| A(Bx) \|_v \leq \| A \| \| Bx \|_v \leq \| A \| \| B \| \| x \|_v$$
$$\frac{\| ABx \|_v}{\| x \|_v} \leq \| A \| \| B \|$$

که ایجاب می‌کند

$$\| AB \| \leq \| A \| \| B \|$$

اکنون نتایج جدول ۱.۷ را به‌طور مفصّلتری شرح می‌دهیم

مثال ۱. از نرم برداری زیر استفاده می‌کنیم

$$\| x \|_1 = \sum_{j=1}^{n} | x_j | \qquad x \in \mathbf{C}^n$$

پس

$$\| Ax \|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} | a_{ij} | | x_j |$$

با تغییر ترتیب مجموعیابی، می‌توانیم جمعوندها را مجزّا کنیم

$$\| Ax \|_۱ \leq \sum_{j=۱}^{n} | x_j | \sum_{i=۱}^{n} | a_{ij} |$$

گیریم

$$c = \underset{۱\leq j\leq n}{\text{Max}} \sum_{i=۱}^{n} | a_{ij} | \qquad (۱۵.۳.۷)$$

پس

$$\| Ax \|_۱ \leq c \| x \|_۱$$

و بنابراین

$$\| A \|_۱ \leq c$$

برای اینکه ثابت کنیم این یک تساوی است x ی پیدا می‌کنیم که:

$$\frac{\| Ax \|_۱}{\| x \|_۱} = c$$

گیریم k اندیس ستونی باشد که برای آن ماکسیمم (۱۵.۳.۷) حاصل شده است. فرض می‌کنیم $x = e^{(k)}$، k امین بردار یکه باشد. پس $\| x \|_۱ = ۱$ و

$$\| Ax \|_۱ = \sum_{i=۱}^{n} \Big| \sum_{j=۱}^{n} a_{ij} x_j \Big| = \sum_{i=۱}^{n} | a_{ik} | = c$$

این ثابت می‌کند که برای نرم برداری $\| \cdot \|_۱$، نرم عملگر چنین است

$$\| A \|_۱ = \underset{۱\leq j\leq n}{\text{Max}} \sum_{i=۱}^{n} | a_{ij} | \qquad (۱۶.۳.۷)$$

این رابطه اغلب نرم ستونی خوانده می‌شود.

۲. برای $\mathbf{C}^n$ با نرم $\| x \|_\infty$، نرم عملگر چنین است

$$\| A \|_\infty = \underset{۱\leq i\leq n}{\text{Max}} \sum_{j=۱}^{n} | a_{ij} | \qquad (۱۷.۳.۷)$$

این نرم سطری A خوانده می‌شود. اثبات این فرمول به مسألهٔ ۲۵ واگذار شده است، گرچه مشابه با اثبات $\| A \|_1$ است.

۳. نرم $\| x \|_2$ در $\mathbf{C}^n$ را به‌کار می‌بریم. از (۱۰.۳.۷) نتیجه می‌گیریم که

$$\| A \|_2 \leq F(A) \qquad (۱۸.۳.۷)$$

در حالت کلی تساوی برقرار نیست. برای مثال، با $A = I$، ماتریس همانی، (۱۰.۳.۷) و (۱۱.۳.۷) را به‌کار می‌بریم تا به‌دست آوریم

$$F(I) = \sqrt{n} \qquad \| I \|_2 = 1$$

ثابت می‌کنیم

$$\| A \|_2 = \sqrt{r_\sigma(A^*A)} \qquad (۱۹.۳.۷)$$

همان‌گونه که قبلاً در جدول ۱.۷ بیان کردیم. ماتریس A^*A ارمیتی است و همهٔ ویژه‌مقدارهای آن، همان‌گونه که در اثبات قضیهٔ ۵.۷ نشان داده شد، نامنفی‌اند. گیریم که ویژه‌مقدارهای آن

$$\lambda_1 \geq \lambda_2 \geq \ldots \leq \lambda_n \geq 0$$

باشند که برحسب چندگانگی آنها شمرده شده‌اند، و گیریم $u^{(1)}, \ldots, u^{(n)}$ ویژه‌بردارهای متناظر با آنها باشند که همچون یک پایهٔ یکا متعامد C^n مرتب شده‌اند.

برای یک $x \in \mathbf{C}^n$،

$$\| Ax \|_2^2 = (Ax, Ax) = (x, A^*Ax)$$

x را به صورت زیر می‌نویسیم

$$x = \sum_{j=1}^{n} \alpha_j u^{(j)} \qquad \alpha_j \equiv (x, u^{(j)}) \qquad (۲۰.۳.۷)$$

پس

$$A^*Ax = \sum_{j=1}^{n} \alpha_j A^*Au^{(j)} = \sum_{j=1}^{n} \alpha_j \lambda_j u^{(j)}$$

و

$$\| Ax \|_2^2 = \left(\sum_{i=1}^{n} \alpha_i u^{(i)}, \sum_{j=1}^{n} \alpha_j \lambda_j u^{(j)} \right) = \sum_{j=1}^{n} \lambda_j \mid \alpha_j \mid^2$$

$$\leq \lambda_1 \sum_{j=1}^{n} \mid \alpha_j \mid^2 = \lambda_1 \| x \|_2^2$$

که از (۲۰.۳.۷) برای محاسبهٔ $\| x \|_۲$ استفاده کرده‌ایم. بنابراین

$$\| A \|_۲^۲ \leq \lambda_۱$$

تساوی با توجه به $x = u^{(۱)}$ به‌دست می‌آید، لذا $\| x \|_۲ = ۱$ و

$$\| Ax \|_۲^۲ = (x, A^*Ax) = (u^{(۱)}, \lambda_۱ u^{(۱)}) = \lambda_۱$$

رابطهٔ (۱۹.۳.۷) از اینجا ثابت می‌شود، زیرا $\lambda_۱ = r_\sigma(A^*A)$. می‌توان نشان داد که A^*A و AA^* ویژه‌مقدارهای ناصفر متساوی دارند (مسألهٔ ۱۹ را ببینید)، بنابراین؛ $r_\sigma(AA^*) = r_\sigma(A^*A)$، که شکل دیگری از فرمول (۱۹.۳.۷) است. این رابطه همچنین ثابت می‌کند که

$$\| A \|_۲ = \| A^* \|_۲ \qquad (۲۱.۳.۷)$$

این تساوی برای نرمهای ماتریسی قبلی صادق نیست.

نسبتاً به آسانی می‌توان نشان داد که اگر A ارمیتی باشد، آنگاه

$$\| A \|_۲ = r_\sigma(A) \qquad (۲۲.۳.۷)$$

برهان درستی این رابطه به مسألهٔ ۲۷ ارجاع شده است.

مثال ماتریس زیر را در نظر می‌گیریم

$$A = \begin{bmatrix} ۱ & -۲ \\ -۳ & ۴ \end{bmatrix}$$

پس

$$\| A \|_۱ = ۶ \qquad \| A \|_۲ = \sqrt{۱۵ + \sqrt{۲۲۱}} \doteq ۵٫۴۶ \qquad \| A \|_\infty = ۷ .$$

به‌عنوان مثالی از نابرابری (۲۳.۳.۷) در قضیهٔ زیر، داریم

$$r_\sigma(A) = \frac{۵ + \sqrt{۳۳}}{۲} \doteq ۵٫۳۷ < \| A \|_۲$$

قضیهٔ ۸.۷ گیریم A یک ماتریس دلخواه مربعی باشد. پس برای هر نرم عملگر ماتریسی،

$$r_\sigma(A) \leq \| A \| \qquad (۲۳.۳.۷)$$

به‌علاوه، اگر $\varepsilon > 0$ داده شده باشد، یک نرم عملگر ماتریسی که با $\| \cdot \|_\varepsilon$ نمایش داده می‌شود وجود دارد به‌طوری که

$$\| A \|_\varepsilon \leq r_\sigma(A) + \varepsilon \qquad (۲۴.۳.۷)$$

برهان برای اثبات (۲۳.۳.۷)، گیریم $\| \cdot \|$ یک نرم ماتریسی سازگار با نرم برداری $\| \cdot \|_v$ باشد. فرض می‌کنیم λ ویژه‌مقداری در $\sigma(A)$ باشد که برای آن

$$| \lambda | = r_\sigma(A)$$

و گیریم x یک ویژه بردار متناظر با آن باشد، $\| x \|_v = 1$. پس

$$r_\sigma(A) = | \lambda | = \| \lambda x \|_v \leq \| Ax \|_v \leq \| A \| \| x \|_v = \| A \|$$

که (۲۳.۳.۷) را ثابت می‌کند.

برهان (۲۴.۳.۷) یک ساختمان غیرنمایان است، و در اَیزکسن و کلر (۱۹۶۶، ص ۱۲) داده شده است. ■

فرع زیر یک نتیجهٔ ساده ولی مهم قضیهٔ ۸.۷ است.

فرع برای یک ماتریس مربعی A، $r_\sigma(A) < 1$ اگر و تنها اگر، به‌ازای یک نرم عملگر ماتریسی داشته باشیم، $\| A \| < 1$. ■

این فرع را می‌توان برای اثبات قضیهٔ ۹.۷ در بخش بعد به‌کار برد، ولی ما ترجیح می‌دهیم که از شکل متعارف ژوردان که در قضیهٔ ۶.۷ داده شده استفاده کنیم. نتیجهٔ (۲۲.۳.۷) و قضیهٔ ۸.۷ نشان می‌دهند که $r_\sigma(A)$ درواقع یک نرم ماتریسی است، و این فرع برای تحلیل نرخ همگرایی بعضی روشهای بارستی که در فصل ۸ برای حل دستگاه معادلات خطی داده شده است به‌کار برده خواهد شد.

۴.۷ قضایای همگرایی و اختلال

قضایای زیر چارچوبی نظری هستند که ما بعداً با استفاده از آنها، به تحلیلهای خطاها در روشهای عددیِ حل دستگاه معادلات خطی می‌پردازیم.

قضیهٔ ۹.۷ گیریم A ماتریس مربعی از مرتبهٔ n باشد. در این صورت A^m به ماتریس صفر همگرا خواهد بود اگر و تنها اگر $r_\sigma(A) < 1$.

برهان قضیهٔ ۶.۷ را به‌عنوان ابزار اصلی به‌کار می‌بریم. گیریم J شکل متعارف ژوردان برای A باشد،

$$P^{-1}AP = J$$

پس

$$A^m = (PJP^{-1})^m = PJ^mP^{-1} \qquad (۱.۴.۷)$$

و $A^m \to 0$ اگر و تنها اگر $J^m \to 0$. با توجه به (۲۷.۲.۷) و (۲۸.۲.۷) یادآور می‌شویم که J را می‌توان به شکل

$$J = D + N$$

نوشت که در آن

$$D = \operatorname{diag}[\lambda_1, \dots, \lambda_n]$$

شامل ویژه‌مقدارهای J (و A) است، و N ماتریسی است که برای آن

$$N^n = 0$$

بنابراین

$$J^m = (D+N)^m = \sum_{j=0}^{m} \binom{m}{j} D^{m-j}N^j$$

و با استفاده از $N^j = 0$ برای $j \geq n$،

$$J^m = \sum_{j=0}^{n} \binom{m}{j} D^{m-j}N^j \qquad (۲.۴.۷)$$

توجه داریم که توانهای D در رابطهٔ زیر صدق می‌کنند

$$m - j \geq m - n \to \infty \qquad \text{وقتی} \qquad m \to \infty \qquad (۳.۴.۷)$$

ما به حدود زیر نیاز داریم: برای هر مقدار مثبت $c < 1$ و هر $r \geq 0$،

$$\lim_{m\to\infty} m^r c^m = 0 \qquad (۴.۴.۷)$$

این رابطه را می‌توان با قاعدهٔ هوپیتال از راه حساب دیفرانسیل انتگرال مقدماتی ثابت کرد.

در (۲.۴.۷)، بدون توجه به بزرگی m، تعداد $n+1$ جملهٔ ثابت، وجود دارند، و ما می‌توانیم همگرایی J^m را با بررسی همگرایی هر یک از جملات آن بررسی کنیم. فرض کنید که $r_\sigma(k) < 1$، می‌دانیم که هر $|\lambda_i| < 1$، $i = 1, 2, \ldots, n$. و برای هر نرم ماتریسی

$$\left\| \binom{m}{j} D^{m-j} N^j \right\| \leq \frac{m^j}{j!} \| N \|^j \| D^{m-j} \|$$

با استفاده از نرم سطری، عبارت فوق با مقدار زیر کراندار می‌شود

$$\frac{1}{j!} \| N \|_\infty^j \, m^j [r_\sigma(A)]^{m-j}$$

که با استفاده از (۳.۴.۷) و (۴.۴.۷) به‌ازای $0 \leq j \leq n$، وقتی $m \to \infty$، به صفر می‌گراید. تا اینجا نیمی از قضیه ثابت شده است، یعنی اگر $r_\sigma(A) < 1$، آنگاه J^m و A^m با توجّه به (۱.۴.۷)، وقتی $m \to \infty$، به صفر همگرا هستند.

فرض کنید $r_\sigma(A) \geq 1$. گیریم λ یک ویژه مقدار A با $|\lambda| \geq 1$، و x ویژه‌بردار متناظر با آن، $x \neq 0$، باشند. پس

$$A^m x = \lambda^m x$$

و روشن است که وقتی $m \to \infty$ این عبارت به صفر همگرا نخواهد بود. بنابراین ممکن نیست که $A^m \to 0$، زیرا که لازم می‌آید $A^m x \to 0$؛ و اثبات کامل می‌شود. ■

قضیهٔ ۱۰.۷ (سری هندسی) گیریم A یک ماتریس مربعی باشد. اگر $r_\sigma(A) < 1$، آنگاه $(I-A)^{-1}$ وجود دارد و می‌توان آن را با سری همگرای زیر بیان کرد

$$(I-A)^{-1} = I + A + A^2 + \ldots + A^m + \ldots \qquad (۵.۴.۷)$$

به عکس، اگر سری (۵.۴.۷) همگرا باشد، آنگاه $r_\sigma(A) < 1$.

برهان فرض می‌کنیم $r_\sigma(A) < 1$. وجود $(I-A)^{-1}$ را با اثبات عبارت هم‌ارز (۳) از قضیهٔ ۲.۷ نشان خواهیم داد. فرض می‌کنیم

$$(I-A)x = 0$$

در این صورت $Ax = x$، و از اینجا لازم می‌آید که اگر $x \neq 0$، ۱، یک ویژه‌مقدار A باشد. ولی ما فرض کردیم $r_\sigma(A) < 1$ پس باید داشته باشیم $x = 0$ که به نتیجهٔ اثبات وجود $(I - A)^{-1}$ می‌رسیم.

ما به اتحاد زیر نیاز داریم

$$(I - A)(I + A + A^2 + \ldots + A^m) = I - A^{m+1} \qquad (۶.۴.۷)$$

که برای هر ماتریس A درست است. اگر طرفین را در $(I - A)^{-1}$ ضرب کنیم خواهیم داشت:

$$I + A + \ldots + A^m = (I - A)^{-1}(I - A^{m+1})$$

طرف چپ دارای حد است اگر طرف راست حد داشته باشد. طبق قضیهٔ ۹.۷، $r_\sigma(A) < 1$ ایجاب می‌کند $A^{m+1} \to 0$ وقتی $m \to \infty$. بنابراین نتیجهٔ (۵.۴.۷) به‌دست می‌آید.

به‌عکس، فرض می‌کنیم که این سری هندسی همگراست و آن را با

$$B = I + A + A^2 + \ldots + A^m + \ldots$$

نشان می‌دهیم. پس $B - AB = B - BA = I$، و بنابراین $I - A$ معکوس دارد، و B معکوس آن است. از طرفین رابطهٔ (۶.۴.۷) حد می‌گیریم، طرف چپ دارای حد $(I - A)B = I$ است، و بنابراین حد طرف راست هم باید I باشد. ولی لازمهٔ آن این است که

$$A^{m+1} \to 0 \qquad \text{وقتی } m \to \infty$$

به موجب قضیهٔ ۹.۷ باید داشته باشیم $r_\sigma(A) < 1$. ■

قضیهٔ ۱۱.۷ گیریم A یک ماتریس مربعی باشد. اگر نرم ماتریسی عملگری، $\| A \| < 1$، آنگاه $(I - A)^{-1}$ وجود دارد و دارای بسط هندسی (۵.۴.۷) است. به‌علاوه

$$\| (I - A)^{-1} \| \leq \frac{1}{1 - \| A \|} \qquad (۷.۴.۷)$$

برهان چون $\| A \| < 1$، از (۲۳.۳.۷) و قضیهٔ ۸.۷ نتیجه می‌شود که $r_\sigma(A) < 1$. غیر از (۷.۴.۷) بقیهٔ حکمها از قضیهٔ ۱۰.۷ نتیجه می‌شوند. برای (۷.۴.۷) فرض می‌کنیم

$$B_m = I + A + \ldots + A^m$$

از (۶.۴.۷) نتیجه می‌شود،

$$B_m = (I - A)^{-1}(I - A^{m+1})$$
$$(I - A)^{-1} - B_m = (I - A)^{-1}[I - (I - A^{m+1})]$$
$$= (I - A)^{-1}A^{m+1} \qquad (۸.۴.۷)$$

با استفاده از عکس نابرابری مثلثی،

$$\left| \| (I - A)^{-1} \| - \| B_m \| \right| \leq \| (I - A)^{-1} - B_m \|$$
$$< \| (I - A)^{-1} \| \| A \|^{m+1}$$

چون طرف راست وقتی $m \to \infty$ به صفر میل می‌کند، داریم

$$\| B_m \| \to \| (I - A)^{-1} \| \quad \text{وقتی} \quad m \to \infty \qquad (۹.۴.۷)$$

با توجّه به تعریف B_m و ویژگیهای نرم ماتریسی،

$$\| B_m \| \leq \| I \| + \| A \| + \ldots + \| A \|^m$$
$$= \frac{1 - \| A \|^{m+1}}{1 - \| A \|} \leq \frac{1}{1 - \| A \|}$$

که از ترکیب با (۹.۴.۷) برهان (۷.۴.۷) پایان می‌یابد. ■

قضیهٔ ۱۲.۷ گیریم A و B ماتریسهای مربعی هم مرتبه باشند. فرض می‌کنیم A ناتکین باشد و

$$\| A - B \| < \frac{1}{\| A^{-1} \|} \qquad (۱۰.۴.۷)$$

پس B نیز ناتکین است،

$$\| B^{-1} \| \leq \frac{\| A^{-1} \|}{1 - \| A^{-1} \| \| A - B \|} \qquad (۱۱.۴.۷)$$

و

$$\| A^{-1} - B^{-1} \| \leq \frac{\| A^{-1} \|^2 \| A - B \|}{1 - \| A^{-1} \| \| A - B \|} \qquad (۱۲.۴.۷)$$

برهان به اتحاد زیر توجه کنید

$$B = A - (A - B) = A[I - A^{-1}(A - B)] \qquad (۱۳.۴.۷)$$

با استفاده از قضیهٔ ۱۱.۷، ماتریس $[I - A^{-1}(A - B)]$، بر اساس نامساوی (۱۰.۴.۷)، ناتکین است که ایجاب می‌کند

$$\| A^{-1}(A - B) \| \leq \| A^{-1} \| \| A - B \| < 1$$

چون B حاصلضرب ماتریسهای ناتکین است، پس خود نیز ناتکین است و

$$B^{-1} = [I - A^{-1}(A - B)]^{-1} A^{-1}$$

با گرفتن نرم از دو طرف و استفاده از قضیهٔ ۱۱.۷، کران (۱۱.۴.۷) به‌دست می‌آید. برای اثبات (۱۲.۴.۷) از

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$$

استفاده می‌کنیم. باز از دو طرف نرم می‌گیریم و (۱۱.۴.۷) را به‌کار می‌بریم. ■

این قضیه از جهات مختلف مهم است. ولی در حال حاضر، مطلب مورد نظر ما بیان این نکته است که تمام اختلالهای به اندازه کافی کوچک یک ماتریس ناتکین، ماتریسهایی ناتکین می‌دهند.

مثال قضیهٔ ۱۱.۷ را با ملاحظهٔ وارونپذیری ماتریس زیر توضیح می‌دهیم

$$A = \begin{bmatrix} ۴ & ۱ & ۰ & ۰ & \cdots & ۰ \\ ۱ & ۴ & ۱ & ۰ & & ۰ \\ ۰ & ۱ & ۴ & ۱ & & \vdots \\ \vdots & & & \ddots & & \\ & & & ۱ & ۴ & ۱ \\ ۰ & \cdots & & ۰ & ۱ & ۴ \end{bmatrix}$$

A را مجدّداً به صورت زیر می‌نویسیم

$$A = ۴(I + B)$$

$$B = \begin{bmatrix} ۰ & \frac{۱}{۴} & ۰ & ۰ & \cdots & ۰ \\ \frac{۱}{۴} & ۰ & \frac{۱}{۴} & ۰ & & ۰ \\ ۰ & & & \ddots & & \vdots \\ \vdots & & & & & \frac{۱}{۴} \\ ۰ & \cdots & & & \frac{۱}{۴} & ۰ \end{bmatrix}$$

با استفاده از نرم سطری (۱۷.۳.۷)، $\| B \|_\infty = \frac{۱}{۲}$. بنابراین طبق قضیهٔ ۱۱.۷، $(I + B)^{-۱}$ وجود دارد و با توجّه به (۷.۴.۷)،

$$\| (I + B)^{-۱} \|_\infty \leq \frac{۱}{۱ - \frac{۱}{۲}} = ۲$$

لذا $A^{-۱}$ موجود است، $A^{-۱} = \frac{۱}{۴}(I + B)^{-۱}$ و

$$\| A^{-۱} \|_\infty \leq \frac{۱}{۲}$$

با استفاده از نرم سطری و نابرابری (۲۳.۳.۷)،

$$r_\sigma(A) \leq ۶ \qquad r_\sigma(A^{-۱}) \leq \frac{۱}{۲}$$

چون ویژه‌مقدارهای $A^{-۱}$ عکس ویژه مقدارهای A هستند (مسألهٔ ۲۷ را ببینید)، و چون A ارمیتی است ویژه‌مقدارهای آن حقیقی‌اند، کران زیر را خواهیم داشت:

$$\text{جمیع مقادیر } \lambda \in \sigma(A) \qquad ۲ \leq |\lambda| \leq ۶$$

برای کرانهای بهتر در این حالت، قضیهٔ دایرهٔ گرشگورین[1] فصل ۹ را ببینید.

بحث در آثار خواندنی

موضوع این فصل جبر خطی است، که به‌خصوص برای استفاده در به‌دست آوردن و تحلیل

1. Gerschgorin

روشهای جبر خطی عددی برگزیده شده است. کتابهای آنتون (۱۹۸۴) و استرنگ (۱۹۸۰) کتابهای درسی در سطح مقدماتی برای جبرخطی دورهٔ کارشناسی هستند. کتاب فرانکلین (۱۹۶۸) یک کتاب مقدماتی در سطح بالاتری برای نظریهٔ ماتریسهاست و کتاب هالموس (۱۹۵۸) یک متن درسی معروفی در جبر خطی مجرد است. کتاب نوبل (۱۹۶۹) یک کتاب درسی جبر خطی کاربردی در زمینه‌های گسترده است. معرفی مبانی جبرخطی در کتابهای فادیوا (۱۹۵۹)، گلوب و وَن‌لون[1] (۱۹۸۲)، پارلت[2] (۱۹۸۰)، استوارت (۱۹۷۳) و ویلکینسن (۱۹۶۵) آمده است، همهٔ این کتابها کلاً برای جبر خطی عددی نوشته شده‌اند. برای اطّلاعات نظری بیشتر با جزئیات بیشتر و در سطح بالاتر، کارهای کلاسیک گانتماخر[3] (۱۹۶۰) و هاوس هولدر (۱۹۶۵) را ببینید. مراجع بیشتر در فهرست مراجع فصلهای ۸ و ۹ داده شده‌اند.

مراجع

Anton, H. (1984). *Elementary Linear Algebra*, 4th ed. Wiley, New York.

Fadeeva, V. (1959). *Computational Methods of Linear Algebra*. Dover, New York.

Franklin, J. (1968). *Matrix Theory*. Prentice-Hall, Englewood Cliffs, N.J.

Gantmacher, F. (1960). *The Theory of Matrices*, vols. I and II. Chelsea, New York.

Golub, G., and C. Van Loan (1983). *Matrix Computations*. Johns Hopkins Press, Baltimore.

Halmos, P. (1958). *Finite-Dimensional Vector Spaces*. Van Nostrand, Princeton, N.J.

Householder, A. (1965). *The Theory of Matrices in Numerical Analysis*. Ginn (Blaisdell), Boston.

Isaacson, E., and H. Keller (1966). *Analysis of Numerical Methods*. Wiley, New York.

Lawson, C., and R. Hanson (1974). *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, N.J.

Noble, B. (1969). *Applied Linear Algebra*. Prentice-Hall, Englewood Cliffs, N.J.

Parlett, B. (1980). *The Symmetric Eigenvalue Problem*. Prentice-Hall, Englewood Cliffs, N.J.

Stewart, G. (1973). *Introduction to Matrix Computations*. Academic Press, New York.

Strang, G. (1980). *Linear Algebra and Its Applications*, 2nd ed. Academic Press, New York.

---

1. Van Loan
2. Parlett
3. Gantmacher

Wilkinson, J. (1965). *The Algebraic Eigenvalue Problem*. Oxford Univ. Press, Oxford, England.

مسائل

۱. تعیین کنید که آیا مجموعه‌های بردارهای زیر وابسته‌اند یا مستقل.

(الف) $(۱,۲,-۱,۳),(۳,-۱,۱,۱),(۱,۹,-۵,۱۱)$

(ب) $(۱,۱,۰),(۰,۱,۱),(۱,۰,۱)$

۲. گیریم A و B و C ماتریسهایی به‌ترتیب از مرتبهٔ $m\times n$، $n\times p$ و $p\times q$ باشند.

(الف) قانون شرکتپذیری $(AB)C=A(BC)$ را ثابت کنید

(ب) ثابت کنید $(AB)^T=B^TA^T$

۳. (الف) دو ماتریس مربعی A و B بنویسید که $AB\neq BA$

(ب) ماتریسهای مربعی A و B با درایه‌های غیرصفر پیدا کنید که $AB=۰$، $BA\neq ۰$.

۴. گیریم A ماتریسی از مرتبهٔ $m\times n$ و r و c به‌ترتیب معرّف رتبه‌های سطری و ستونی A باشند. ثابت کنید که $r=c$. **راهنمایی:** برای راحتی، فرض کنید که اولین r سطر A مستقل‌اند و بقیهٔ سطرها وابسته به این rسطرند و به همین ترتیب برای اولین c ستون A. گیریم $\hat{A}$ معرّف ماتریس $r\times n$ حاصل از حذف $m-r$ سطر آخر A باشد و $\hat{r}$ و $\hat{c}$ به‌ترتیب رتبه‌های سطری و ستونی $\hat{A}$ باشند. روشن است که $\hat{r}=r$. همچنین ستونهای $\hat{A}$ عناصری از C^r هستند که دارای بعد r است، و بنابراین باید داشته باشیم $\hat{c}\leq r$. نشان دهید که $\hat{c}=c$، و بنابراین ثابت می‌شود که $c\leq r$. نابرابری $r\leq c$ از همین استدلال برای A^T نتیجه می‌شود و این دو نابرابری با هم نتیجه می‌دهند که $r=c$.

۵. هم‌ارزی احکام (۱) – (۴) و (۶) قضیهٔ ۲.۷ را ثابت کنید.

راهنمایی: قضیهٔ ۱.۷، نتیجهٔ مسألهٔ ۴، و تجزیهٔ (۶.۱.۷) را به‌کار برید.

۶. گیریم

$$f_n(x)=\det\begin{bmatrix} x & ۱ & ۰ & \dots & ۰\\ ۱ & x & ۱ & ۰ & ۰\\ ۰ & ۱ & x & ۱ & \vdots\\ \vdots & & & \ddots & ۰\\ & & & & ۱\\ ۰ & \dots & & ۰ & ۱ & x \end{bmatrix}$$

و مرتبهٔ ماتریس n باشد. همچنین تعریف می‌کنیم $f_0(x) \equiv 1$.

(الف) نشان دهید

$$f_{n+1}(x) = xf_n(x) - f_{n-1}(x) \qquad n \geq 1$$

(ب) نشان دهید

$$f_n(x) = S_n\left(\frac{x}{2}\right) \qquad n \geq 0$$

که $S_n(x)$ چندجمله‌یی نوع دوّم چبیشف از درجهٔ n است (مسألهٔ ۲۴ فصل ۴ را ببینید).

۷. گیریم A یک ماتریس مربعی از مرتبهٔ n با درایه‌های حقیقی باشد. تابع

$$q(x_1, \ldots, x_n) = (Ax, x) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j \qquad x \in \mathbf{R}^n$$

صورت درجهٔ دوّم حاصل از A نامیده می‌شود. این عبارت یک چندجمله‌یی درجهٔ دوّمی است با n متغیر x_1، ...، x_n و در ماکسیمم‌سازی یا مینیمم‌سازی تابع n متغیره ظاهر می‌شود.

(الف) ثابت کنید که اگر A یک ماتریس پادمتقارن باشد، آنگاه $q(x) \equiv 0$.

(ب) برای یک ماتریس مربعی A در حالت کلی، $A_1 = \frac{1}{2}(A + A^T)$ و $A_2 = \frac{1}{2}(A - A^T)$ را تعریف می‌کنیم. پس $A = A_1 + A_2$. نشان دهید که A_1 متقارن است و به‌ازای جمیع مقادیر $x \in \mathbf{R}^n$، $(Ax, x) = (A_1x, x)$. این رابطه نشان می‌دهد که، بدون از دست‌دادن کلیّت، ماتریس ضرایب A را برای یک صورت درجهٔ دوّم همیشه می‌توان متقارن فرض کرد.

۸. بردارهای متعامد زیر داده شده‌اند

$$u^{(1)} = (1, 2, -1) \qquad u^{(2)} = (1, 1, 3)$$

یک بردار سوّم $u^{(3)}$ پیدا کنید که $\{u^{(1)}, u^{(2)}, u^{(3)}\}$ یک پایهٔ $\mathbf{R}^3$ باشد. این پایه را به صورت پایه‌ای یکا در آورید.

۹. برای بردارستونی $w \in \mathbf{C}^n$ با $\| w \|_2 = \sqrt{w^*w} = 1$، ماتریس $n \times n$ زیر را تعریف می‌کنیم

$$A = I - 2ww^*$$

(الف) برای حالت خاص $w = [\frac{1}{3}, \frac{2}{3}, \frac{2}{3}]^T$، ماتریس A را پیدا کنید. تحقیق کنید که این ماتریس متقارن و متعامد است.

(ب) نشان دهید که در حالت کلی، چنین ماتریسهایی ارمیتی و یکانی هستند.

۱۰. گیریم W یک زیرفضای $\mathbf{R}^n$ باشد. برای $x \in \mathbf{R}^n$ تعریف می‌کنیم

$$\rho(x) = \inf_{y \in W} \| x - y \|_2$$

گیریم $\{u_1, \ldots, u_m\}$ یک پایهٔ یکا متعامد W با بعد m باشد. این پایه را به یک پایهٔ یکا متعامد $\{u_1, \ldots, u_m, \ldots, u_n\}$ تمام $\mathbf{R}^n$ گسترش دهید.

(الف) نشان دهید

$$\rho(x) = \left[\sum_{j=m+1}^{n} | (x, u_j) |^2 \right]^{1/2}$$

که به‌طور یکتا در

$$y = Px \qquad P = \sum_{j=1}^{m} u_j u_j^T$$

به اینفیمم می‌رسد. نقطهٔ y **تصویر قائم** x روی W خوانده می‌شود.

(ب) نشان دهید $P^2 = P$. چنین ماتریسی را **ماتریس تصویر** خوانند.

(ج) نشان دهید $P^T = P$

(د) نشان دهید برای هر $x, z \in R^n$، $(Px, z - Pz) = 0$

(ه) نشان دهید که به‌ازای جمیع مقادیر $x \in \mathbf{R}^n$، $\| x \|_2^2 = \| Px \|_2^2 + \| x - Px \|_2^2$. این شکلی از قضیه فیثاغورس است.

۱۱. ویژه‌مقدارها و ویژه‌بردارهای ماتریسهای زیر را به‌دست آورید

(الف) $\begin{bmatrix} 1 & 6 \\ 1 & 1 \end{bmatrix}$ (ب) $\begin{bmatrix} 1 & 1 \\ -1 & 3 \end{bmatrix}$

۱۲. گیریم $y \neq 0$ در $\mathbf{R}^n$ باشد و $A = yy^T$ را یک ماتریس $n \times n$ تعریف می‌کنیم. نشان دهید که $\lambda = 0$ ویژه مقداری دقیقاً از چندگانگی $n - 1$ است. تنها ویژه مقدار ناصفر آن چیست؟

۱۳. گیریم U یک ماتریس یکانی $n \times n$ باشد.

(الف) نشان دهید برای هر $x \in \mathbf{C}^n$، $\| Ux \|_2 = \| x \|_2$. با استفاده از این ویژگی نشان دهید که فاصلهٔ x از y برابر فاصلهٔ Ux از Uy است، که نشان می‌دهد تبدیلات یکانی بر $\mathbf{C}^n$ فاصلهٔ بین نقاط را حفظ می‌کنند.

(ب) گیریم U متعامد باشد، نشان دهید

$$(Ux, Uy) = (x, y) \qquad x, y \in \mathbf{R}^n$$

این رابطه نشان می‌دهد که تبدیلات متعامد $\mathbf{R}^n$، زاویهٔ بین خطوط را هم، آن‌گونه که در (۱۲.۱.۷) تعریف شد، حفظ می‌نمایند.

(ج) نشان دهید که همهٔ ویژه‌مقدارهای یک ماتریس یکانی طول واحد دارند.

۱۴. گیریم A یک ماتریس ارمیتی از مرتبهٔ n باشد. آن را **معین مثبت** خوانند اگر و تنها اگر، برای هر $x \neq 0$ در $\mathbf{C}^n$، $(Ax, x) > 0$. نشان دهید که A معین مثبت است اگر و تنها اگر همهٔ ویژه‌مقدارهای آن حقیقی و مثبت باشند.

راهنمایی: قضیهٔ ۴.۷ را به‌کار برید و برای بیان یک بردار دلخواه $x \in \mathbf{R}^n$، (Ax, x) را برحسب یک پایه از ویژه‌بردارها بسط دهید.

۱۵. گیریم A حقیقی و متقارن باشد و ویژه‌مقدارهای λ_1، ...، λ_n برحسب چندگانگی آنها تکرار شده باشند. با استفاده از یک پایهٔ یکا متعامد از ویژه‌بردارها نشان دهید که صورت درجهٔ دوم مسألهٔ ۷، $q(x) = (Ax, x)$، $x \in \mathbf{R}^n$، را می‌توان به شکل ساده‌تر زیر بدل کرد.

$$q(x) = \sum_{j=1}^{n} \alpha_j^2 \lambda_j$$

با $\{\alpha_j\}$ که از روی x تعیین شده است. با استفاده از این رابطه، نمودارهای ممکن

$$(Ax, x) = \text{ثابت}$$

را وقتی A از مرتبهٔ ۳ باشد پیدا کنید.

۱۶. فرض می‌کنیم A حقیقی، متقارن، معین مثبت و از مرتبهٔ n باشد. تعریف می‌کنیم

$$f(x) = \frac{1}{2} x^T A x - b^T x \qquad x, b \in \mathbf{R}^n$$

نشان دهید که مینیمم یکتای $f(x)$ از حل $Ax = b$ به‌ازای $\alpha = A^{-1}b$ به‌دست می‌آید.

۱۷. گیریم $f(x)$ تابعی حقیقی مقدار از $x \in R^n$ باشد، و فرض می‌کنیم $f(x)$ سه بار پیوسته مشتقپذیر نسبت به مؤلفه‌های x باشد. تعمیم قضیهٔ تیلر، ۵.۱ فصل ۱ به n متغیر را برای به‌دست آوردن

$$\begin{aligned} f(x) =& f(\alpha) + (x - \alpha)^T \nabla f(\alpha) \\ &+ \frac{1}{2}(x - \alpha)^T H(\alpha)(x - \alpha) + O(\| x - \alpha \|^3) \end{aligned}$$

به‌کار برید. در اینجا

$$\nabla f(x) = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right]^T$$

گرادیان f و

$$H(x) = \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right] \qquad 1 \le i, j \le n$$

ماتریس هسی $f(x)$ است. جملهٔ آخر نشان می‌دهد که به‌ازای مقادیر x نزدیک به α جملات باقیمانده از مضربی از $\| x - \alpha \|^3$ کوچکترند.

اگر بخواهیم α ماکسیمم یا مینیمم موضعی باشد، آنگاه یک شرط لازم برای آن، برقراری $\nabla f(x) = 0$ است. فرض می‌کنیم $\nabla f(\alpha) = 0$، نشان دهید α یک مینیمم موضعی اکید (یا یکتای) $f(x)$ است اگر و تنها اگر $H(\alpha)$ معین مثبت باشد، [توجه کنید که $H(x)$ همیشه متقارن است].

۱۸. رابطهٔ (۲۸.۲.۷) را ثابت کنید.

۱۹. نمادی را که در قضیهٔ ۵.۷ در تجزیهٔ تکین ـ مقدار ماتریس A به‌کار برده شد به‌خاطر آورید.

(الف) نشان دهید μ_1^2، ...، μ_r^2 ویژه‌مقدارهای ناصفر A^*A و AA^* هستند که به‌ترتیب با ویژه‌بردارهای، $U^{(1)}$، ...، $U^{(r)}$ و $V^{(1)}$، ...، $V^{(r)}$ متناظرند. بردار $U^{(j)}$ ستون jام U را مشخص می‌کند، همین‌طور است برای $V^{(j)}$ و V.

(ب) نشان دهید $AU^{(j)} = \mu_j V^{(j)}$، $A^*V^{(j)} = \mu_j U^{(j)}$، $1 \le j \le r$.

(ج) ثابت کنید $r = \text{rank}(A)$.

۲۰. برای هر چندجمله‌یی $p(x) = b_0 + b_1 x + \ldots + b_m x^m$ و برای هر ماتریس مربعی A تعریف می‌کنیم

$$p(A) = b_0 I + b_1 A + \ldots + b_m A^m$$

گیریم A ماتریسی باشد که شکل متعارف ژوردان آن ماتریس قطری

$$P^{-1}AP = D = \text{diag}[\lambda_1, \ldots, \lambda_n]$$

است. $f_A(\lambda)$ چندجمله‌یی مشخصهٔ A است، ثابت کنید $f_A(A) = 0$. (این قضیّه، **قضیه کیلی-همیلتن** است. و برای همهٔ ماتریسهای مربعی درست است، نه فقط برای ماتریسهایی که دارای شکل قطری متعارف ژوردان هستند).

راهنمایی: برای ساده‌کردن $f_A(A)$، $A = PDP^{-1}$ را به‌کار برید.

۲۱. روابط زیر را ثابت کنید

(الف) $\| x \|_\infty \le \| x \|_1 \le n \| x \|_\infty$

(ب) $\| x \|_\infty \leq \| x \|_2 \leq \sqrt{n} \| x \|_\infty$

(ج) $\| x \|_2 \leq \| x \|_1 \leq \sqrt{n} \| x \|_2$

۲۲. گیریم A ماتریس حقیقی ناتکینی از مرتبهٔ n، و $\| \cdot \|_v$ یک نرم برداری در $\mathbf{R}^n$ باشد. تعریف می‌کنیم

$$\| x \|_* = \| Ax \|_v \qquad x \in \mathbf{R}^n$$

نشان دهید $\| \cdot \|_*$ یک نرم برداری روی $\mathbf{R}^n$ است.

۲۳. نشان دهید

$$\lim_{p \to \infty} \left[\sum_{j=1}^{n} | x_i |^p \right]^{1/p} = \max_{1 \leq i \leq n} | x_i | \qquad x \in \mathbf{C}^n$$

این رابطه استفاده از نماد $\| x \|_\infty$ را برای طرف راست موجّه می‌سازد.

۲۴. برای هر نرم ماتریسی نشان دهید که (الف) $\| I \| \geq 1$، و (ب) $\| A^{-1} \| \geq (1/\| A \|)$. برای هر نرم عملگر، از (۱۱.۳.۷) فوراً نتیجه می‌شود که $\| I \| = 1$.

۲۵. فرمول (۱۷.۳.۷) را برای نرم عملگر ماتریسی $\| A \|_\infty$ به‌دست آورید.

۲۶. یک نرم برداری در $\mathbf{R}^n$ را با

$$\| x \| = \frac{1}{n} \sum_{j=1}^{n} | x_j | \qquad x \in \mathbf{R}^n$$

تعریف می‌کنیم. نرم عملگر ماتریسی مربوط به این نرم برداری چیست؟

۲۷. گیریم A یک ماتریس مربعی از مرتبهٔ $n \times n$ باشد.

(الف) اگر ویژه‌مقدارها و ویژه بردارهای A داده شده باشند آنها را برای ماتریسهای زیر پیدا کنید.

(۱) A^m، $m \geq 2$ (۲) A^{-1}، با فرض ناتکینی A، (۳) $A + cI$، $c =$ ثابت.

(ب) ثابت کنید اگر A ارمیتی باشد، آنگاه $\| A \|_2 = r_\sigma(A)$.

(ج) برای A ی دلخواه و U یکانی از همان مرتبه، نشان دهید $\| A \|_2 = \| UA \|_2 = \| AU \|_2$.

۲۸. گیریم A ماتریس مربعی از مرتبهٔ $n \times n$ باشد.

(الف) نشان دهید برای هر ماتریس یکانی U، $F(AU) = F(UA) = F(A)$.

(ب) اگر A ارمیتی باشد، نشان دهید که

$$F(A) = \sqrt{\lambda_1^2 + \ldots + \lambda_n^2}$$

که $\lambda_1, \ldots, \lambda_n$ ویژه‌مقدارهای A هستند که برحسب چندگانگی تکرار شده‌اند. به‌علاوه

$$\frac{1}{\sqrt{n}} F(A) \leq \| A \|_2 \leq F(A)$$

۲۹. نماد قضیهٔ ۵.۷ را به یاد آورده نشان دهید

$$\| A \|_2 = \mu_1 \qquad F(A) = \sqrt{\mu_1^2 + \ldots + \mu_r^2}$$

۳۰. گیریم A از مرتبهٔ $n \times n$ باشد. نشان دهید

$$| \text{اثر } (A) | \leq n r_\sigma(A)$$

اگر A متقارن و معین مثبت باشد، نشان دهید

$$\text{اثر } (A) \geq r_\sigma(A)$$

۳۱. نشان دهید که سری نامتناهی

$$I + A + \frac{A^2}{2!} + \ldots + \frac{A^n}{n!} + \ldots$$

برای هر ماتریس مربعی A همگراست. مجموع سری را با e^A نمایش می‌دهیم

(الف) اگر $A = P^{-1} B P$، نشان دهید $e^A = P^{-1} e^B P$

(ب) گیریم $\lambda_1, \ldots, \lambda_n$ ویژه‌مقدارهای A برحسب چندگانگی تکرار شده باشند، نشان دهید ویژه‌مقدارهای e^A عبارت‌اند از $e^{\lambda_1}, \ldots, e^{\lambda_n}$.

۳۲. ماتریس زیر را در نظر می‌گیریم

$$A = \begin{bmatrix} 6 & 1 & 1 & 0 & & \ldots & & & 0 \\ 1 & 6 & 1 & 1 & 0 & & & & \vdots \\ 1 & 1 & 6 & 1 & 1 & 0 & & & \\ 0 & 1 & 1 & 6 & 1 & 1 & 0 & & 0 \\ \vdots & & & & & & & & 1 \\ & & & & & & 1 & 1 & 6 & 1 \\ 0 & & \ldots & & & & 0 & 1 & 1 & 6 \end{bmatrix}$$

نشان دهید A ناتکین است. یک کران بالا برای $\| A^{-1} \|_\infty$ و $\| A^{-1} \|_2$ پیدا کنید.

۳۳. در به‌دست‌آوردن درونیاب برازای درجهٔ ۳ در بخش ۷.۳ از فصل ۳، لازم بود که دستگاه خطی $AM = D$ از (۲۱.۷.۳) را حل کنیم که در آن

$$A = \begin{bmatrix} \frac{h_1}{3} & \frac{h_1}{6} & 0 & \cdots & 0 \\ \frac{h_1}{6} & \frac{h_1 + h_2}{3} & \frac{h_2}{6} & & \vdots \\ \vdots & & \ddots & & \\ & & \frac{h_{m-1}}{6} & \frac{h_{m-1} + h_m}{3} & \frac{h_m}{6} \\ 0 & \cdots & 0 & \frac{h_m}{6} & \frac{h_m}{3} \end{bmatrix}$$

و همهٔ h_i ها، $i = 1, \ldots, m$ مثبت‌اند. با استفاده از یک یا چند قضیّهٔ بخش ۴.۷ نشان دهید A ناتکین است. به‌علاوه، کرانهای

$$\frac{1}{6}\text{Min}(h_i) \leq |\lambda| \leq \text{Max}\, h_i \qquad \lambda \in \sigma(A)$$

را برای ویژه‌مقدارهای A به‌دست آورید.

۳۴. گیریم A یک ماتریس مربعی با $A^m = 0$ برای $m \geq 2$ باشد. نشان دهید که $I - A$ ناتکین است. چنین ماتریس A را **پوچتوان** خوانند.

۸

حل عددی دستگاههای معادلات خطی

دستگاههای معادلات خطی در زمینه‌های وسیعی، هم مستقیماً در مدلسازی وضعیّتهای فیزیکی و هم غیرمستقیم در حل عددی سایر مدلهای ریاضی ظاهر می‌شوند. این کاربردها تقریباً در تمام زمینه‌های فیزیکی، زیست‌شناسی و علوم‌اجتماعی صورت می‌گیرند. به‌علاوه، دستگاههای خطی در مباحث زیر دخالت دارند: نظریهٔ بهینه‌سازی، حل دستگاه معادلات ناخطی، تقریب توابع، حل عددی مسائل مقدار مرزی برای معادلات دیفرانسیل معمولی، معادلات دیفرانسیل جزئی و معادلات انتگرالی، استنباط آماری؛ و مسائل متعدد دیگر. به دلیل اهمیّت بسیار گستردهٔ دستگاههای خطی، پژوهشهای فراوانی در حل عددی آنها به‌عمل آمده است. الگوریتمهای عالی برای عادیترین نوع مسائل دستگاههای خطی ایجاد شده و بعضی از آنها در این فصل تعریف، تحلیل و توضیح داده شده‌اند.

عادیترین نوع مسأله، حل دستگاه خطی مربعی

$$Ax = b$$

از مرتبهٔ متوسّط، با ضرایبی است که اغلب ناصفرند. این‌گونه دستگاههای خطی، از هر مرتبه، **چگال** خوانده می‌شوند. برای چنین دستگاههایی، ماتریس ضرایب A معمولاً باید در حافظهٔ اصلی رایانه نگهداری شود تا بتوان دستگاه را به‌طورکارا حل کرد، و بنابراین محدودیتهای حافظه در اغلب

رایانه‌ها، مرتبهٔ دستگاه را محدود می‌نماید. با کاهش سریع هزینهٔ حافظهٔ رایانه‌ها، دستگاههای خطّی خیلی بزرگی را می‌توان در بعضی ماشینها جا داد، ولی انتظار می‌رود برای اغلب ماشینهای کوچکتر، کرانهای بالای عملی از مرتبه‌ای به بزرگی ۱۰۰ تا ۵۰۰ برسد. اغلب الگوریتمها برای حل چنین دستگاههای چگال، براساس **حذف گاوسی** ساخته شده‌اند که در بخش ۱.۸ تعریف شده‌اند. از جنبهٔ نظری، این یک روش مستقیم است که اگر خطای گردکردن نادیده گرفته شود، پس از یک تعداد متناهی مرحلهٔ محاسباتی، جواب درست از حذف گاوسی به‌دست می‌آید. جرح و تعدیل برای بهبود رفتار خطا در حذف گاوسی، شکلهای متنوع بعضی دسته‌های خاص ماتریسها، و تحلیلهای خطا در بخش ۲.۸ تا ۵.۸ داده شده‌اند.

نوع مهم دیگر مسأله، حل $Ax = b$ است، وقتی که A مربعی، با درایه‌های تُنُک، و از مرتبهٔ بزرگ باشد. ماتریس تُنُک ماتریسی است که بیشتر درایه‌های آن صفر باشند. چنین دستگاههایی به شکلهای گوناگونی پیدا می‌شوند، ولی ما بحث خود را به آنهایی که الگویی ساده و معلوم برای عناصر غیرصفر دارند، محدود می‌سازیم. این دستگاهها معمولاً در حل عددی معادلات دیفرانسیل جزئی ظاهر می‌شوند، و یک مثال در بخش ۸.۸ داده شده است. به علت مرتبهٔ بزرگ اغلب دستگاههای معادلات خطی تُنُک، گاهی به بزرگی 10^5 یا بیشتر، دستگاه خطی را نمی‌توان با روشهایی مستقیم مانند حذف گاوسی حل کرد. روشهای بارستی، روشهای ارجح‌اند و این روشها در بخش ۶.۸ تا ۹.۸ معرفی شده‌اند.

برای حل دستگاههای مربعی چگال از مرتبهٔ متوسط، اغلب مراکز رایانه‌یی مجموعه برنامه‌هایی دارند که می‌توان برای انواع مسائل به‌کار برد. دانشجویان باید با این برنامه‌ها در مرکز رایانهٔ دانشگاه خود آشنا شده و آنها را برای بیشتر روشن‌کردن مطالب این فصل به‌کار برند. یک بستهٔ نرم‌افزاری عالی به نام LINPACK ساخته شده است، که در دونگارا[1] و همکاران (۱۹۷۹)، توضیح داده شده است. این نرم‌افزار به‌طور گسترده در دسترس است و ما باز هم در این فصل به آن مراجعه خواهیم کرد.

۱.۸ حذف گاوسی

حذف گاوسی نام رسمی روش حل دستگاههای معادلات خطی با حذف تدریجی مجهولات و تبدیل آنها به دستگاههای مرتبهٔ پایینتر است، این روشی است که بیشتر در دبیرستان یا در درس جبر خطی دورهٔ کارشناسی (که اغلب با تولید شکل سطری ـ پلکانی ماتریس بستگی دارد) تدریس می‌شود. در اینجا یک تعریف دقیق حذف گاوسی داده شده است، که برای کاربرد در رایانه و تحلیل خطای گردکردن که در محاسبهٔ آن پیش می‌آید، لازم است.

1. Dongarra

برای حل $Ax = b$، آن را به یک دستگاه هم‌ارز $Ux = g$ که U یک ماتریس بالامثلثی است تبدیل می‌کنیم. این دستگاه با فرایند **جایگذاری پسرو** به‌سادگی حل می‌شود. دستگاه خطّی اصلی را با $A^{(۱)}x = b^{(۱)}$ نشان می‌دهیم

$$A^{(۱)} = [a_{ij}^{(۱)}] \qquad b^{(۱)} = [b_۱^{(۱)}, \dots, b_n^{(۱)}]^T \qquad ۱ \le i, j \le n$$

که در آن n مرتبهٔ دستگاه است. این دستگاه را به شکل یک دستگاه مثلثی $Ux = g$ بدل می‌کنیم بدین نحو که با افزودن مضاربی از یک معادله به معادلهٔ دیگر مجهولی را از معادلهٔ دوّم حذف می‌نماییم. از عملیات سطری دیگر در تبدیلهایی که در بخشهای آینده آمده‌اند، استفاده شده است. برای آنکه ارائهٔ مطلب ساده باشد، بعضی فرضهای فنی را در تعریف الگوریتم به‌کار می‌بریم؛ این فرضها در بخش بعد حذف می‌شوند.

الگوریتم حذف گاوسی

مرحلهٔ ۱ فرض می‌کنیم $a_{۱۱}^{(۱)} \neq ۰$. ضریبهای سطری را چنین تعریف می‌کنیم

$$m_{i۱} = \frac{a_{i۱}^{(۱)}}{a_{۱۱}^{(۱)}} \qquad i = ۲, ۳, \dots, n$$

این ضریبها برای حذف جملهٔ $x_۱$ از معادلات ۲ تا n به‌کار می‌روند. تعریف می‌کنیم

$$a_{ij}^{(۲)} = a_{ij}^{(۱)} - m_{i۱}a_{۱j}^{(۱)} \qquad i, j = ۲, \dots, n$$
$$b_i^{(۲)} = b_i^{(۱)} - m_{i۱}b_۱^{(۱)} \qquad i = ۲, \dots, n$$

همچنین سطرهای اول A و b دست‌نخورده باقی می‌مانند و در ستون اول $A^{(۱)}$، همهٔ عناصر زیرقطر، صفر گذاشته می‌شود.
دستگاه $A^{(۲)}x = b^{(۲)}$ بدین شکل است

$$\begin{bmatrix} a_{۱۱}^{(۱)} & a_{۱۲}^{(۱)} & \dots & a_{۱n}^{(۱)} \\ ۰ & a_{۲۲}^{(۲)} & \dots & a_{۲n}^{(۲)} \\ \vdots & \vdots & & \\ ۰ & a_{n۲}^{(۲)} & \dots & a_{nn}^{(۲)} \end{bmatrix} \begin{bmatrix} x_۱ \\ x_۲ \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_۱^{(۱)} \\ b_۲^{(۲)} \\ \vdots \\ b_n^{(۲)} \end{bmatrix}$$

حذف مجهولات را با رفتن به ستونهای ۲ و ۳ و غیره ادامه می‌دهیم. و این وضع به‌طورکلّی در زیر بیان شده است.

مرحلهٔ k: گیریم $1 \le k \le n-1$. فرض می‌کنیم $A^{(k)}x = b^{(k)}$ ساخته شده و x_1، x_2، ...، x_{k-1} در مراحل پی‌درپی حذف شده‌اند و $A^{(k)}$ دارای شکل زیر است

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & & \cdots & & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & & & & a_{2n}^{(2)} \\ \vdots & & \ddots & & & \vdots \\ 0 & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \vdots & & & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{bmatrix}$$

فرض می‌کنیم $a_{kk}^{(k)} \neq 0$. ضرایبها را چنین تعریف می‌کنیم

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \qquad i = k+1, \ldots, n \tag{۱.۱.۸}$$

از این ضرایبها برای حذف مجهول x_k از معادلات $k+1$ تا n استفاده می‌کنیم. تعریف می‌کنیم

$$\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)} \\ b_i^{(k+1)} &= b_i^{(k)} - m_{ik}b_k^{(k)} \qquad i,j = k+1, \ldots, n \end{aligned} \tag{۲.۱.۸}$$

سطرهای قبلی ۱ تا k دست‌نخورده می‌مانند و در ستون kام زیر عنصر قطر، همه‌جا صفر گذارده می‌شود. با ادامهٔ این شیوه، پس از $n-1$ مرحله، $A^{(n)}x = b^{(n)}$ را به‌دست می‌آوریم:

$$\begin{bmatrix} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & & \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{nn}^{(n)} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ \vdots \\ \\ b_n^{(n)} \end{bmatrix}$$

برای راحتی در نمادگذاری، می‌گیریم $U = A^{(n)}$ و $g = b^{(n)}$. دستگاه $Ux = g$ یک دستگاه بالا مثلثی و به‌سادگی قابل حل است. ابتدا داریم

$$x_n = \frac{g_n}{u_{nn}}$$

و سپس

$$x_k = \frac{۱}{u_{kk}} \left[g_k - \sum_{j=k+۱}^{n} u_{kj} x_j \right] \qquad k = n-۱, n-۲, \ldots, ۱ \quad (۳.۱.۸)$$

بدین صورت الگوریتم حذف گاوسی کامل می‌شود.

مثال دستگاه خطی زیر را حل کنید

$$\begin{aligned} x_۱ + ۲x_۲ + x_۳ &= ۰ \\ ۲x_۱ + ۲x_۲ + ۳x_۳ &= ۳ \\ -x_۱ - ۳x_۲ &= ۲ \end{aligned} \qquad (۴.۱.۸)$$

برای ساده کردن نمادگذاری، توجه داریم که مجهولات $x_۱$، $x_۲$ و $x_۳$ تا آخرین مرحله، در الگوریتم وارد نمی‌شوند. بنابراین دستگاه خطی بالا را با **ماتریس افزودهٔ** زیر نشان می‌دهیم:

$$[A|b] = \left[\begin{array}{ccc|c} ۱ & ۲ & ۱ & ۰ \\ ۲ & ۲ & ۳ & ۳ \\ -۱ & -۳ & ۰ & ۲ \end{array}\right]$$

عملیات سطری در این ماتریس افزوده اجرا و مجهولات در آخرین مرحله داده شده‌اند. در نمودار زیر، ضرایب درکنار پیکان، متناظر با تغییراتی که ایجاد می‌نمایند، داده شده‌اند

$$\left[\begin{array}{ccc|c} ۱ & ۲ & ۱ & ۰ \\ ۲ & ۲ & ۳ & ۳ \\ -۱ & -۳ & ۰ & ۲ \end{array}\right] \xrightarrow[m_{۳۱}=-۱]{m_{۲۱}=۲} \left[\begin{array}{ccc|c} ۱ & ۲ & ۱ & ۰ \\ ۰ & -۲ & ۱ & ۳ \\ ۰ & -۱ & ۱ & ۲ \end{array}\right]$$

$$\downarrow m_{۳۲}=\frac{۱}{۲}$$

$$[U|g] \equiv \left[\begin{array}{ccc|c} ۱ & ۲ & ۱ & ۰ \\ ۰ & -۲ & ۱ & ۳ \\ ۰ & ۰ & \frac{۱}{۲} & \frac{۱}{۲} \end{array}\right]$$

از حل $Ux = g$ داریم

$$x_۳ = ۱ \qquad x_۲ = -۱ \qquad x_۱ = ۱$$

تجزیهٔ مثلثی ماتریس خیلی بجاست که ضرایب m_{ij} را حفظ کنیم، زیرا اغلب $Ax = b$ را با یک A ولی با بردارهای مختلف b حل می‌کنیم. در رایانه، عناصر $a_{ij}^{(k+1)}$، $j \geq i$، همیشه در حافظهٔ $a_{ij}^{(k)}$ ذخیره می‌شوند. عناصر زیر قطر صفر شده‌اند، و این یک حافظهٔ مناسب برای عناصر m_{ij} در اختیار می‌گذارد. m_{ij} را در محلی که ابتدا برای ذخیرهٔ a_{ij}، $i > j$، به‌کار برده شده بود ذخیره می‌نماییم.

باز یک دلیل دیگر وجود دارد که به ضرایب m_{ij} به‌عنوان عناصر یک ماتریس نگاه می‌کنیم. ابتدا ماتریس پایین مثلثی زیر را معرفی می‌کنیم

$$L = \begin{bmatrix} ۱ & ۰ & ۰ & \dots & ۰ \\ m_{۲۱} & ۱ & ۰ & \dots & ۰ \\ \vdots & & \ddots & & \vdots \\ m_{n۱} & m_{n۲} & \dots & ۰ & ۱ \end{bmatrix}$$

قضیهٔ ۱.۸ اگر L و U ماتریسهای پایین مثلثی و بالا مثلثی باشند که قبلاً در استفاده از حذف گاوسی تعریف شده‌اند، آنگاه

$$A = LU \qquad (۵.۱.۸)$$

برهان این اثبات اساساً یک محاسبهٔ جبری است که با استفاده از تعاریف (۱.۱.۸) و (۲.۱.۸) صورت می‌گیرد. برای تجسّم درایه‌های $(LU)_{ij}$، از فرمول برداری زیر استفاده می‌نماییم

$$(LU)_{ij} = [m_{i۱}, \dots, m_{i,i-۱}, ۱, ۰, \dots, ۰] \begin{bmatrix} u_{۱j} \\ \vdots \\ u_{jj} \\ ۰ \\ \vdots \\ ۰ \end{bmatrix}$$

برای $i \leq j$،

$$\begin{aligned}(LU)_{ij} &= m_{i1}u_{1j} + m_{i2}u_{2j} + \ldots + m_{i,i-1}u_{i-1,j} + u_{ij} \\ &= \sum_{k=1}^{i-1} m_{ik}a_{kj}^{(k)} + a_{ij}^{(i)} \\ &= \sum_{k=1}^{i-1}[a_{ij}^{(k)} - a_{ij}^{(k+1)}] + a_{ij}^{(i)} \\ &= a_{ij}^{(1)} = a_{ij}\end{aligned}$$

برای $i > j$،

$$\begin{aligned}(LU)_{ij} &= m_{i1}u_{1j} + \ldots + m_{ij}u_{jj} \\ &= \sum_{k=1}^{j-1} m_{ik}a_{kj}^{(k)} + m_{ij}a_{jj}^{(j)} \\ &= \sum_{k=1}^{j-1}[a_{ij}^{(k)} - a_{ij}^{(k+1)}] + a_{ij}^{(j)} \\ &= a_{ij}^{(1)} = a_{ij}\end{aligned}$$

و اثبات کامل می‌شود. ■

تجزیهٔ (۸.۱.۵) نتیجهٔ مهمی است که از آن در بسط انواع روشهای حذف گاوسی برای رده‌های خاصّی از ماتریسها، استفادهٔ وسیعی می‌شود. ولی در حال حاضر ما فقط فرع زیر را مورد نظر قرار می‌دهیم.

فرع با همان ماتریسهای A، L و U در قضیهٔ ۱.۸،

$$\begin{aligned}\det(A) &= u_{11}u_{22}\ldots u_{nn} \\ &= a_{11}^{(1)}a_{22}^{(2)}\ldots a_{nn}^{(n)}\end{aligned}$$

■

برهان طبق قاعدهٔ حاصلضرب دترمینانها

$$\det(A) = \det(L)\det(U)$$

چون L و U مثلثی هستند، دترمینان آنها برابر حاصلضرب عناصر قطری آنهاست. و نتیجهٔ مطلوب به‌سادگی حاصل می‌شود، زیرا $\det(L) = ۱$ ■

مثال برای دستگاه (۴.۱.۸) از مثال قبلی،

$$L = \begin{bmatrix} ۱ & ۰ & ۰ \\ ۲ & ۱ & ۰ \\ -۱ & \frac{۱}{۲} & ۱ \end{bmatrix} \qquad U = \begin{bmatrix} ۱ & ۲ & ۱ \\ ۰ & -۲ & ۱ \\ ۰ & ۰ & \frac{۱}{۲} \end{bmatrix}$$

به‌سادگی تحقیق می‌شود که $A = LU$. همچنین $\det(A) = \det(U) = -۱$.

شمارش عملیات. برای تحلیل تعداد عملیات لازم در حل $Ax = b$ با استفاده از حذف گاوسی، پیدایش L و U را از A و تبدیل b به g و بالاخره پیداکردن جواب x را به‌طور جداگانه در نظر می‌گیریم.
۱. محاسبهٔ L و U. در مرحلهٔ ۱، $(n-۱)$ تقسیم برای محاسبهٔ مضارب $m_{i۱}$، $۲ \le i \le n$، به‌کار برده شده بود. سپس $(n-۱)^۲$ ضرب و $(n-۱)^۲$ جمع به‌کار رفته بود تا عناصر جدید $a_{ij}^{(۲)}$ به‌دست آیند. شمارش را می‌توانیم به همین شیوه برای سایر مراحل ادامه دهیم.
نتایج در جدول ۱.۸ خلاصه شده است. مقدار حاصل جمع هر ستون با استفاده از اتحادهای زیر به‌دست آمده است

$$\sum_{j=۱}^{p} j = \frac{p(p+۱)}{۲} \qquad \sum_{j=۱}^{p} j^۲ = \frac{p(p+۱)(۲p+۱)}{۶} \qquad p \ge ۱$$

معمولاً، تعداد ضرب و تقسیمهاست که روی‌هم‌رفته به‌عنوان تعداد عملیات در روش حذف گاوسی به‌کار می‌رود. در رایانه‌های قدیمی، جمع بسیار سریعتر از ضرب و تقسیم انجام می‌شد و

جدول ۱.۸ شمارش عملیات برای تجزیهٔ یک ماتریس به LU

مرحلهٔ k	جمع	ضرب	تقسیم
۱	$(n-۱)^۲$	$(n-۱)^۲$	$n-۱$
۲	$(n-۲)^۲$	$(n-۲)^۲$	$n-۲$
⋮	⋮	⋮	⋮
$n-۱$	۱	۱	۱
حاصل جمع	$\frac{n(n-۱)(۲n-۱)}{۶}$	$\frac{n(n-۱)(۲n-۱)}{۶}$	$\frac{n(n-۱)}{۲}$

بنابراین در محاسبهٔ هزینهٔ بسیاری از الگوریتمها جمعها نادیده گرفته می‌شدند. ولی، در رایانه‌های جدید، زمانهای انجام جمع، ضرب و تقسیم از لحاظ اندازه خیلی به‌هم نزدیک‌اند. برای سهولت در نمادگذاری، فرض می‌کنیم $MD(\cdot)$ و $AS(\cdot)$ به ترتیب معرّف تعداد ضرب و تقسیمها، و تعداد جمع و تفریقها برای محاسبهٔ کمیت داخل پرانتزها باشند.

برای تجزیهٔ A به LU، داریم

$$\begin{aligned} MD(LU) &= \frac{n(n^2-1)}{3} \doteq \frac{n^3}{3} \\ AS(LU) &= \frac{n(n-1)(2n-1)}{6} \doteq \frac{n^3}{3} \end{aligned} \tag{۶.۱.۸}$$

براوردهای نهایی برای مقادیر بزرگتر n معتبرند.

۲. تبدیل b به $g = b^{(n)}$:

$$\begin{aligned} MD(g) &= (n-1) + (n-2) + \ldots + 1 = \frac{n(n-1)}{2} \\ AS(g) &= \frac{n(n-1)}{2} \end{aligned} \tag{۷.۱.۸}$$

۳. حل $Ux = g$

$$MD(x) = \frac{n(n+1)}{2} \qquad AS(x) = \frac{n(n-1)}{2} \tag{۸.۱.۸}$$

۴. حل $Ax = b$. مراحل ۱ تا ۳ را ترکیب می‌کنیم تا به‌دست آوریم

$$\begin{aligned} MD(LU, x) &= \frac{n^3}{3} + n^2 - \frac{n}{3} \doteq \frac{1}{3}n^3 \\ AS(LU, x) &= \frac{n(n-1)(2n+5)}{6} \doteq \frac{1}{3}n^3 \end{aligned} \tag{۹.۱.۸}$$

تعداد جمعها همیشه در همان حدود تعداد ضرب و تقسیمهاست، و لذا از این پس ما فقط آخری را در نظر می‌گیریم. به اولین چیزی که باید توجه کنیم این است که حل $Ax = b$، در مقایسه با عمل ظاهراً ساده‌ای مثل ضرب دو ماتریس $n \times n$، بسیار ناچیز است. ضرب دو ماتریس نیاز به n^3 عمل دارد و حل $Ax = b$ فقط در حدود $\frac{1}{3}n^3$ عمل در برخواهد داشت.

دوّم اینکه، هزینهٔ اصلی حل $Ax = b$ در تجزیهٔ $A = LU$ است. وقتی این را پیدا کردیم، فقط n^2 عمل دیگر برای حل $Ax = b$ لازم است. وقتی یک بار $Ax = b$ حل شده باشد حل دستگاههای اضافی دیگر با همان ماتریس ضرایب نسبتاً ناچیز خواهد بود، به شرطی که تجزیهٔ LU ذخیره شده باشد.

بالاخره، روش حذف گاوسی بسیار ناچیزتر از **قاعدهٔ کرامر** است، که از دترمینان استفاده می‌کند و اغلب در درس جبر خطی تدریس می‌شود [برای مثال، بخش ۴.۲ آنتون (۱۹۸۴) را ببینید]. اگر دترمینان‌ها در قاعدهٔ کرامر از راه **بسط به دترمینان‌های جزئی** محاسبه شوند، تعداد عملیات $(n+1)!$ خواهد شد. برای $n = ۱۰$، حذف‌گاوسی ۴۳۰ عمل و قاعدهٔ کرامر ۳۹۹۱۶۸۰۰ عمل خواهد داشت. باید تأکید کنیم که قاعدهٔ کرامر یک وسیلهٔ محاسبهٔ عملی نیست و فقط به‌عنوان یک وسیلهٔ ریاضیات نظری باید به آن بنگریم.

۵. وارون A. وارون A، معمولاً لازم نیست ولی می‌توان با روش حذف گاوسی آن را پیدا کرد. پیداکردن A^{-1} هم‌ارز حل معادلهٔ $AX = I$ است که X یک ماتریس $n \times n$ مجهول است. اگر X و I را برحسب ستونهای آنها بنویسیم.

$$X = [x^{(1)}, \ldots, x^{(n)}] \qquad I = [e^{(1)}, \ldots, e^{(n)}]$$

آنگاه حل $AX = I$ هم‌ارز حل n دستگاه

$$Ax^{(1)} = e^{(1)}, \ldots, \; Ax^{(n)} = e^{(n)} \qquad (۱۰.۱.۸)$$

است که همگی دارای یک ماتریس ضرایب A هستند. با استفاده از (۱) ـ (۳)

$$MD(A^{-1}) = \frac{۳}{۴}n^{۳} - \frac{n}{۳} \doteq \frac{۴}{۳}n^{۳}$$

محاسبهٔ A^{-1} چهار برابر حلّ $Ax = b$ برای یک بردار تنهای b، هزینه خواهد داشت، نه n بار که ممکن است در ابتدا به‌نظر آید. با توجه دقیق به جزئیات فرایند وارون‌سازی، و استفاده از شکل خاص بردارهای سمت راست $e^{(1)}$، ...، $e^{(n)}$، ممکن است هزینه دقیقاً به

$$MD(A^{-1}) = n^{۳} \qquad (۱۱.۱.۸)$$

عمل کاهش یابد.

ولی، باز هم در بسیاری از موارد، اتلاف وقت است که در حل $Ax = b$ وارون A، یعنی A^{-1}، را به‌دست آوریم. و هیچ مزیّتی در ذخیرهٔ A^{-1} به جای تجزیهٔ LU برای حل دستگاههای دیگر $Ax = b$، وجود ندارد. در هر دو حالت تعداد ضرب و تقسیمهای لازم برای حل $Ax = b$ دقیقاً $n^{۲}$ است.

۲.۸ محورگیری و مقیاس‌دهی در حذف گاوسی

در هر مرحله از فرایند حذف در بخش اخیر، عنصر محور مربوط $a_{kk}^{(k)}$ را ناصفر فرض کردیم. برای حذف این فرض، هر مرحله از فرایند حذف را، با تعویض سطرها با هم و قراردادن یک عنصر ناصفر در جای محور، آغاز می‌کنیم. اگر چنین عنصری موجود نباشد، این ماتریس باید تکین باشد که با فرض تناقض دارد.

ولی، این کافی نیست که فقط بخواهیم عنصر محور ناصفر باشد. اغلب یک عنصر صفر به علت خطاهای گردکردن در محاسبهٔ آن ناصفر می‌شود. استفاده از چنین عنصری به‌عنوان عنصر محور به خطاهای بزرگی در محاسبات بعدی در ماتریس منجر خواهد شد. برای اجتناب از این امر و به دلایل دیگری مربوط به انتشار خطاهای گردکردن، ما **محورگیری جزئی و محورگیری کلی** را معرفی می‌نماییم.

تعریف ۱. محورگیری جزئی. برای $1 \le k \le n-1$، در فرایند حذف گاوسی در مرحلهٔ kام، می‌گیریم

$$c_k = \underset{k \le i \le n}{\text{Max}} \mid a_{ik}^{(k)} \mid \qquad (۱.۲.۸)$$

گیریم i کوچکترین اندیس سطری باشد، $i \ge k$، که برای آن ماکسیمم c_k به‌دست آمده است. اگر $i > k$، آنگاه سطرهای k و i در A و b را با هم تعویض کرده و به مرحلهٔ kام فرایند حذف می‌پردازیم. همهٔ ضریبها اکنون در

$$\mid m_{ik} \mid \le 1 \qquad i = k+1, \ldots, n \qquad (۲.۲.۸)$$

صدق می‌کنند و این امر به جلوگیری از بسیار زیادشدن تفاوت اندازه‌های عناصر $A^{(k)}$ کمک خواهد کرد، و بنابراین امکان خطاهای کاهش بزرگ در ارقام بامعنی را کمتر خواهد نمود.

۲. محورگیری کلی. تعریف می‌کنیم

$$c_k = \underset{k \le i,j \le n}{\text{Max}} \mid a_{ij}^{(k)} \mid$$

سطرهای A و b و ستونهای A را تعویض می‌کنیم تا عنصری که ماکسیمم c_k را دارد در جای محور قرار گیرد. باید توجه کرد که با تعویض یک ستون، ترتیب مجهولات عوض می‌شود. پس از اتمام حذف و فرایند جایگذاری پسرو، این تعویض باید عکس شود.

ثابت شده است که محورگیری کلی موجب می‌شود که خطای گردکردن در حذف گاوسی، در مقایسه با موقعی که از هیچ محورگیری استفاده نمی‌شود، با سرعت کم و قابل‌قبولی انتشار یابد. نتایج

نظری در استفاده از محورگیری جزئی کاملاً به آن خوبی نیست، ولی تقریباً در تمام مسائل عملی، رفتار خطا اساساً مانند رفتار خطا در محورگیری کلی است. با مقایسهٔ زمانهای عملیات، محورگیری کلی یک شیوه گرانتری است و بنابراین، در بیشتر الگوریتم‌های عملی، محورگیری جزئی به‌کار می‌رود. از این پس وقتی کلمهٔ **محورگیری** را به‌کار می‌بریم منظورمان **محورگیری جزئی** است. کلّ مسألهٔ انتشار خطای گردکردن در روش حذف گاوسی توسط ج. ه. ویلکینسن عمیقاً تحلیل شده است [مثلاً، (صفحات ۲۰۹ـ۲۲۰) ویلکینسن (۱۹۵۶) را ببینید]، و بعضی از نتیجه‌گیریهای او در بخش ۴.۸ ارائه شده است.

مثال تأثیر به‌کار بردن محورگیری را با حل دستگاه زیر نشان می‌دهیم:

$$\begin{aligned} ۰٫۷۲۹x + ۰٫۸۱y + ۰٫۹z &= ۰٫۶۸۶۷ \\ x + y + z &= ۰٫۸۳۳۸ \\ ۱٫۳۳۱x + ۱٫۲۱y + ۱٫۱z &= ۱٫۰۰۰ \end{aligned} \qquad (۳.۲.۸)$$

جواب دقیق که تا چهار رقم معنی‌دار گرد شده چنین است:

$$x = ۰٫۲۲۴۵ \quad y = ۰٫۲۸۱۴ \quad z = ۰٫۳۲۷۹ \qquad (۴.۲.۸)$$

از حساب اعشاری با ممیز شناور، با چهار رقم اعشاری، برای حل دستگاه خطی استفاده خواهد شد. دلیل استفاده از این حساب نشان‌دادن اثر کار با یک تعداد متناهی رقم است، در حالی که بتوان اندازهٔ ارقام را معقول نگهداشت. نماد ماتریس افزوده برای نمایش دستگاه (۳.۲.۸) به‌کار برده می‌شود، درست مانند آنچه در مثال قبل (۴.۱.۸) عمل شد.

۱. حل بدون محورگیری

$$\left[\begin{array}{ccc|c} ۰٫۷۲۹۰ & ۰٫۸۱۰۰ & ۰٫۹۰۰۰ & ۰٫۶۸۶۷ \\ ۱٫۰۰۰ & ۱٫۰۰۰ & ۱٫۰۰۰ & ۰٫۸۳۳۸ \\ ۱٫۳۳۱ & ۱٫۲۱۰ & ۱٫۱۰۰ & ۱٫۰۰۰ \end{array}\right]$$

$$\downarrow \begin{array}{l} m_{۲۱}=۱٫۳۷۲ \\ m_{۳۱}=۱٫۸۲۶ \end{array}$$

$$\left[\begin{array}{ccc|c} ۰٫۷۲۹۰ & ۰٫۸۱۰۰ & ۰٫۹۰۰۰ & ۰٫۶۸۶۷ \\ ۰٫۰ & -۰٫۱۱۱۰ & -۰٫۲۳۵۰ & -۰٫۱۰۸۴ \\ ۰٫۰ & -۰٫۲۶۹۰ & -۰٫۵۴۳۰ & -۰٫۲۵۴۰ \end{array}\right]$$

$$\downarrow m_{۳۲}=۲٫۴۲۳$$

$$\left[\begin{array}{cccc|c} ۰٫۷۲۹۰ & ۰٫۸۱۰۰ & ۰٫۹۰۰۰ & & ۰٫۶۸۶۷ \\ ۰٫۰ & -۰٫۱۱۱۰ & -۰٫۲۳۵۰ & & -۰٫۱۰۸۴ \\ ۰٫۰ & ۰٫۰ & ۰٫۲۶۴۰ & & ۰٫۰۰۸۷۰۰ \end{array}\right]$$

جواب چنین است

$$x = ۰٫۲۲۵۱ \quad y = ۰٫۲۷۹۰ \quad z = ۰٫۳۲۹۵ \qquad (۵.۲.۸)$$

۲. حل با محورگیری. برای نشان‌دادن تعویض سطرهای i و j، نماد $r_i \longleftrightarrow r_j$ را به‌کار خواهیم برد.

$$\left[\begin{array}{ccc|c} ۰٫۷۲۹۰ & ۰٫۸۱۰۰ & ۰٫۹۰۰۰ & ۰٫۶۸۶۷ \\ ۱٫۰۰۰۰ & ۱٫۰۰۰۰ & ۱٫۰۰۰۰ & ۰٫۸۳۳۸ \\ ۱٫۳۳۱ & ۱٫۲۱۰ & ۱٫۱۰۰ & ۱٫۰۰۰ \end{array}\right]$$

$$r_۱ \leftrightarrow r_۲ \downarrow \begin{array}{l} m_{۲۱}=۰٫۷۵۱۳ \\ m_{۳۱}=۰٫۵۴۷۷ \end{array}$$

$$\left[\begin{array}{ccc|c} ۱٫۳۳۱ & ۱٫۲۱۰ & ۱٫۱۰۰ & ۱٫۰۰۰ \\ ۰٫۰ & ۰٫۰۹۰۹۰ & ۰٫۱۷۳۶ & ۰٫۰۸۲۵۰ \\ ۰٫۰ & ۰٫۱۴۷۳ & ۰٫۲۹۷۵ & ۰٫۱۳۹۰ \end{array}\right]$$

$$r_۲ \leftrightarrow r_۳ \downarrow m_{۳۲}=۰٫۶۱۷۱$$

$$\left[\begin{array}{ccc|c} ۱٫۳۳۱ & ۱٫۲۱۰ & ۱٫۱۰۰ & ۱٫۰۰۰ \\ ۰٫۰ & ۰٫۱۴۷۳ & ۰٫۲۹۷۵ & ۰٫۱۳۹۰ \\ ۰٫۰ & ۰٫۰ & -۰٫۰۱۰۰۰ & -۰٫۰۰۳۲۸۰ \end{array}\right]$$

جواب چنین است:

$$x = ۰٫۲۲۴۶ \quad y = ۰٫۲۸۱۲ \quad z = ۰٫۳۲۸۰ \qquad (۶.۲.۸)$$

خطا در (۵.۲.۸) برحسب اینکه کدام مؤلفهٔ جواب در نظرگرفته شود، از هفت تا شانزده برابر بزرگتر از خطا در (۶.۲.۸) است. نتایج (۶.۲.۸) یک رقم معنی‌دار بیش از نتایج (۵.۲.۸) دارند. این نکته اثر مثبت استفاده از محورگیری را در رفتار خطا، در روش حذف گاوسی، نشان می‌دهد.

محورگیری نتیجهٔ تجزیه به عوامل (۵.۱.۸) راکه در قضیهٔ ۱.۸ داده شده عوض می‌کند. ولی

در یک شکل اصلاح‌شده، نتیجه باز هم درست است. اگر تعویض سطرها که بر اثر محورگیری القا شده است، قبل از عمل حذف در A انجام شده باشد، دیگر محورگیری لزومی نخواهد داشت. تعویض سطرها در A را می‌توان با ضرب قدامی (premultiplication) A در **ماتریس جایگشت** مناسب P، به‌صورت PA به‌دست آورد. در این صورت حذف گاوسی روی PA به

$$LU = PA \qquad (۷.۲.۸)$$

می‌انجامد که U ماتریس بالا مثلثی حاصل از فرایند حذف با محورگیری است. ماتریس پایین مثلثی L را می‌توان با استفاده از ضرایب در حذف گاوسی با محورگیری، ساخت. ما از ذکر جزئیات صرف‌نظر می‌کنیم، زیرا ساخت عملی مهم نیست.

مثال با توجّه به مثال قبل که با محورگیری حل شد ماتریسهای زیر را تشکیل می‌دهیم

$$L = \begin{bmatrix} ۱ر۰۰۰ & ۰ر۰ & ۰ر۰ \\ ۰ر۵۴۷۷ & ۱ر۰۰۰ & ۰ر۰ \\ ۰ر۷۵۱۳ & ۰ر۶۱۷۱ & ۱ر۰۰۰ \end{bmatrix} \qquad U = \begin{bmatrix} ۱ر۳۳۱ & ۱ر۲۱۰ & ۱ر۱۰۰ \\ ۰ر۰ & ۰ر۱۴۷۳ & ۰ر۲۹۷۵ \\ ۰ر۰ & ۰ر۰ & -۰ر۰۱۰۰۰ \end{bmatrix}$$

از ضرب این دو

$$LU = \begin{bmatrix} ۱ر۳۳۱ & ۱ر۲۱۰ & ۱ر۱۰۰ \\ ۰ر۷۲۸۹ & ۰ر۸۱۰۰ & ۰ر۹۰۰۰ \\ ۱ر۰۰۰ & ۱ر۰۰۰ & ۱ر۰۰۰ \end{bmatrix} = PA \qquad P = \begin{bmatrix} ۰ & ۰ & ۱ \\ ۱ & ۰ & ۰ \\ ۰ & ۱ & ۰ \end{bmatrix}$$

نتیجهٔ PA ماتریس A است که در آن ابتدا، سطرهای ۱ و ۳، و سپس سطرهای ۲ و ۳ تعویض شده‌اند. این مثالی از (۷.۲.۸) است.

مقیاس‌دهی به‌طور تجربی مشاهده شده است که اگر عناصر ماتریس ضرایب A از نظر اندازه زیاد متفاوت باشند، خطاهای کاهش در ارقام بامعنی بزرگی احتمالاً وارد محاسبه می‌شوند و انتشار خطاهای گردکردن بدتر می‌شود. برای اجتناب از این مسأله، معمولاً به ماتریس A **مقیاس می‌دهیم** تا عناصر کمتر متفاوت باشند. این کار معمولاً از ضرب سطرها و ستونها در ثابتهای مناسب انجام می‌گیرد. موضوع مقیاس‌دهی فعلاً به خوبی معلوم نیست، به‌ویژه اینکه چگونه می‌توان تضمین کرد که اثر خطاهای گردکردن در حذف گاوسی، با چنین مقیاس‌دهی، کوچک می‌شوند. از تجربهٔ محاسباتی به‌نظر می‌رسد که اغلب تمام سطرها باید طوری مقیاس‌دهی شوند که از نظر اندازه

تقریباً مساوی شوند. به‌علاوه، به تمام ستونها می‌توان مقیاس داد تا از نظر قدرمطلق تقریباً به یک اندازه درآیند. این عمل اخیر، هم‌ارز است با اینکه به مؤلفه‌های مجهول، یعنی x_i ها مقیاس دهیم؛ و آن را می‌توان اغلب این طور تعبیر نمود که بگوییم x_i ها باید با واحدهایی قابل مقایسه با هم، اندازه‌گیری شوند.

هیچ شیوهٔ از پیش تعیین شده‌ای برای انتخاب عاملهای مقیاس‌دهی در دست نیست که بتواند تنها بر پایهٔ اطلاع از A و b، اثر انتشار خطای گردکردن را همیشه کاهش دهد. آنچه در استوارت (۱۹۷۷) در استفادهٔ کلی از مقیاس‌دهی آمده است، آن‌گونه که در پاراگراف قبلی توضیح داده شد، تا اندازه‌ای قابل انتقاد است. او پیشنهاد می‌نماید که فاکتورهای مقیاس‌دهی طوری انتخاب شوند که در ماتریس با مقیاسهای جدید، خطاهای درایه‌ها تا حدی برابر باشند. وقتی گردکردن تنها منبع خطا باشد، این عمل به یک شیوهٔ مقیاس‌دهی می‌انجامد که تمام عناصر را تقریباً به یک اندازه می‌کند. در برنامه‌های LIN-PACK، مقیاس‌دهی دخالت داده نمی‌شود، ولی یک شیوه به همان طریق استوارت توصیه می‌شود. [برای بحث مفصلتر دونگارا و همکاران (۱۹۷۹، صص ۱۷ـ۱۱۲) را ببینید. برای بحثهای دیگر مقیاس‌دهی فورسایت و مولر (۱۹۶۷، فصل ۱۱) وگلوب وون‌لون (۱۹۸۳، صص ۷۲ـ۷۴) را ببینید].

اگر B معرّف نتیجهٔ مقیاس‌دهی سطری و ستونی A باشد، آنگاه

$$B = D_1 A D_2$$

که D_1 و D_2 ماتریسهای قطری هستند با درایه‌های متشکل از ثابتهای مقیاس‌دهی. برای حل $Ax = b$ مشاهده می‌کنیم که

$$D_1 A D_2 (D_2^{-1} x) = D_1 b$$

بنابراین جواب x را از حل

$$Bz = D_1 b \qquad x = D_2 z \qquad (۸.۲.۸)$$

به‌دست می‌آوریم. بقیّهٔ بحث، به مقیاس‌دهی سطری محدود شده است، زیرا نوعی از آن تاحدّی مورد موافقت همگان قرار گرفته است.

معمولاً سعی می‌کنیم ضرایب را طوری انتخاب نماییم که داشته باشیم

$$\max_{1 \le j \le n} | b_{ij} | \doteq 1 \qquad i = 1, \ldots, n \qquad (۹.۲.۸)$$

که $B = [b_{ij}]$ نتیجهٔ مقیاس‌دهی A است. ساده‌ترین راه این است که چنین تعریف کنیم

$$s_i = \max_{1 \le j \le n} | a_{ij} | \qquad i = 1, \ldots, n$$

$$b_{ij} = \frac{a_{ij}}{s_i}, \qquad j = ۱, \ldots, n \qquad (۱۰.۲.۸)$$

ولی چون این روش خطای گردکردن اضافه‌تری برای هر عنصر ماتریس ضرایب ایجاد می‌نماید، از دو روش دیگر بیشتر استفاده می‌شوند.

۱. تبدیل مقیاس با استفاده از پایهٔ عددی رایانه. گیریم β معرّف پایه‌ای باشد که در حساب رایانه‌یی به‌کار برده می‌شود، برای مثال، $\beta = ۲$ در ماشینهای دودویی. گیریم r_i کوچکترین عدد صحیحی باشد که برای آن $\beta^{r_i} \geq s_i$. ماتریس مقیاس‌دهی‌شدهٔ B را به شکل زیر تعریف می‌کنیم

$$b_{ij} = \frac{a_{ij}}{\beta^{r_i}} \qquad i, j = ۱, \ldots, n \qquad (۱۱.۲.۸)$$

هیچ گردکردنی در تعریف b_{ij} وجود ندارد، فقط تغییر نما در شکل با ممیز شناور a_{ij} رخ می‌دهد. مقادیر B در رابطهٔ زیر صدق می‌کنند

$$\beta^{-۱} < \underset{۱ \leq j \leq n}{\text{Max}} \mid b_{ij} \mid \leq ۱$$

و بنابراین (۹.۲.۸) به‌خوبی برقرار می‌گردد.

۲. مقیاس‌دهی ضمنی. چنانچه محورگیری در حذف گاوسی به‌کار رفته باشد، استفاده از مقیاس‌دهی معمولاً انتخاب عناصر محورگیری را عوض می‌کند. و فقط به دلیل همین تغییر عناصر محورگیری است که نتایج در حذف گاوسی عوض می‌شوند. این توجه را مدیون ف. باوئر هستیم [در صفحهٔ ۳۸ فورسایت و مولر (۱۹۶۷)] که می‌گوید اگر مقیاس‌دهی (۱۱.۲.۸) به‌کار برده شود، و اگر عناصر محورگیری همانهایی انتخاب شوند که در حل $Ax = b$ به‌کار رفته‌اند، آنگاه حل (۸.۲.۸) دقیقاً همان مقدار محاسبه‌شدهٔ x را خواهد داد. بنابراین تنها انتخاب عناصر محورگیری است که اهمیت دارد.

برای مقیاس‌دهی ضمنی، کار با ماتریس A را ادامه می‌دهیم، ولی برای تعیین عنصر محورگیری در مرحلهٔ kـام الگوریتم حذف گاوسی c_k را با

$$c_k = \underset{k \leq i \leq n}{\text{Max}} \frac{\mid a_{ik}^{(k)} \mid}{s_i^{(k)}} \qquad (۱۲.۲.۸)$$

تعریف می‌نماییم که به‌جای تعریف (۱.۲.۸) که در تعریف محورگیری جزئی استفاده شده، به‌کار می‌رود. کوچکترین اندیس $i \geq k$ را انتخاب می‌کنیم که c_k در (۱۲.۲.۸) را به‌دست دهد، و اگر $i \neq k$، سطرهای i و k را تعویض می‌کنیم. سپس الگوریتم حذف بخش ۱.۸ را مانند قبل ادامه می‌دهیم. به‌نظر می‌آید که این شکل مقیاس‌دهی در بین الگوریتمهای منتشر شده که در آنها از مقیاس‌دهی استفاده شده از همه متداولتر است.

الگوریتمی برای حذف گاوسی ابتدا الگوریتمی به‌نام *Factor* برای تجزیهٔ مثلثی ماتریس A خواهیم داد. در این الگوریتم، حذف گاوسی با محورگیری جزئی توأم با مقیاس‌دهی ضمنی (۱۰.۲.۸) و (۱۲.۲.۸)، به‌کار برده می‌شود. سپس یک الگوریتم به‌نام *Solve* برای استفاده از نتایج *Factor*، در حل دستگاه خطی $Ax = b$ خواهیم داد. دلیل تجزیهٔ فرایند حذف به این دو مرحله این است که اغلب می‌خواهیم چندین دستگاه $Ax = b$ را با یک ماتریس A، ولی با مقادیر مختلف b، حل کنیم.

الگوریتم *Factor* (A, n, Pivot, det, ier)

۱. ملاحظات: A یک ماتریس $n \times n$ است که با استفاده از تجزیهٔ LU به عوامل صورت می‌گیرد. حذف گاوسی همراه با محورگیری جزیی و مقیاس‌دهی ضمنی در سطرها به‌کار برده می‌شود. پس از تکمیل الگوریتم، ماتریس بالامثلثی U، در قسمت بالایی مثلث A ذخیره می‌شود؛ و ضریبهای (۱.۱.۸) که L را زیر قطرش تشکیل می‌دهد در جاهای متناظر در A ذخیره می‌شوند. بردار محور[1] شامل ثبت تمام تعویضهای سطری خواهد بود. اگر $\text{Pivot}(k) = k$، آنگاه هیچ تعویضی در مرحلهٔ k ام فرایند حذف، صورت نگرفته است. ولی اگر $\text{Pivot}(k) = i \neq k$، آنگاه در مرحلهٔ k ام فرایند حذف، سطرهای i و k باهم تعویض شده‌اند. متغیر det در خروج شامل مقدار دترمینان A خواهد بود.

متغیر ier یک شاخص خطاست. اگر $\text{ier} = ۰$، این زیربرنامه به نحو رضایتبخشی انجام شده است. ولی اگر $\text{ier} = ۱$، ماتریس A تکین است، بدین معنی که تمام عناصر محورگیری ممکن در مرحله‌ای از فرایند حذف، صفر بوده‌اند. در این حالت تمام محاسبات متوقف شده و از زیربرنامه خارج شده‌ایم. هیچ اقدامی برای کنترل دقّت تجزیهٔ محاسبه شدهٔ A به عمل نیامده است و A ممکن است بدون اینکه متوجه شده باشیم تقریباً تکین بوده باشد.

۲. $\det := ۱$

۳. $i = ۱, \ldots, n$ ،$s_i = \underset{۱ \le j \le n}{\text{Max}} \mid a_{i,j} \mid$

۴. برای $k = ۱, \ldots, n - ۱$ اعمال را تا مرحلهٔ ۱۶ انجام دهید.

۵. $c_k := \underset{k \le i \le n}{\text{Max}} \mid \frac{a_{ik}}{s_i} \mid$

۶. $i_۰$ را کوچکترین زیرنمایهٔ $i \ge k$ بگیرید که ماکسیمم آن در مرحلهٔ ۵ حاصل شده است.
$\text{Pivot}(k) := i_۰$.

۷. اگر $c_k = ۰$، آنگاه $\det := ۰$، $\text{ier} := ۱$ و از الگوریتم خارج شوید.

۸. اگر $k = i_۰$، به مرحلهٔ ۱۱ بروید.

۹. $\det := -\det$

1. vector Pivot

۱۰. a_{kj} را با $a_{i,j}$برای $j = k, \ldots, n$، عوض کنید و s_k را با s_i.

۱۱. برای $i = k+۱, \ldots, n$ تا مرحلهٔ ۱۴ کار را ادامه دهید.

۱۲. $a_{ik} := m_i := \dfrac{a_i k}{a_{kk}}$

۱۳. $a_{ij} := a_{ij} - m_i a_{kj}$، $j = k+۱, \ldots, n$

۱۴. حلقه را در i ختم کنید.

۱۵. $\det := a_{kk} \cdot \det$

۱۶. حلقه را در k ختم کنید.

۱۷. ier := ۰ و از الگوریتم خارج شوید.

الگوریتم $Solve\ (A,\ n\ ,b,\text{Pivot})$

۱. **ملاحظات:** با این الگوریتم دستگاه خطی $Ax = b$ حل می‌شود. فرض بر این است که ماتریس اصلی A با استفاده از الگوریتم $Factor$ تجزیه شده است که تعویض سطرها در Pivot ثبت شده است. هنگام خروج، جواب در b ذخیره می‌شود. ماتریس A و بردار محور دست‌نخورده می‌مانند.

۲. برای $k = ۱, ۲, \ldots, n-۱$ تا مرحلهٔ ۵ جلو بروید.

۳. اگر $i := \text{pivot}(k) \neq k$ آنگاه b_i و b_k را با هم عوض کنید.

۴. $b_i := b_i - a_{ik} b_k$، $i = k+۱, \ldots, n$

۵. حلقهٔ را در k قطع کنید.

۶. $b_n := b_n / a_{nn}$

۷. برای $i = n-۱, \ldots, ۱$ تا مرحلهٔ ۹ عمل کنید.

۸. $b_i := \dfrac{۱}{a_{ii}} \left\{ b_i - \sum_{j=i+۱}^{n} a_{ij} b_j \right\}$

۹. حلقهٔ را در i ختم کنید.

۱۰. از الگوریتم خارج شوید.

مثال قبلی (۳.۲.۸) را به‌عنوان یک توضیح به‌کار می‌بریم. استفاده از تبدیل مقیاس ضمنی در این حالت نیاز به تغییر انتخاب عناصر محور، در محورگیری جزئی، ندارد. الگوریتمهایی شبیه $Factor$ و $Solve$ در منابع چندی داده شده‌اند [فورسایت و مولر (۱۹۶۷) و فصل اول دونگارا و همکاران (۱۹۷۹) را برای نسخه‌های تکمیل‌شدهٔ این الگوریتمها ببینید]. در برنامه‌های LINPACK اطلاعاتی در مورد شرط یا پایداری مسألهٔ $Ax = b$ و دقت جواب محاسبه شده به‌دست می‌آید.

یک جنبهٔ مهم برنامه‌های LINPACK استفاده از زیرروالهای اساسی جبر خطی (BLAS)[1] است. این زیربرنامه‌ها عملیاتی ساده از قبیل حاصلضرب نقطه‌ای دو بردار، یا جمع مضربی از یک بردار با بردار دیگر را انجام می‌دهند. برنامه‌های LINPACK این BLASها را به جای خیلی از حلقه‌های داخلی یک روش به‌کار می‌برند. در صورت تمایل BLAS را می‌توان برای هر رایانه اپتیمم نمود؛ بنابراین، با حفظ استقلال برنامهٔ منبع از ماشین می‌توان اجرای برنامه‌های اصلی LINPACK را نیز بهبود بخشید. برای یک بحث کاملتر در مورد BLAS به لاوسن[2] و همکاران (۱۹۷۹) مراجعه کنید.

۳.۸ صورتهای دیگر حذف گاوسی

صورتهای متفاوت زیادی از حذف گاوسی وجود دارند. بعضی صورتها، اصلاحات یا ساده‌سازیهایی هستند که براساس ویژگیهای خاص رده‌هایی از ماتریسها صورت می‌گیرند، مثل، ماتریسهای متقارن، ماتریسهای معین‌مثبت. صورتهای دیگر، راههای بازنویسی حذف گاوسی، به شکل فشرده‌تر، گاهی برای استفاده از فن‌های بخصوصی برای کاهش خطا هستند. ما فقط بعضی از این صورتها را بررسی می‌کنیم و سپس برای بقیه، مراجعی ذکر خواهیم کرد.

روش گاوس ـ یوردان این روند عیناً از نظر حذف معمولی ازجمله امکان استفاده از محورگیری و مقیاس‌دهی است. تفاوت این روند در حذف مجهول در معادلات بالای قطر و معادلات پایین قطر است. در مرحلهٔ k ام الگوریتم حذف، عنصر محور را مانند قبل انتخاب می‌نماییم. سپس تعریف می‌کنیم

$$a_{kj}^{(k+1)} = \frac{a_{kj}^{(k)}}{a_{kk}^{(k)}} \qquad j = k, \ldots, n$$
$$b_k^{(k+1)} = \frac{b_k^{(k)}}{a_{kk}^{(k)}}$$

مجهول x_k را در معادلات بالا و پایین معادلهٔ k ام حذف می‌نماییم. تعریف می‌کنیم

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - a_{ik}^{(k)} a_{kj}^{(k+1)}$$
$$b_i^{(k+1)} = b_i^{(k)} - a_{ik}^{(k)} b_k^{(k+1)} \qquad (۱.۳.۸)$$

برای $j = k, \ldots, n$ و $i = 1, \ldots, n$ و $i \neq k$. روش گاوس ـ یوردان هم‌ارز با استفاده از

1. Basic Linear Algebra Subroutines 2. Lawson

صورت ساده شدهٔ پلکانی ـ سطری درکتابهای جبرخطی است. [برای مثال، صفحات ۸ـ۹، آنتون (۱۹۸۴) را ببینید].

در این روند، ماتریس افزودهٔ $[A \mid b]$ به $[I \mid b^{(n)}]$ بدل می‌شود، به‌طوری که در پایان روند حذف پیشین $x = b^{(n)}$. حل $Ax = b$ با این روش به تعداد

$$\frac{n(n-1)^2}{2} \doteq \frac{n^3}{2} \qquad (2.3.8)$$

ضرب و تقسیم نیاز دارد. این تعداد ۵۰ درصد بیشتر از روش حذف معمولی است؛ در نتیجه برای حل دستگاههای خطی معمولاً نباید از روش گاوس ـ یوردان استفاده کرد. ولی، این روش را می‌توان برای تولید یک برنامهٔ وارونیابی ماتریس که از مینیمم حافظه استفاده می‌کند، به‌کار برد. با بهره‌گیری از مزیّت ویژهٔ ساختار خاص طرف راست معادلهٔ $AX = I$، با روش گاوس ـ یوردان می‌توان جواب $X = A^{-1}$ را فقط با n محل حافظهٔ اضافی، به جای n^2 محل حافظهٔ اضافی معمولی، محاسبه کرد. محورگیری جزئی و مقیاس‌دهی ضمنی را کماکان می‌توان به‌کار برد.

روشهای فشرده ممکن است مستقیماً از ماتریس A به تجزیهٔ LU ی آن رسید و این کار را می‌توان با محورگیری جزئی و مقیاس‌دهی توأم کرد. اگر امکان محورگیری را موقتاً ندیده بگیریم، نتیجهٔ

$$A = LU \qquad (3.3.8)$$

ما را به مجموعه‌ای از فرمول‌های بازگشتی برای عناصر L و U می‌رساند.

چنانچه فقط بخواهیم که L و U به ترتیب پایین مثلثی و بالا مثلثی باشند، یک نایکتایی در انتخاب L و U وجود خواهد داشت. اگر A ناتکین باشد، و اگر دو تجزیه داشته باشیم

$$A = L_1 U_1 = L_2 U_2 \qquad (4.3.8)$$

آنگاه

$$L_2^{-1} L_1 = U_2 U_1^{-1} \qquad (5.3.8)$$

معکوس و حاصلضربهای ماتریسهای پایین‌مثلثی بازهم پایین‌مثلثی‌اند و همینطور است برای ماتریسهای بالامثلثی. طرفهای چپ و راست (۵.۳.۸) به‌ترتیب پایین‌مثلثی و بالامثلثی هستند. بنابراین باید با یک ماتریس قطری، که آن را D می‌نامیم، برابر باشند، و

$$L_1 = L_2 D \qquad U_1 = D^{-1} U_2 \qquad (6.3.8)$$

انتخاب D مستقیماً درگرو انتخاب عناصر قطری L و U است و وقتی آنها انتخاب شدند، D به‌طور یکتا معین می‌شود.

اگر بخواهند تمام عناصر قطری L برابر یک باشند، تجزیهٔ $A = LU$ همان تجزیه‌ای است که در حذف گاوسی، در بخش ۱.۸ به‌دست آمده است. روش فشردهٔ متناظر، فرمولهای صریحی برای l_{ij} و u_{ij} به‌دست می‌دهد که به روش **دولیتل**[1] معروف شده است. اگر بخواهیم تمام عناصر قطری U یک باشند، روش فشردهٔ متناظر برای محاسبهٔ $A = LU$ **روش کروت**[2] نامیده می‌شود. فقط یک ماتریس قطری ضربی آن را از روش دولیتل متمایز می‌سازد. برای الگوریتمی که از الگوریتم کروت برای تجزیهٔ (۳.۳.۸) با محورگیری جزئی و مقیاس‌دهی ضمنی استفاده می‌کند، برنامهٔ *unsymdet* ویلکینسن و راینش (۱۹۷۱، صص ۹۳ـ۱۱۰) را ببینید. در بعضی موردها، روش کروت نسبت به روش معمولی دولیتل ارجحیّت دارد.

امتیاز اصلی فرمول فشرده آن است که عناصر l_{ij} و u_{ij} همه در حاصلضربهای داخلی ظاهر می‌شوند، همان‌گونه که در فرمولهای (۱۴.۳.۸) ـ (۱۵.۳.۸) در تجزیهٔ ماتریس متقارن معین مثبت، در زیر دیده می‌شود. این حاصلضربهای داخلی و شاید تقسیم پایانی را می‌توان با استفاده از حساب با دقت مضاعف گردآوری کرد و سپس با یک دقت ساده گرد کرد. این طریق محاسبهٔ حاصلضربهای داخلی در فصل یک، قبل از فرمول خطای (۱۹.۵.۱) مورد بحث قرار گرفته است. با این استفادهٔ محدود از حساب با دقت مضاعف، می‌توان دقت عاملهای L و U را افزایش داد و این کار با روش حذف معمولی ممکن نخواهد بود مگر آنکه تمام عملیات و ذخیره‌سازی با حساب دقت مضاعف انجام شوند [برای بحث کامل این روشهای فشرده ویلکینسن (۱۹۶۵، صص ۱۲۸ـ۲۲۱) و گلوب و ون‌لون (۱۹۸۳، بخش ۵.۱) را ببینید].

روش چولسکی[3] گیریم A یک ماتریس متقارن، معین مثبت از مرتبهٔ n باشد. ماتریس A معین مثبت است اگر به ازای تمام مقادیر $x \in \mathbf{R}^n$ و $x \neq 0$

$$(Ax, x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j > 0 \qquad (۷.۳.۸)$$

بعضی از ویژگیهای ماتریسهای معین مثبت در مسألهٔ ۱۴ فصل ۷ و مسائل ۹ و ۱۱ و ۱۲ این فصل داده شده‌اند. ماتریسهای متقارن معین مثبت در کاربردهای متنوّع و گسترده‌ای ظاهر می‌شوند.

برای یک چنین ماتریس A، تجزیه به عوامل بسیار مناسبی وجود دارد که می‌توان بدون احتیاج به محورگیری و مقیاس‌دهی آن را انجام داد. این تجزیه **روش چولسکی** خوانده می‌شود،

1. Doolittle 2. Crout 3. Cholesky

که می‌گوید می‌توانیم یک ماتریس حقیقی پایین مثلثی L پیدا کنیم به‌طوری که

$$A = LL^T \qquad (۸.۳.۸)$$

در این روش به جای $n^۲$ محل حافظه برای L فقط به $\frac{۱}{۲}n(n+۱)$ محل حافظه نیاز است و تعداد عملیات، به‌جای $\frac{۱}{۳}n^۳$ در حدود $\frac{۱}{۶}n^۳$ است، که در تجزیهٔ معمولی مورد نیاز بود.

برای اثبات امکان (۸.۳.۸)، یک راه به‌دست آوردن L بر پایهٔ استقرا را بیان می‌کنیم. فرض می‌کنیم که نتیجه برای ماتریسهای متقارن معین مثبت از مرتبهٔ نابزرگتر از $n-۱$ درست باشد. نشان می‌دهیم که نتیجه برای تمام این‌گونه ماتریسهای مرتبهٔ n نیز درست است. ماتریس مطلوب L از مرتبهٔ n را به صورت

$$L = \begin{bmatrix} \hat{L} & ۰ \\ \gamma^T & x \end{bmatrix}$$

می‌نویسیم که $\hat{L}$ یک ماتریس مربعی از مرتبهٔ $n-۱$ است، $\gamma \in \mathbf{R}^{n-۱}$ و x یک اسکالر است. L باید طوری انتخاب شود که در $A = LL^T$ صدق کند:

$$\begin{bmatrix} \hat{L} & ۰ \\ \gamma^T & x \end{bmatrix} \begin{bmatrix} \hat{L}^T & \gamma \\ ۰ & x \end{bmatrix} = A = \begin{bmatrix} \hat{A} & c \\ c^T & d \end{bmatrix} \qquad (۹.۳.۸)$$

که $\hat{A}$ از مرتبهٔ $n-۱$، $c \in \mathbf{R}^{n-۱}$ و $d = a_{nn}$ حقیقی است. چون (۷.۳.۸) برای A درست است x_n را صفر می‌گیریم، تا گزارهٔ مشابهی برای $\hat{A}$ به‌دست آوریم که نشان دهد $\hat{A}$ نیز معین مثبت و متقارن است. به‌علاوه، $d > ۰$، زیرا کافی است در (۷.۳.۸) بگیریم $x_۱ = x_۲ = \ldots = x_{n-۱} = ۰$ و $x_n = ۱$. از ضرب در (۹.۳.۸)، $\hat{L}$ را با فرض استقرا که باید در

$$\hat{L}\hat{L}^T = \hat{A} \qquad (۱۰.۳.۸)$$

صدق کند انتخاب می‌کنیم، سپس γ را از حل

$$\hat{L}\gamma = c \qquad (۱۱.۳.۸)$$

به‌دست می‌آوریم. γ به‌دست می‌آید زیرا از $\det(\hat{A}) = [\det[\hat{L}]^۲]$ نتیجه می‌شود که $\hat{L}$ ناتکین است. بالاخره، x باید در رابطهٔ زیر صدق کند

$$\gamma^T\gamma + x^۲ = d \qquad (۱۲.۳.۸)$$

برای آنکه به مثبت‌بودن x^2 پی ببریم، دترمینان طرفین (۹.۳.۸) را تشکیل می‌دهیم و نتیجه می‌گیریم

$$[\det(\hat{L})^2]x^2 = \det(A) \qquad (۱۳.۳.۸)$$

چون دترمینان A حاصلضرب ویژه مقدارهای A است، و چون ویژه‌مقدارهای ماتریس متقارن معین مثبت، مثبت‌اند (مسألهٔ ۱۴ فصل ۷ را ببینید)، $\det(A)$ مثبت است. همچنین با فرض استقرا، $\hat{L}$ حقیقی است. بنابراین x^2 در (۱۳.۳.۸) مثبت است، و x را ریشهٔ دوّم مثبت آن می‌گیریم. چون رابطهٔ (۸.۳.۸) به‌طور نمایان برای ماتریس از مرتبهٔ $n = 1$ درست است، پس اثبات تجزیهٔ (۸.۳.۸) کامل شده است. برای یک روش دیگر، بخش ۲.۵، گلوب و ون‌لون (۱۹۸۳) را ببینید.

یک راه ساختن عملی L را می‌توان بر مبنای (۹.۳.۸) ـ (۱۲.۳.۸) انجام داد، ولی ما روشی را ارائه می‌دهیم که بر پایهٔ پیداکردن مستقیم عناصر L نهاده شده است. گیریم $L = [l_{ij}]$، $l_{ij} = 0$ برای $j > i$. ساختن L را بدین شکل آغاز می‌کنیم که سطر اول L را در ستون اول L^T ضرب می‌نماییم تا به‌دست آوریم

$$l_{11}^2 = a_{11}$$

چون A معین مثبت است، $a_{11} > 0$ و $l_{11} = \sqrt{a_{11}}$. سطر دوّم L را در دو ستون اوّل L^T ضرب می‌کنیم تا به‌دست آوریم

$$l_{21}l_{11} = a_{21} \qquad l_{21}^2 + l_{22}^2 = a_{22}$$

باز هم می‌توانیم دستگاه فوق را نسبت به مجهولهای l_{21} و l_{22} حل کنیم. به‌طور کلی[1] برای $i = 1, 2, \ldots, n$

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}}{l_{jj}} \qquad j = 1, \ldots, i-1 \qquad (۱۴.۳.۸)$$

$$l_{ii} = \left[a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2\right]^{1/2} \qquad (۱۵.۳.۸)$$

مقدار زیر رادیکال همان جملهٔ x^2 در دستور به‌دست آمدهٔ قبلی (۱۲.۳.۸) است و l_{ii} حقیقی و مثبت است. برای برنامه‌هایی که روش چولسکی را اجرا می‌کنند،

۱. توجه کنید به‌ازای $i = 1, 2, \ldots, n$، $l_{i1} = \frac{a_{i1}}{l_{11}}$ ـم.

دونگارا و همکاران (۱۹۷۹، فصل ۳) و ویلکینسن و راینش[1] (۱۹۷۱، صص ۱۰ـ۳۰) را ببینید.

به حاصلضربهای داخلی در (۱۴.۳.۸) و (۱۵.۳.۸) توجه نمایید. این حاصلضربهای داخلی را می‌توان با دقت مضاعف محاسبه کرد. و تعداد خطاهای گردکردن را مینیمم نمود، در این صورت خطا در عناصر l_{ij} خیلی کمتر از موقعی خواهد بود که با دقت ساده حساب شوند. همچنین توجه نمایید که عناصر L نسبت به A همچنان کراندار باقی می‌مانند، زیرا (۱۵.۳.۸) با استفاده از رابطهٔ زیر یک کران برای عناصر سطر i ام به‌دست می‌دهد

$$l_{i1}^2 + \ldots + l_{ii}^2 = a_{ii} \qquad (۱۶.۳.۸)$$

با یک اصلاح جزئی در (۸.۳.۸) می‌توان از به‌کاربردن ریشه‌های دوّم در (۱۵.۳.۸) ی روش چولسکی دوری کرد. یک ماتریس قطری D و یک ماتریس پایین‌مثلثی $\tilde{L}$ که عناصر قطری آن همه ۱ باشند، پیدا می‌کنیم به‌طوری که

$$A = \tilde{L}D\tilde{L}^T \qquad (۱۷.۳.۸)$$

این تجزیه به عوامل را می‌توان با همان تعدادِ عملیاتِ روش چولسکی، در حدود $\frac{1}{6}n^3$، بدون هیچ ریشه‌گیری انجام داد. برای بحث بیشتر و یک برنامه، ویلکینسن و راینش (۱۹۷۱، صص ۱۰ـ۳۰) را ببینید.

مثال ماتریس مرتبهٔ ۳ ی هیلبرت در زیر را در نظر می‌گیریم

$$A = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix} \qquad (۱۸.۳.۸)$$

در تجزیهٔ چولسکی،

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2\sqrt{3}} & 0 \\ \frac{1}{3} & \frac{1}{2\sqrt{3}} & \frac{1}{6\sqrt{5}} \end{bmatrix}$$

1. Reinsch

و در (۱۷.۳.۸)،

$$\tilde{L} = \begin{bmatrix} ۱ & ۰ & ۰ \\ \frac{۱}{۲} & ۱ & ۰ \\ \frac{۱}{۳} & ۱ & ۱ \end{bmatrix} \qquad D = \begin{bmatrix} ۱ & ۰ & ۰ \\ ۰ & \frac{۱}{۱۲} & ۰ \\ ۰ & ۰ & \frac{۱}{۱۸۰} \end{bmatrix}$$

برای بسیاری از دستگاههای خطی درکاربردها، ماتریس ضرایب A **نواری** است یعنی برای یک مقدارکوچک مثبت m

$$a_{ij} = ۰ \qquad \text{اگر} \quad |\, i-j \,| > m \tag{۱۹.۳.۸}$$

الگوریتمهای قبلی در چنین حالتی ساده خواهند شد و صرفه‌جویی قابل‌ملاحظه‌ای در زمان محاسبات به‌وجود خواهد آمد. برای چنین الگوریتم‌هایی وقتی A متقارن و معین مثبت باشد برنامه‌های LINPACK در دونگارا و همکاران (۱۹۷۹، فصل ۴) را ببینید.

دستگاههای سه قطری ماتریس $A = [a_{ij}]$ **سه قطری** است اگر

$$a_{ij} = ۰ \qquad \text{برای} \quad |\, i-j \,| > ۱ \tag{۲۰.۳.۸}$$

که شکل زیر را به‌دست می‌دهد

$$A = \begin{bmatrix} a_۱ & c_۱ & ۰ & ۰ & \dots & ۰ \\ b_۲ & a_۲ & c_۲ & ۰ & & \vdots \\ ۰ & b_۳ & a_۳ & c_۳ & & \\ & & & \ddots & & \\ \vdots & & & b_{n-۱} & a_{n-۱} & c_{n-۱} \\ ۰ & \dots & & ۰ & b_n & a_n \end{bmatrix} \tag{۲۱.۳.۸}$$

ماتریسهای سه‌قطری درکاربردهای متنوعی ظاهر می‌شوند. دستگاه خطی (۲۲.۷.۳) در توابع برازا در بخش ۷.۳ از فصل ۳ را به خاطر آورید. به‌علاوه، بسیاری از روشهای عددیِ حل مسائل مقدار مرزی در معادلات دیفرانسیل معمولی و معادلات دیفرانسیل جزئی متضمّن حل دستگاههای سه قطری‌اند. تقریباً تمام این کاربردها ماتریسهای سه‌قطری به‌دست می‌دهند که تجزیهٔ LU ی

آنها را می‌توان بدون محورگیری انجام داد و در نتیجه برای آنها افزایش زیادِ خطا پیش نمی‌آید. فرضهای دقیق برای A در قضیهٔ ۲.۸ در ذیل داده شده است.

با توجه به تجزیهٔ $A = LU$ بدون محورگیری، می‌بینیم که بیشتر عناصر L و U صفرند و برای تجزیه به فرمول کلی زیر می‌رسیم:

$$A = LU = \begin{bmatrix} \alpha_1 & 0 & \dots & & 0 \\ b_2 & \alpha_2 & 0 & & \vdots \\ 0 & b_3 & \alpha_3 & & \\ \vdots & & \ddots & & \\ 0 & & \dots & b_n & \alpha_n \end{bmatrix} \begin{bmatrix} 1 & \gamma_1 & 0 & & \dots & 0 \\ 0 & 1 & \gamma_2 & & 0 & \vdots \\ \vdots & & & \ddots & & \\ & & & & 1 & \gamma_{n-1} \\ 0 & & \dots & & 0 & 1 \end{bmatrix}$$

می‌توانیم از ضرب دو ماتریس، یک راه بازگشتی، برای محاسبهٔ $\{\alpha_i\}$ و $\{\gamma_i\}$ به‌دست آوریم

$$\begin{aligned} a_1 &= \alpha_1 \qquad \alpha_1\gamma_1 = c_1 \\ a_i &= \alpha_i + b_i\gamma_{i-1} \qquad i = 2, \dots, n \\ \alpha_i\gamma_i &= c_i \qquad i = 2, 3, \dots, n-1 \end{aligned} \tag{۲۲.۳.۸}$$

این معادلات را می‌توان حل کرد و به‌دست آورد

$$\begin{aligned} &\alpha_1 = a_1 \qquad \gamma_1 = \frac{c_1}{\alpha_1} \\ &\alpha_i = a_i - b_i\gamma_{i-1} \qquad \gamma_i = \frac{c_i}{\alpha_i} \qquad i = 2, 3, \dots, n-1 \\ &\alpha_n = a_n - b_n\gamma_{n-1} \end{aligned} \tag{۲۳.۳.۸}$$

برای حل $LUx = f$، می‌گیریم $Ux = z$ و $Lz = f$. پس

$$\begin{aligned} z_1 &= \frac{f_1}{\alpha_1} \qquad z_i = \frac{f_i - b_i z_{i-1}}{\alpha_i} \qquad i = 2, 3, \dots, n \\ x_n &= z_n \qquad x_i = z_i - \gamma_i x_{i+1} \qquad i = n-1, n-2, \dots, 1 \end{aligned} \tag{۲۴.۳.۸}$$

مقادیر ثابت در (۲۳.۳.۸) را می‌توان برای استفادهٔ بعدی در حل دستگاه خطی $Ax = f$، برای هر f طرف راست دستگاه که خواسته شده باشد، ذخیره نمود.

با شمردن فقط ضربها و تقسیمها، تعداد عملیات در محاسبهٔ L و U برابر $2n - 2$ است؛ برای حل $Ax = f$، $3n - 2$ عمل دیگر لازم است. بنابراین به $5n - 4$ عمل برای حل $Ax = f$

در اولین بار نیاز داریم، برای هر f دیگر، با همان A، فقط $3n-2$ عمل لازم خواهد بود. این روشی بسیار سریع است. برای توضیح، توجه نمایید که A^{-1} معمولاً چگال است و بیشتر درایه‌های آن ناصفرند؛ بنابراین محاسبه $x=A^{-1}f$ به n^2 عمل نیاز دارد. در بسیاری از کاربردها، n ممکن است بزرگتر از ۱۰۰۰ باشد و بنابراین یک صرفه‌جویی قابل توجهی در استفاده از (۲۳.۳.۸) ـ (۲۴.۳.۸) در مقایسه با سایر روشهای حل دستگاه، وجود دارد.

برای توجیه تجزیهٔ قبلی A، بخصوص نشان‌دادن اینکه تمام ضرایب $\alpha_i \neq 0$، قضیهٔ زیر را داریم:

قضیهٔ ۲.۸ فرض می‌کنیم که ضرایب $\{a_i, b_i, c_i\}$ در (۲۱.۳.۸) در شرایط زیر صدق می‌کنند:

۱. $|a_1| > |c_1| > 0$

۲. $|a_i| \geq |b_i| + |c_i|$، $b_i, c_i \neq 0$، $i = 2, \ldots, n-2$

۳. $|a_n| > |b_n| > 0$

در این صورت A ناتکین است و

$$|\gamma_i| < 1 \qquad i = 1, \ldots, n-1$$

$$|a_i| - |b_i| < |\alpha_i| < |a_i| + |b_i| \qquad i = 2, \ldots, n$$

برهان برای اثبات، (صفحهٔ ۵۷) آیزکسون و کلر (۱۹۶۶) را ببینید. ملاحظه می‌کنید که کران آخری نشان می‌دهد که $|c_i| < |\alpha_i|$، برای $i = 2, \ldots, n-2$. بنابراین ضرایب L و U کراندار باقی می‌مانند، و هیچ مقسوم‌علیهی که تقریباً صفر باشد به‌کار نمی‌رود مگر به دلیل خطای گردکردن.

شرط مخالف صفربودن c_i و b_i اساسی نیست. برای مثال، اگر یکی از b_i ها مساوی صفر باشد دستگاه خطی را می‌توان به دو دستگاه جدید تجزیه کرد، یکی از مرتبهٔ $i-1$ و دیگری از مرتبهٔ $n-i+1$. برای مثال اگر

$$A = \begin{bmatrix} a_1 & c_1 & 0 & 0 \\ b_2 & a_2 & c_2 & 0 \\ 0 & 0 & a_3 & c_3 \\ 0 & 0 & b_4 & a_4 \end{bmatrix}$$

آنگاه $Ax = f$ را با تبدیل به دو دستگاه خطی زیر حل می‌کنیم

$$\begin{bmatrix} a_3 & c_3 \\ b_4 & a_4 \end{bmatrix} \begin{bmatrix} x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} f_3 \\ f_4 \end{bmatrix} \qquad \begin{bmatrix} a_1 & c_1 \\ b_2 & a_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 - c_2 x_3 \end{bmatrix}$$

و اثبات کامل می‌شود. ■

مثال ماتریس ضرایب درونیابی برازای (۲۲.۷.۳) از فصل ۳ را در نظر می‌گیریم. در آن ماتریس h_i را ثابت می‌گیریم و سپس از $h/۶$ در هر سطر فاکتور می‌گیریم. اگر توجه خود را به ماتریس مرتبهٔ چهار محدود کنیم، ماتریس حاصل چنین خواهد شد:

$$A = \begin{bmatrix} ۲ & ۱ & ۰ & ۰ \\ ۱ & ۴ & ۱ & ۰ \\ ۰ & ۱ & ۴ & ۱ \\ ۰ & ۰ & ۱ & ۲ \end{bmatrix}$$

با استفاده از روش (۲۳.۳.۸)، تجزیهٔ LU به شکل زیر خواهد بود

$$L = \begin{bmatrix} ۲ & ۰ & ۰ & ۰ \\ ۱ & \frac{۷}{۲} & ۰ & ۰ \\ ۰ & ۱ & \frac{۲۶}{۷} & ۰ \\ ۰ & ۰ & ۱ & \frac{۴۵}{۲۶} \end{bmatrix} \qquad U = \begin{bmatrix} ۱ & \frac{۱}{۲} & ۰ & ۰ \\ & ۱ & \frac{۲}{۷} & ۰ \\ ۰ & & ۱ & \frac{۷}{۲۶} \\ ۰ & ۰ & ۰ & ۱ \end{bmatrix}$$

که مثال را کامل می‌کند. و باید اشاره کرد که حل مسألهٔ درونیابی برازای درجهٔ سوّمی که در بخش ۷.۳ فصل ۳، توضیح داده شد مشکل نیست.

۴.۸ تحلیل خطا

تحلیل خطای روشهای حل $Ax = b$ را با بررسی پایداری جواب x نسبت به اختلالات کوچک b ی سمت راست معادله، آغاز می‌کنیم. طرح کلی بخش ۶.۱ فصل یک را دنبال، و به‌ویژه، ضریب وضعیت در (۶.۶.۱) را مطالعه می‌کنیم.

گیریم $Ax = b$، از مرتبهٔ n، به‌طور یکتا حلپذیر باشد، و جواب مسألهٔ اختلال‌یافتهٔ زیر را در نظر می‌گیریم

$$A\tilde{x} = b + r \qquad (۱.۴.۸)$$

فرض می‌کنیم $e = \tilde{x} - x$، و از تفاضل $Ax = b$ از رابطهٔ بالا به‌دست می‌آوریم

$$Ae = r \qquad e = A^{-۱}r \qquad (۲.۴.۸)$$

برای بررسی پایداری $Ax = b$، مانند (۶.۶.۱)، می‌خواهیم کمیّت

$$\frac{\| e \|}{\| x \|} \div \frac{\| r \|}{\| b \|} \qquad (۳.۴.۸)$$

را کراندار کنیم وقتی r تمام عناصری از $\mathbf{R}^n$ را که نسبت به b کوچک‌اند، اختیار می‌کند.

از (۲.۴.۸) نرم می‌گیریم تا حاصل شود

$$\| r \| \leq \| A \| \| e \| \qquad \| e \| \leq \| A^{-1} \| \| r \|$$

نابرابری اول را بر $\| x \| \| A \|$ و نابرابری دوّم را بر $\| x \|$ تقسیم می‌نماییم، پس

$$\frac{\| r \|}{\| A \| \| x \|} \leq \frac{\| e \|}{\| x \|} < \frac{\| A^{-1} \| \| r \|}{\| x \|}$$

نرم ماتریسی در اینجا، همان نرم عملگر ماتریسی است که توسط نرم برداری القا شده است. با استفاده از کرانهای

$$\| b \| \leq \| A \| \| x \| \qquad \| x \| \leq \| A^{-1} \| \| b \|$$

به‌دست می‌آوریم

$$\frac{1}{\| A \| \| A^{-1} \|} \cdot \frac{\| r \|}{\| b \|} \leq \frac{\| e \|}{\| x \|} \leq \| A \| \| A^{-1} \| \cdot \frac{\| r \|}{\| b \|} \qquad (۴.۴.۸)$$

با توجه به (۳.۴.۸)، این نتیجه معرفی **ضریب وضعیت** A:

$$\mathrm{Cond}(A) = \| A \| \| A^{-1} \| \qquad (۵.۴.۸)$$

را توجیه می‌کند. برای هر A ی داده شده، انتخابهایی برای b و r وجود دارند به‌طوری که یکی از نابرابریها در (۴.۴.۸) را به برابری تبدیل کنند. این دلیل دیگری برای معرفی ضریب وضعیت A است وقتی (۳.۴.۸) در نظر گرفته می‌شود. اثبات را به مسألهٔ ۲۰ واگذار می‌نماییم.

کمیّت ضریب وضعیت A با نرمی که به‌کار برده شود تغییر می‌نماید، ولی همیشه دارای کران پایین ۱ است، زیرا

$$1 \leq \| I \| = \| AA^{-1} \| \leq \| A \| \| A^{-1} \| = \mathrm{cond}(A)$$

اگر ضریب وضعیّت نزدیک ۱ باشد، با توجه به (۴.۴.۸) معلوم می‌شود که اختلال نسبی کوچک در b به اختلال نسبی کوچک در جواب x می‌انجامد. ولی اگر $\mathrm{cond}(A)$ بزرگ باشد، (۴.۴.۸)

بیان می‌کند که ممکن است اختلالات نسبی کوچکی برای b وجود داشته باشند که اختلالات نسبی بزرگی در x ایجاد نمایند.

چون (۵.۴.۸) با انتخاب نرم تغییر می‌کند، گاهی تعریف دیگری برای ضریب وضعیت به‌کار می‌بریم که مستقل از نرم است. از قضیهٔ ۸.۷، فصل ۷،

$$\text{cond}(A) \geq r_\sigma(A) r_\sigma(A^{-1})$$

چون ویژه‌مقدارهای A^{-1}، معکوس ویژه‌مقدارهای A هستند، نتیجهٔ زیر را خواهیم داشت

$$\text{cond}(A) \geq \frac{\underset{\lambda \in \sigma(A)}{\text{Max}} \mid \lambda \mid}{\underset{\lambda \in \sigma(A)}{\text{Min}} \mid \lambda \mid} \equiv \text{cond}(A)_* \qquad (۶.۴.۸)$$

که $\sigma(A)$ معرّف مجموعهٔ همهٔ ویژه‌مقدارهای A است.

مثال دستگاه خطی زیر را در نظر می‌گیریم

$$\begin{aligned} ۷x_۱ + ۱۰x_۲ &= b_۱ \\ ۵x_۱ + ۷x_۲ &= b_۲ \end{aligned} \qquad (۷.۴.۸)$$

برای ماتریس ضرایب

$$A = \begin{bmatrix} ۷ & ۱۰ \\ ۵ & ۷ \end{bmatrix}, \qquad A^{-۱} = \begin{bmatrix} -۷ & ۱۰ \\ ۵ & -۷ \end{bmatrix}$$

گیریم ضریب وضعیت در (۵.۴.۸) وقتی از نرم ماتریس $\| \cdot \|_p$ تولید شده است. با $\text{cond}(A)_p$ نمایش داده شود در این مثال

$$\text{cond}(A)_۱ = \text{cond}(A)_\infty = (۱۷)(۱۷) = ۲۸۹$$

$$\text{cond}(A)_۲ \doteq ۲۲۳ \qquad \text{cond}(A)_* \doteq ۱۹۸$$

این ضرایب وضعیت همگی بیان می‌کنند که (۷.۴.۸) ممکن است نسبت به تغییرات b ی طرف راست معادله حساس باشد. برای نشان دادن این امکان، حالت خاص زیر را در نظر می‌گیریم

$$\begin{aligned} ۷x_۱ + ۱۰x_۲ &= ۱ \\ ۵x_۱ + ۷x_۲ &= ۰٫۷ \end{aligned}$$

که دارای جواب

$$x_۱ = ۰, \qquad x_۲ = ۰ر۱$$

است. برای دستگاه اختلال یافتهٔ

$$۷\tilde{x}_۱ + ۱۰\tilde{x}_۲ = ۱ر۰۱$$

$$۵\tilde{x}_۱ + ۷\tilde{x}_۲ = ۰ر۶۹$$

جواب چنین است

$$\tilde{x}_۱ = -۰ر۱۷ \qquad \tilde{x}_۲ = ۰ر۲۲$$

تغییرات نسبی در x، در مقایسه با تغییرات نسبی در b ی طرف راست معادله، بسیار بزرگ‌اند.

یک دستگاه خطی که جوابش نسبت به تغییرات نسبی جزئی b ی طرف راست معادله ناپایدار باشد **بدوضع** خوانده می‌شود. دستگاه قبلی (۷.۴.۸)، تا حدّی بدوضع است، به‌ویژه وقتی که فقط در حل آن از حساب سه یا چهار رقم با ممیز شناور استفاده شده باشد. ضرایب وضعیت $\mathrm{cond}(A)$ و $\mathrm{cond}(A)_*$ شاخصهای نسبتاً خوبی برای نشان‌دادن بدوضعی هستند. از آنجا که این اختلالها با عامل ۱۰ بزرگ می‌شوند. احتمال دارد که دقّت در جواب با یک رقم کمتر به‌دست آید.

به‌طور کلی، اگر $\mathrm{cond}(A)_*$ بزرگ باشد، برای بعضی مقادیر b دستگاه $Ax = b$ نسبت به تغییرات r در b، خیلی حساس خواهد بود. گیریم λ_l و λ_u ویژه مقدارهای A باشند به‌طوری که

$$|\lambda_l| = \underset{\lambda\in\sigma(A)}{\mathrm{Min}} |\lambda| \qquad |\lambda_u| = \underset{\lambda\in\sigma(A)}{\mathrm{Max}} |\lambda|$$

و بنابراین

$$\mathrm{cond}(A)_* = \left|\frac{\lambda_u}{\lambda_l}\right| \qquad (۸.۴.۸)$$

گیریم x_l و x_u ویژه بردارهای متناظر باشند با $\|x_u\|_\infty = ۱ = \|x_l\|_\infty$. پس

$$Ax = \lambda_u x_u$$

دارای جواب $x = x_u$ است. و دستگاه

$$A\tilde{x} = \lambda_u x_u + \lambda_l x_l = \lambda_u\left[x_u + \frac{۱}{\mathrm{cond}(A)_*}x_l\right]$$

دارای جواب زیر است

$$\tilde{x} = x_u + x_l$$

اگر $\text{cond}(A)_*$ بزرگ باشد، طرف راست فقط یک اختلال نسبی کوچک دارد؛

$$\frac{\| r \|_\infty}{\| b \|_\infty} = \frac{1}{\text{cond}(A)_*} \qquad (۹.۴.۸)$$

ولی جواب، اختلال نسبی بسیار بزرگتر زیر را دارد

$$\frac{\| \tilde{x} - x \|_\infty}{\| x \|_\infty} = \frac{\| x_l \|_\infty}{\| x_u \|_\infty} = 1 \qquad (۱۰.۴.۸)$$

دستگاههایی وجود دارند که در عمل بدوضع نیستند ولی ضرایب وضعیت قبلی برای آنها کاملاً بزرگ‌اند. برای مثال در ماتریس

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-10} \end{bmatrix}$$

ضرایب وضعیت $\text{cond}(A)_p$ و $\text{cond}(A)_*$ همگی برابر 10^{10} هستند. ولی معمولاً این ماتریس بدوضع تلقی نمی‌شود. مشکل در استفاده از نرم‌ها برای اندازه‌گیری تغییرات بردار است، به جای آنکه به هر مؤلفه جداجدا توجه شود. اگر مقیاس‌دهی در ماتریس ضرایب و بردار مجهول اجرا شود، این مسأله معمولاً پیش نمی‌آید و در آن صورت ضرایب وضعیت معمولاً پیشگوی دقیقی برای بدوضعی می‌شوند.

در یک توجیه نهایی برای استفاده از $\text{cond}(A)$ به‌عنوان ضریب وضعیت، قضیهٔ زیر را می‌آوریم.

قضیهٔ ۳.۸ (گاستینل[1]) گیریم A ماتریس ناتکینی از مرتبهٔ n باشد و $\| \cdot \|$ معرف یک نرم ماتریس عملگر در این صورت

$$\frac{1}{\text{cond}(A)} = \text{Min} \left\{ \frac{\| A - B \|}{\| A \|} \middle| B \text{ یک ماتریس تکین} \right\} \qquad (۱۱.۴.۸)$$

که $\text{cond}(A)$ همان است که در (۵.۴.۸) تعریف شده است.

برهان کاهان[2] (۱۹۶۶، ص ۷۷۵) را ببینید. ■

این قضیه بیان می‌کند که از لحاظ خطای نسبی، A را می‌توان به‌خوبی با یک ماتریس تکین تقریب زد اگر و تنها اگر $\text{cond}(A)$ خیلی بزرگ باشد. و از دید ما، ماتریس تکین B در حدّ اعلای

1. Gastinel 2. Kahan

جدول ۲.۸ ضرایب وضعیت برای ماتریس هیلبرت

n	$\text{cond}(H_n)_*$	n	$\text{cond}(H_n)_*$
۳	۵٫۲۴E + ۲	۷	۴٫۷۵E + ۸
۴	۱٫۵۵E + ۴	۸	۱٫۵۳E + ۱۰
۵	۴٫۷۷E + ۵	۹	۴٫۹۳E + ۱۱
۶	۱٫۵۰E + ۷	۱۰	۱٫۶۰E + ۱۳

بدوضعی است. با ویژه‌بردار متناظر با ویژه‌مقدار $\lambda = 0$ اختلالاتی ناصفر برای جواب وجود دارند که متناظر با اختلال صفر در b، طرف راست معادله هستند. مهمتر از آن، مقادیری از b وجود دارند که برای آنها $Bx = b$ دیگر حلپذیر نیست.

ماتریس هیلبرت ماتریس هیلبرت از مرتبهٔ n چنین تعریف می‌شود

$$H_n = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & & \frac{1}{n+1} \\ \vdots & & & & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & & \frac{1}{2n-1} \end{bmatrix} \qquad (۱۲.۴.۸)$$

این ماتریس به‌طور طبیعی در حل مسألهٔ تقریب باکمترین مربعات پیوسته ظاهر می‌شود. اصل آن در نزدیک به انتهای بخش ۳.۴ از فصل ۴ آمده، همراه با دستگاه خطی حاصل که در (۱۴.۳.۴) داده شده است. همان‌طورکه در بخش ۳.۴ اشاره شد و به دنبال (۹.۶.۱) در بخش ۴.۱ فصل ۱ نشان داده شد، ماتریس هیلبرت بسیار بدوضع است و با افزایش n بدوضعتر می‌شود. بدین لحاظ این، یک مثال عددی موردپسندی، برای بررسی برنامه‌های حل دستگاههای معادلات خطی، و تعیین حدود کارایی برنامه هنگام برخورد با مسائل بدوضع، بوده است. جدول ۲.۸ ضریب وضعیت $\text{cond}(H_n)_*$ را برای چند مقدار n نشان می‌دهد. ماتریس معکوس $H_n^{-1} = [\alpha_{ij}^{(n)}]$ به‌صورت صریح معلوم است:

$$\alpha_{ij}^{(n)} = \frac{(-1)^{i+j}(n+i-1)!(n+j-1)!}{(i+j-1)[(i-1)!(j-1)!]^2(n-i)!(n-j)!} \qquad 1 \le i,j \le n \qquad (۱۳.۴.۸)$$

برای اطلاع بیشتر از H_n، از جمله یک فرمول مجانبی برای $\text{Cond}(H_n)_*$، گرگوری[1] و کارنی[2] (۱۹۶۹، صص ۳۳ـ۳۸ و ۶۶ـ۷۳) را ببینید.

1. Gregory 2. Karney

اگرچه ماتریس هیلبرت به‌عنوان مثال زیاد به‌کار برده می‌شود، ولی باید مواظب باشیم و ببینیم که جواب درست چیست. گیریم $\overline{H}_n$ معرّف H_n پس از ورود به حساب متناهی یک رایانه باشد. برای یک برنامهٔ وارون ماتریس، نتایج برنامه باید با $\overline{H}_n^{-1}$ مقایسه شود، نه با H_n^{-1}؛ این دو ماتریس وارون ممکن است کاملاً متفاوت باشند. برای مثال، اگر از حساب ممیز شناور چهار رقم اعشاری با گردکردن استفاده کنیم، آنگاه

$$\overline{H}_3 = \begin{bmatrix} 1.000 & 0.5000 & 0.3333 \\ 0.5000 & 0.3333 & 0.2500 \\ 0.3333 & 0.2500 & 0.2000 \end{bmatrix} \qquad (14.4.8)$$

گردکردن فقط در بسط $\frac{1}{3}$ به شکل کسر اعشاری صورت گرفته است. در این صورت

$$H_3^{-1} = \begin{bmatrix} 9.000 & -36.00 & 30.00 \\ -36.00 & 192.0 & -180.0 \\ 30.00 & -180.0 & 180.0 \end{bmatrix}$$

$$\overline{H}_3^{-1} = \begin{bmatrix} 9.062 & -36.32 & 30.30 \\ -36.32 & 193.7 & -181.6 \\ 30.30 & -181.6 & 181.5 \end{bmatrix} \qquad (15.4.8)$$

هر برنامهٔ وارون ماتریس، که برای $\overline{H}_3$ به‌کار برده شده باشد جوابش باید با $\overline{H}_3^{-1}$ مقایسه شود، نه با H_3^{-1}. ما به این مسأله بعداً در بخش ۵.۸ باز می‌گردیم.

کرانهای خطا اکنون آثار خطای گردکردن روی جواب $\hat{x}$ از $Ax = b$ را که با حذف گاوسی به‌دست آمده است، در نظر می‌گیریم. مطلب را با ارائهٔ قضیه‌ای که خطا را کراندار می‌کند، وقتی b و A به اندازهٔ کوچکی تغییر یافته‌اند، آغاز می‌کنیم. این خود قضیهٔ مفیدی است، و برای تحلیل خطای حذف گاوسی که بعداً خواهد آمد، لازم است.

قضیهٔ ۴.۸ دستگاه $Ax = b$ را، وقتی که A ناتکین است، در نظر می‌گیریم. فرض می‌کنیم δA و δb اختلالات A و b باشند و

$$\| \delta A \| < \frac{1}{\| A^{-1} \|} \qquad (16.4.8)$$

پس $A + \delta A$ ناتکین است. و اگر δx را به‌طور ضمنی با رابطهٔ

$$(A+\delta A)(x+\delta x)=b+\delta b \qquad (۱۷.۴.۸)$$

تعریف کنیم، آنگاه

$$\frac{\| \delta x \|}{\| x \|} \leq \frac{\text{cond}(A)}{۱-\text{cond}(A)\frac{\| \delta A \|}{\| A \|}} \cdot \left\{ \frac{\| \delta A \|}{\| A \|} + \frac{\| \delta b \|}{\| b \|} \right\} \qquad (۱۸.۴.۸)$$

برهان ابتدا باید توجه داشت که δA معرّف ماتریسی است که در رابطهٔ (۱۶.۴.۸) صدق می‌کند، نه یک ثابت δ که در ماتریس A ضرب شده است، و همچنین است برای δb و δx. با استفاده از (۱۶.۴.۸) ناتکینی $A+\delta A$ بلافاصله از قضیهٔ ۱۲.۷ فصل ۷ نتیجه می‌شود. با توجه به (۱۱.۴.۷)،

$$\| (A+\delta A)^{-۱} \| \leq \frac{\| A^{-۱} \|}{۱- \| A^{-۱} \| \| \delta A \|} \qquad (۱۹.۴.۸)$$

از حل (۱۷.۴.۸) نسبت به δx و استفاده از $Ax=b$،

$$(A+\delta A)\delta x + Ax + (\delta A)x = b+\delta b$$
$$\delta x = (A+\delta A)^{-۱}[\delta b - (\delta A)x]$$

با استفاده از (۱۹.۴.۸) و تعریف (۵.۴.۸) برای $\text{cond}(A)$،

$$\| \delta x \| \leq \frac{\text{cond}(A)}{۱-\text{cond}(A)\frac{\| \delta A \|}{\| A \|}} \cdot \left\{ \frac{\| \delta b \|}{\| A \|} + \| x \| \frac{\| \delta A \|}{\| A \|} \right\}$$

طرفین را بر $\| x \|$ تقسیم و از $\| x \| \leq \| A \| \| b \|$ استفاده می‌کنیم تا (۱۸.۴.۸) حاصل شود. ■

تحلیل اثر خطاهای گردکردن در حذف گاوسی منسوب به ج. ه. ویلکینسن است و می‌توان آن را در ویلکینسن (۱۹۶۳، صص ۹۴ـ۹۹)، (۱۹۶۵، صص ۲۰۹ـ۲۱۶)، فورسایت و مولر (۱۹۶۷، فصل ۲۱)، و گلوب و ون‌لون (۱۹۸۳، فصل ۴) یافت. گیریم $\hat{x}$ جواب محاسبه‌شدهٔ $Ax=b$ باشد. محاسبهٔ مستقیم اثرات گردکردن x در هر مرحله، به‌عنوان وسیله‌ای برای به‌دست آوردن یک کران $\|\hat{x}-x\|$، دشوار است. گرچه نمایان نیست، ولی ساده‌تر آن است که $\hat{x}$ و الگوریتم حذف را در نظر گرفته به عقب برگردیم و نشان دهیم که $\hat{x}$جواب دقیق دستگاه

$$(A+\delta A)\hat{x}=b$$

است که در آن می‌توان کرانهای δA را تعیین کرد. این شیوه به **تحلیل پسرو خطا** معروف است. در این صورت می‌توانیم قضیهٔ قبلی ۴.۸ را برای کراندارکردن $\| x - \hat{x} \|$ به‌کار بریم. در قضیهٔ زیر، نرم ماتریس $\| A \|_\infty$، نرم سطری (۱۷.۳.۷) است که بر اثر نرم برداری $\| x \|_\infty$ القا شده است.

قضیهٔ ۵.۸ گیریم A از مرتبهٔ n و ناتکین باشد، و فرض می‌کنیم از محورگیری جزئی یا کلی در فرایند حذف استفاده شده باشد. تعریف می‌کنیم

$$\rho = \frac{1}{\| A \|_\infty} \underset{1 \le i,j,k \le n}{\text{Max}} | a_{ij}^{(k)} | \qquad (۲۰.۴.۸)$$

گیریم u واحد گردکردن در رایانهٔ مورد استفاده باشد [(۱۱.۲.۱) و (۱۲.۲.۱) را برای تعریف u ببینید]،

۱. ماتریسهای L و U که با حذف گاوسی محاسبه شده‌اند در روابط زیر صدق می‌کنند:

$$LU = A + E$$

$$\| E \|_\infty \le n^3 \rho \| A \|_\infty u \qquad (۲۱.۴.۸)$$

۲. $\hat{x}$، جواب تقریبی $Ax = b$ که با استفاده از حذف گاوسی محاسبه شده در رابطهٔ زیر صدق می‌کند

$$(A + \delta A)\hat{x} = b \qquad (۲۲.۴.۸)$$

با

$$\frac{\| \delta A \|_\infty}{\| A \|_\infty} \le \{1.01(n^3 + 3n^2)\rho u\} \qquad (۲۳.۴.۸)$$

۳. با استفاده از قضیهٔ ۴.۸،

$$\frac{\| x - \hat{x} \|_\infty}{\| x \|_\infty} \le \frac{\text{cond}(A)_\infty}{1 - \text{cond}(A)_\infty \dfrac{\| \delta A \|_\infty}{\| A \|_\infty}} [1.01(n^3 + 3n^2)\rho u] \qquad (۲۴.۴.۸)$$

برهان اثبات ۱ و ۲ در (فصل ۲۱) فورسایت و مولر (۱۹۶۷) داده شده است. صورتهای مختلف این قضایا در (فصل ۴) گلوب و ون‌لون (۱۹۸۳) آمده است. ■

به روش تجربی، کران (۲۴.۴.۸)، به علّت حذف خطاهای گردکردن با اندازه و علامت متغیّر، بسیار بزرگ است. به گفته ویلکینسن (۱۹۶۳، ص ۱۰۸)، یک کران تجربی بهتر در اغلب حالات

$$\frac{\| \delta A \|_\infty}{\| A \|_\infty} \le nu \qquad (۲۵.۴.۸)$$

است. فرمول (۲۴.۴.۸) اهمیت اندازهٔ $\mathrm{cond}(A)$ را نشان می‌دهد.

کمیت ρ در کرانها را می‌توان در خلال فرایند حذف محاسبه کرد و می‌توان آن را از پیش نیز کراندار کرد. برای محورگیری کامل، یک کران از پیش تعیین شده چنین است

$$\rho \leq ۱٫۸n^{(\ln n)/۴} \qquad n \geq ۱$$

و حدس زده‌اند که برای یک مقدار c، $\rho \leq cn$. برای محورگیری جزئی، یک کران از پیش تعیین شده $۲^{n-۱}$ است و مثالهای غیرعادی دالّ‌بر امکان این امر وجود دارد. با این همه، در تمام مطالعات تجربی تا به امروز، ρ با یک عدد نسبتاً کوچک مستقل از n کراندار شده‌است. به لحاظ اختلاف در کرانهای نظری برای ρ، گاهی محورگیری کلی را ترجیح می‌دهند. ولی، در عمل، رفتار خطا با محورگیری جزئی به همان خوبی محورگیری کلی است. به‌علاوه، محورگیری کلی، نیاز به مقایسه‌های خیلی بیشتری در هر مرحله از فرایند حذف خواهد داشت. نتیجتاً محورگیری جزئی روشی است که در همهٔ برنامه‌های جدید رایانه‌یی حذف گاوسی به‌کار می‌رود.

یکی از مهمترین نتایج تحلیل بالا این است که نشان می‌دهد حذف گاوسی یک فرایند بسیار پایدار است، فقط مشروط بر آنکه ماتریس A خیلی بدوضع نباشد. از نظر تاریخی، پژوهشگران در اوایل دههٔ ۱۹۵۰، از پایداری حذف گاوسی برای دستگاههای بزرگ، مثلاً $n \geq ۱۰$، مطمئن نبودند، ولی این مسأله اکنون حل شده است.

اندازهٔ مانده در جواب محاسبه شدهٔ $\hat{x}$، یعنی

$$r = b - A\hat{x} \qquad (۲۶.۴.۸)$$

گاهی به اشتباه به اندازهٔ خطای $x - \hat{x}$ ارتباط داده شده است. درواقع، خطای $\hat{x}$ ممکن است بزرگ و r کوچک باشد و این حالت معمولاً در مسائل بدوضع به‌وجود می‌آید. از (۲۶.۴.۸) و $Ax = b$،

$$r = A(x - \hat{x})$$

$$x - \hat{x} = A^{-۱}r \qquad (۲۷.۴.۸)$$

بنابراین اگر $A^{-۱}$ عناصر بزرگی داشته باشد، $x - \hat{x}$ ممکن است بسیار بزرگتر از r شود.

در عمل، ماندهٔ r حتی برای مسائل بدوضع خیلی کوچک است. برای اینکه ببینیم چرا این امر اتفاق می‌افتد، از (۲۲.۴.۸) استفاده می‌کنیم تا به‌دست آوریم

$$r = (\delta A)\hat{x}$$

$$\| r \|_\infty \leq \| \delta A \|_\infty \| \hat{x} \|_\infty$$

$$\frac{\| r \|_\infty}{\| A \|_\infty \| \hat{x} \|_\infty} \leq \frac{\| \delta A \|_\infty}{\| A \|_\infty} \qquad (۲۸.۴.۸)$$

کرانهای $\| A \| / \| \delta A \|$ در (۲۳.۴.۸) یا در (۲۵.۴.۸) مستقل از بدوضعی یا خوش‌وضعی مسأله هستند. بنابراین $\| r \|_\infty$ نسبت به $\| \hat{x} \| \| A \|$، معمولاً کوچک است. جملهٔ اخیر اغلب نزدیک به $\| b \|$ یا به اندازهٔ آن است، زیرا $b = Ax$، پس $\| r \|$ نسبت به $\| b \|$ کوچک خواهد بود. به‌عنوان آخرین نکته در مورد اندازهٔ مانده، مسائلی هستند که در آنها نیاز به اینکه $x - \hat{x}$ کوچک باشد، نیست مهم آن است که فقط r کوچک باشد، در چنین حالتهایی، بدوضعی معنای خود را نخواهد داشت.

کرانهای (۱۸.۴.۸) و (۲۴.۴.۸) اهمیّت $\text{cond}(A)$ را در تعیین خطا، نشان می‌دهند. به‌طور کلی اگر $\text{cond}(A) \doteq ۱۰^m$، برای یک $m \geq ۰$، آنگاه در حدود m رقم از دقت در محاسبهٔ $\hat{x}$ از تعداد ارقامی که در حساب رایانه‌یی به‌کار رفته، از دست خواهد رفت. بنابراین اندازه‌گیری $\text{cond}(A) = \| A \| \| A^{-۱} \|$ مطلوب خواهد بود. محاسبهٔ $\| A \|$ آسان و ارزان است و $\| A^{-۱} \|$ مسألهٔ اصلی در محاسبهٔ $\text{cond}(A)$ است. محاسبهٔ $A^{-۱}$ نیاز به $n^۳$ عمل دارد و این راه بسیار گرانی برای محاسبهٔ $\| A^{-۱} \|$ است. یک روش ارزانتر، با $O(n^۲)$ عمل دربستهٔ نرم‌افزاری LINPACK وجود دارد.

برای هر دستگاه $Ay = d$،

$$y = A^{-۱} d$$

$$\| y \| \leq \| A^{-۱} \| \| d \|$$

$$\| A^{-۱} \| \geq \frac{\| y \|}{\| d \|} \qquad (۲۹.۴.۸)$$

می‌خواهیم d را طوری انتخاب کنیم تا این نسبت، در حد ممکن بزرگ باشد. می‌نویسیم $A = LU$ که LU از حذف گاوسی به‌دست آمده است. پس حل $Ay = d$ هم‌ارز با حل

$$Lw = d \qquad Uy = w$$

است. وقتی $Lw = d$ را حل می‌کنیم، d را طوری پیدا می‌کنیم که w هرچه بیشتر بزرگ و $\| d \|_\infty = ۱$ نگهداشته شود. آنگاه $Uy = w$ را نسبت به y حل می‌کنیم. این عمل در (۲۹.۴.۸)، کرانی بهتر از d یی که به‌طور تصادفی انتخاب شده به‌دست خواهد داد. یک الگوریتم برای انتخاب d در گلوب و ون‌لون (۱۹۸۳، ص ۷۷) داده شده است. الگوریتم LINPACK تعمیمی پیچیده‌تر از الگوریتم قبلی است. برای توضیح گلوب و ون‌لون (۱۹۸۳، ص ۷۸) را ببینید.

کرانهای پسین خطا با کرانهای خطا برای وارون حساب‌شدهٔ C از ماتریس داده‌شدهٔ A شروع می‌کنیم. ماتریس مانده را به شکل زیر تعریف می‌کنیم

$$R = I - CA$$

قضیهٔ ۶.۸ اگر $\| R \| < ۱$، آنگاه A و C ناتکین هستند و

$$\frac{\| R \|}{\| A \| \| C \|} \leq \frac{\| A^{-۱} - C \|}{\| C \|} \leq \frac{\| R \|}{۱ - \| R \|} \qquad (۳۰.۴.۸)$$

برهان چون $\| R \| < ۱$، با استفاده از قضیهٔ ۱۱.۷ فصل ۷، $I - R$ ناتکین است، و

$$\| (I - R)^{-۱} \| \leq \frac{۱}{۱ - \| R \|}$$

ولی

$$I - R = CA \qquad (۳۱.۴.۸)$$

$$۰ \neq \det(I - R) = \det(CA) = \det(C)\det(A)$$

و بنابراین $\det(A)$ و $\det(C)$ هر دو مخالف صفرند. این نشان می‌دهد که A و C هر دو ناتکین‌اند.

برای کران پایین در (۳۰.۴.۸)،

$$R = I - CA = (A^{-۱} - C)A$$

$$\| R \| \leq \| A^{-۱} - C \| \| A \|$$

و تقسیم بر $\| C \| \| A \|$ قضیه را ثابت می‌کند. برای کران بالا، (۳۱.۴.۸) ایجاب می‌کند که

$$(I - R)^{-۱} = A^{-۱}C^{-۱}$$

$$A^{-۱} = (I - R)^{-۱}C \qquad (۳۲.۴.۸)$$

برای خطا در C،

$$A^{-۱} - C = (I - CA)A^{-۱} = RA^{-۱} = R(I - R)^{-۱}C$$

$$\| A^{-۱} - C \| \leq \frac{\| R \| \| C \|}{۱ - \| R \|}$$

که برهان کامل می‌شود. ■

جنبهٔ نظری این قضیه معمولاً مهمتر از جنبهٔ عملی آن است. همان‌گونه که قبلاً در بخش ۱.۸ اشاره شد، ماتریسهای وارون لازم نیست برای حل دستگاه خطی حساب شوند. و در نتیجه، به‌ندرت نیاز واقعی به کران خطا از نوع قبل، خواهد بود. استثنای عمده وقتی است که C تقریبی باشد که به وسیله‌ای غیر از حذف گاوسی، بیشتر از راههای نظری، به‌دست آمده باشد. در آن صورت، چنین وارونهای تقریبی، برای حل $Ax = b$، با روش تصحیح ماندهٔ (۳.۵.۸)، که در بخش بعد توضیح داده خواهد شد، به‌کار برده می‌شوند. در چنین حالتی، کران (۳۰.۴.۸) می‌تواند اطلاعات سودمندی دربارهٔ C به ما بدهد.

فرع گیریم A، C و R همانهایی باشند که در قضیهٔ ۶.۸ داده شده‌اند. فرض می‌کنیم $\hat{x}$ یک جواب تقریبی برای $Ax = b$ باشد و $r = b - A\hat{x}$. در این صورت

$$\| x - \hat{x} \| \leq \frac{\| Cr \|}{1 - \| R \|} \qquad (۳۳.۴.۸)$$

■

برهان داریم

$$r = b - A\hat{x} = Ax - A\hat{x} = A(x - \hat{x})$$

$$x - \hat{x} = A^{-1}r = (I - R)^{-1}Cr \qquad (۳۴.۴.۸)$$

با استفاده از (۳۲.۴.۸) که در معادلهٔ آخر به‌کار رفته است. اگر از معادلهٔ اخیر نرم بگیریم (۳۳.۴.۸) به‌دست می‌آید. ■

معلوم شده است که این کران (۳۳.۴.۸) کاملاً دقیق است، به‌ویژه در مقایسه با بعضی از کرانهایی که اغلب به‌کار می‌روند. برای یک بحث کامل دربارهٔ کرانهای محاسبه‌پذیر خطا، از جمله تعدادی مثال، اِرد ولینچ[1] (۱۹۷۵) را ببینید.

پیداکردن کران خطای (۳۳.۴.۸) نسبتاً گران تمام می‌شود. اگر فرض کنیم که $\hat{x}$ از حذف گاوسی به‌دست آمده است، $n^3/3$ عمل برای محاسبهٔ $\hat{x}$ و تجزیهٔ A به LU به‌کار رفته است. برای محاسبهٔ $C \doteq A^{-1}$ از راه حذف، حداقل $\frac{2}{3}n^3$ عمل اضافی لازم است، به‌دست آوردن CA نیاز به n^3 عمل ضرب دارد، و به‌دست آوردن Cr نیاز به n^2 ضرب. بنابراین کران خطا حداقل یک افزایش پنج برابر عمل در بر خواهد داشت. معمولاً ترجیح داده می‌شود که خطا از حل تقریبی معادلهٔ خطای

$$A(x - \hat{x}) = r$$

1. Aird and Lynch

با استفاده از LU که قبلاً ذخیره شده براورد شود. این امر نیاز به n^2 عمل برای محاسبهٔ r و n^2 عمل دیگر برای حل دستگاه خطی دارد. مگر وقتی که ماتریس $R = I - CA$ دارای نرم نزدیک به یک باشد، این روش یک خطای بسیار قابل قبولی به‌دست می‌دهد. این مطلب در بخش بعد دنبال و توضیح داده شده است.

۵.۸ روش تصحیح مانده

فرض می‌کنیم که $Ax = b$ حل شده، جواب تقریبی $\hat{x} \equiv x^{(0)}$ به‌دست آمده است. همچنین گیریم تجزیهٔ LU و تعداد تمام تعویضهای سطری یا ستونی ذخیره شده است. کمیت زیر را حساب می‌کنیم

$$r^{(0)} = b - Ax^{(0)} \qquad (۱.۵.۸)$$

تعریف می‌کنیم $e^{(0)} = x - x^{(0)}$. در این صورت مانند قبل در (۳۴.۴.۸)،

$$Ae^{(0)} = r^{(0)}$$

این دستگاه را با تجزیهٔ LU ی ذخیره شده حل کرده جواب تقریبی نتیجه را $\hat{e}^{(0)}$ می‌نامیم. یک جواب تقریبی جدید

$$x^{(1)} = x^{(0)} + \hat{e}^{(0)} \qquad (۲.۵.۸)$$

را برای $Ax = b$ تعریف می‌کنیم این فرایند را می‌توانیم تکرار و $x^{(2)}, \ldots$ را، با خطایی که مرتب کاهش می‌یابد، حساب کنیم. محاسبهٔ $r^{(0)}$، نیاز به n^2 عمل، و محاسبهٔ $\hat{e}^{(0)}$، به n^2 عمل دیگر نیاز دارد. بنابراین محاسبهٔ مقادیر بهبودیافتهٔ $x^{(1)}$، $x^{(2)}$، ...، در مقایسه با محاسبهٔ مقدار اصلی $x^{(0)}$، گران نخواهد بود. این روش را روش **بهبود بارستی** یا روش **تصحیح مانده** نیز می‌نامند.

به‌دست آوردن مقادیر دقیق برای $r^{(0)}$ بسیار اهمیت دارد. چون $x^{(0)}$ جواب تقریبی $Ax = b$ است، $r^{(0)}$ معمولاً در محاسبه‌اش خطاهای کاهش دقت خواهد داشت، وقتی $Ax^{(0)}$ و b تقریباً تا آخرین درجهٔ دقت حساب ماشین حساب شده‌اند، بنابراین در به‌دست آوردن مقادیر دقیق برای $r^{(0)}$، معمولاً باید به سمت حساب با دقت بالاتر برویم. اگر از همان درجه دقتی که برای حساب LU و $x^{(0)}$ به‌کار برده شده برای محاسبهٔ $r^{(0)}$ استفاده کنیم، این نتیجهٔ غیردقیق $r^{(0)}$ معمولاً به $\hat{e}^{(0)}$ ی منجر خواهد شد که یک تقریب ضعیف برای $e^{(0)}$ خواهد بود. در حساب با دقت ساده،

$r^{(0)}$ را با دقت مضاعف محاسبه می‌کنیم. ولی اگر محاسبات قبلاً با دقت مضاعف انجام شده باشد، معمولاً مشکل است که به دقت بالاتر برسیم.

مثال دستگاه $Ax = b$ را با $A = \overline{H}_4$ که در (۱۴.۴.۸) آمده با حساب ممیز شناور با چهار رقم اعشاری و گردکردن حل کنید. برای طرف راست از

$$b = [1, 0, 0]^T$$

استفاده می‌کنیم. جوابِ درست اولین ستون $\overline{H}_4^{-1}$ است، که با توجه به (۱۵.۴.۸)، تا چهار رقم معنی‌دار چنین است

$$x = [9{,}062, -36{,}32, 30{,}30]^T$$

با استفاده از حذف و محورگیری جزئی،

$$x^{(0)} = [8{,}968, -35{,}77, 29{,}77]^T$$

ماندهٔ $r^{(0)}$ با حساب با دقت مضاعف به‌دست آمده و سپس تا چهاررقم با معنی گرد شده است. مقادیر به‌دست آمده چنین‌اند

$$r^{(0)} = [-0{,}00\,5341, -0{,}00\,4359, -0{,}00\,5344]^T$$

از حل دستگاه $Ae^{(0)} = r^{(0)}$ با تجزیهٔ ذخیره‌شدهٔ LU،

$$\hat{e}^{(0)} = [0{,}09216, -0{,}5442, 0{,}5239]^T$$
$$x^{(1)} = [9{,}060, -36{,}31, 30{,}29]^T$$

از تکرار این عملیات،

$$r^{(1)} = [-0{,}000\,6570, -0{,}000\,3770, -0{,}000\,1980]^T$$
$$\hat{e}^{(1)} = [0{,}00\,1707, -0{,}0\,1300, 0{,}0\,1241]^T$$
$$x^{(2)} = [9{,}062, -36{,}32, 30{,}30]^T$$

بردار $x^{(2)}$ تا چهاررقم بامعنی دقیق است. ضمناً توجه کنید که $x^{(1)} - x^{(0)} = \hat{e}^{(0)}$ یک پیشگوی دقیق برای خطای $e^{(0)}$ در $x^{(0)}$ است.

فرمولها را می‌توان بسط داده برآورد نمود که چند بارست برای پیداکردن جواب x با دقت کامل، لازم است. برای بحث در مورد مطالب مربوطه و الگوریتم‌هایی که از این روش استفاده می‌کنند دونگارا و همکاران (۱۹۷۹، ص ۱۹)، فورسایت و مولر (۱۹۶۷، فصلهای ۱۳، ۱۶، ۱۷)، گلوب و ون‌لون (۱۹۸۳) و ویلکینسن و راینش (۱۹۷۱، صص ۹۳ـ۱۱۰) را ببینید.

روش دیگر تصحیح مانده مواردی وجود دارند که می‌توانیم یک وارون تقریبی C را برای ماتریس داده‌شدهٔ A محاسبه نماییم. این عمل معمولاً با بررسی دقیق ساختار A، و سپس استفاده از تکنیکهای متنوع تقریب در براورد A^{-1} انجام می‌گیرد. بدون توجه به منشأ C، نشان خواهیم داد که چگونه آن را به گونه‌ای بارستی برای حل $Ax = b$ به‌کار می‌بریم.

گیریم $x^{(0)}$ یک حدس اولیه باشد و تعریف می‌کنیم $r^{(0)} = b - Ax^{(0)}$. مانند قبل، $A(x - x^{(0)}) = r^{(0)}$، $x^{(1)}$ را به‌طور ضمنی با رابطهٔ زیر تعریف می‌کنیم

$$x^{(1)} - x^{(0)} = Cr^{(0)}$$

به‌طور کلی، تعریف می‌کنیم

$$r^{(m)} = b - Ax^{(m)} \qquad x^{(m+1)} = x^{(m)} + Cr^{(m)} \qquad m = 0, 1, 2 \ldots \qquad (۳.۵.۸)$$

اگر C یک تقریب خوبی برای A^{-1} باشد، همان‌گونه که در تحلیل زیر نشان داده خواهد شد این بارست به‌سرعت همگرا می‌شود.

ابتدا یک فرمول بازگشتی برای خطا به‌دست می‌آوریم:

$$\begin{aligned} x - x^{(m+1)} &= x - x^{(m)} - Cr^{(m)} = x - x^{(m)} - C[b - Ax^{(m)}] \\ &= x - x^{(m)} - C[Ax - Ax^{(m)}] \\ x - x^{(m+1)} &= (I - CA)(x - x^{(m)}) \end{aligned} \qquad (۴.۵.۸)$$

با استقرا

$$x - x^{(m)} = (I - CA)^m (x - x^{(0)}) \qquad m \geq 0 \qquad (۵.۵.۸)$$

اگر برای یک نرم ماتریسی،

$$\| (I - CA) \| < 1 \qquad (۶.۵.۸)$$

آنگاه با استفاده از نرم برداری وابسته

$$\| x - x^{(m)} \| \leq \| I - CA \|^m \| x - x^{(0)} \| \qquad (۷.۵.۸)$$

و این رابطه با هر انتخابی برای $x^{(0)}$ هرگاه $m \to \infty$، به صفر همگرا خواهد شد. به‌طورکلیتر، برای هر انتخاب $x^{(0)}$، $x^{(m)}$ به x همگرا خواهد بود اگر و تنها اگر

$$(I - CA)^m \to 0 \qquad m \to \infty \quad \text{وقتی}$$

و با توجه به قضیهٔ ۹.۷ از فصل ۷، این عبارت هم‌ارز است با

$$r_\sigma(I - CA) < 1 \qquad (۸.۵.۸)$$

برای شعاع خاص $I - CA$. این رابطه را ممکن است، حتی وقتی که (۶.۵.۸) برای یک نرم ماتریسی معمولی برقرار باشد، ثابت کرد. همچنین توجه نمایید که

$$I - AC = A(I - CA)A^{-1}$$

و بنابراین $I - AC$ و $I - CA$ ماتریسهای متشابه‌اند و دارای ویژه‌مقدارهای برابرند. اگر

$$\| I - AC \| < 1 \qquad (۹.۵.۸)$$

آنگاه (۸.۵.۸) درست است حتی اگر (۶.۵.۸) درست نباشد، و بازهم همگرایی برقرار است.

عبارت (۴.۵.۸) نشان می‌دهد که نرخ همگرایی $x^{(m)}$ به x خطی است:

$$\| x - x^{(m+1)} \| \leq c \| x - x^{(m)} \| \qquad m \geq 0 \qquad (۱۰.۵.۸)$$

با مقدار مجهول $c < 1$. ثابت c اغلب از راه محاسبه با رابطهٔ زیر براورد می‌شود

$$c = \text{Max} \frac{\| x^{(m+2)} - x^{(m+1)} \|}{\| x^{(m+1)} - x^{(m)} \|} \qquad (۱۱.۵.۸)$$

که ماکسیمم روی بعضی یا تمام بارستهایی که محاسبه شده، گرفته می‌شود. این رابطه دقیق نیست و از فرمول زیر حاصل شده است

$$x^{(m+2)} - x^{(m+1)} = (I - CA)(x^{(m+1)} - x^{(m)}) \qquad (۱۲.۵.۸)$$

برای اثبات این تساوی، تنها (۴.۵.۸) را به‌کار می‌بریم و فرمولها را برای مقادیر پیاپی m از هم کم می‌کنیم.

اگر فرض کنیم (۱۰.۵.۸) برای بارستهایی که محاسبه می‌کنیم معتبر باشد، و اگر یک براوردی برای c داشته باشیم، آنگاه می‌توانیم یک کران خطا بسازیم.

$$\begin{aligned} \| x^{(m+1)} - x^{(m)} \| &= \| [x - x^{(m)}] - [x - x^{(m+1)}] \| \\ &\geq \| x - x^{(m)} \| - \| x - x^{(m+1)} \| \\ &\geq \| x - x^{(m)} \| - c \| x - x^{(m)} \| \\ \| x - x^{(m)} \| &\leq \frac{1}{1-c} \| x^{(m+1)} - x^{(m)} \| \\ \| x - x^{(m+1)} \| &\leq \frac{c}{1-c} \| x^{(m+1)} - x^{(m)} \| \end{aligned} \qquad (۱۳.۵.۸)$$

برای بارستهایی که همگرایی آهسته دارند [با $c \doteq 1$]، این کران مهم است، زیرا در این صورت $\| x^{(m+1)} - x^{(m)} \|$ ممکن است بسیار کوچکتر از $\| x - x^{(n)} \|$ باشد. همچنین به‌دست آوردن کران در بخش ۵.۲ از فصل دوم را به‌یاد آورید. یک کران مشابه، (۵.۵.۲)، برای خطا در یک روش همگرای خطی به‌دست آمده بود.

مثال تعریف می‌کنیم $A(\varepsilon) = A_0 + \varepsilon B$، با

$$A_0 = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \qquad B = \begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix}$$

به‌عنوان یک وارون تقریبی برای $A(\varepsilon)$، از

$$A(\varepsilon)^{-1} \doteq C = A_0^{-1} = \begin{bmatrix} \frac{3}{4} & -\frac{1}{2} & \frac{1}{4} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ \frac{1}{4} & -\frac{1}{2} & \frac{3}{4} \end{bmatrix}$$

استفاده می‌کنیم. دستگاه $A(\varepsilon)x = b$ را می‌توان با روش تصحیح ماندهٔ (۳.۵.۸) حل کرد. برای تحلیل همگرایی

$$\begin{aligned} I - CA(\varepsilon) &= I - A_0^{-1}[A_0 + \varepsilon B] = -\varepsilon A_0^{-1} B \\ &= -\varepsilon \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ -\frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{4} & -\frac{1}{2} & -\frac{1}{4} \end{bmatrix} \end{aligned}$$

همگرایی وقتی تضمین می‌شود که

$$\| I - CA(\varepsilon) \|_\infty = | \varepsilon | < 1$$

و از (۴.۵.۸)

$$\| x - x^{(m+1)} \|_\infty \leq | \varepsilon | \| x - x^{(m)} \|_\infty \qquad m \geq 0$$

موارد زیادی از نوع این مثال وجود دارند. ممکن است لازم باشد دستگاههای خطی در شکل کلی $A(\varepsilon)x = b$ را برای هر ε نزدیک به صفر حل کنیم. برای صرفه‌جویی در وقت، یا $A(0)^{-1}$ یا LU، تجزیهٔ $A(0)$، را به‌دست می‌آوریم. در این صورت این را به‌عنوان یک وارون تقریبی برای $A(\varepsilon)$ به‌کار می‌بریم، و $A(\varepsilon)x = b$ را با استفاده از روش تصحیح مانده حل می‌کنیم.

۶.۸ روشهای بارستی

همان‌گونه که در مقدمهٔ این فصل اشاره شد، بسیاری از دستگاهها خیلی بزرگتر از آن هستند که با روشهای مستقیم، بر پایهٔ حذف گاوسی، حل شوند. برای این دستگاهها، روشهای بارستی، اغلب تنها راه حل ممکن و همچنین در بسیاری از حالات، سریعتر از حذف هستند. وسیعترین زمینه برای کاربرد روشهای بارستی دستگاههای خطی است که از حل عددی معادلات دیفرانسیل جزئی حاصل می‌شوند. دستگاههای از مراتب 10^3 تا 10^5 غیرمعمول نیستند، ولی تقریباً همهٔ ضرایب دستگاه صفرند. به‌عنوان یک مثال از چنین مسائلی، حل عددی معادلهٔ پواسن است که در بخش ۸.۸ مطالعه شده است. خواننده، در صورتی که بخواهد می‌تواند آن بخش را با این بخش با هم بخواند.

گذشته از بزرگ‌بودن، دستگاههای خطی که باید حل شوند، $Ax = b$، اغلب چندین ویژگی مهم دیگر دارند. معمولاً **تنک** هستند، یعنی درصد کوچکی از ضرایب ناصفرند. ضرایب ناصفر به شیوهٔ خاصی در A ظاهر می‌شوند، و معمولاً فرمول ساده‌ای برای تولید ضرایب a_{ij}، در صورت لزوم، وجود دارد تا مجبور به ذخیره‌کردن آنها نباشیم. به‌عنوان یک نتیجهٔ این ویژگیها، توجه به فضای حافظه برای بردارهای x و b ممکن است بسیار مهمتر از رعایت فضای حافظه برای A باشد. ماتریسهای A اغلب دارای ویژگیهای خاصی هستند، که در این بخش و دو بخش دیگر مورد بحث قرار خواهند گرفت.

موضوع را با تعریف و تحلیل دو روش بارستی کلاسیک آغاز می‌کنیم؛ به دنبال آن، یک چارچوب مجرد کلی برای مطالعهٔ روشهای بارستی ارائه می‌کنیم. ویژگیهای خاص دستگاه خطی $Ax = b$، هنگام تدوین روش بارستی برای حل آن، بسیار مهم‌اند. نتایج این بخش فقط آغاز طرح یک روش برای هر زمینهٔ خاص کاربردی خواهند بود.

روش گاوس ـ ژاکوبی (جایگزینی همزمان) دستگاه $Ax = b$ را به شکل زیر بازنویسی می‌نماییم

$$x_i = \frac{1}{a_{ii}} \left\{ b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j \right\} \qquad i = 1, 2, \ldots, n \qquad (۱.۶.۸)$$

با فرض $a_{ii} \neq 0$. بارست را این طور تعریف می‌نماییم

$$x_i^{(m+1)} = \frac{1}{a_{ii}} \left\{ b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(m)} \right\} \qquad i = 1, , \ldots, n \qquad m \geq 0 \qquad (۲.۶.۸)$$

و فرض می‌کنیم که حدسهای اوّلیهٔ $x_i^{(0)}$، $i = 1, \ldots, n$، داده شده‌اند. صورتهای دیگری از این روش وجود دارند. برای مثال، بسیاری از مسائل طبیعتاً به‌صورت

$$(I - B)x = b$$

داده شده‌اند، بنابراین معمولاً ابتدا بارستِ

$$x^{(m+1)} = b + Bx^{(m)} \qquad m \geq 0 \qquad (۳.۶.۸)$$

را در نظر می‌گیریم. اولین تحلیل خطا به (۲.۶.۸) محدود می‌شود، ولی همین تحلیل را می‌توان برای (۳.۶.۸) به‌کار برد.

برای تحلیل همگرایی گیریم $e^{(m)} = x - x^{(m)}$، $m \geq 0$. با کم‌کردن (۲.۶.۸) از (۱.۶.۸) داریم:

$$e_i^{(m+1)} = -\sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{a_{ij}}{a_{ii}} e_j^{(m)} \qquad i = 1, \ldots, n \qquad m \geq 0 \qquad (۴.۶.۸)$$

$$| e_i^{(m+1)} | \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \| e^{(m)} \|_\infty$$

تعریف می‌کنیم

$$\mu = \operatorname*{Max}_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \qquad (۵.۶.۸)$$

پس

$$| e_i^{(m+1)} | \leq \mu \| e^{(m)} \|_\infty$$

و چون سمت راست مستقل از i است،

$$\| e^{(m+1)} \|_\infty \leq \mu \| e^{(m)} \|_\infty \qquad (۶.۶.۸)$$

اگر $\mu < 1$، وقتی m با یک نرخ خطی که با μ کراندار می‌شود به بینهایت میل کند، $e^{(m)} \to 0$ و

$$\| e^{(m)} \|_\infty \leq \mu^m \| e^{(0)} \|_\infty \qquad (۷.۶.۸)$$

برای آنکه $\mu < 1$ درست باشد، ماتریس A باید **غالب قطری** باشد، یعنی، باید در شرط زیر صدق کند.

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} | < | a_{ii} | \qquad i = 1, 2, \ldots, n \qquad (۸.۶.۸)$$

چنین ماتریسهایی در کاربردهای متعددی پیدا می‌شوند و اغلب ماتریسِ وابستهٔ آنها تُنک است.

برای یک نتیجهٔ کلّیتر، (۴.۶.۸) را به شکل

$$e^{(m+1)} = M e^{(m)} \qquad m \geq 0 \qquad (۹.۶.۸)$$

می‌نویسیم.

$$M = - \begin{bmatrix} 0 & \frac{a_{12}}{a_{11}} & \cdots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & 0 & \frac{a_{23}}{a_{22}} & \cdots & \frac{a_{2n}}{a_{22}} \\ \vdots & & \ddots & & \vdots \\ \frac{a_{n1}}{a_{nn}} & & \cdots & & 0 \end{bmatrix}$$

به استقرا

$$e^{(m)} = M^m e^{(0)} \qquad (۱۰.۶.۸)$$

اگر بخواهیم وقتی $m \to \infty$، $e^{(m)}$، مستقل از انتخاب $x^{(0)}$ (و بنابراین مستقل از $e^{(0)}$) به 0 میل کند، لازم و کافی است که

$$M^m \to 0 \qquad \text{وقتی } m \to \infty$$

یا هم‌ارز با آن، بنابر قضیهٔ ۹.۷ از فصل ۷،

$$r_\sigma(M) < ۱ \qquad (۱۱.۶.۸)$$

شرط $\mu < ۱$ تنها شرط کوچکتر از یک بودن نرم سطری M، $\| M \|_\infty < ۱$، است و این (۱۱.۶.۸) را ایجاب می‌کند. ولی اکنون می‌بینیم که اگر برای هر عملگر نرم ماتریسی، $\| M \| < ۱$، آنگاه $e^{(m)} \to ۰$.

مثال حل $Ax = b$ را به روش گاوس ـ ژاکوبی با ماتریسهای زیر در نظر می‌گیریم

$$A = \begin{bmatrix} ۱۰ & ۳ & ۱ \\ ۲ & -۱۰ & ۳ \\ ۱ & ۳ & ۱۰ \end{bmatrix} \quad b = \begin{bmatrix} ۱۴ \\ -۵ \\ ۱۴ \end{bmatrix} \quad x^{(۰)} = \begin{bmatrix} ۰ \\ ۰ \\ ۰ \end{bmatrix} \qquad (۱۲.۶.۸)$$

اگر معادلهٔ iام را نسبت به x_i حل کنیم، داریم $x = g + Mx$،

$$M = \begin{bmatrix} ۰ & -۰ر۳ & -۰ر۱ \\ ۰ر۲ & ۰ & ۰ر۳ \\ -۰ر۱ & -۰ر۳ & ۰ \end{bmatrix} \quad g = \begin{bmatrix} ۱ر۴ \\ ۰ر۵ \\ ۱ر۴ \end{bmatrix}$$

جواب درست $x = [۱, ۱, ۱]^T$ است. برای وارسی همگرایی، توجه داریم که $\| M \|_\infty = ۰ر۵$ و $\| M \|_۱ = ۰ر۶$. بنابراین

$$\| e^{(m+۱)} \|_\infty \leq ۰ر۵ \| e^{(m)} \|_\infty \qquad m \geq ۰ \qquad (۱۳.۶.۸)$$

حکم مشابهی برای $\| e^{(m+۱)} \|_۱$ برقرار است. بنابراین همگرایی تضمین می‌شود و خطاها در هر بار ست اقلاً با ضریب $\frac{۱}{۲}$ کاهش می‌یابند. نتایج عددی واقعی در جدول ۳.۸ داده شده‌اند و مؤیّد نتیجهٔ (۱۳.۶.۸) هستند. ستون آخر عبارت است از:

$$\text{Ratio} \equiv \frac{\| e^{(m)} \|_\infty}{\| e^{(m-۱)} \|_\infty} \qquad (۱۴.۶.۸)$$

این رابطه نشان می‌دهد که ممکن است همگرایی از یک مرحله به مرحلهٔ دیگر تغییر کند در حالی که در (۱۳.۶.۸) یا به‌طور کلیتر در (۶.۶.۸) صدق می‌کند.

جدول ۳.۸ نتایج عددی برای روش گاوس ـ ژاکوبی

m	$x_1^{(m)}$	$x_2^{(m)}$	$x_3^{(m)}$	$\| e^{(m)} \|_\infty$	نسبت
۰	۰	۰	۰	۱٫۰	
۱	۱٫۴	۰٫۵	۱٫۴	۰٫۵	۰٫۵
۲	۱٫۱۱	۱٫۲۰	۱٫۱۱	۰٫۲	۰٫۴
۳	۰٫۹۲۹	۱٫۰۵۵	۰٫۹۲۹	۰٫۰۷۱	۰٫۳۶
۴	۰٫۹۹۰۶	۰٫۹۶۴۵	۰٫۹۹۰۶	۰٫۰۳۵۵	۰٫۵۰
۵	۱٫۰۱۱۵۹	۰٫۹۹۵۳	۱٫۰۱۱۵۹	۰٫۰۱۱۵۹	۰٫۳۳
۶	۱٫۰۰۰۲۵۱	۱٫۰۰۵۷۹۵	۱٫۰۰۰۲۵۱	۰٫۰۰۵۷۹۵	۰٫۵۰

روش گاوس ـ زایدل[1] (جایگزینی پی‌درپی) با استفاده از (۱.۶.۸)، تعریف می‌کنیم

$$x_i^{(m+1)} = \frac{1}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(m)} \right\} \qquad i = 1, 2, \ldots, n \tag{۱۵.۶.۸}$$

هر مؤلفهٔ جدید $x_i^{(m+1)}$ بلافاصله در محاسبهٔ مولفهٔ بعدی به‌کارگرفته می‌شود. این امر برای محاسبهٔ رایانه‌یی بسیار مناسب است، زیرا مقدار جدید بلافاصله در جایی که مقدار قدیم قرار داشته ذخیره می‌شود، و این کار جای لازم برای ذخیره‌سازی را مینیمم می‌نماید. روش گاوس ـ زایدل در مقایسه با روش گاوس ـ ژاکوبی، به نصف جا برای ذخیرهٔ x نیاز دارد.

برای تحلیل خطا، (۱۵.۶.۸) را از (۱.۶.۸) کم می‌کنیم.

$$e_i^{(m+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} e_j^{(m+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} e_j^{(m)} \qquad i = 1, 2, \ldots, n \tag{۱۶.۶.۸}$$

تعریف می‌کنیم

$$\alpha_i = \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right|, \qquad \beta_i = \sum_{j=i+1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \qquad i = 1, \ldots, n$$

با $\alpha_1 = \beta_n = 0$. از همان تعریف (۵.۶.۸) برای μ که برای روش ژاکوبی به‌کار رفت استفاده می‌کنیم،

$$\mu = \operatorname*{Max}_{1 \le i \le n} (\alpha_i + \beta_i)$$

فرض می‌کنیم $\mu < 1$. سپس تعریف می‌کنیم

$$\eta = \operatorname*{Max}_{1 \le i \le n} \frac{\beta_i}{1 - \alpha_i} \tag{۱۷.۶.۸}$$

1. Seidel

از (۱۶.۶.۸)،

$$| e_i^{(m+1)} | \leq \alpha_i \| e^{(m+1)} \|_\infty + \beta_i \| e^{(m)} \|_\infty \qquad i = 1, \ldots, n \quad (۱۸.۶.۸)$$

گیریم k اندیسی باشد که برای آن

$$\| e^{(m+1)} \|_\infty = | e_k^{(m+1)} |$$

در این صورت با $i = k$ در (۱۸.۶.۸)،

$$\| e^{(m+1)} \|_\infty \leq \alpha_k \| e^{(m+1)} \|_\infty + \beta_k \| e^{(m)} \|_\infty$$

$$\| e^{(m+1)} \|_\infty \leq \frac{\beta_k}{1 - \alpha_k} \| e^{(m)} \|_\infty$$

و بنابراین

$$\| e^{(m+1)} \|_\infty \leq \eta \| e^{(m)} \|_\infty \qquad (۱۹.۶.۸)$$

چون برای هر i،

$$(\alpha_i + \beta_i) - \frac{\beta_i}{1 - \alpha_i} = \frac{\alpha_i[1 - (\alpha_i + \beta_i)]}{1 - \alpha_i} \geq \frac{\alpha_i}{1 - \alpha_i}[1 - \mu] \geq 0$$

داریم

$$\eta \leq \mu < 1 \qquad (۲۰.۶.۸)$$

که ترکیب آن با (۱۹.۶.۸)، همگرایی $e^{(m)} \to 0$ را وقتی $m \to \infty$، نشان می‌دهد. در این روش هم نرخ همگرایی خطی خواهد بود، ولی نسبت به روش ژاکوبی، همگرایی سریعتر است.

مثال دستگاه (۱۲.۶.۸) از مثال قبل را بگیرید و آن را با روش گاوس ـ زایدل حل کنید. با یک محاسبهٔ ساده از (۱۷.۶.۸) و (۱۲.۶.۸)،

$$\eta = 0.4$$

نتایج عددی در جدول ۴.۸ داده شده‌اند. سرعت همگرایی به مراتب بهتر از سرعت همگرایی روش گاوس ـ ژاکوبی مثال قبل است که در جدول ۳.۸ داده شده است. به‌نظر می‌آید که مقادیر نسبت همگرایی حدود ۰.۱۸ است.

چارچوب کلی برای روشهای بارستی برای حل $Ax = b$ یک شکل دوپاره‌شدهٔ A را درنظر می‌گیریم:

$$A = N - P \qquad (۲۱.۶.۸)$$

جدول ۴.۸ نتایج عددی برای روش گاوس ـ زایدل

m	$x_1^{(m)}$	$x_2^{(m)}$	$x_3^{(m)}$	$\| e^{(m)} \|_\infty$	نسبت
۰	۰	۰	۰	۱	
۱	۱٫۴	۰٫۷۸	۱٫۰۲۶	۰٫۴	۰٫۴
۲	۱٫۰۶۳۴۰۰	۱٫۰۲۰۴۸۰	۰٫۹۸۷۵۱۶	۶٫۳۴E − ۲	۰٫۱۶
۳	۰٫۹۹۵۱۰۴	۰٫۹۹۵۲۷۶	۱٫۰۰۱۹۰۷	۴٫۹۰E − ۳	۰٫۰۷۷
۴	۱٫۰۰۱۲۲۷	۱٫۰۰۰۸۱۷	۰٫۹۹۹۶۳۲	۱٫۲۳E − ۳	۰٫۲۵
۵	۰٫۹۹۹۷۹۲	۰٫۹۹۹۸۴۸	۱٫۰۰۰۰۶۶	۲٫۰۸E − ۴	۰٫۱۷
۶	۱٫۰۰۰۰۳۹	۱٫۰۰۰۰۲۸	۰٫۹۹۹۹۸۸	۳٫۹۰E − ۵	۰٫۱۹

و $Ax = b$ را به‌صورت زیر می‌نویسیم

$$Nx = b + Px \qquad (۲۲.۶.۸)$$

ماتریس N به گونه‌ای انتخاب شده است که دستگاه خطی $Nz = f$، برای هر f، «به‌سادگی حلپذیر» باشد. مثلاً N ممکن است قطری، سه قطری یا مثلثی باشد. روش بارستی را با $x^{(۰)}$ داده‌شده به صورت زیر تعریف می‌کنیم

$$Nx^{(m+۱)} = b + Px^{(m)} \qquad m \geq ۰ \qquad (۲۳.۶.۸)$$

مثال ۱. روش ژاکوبی

$$N = \operatorname{diag}[a_{۱۱}, a_{۲۲}, \ldots, a_{nn}] \qquad P = N - A$$

۲. روش گاوس ـ زایدل

$$N = \begin{bmatrix} a_{۱۱} & ۰ & & \cdots & ۰ \\ a_{۲۱} & a_{۲۲} & ۰ & \cdots & ۰ \\ \vdots & & \ddots & & \vdots \\ a_{n۱} & & \cdots & & a_{nn} \end{bmatrix} \qquad (۲۴.۶.۸)$$

برای تحلیل خطا، (۲۳.۶.۸) را از (۲۲.۶.۸) کم می‌کنیم تا به‌دست آوریم

$$Ne^{(m+۱)} = Pe^{(m)}$$

$$e^{(m+۱)} = Me^{(m)} \qquad M = N^{-۱}P \qquad (۲۵.۶.۸)$$

با استقرا

$$e^{(m)} = M^m e^{(0)} \qquad m \geq 0 \qquad (۲۶.۶.۸)$$

برای آنکه برای حدس اولیهٔ دلخواه $x^{(0)}$، (و بنابراین هر مقدار دلخواه $e^{(0)}$) وقتی n به بینهایت میل می‌کند، $e^{(m)} \to 0$، لازم و کافی است که

$$M^m \to 0 \qquad \text{وقتی} \quad m \to \infty$$

یا هم‌ارز با آن از قضیهٔ ۹.۷

$$r_\sigma(M) < 1 \qquad (۲۷.۶.۸)$$

این چارچوب کلی برای روشهای بارستی، از آیزکسون و کلر (۱۹۶۶، صص ۶۱ـ۸۱) اقتباس شده است.

شرط (۲۷.۶.۸) قبلاً در (۱۱.۶.۸) برای روش گاوس ـ ژاکوبی به‌دست آمده بود. در روش گاوس ـ زایدل کار با ماتریس $N^{-1}P$ که N با (۲۴.۶.۸) داده شده، دشوارتر است. باید مقادیری از λ را به‌دست آوریم که برای آن

$$\det(\lambda I - N^{-1}P) = 0$$

یا هم‌ارز آن

$$\det(\lambda N - P) = 0 \qquad (۲۸.۶.۸)$$

برای استفاده در حل عددی معادلات دیفرانسیل جزئی در بخش ۸.۸، تحلیل همگرایی بالا با روش گاوس ـ زایدل مناسب نیست. ثابتهای μ و η در (۵.۶.۸) و (۱۷.۶.۸) هر دو مساوی ۱ می‌شوند، با اینکه به‌طور تجربی، این روش باز هم همگراست. برای پرداختن به بسیاری از این دستگاهها، اغلب قضیه مهم زیر را به‌کار می‌برند.

قضیهٔ ۷.۸ گیریم A ارمیتی با عناصر قطری مثبت باشد. در این صورت روش گاوس ـ زایدل برای حل $Ax = b$، برای هر انتخاب $x^{(0)}$ همگرا می‌شود، اگر و تنها اگر A معین مثبت باشد.

برهان اثبات در آیزکسون و کلر (۱۹۶۶، صص ۷۰ـ۷۱) داده شده است. برای تعریف ماتریس معین مثبت، مسألهٔ ۱۴ فصل ۷ را به یاد آورید. این قضیه در بخش ۸.۸ توضیح داده شده است. ■

روشهای دیگر بارستی بهترین روشهای بارستی روشهایی هستند که بر پایهٔ یک اطلاّع کامل

از مسأله‌ای که باید حل شود، جنبه‌های خاص آن در طرح برنامهٔ بارستی به‌حساب آیند. این کار معمولاً با نگاه‌کردن به شکل ماتریس و منشاء دستگاه خطی صورت می‌گیرد.

ماتریس A ممکن است شکل خاصی داشته باشد که به روش بارستی ساده‌ای منجر شود. به‌عنوان مثال، فرض کنید A یک شکل سه قطری بلوکی دارد:

$$A = \begin{bmatrix} B_1 & C_1 & & & \cdots & 0 \\ A_2 & B_2 & C_2 & & & \\ \vdots & & & \ddots & & \vdots \\ 0 & & \cdots & & A_r & B_r \end{bmatrix} \qquad (۲۹.۶.۸)$$

ماتریسهای A_i، B_i و C_i مربعی از مرتبهٔ m و A از مرتبهٔ $n = rm$ است. به ازای $x, b \in \mathbf{R}^n$، بردارهای x و b را به شکل افراز شده مثل:

$$x = \begin{bmatrix} x_{(1)} \\ \vdots \\ x_{(r)} \end{bmatrix} \qquad b = \begin{bmatrix} b_{(1)} \\ \vdots \\ b_{(r)} \end{bmatrix} \qquad x_{(i)}, b_{(i)} \in \mathbf{R}^m$$

می‌نویسیم. در این صورت $Ax = b$ به شکل زیر نوشته می‌شود

$$\begin{aligned} B_1 x_{(1)} + C_1 x_{(2)} &= b_{(1)} \\ A_i x_{(i-1)} + B_i x_{(i)} + C_i x_{(i+1)} &= b_{(i)} \qquad 2 \le i \le r-1 \\ A_r x_{(r-1)} + B_r x_{(r)} &= b_{(r)} \end{aligned} \qquad (۳۰.۶.۸)$$

فرض می‌کنیم که دستگاههای خطی

$$B_j x_{(j)} = d_{(j)} \qquad 1 \le j \le r \qquad (۳۱.۶.۸)$$

به‌سادگی، شاید مستقیماً برای تمام $d_{(j)}$ های طرفهای راست حلپذیر باشند. مثلاً اغلب همهٔ B_i ها یک ماتریس ثابت سه‌قطری T هستند که برای آنها شیوهٔ به‌کاررفته در (۲۰.۳.۸) ـ (۲۴.۳.۸) را می‌توان به‌کار برد.

می‌توان روشی از نوع روش ژاکوبی برای (۳۰.۶.۸) به‌کار برد

$$\begin{aligned} B_1 x_{(1)}^{(v+1)} &= b_{(1)} - C_1 x_{(2)}^{(v)} \\ B_i x_{(i)}^{(v+1)} &= b_{(i)} - A_i x_{(i-1)}^{(v)} - C_i x_{(i+1)}^{(v)} \qquad 2 \le i \le r-1 \\ B_r x_{(r)}^{(v+1)} &= b_{(r)} - A_r x_{(r-1)}^{(v)} \end{aligned} \qquad (۳۲.۶.۸)$$

تحلیل همگرایی پیچیده‌تر از تحلیل همگرایی در روشهای گاوس ـ ژاکوبی و گاوس ـ زایدل است؛ به بعضی از نتایج در مسألهٔ ۲۹ اشاره شده است. روشهای مشابهی برای دستگاههای خطی که از حل بعضی معادلات دیفرانسیل جزئی به‌دست می‌آیند، به‌کار برده شده‌اند، که در بخش ۸.۸ به آنها اشاره شده است.

یک جنبهٔ مهم دیگر حل دستگاههای خطی $Ax = b$، نگریستن به منشاء پیدایش آنهاست. در بسیاری از موارد یک معادلهٔ دیفرانسیل یا انتگرالی مثل

$$\mathscr{A}x = y \qquad (۳۳.۶.۸)$$

داریم که x و y تابع‌اند. اگر این معادله را به صورت گسسته بنویسیم، یک خانوادهٔ مسائل

$$A_n x_n = y_n \qquad x_n, y_n \in \mathbf{R}^n \qquad (۳۴.۶.۸)$$

با A_n از مرتبهٔ n حاصل می‌شود. وقتی $n \to \infty$، جوابهای x_n دستگاه (۳۴.۶.۸)، (به تعبیری) به جواب x (۳۳.۶.۸) میل می‌کند. بنابراین دستگاههای خطی در (۳۴.۶.۸) بستگی نزدیکی با هم دارند. به‌عنوان مثال، به تعبیری، برای m و n به اندازهٔ کافی بزرگ $A_m^{-1} \doteq A_n^{-1}$، اگرچه مرتبهٔ ماتریسها مختلف‌اند. به این مطلب می‌توان معنای دقیقتری داد که ما را به روشهای بارستی دیگری هدایت می‌کند که در آنجا، حل دستگاههای بزرگ خطی با حل دستگاههای کوچکتر امکان‌پذیر می‌شود. اخیراً، چنین روشهایی به نام **روشهای چندشبکه‌یی** به‌وجود آمده‌اند که کاربردهایی به‌ویژه در حل معادلات دیفرانسیل جزئی دارند [هاکبوش[1] و تروتنبرگ[2] (۱۹۸۲) را ببینید]. برای روشهای بارستی در حل معادلات انتگرالی، پیشرفتهای مربوطه ولی متفاوت را در اتکینسن (۱۹۷۶، بخش ۲ فصل ۴) را ببینید. روشهای چندشبکه‌یی، روشهای بارستیِ بسیار مؤثر و کارا برای معادلات دیفرانسیل و انتگرالی هستند.

۷.۸ پیشگویی خطا و شتاب

با توجه به (۲۵.۶.۸)، رابطهٔ خطای زیر را داریم

$$x - x^{(m+1)} = M(x - x^{(m)}) \qquad m \geq 0 \qquad (۱.۷.۸)$$

نحوهٔ همگرایی $x^{(m)}$ به x، بسته به ویژه‌مقدارها و ویژه‌بردارهای M، ممکن است کاملاً پیچیده باشد. ولی در بیشتر حالات عملی، رفتار خطا خیلی ساده است: اندازهٔ $\| x - x^{(m)} \|_\infty$ در هر مرحله با یک عامل تقریباً ثابت کاهش می‌یابد، و

$$\| x - x^{(m+1)} \|_\infty \leq c \| x - x^{(m)} \|_\infty \qquad (۲.۷.۸)$$

1. Hackbusch 2. Trottenberg

که در آن $c < ۱$ و خیلی به $r_\sigma(M)$ وابسته است. برای اندازه‌گیری این ثابت c، با توجه به (۱.۷.۸) داریم

$$x^{(m+۱)} - x^{(m)} = e^{(m)} - e^{(m+۱)} = Me^{(m-۱)} - Me^{(m)}$$

$$x^{(m+۱)} - x^{(m)} = M(x^{(m)} - x^{(m-۱)}) \qquad m \geq ۰ \tag{۳.۷.۸}$$

که موجب می‌شود از

$$c \doteq \frac{\| x^{(m+۱)} - x^{(m)} \|_\infty}{\| x^{(m)} - x^{(m-۱)} \|_\infty} \tag{۴.۷.۸}$$

یا برای اطمینان بیشتر، از ماکسیمم چندین نسبت پی‌درپی از نوع فوق استفاده کنیم. در بسیاری از کاربردها، این نسبت برای مقادیر بزرگ m ، تقریباً ثابت است.

وقتی این ثابت c به‌دست آمد، و با فرض (۲.۷.۸)، می‌توانیم خطا در $x^{(m+۱)}$ را با استفاده از (۱۳.۵.۸) کراندار نماییم.

$$\| x - x^{(m+۱)} \|_\infty \leq \frac{c}{۱ - c} \| x^{(m+۱)} - x^{(m)} \|_\infty \tag{۵.۷.۸}$$

این کران وقتی $c \doteq ۱$، مسأله‌ساز می‌شود و همگرایی کُند خواهد بود. در این حالت، تفاضل $\| x^{(m+۱)} - x^{(m)} \|_\infty$ ممکن است از خطای واقعی $\| x - x^{(m+۱)} \|_\infty$ خیلی کوچکتر باشد.

مثال دستگاه خطی (۵.۸.۸) از بخش ۸.۸ با استفاده از روش گاوس ـ زایدل حل شده بود. برای هم‌آهنگی با (۵.۸.۸)، بردار مجهول را u گرفته‌ایم. در (۴.۸.۸)، تابع $f = x^۲y^۲$، و در (۵.۸.۸) تابع $g = ۲(x^۲ + y^۲)$ است. ناحیهٔ جواب $۰ \leq x, y \leq ۱$ و اندازهٔ شبکه در هر جهت $h = \frac{۱}{۱۶}$ است. این مفروضات، یک مرتبهٔ ۲۲۵ برای دستگاه خطی (۵.۸.۸) به‌دست می‌داد. حدس اولیهٔ $u^{(۰)}$ در این بارست، بر پایهٔ درونیابی **دوخطی** $f = x^۲y^۲$ بر ناحیهٔ $۰ \leq x, y \leq ۱$ نهاده شده است. [(۱۷.۸.۸) را ببینید]. نمونه‌ای از نتایج عددی در جدول ۵.۸ داده شده است. ستون نسبت از روی (۴.۷.۸) محاسبه شده و ستون براورد خطا از روی (۵.۷.۸)؛ و ستون خطا، خطای درست $\| u - u^{(m)} \|_\infty$ است.

همان‌گونه که از جدول می‌توان دید، همگرایی کاملاً آهسته است که نیاز به (۵.۷.۸) را به جای مقدار بسیار بزرگتر $\| u^{(m)} - u^{(m-۱)} \|_\infty$ توجیه می‌نماید. وقتی $m \to \infty$، مقدار نسبت به ۰٫۹۶۲ همگرا می‌شود، و براورد خطای (۵.۷.۸) یک براوردگر دقیق برای خطای درست بارستی است.

سرعت همگرایی اکنون ببینیم که چند بارست باید حساب کنیم تا به خطای مطلوب برسیم. و چه موقع روش بارستی در حل $Ax = b$ بر روش حذف گاوسی برتری دارد؟ مقداری از m

جدول ۵.۸ مثالی از بارست گاوس ـ زایدل

m	$\| u^{(m)} - u^{(m-1)} \|_\infty$	نسبت	براورد خطا	خطا
۲۰	۱٫۲۰E − ۳	۰٫۹۶۶	۳٫۴۲E − ۲	۳٫۰۹E − ۲
۲۱	۱٫۱۶E − ۳	۰٫۹۶۶	۳٫۲۴E − ۲	۲٫۹۸E − ۲
۲۲	۱٫۱۲E − ۳	۰٫۹۶۵	۳٫۰۸E − ۲	۲٫۸۶E − ۲
۲۳	۱٫۰۸E − ۳	۰٫۹۶۵	۲٫۹۳E − ۲	۲٫۷۶E − ۲
۲۴	۱٫۰۴E − ۳	۰٫۹۶۴	۲٫۸۰E − ۲	۲٫۶۵E − ۲
۶۰	۲٫۶۰E − ۴	۰٫۹۶۲	۶٫۵۸E − ۳	۶٫۵۸E − ۳
۶۱	۲٫۵۰E − ۴	۰٫۹۶۲	۶٫۳۳E − ۳	۶٫۳۳E − ۳
۶۲	۲٫۴۱E − ۴	۰٫۹۶۲	۶٫۰۹E − ۳	۶٫۰۹E − ۳

را پیدا می‌کنیم که برای آن

$$\| x - x^{(m)} \|_\infty \leq \varepsilon \| x - x^{(0)} \|_\infty \qquad (۶.۷.۸)$$

که ε ضریب داده‌شده‌ای است که خطای اولیه باید با آن کاهش یابد. تحلیل را بر پایهٔ فرض (۲.۷.۸) می‌گذاریم. معمولاً ثابت c تقریباً با $r_\sigma(M)$ برابر است که M همان است که در (۱.۷.۸) داده شده است.

از رابطهٔ (۲.۷.۸) چنین نتیجه می‌شود

$$\| x - x^{(m)} \|_\infty \leq c^m \| x - x^{(0)} \|_\infty \qquad m \geq 0 \qquad (۷.۷.۸)$$

بنابراین کوچکترین مقدار m را که برای آن

$$c^m \leq \varepsilon$$

می‌یابیم. از حل این رابطه به‌دست می‌آید

$$m \geq \frac{-\ln \varepsilon}{R(c)} = m^* \qquad R(c) = -\ln c \qquad (۸.۷.۸)$$

دو برابرکردن $R(c)$ منجر به نصف‌کردن تعداد بارستهایی خواهد شد که باید محاسبه شود.

برای آنکه به این نتیجه معنی بیشتری بدهیم، آن را در حل یک دستگاه خطی چگال به روش بارستی به‌کار می‌بریم. فرض می‌کنیم که روش گاوس ـ ژاکوبی یا گاوس ـ زایدل برای حل

جدول ۶.۸ مثال از شمارش بارستی

c	$R(c)$	m^*
۰٫۹	۰٫۱۰۵	۱۳۱
۰٫۸	۰٫۲۲۳	۶۲
۰٫۶	۰٫۵۱۱	۲۷
۰٫۴	۰٫۹۱۶	۱۵
۰٫۲	۱٫۶۱	۹

$Ax = b$، با دقت ساده در رایانهٔ بزرگ IBM، یعنی، در حدود ۶ رقم با معنی، به‌کار برده شده باشد. فرض می‌کنیم $x^{(0)} = 0$، و می‌خواهیم m را طوری پیدا کنیم که

$$\frac{\| x - x^{(m)} \|_\infty}{\| x \|_\infty} \leq 10^{-6} = \varepsilon \qquad (۹.۷.۸)$$

فرض می‌کنیم که A از مرتبهٔ n باشد، تعداد عملها (ضرب و تقسیم) در هر بارست n^2 است. برای به‌دست آوردن نتیجهٔ (۹.۷.۸)، تعداد بارستنهای لازم عبارت است از

$$m^* = \frac{6 \ln_e 10}{R(c)}$$

و تعداد عملها چنین است

$$m^* n^2 = (6 \ln_e 10) \frac{n^2}{R(c)}$$

اگر حذف گاوسی برای حل $Ax = b$ با همان دقت به‌کار برده شده باشد، تعداد عملها در حدود $\frac{n^3}{3}$ است. روش بارستی از روش حذف گاوسی کاراتر خواهد بود اگر

$$m^* n^2 < \frac{n^3}{3}$$
$$m^* < \frac{n}{3} \qquad (۱۰.۷.۸)$$

مثال ماتریس A از مرتبهٔ $n = 51$ را در نظر می‌گیریم. پس روش بارستی کاراتر است اگر $m^* < 17$. جدول ۶.۸ مقادیر m^* را به ازای مقادیر مختلف c به‌دست می‌دهد. برای $c \leq 0.44$ ، روش بارستی از روش حذف گاوسی کاراتر خواهد بود. و اگر دقتی کمتر از درجهٔ دقت کامل مورد نظر باشد، روش بارستی حتی برای cهای بزرگتر نیز کاراتر است. در عمل، معمولاً از یک حدس اولیهٔ $x^{(0)}$ نیز آگاهی داریم که از $x^{(0)} = 0$ بهتر است، که تعداد بارستنهای لازم را بازهم کاهش خواهد داد.

استفادهٔ عمده از روشهای بارستی در حل دستگاههای بزرگِ تُنک است، که در آنها حالت حذف گاوسی اغلب ناممکن است. و حتّی وقتی که حذف گاوسی هم ممکن باشد، روش بارستی باز ممکن است ارجح باشد. چند مثال از این دستگاهها در بخش ۸.۸ داده شده است.

روشهای شتابی اکثر روشهای بارستی الگوی منظّمی دارند که در آنها خطا کاهش می‌یابد. این الگو را اغلب می‌توان برای تسریع همگرایی به‌کار برد، درست همان‌گونه که در فصلهای پیشین برای روشهای عددی دیگر انجام گرفته است. به‌جای آنکه یک نظریهٔ کلی برای شتاب روشهای بارستی در حل $Ax = b$ بیاوریم، فقط شتاب برای روش گاوس ـ زایدل را شرح خواهیم داد، که یکی از حالتهای عمدهٔ مطلوب در کاربردهاست.

تعریف (۱۵.۶.۸) در روش گاوس ـ زایدل را به یاد می‌آوریم. یک **پارامتر شتاب** ω معرفی می‌کنیم، و (۱۵.۶.۸) را با یک تغییر جزیی در نظر می‌گیریم:

$$z_i^{(m+1)} = \frac{1}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(m)} \right\}$$

$$x_i^{(m+1)} = \omega z_i^{(m+1)} + (1 - w) x_i^{(m)} \qquad i = 1, \ldots, n \qquad m \geq 0 \qquad (۱۱.۷.۸)$$

حالت $m = 1$ همان روش گاوس ـ زایدل معمولی است. منظور از شتاب‌دادن، انتخاب بهینهٔ یک ترکیب خطی از بارست قبلی و بارست گاوس ـ زایدل معمولی است. روش (۱۱.۷.۸)، با انتخاب بهینهٔ ω روش ***SOR*** خوانده می‌شود که مخفّف اصطلاح تاریخی ***successive over relaxation*** است.

برای اینکه بفهمیم چگونه ω باید انتخاب شود، (۱۱.۷.۸) را به شکل ماتریسی می‌نویسیم. A را به شکل زیر تجزیه می‌کنیم

$$A = D + L + U$$

که $D = dig[a_{11}, \ldots, a_{nn}]$، L ماتریس پایین مثلثی، U ماتریس بالامثلثی است و قطرهای اصلی L و U هر دو صفرند. در این صورت (۱۱.۷.۸) چنین خواهد شد:

$$z^{(m+1)} = D^{-1}[b - Lx^{(m+1)} - Ux^{(m)}]$$

$$x^{(m+1)} = \omega z^{(m+1)} + (1 - \omega) x^{(m)} \qquad m \geq 0$$

از حذف $z^{(m+1)}$ و حل نسبت به $x^{(m+1)}$ داریم

$$[I + \omega D^{-1} L] x^{(m+1)} = \omega D^{-1} b + [(1 - \omega) I - \omega D^{-1} U] x^{(m)}$$

برای خطا،

$$e^{(m+1)} = M(\omega)e^{(m)} \qquad m \geq 0 \tag{۱۲.۷.۸}$$

$$M(\omega) = [I + \omega D^{-1}L]^{-1}[(1-\omega)I - \omega D^{-1}U] \tag{۱۳.۷.۸}$$

پارامتر ω باید طوری انتخاب شود که $r_\sigma(M(\omega))$ را مینیمم کند، تا اینکه $x^{(m)}$ هرچه سریعتر به x همگرا شود. این مقدار بهینه را ω^* می‌نامیم.

محاسبهٔ ω^* جز در ساده‌ترین حالتها دشوار است، و معمولاً فقط به‌طور تقریبی، با انتخاب چند مقدار برای ω و توجه به اثر آن روی سرعت همگرایی به‌دست می‌آید. علی‌رغم مسألهٔ محاسبهٔ ω^*، افزایش به‌دست آمده در سرعت همگرایی $x^{(m)}$ به x، بسیار چشمگیر است و سعی در محاسبهٔ ω^* کاملاً ارزش دارد. این موضوع در بخش بعد نشان داده شده است.

مثال ما شتاب (۱۱.۷.۸) در مثال قبلی را برای روش گاوس ـ زایدل که به دنبال (۵.۷.۸) آمده به‌کار می‌بریم. مقدار پارامتر بهینهٔ شتاب $\omega^* \doteq$ ۱٫۶۷۳۵ است. یک بحث کامل از روش SOR برای حل دستگاه خطی که در حل معادلات دیفرانسیل جزیی پیش می‌آید در بخش بعدی ارائه شده است. حدس اولیه مانند حدس قبل بوده است. نتایج در جدول ۷.۸ داده شده‌اند.

نتایج نشان می‌دهند که نرخ همگرایی بسیار سریعتر از نرخ همگرایی در روش گاوس ـ زایدل است. به‌عنوان مثال، با روش گاوس ـ زایدل، داریم $\| u - u^{(۲۲۸)} \|_\infty =$ ۹٫۷۰E − ۶ . در روش SOR داریم، $\| u - u^{(۳۲)} \|_\infty =$ ۸٫۷۱E − ۶ . ولی، همان‌طور که می‌توان از مقادیر نسبت ملاحظه کرد، رفتار منظم همگرایی بارستنها را از دست داده‌ایم. مقدار c که در آزمون خطای (۵.۷.۸) به‌کار برده شده باید با دقّتی بیشتر از نسبت c که در جدول به‌کار برده شده، انتخاب شود. می‌توانید از یک میانگین یا ماکسیمم چندین مقدار پی‌درپی نسبت استفاده نمایید.

۸.۸ حل عددی معادلهٔ پواسن

مهمترین کاربرد روشهای بارستی خطی، در دستگاههای خطی بزرگی است که از حل عددی معادلات دیفرانسیل جزئی به روش تفاضلات متناهی، به‌وجود می‌آیند. برای نشان‌دادن این موضوع، مسألهٔ دیریکله را برای معادلهٔ پواسن بر یک مربع واحد در صفحهٔ xy حل می‌کنیم:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = g(x,y) \qquad 0 < x, y < 1 \tag{۱.۸.۸}$$

$$u(x,y) = f(x,y) \qquad (x,y) \text{ یک نقطهٔ مرزی}$$

جدول ۷.۸ مثال روش SOR (۱۱.۷.۸)

m	$\| u^{(m)} - u^{(m-1)} \|_\infty$	نسبت	براورد خطا	خطا
۲۱	۲٫۰۶E − ۴	۰٫۶۹۳	۴٫۶۵E − ۴	۳٫۶۴E − ۴
۲۲	۱٫۳۵E − ۴	۰٫۶۵۷	۲٫۵۹E − ۴	۲٫۶۵E − ۴
۲۳	۸٫۷۶E − ۵	۰٫۶۴۸	۱٫۶۱E − ۴	۱٫۸۷E − ۴
۲۴	۵٫۱۱E − ۵	۰٫۵۸۴	۷٫۱۷E − ۵	۱٫۳۹E − ۴
۲۵	۳٫۴۸E − ۵	۰٫۶۸۰	۷٫۴۰E − ۵	۱٫۰۶E − ۴
۲۶	۲٫۷۸E − ۵	۰٫۸۰۰	۱٫۱۱E − ۴	۸٫۰۴E − ۵
۲۷	۲٫۴۶E − ۵	۰٫۸۸۴	۱٫۸۷E − ۴	۶٫۱۵E − ۵
۲۸	۲٫۰۷E − ۵	۰٫۸۴۲	۱٫۱۱E − ۴	۴٫۱۶E − ۵

توابع $g(x,y)$ و $f(x,y)$ داده شده‌اند و باید $u(x,y)$ را پیدا کنیم.

برای $N > 1$ تعریف می‌کنیم $h = 1/N$، و

$$(x_j, y_k) = (jh, kh) \qquad 0 \le j,k \le N$$

این نقاط را نقاط گرهی یا نقاط شبکه‌یی خوانند (شکل ۱.۸ را ببینید). برای تقریب‌زدن (۱.۸.۸) تقریبهایی را برای مشتقات دوّم به‌کار می‌بریم. برای یک تابع چهار بار پیوسته مشتقپذیر $G(x)$ در $[x-h, x+h]$، نتایج (۱۷.۷.۵) و (۱۸.۷.۵) از بخش (۷.۵) چنین به‌دست می‌دهند،

$$G''(x) = \frac{G(x+h) - 2G(x) + G(x-h)}{h^2} - \frac{h^2}{12} G^{(4)}(\xi) \qquad x-h \le \xi \le x+h \tag{۲.۸.۸}$$

وقتی تقریب فوق را در (۱.۸.۸) در هر نقطهٔ شبکه‌یی داخلی به‌کار بریم، به‌دست می‌آوریم

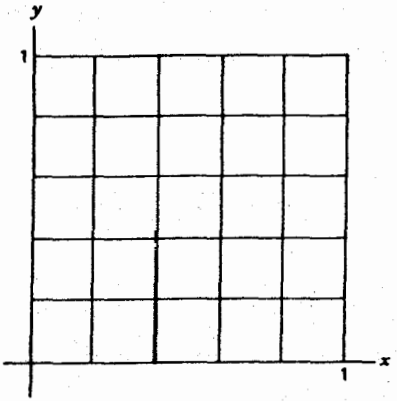
$$\frac{u(x_{j+1},y_k) - 2u(x_j,y_k) + u(x_{j-1},y_k)}{h^2} + \frac{u(x_j,y_{k+1}) - 2u(x_j,y_k) + u(x_j,y_{k-1})}{h^2} = g(x_j,y_k) + \frac{h^2}{12}\left\{ \frac{\partial^4 u(\xi,y_k)}{\partial x^4} + \frac{\partial^4 u(x_j,\eta)}{\partial y^4} \right\} \tag{۳.۸.۸}$$

برای مقداری از $x_{j+1} \le \xi \le x_{j-1}$، $y_{k-1} \le \eta \le y_{k+1}$ و $1 \le j,k \le N-1$.

برای تقریب عددی $u_h(x,y)$ مربوط به (۱.۸.۸)، گیریم

$$u_h(x_j,y_k) = f(x_j,y_k) \qquad (x_j,y_k) \text{ یک نقطهٔ گرهی مرزی} \tag{۴.۸.۸}$$

در همهٔ نقاط شبکه‌یی داخلی، خطاهای برشی طرف راست در (۳.۸.۸) را حذف و مسأله را



شکل ۱.۸ شبکهٔ تفضل متناهی

نسبت به جواب تقریبی $u_h(x_j, y_k)$ حل می‌کنیم:

$$u_h(x_j,y_k)=\frac{۱}{۴}\{u_h(x_{j+۱},y_k)+u_h(x_j,y_{k+۱})+u_h(x_{j-۱},y_k)+u_h(x_j,y_{k-۱})\}$$
$$-\frac{h^۲}{۴}g(x_j,y_k) \qquad ۱\le j,k\le N-۱ \qquad (۵.۸.۸)$$

تعداد معادلات در (۴.۸.۸) ـ (۵.۸.۸) برابر است با تعداد مجهولات، $(N+۱)^۲$.

قضیهٔ ۸.۸ برای هر $N \ge ۲$، دستگاه خطی (۴.۸.۸) ـ (۵.۸.۸) دارای یک جواب یکتای $\{u_h(x_j, y_k) \mid ۰ \le j,k \le N\}$ است. اگر $u(x,y)$، جواب (۱.۸.۸)، چهاربار پیوسته مشتقپذیر باشد، آنگاه

$$\max_{۰\le j,k\le n} \mid u(x_j,y_k)-u_h(x_j,y_k) \mid \le ch^۲ \qquad (۶.۸.۸)$$

$$c=\frac{۱}{۲۴}\left\{\max_{۰\le x,y\le ۱}\left|\frac{\partial^۴u(x,y)}{\partial x^۴}\right|+\max_{۰\le x,y\le ۱}\left|\frac{\partial^۴u(x,y)}{\partial y^۴}\right|\right\}$$

برهان ۱. حلپذیری یکتای (۴.۸.۸) ـ (۵.۸.۸) را با استفاده از قضیهٔ ۲.۷، فصل ۷، ثابت می‌کنیم. دستگاه همگن زیر را در نظر می‌گیریم

$$v_h(x_j,y_k)=\frac{۱}{۴}\,[v_h(x_{j+۱},y_k)+v_h(x_j,y_{k+۱})+\ v_h(x_{j-۱},y_k)$$
$$+v_h(x_j,y_{k-۱})] \qquad ۱\le j,k\le N-۱ \qquad (۷.۸.۸)$$
$$v_h(x_j,y_k)=۰ \qquad (x_j,y_k) \text{ یک نقطه مرزی} \qquad (۸.۸.۸)$$

با نشان‌دادن اینکه این دستگاه فقط جواب نمایان $v_h(x_j, y_k) \equiv 0$ را دارد، از قضیهٔ ۲.۷ نتیجه می‌شود که دستگاه ناهمگن (۴.۸.۸) ـ (۵.۸.۸) فقط یک جواب یکتا دارد.

گیریم

$$\alpha = \underset{0 \le j,k \le N}{\text{Max}} v_h(x_j, y_k)$$

با توجه به (۸.۸.۸)، $\alpha \ge 0$. فرض می‌کنیم $\alpha > 0$. پس یک نقطهٔ گرهی داخلی $(\bar{x}_j, \bar{y}_k)$ باید وجود داشته باشد که در آن این ماکسیمم رخ دهد. ولی با استفاده از (۷.۸.۸)، $v_h(\bar{x}_j, \bar{y}_k)$ میانگین مقادیر v_h در چهار نقطهٔ مجاور $(\bar{x}_j, \bar{y}_k)$ است. تنها راهی که میانگین‌بودن با ماکسیمم‌بودن $(\bar{x}_j, \bar{y}_k)$ سازگار باشد آن است که v_h در چهار نقطهٔ شبکه‌یی مجاور نیز برابر α باشد. همین استدلال را برای نقاط مجاور هم ادامه می‌دهیم. چون فقط تعدادی متناهی نقطهٔ شبکه‌یی وجود دارند، سرانجام برای یک نقطهٔ مرزی (x_j, y_h) داریم $v_h(x_j, y_h) = \alpha$. ولی در این صورت $\alpha > 0$ با شرط (۸.۸.۸) تناقض پیدا می‌کند. بنابراین ماکسیمم $v_h(x_j, y_k)$ برابر صفر است. یک استدلال مشابه نشان می‌دهد که مینیمم $v_h(x_j, y_k)$ نیز برابر صفر است. این نتایج وقتی با هم در نظر گرفته شوند معلوم می‌شود که تنها جواب (۷.۸.۸) ـ (۸.۸.۸)، $v_h(x_j, y_k) \equiv 0$ است.

۲. برای ملاحظهٔ همگرایی $u_h(x_j, y_k)$ به $u(x_j, y_k)$، تعریف می‌کنیم

$$e_h(x_j, y_k) = u(x_j, y_k) - u_h(x_j, y_k)$$

با کم‌کردن (۵.۸.۸) از (۳.۸.۸) به‌دست می‌آوریم

$$e_h(x_j, y_k) = \frac{1}{4}\left[e_h(x_{j+1}, y_k) + e_h(x_j, y_{k+1}) + e_h(x_{j-1}, y_k) + e_h(x_j, y_{k-1})\right] - \frac{h^2}{12}\left[\frac{\partial^4 u(\xi_j, y_k)}{\partial x^4} + \frac{\partial^4 u(x_j, \eta_k)}{\partial y^4}\right] \qquad (۹.۸.۸)$$

و از (۴.۸.۸)،

$$e_h(x_j, y_k) = 0 \qquad (x_j, y_k) \text{ یک نقطهٔ مرزی} \qquad (۱۰.۸.۸)$$

با این دستگاه می‌توان مشابه با قسمت (۱) عمل کرد، نتیجهٔ (۶.۸.۸) به‌دست خواهد آمد. ولی چون این امر برای بحث دستگاههای خطی اساسی نیست، از استدلال آن صرف‌نظر می‌شود [آیزکسون و کلر (۱۹۶۶، صص ۴۴۷ـ۴۵۰) را ببینید]. ■

مثال مسألهٔ زیر را حل کنید

$$\frac{\partial^۲ u}{\partial x^۲}+\frac{\partial^۲ u}{\partial y^۲}=۰ \qquad ۰\le x,y\le ۱ \tag{۱۱.۸.۸}$$
$$u(۰,y)=\cos(\pi y) \qquad u(۱,y)=e^{\pi}\cos(\pi y)$$
$$u(x,۰)=e^{\pi x} \qquad u(x,۱)=-e^{\pi x}$$

جواب درست $u(x,y)=e^{\pi x}\cos(\pi y)$ است.

نتایج عددی برای چندین مقدار N در جدول ۸.۸ داده شده است. خطا، در همهٔ نقاط شبکه ماکسیمم است، و ستون نسبت عامل کاهش خطای ماکسیمم وقتی h نصف شود را می‌دهد. از لحاظ نظری، به موجب (۶.۸.۸)، این عامل باید برابر ۴٫۰ باشد. نتایج عددی این را تایید می‌کند.

راه حلّ بارستی چون روش گاوس ـ زایدل معمولاً سریعتر از روش گاوس ـ ژاکوبی است، ما فقط همان روش اولی را در نظر می‌گیریم. برای $k=۱,۲,\dots,N-۱$ تعریف می‌کنیم

$$u_h^{(m+۱)}(x_j,y_k)=\frac{۱}{۴}[u_h^{(m)}(x_{j+۱},y_k)+u_h^{(m)}(x_j,y_{k+۱})+u_h^{(m+۱)}(x_{j-۱},y_k)$$
$$+u_h^{(m+۱)}(x_j,y_{k-۱})]-\frac{h^۲}{۴}g(x_j,y_k) \qquad j=۱,۲,\dots,N-۱ \tag{۱۲.۸.۸}$$

برای نقاط مرزی، از رابطهٔ زیر استفاده می‌کنیم

$$u_h^{(m)}(x_j,y_k)=f(x_j,y_k) \qquad \text{برای هر } m\ge ۰$$

مقادیر $u_h^{(m+۱)}(x_j,y_k)$ سطر به سطر حساب می‌شوند، ابتدا، از سطر آخر نقاط شبکه تا سطر اول نقاط شبکه. و در هر سطر، از چپ به راست.

جدول ۸.۸ حل عددی (۱۱.۸.۸)

N	$\| u-u_h \|_\infty$	نسبت
۴	۰٫۱۴۴	
۸	۰٫۰۳۹۰	۳٫۷
۱۶	۰٫۰۱۰۲	۳٫۸
۳۲	۰٫۰۰۲۶۰	۳٫۹
۶۴	۰٫۰۰۰۶۵۴	۴٫۰

برای بارست (۱۲.۸.۸)، تحلیل همگرایی باید بر پایهٔ قضیهٔ ۷.۸ صورت گیرد. به آسانی می‌توان نشان داد که این ماتریس متقارن است و بنابراین ویژه‌مقدارها حقیقی‌اند. به‌علاوه می‌توان نشان داد که ویژه‌مقدارها در بازهٔ $0 < \lambda < 2$ قرار دارند. با توجه به این نکته و مسألهٔ ۱۴، فصل ۷، این ماتریس معین مثبت است. چون همهٔ ضرایب قطری ماتریس مثبت‌اند، از قضیهٔ ۷.۸ نتیجه می‌شود که روش گاوس ـ زایدل همگرا خواهد بود. برای نشان‌دادن اینکه ویژه‌مقدارها در بازهٔ $0 < \lambda < 2$ قرار دارند آیزکسون و کلر (۱۹۶۶، صص ۴۴۸ـ۴۵۹) یا مسألهٔ ۳ فصل ۹ را ببینید.

محاسبهٔ سرعت همگرایی $r_\sigma(M)$ از روی (۲۸.۶.۸) آسان نیست. استدلال کاملاً پیچیده است، و فقط خواننده را به بحث کامل آن در آیزکسون و کلر (۱۹۶۶، صص ۴۶۳ـ۴۷۰)، شامل مطالب مربوط به روش شتاب گاوس ـ زایدل ارجاع می‌دهیم. می‌توان نشان داد که

$$r_\sigma(M) = 1 - \pi^2 h^2 + O(h^4) \qquad (۱۳.۸.۸)$$

روش گاوس ـ زایدل همگراست ولی سرعت همگرایی، حتی برای مقدار نسبتاً کوچک h خیلی آهسته است. این مطلب در جدول ۵.۸ بخش ۷.۸ نشان داده شده است.

برای شتاب‌دادن به روش گاوس ـ زایدل از رابطهٔ زیر استفاده می‌کنیم:

$$\begin{aligned} v_h^{(m+1)}(x_j, y_k) &= \frac{1}{4}[u_h^{(m)}(x_{j+1}, y_k) + u_h^{(m)}(x_j, y_{k+1}) \\ &\quad + u_h^{(m+1)}(x_{j-1}, y_k) + u_h^{(m+1)}(x_j, y_{k-1})] - \frac{h^2}{4} g(x_j, y_k) \\ u_h^{(m+1)}(x_j, y_k) &= \omega v_h^{(m+1)}(x_j, y_k) \\ &\quad + (1-\omega) u_h^{(m)}(x_j, y_k) \qquad j = 1, \ldots, N-1 \end{aligned} \qquad (۱۴.۸.۸)$$

پارامتر بهینهٔ شتاب عبارت است از

$$\begin{aligned} \omega^* &= \frac{2}{1 + \sqrt{1 - \xi^2}} \\ \xi &= 1 - 2\sin^2\left(\frac{\pi}{2N}\right) \end{aligned} \qquad (۱۵.۸.۸)$$

نرخ همگرایی متناظر، چنین است.

$$r_\sigma(M(\omega^*)) = \omega^* - 1 = 1 - 2h\pi + O(h^2) \qquad (۱۶.۸.۸)$$

این نرخ همگرایی خیلی بهتر از نرخی است که با (۱۳.۸.۸)، داده شده است. روش شتاب داده‌شدهٔ (۱۴.۸.۸) گاوس ـ زایدل با مقدار بهینهٔ ω^* در (۱۵.۸.۸)، روش ***SOR*** خوانده می‌شود. نام

SOR مخفف ***successive over relaxation*** است، نامی است بر پایهٔ تعبیر فیزیکی این روش، که ابتدا در پیدا کردن آن به‌کار برده شده است.

مثال مثال قبلی (۱۱.۸.۸) را به‌یاد آورید. این مثال هم با روش گاوس ـ زایدل حل شده بود و هم با روش SOR. حدس اولیه برای بارست، از فرمول درونیابی دو خطی برای داده‌های مرزی f، یعنی

$$u_h^{(0)}(x,y) = (1-x)f(0,y) + xf(1,y) + (1-y)f(x,0) + yf(x,1)$$
$$-[(1-y)(1-x)f(0,0) + (1-y)xf(1,0)$$
$$+y(1-x)f(0,1) + xyf(1,1)] \qquad (17.8.8)$$

در تمام نقاط شبکه‌یی داخلی گرفته شده‌است. آزمون خطا برای توقف بارست چنین بوده است،

$$\underset{1\le j,k\le N-1}{\text{Max}} \mid u_h(x_j,y_k) - u_h^{(m)}(x_j,y_k) \mid \le \varepsilon$$

که $\varepsilon > 0$ داده شده و طرف راست (۵.۷.۸) برای پیشگویی خطا در بارست به‌کار برده شده است. نتایج عددی برای تعداد لازم بارستها در جدول ۹.۸ داده شده‌اند. روش SOR به بارستهای بسیار کمتری نسبت به روش گاوس ـ زایدل برای مقادیر کوچکتر h نیاز دارد.

از بخش قبل یادآوری می‌کنیم که تعداد بارستهای لازم، که m^* نامیده شده بود، برای کاهش خطای بارستی با عامل ε، متناسب است با $1/\ln(c)$، که c نسبتی است که خطای بارستی در هر مرحله کاهش می‌یابد. برای روشهای این بخش، c را، $c = r_\sigma(M)$ می‌گیریم. اگر بنویسیم $r_\sigma(M) = 1-\delta$، آنگاه

$$\frac{1}{-\ln r_\sigma(M)} = \frac{1}{-\ln(1-\delta)} \doteq \frac{1}{\delta}$$

جدول ۹.۸ تعداد بارستهای لازم برای حل (۵.۸.۸)

N	ε	گاوس ـ زایدل	SOR
۸	۰٫۰۱	۲۵	۱۲
۸	۰٫۰۰۱	۴۰	۱۶
۱۶	۰٫۰۰۱	۱۴۲	۳۲
۳۲	۰٫۰۰۱	۴۹۵	۶۵
۸	۰٫۰۰۰۱	۵۴	۱۸
۱۶	۰٫۰۰۰۱	۲۰۱	۳۵
۳۲	۰٫۰۰۰۱	۷۳۳	۷۱

وقتی δ نصف شود، تعداد بارستهای محاسبه دو برابر می‌شود. در روش گاوس ـ زایدل

$$\frac{1}{\ln r_\sigma(M)} \doteq \frac{1}{\pi^2 h^2}$$

وقتی h نصف شود تعداد بارستهای محاسبات با ضریب ۴ افزایش می‌یابد. در روش SOR

$$\frac{1}{\ln r_\sigma(M(\omega^*))} \doteq \frac{1}{2\pi h}$$

و وقتی که h نصف شود، تعداد بارستها، دو برابر می‌شود. این دو نتیجه برای داده‌های $\varepsilon = 10^{-3}$ و $\varepsilon = 10^{-4}$ در جدول ۹.۸ داده شده‌اند.

ملاحظه می‌کنید که دو برابرکردن N در هر روش تعداد معادلاتی را که باید حل شوند ۴ برابر افزایش می‌دهد و لذا کار در هر بارست نیز به همین اندازه افزایش می‌یابد. استفاده از SOR کار را بسیار تقلیل می‌دهد اگرچه باز هم وقتی N بزرگ باشد زیاد است.

۹.۸ روش گرادیان مزدوج

روش بارستی که در این بخش ارائه شد، در دههٔ ۱۹۵۰ به‌وجود آمده، ولی شهرت عمدهٔ خود را در سالهای اخیر به‌دست آورده است، به‌ویژه در حل دستگاههای خطی که از حل عددی معادلات دیفرانسیل جزئی خطی به‌دست آمده‌اند. نوشته‌ها دربارهٔ روش گرادیان مزدوج (روش CG) بسیار گسترده و پیچیده شده‌اند و با سایر مباحث جبر خطی مربوط هستند. بنابراین به لحاظ کمبود جا، می‌توانیم فقط به یک مقدمهٔ کوتاه، شامل تعریف روش گرادیان مزدوج و ذکر بعضی نتایج نظری اساسی مربوط به آن اکتفا کنیم.

روش گرادیان مزدوج با روشهای دیگر این فصل از این نظر تفاوت دارد که این روش برای حل مسائل غیرخطی پایه‌گذاری شده است؛ در واقع، روش گرادیان مزدوج روشی است که برای مینیمم‌سازی توابع غیرخطی مورد استفاده قرار گرفته است. در دستگاه خطی که باید حل شود،

$$Ax = b \qquad (۱.۹.۸)$$

فرض می‌شود ماتریس ضرایب A حقیقی، متقارن، و معین مثبت است. حل این دستگاه هم‌ارز مینیمم‌سازی تابع زیر است:

$$f(x) = \frac{1}{2} x^T Ax - b^T x \qquad x \in \mathbf{R}^n \qquad (۲.۹.۸)$$

جواب یکتای x^* معادلهٔ $Ax = b$ نیز مینیمم‌کنندهٔ یکتای $f(x)$ است، وقتی x در $\mathbf{R}^n$ تغییر می‌کند. برای مشاهدهٔ این موضوع، ابتدا نشان می‌دهیم که

$$f(x) = E(x) - \frac{1}{2}b^T x^* \qquad (۳.۹.۸)$$
$$E(x) \equiv \frac{1}{2}(x^* - x)^T A(x^* - x)$$

با استفاده از $Ax^* = b$، اثبات سرراست است. تفاوت $E(x)$ و $f(x)$ یک مقدار ثابت است، پس هر دو یک مینیمم‌ساز دارند. به موجب معین مثبت‌بودن A، $E(x)$ فقط در $x = x^*$ مینیمم می‌شود، و بنابراین $f(x)$ نیز فقط در x^* مینیمم است.

یک روش بارستی معروف برای پیداکردن مینیمم یک تابع غیرخطی، **تندترین کاهش** است، که به اختصار در بخش ۱۲.۲ توضیح داده شده است. برای مینیمم‌کردن $f(x)$ با این روش، فرض می‌کنیم که یک حدس اولیهٔ x_0 داده شده است. مسیری را انتخاب می‌کنیم که در x_0، $f(x)$ تندترین کاهش را داشته باشد. این مسیر با $g_0 = -\nabla f(x_0)$، گرادیان $f(x)$ در x_0 با علامت منفی، داده می‌شود:

$$g(x_0) \equiv g_0 = b - Ax_0 \qquad (۴.۹.۸)$$

سپس، مسألهٔ مینیمم‌سازی یک‌بعدی

$$\underset{0 \le \alpha \le \infty}{\text{Min}} f(x_0 + \alpha g_0)$$

را حل می‌کنیم و جواب آن را α_1 می‌نامیم. با استفاده از این جواب بارست جدید زیر را تعریف می‌کنیم

$$x_1 = x_0 + \alpha_1 g_0 \qquad (۵.۹.۸)$$

این فرایند را به‌طور استقرایی ادامه می‌دهیم. روش تندترین کاهش همگرا می‌شود ولی همگرایی معمولاً خیلی آهسته است. تدبیر بهینهٔ موضعی، با استفاده از امتداد تندترین کاهش، یک تدبیر خوب برای پیداکردن یک امتداد بهینه در پیداکردن مینیمم کلی نخواهد بود. در مقایسه، روش گرادیان مزدوج سریعتر خواهد بود، و اگر فرض کنیم خطای گردکردن وجود ندارد، به بیشتر از n بارست نیاز نخواهد داشت.

برای بقیهٔ این بخش، فرض می‌کنیم که حدس اوّلیه $x_0 = 0$ است. اگر صفر نباشد، همیشه می‌توانیم مسألهٔ کمی تغییریافتهٔ زیر را حل کنیم

$$Az = b - Ax_0$$

با نشان‌دادن جواب آن با z^*، داریم $x^* \doteq x_0 + z^*$. یک حدس اولیه $z_0 = 0 \doteq z^*$ متناظر با $x^* \doteq x_0$ در مسألهٔ اصلی خواهد بود. از این پس فرض می‌کنیم $x_0 = 0$.

روشهای امتداد مزدوج. فرض می‌کنیم A، $n \times n$ باشد، گوییم در $\mathbf{R}^n$ یک مجموعه از بردارهای ناصفر p_1، ...، p_n، A ـ **مزدوج** است اگر

$$p_i^T A p_j = 0 \qquad 1 \le i, j \le n \qquad i \ne j \tag{۶.۹.۸}$$

بردارهای p_j اغلب **امتدادهای مزدوج** نامیده می‌شوند. با معرفی یک حاصلضرب داخلی و یک نرم جدید برای $\mathbf{R}^n$، می‌توان تعریف هندسی هم‌ارزی ارائه داد:

$$\begin{aligned} (x, y)_A &= x^T A y \\ \| x \|_A &= \sqrt{(x, x)_A} = \sqrt{x^T A x} \qquad x \in \mathbf{R}^n \end{aligned} \tag{۷.۹.۸}$$

شرط (۶.۹.۸) هم‌ارز آن است که خواسته شود p_1، ...، p_n یک پایهٔ متعامد برای R^n نسبت به ضرب داخلی $(0, 0)_A$ باشد. از این‌رو، همچنین می‌گوییم که $\{p_1, \ldots, p_n\}$، A ـ متعامدند اگر در (۶.۹.۸) صدق کنند. مشاهده می‌شود که تابع $E(x)$ در (۳.۹.۸) با نرم $\| \cdot \|_A$ چنین است:

$$E(x) = \frac{1}{2} \| x^* - x \|_A^2 \tag{۸.۹.۸}$$

که به نحو روشنتری یک اندازه برای خطای $x^* - x$ است. رابطهٔ بین $\| x \|_2$ و $\| x \|_A$ در مسألهٔ ۳۶ بررسی شده است.

اگر یک مجموعه از امتدادهای مزدوج $\{p_1, \ldots, p_n\}$ داده شده باشد، حل $Ax = b$ سرراست است. گیریم

$$x^* = \alpha_1 p_1 + \ldots + \alpha_n p_n$$

با استفاده از (۶.۹.۸)،

$$\alpha_k = \frac{p_k^T A x^*}{p_k^T A p_k} = \frac{p_k^T b}{p_k^T A p_k} \qquad k = 1, \ldots, n. \tag{۹.۹.۸}$$

ما این فرمول را برای x^* به‌کار می‌بریم تا **روش بارستی امتداد مزدوج** را معرفی کنیم. گیریم $x_0 = 0$،

$$x_k = \alpha_1 p_1 + \ldots + \alpha_k p_k \qquad 1 \le k \le n \tag{۱۰.۹.۸}$$

ماندهٔ x_k در $Ax = b$ را به شکل زیر معرفی می‌کنیم:

$$r_k = b - A x_k = - \bigtriangledown f(x_k)$$

به‌سادگی دیده می‌شود که $r_0 = b$ و

$$x_k = x_{k-1} + \alpha_k p_k \qquad r_k = r_{k-1} - \alpha_k A p_k \qquad k = 1, \ldots, n \quad (11.9.8)$$

به‌ازای $k = n$ داریم $x_n = x^*$ و $r_n = 0$ و x_k ممکن است به‌ازای $k < n$ برابر x^* باشد.

لم ۱ بردار r_k بر p_1، ...، p_k متعامد است، یعنی، $r_k^T p_i = 0$، $1 \le i \le k$.
برهان مسألهٔ ۳۷ را ببینید.

لم ۲(الف) مسألهٔ مینیمم‌سازی

$$\operatorname*{Min}_{-\infty<\alpha<\infty} f(x_{k-1} + \alpha p_k)$$

با $\alpha = \alpha_k$ به‌طور یکتا حل می‌شود و مقدار مینیمم $f(x_k)$ به‌دست می‌آید.
(ب) گیریم $\mathscr{S}_k = \operatorname{Span}\{p_1, \ldots, p_k\}$، زیرفضای kـ بعدی باشد که از $\{p_1, \ldots, p_k\}$ به‌وجود می‌آید. در این صورت مسألهٔ

$$\operatorname*{Min}_{x \in \mathscr{S}_k} f(x)$$

به‌طور یکتا با $x = x_k$ حل می‌شود و مقدار مینیمم، $f(x_k)$ را به‌دست می‌دهد.

برهان (الف) : $\varphi(\alpha) \equiv f(x_{k-1} + \alpha p_k)$ را بسط می‌دهیم:

$$\varphi(\alpha) = f(x_{k-1}) + \alpha p_k^T A x_{k-1} + \frac{1}{2}\alpha^2 p_k^T A p_k - \alpha b^T p_k$$

جملهٔ $p_k^T A x_{k-1}$ برابر صفر است، زیرا $x_{k-1} \in \mathscr{S}_{k-1}$ و p_k، Aـ متعامد بر $\mathscr{S}_{k-1}$ است. از حل $\varphi'(\alpha) = 0$، $\alpha = \alpha_k$ نتیجه می‌شود که برهان را کامل می‌کند.
(ب) برای هر $h \in \mathscr{S}_k$، تابع $f(x_k + h)$ را بسط می‌دهیم:

$$f(x_k + h) = f(x_k) + h^T A x_k + \frac{1}{2} h^T A h - h^T b$$
$$= f(x_k) + \frac{1}{2} h^T A h - h^T r_k$$

با توجه به لم ۱ و فرض $h \in \mathscr{S}_k$ نتیجه می‌شود که $h^T r_k = 0$. بنابراین چون A معین مثبت است،

$$f(x_k + h) = f(x_k) + \frac{1}{2} h^T A h \ge f(x_k)$$

اگر $h = 0$ گرفته شود، مینیمم به‌طور یکتا در $\mathscr{S}_k$ به‌دست می‌آید که (ب) را ثابت می‌کند. ■

لم ۲ یک ویژگی بهینگی برای روشهای امتدادهای مزدوج را به‌دست می‌دهد که با (۱۱.۹.۸) و (۹.۹.۸) تعریف شده‌اند. مسأله، دانستن چگونگی انتخاب امتدادهای مزدوج $\{p_j\}$ است. انتخابهای زیادی وجود دارند که بعضی به روشهای معلوم برای حل مستقیم $Ax = b$ می‌انجامند [استوارت[1] (۱۹۷۳b) را ببینید].

روش گرادیان مزدوج. ما راهی ارائه می‌دهیم که همزمان امتدادهای $\{p_k\}$ و بارستهای $\{x_k\}$ را تولید کند. برای اولین امتداد p_1، امتداد تندترین کاهش را به‌کار می‌بریم:

$$p_1 = -\nabla f(x_0) = r_0 = b \qquad (۱۲.۹.۸)$$

زیرا $x_0 = 0$. یک ساختمان استقرایی برای بقیهٔ امتدادها می‌دهیم. فرض می‌کنیم x_1، ...، x_k همراه با امتدادهای p_1، ...، p_k تولید شده باشند. یک جهت جدید p_{k+1} باید انتخاب شود که با p_1، ...، p_k، A ـ مزدوج باشد. همچنین فرض می‌کنیم $x_k \neq x^*$ و بنابراین $r_k \neq 0$؛ در غیر این صورت جواب x^* به‌دست آمده و کار تمام شده است.

به موجب لم ۱، r_k بر $\mathscr{S}_k$ متعامد است، و بنابراین r_k، به $\mathscr{S}_k$ تعلق ندارد. r_k را برای تولید p_{k+1}، با انتخاب یک مولفهٔ r_k، به‌کار می‌بریم. به دلایلی که ذکر آن در اینجا دشوار است، کافی است که

$$p_{k+1} = r_k + \beta_{k+1} p_k \qquad (۱۳.۹.۸)$$

را درنظر بگیریم. پس شرط $p_k^T A p_{k+1} = 0$ ایجاب می‌کند که

$$\beta_{k+1} = -\frac{p_k^T A r_k}{p_k^T A p_k} \qquad (۱۴.۹.۸)$$

مخرج مخالف صفر است زیرا A معین مثبت است و $p_k \neq 0$. می‌توان نشان داد [لوئنبرگر (۱۹۸۴)، ص ۲۴۵] که این تعریف p_{k+1} در رابطهٔ زیر نیز صدق می‌کند

$$p_j^T A p_{k+1} = 0 \qquad j = 1, 2, \ldots, k-1 \qquad (۱۵.۹.۸)$$

که نشان می‌دهد $\{p_1, \ldots, p_{k+1}\}$، A ـ مزدوج است.

روش گرادیان مزدوج عبارت است از انتخاب $\{p_k\}$ از (۱۲.۹.۸) ـ (۱۴.۹.۸) و $\{x_k, r_k\}$ از (۱۱.۹.۸) و (۹.۹.۸). اگر از خطاهای گردکردن صرفنظر شود، این روش با n بارست یا کمتر

1. Stewart

همگرا می‌شود. سرعت واقعی همگرایی تا اندازهٔ زیادی با ویژه‌مقدارهای A تغییر می‌کند. تحلیل خطای روش گرادیان مزدوج بر پایهٔ قضیهٔ بهینگی زیر استوار است:

قضیهٔ ۹.۸ بارستهای $\{x_k\}$ در روش گرادیان مزدوج در رابطهٔ زیر صدق می‌کنند

$$\| x^* - x_k \|_A = \min_{\deg(q)<k} \| x^* - q(A)b \|_A \qquad (۱۶.۹.۸)$$

برهان برای یک چندجمله‌یی $q(\lambda)$ ، نماد $q(A)$ یک عبارت ماتریسی است که در آن به‌جای هر توان λ^i، A^i گذاشته شده است. برای مثال

$$q(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 \Rightarrow q(A) = a_0 I + a_1 A + a_2 A^2$$

برهان (۱۶.۹.۸) در لوئنبرگر (۱۹۸۴، ص ۲۴۶) داده شده است. ■

با استفاده از این قضیه، تعدادی از نتیجه‌های خطا داده می‌شوند، که با ویژگیهای ماتریس A تغییر می‌یابند. برای مثال گیریم که ویژه مقدارهای A چنین باشند

$$0 < \lambda_1 \leq \ldots \leq \lambda_n \qquad (۱۷.۹.۸)$$

که برحسب چندگانگی تکرار شده‌اند، وگیریم v_1، ...، v_n یک پایهٔ یکا متعامد مربوط به ویژه‌بردارها باشند، می‌نویسیم

$$x^* = \sum_{j=1}^{n} c_j v_j \qquad b = Ax^* = \sum_{j=1}^{n} c_j \lambda_j v_j \qquad (۱۸.۹.۸)$$

پس

$$q(A)b = \sum_{j=1}^{n} c_j \lambda_j q(\lambda_j) v_j$$

$$\| x^* - q(A)b \|_A = \left[\sum_{j=1}^{n} c_j^2 \lambda_j [1 - \lambda_j q(\lambda_j)]^2 \right]^{1/2} \qquad (۱۹.۹.۸)$$

هر انتخاب چندجمله‌یی $q(\lambda)$ از درجهٔ کمتر از k یک کران برای $\| x^* - x_k \|_A$ به‌دست می‌دهد. یکی از بهترین کرانهایی که می‌شناسیم

$$\| x^* - x_k \|_A \leq 2 \left[\frac{1 - \sqrt{c}}{1 + \sqrt{c}} \right]^k \| x^* \|_A \qquad (۲۰.۹.۸)$$

است که در آن $c = \lambda_1/\lambda_n$، معکوس ضریب وضعیت $\text{cond}(A)_2$ است. این یک کران عادی است، که یک همگرایی ضعیف برای مسائل بدوضع به‌دست می‌دهد. طرح کلی اثبات آن در لوئنبرگ (۱۹۸۴، ص ۲۵۸، مسألهٔ ۱۰) آمده است.. کرانهای دیگر را می‌توان بر پایهٔ رفتار ویژه‌مقدارهای $\{\lambda_i\}$ و ضرایب c_i در (۱۸.۹.۸) به‌دست آورد. در بسیاری از کاربردها که λ_i زیاد تغییر می‌نماید، اغلب اتفاق می‌افتد که c_j برای λ_j ی کوچکتر کاملاً نزدیک صفر باشد. در این صورت فرمول (۱۹.۹.۸) را می‌توان دستکاری کرد تا یک کران بهبودیافته نسبت به (۲۰.۹.۸) به‌دست آید. در موارد دیگر، ممکن است تعدادی از ویژه‌مقدارها حول یک تعداد از مقادیر یکی شوند و در این صورت (۱۹.۹.۸) را می‌توان برای نشان‌دادن همگرایی با یک k ی کوچکتر به‌کار برد. برای قضایای دیگر، لوئنبرگ (۱۹۸۴، ص ۳۵۰)، جنینگز[1] (۱۹۷۷)، و وان در اسلوییس[2] و وان در هورست[3] (۱۹۸۶) را ببینید.

فرمولهای $\{\alpha_j, \beta_j\}$ در تعریف روش CG را بازهم ممکن است بهترکرد تا فرمولهای ساده‌تر و کاراتری به‌دست دهند. ما آنها را به صورت زیر می‌آوریم.

الگوریتم $\text{CG}(A, b, x, n)$

۱. ملاحظه: در این الگوریتم جواب $Ax = b$ با استفاده از روش گرادیان مزدوج حساب می‌شود.

۲. $x_0 := 0$، $r_0 := b$، $p_0 := 0$

۳. برای $k = 0, \ldots, n-1$، تا مرحلهٔ ۷ جلو بروید.

۴. اگر $r_k = 0$، بگیرید $x = x_k$ و خارج شوید.

۵. برای $k = 0$، $\beta_1 := 0$ و

برای $k > 0$، $\beta_{k+1} := r_k^T r_k / r_{k-1}^T r_{k-1}$

$$p_{k+1} := r_k + \beta_{k+1} p_k$$

۶. $\alpha_{k+1} := r_k^T r_k / p_{k+1}^T A p_{k+1}$

$$x_{k+1} := x_k + \alpha_{k+1} p_{k+1}$$

$$r_{k+1} := b - A x_{k+1}$$

۷. حلقه را در k تمام کنید.

۸. $x := x_n$ خارج شوید.

در این الگوریتم مسائل مربوط به حساب دقت متناهی در نظر گرفته نمی‌شود. برای یک الگوریتم کاملتر به ویلکینسن و راینش (۱۹۷۱، ۵۷ـ۶۹) مراجعه کنید.

1. Jennings 2. Van der Sluis 3. Van der Horst

بحث ما دربارهٔ روش گرادیان مزدوج خیلی نزدیک روشی است که در لوئنبرگر (۱۹۸۴، فصل ۸) آمده است. برای روش دیگری با انگیزهٔ هندسی بیشتر گلوب و وَن‌لون (۱۹۸۳، بخش ۲.۱۰) را ببینید. اینان همچنین به منابع وسیعی از نوشته‌ها ارجاع داده‌اند.

مثال به‌عنوان یک آزمون ساده، ماتریس مرتبهٔ ۵ زیر را به‌کار می‌بریم

$$A = \begin{bmatrix} ۵ & ۴ & ۳ & ۲ & ۱ \\ ۴ & ۵ & ۴ & ۳ & ۲ \\ ۳ & ۴ & ۵ & ۴ & ۳ \\ ۲ & ۳ & ۴ & ۵ & ۴ \\ ۱ & ۲ & ۳ & ۴ & ۵ \end{bmatrix} \qquad (۲۱.۹.۸)$$

کوچکترین و بزرگترین ویژه‌مقدارها به ترتیب، $\lambda_۱ \doteq ۰٫۵۴۸۴$ و $\lambda_۵ \doteq ۱۷٫۱۷۷۸$ هستند. برای کران خطای (۲۰.۹.۸)، $c \doteq ۰٫۰۳۱۹۲۵$، و

$$\frac{۱ - \sqrt{c}}{۱ + \sqrt{c}} \doteq ۰٫۶۹۷$$

برای دستگاه خطی می‌گیریم

$$b \doteq [۷٫۹۳۸۰, ۱۲٫۹۷۶۳, ۱۷٫۳۰۵۷, ۱۹٫۴۳۳۲, ۱۸٫۴۱۹۶]^T$$

که به جواب درست

$$x^* \doteq [-۰٫۳۲۲۷, ۰٫۳۵۴۴, ۱٫۱۰۱۰, ۱٫۵۷۰۵, ۱٫۶۸۹۷]^T$$

می‌رسیم. نتایج با استفاده از روش گرادیان مزدوج، همراه با کران خطای (۲۰.۹.۸) در جدول ۱۰.۸ داده شده است. همان‌گونه که قبلاً بیان شد، کران (۲۰.۹.۸) خیلی عادی است.

همان‌گونه که انتظار می‌رفت، مانده‌ها کاهش یافته‌اند. ولی توجّه به طریقی که امتدادهای $\{p_k\}$ ساخته شده‌اند، ایجاب می‌کند که به‌دست آوردن امتدادهای دقیق p_k، برای k های بزرگتر، به علّت تعداد کمتر ارقام دقیق مانده‌های r_k، احتمالاً مشکل شود. برای بحث در این مورد، گلوب و وَن‌لون (۱۹۸۳، ص ۳۷۳) را ببینید، که شامل منابع بیشتری از نوشته‌ها در باب این مسأله نیز هست.

جدول ۱۰.۸ مثال از روش گرادیان مزدوج

k	$\| r_k \|_\infty$	$\| x - x_k \|_\infty$	$\| x - x_k \|_A$	کران (۲۰.۹.۸)
۱	۴٫۲۷	۸٫۰۵E − ۱	۲٫۶۲	۱۲٫۷
۲	۸٫۹۸E − ۲	۷٫۰۹E − ۲	۱٫۳۱E − ۱	۸٫۸۳
۳	۲٫۷۵E − ۳	۳٫۶۹E − ۳	۴٫۷۸E − ۳	۶٫۱۵
۴	۷٫۵۹E − ۵	۱٫۳۸E − ۴	۱٫۶۶E − ۴	۴٫۲۹
۵	≐ ۰	≐ ۰	≐ ۰	۲٫۹۹

روش گرادیان مزدوج مشروط کران (۲۰.۹.۸) نشان می‌دهد یا به‌نظر می‌آید که ایجاب می‌کند بارستهای CG حتی برای روشهایی با ضریب وضعیت معمولی چون $\text{cond}(A)_۲ = ۱/c = ۱۰۰$ خیلی آهسته همگرا شوند. برای افزایش نرخ همگرایی، یا حداقل تضمین یک نرخ سریع همگرایی، مسألهٔ $Ax = b$ را به یک مسألهٔ هم‌ارز با ضریب وضعیت کوچکتر تبدیل می‌کنند. کران (۲۰.۹.۸) کوچکتر خواهد شد، و انتظار می‌رود که بارستها با سرعت بیشتری همگرا شوند.

برای یک ماتریس ناتکین Q، $Ax = b$ را به وسیلهٔ

$$(Q^{-۱}AQ^{-T})(Q^T x) = Q^{-۱}b \qquad (۲۲.۹.۸)$$

با $Q^{-T} \equiv (Q^T)^{-۱}$ تبدیل می‌کنیم. می‌نویسیم

$$\tilde{A} = Q^{-۱}AQ^{-T}, \quad \tilde{x} = Q^T x, \quad \tilde{b} = Q^{-۱}b \qquad (۲۳.۹.۸)$$

در این صورت (۲۲.۹.۸) به‌صورت $\tilde{A}\tilde{x} = \tilde{b}$ درآید. ماتریس Q باید طوری انتخاب شود که $\text{cond}(\tilde{A})_۲$ به‌طور چشمگیری از $\text{cond}(A)_۲$ کوچکتر باشد. روش واقعی CG دقیقاً برای حل $\tilde{A}\tilde{x} = \tilde{b}$ به‌کار گرفته نمی‌شود، بلکه الگوریتم CG کمی تعدیل می‌شود. برای الگوریتم حاصل، وقتی ماتریس Q متقارن باشد، گلوب و وَن‌لون (۱۹۸۳، ص ۳۷۴) را ببینید.

پیداکردن Q نیاز به تحلیل دقیق مسألهٔ اصلی $Ax = b$ و درک ساختار A برای انتخاب Q دارد. از (۲۳.۹.۸) داریم:

$$A = Q\tilde{A}Q^T$$

که $\tilde{A}$ باید طوری انتخاب شود که اندازهٔ ویژه‌مقدارها نزدیک به ۱ باشد. مثلاً اگر $\tilde{A}$ نزدیک به ماتریس همانی I باشد، آنگاه $A \doteq QQ^T$. این تجزیه را می‌توان با تجزیه مثلثی چولسکی انجام داد. عاملهای تقریبی چولسکی گاهی در تعریف شرط گزارها به‌کار می‌روند. برای آشنایی با مسألهٔ انتخاب شرط گزارها، گلوب و وَن‌لون (۱۹۸۳، بخش ۳.۱۰) و اکسلسون[1] (۱۹۸۵) را ببینید.

1. Axelson

بحث در آثار خواندنی

منابعی که تأثیر زیادی در ارائهٔ حذف گاوسی و مباحث دیگر این فصل دارند کتابهای فورسایت و مولر (۱۹۶۷) و گلوب و وَن‌لون (۱۹۸۳)، آیزکسن و کلر (۱۹۶۶) و ویلکینسن (۱۹۶۳، ۱۹۶۵) همراه با مقالهٔ کاهان (۱۹۶۶) هستند. برداشتهای کلّی بسیار خوب دیگر در کنت و دبور[1] (۱۹۸۰)، نوبل (۱۹۶۹)، رایس[2] (۱۹۸۱) و استوارت (۱۹۷۳a) دیده می‌شوند. درامد مقدماتی‌تری در آنتون (۱۹۸۴) و استرنگ (۱۹۸۰) داده شده‌اند.

بهترین برنامه‌ها برای حلّ مستقیم صورتهای کلّی و خاص دستگاههای خطی با اندازه‌های کوچک تا متوسط، مبتنی بر صورتهایی هستند که در بسته‌های LINPACK آمده و در دونگارا و همکاران (۱۹۷۹) تشریح شده است. این برنامه‌ها کاملاً قابل‌اجرا هستند و در هر دو حساب با دقّت ساده و دقّت مضاعف هم برای حساب با اعداد حقیقی و هم حساب با اعداد مختلط موجودند. درکنار حل این دستگاهها، با این برنامه‌ها می‌توان ضریب وضعیت ماتریس مورد نظر را نیز براورد کرد. برنامه‌های معادلات خطی در IMSL و NAG انواع مختلف و پیشرفتهٔ همان برنامه‌های LINPACK هستند.

جنبهٔ دیگر LINPACK استفاده از **زیر روالهای اساسی جبرخطی** (BLAS) است. اینها زیربرنامه‌هایی در سطح پایین‌اند که عملیات اساسی برداری مانند حاصلضرب عددی دو بردار و مجموع دو بردار را انجام می‌دهند. اینها به زبان فرترن، به صورت قسمتی از LINPACK موجودند؛ ولی با ارائه کُد زبان اسمبلی آنها، اغلب می‌توان کارایی برنامه‌های اصلی LINPACK را به‌طور قابل ملاحظه‌ای بهبود بخشید. برای یک بحث کلیتر در مورد BLAS ، لاوسن و همکاران (۱۹۷۹) را ببینید. برنامه‌های LINPACK به فراوانی در دسترس‌اند و در پیشرفت برنامه‌های معادلات جبری در سایر بسته‌های نرم‌افزاری مؤثر بوده‌اند.

نوشته‌های زیادی در مورد حل دستگاههای خطی که از حل عددی معادلات دیفرانسیل جزئی به‌دست می‌آیند وجود دارد. برای متون کلّی در حل عددی معادلات دیفرانسیل جزئی، بیرکهاف و لینچ (۱۹۸۴)، فورسایت و وازوو[3] (۱۹۶۰)، گلادول[4] و ویت[5] (۱۹۷۹)، لاپیدوس و پیندر[6] (۱۹۸۱) و ریشتمایر و مورتون[7] (۱۹۶۷) را ببینید. برای متنهایی که به روشهای بارستی کلاسیک برای حل دستگاههای خطی حاصل از حل عددی معادلات دیفرانسیل جزئی اختصاص یافته، هاگمن[8] و یانگ (۱۹۸۱) و وارگا[9] (۱۹۶۲) را ببینید. برای شیوه‌های دیگر که اخیراً مورد توجه قرار گرفته اسوارتس‌تراوبر[10] (۱۹۸۴)، اسوارتس‌تراوبر و سویت (۱۹۷۹)، جورج و لیو[11] (۱۹۸۱) و هاکبوش و تروتنبرگ[12] (۱۹۸۲) را ببینید.

---

1. Conte and de Boor
2. Rice
3. Wasow
4. Gladwell
5. Wait
6. Lapidus and Pinder
7. Richtmyer and Morton
8. Hageman
9. Varga
10. Swarztrauber
11. George and Liu
12. Hackbusch and Trottenberg

حل عددی معادلات دیفرانسیل جزیی منبع درصد زیادی از دستگاههای خطی تُنک است که در عمل حل شده‌اند. ولی، دستگاههای خطی تُنک از مرتبهٔ بزرگ درکاربردهای دیگر نیز وجود دارند [مثلاً داف[1] (۱۹۸۱) را ببینید]. روشهای گوناگون زیادی برای حل دستگاههای بزرگ تُنک وجود دارند، که دربارهٔ بعضی از آنها در بخشهای ۶.۸ تا ۸.۸ بحث کردیم. روشهای مستقیم و بارستی دیگر، که به ساختار ماتریس بستگی دارند، در دسترس‌اند. برای یک نمونه از تحقیقات جاری در این زمینهٔ بسیار فعال، بررسی داف (۱۹۷۷)، گزارش بیورک[2] و همکاران (۱۹۸۱)، داف (۱۹۸۱)، داف و استوارت (۱۹۷۹)، اوانز[3] (۱۹۸۵) وکتابهای درسی جورج و لیو (۱۹۸۱) و پیسانتسکی[4] (۱۹۸۴) را ببینید. بسته‌های نرم‌افزاری چندی برای حل انواع مختلف دستگاههای تُنک وجود دارند، که بعضاً درکتابهای قبلی آمده‌اند. برای یک فهرست کلی از بسیاری از بسته‌های موجود، گردآوری هیث[5] (۱۹۸۲) را ببینید. برای روشهای بارستی در حل دستگاههای مربوط به حل معادلات دیفرانسیل جزئی، درکتابهای وارگا (۱۹۶۲) و هاگمن و یانگ (۱۹۸۱) بسیاری از روشهای کلاسیک مورد بحث قرار گرفته‌اند.

معادلات انتگرالی به دستگاههای خطی چگال منجر می‌شوند، و انواع دیگر روشهای بارستی برای حل آنها به‌کار برده شده‌اند. برای بعضی روشهای کاملاً موفق، آتکینسون (۱۹۷۶، فصل ۴) را ببینید.

روش گرادیان مزدوج با هستنسن[6] و اشتیفل[7] (۱۹۵۲) آغاز شده است، و استفاده از آن در حل معادلات دیفرانسیل جزیی و انتگرالی هنوز در حال گسترش است. برای بحثهای مفصّلتر مربوط به روش جهتهای مزدوج برای سایر روشهای عددی هستنس (۱۹۸۰) و استوارت (۱۹۷۳b) را ببینید. برای مراجعه به نوشته‌های تازه، از جمله بحثهایی در روش گرادیان مزدوج مشروط اکسلسون (۱۹۸۵)، اکسلسون و لیندزکوگ[8] (۱۹۸۶) و (بخشهای ۲.۱۰ و ۳.۱۰) گلوب و ون‌لون (۱۹۸۳) را ببینید. یک تعمیم برای دستگاههای نامتقارن در آیزن‌اشتات[9] و همکاران (۱۹۸۳) پیشنهاد شده است.

یکی از مهمترین عواملی که جهت پژوهشهای آینده در جبر خطی عددی را مشخص خواهد نمود استفادهٔ روزافزون از رایانه‌های برداری و پردازندهٔ موازی است. ماشینهای برداری، مانند CRAY-2 هنگام عملیات اساسی روی کمّیتهای برداری، مانند آنچه که در BLAS از LINPACK به‌کار برده‌شده، به نحو احسن کار می‌کنند. در سالهای اخیر، زمان دسترسی به استفاده از این ماشینها در شبکه‌های رایانه‌یی جهانی که اخیراً توسعه یافته بسیار افزوده شده است. این موضوع، مقیاس بسیاری از مسائل فیزیکی را که می‌توان به آنها پرداخت، تغییر داده است، و همین امر نیاز بیشتری، به برنامه‌های هرچه کاراتر رایانه‌یی را، در حل دستهٔ وسیعی از دستگاههای خطی، موجب شده است. استفاده از رایانه‌های موازی حتی جدیدتر است، و تنها در اواسط تا اواخر دهه ۱۹۸۰ متداول

1. Duff 2. Björck 3. Evans 4. Pissanetsky
5. Heath 6. Hestenson 7. Stiefel
8. Axelsson and Lindskog 9. Eisenstat

شده است. معماریهای خیلی متنوعی برای این‌گونه ماشینها وجود دارد. بعضی دارای پردازنده‌های چندتایی هستند که از یک حافظه، با طرحهای گوناگون مشترکاً استفاده می‌کنند؛ در بعضی دیگر، هر پردازنده دارای حافظهٔ اختصاصی است که به راههای گوناگون با پردازنده‌های دیگر مرتبط است. رایانه‌های موازی، در مقایسه با رایانه‌هائی که ما مطالعه کرده‌ایم، اغلب به الگوریتمهای عددی کاملاً متفاوتی منجر می‌شوند، الگوریتمهایی که در آن می‌توان از پردازنده‌های همزمان، که در یک مسأله به‌کار می‌آیند استفاده کرد. آثار مکتوب در این موارد کم است، ولی به سرعت در حال افزایش است. به‌عنوان یک بررسی برای حل دستگاههای خطی مرتبط با معادلات دیفرانسیل جزئی، هم در رایانه‌های برداری، هم در رایانه‌های موازی، اورتگا و فویکت (۱۹۸۵) را ببینید. برای یک کتاب درسی برنامه‌ریزی شده برای حل دستگاههای خطی، اورتگا (۱۹۸۷) را ببینید.

مراجع

Aird, T., and R. Lynch (1975). Computable accurate upper and lower error bounds for approximate solutions of linear algebraic systems, *ACM Trans. Math. Softw.* **1**, 217–231.

Anton, H. (1984). *Elementary Linear Algebra*, 4th ed. Wiley, New York.

Atkinson, K. (1976). *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind*. SIAM Pub., Philadelphia.

Axelsson, O. (1985). A survey of preconditioned iterative methods for linear systems of algebraic equations, *BIT* **25**, 166–187.

Axelsson, O., and G. Lindskop (1986). On the rate of convergence of the preconditioned conjugate gradient method, *Numer. Math.* **48**, 499–523.

Birkhoff, G., and R. Lynch (1984). *Numerical Solution of Elliptic Problems*. SIAM Pub., Philadelphia.

Björck, A., R. Plemmons, and H. Schneider, Eds. (1981). *Large Scale Matrix Problems*. North-Holland, Amsterdam.

Concus, P., G. Golub, and D. O'Leary (1984). A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations. In *Studies in Numerical Analysis*, G. Golub, Ed. Mathematical Association of America, pp. 178–198.

Conte, S., and C. de Boor (1980). *Elementary Numerical Analysis*, 3rd ed. McGraw-Hill, New York.

Dongarra, J., J. Bunch, C. Moler, and G. Stewart (1979). *Linpack User's Guide*. SIAM Pub., Philadelphia.

Dorr, F. (1970). The direct solution of the discrete Poisson equations on a

rectangle, *SIAM Rev.* **12**, 248–263.

Duff, I. (1977). A survey of sparse matrix research, *Proc. IEEE* **65**, 500–535.

Duff, I., Ed. (1981). *Sparse Matrices and Their Uses*. Academic Press, New York.

Duff, I., and G. Stewart, Eds. (1979). *Sparse Matrix Proceedings 1978*. SIAM Pub., Philadelphia.

Eisenstat, S., H. Elman, and M. Schultz (1983). Variational iterative methods for nonsymmetric systems of linear equations, *SIAM J. Numer. Anal.* **20**, 345–357.

Evans, D., Ed. (1985). *Sparsity and Its Applications*. Cambridge Univ. Press, Cambridge, England.

Forsythe, G., and C. Moler (1967). *Computer Solution of Linear Algebraic Systems*. Prentice-Hall, Englewood Cliffs, N.J.

Forsythe, G., and W. Wasow (1960). *Finite Difference Methods for Partial Differential Equations*. Wiley, New York.

George, A., and J. Liu (1981). *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall, Englewood Cliffs, N.J.

Gladwell, I., and R. Wait, Eds. (1979). *A Survey of Numerical Methods for Partial Differential Equations*. Oxford Univ. Press, Oxford, England.

Golub, G., and C. Van Loan (1983). *Matrix Computations*. Johns Hopkins Press. Baltimore.

Gregory, R., and D. Karney (1969). *A Collection of Matrices for Testing Computational Algorithms*. Wiley, New York.

Hackbusch, W., and U. Trottenberg, Eds. (1982). *Multigrid Methods, Lecture Notes Math. 960*. Springer-Verlag, New York.

Hageman, L., and D. Young (1981). *Applied Iterative Methods*. Academic Press. New York.

Heath, M., Ed. (1982). *Sparse Matrix Software Catalog*. Oak Ridge National Laboratory, Oak Ridge, Tenn. (Published in connection with the Sparse Matrix Symposium 1982.)

Hestenes, M. (1980). *Conjugate Direction Methods in Optimization*. Springer-Verlag, New York.

Hestenes, M., and E. Stiefel (1952). Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Stand.* **49**, 409–439.

Isaacson, E., and H. Keller (1966). *Analysis of Numerical Methods*. Wiley, New York.

Jennings, A. (1977). Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method, *J. Inst. Math. Its Appl.* **20**, 61–72.

Kahan, W. (1966). Numerical linear algebra, *Can. Math. Bull.* **9**, 756–801.

Lapidus, L., and G. Pinder (1982). *Numerical Solution of Partial Differential*

Equations in Science and Engineering. Wiley, New York.

Lawson, C., and R. Hanson (1974). *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, N.J.

Lawson, C., R. Hanson, D. Kincaid, and F. Krogh (1979). Basic linear algebra subprograms for Fortran usage, *ACM Trans. Math. Softw.* **5**, 308–323.

Luenberger, D. (1984). *Linear and Nonlinear Programming*, 2nd ed. Addison-Wesley, Reading, Mass.

Noble, B. (1969). *Applied Linear Algebra*. Prentice-Hall, Englewood Cliffs, N.J.

Ortega, J. (1987). *Parallel and Vector Solution of Linear Systems*. Preprint, Univ. of Virginia, Charlottesville.

Ortega, J., and R. Voigt (1985). Solution of partial differential equations on vector and parallel computers, *SIAM Rev.* **27**, 149–240.

Pissanetzky, S. (1984). *Sparse Matrix Technology*. Academic Press, New York.

Rice, J. (1981). *Matrix Computations and Mathematical Software*. McGraw-Hill, New York.

Richtmyer, R., and K. Morton (1967). *Difference Methods for Initial Value Problems*, 2nd ed. Wiley, New York.

Stewart, G. (1973a). *Introduction to Matrix Computations*. Academic Press, New York.

Stewart, G. (1973b). Conjugate direction methods for solving systems of linear equations, *Numer. Math.* **21**, 284–297.

Stewart, G. (1977). Research, development, and LINPACK. In *Mathematical Software III*, John Rice (Ed.). Academic Press, New York.

Stone, H. (1968). Iterative solution of implicit approximations of multidimensional partial differential equations, *SIAM J. Numer. Anal.* **5**, 530–558.

Strang, G. (1980). *Linear Algebra and Its Applications*, 2nd ed. Academic Press, New York.

Swarztrauber, P. (1984). Fast Poisson solvers. In *Studies in Numerical Analysis*, G. Golub, Ed. Mathematical Association of America, pp. 319–370.

Swarztrauber, P., and R. Sweet (1979). Algorithm 541: Efficient Fortran subprograms for the solution of separable elliptic partial differential equations, *ACM Trans. Math. Softw.* **5**, 352–364.

Van der Sluis, A., and H. van der Vorst (1986). The rate of convergence of conjugate gradients, *Numer. Math.* **48**, 543–560.

Varga, R. (1962). *Matrix Analysis*. Prentice-Hall, Englewood Cliffs, N.J.

Wilkinson, J. (1963). *Rounding Errors in Algebraic Processes*. Prentice-Hall, Englewood Cliffs, N.J.

Wilkinson, J. (1965). *The Algebraic Eigenvalue Problem*. Oxford Univ. Press, Oxford, England.

Wilkinson, J., and C. Reinsch (1971). *Linear Algebra, Handbook for Automatic Computation*, Vol. 2. Springer-Verlag, New York.

مسائل

۱. دستگاههای $Ax = b$ در زیر را با حذف گاوسی بدون محورگیری حل کنید. مانند (۵.۱.۸) $A = LU$ را پیدا کنید.

(الف)
$$A = \begin{bmatrix} ۱ & ۱ & -۱ \\ ۱ & ۲ & -۲ \\ -۲ & ۱ & ۱ \end{bmatrix} \qquad b = \begin{bmatrix} ۱ \\ ۰ \\ ۱ \end{bmatrix}$$

(ب)
$$A = \begin{bmatrix} ۴ & ۳ & ۲ & ۱ \\ ۳ & ۴ & ۳ & ۲ \\ ۲ & ۳ & ۴ & ۳ \\ ۱ & ۲ & ۳ & ۴ \end{bmatrix} b = \begin{bmatrix} ۱ \\ ۱ \\ -۱ \\ -۱ \end{bmatrix}$$

(ج)
$$A = \begin{bmatrix} ۱ & -۱ & ۱ & -۱ \\ -۱ & ۳ & -۳ & ۳ \\ ۲ & -۴ & ۷ & -۷ \\ -۳ & ۷ & -۱۰ & ۱۴ \end{bmatrix} \qquad b = \begin{bmatrix} ۰ \\ ۲ \\ -۲ \\ ۸ \end{bmatrix}$$

۲. دستگاه خطی زیر را در نظر بگیرید و تحقیق کنید که جواب آن، $x_۱ = ۲٫۶$، $x_۲ = -۳٫۸$ و $x_۳ = -۵٫۰$ است.

$$\begin{aligned} ۶x_۱ + ۲x_۲ + ۲x_۳ &= -۲ \\ ۲x_۱ + \frac{۲}{۳}x_۲ + \frac{۱}{۳}x_۳ &= ۱ \\ x_۱ + ۲x_۲ - x_۳ &= ۰ \end{aligned}$$

(الف) با استفاده از حساب اعشاری ممیز شناور چهاررقمی، با گردکردن، دستگاه فوق را با حذف گاوسی بدون محورگیری حل کنید.

(ب) بخش (الف) را با استفاده از محورگیری جزیی تکرار کنید. برای انجام اعمال حساب، به یاد داشته باشید که پس از هر محاسبه اعداد را تا چهاررقم بامعنی گرد کنید، درست همان‌طور که در رایانه عمل می‌شود.

۳. (الف) الگوریتم‌های *Factor* و *Solve* در بخش ۲.۸، یا برنامه‌های مشابه داده‌شده در فورسایت و مولر (۱۹۶۷، فصلهای ۱۶ و ۱۷) (۱۹۶۷) را انجام دهید.

(ب) برای آزمون برنامه، دستگاه $Ax = b$ از مرتبهٔ n را با $A = [a_{ij}]$ که به‌صورت زیر تعریف شده است حل کنید

$$a_{ij} = \text{Max}(i, j)$$

همچنین $b = [۱, ۱, \dots, ۱]^T$ را تعریف می‌کنیم. جوابِ درست $x = [۰, \dots, ۰, (۱/n)]^T$ است. این ماتریس از گرگوری و کارنی (۱۹۶۹، ص ۴۲) اقتباس شده است.

۴. حل معادلهٔ انتگرالی زیر را با گسسته‌ساختن انتگرال با قاعدهٔ انتگرالگیری میانگاهی (۱۸.۲.۵) در نظر می‌گیریم

$$\lambda x(s) - \int_{۰}^{۱} \cos(\pi st)x(t)dt = ۱ \qquad ۰ \le s \le ۱$$

به عبارت دقیقتر، گیریم $n > ۰$، $h = ۱/n$، $t_i = (i - ۱/۲)h$ به ازای $i = ۱, \dots, n$. برای مقادیر تقریبی $x(t_۱)$، ... ، $x(t_n)$ دستگاه خطی زیر را حل می‌کنیم

$$\lambda z_i - \sum_{j=۱}^{n} h \cos(\pi t_i t_j) z_j = ۱ \qquad i = ۱, \dots, n$$

این دستگاه خطی را با $(\lambda I - K_n)z = b$ نمایش می‌دهیم که K_n از مرتبهٔ $n \times n$ و به‌شکل زیر است:

$$(K_n)_{ij} = h \cdot \cos(\pi t_i t_j) \qquad b_i = ۱ \qquad ۱ \le i, j \le n$$

برای مقدار به اندازهٔ کافی بزرگ n، $z_i \doteq x(t_i)$، $۱ \le i \le n$. مقدار λ ناصفر است و در اینجا فرض شده است که λ ویژه‌مقدار K_n نیست.

$(\lambda I - K_n)z = b$ را برای چندین مقدار n، مثلاً $n = ۲, ۴, ۸, ۳۲, ۶۴$ حل کرده، جوابهای بردار z را ثبت کنید. چنانچه ممکن باشد نمودار این جوابها را به‌دست آورید تا اطلاعاتی راجع به تابع جواب $x(s)$ از معادلهٔ انتگرالی اصلی به‌دست آورید. از $\lambda = ۴, ۲, ۱, ۰٫۵$ استفاده کنید.

۵. (الف) حل $Ax = b$ را با A و b مختلط و A از مرتبهٔ n در نظر می‌گیریم. این مسأله را به حل یک دستگاه حقیقی مربعی از مرتبهٔ $۲n$ تبدیل کنید.

راهنمایی: بنویسید $A = A_۱ + iA_۲$، $b = b_۱ + ib_۲$، $x = x_۱ + ix_۲$ که $A_۱$، $A_۲$، $b_۱$، $b_۲$، $x_۱$، $x_۲$ همگی حقیقی‌اند. معادلاتی را که باید $x_۱$، $x_۲$ در آنها صدق کنند تعیین کنید.

(ب) حافظهٔ مورد نیاز و تعداد عملیات روش (الف) را برای حل یک دستگاه مختلط $Ax = b$ تعیین نمایید. این نتایج را با نتایجی که از حل مستقیم $Ax = b$ به روش حذف گاوسی و حساب مختلط به‌دست می‌آید، مقایسه کنید. به هزینهٔ بیشتر اعمال حساب مختلط توجه نمایید.

۶. گیریم A و B و Cبه‌ترتیب ماتریسهایی از مراتب $m \times n$، $n \times p$ و $p \times q$ باشند. تعداد عملیات محاسبهٔ $A(BC)$ و $(AB)C$ را شمارش کنید. مثالهایی بیاورید که در آنها یک ترتیب محاسبه بر ترتیب محاسبهٔ دیگر رجحان داشته باشد.

۷. (الف) نشان دهید که تعداد ضرب و تقسیمهای روش گاوس ـ ژوردان در بخش ۳.۸ در حدود $\frac{1}{2}n^3$ است.

(ب) نشان دهید که چگونه روش گاوس ـ ژوردان، با محورگیری جزئی را می‌توان برای وارون‌کردن ماتریس $n \times n$ فقط با $n(n+1)$ جا در حافظه، به‌کار برد. آیا محورگیری کلی را می‌توان به‌کار برد؟

۸. هریک از برنامه‌های مسألهٔ ۳ (الف) یا روش گاوس ـ ژوردان را برای وارون‌کردن ماتریسهای مسائل ۱ و ۳ (ب) به‌کار برید.

۹. ثابت کنید که اگر $A = LL^T$، و L حقیقی و ناتکین باشد، A متقارن و معین مثبت است.

۱۰. با استفاده از روش چولسکی، تجزیهٔ $A = LL^T$ را برای ماتریسهای زیر به‌دست آورید

(الف)
$$\begin{bmatrix} ۴٫۵ & -۳٫۰ & ۲٫۲۵ \\ -۱۰٫۰ & ۵٫۰ & -۳٫۰ \\ ۳۴٫۰ & -۱۰٫۰ & ۴٫۵ \end{bmatrix}$$

(ب)
$$\begin{bmatrix} -۳ & ۱۵ & -۱۸ & ۱۵ \\ ۴ & -۱۸ & ۲۴ & -۱۸ \\ -۳ & ۱۸ & -۱۸ & ۱۵ \\ ۱ & -۳ & ۴ & -۳ \end{bmatrix}$$

۱۱. گیریم $A = LU = LDM$، و همهٔ l_{ii} و m_{ii} ها مساوی ۱ باشند و D قطری باشد. به‌علاوه فرض می‌کنیم A متقارن است. نشان دهید که $M = L^T$ و بنابراین $A = LDL^T$. نشان دهید که A معین مثبت است اگر و تنها اگر همهٔ d_{ii} ها مثبت باشند.

۱۲. گیریم A حقیقی، متقارن، معین مثبت و از مرتبهٔ n باشد. حل $Ax = b$ با استفاده از حذف گاوسی بدون محورگیری را درنظر می‌گیریم. منظور ما از این مسأله این است که ثابت کنیم محورها ناصفرند.

(الف) نشان دهید که همهٔ عناصر قطری در $a_{ii} > ۰$ صدق می‌کنند. این امر نشان می‌دهد که $a_{۱۱}$ را می‌توان یک عنصر محور گرفت.

(ب) پس از حذف $x_۱$ از معادله‌های ۲ تا n، فرض می‌کنیم ماتریس نتیجهٔ $A^{(۲)}$ بدین شکل باشد.

$$A^{(۲)} = \begin{bmatrix} a_{۱۱} & a_{۱۲} & \dots & a_{۱n} \\ ۰ & & & \\ \vdots & & \hat{A}^{(۲)} & \\ ۰ & & & \end{bmatrix}$$

نشان دهید که $\hat{A}^{(2)}$ متقارن و معین مثبت است.

این شیوهٔ عمل را می‌توان به‌طور استقرایی در هر مرحله از فرایند حذف ادامه داد و بنابراین وجود محورهای ناصفر در هر مرحله توجیه می‌شود.

راهنمایی: برای اثبات معین مثبت‌بودن $\hat{A}^{(2)}$، ابتدا اتحاد زیر را به‌ازای هر انتخاب x_1، ...، x_n ثابت کنید

$$\sum_{i,j=2}^{n} a_{ij}^{(2)} x_i x_j = \sum_{i,j=1}^{n} a_{ij} x_i x_j - a_{11} \left[x_1 + \sum_{j=2}^{n} \frac{a_{j1}}{a_{11}} x_j \right]^2$$

سپس x_1 مناسب را اختیار نمایید.

۱۳. به‌عنوان یک راه دیگر برای ارائهٔ یک روش فشرده در به‌دست آوردن تجزیهٔ A به LU، روش ماتریس جهتدار زیر را در نظر بگیرید. بنویسید

$$A = \begin{bmatrix} \hat{A} & d \\ c^T & \alpha \end{bmatrix} \qquad c, d \in \mathbf{R}^{n-1} \qquad \alpha \in \mathbf{R}$$

و $\hat{A}$ ماتریس مربع از مرتبهٔ $n-1$ است. فرض کنید A ناتکین است. به‌عنوان یک مرحله در فرایند استقرا، فرض کنید که $\hat{A} = \hat{L}\hat{U}$ معلوم است. $A = LU$ را به شکل زیر پیدا کنید

$$A = \begin{bmatrix} \hat{L} & 0 \\ m^T & 1 \end{bmatrix} \begin{bmatrix} \hat{U} & q \\ 0 & \gamma \end{bmatrix} \qquad m, q \in \mathbf{R}^{n-1} \qquad \gamma \in \mathbf{R}$$

نشان دهید که می‌توان q، m و γ را پیدا کرد و توضیح دهید چگونه. (این روش برای ماتریس اصلی A به‌کار می‌رود، با تجزیهٔ هر زیرماتریس اصلی درگوشهٔ چپ بالا به‌ترتیب صعودی).

۱۴. با استفاده از الگوریتم (۲۳.۳.۸) ـ (۲۴.۳.۸) برای حل دستگاههای سه‌قطری، معادلهٔ $Ax = b$ را حل کنید که در آن،

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ 1 & 2 & -1 & 0 & 0 \\ 0 & 1 & 2 & -1 & 0 \\ 0 & 0 & 1 & 2 & -1 \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix} \qquad b = \begin{bmatrix} 3 \\ -2 \\ 2 \\ -2 \\ 1 \end{bmatrix}$$

تحقیق کنید که فرضها و حکمهای قضیهٔ ۲.۸ در این مثال صادق‌اند.

۱۵. ماتریسی سه‌قطری از مرتبهٔ n به شکل زیر را تعریف می‌کنیم

$$A_n = \begin{bmatrix} ۲ & -۱ & ۰ & & \cdots & ۰ \\ -۱ & ۲ & -۱ & ۰ & & \\ ۰ & -۱ & ۲ & -۱ & & \vdots \\ \vdots & & & \ddots & & -۱ \\ ۰ & & \cdots & & -۱ & ۲ \end{bmatrix}$$

یک فرمول کلی برای $A_n = LU$ پیدا کنید.

راهنمایی: حالتهای $n = ۲, ۳, ۵$ را درنظر بگیرید و سپس الگوی کلی را حدس بزنید و درستی آن را بررسی کنید.

۱۶. زیروالی برای حل دستگاههای سه‌قطری با استفاده از (۲۳.۳.۸) ـ (۲۴.۳.۸) بنویسید. درستی آن را با استفاده از مثالهای فصل ۱۴ و ۱۵ امتحان کنید. در گرگوری و کارنی (۱۹۶۹، فصل ۲) نیز تعدادی دستگاههای سه‌قطری وجود دارند که وارونهای درست آنها معلوم‌اند.

۱۷. خانواده‌هایی از دستگاههای خطی $A_k x = b$ وجود دارند که در آنها A_k به طریقی ساده به ماتریس $A_{k+۱}$ تبدیل می‌شود و ممکن است ساده‌تر باشد که برای تجزیهٔ $A_{k+۱}$ به LU، از تبدیل تجزیهٔ A_k استفاده شود. به‌عنوان یک مثال که در روش سادکی برای برنامه‌ریزی خطی پیش می‌آید، فرض می‌کنیم $A_۱ = [a_۱, \ldots, a_n]$ و $A_۲ = [a_۲, \ldots, a_{n+۱}]$، که هر $a_j \in \mathbf{R}^n$. فرض می‌کنیم که $A_۱ = L_۱U_۱$ معلوم و $L_۱$ پایین‌مثلثی و $U_۱$ بالامثلثی باشد. یک راه ساده پیدا کنید که تجزیهٔ $A_۲ = L_۲U_۲$ را از تجزیهٔ $A_۱$ به‌دست دهد، با فرض اینکه محورگیری لازم نباشد.

راهنمایی: با استفاده از $L_۱u_i = a_i$، $۱ \le i \le n$، بنویسید

$$A_۲ = L_۱[u_۲, u_۳, \ldots, u_n, L_۱^{-۱}a_{n+۱}] \equiv L_۱\tilde{U}$$

نشان دهید که $\tilde{U}$ را می‌توان به‌سادگی به یک شکل بالامثلثی $U_۲$ تبدیل کرد، و این تبدیل متناظر با برگرداندن $L_۱$ به شکل مطلوب $L_۲$ است. به عبارت دقیقتر، $U_۲ = M\tilde{U}$، $L_۲ = L_۱M^{-۱}$. نشان دهید که هزینهٔ عمل پیداکردن $A_۲ = L_۲U_۲$ برابر $O(n^۲)$ است.

۱۸. (الف) ضریبهای وضعیت $\mathrm{cond}(A)_p$ را به‌ازای $p = ۱, ۲, \infty$ برای ماتریس زیر پیدا کنید

$$A = \begin{bmatrix} ۱۰۰ & ۹۹ \\ ۹۹ & ۹۸ \end{bmatrix}$$

(ب) ویژه‌مقدارها و ویژه‌بردارهای A را پیدا کنید، و آنها را برای روشن‌کردن تبصره‌هایی که به دنبال (۸.۴.۸) در بخش ۴.۸ آمده‌اند به‌کار برید.

۱۹. ثابت کنید که اگر A یکانی باشد، آنگاه $\text{cond}(A)_* = ۱$.

۲۰. نشان دهید که برای هر A، کران بالای (۴.۸.۸) را با انتخابهای مناسب b و r می‌توان به‌دست آورد.

راهنمایی: به موجب تعریفهای $\| A \|$ و $\| A^{-۱} \|$ در بخش ۳.۷، بردارهای $\hat{x}$ و $\hat{r}$ وجود دارند که برای آنها $\| \hat{x} \| \| A \| = \| A\hat{x} \|$ و $\| \hat{r} \| \| A^{-۱} \| = \| A^{-۱}\hat{r} \|$. با استفاده از این روابط ساختن تساوی در کران بالای (۴.۴.۸) را کامل کنید.

۲۱. ضریب وضعیت $\text{cond}(A)_*$ مربوط به (۶.۴.۸) ممکن است برای ماتریسهای بدوضع A کاملاً کوچک باشد. برای مشاهدهٔ این مطلب، ماتریس $n \times n$ در زیر را تعریف می‌کنیم:

$$A_R = \begin{bmatrix} ۱ & -۱ & -۱ & \dots & -۱ \\ ۰ & ۱ & -۱ & \dots & -۱ \\ \vdots & & \ddots & & \vdots \\ & & & ۱ & -۱ \\ ۰ & & \dots & ۰ & ۱ \end{bmatrix}$$

به سادگی دیده می‌شود که $\text{cond}(A)_* = ۱$. تحقیق کنید که $A_n^{-۱}$ با ماتریس بالامثلثی $B = [b_{ij}]$ داده می‌شود که در آن $b_{ii} = ۱$،

$$b_{ij} = ۲^{j-i-۱} \qquad i < j \le n$$

$\text{cond}(A)_\infty$ را محاسبه کنید.

۲۲. مانند بخش ۲.۸، فرض می‌کنیم $H_n = [۱/(i+j-۱)]$ معرّف ماتریس مرتبهٔ n هیلبرت باشد و گیریم $\overline{H}_n$ ماتریسی باشد که از واردکردن H_n به رایانهٔ شما در حساب با دقّت ساده، به‌دست آمده است. برای مقایسهٔ $H_n^{-۱}$ و $\overline{H}_n^{-۱}$، ماتریس $\overline{H}_n$ را، با افزودن صفرهایی به‌جزء اعشاری هر درایه، به یک ماتریس با دقت مضاعف بدل کنید. سپس از یک برنامهٔ رایانه‌یی وارون‌سازی ماتریس با دقّت مضاعف استفاده کنید تا $\overline{H}_n^{-۱}$ به‌طور عددی محاسبه شود. این کار یک مقدار دقیق برای $\overline{H}_n^{-۱}$ با دقت ساده، برای مقادیر کوچک n، به‌دست خواهد داد. پس از به‌دست آوردن $\overline{H}_n^{-۱}$، آن را با $H_n^{-۱}$ که در (۱۳.۴.۸) یا گرگوری و کارنی (۱۹۶۹، صص ۳۴ـ۳۷) داده‌شده مقایسه نمایید.

۲۳. با استفاده از برنامه‌های مسألهٔ ۳ یا برنامه‌های SGECO و SGESL از LINPACK، $\overline{H}_n x = b$ را برای چندین مقدار n حل کنید. از $b = [۱, -۱, ۱, -۱, \dots]^T$ استفاده کنید و جواب درست را با به‌کاربردن $\overline{H}_n^{-۱}$ از مسألهٔ ۲۲ محاسبه کنید. دربارهٔ نتایج خود اظهارنظر کنید.

۲۴. با استفاده از روش تصحیح مانده، که در ابتدای بخش ۵.۸ توضیح داده شد، جوابهای دستگاههای خطی مسألهٔ ۲۳ را با استفاده از حساب با دقّت ساده به‌دست آورید. مانده‌ها و تصحیحات را حساب کنید، نرخ کاهش در جملات تصحیح را وقتی n افزایش می‌یابد بررسی کنید. کوشش کنید نتایج خود را توضیح دهید.

۲۵. دستگاه خطی مسألهٔ ۴ را برای حل تقریبی یک معادلهٔ انتگرالی درنظر می‌گیریم. گاهی اوقات می‌خواهیم چنین دستگاهی را به‌ازای چندین مقدار λ که بهم نزدیک هستند حل کنیم. برنامه‌ای بنویسید که ابتدا دستگاه را به‌ازای $\lambda_0 = 0.4$ حل و سپس تجزیهٔ $\lambda_0 I - K$ به LU را ذخیره کند. برای حل $(\lambda I - K)x = b$ با مقادیر دیگر λ نزدیک به λ_0 روش تصحیح ماندهٔ (۳.۵.۸) را با $C = [LU]^{-1}$ به‌کار برید. برای مثال، این دستگاه را به‌ازای $\lambda = 0.4, 0.5, 0.5, 0.1$ حل کنید. در هر حالت بارستها را چاپ کرده و نسبت در (۱۱.۵.۸) را محاسبه کنید. در رفتار بارستها وقتی λ افزایش می‌یابد اظهارنظر کنید.

۲۶. گیریم دستگاه $Ax = b$ با

$$A = \begin{bmatrix} 4 & -1 & 0 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 \\ -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 0 & -1 & 4 \end{bmatrix} \qquad b = \begin{bmatrix} 2 \\ 1 \\ 2 \\ 2 \\ 1 \\ 2 \end{bmatrix}$$

دارای جواب $x = [1, 1, 1, 1, 1, 1]^T$ باشد. دستگاه را با استفاده از روش بارستی گاوس ـ ژاکوبی و سپس با روش گاوس ـ زایدل حل کنید. حدس اولیهٔ $x^{(0)} = 0$ را به‌کار برید. به نرخ کاهش خطای بارستی توجه کنید. جواب را با دقت $\varepsilon = 0.0001$ پیدا کنید.

۲۷. گیریم A و B از مرتبهٔ n باشند و A ناتکین باشد. حل دستگاه زیر را در نظر بگیرید

$$Az_1 + Bz_2 = b_1 \qquad Bz_1 + Az_2 = b_2$$

با $z_1, z_2, b_1, b_2 \in \mathbf{R}^n$.

(الف) شرایط لازم و کافی برای همگرایی روش بارستی زیر را پیدا کنید

$$Az_1^{(m+1)} = b_1 - Bz_2^{(m)} \qquad Az_2^{(m+1)} = b_2 - Bz_1^{(m)} \qquad m \geq 0$$

(ب) قسمت (الف) را برای روش بارستی زیر تکرار کنید

$$Az_1^{(m+1)} = b_1 - Bz_2^{(m)} \qquad Az_2^{(m+1)} = b_2 - Bz_1^{(m+1)} \qquad m \geq 0$$

نرخهای همگرایی دو روش را با هم مقایسه نمایید.

۲۸. برای معادلهٔ خطای (۲۵.۶.۸) نشان دهید که $r_\sigma(M) < ۱$، اگر برای نرم ماتریسی مناسبی، رابطهٔ

$$\| P \| < \frac{۱}{۲ \| A^{-۱} \|}$$

برقرار باشد.

۲۹. برای بارست دستگاه سه‌قطری بلوکی، که در (۳۰.۶.۸) داده شده است، همگرایی را تحت مفروضات زیر ثابت کنید.

$$\| C_۱ \| < \frac{۱}{\| B_۱^{-۱} \|}; \quad \| A_i \| + \| C_i \| < \frac{۱}{\| B_i^{-۱} \|},$$
$$۲ \le i \le r-۱; \quad \| A_r \| < \frac{۱}{\| B_r^{-۱} \|}$$

کران نرخ همگرایی را پیدا کنید.

۳۰. ماتریس A_n از مسألهٔ ۱۵ را به خاطر آورید و دستگاه خطی $A_n x = b$ را درنظر بگیرید. این دستگاه از این نظر مهم است که در تقریب استانداردِ تفاضلات متناهی (۳۰.۱۱.۶) در مسألهٔ مقدار مرزی دو نقطه‌یی ظاهر می‌شود

$$y''(x) = f(x, y(x)) \qquad \alpha < x < \beta \qquad y(\alpha) = a_۰ \qquad y(\beta) = a_۱$$

این دستگاه همچنین به لحاظ ظاهرشدن در تحلیل روشهای بارستی برای حل گسسته‌سازی معادلهٔ پواسن به صورت (۵.۸.۸) نیز اهمیت دارد. در این راستا، روش ژاکوبی را برای حل $A_n x = b$ به‌طور بارستی درنظر بگیرید. همگرایی روش ژاکوبی را با نشان‌دادن $r_\sigma(M) < ۱$ برای ماتریس مناسب M، نشان دهید.

راهنمایی: نتایج مسألهٔ ۶ از فصل ۷ را به‌کار برید.

۳۱. روشهای بارستی همگرا ممکن است رفتار غیرعادی داشته باشند، به‌عنوان مثال دستگاه زیر را در نظر بگیرید

$$x^{(k+۱)} = b + Ax^{(k)} \qquad k \ge ۰$$

با

$$A = \begin{bmatrix} \lambda & c \\ ۰ & -\lambda \end{bmatrix}, \qquad b \in \mathbf{R}^۲$$

با فرض $|\lambda| < ۱$، می‌دانیم که $(I-A)^{-۱}$ موجود، و برای تمام مقادیر اولیهٔ $x^{(۰)}$، $x^{(k)} \to x^* = (I-A)^{-۱}b$. فرمولهای صریحی برای A^k، $x^* - x^{(k)}$ و $x^{(k+۱)} - x^{(k)}$ پیدا کنید. با وفق‌دادن c با λ نشان دهید که امکان دارد $\| x^* - x^{(k)} \|_\infty$ هنگام همگرایی به صفر، به‌طور تناوبی افزایش و کاهش یابد. به مقادیر متناظر $\| x^{(k+۱)} - x^{(k)} \|_\infty$ توجه کنید. برای سادگی در تمام محاسبات $x^{(۰)} = ۰$ را اختیار کنید.

۳۲. (الف) گیریم $C_۰$ یک وارون تقریبی A باشد. تعریف می‌کنیم، $R_۰ = I - AC_۰$ و فرض می‌کنیم $\| R_۰ \| < ۱$. روش بارستی زیر را تعریف می‌کنیم

$$C_{m+۱} = C_m(I + R_m) \qquad R_{m+۱} = I - AC_{m+۱} \qquad m \geq ۰$$

که روش بارستی معروفی برای محاسبهٔ وارون $A^{-۱}$ است. همگرایی C_m به $A^{-۱}$ را ابتدا با مربوط‌ساختن خطای $A^{-۱} - C_m$ به ماندهٔ R_m، نشان دهید. و سپس رفتار ماندهٔ R_m را با نشان‌دادن $R_{m+۱} = R_m^۲$، $m \geq ۰$، بررسی کنید.

(ب) C_m را به بسط زیر مربوط سازید

$$A^{-۱} = C_۰(I - R_۰)^{-۱} = C_۰ \sum_{j=۰}^{\infty} R_۰^j$$

به رابطهٔ این روش وارون‌کردن A با روش بارستی (۶.۰.۲) از فصل ۲ در محاسبهٔ $۱/a$ به‌ازای مقادیر ناصفر a توجه کنید. همچنین مسألهٔ ۱، فصل ۲، را ملاحظه نمایید.

۳۳. برنامه‌های حل بارستی گسسته‌سازی (۵.۸.۸) معادلهٔ پواسون بر مربع واحد را اجرا کنید. برای آنکه وضعیتی داشته باشید که در آن یک جواب درست دستگاه خطی در دست باشد، معادله‌های پواسنی را انتخاب کنید که در آنها هیچ خطای گسسته‌سازی در رفتن به (۵.۸.۸) وجود نداشته باشد. این مطلب زمانی درست خواهد بود که خطاهای برشی در (۳.۸.۸) متحداً برابر صفر باشند، مثلاً $u(x,y) = x^۲y^۲$.

(الف) (۵.۸.۸) را با روش ژاکوبی حل کنید. به خطای واقعی $\| x - x^{(\nu)} \|_\infty$ در هر بارست و همچنین به $\| x^{(\nu+۱)} - x^{(\nu)} \|_\infty$ توجه نمایید. ثابت c در (۲.۷.۸) را براورد نمایید و کران خطای براوردشدهٔ (۵.۷.۸) را محاسبه کنید. نتیجه را با خطای واقعی بارستی مقایسه نمایید.

(ب) مسأله را با روش گاوس ـ زایدل مجدداً حل کنید. همچنین نرخ بارستی c را با مقدار براوردشدهٔ (۱۳.۸.۸) مقایسه کنید.

(ج) روش SOR را به‌کار برید، از پارامتر شتاب اپتیمال ω^* در (۱۵.۸.۸) استفاده کنید.

۳۴. (الف) گسسته‌سازی معادلهٔ پواسن در (۵.۹.۸) را برای معادلهٔ زیر تعمیم دهید

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - c(x,y)u = g(x,y) \qquad 0 < x, y < 1$$

با $u = f(x,y)$ که مانند قبل روی مرز، اختیار شده است.

(ب) فرض می‌کنیم $c(x,y) \geq 0$، $0 \leq x, y \leq 1$. قسمت (الف) از برهان قضیهٔ ۸.۸ را تعمیم داده نشان دهید که دستگاه خطی قسمت (الف) دارای جواب یکتاست.

۳۵. الگوریتم گرادیان مزدوج بخش ۹.۸ را پیاده کنید. آن را برای دستگاههای مسائل ۱ و ۳ و ۴ امتحان کنید. هرگاه ممکن باشد، به منظور امتحان‌کردن، دستگاههایی با جواب درست معلوم را به‌کار برید. با استفاده از آن، خطاهای واقعی را در هر بارست محاسبه و بررسی کنید که با چه سرعتی کاهش می‌یابند. برای دستگاه خطی در مسألهٔ ۴، که بر پایهٔ حل یک معادلهٔ انتگرالی حاصل شده است، دستگاه را برای چندین مقدار n حل کنید. دربارهٔ نتایج خود اظهارنظر کنید.

۳۶. نرم برداری $\| x \|_A$ ی از (۷.۹.۸) را که A متقارن و معین مثبت است به یاد آورید. گیریم ویژه‌مقدارهای A چنین نشان داده شده باشند:

$$0 < \lambda_1 \leq \lambda_2 \ldots \leq \lambda_n$$

نشان دهید که

$$\sqrt{\lambda_1} \| x \|_2 \leq \| x \|_A \leq \sqrt{\lambda_n} \| x \|_2$$

که هر دو تساوی با انتخابهای مناسبی از x به‌دست می‌آیند.

راهنمایی: از یک پایهٔ یکا متعامد از ویژه بردارهای A استفاده کنید.

۳۷. لم ۱ را که پس از (۱۱.۹.۸) آمده است ثابت کنید.

راهنمایی: از استقرای ریاضی روی k استفاده کنید. آن را برای $k = 1$ ثابت و سپس فرض کنید که برای $k \leq l$ درست است و آن را برای $k = l + 1$ ثابت کنید. اثبات را به دو بخش بشکنید. (۱): $p_i^T r_{l+1} = 0$ برای $i \leq l$، و (۲) $p_{l+1}^T r_{l+1} = 0$.

۳۸. گیریم A متقارن، معین مثبت و از مرتبهٔ $n \times n$ باشد. فرض می‌کنیم $U = \{u_1, \ldots, u_n\}$ یک مجموعه از بردارهای ناصفر در R^n باشد. در این صورت اگر U هم یک مجموعهٔ متعامد و هم یک مجموعهٔ A ـ متعامد باشد، آنگاه برای مقادیر مناسب $\lambda_i > 0$، $i = 1, \ldots, n$، $Au_i = \lambda u_i$. برعکس، همیشه می‌توان یک مجموعه از ویژه‌بردارهای $\{u_1, \ldots, u_n\}$ از A را انتخاب نمود که هم متعامد و هم A ـ متعامد باشند.

۳۹. گیریم A متقارن، معین مثبت و از مرتبهٔ n باشد. فرض می‌کنیم $\{v_1, \ldots, v_n\}$ یک

مجموعهٔ A ـ متعامد در $\mathbf{R}^n$ باشد که همهٔ v_i ها مخالف صفرند. تعریف می‌کنیم

$$Q_j = \frac{v_j v_j^T A}{v_j^T A v_j} \qquad j = ۱, \ldots, n$$

که ویژگیهای زیر را برای Q_j نشان می‌دهد.

۱. $Q_j v_i = ۰$ اگر $i \neq j$؛ و $Q_j v_j = v_j$

۲. $Q_j^۲ = Q_j$

۳. $(x, Q_j y)_A = (Q_j x, y)_A$ برای تمام مقادیر $x, y \in \mathbf{R}^n$

۴. $(Q_j x, (I - Q_j) y)_A = ۰$ برای تمام مقادیر $x, y \in \mathbf{R}^n$

۵. $\| x \|_A^۲ = \| Q_j x \|_A^۲ + \| (I - Q_j) x \|_A^۲$ برای تمام مقادیر $x \in \mathbf{R}^n$

ویژگیهای ۵.۲ می‌گویند که Q_j یک تصویر قائم در فضای برداری $\mathbf{R}^n$ با ضرب داخلی $(۰,۰)_A$ است. تعریف می‌کنیم

$$S_k = \mathrm{Span}\{v_۱, \ldots, v_k\}$$

نشان دهید که جواب مسألهٔ مینیمم‌سازی

$$\operatorname*{Min}_{y \in S_k} \| x - y \|_A$$

با رابطهٔ زیر داده می‌شود

$$y = \left[\sum_{j=۱}^{k} Q_j \right] x \equiv P_k x$$

ماتریس P_k در ویژگیهای ۵.۲ نیز صدق می‌کند.

۹

مسألهٔ ویژه‌مقدار ماتریس

ما در این فصل مسألهٔ محاسبهٔ ویژه‌مقدارها و ویژه‌بردارهای یک ماتریس مربعی را بررسی می‌کنیم. این مسأله در شماری از زمینه‌ها پیش می‌آید و ماتریسهای حاصل ممکن است شکلهای مختلف داشته باشند. این ماتریسها ممکن است تُنک یا فشرده باشند، ممکن است مرتبه و ساختارهای بسیار متغیّر داشته و اغلب متقارن باشند. به‌علاوه آنچه باید محاسبه شود ممکن است آنقدر تغییر کند که بر انتخاب روشی که باید به‌کارگرفته شود اثر بگذارد. اگر تنها محاسبهٔ چند ویژه‌مقدار موردنظر باشد، روش عددی با روشی که محاسبهٔ تمام ویژه‌مقدارها خواسته شده باشد متفاوت خواهد بود.

مسألهٔ کلی پیداکردن تمام ویژه‌مقدارها و ویژه‌بردارهای یک ماتریس نامتقارن A ممکن است نسبت به اختلالات درایه‌های A، کاملاً ناپایدار باشد، و این امر طرح روشهای کلی و برنامه‌های رایانه‌یی را مشکلتر می‌سازد. ویژه‌مقدارهای یک ماتریس متقارن A نسبت به اختلالات A کاملاً پایدارند. این موضوع، همراه با ناپایداری ممکن در ماتریسهای نامتقارن، در بخش ۱.۹ بررسی شده است. به علّت پایداری بیشتر مسألهٔ ویژه‌مقدار ماتریسهای متقارن و به لحاظ پیدایش فراوان آنها، روشهای زیادی به‌خصوص برای آنها ایجاد شده‌اند. این موضوع تأکید اصلی ما در این فصل خواهد بود، اگرچه روشهایی برای ویژه‌مقدارهای ماتریسهای نامتقارن نیز مورد بحث قرار گرفته‌اند.

معمولاً ویژه‌مقدارهای ماتریس ابتدا محاسبه می‌شوند و سپس، در صورت لزوم، از آنها در

پیداکردن ویژه‌بردارها استفاده می‌شود. استثنای عمده بر این قاعده، **روش توانی** است که در بخش ۲.۹ توضیح داده شده است و در محاسبهٔ ویژه‌مقدار غالب منفردِ یک ماتریس، مفید است. روال معمولی برای محاسبهٔ ویژه‌مقدارهای یک ماتریس A شامل دو مرحله است. ابتدا، با استفاده از تبدیلات تشابهی، A به شکل ساده‌تری، که معمولاً برای ماتریسهای متقارن، سه‌قطری است، بدل می‌شود. و سپس، این ماتریسهای ساده برای محاسبهٔ ویژه‌مقدارها و در صورت نیاز برای ویژه‌بردارها، به‌کار برده می‌شوند. شکل اصلی تبدیلات تشابهی که مورد استفاده قرار می‌گیرند ماتریسهای یکانی یا متعامد خاصی هستند، که در بخش ۳.۹ دربارهٔ آنها بحث شده است. برای محاسبهٔ ویژه‌مقدارهای ماتریس سه قطری متقارن، نظریهٔ دنباله‌های اِستورم[1] در بخش ۴.۹ معرفی شده و الگوریتم QR در بخش ۵.۹ مورد بحث قرارگرفته است. هنگامی که ویژه‌مقدارها محاسبه شده باشند، قویترین تکنیک برای محاسبهٔ ویژه‌بردارها روش **بارست معکوس** است. این روش در بخش ۶.۹ بحث و توضیح داده شده است. لازم به تذکر است که ما کلمات **متقارن و نامتقارن** را عموماً به‌کار می‌بریم، و حال آنکه معمولاً باید فقط در ارتباط با ماتریسهای حقیقی به‌کار برده شوند. برای ماتریسهای مختلط همیشه باید به‌ترتیب واژه‌های **ارمیتی** و **ناارمیتی** را جایگزین کنیم.

بیشتر روشهای عددی که در حال حاضر به‌کار برده می‌شوند از ۱۹۵۰ به بعد گسترش یافته‌اند. پیاده‌کردن آنها به‌صورت برنامه‌های رایانه‌یی چندان آسان نیست، به‌ویژه برای آن روشهایی که در ماتریسهای نامتقارن به‌کار برده می‌شوند. از نیمهٔ دههٔ ۱۹۶۰ الگوریتمهایی برای تعدادی از مسائل ویژه‌مقداری ماتریس به زبان ALGOL در مجلهٔ ریاضیات کاربردی[2] انتشار یافته‌اند. این برنامه‌ها به‌نحوی گسترده مورد آزمایش قرارگرفته و براساس آزمونها و نظریه‌های جدید اصلاح شده‌اند. این الگوریتمها درکتاب ویلکینسن و راینش (۱۹۷۱، قسمت دوّم) گردآوری شده‌اند. در یک پروژهٔ بخش ریاضیات کاربردی آزمایشگاه ملّی آرگون[3]، این برنامه‌ها به زبان فورترن ترجمه شده و آزمایشها و اصلاحات بیشتری در آنها صورت گرفته است. این بسته برنامه‌ها EISPACK نامیده می‌شود و می‌توان آنها را از آزمایشگاه ملی آرگون و منابع دیگر تهیه نمود (به پیوست مراجعه شود). شرح کامل این بسته، شامل تمام برنامه‌ها در اسمیت و همکاران (۱۹۷۶) وگاربو[4] و همکاران (۱۹۷۷) داده شده است.

۱.۹ جای ویژه‌مقدار، خطا، و قضایای پایداری

موضوع را با ارائهٔ چند قضیه برای جایابی و کراندارکردن ویژه‌مقدارهای یک ماتریس A آغاز می‌نماییم. برای کران بالا، قضیهٔ ۸.۷ از فصل ۷ را یادآوری می‌نماییم که برای هر نرم ماتریسی،

$$\max_{\lambda \in \sigma(A)} |\lambda| \leq \| A \| \qquad (۱.۱.۹)$$

1. Sturm 2. Mathematik Numerische 3. Argonne 4. Garbow

نماد $\sigma(A)$ معرّف مجموعهٔ تمام ویژه‌مقدارهای A است. قضیهٔ بعدی یک روش محاسبه‌ای ساده در براوردهای بهتر جای ویژه‌مقدارهای A به‌دست می‌دهد.

برای $A = [a_{ij}]$ از مرتبه n، تعریف می‌کنیم

$$r_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} | \qquad i = 1, 2, \ldots, n \tag{۲.۱.۹}$$

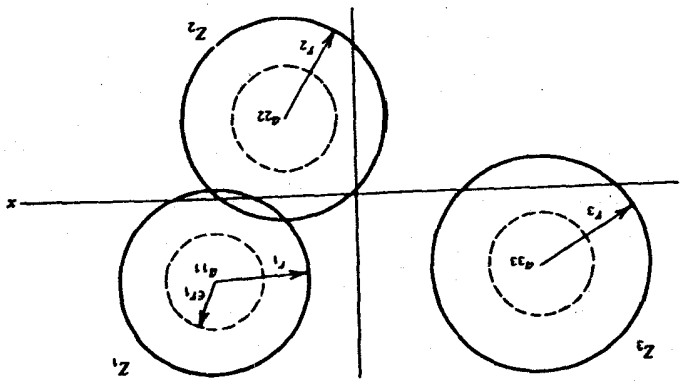
و گیریم Z_i معرّف دایره‌ای به مرکز a_{ii} و شعاع r_i در صفحهٔ مختلط باشد:

$$Z_i = \{z \in \mathbf{C} | \; | z - a_{ii} | \leq r_i\} \tag{۳.۱.۹}$$

قضیهٔ ۱.۹ (گرشگورین[1]) گیریم A دارای مرتبهٔ n باشد و λ یک ویژه‌مقدار A. آنگاه λ به یکی از دایره‌های Z_i تعلق دارد. به‌علاوه اگر m تا از این دایره‌ها یک مجموعهٔ همبند S تشکیل دهند که از $n - m$ دایرهٔ باقیمانده مجزّا باشند، آنگاه S دقیقاً m ویژه‌مقدار A را در بر دارد که برحسب چندگانی‌بودن ریشه‌های چندجمله‌یی مشخصهٔ A شمارش شده‌اند.

چون A و A^T ویژه‌مقدارها و چندجمله‌ییهای مشخصهٔ واحدی دارند، این قضایا، اگر برای تعریف شعاعها در (۲.۱.۹)، از مجموعیابی روی ستونها به جای سطرها، استفاده شود، نیز معتبرند.

برهان شکل ۱.۹ تصویری است از صفحهٔ مختلط که چگونگی دوایر را برای یک ماتریس مختلط از مرتبهٔ ۳ نشان می‌دهد. دوایر با خط پر آنهایی هستند که با (۳.۱.۹) داده شده‌اند و دوایر خط‌چین



شکل ۱.۹ مثال قضیهٔ دایرهٔ گرشگورین

1. Gerschgorin

بعداً در برهان پیش می‌آیند. به موجب این قضیه، باید یک ویژه‌مقدار در $Z_۳$ باشد و دو ویژه‌مقدار در اجتماع $Z_۱$ و $Z_۲$.

گیریم λ یک ویژه‌مقدار A باشد و x ویژه‌بردار متناظر با آن. فرض می‌کنیم k، زیرنمایهٔ یک مؤلفهٔ x باشد که برای آن

$$| x_k | = \max_{۱\le i\le n} | x_i | = \| x \|_\infty$$

در این صورت از مؤلفهٔ kام رابطهٔ $Ax = \lambda x$ نتیجه می‌شود

$$\sum_{j=۱}^{n} a_{kj}x_j = \lambda x_k$$

$$(\lambda - a_{kk})x_k = \sum_{\substack{j=۱\\ j\neq k}}^{n} a_{kj}x_j$$

$$| \lambda - a_{kk} |\, | x_k | \le \sum_{\substack{j=۱\\ j\neq k}}^{n} | a_{kj} |\, | x_j | \le r_k \| x \|_\infty$$

با حذف $\| x \|_\infty$ از طرفین، اولین قسمت قضیه اثبات می‌شود.

تعریف می‌کنیم

$$D = \operatorname{diag}[a_{۱۱}, a_{۲۲}, \dots, a_{nn}] \qquad E = A - D$$

برای $۰ \le \epsilon \le ۱$ تعریف می‌کنیم

$$A(\epsilon) = D + \epsilon E \tag{۴.۱.۹}$$

و ویژه‌مقدارهای آن را با $\lambda_۱(\epsilon)$، ...، $\lambda_n(\epsilon)$ نشان می‌دهیم. توجه نمایید که $A(۱) = A$ که همان ماتریس اصلی است. ویژه‌مقدارها ریشه‌های چندجمله‌یی مشخصهٔ

$$f_\epsilon(\lambda) \equiv \det[A(\epsilon) - \lambda I]$$

هستند. چون ضرایب $f_\epsilon(\lambda)$ توابع پیوسته از ϵ هستند و چون ریشه‌های هر چندجمله‌یی، توابعی پیوسته از ضرایب آن هستند [هنریچی (۱۹۷۴)، ص ۲۸۱ را ببینید]، پس $\lambda_n(\epsilon), \dots, \lambda_۱(\epsilon)$ توابعی پیوسته از ϵ هستند. اگر پارامتر ϵ تغییر کند هریک از ویژه‌مقدارها مانند $\lambda_i(\epsilon)$ در صفحهٔ مختلط تغییر می‌نماید و راهی از $\lambda_i(۰)$ به $\lambda_i(۱)$ رسم می‌کند.

با توجّه به قسمت اول قضیه، می‌دانیم که ویژه‌مقدارهای $\lambda_i(\epsilon)$ در دایره‌های

$$Z_i(\epsilon) = \{z \in \mathbf{C} \mid\mid z - a_{ii} \mid \leq \epsilon r_i\} \qquad i = ۱, \ldots, n \qquad (۵.۱.۹)$$

جا دارند، r_i مانند قبل با (۲.۱.۹) تعریف می‌شود. مثالهای این دایره‌ها در شکل ۱.۹ با دایره‌های خط‌چین داده شده‌اند. این دایره‌ها وقتی ϵ از ۱ به ۰ می‌رود کاهش می‌یابند و ویژه‌مقدارهای $\lambda_i(\epsilon)$ باید در داخل این دایره‌ها باقی بمانند. وقتی $\epsilon = ۰$، ویژه‌مقدارها

$$\lambda_i(۰) = a_{ii}$$

خواهند شد. با توجه به مسیر $\lambda_i(\epsilon)$، $۰ \leq \epsilon \leq ۱$، این راه که از $\epsilon = ۰$ شروع می‌شود باید در دایرهٔ $Z_i(۱)$ باقی بماند. بنابراین اگر m تا از این دایره‌های $Z_i(۱)$ یک مجموعهٔ همبند S، مجزا از $n - m$ دایرهٔ باقیمانده، تشکیل دهند، آنگاه S باید دقیقاً شامل m ویژه‌مقدار $\lambda_i(۱)$ باشد، زیرا شامل m ویژه‌مقدار $\lambda_i(۰)$ است. و این قسمت دوّم قضیه را اثبات می‌کند.

از آنجا که

$$\det[A - \lambda I] = \det[A - \lambda I]^T = \det[A^T - \lambda I]$$

خواهیم داشت $\sigma(A) = \sigma(A^T)$. لذا این قضیه را برای سطرهای A^T به‌کار می‌بریم تا آن را برای ستونهای A اثبات نماییم؛ و اثبات قضیه کامل می‌شود. ■

از این قضیه می‌توان به راههای مختلف استفاده کرد، ما فقط دو مثال عددی ساده ذکر می‌کنیم.

مثال ماتریس

$$A = \begin{bmatrix} ۴ & ۱ & ۰ \\ ۱ & ۰ & -۱ \\ ۱ & ۱ & -۴ \end{bmatrix}$$

را درنظر می‌گیریم. مطابق با قضیهٔ قبل، ویژه‌مقدارها باید در دایره‌های

$$\mid \lambda - ۴ \mid \leq ۱ \qquad \mid \lambda \mid \leq ۲ \qquad \mid \lambda + ۴ \mid \leq ۲ \qquad (۶.۱.۹)$$

قرار داشته باشند. چون دایرهٔ اول از بقیه مجزاست، باید یک ریشهٔ تنها در این دایره واقع باشد. چون ضرایب

$$f(\lambda) = \det \mid A - \lambda I \mid$$

حقیقی‌اند، ویژه‌مقدارهای مختلط، اگر وجود داشته باشند، باید جفتهای مزدوج باشند. و این مطلب، با استفاده از (۶.۱.۹)، به‌سادگی ایجاب می‌کند که یک ویژه‌مقدار حقیقی در بازهٔ $[۳,۵]$ وجود داشته باشد. دو دایرهٔ باقیمانده در تنها نقطهٔ $(۰، ۲-)$ برهم مماس می‌شوند. با استفاده از برهان قبلی، ویژه‌مقدارهای واقع در این دو دایره باید حقیقی باشند. و با توجه به تعریف $A(\epsilon)$، $\epsilon < ۱$، در (۴.۱.۹) یک ویژه‌مقدار در $[-۲, -۶]$ و یکی در $[۲, -۲]$ وجود دارد.چون به‌آسانی دیده می‌شود که $\lambda = -۲$ یک ویژه‌مقدار نیست، می‌توانیم نتیجه بگیریم که A در هریک از بازه‌های،

$$[-۶, -۲), (-۲, ۲], [۳, ۵]$$

یک ویژه‌مقدار دارد. ویژه‌مقدارهای واقعی عبارت‌اند از:

$$-۳٫۷۶۰۱۰, -۰٫۴۴۲۹۳۱, ۴٫۲۰۳۰۳$$

مثال ماتریس مرتبهٔ n زیر را درنظر می‌گیریم

$$A = \begin{bmatrix} ۴ & ۱ & ۰ & & & \cdots & ۰ \\ ۱ & ۴ & ۱ & ۰ & & & \\ ۰ & ۱ & ۴ & ۱ & & & \vdots \\ \vdots & & & & \ddots & & \\ & & & & ۱ & ۴ & ۱ \\ ۰ & \cdots & & & ۰ & ۱ & ۴ \end{bmatrix} \qquad (۷.۱.۹)$$

چون A متقارن است، همهٔ ویژه‌مقدارهای آن حقیقی‌اند. شعاعهای r_iی (۲.۱.۹) همگی یا ۱ هستند یا ۲، و مرکز تمام دایره‌ها $a_{ii} = ۴$ است. بنابراین به موجب قضیهٔ قبل، ویژه‌مقدارها باید همگی در بازهٔ $[۲,۶]$ واقع باشند. چون ویژه‌مقدارهای $A^{-۱}$ عکس ویژه‌مقدارهای A هستند، به‌ازای تمام این ویژه‌مقدارهای $A^{-۱}$ که با μ نمایش می‌دهیم، باید داشته باشیم

$$\frac{۱}{۶} \leq \mu \leq \frac{۱}{۲}$$

با استفاده از نرم ماتریسی (۲۲.۳.۷) که از نرم برداری اقلیدسی به‌دست آمده است داریم:

$$\| A^{-1} \|_2 = r_\sigma(A^{-1}) \leq \frac{1}{2}$$

که مستقل از اندازهٔ n است.

کرانها برای ویژه‌مقدارهای اختلال‌یافته دریک ماتریس داده‌شدهٔ A، می‌خواهیم اختلالی ایجاد و سپس تأثیر آن را بر ویژه‌مقدارهای A مشاهده کنیم. کرانهای تحلیلی برای اختلالات روی ویژه‌مقدارها را، براساس اختلالات روی ماتریس A به‌دست می‌آوریم. این کرانها همچنین تعریفی برای ضریب وضعیت به ذهن القا می‌کنند که می‌توان آن را برای بیان درجهٔ پایداری یا ناپایداری در ویژه‌مقدارها به‌کار برد.

برای آنکه استدلال را خیلی ساده کنیم، فرض می‌کنیم که شکل متعارفِ ژوردانِ ماتریس A قطری باشد (قضیه ۶.۷ را ببینید). یعنی برای ماتریس ناتکینی چون P:

$$P^{-1}AP = \text{diag}[\lambda_1, \ldots, \lambda_n] \equiv D \qquad (۸.۱.۹)$$

ستونهای P ویژه‌بردارهای A متناظر با ویژه‌مقدارهای λ_1، λ_2، ...، λ_n هستند. ماتریسهایی که در (۸.۱.۹) صدق می‌کنند مهمترین حالتها در عمل هستند. برای یک بحث مختصر دربارهٔ حالتی که در آن شکل متعارف ژوردان قطری نباشد، مبحث آخر این بخش را ملاحظه کنید.

همچنین نیازمندیم که ویژگی خاصّی را برای نرمهای ماتریسی که به‌کاربرده می‌شوند بپذیریم. برای هر ماتریس قطری

$$G = \text{diag}[g_1, \ldots, g_n]$$

باید داشته باشیم

$$\| G \| = \underset{1 \leq i \leq n}{\text{Max}} \mid g_i \mid \qquad (۹.۱.۹)$$

کلیهٔ نرم‌های ماتریس عملگر القاشده از نرمهای برداری $\| x \|_p$، $1 \leq p \leq \infty$، این ویژگی را دارا هستند. اکنون می‌توانیم قضیّهٔ زیر را بیان کنیم.

قضیهٔ ۲.۹ (باوئر ـ فیکه[1]) گیریم A یک ماتریس با شکل متعارف قطری ژوردان مانند (۸.۱.۹) باشد. و فرض می‌کنیم که نرم ماتریس در (۹.۱.۹) صدق کند. گیریم $A + E$ اختلال‌یافتهٔ A باشد و λ یک ویژه‌مقدار $A + E$. در این صورت

$$\underset{1 \leq i \leq n}{\text{Min}} \mid \lambda - \lambda_i \mid \leq \| P \| \| P^{-1} \| \| E \| \qquad (۱۰.۱.۹)$$

1. Bauer-Fike

برهان اگر λ ویژه‌مقدار A نیز باشد، آنگاه (۱.۱.۹) بالبداهه درست است. پس فرض می‌کنیم $\lambda \neq \lambda_1, \lambda_2 \dots, \lambda_n$، و x یک ویژه‌بردار $A+E$ متناظر با λ باشد. در این صورت

$$(A+E)x = \lambda x$$
$$(\lambda I - A)x = Ex$$

A را از (۸.۱.۹) به‌دست آورده در رابطهٔ اخیر می‌گذاریم

$$(\lambda I - PDP^{-1})x = Ex$$
$$(\lambda I - D)(P^{-1}x) = (P^{-1}EP)(P^{-1}x)$$

چون $\lambda \neq \lambda_1, \lambda_2, \dots, \lambda_n$، $\lambda I - D$ ناتکین است،

$$(\lambda I - D)^{-1} = \text{diag}[(\lambda - \lambda_1)^{-1}, \dots, (\lambda - \lambda_n)^{-1}]$$

پس

$$P^{-1}x = (\lambda I - D)^{-1}(P^{-1}EP)(P^{-1}x)$$
$$\| P^{-1}x \| \leq \| (\lambda I - D)^{-1} \| \| P^{-1}EP \| \| P^{-1}x \|$$

با حذف $\| P^{-1}x \|$ و استفاده از (۹.۱.۹)

$$1 \leq [\text{Max} \mid \lambda - \lambda_i \mid^{-1}] \| P^{-1} \| \| P \| \| E \|$$

این رابطه هم‌ارز با (۱۰.۱.۹) بوده و برهان کامل می‌شود. ■

فرع اگر A ارمیتی باشد، و اگر $A+E$ یک شکل اختلال‌یافتهٔ A باشد، آنگاه برای هر ویژه‌مقدار λ ی $A+E$،

$$\underset{1 \leq i \leq n}{\text{Min}} \mid \lambda - \lambda_i \mid \leq \| E \|_2 \qquad (۱۱.۱.۹)$$

■

برهان چون A ارمیتی است ماتریس P را می‌توان به گونه‌ای انتخاب کرد که یکانی باشد. و با استفاده از نرم عملگر (۱۹.۳.۷) که از نرم برداری اقلیدسی القا شده، $\| P \|_2 = \| P^{-1} \|_2 = 1$ (به مثال ۱۳ فصل ۷ نگاه کنید). و در اینجا برهان کامل می‌شود. ■

حکم (۱۱.۱.۹) ثابت می‌کند که اختلالهای کوچک یک ماتریس ارمیتی، چنانچه در مقدمهٔ این فصل گفته شده بود، به اختلالهای کوچک ویژه‌مقدارها می‌انجامد. باید توجه داشت که خطای نسبی در بعضی یا همهٔ ویژه‌مقدارها ممکن است بزرگ باشد، و این حالت معمولاً زمانی رخ می‌دهد که اندازه‌های ویژه‌مقدارهای ماتریس خیلی زیاد تغییر کنند.

مثال ماتریس هیلبرت از مرتبهٔ ۳ را درنظر می‌گیریم،

$$H_3 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix} \qquad (۱۲.۱.۹)$$

ویژه‌مقدارهای آن با هفت رقم بامعنی عبارت‌اند از:

$$\lambda_1 = ۱٫۴۰۸۳۱۹ \quad \lambda_2 = ۰٫۱۲۲۳۲۷۱ \quad \lambda_3 = ۰٫۰۰۲۶۸۷۳۴۰ \qquad (۱۳.۱.۹)$$

اکنون ماتریس اختلال‌یافتهٔ $\hat{H}_3$ را که معرّف ماتریس H_3 تا چهار رقم بامعنی است درنظر می‌گیریم:

$$\hat{H}_3 = \begin{bmatrix} ۱٫۰۰۰ & ۰٫۵۰۰۰ & ۰٫۳۳۳۳ \\ ۰٫۵۰۰۰ & ۰٫۳۳۳۳ & ۰٫۲۵۰۰ \\ ۰٫۳۳۳۳ & ۰٫۲۵۰۰ & ۰٫۲۰۰۰ \end{bmatrix} \qquad (۱۴.۱.۹)$$

ویژه‌مقدارهای آن تا ۷ رقم بامعنی عبارت‌اند از:

$$\hat{\lambda}_1 = ۱٫۴۰۸۲۹۴ \quad \hat{\lambda}_2 = ۰٫۱۲۲۳۴۱۵ \quad \hat{\lambda}_3 = ۰٫۰۰۲۶۶۴۴۸۹ \qquad (۱۵.۱.۹)$$

برای تحقیق در صحت (۱۱.۱.۹) برای این حالت، به‌سادگی می‌توان مقدار

$$\| E \|_2 = r_\sigma(E) = \frac{1}{3} \times ۱۰^{-۴} \doteq ۰٫۰۰۰۰۳۳$$

را محاسبه نمود. برای خطاها و خطاهای نسبی در (۱۵.۱.۹)،

$$\lambda_1 - \hat{\lambda}_1 = ۰٫۰۰۰۰۲۴۹ \qquad \mathrm{Rel}(\hat{\lambda}_1) = ۰٫۰۰۰۰۱۷۷$$
$$\lambda_2 - \hat{\lambda}_2 = -۰٫۰۰۰۰۱۴۴ \qquad \mathrm{Rel}(\hat{\lambda}_2) = -۰٫۰۰۰۱۱۸$$
$$\lambda_3 - \hat{\lambda}_3 = ۰٫۰۰۰۰۲۲۹ \qquad \mathrm{Rel}(\hat{\lambda}_3) = ۰٫۰۰۸۵$$

همهٔ خطاها در (۱۱.۱.۹) صدق می‌کنند. ولی خطای نسبی $\hat{\lambda}_۳$ در مقایسه با اختلالات نسبی $\hat{H}_۳$ کاملاً چشمگیر است.

برای یک ماتریس نامتقارن A با P یی که در (۸.۱.۹) داده‌شده، عدد

$$K(A) = \| P \| \| P^{-۱} \|$$

ضریب وضعیت مسألهٔ ویژه‌مقدار برای A خوانده می‌شود. این تعریف براساس کران (۱۰.۱.۹) برای اختلالات ویژه‌مقدارهای ماتریس اختلال‌یافتهٔ A داده شده است. انتخاب دیگر، که حتی از نظر محاسبه مشکلتر است، استفاده از

$$K(A) = \text{Inf} \| P \| \| P^{-۱} \| \qquad (۱۶.۱.۹)$$

است که در آن اینفیموم روی تمام‌ماتریسهای P، که در رابطهٔ (۸.۱.۹) صدق می‌کنند، و تمام نرمهای ماتریسی که در رابطهٔ (۹.۱.۹) صدق می‌کنند، محاسبه شده است. دلیل درنظرگرفتن ضرایب وضعیت این است که در ماتریسهای نامتقارن A، اختلالات کوچک E ممکن است به اختلالات نسبی بزرگی در ویژه‌مقدارهای A منجر شوند.

مثال برای روشن‌کردن مسائل آسیب‌شناختی که با ماتریسهای نامتقارن ظاهر می‌شوند، درنظر می‌گیریم:

$$A = \begin{bmatrix} ۱۰۱ & -۹۰ \\ ۱۱۰ & -۹۸ \end{bmatrix} \qquad A + E = \begin{bmatrix} ۱۰۱ - \epsilon & -۹۰ - \epsilon \\ ۱۱۰ & -۹۸ \end{bmatrix} \qquad (۱۷.۱.۹)$$

ویژه‌مقدارهای A برابر $\lambda = ۱, ۲$ و ویژه‌مقدارهای $A + E$:

$$\lambda = \frac{۳ - \epsilon \pm \sqrt{۱ - ۸۲۸\epsilon + \epsilon^۲}}{۲}$$

هستند. برای درک بهتر به‌عنوان مثال ϵ را ۰٫۰۰۱ انتخاب می‌کنیم. پس

$$A + E = \begin{bmatrix} ۱۰۰٫۹۹۹ & -۹۰٫۰۰۱ \\ ۱۱۰ & -۹۸ \end{bmatrix}$$

و ویژه‌مقدارها عبارت‌اند از:

$$\lambda \doteq ۱٫۲۹۸, ۱٫۷۰۱ \qquad (۱۸.۱.۹)$$

از این مسأله نباید نتیجه بگیریم که ماتریسهای نامتقارن بدوضع‌اند. بسیاری حالتها در عمل کاملاً خوش‌وضع هستند. ولی برای نوشتن یک الگوریتم کلی، ما همیشه در پی آن هستیم که هر چه ممکن است حالتهای بیشتری را پوشش دهیم، و این مثال نشان می‌دهد که این امر احتمالاً برای ردهٔ همهٔ ماتریسهای نامتقارن دشوار است.

برای ماتریسهای متقارن، قضیهٔ (۱۱.۱.۹) را به چندین راه می‌توان بهبود بخشید. یکی **توصیف مینیماکس** برای ویژه‌مقدارهای ماتریسهای متقارن است. برای بحث در این مورد وکرانهای خطای حاصله، به پارلت[1] (۱۹۸۰، بخش ۲.۱۰) یا ویلکینسن (۱۹۶۵، ص ۱۰۱) مراجعه کنید. به‌جای آن قضیّهٔ زیر را می‌دهیم که برای تحلیل خطای روشهایی که بعداً می‌آیند مفیدتر خواهد بود.

قضیهٔ ۳.۹ (ویلانت‌ـهوفمن[2]) گیریم A و E ماتریسهای حقیقی، متقارن و از مرتبهٔ n باشند و $\hat{A} = A + E$ را تعریف می‌کنیم. گیریم λ_i و $\hat{\lambda}_i$، $i = ۱, \ldots, n$، به‌ترتیب ویژه‌مقدارهای A و $\hat{A}$ باشند که به‌طور صعودی مرتب شده‌اند، در این صورت

$$\left[\sum_{j=۱}^{n}(\lambda_j - \hat{\lambda}_j)^۲\right]^{۱/۲} \leq F(E) \qquad (۱۹.۱.۹)$$

که در آن $F(E)$ نرم فروبنیوس E است که در (۱۰.۳.۷) تعریف شده بود.

برهان ویلکینسن (۱۹۶۵، صص ۱۰۴ـ۱۰۸) را ببینید. ■

این قضیّه برای کراندارکردن اثر خطاهای گردکردن ناشی از تبدیل یک ماتریس متقارن به ماتریس سه قطری بعداً مورد استفاده قرار خواهد گرفت.

یک کران خطای قابل‌محاسبه برای ماتریسهای متقارن گیریم A ماتریس متقارنی باشد که برای آن ویژه‌مقدار تقریبی λ و ویژه‌بردار تقریبی x محاسبه شده است. ماندهٔ

$$\eta = Ax - \lambda x \qquad (۲۰.۱.۹)$$

را تعریف می‌کنیم. چون A متقارن است، یک ماتریس یکانی U موجود است که برای آن

$$U^*AU = \mathrm{diag}[\lambda_۱, \ldots, \lambda_n] \equiv D \qquad (۲۱.۱.۹)$$

سپس نشان خواهیم داد که

1. Parlett 2. Weilant-Hoffman

$$\underset{1\le i\le n}{\text{Min}} \mid \lambda - \lambda_i \mid \le \frac{\| \eta \|_2}{\| x \|_2} \qquad (۲۲.۱.۹)$$

با استفاده از (۲۱.۱.۹)

$$\eta = UDU^* x - \lambda x$$

$$U^*\eta = DU^* x - \lambda U^* x = (D - \lambda I)U^* x$$

اگر λ یک ویژه‌مقدار A باشد، درستی (۲۲.۱.۹) روشن است. بنابراین اگر فرض کنیم $\lambda \neq \lambda_1, \ldots, \lambda_n$، از کلیت مسأله کاسته نخواهد شد. از این‌رو $D - \lambda I$ ناتکین بوده، و

$$U^* x = (D - \lambda I)^{-1} U^* \eta$$

$$\| U^* x \|_2 \le \| (D - \lambda I)^{-1} \|_2 \| U^* \eta \|_2$$

مسألهٔ ۱۳ فصل ۷ را به یاد می‌آوریم، که ایجاب می‌کند

$$\| U^* x \|_2 = \| x \|_2 \qquad \| U^* \eta \|_2 = \| \eta \|_2$$

پس با استفاده از تعریف نرم ماتریسی، داریم

$$\| x \|_2 \le [\underset{1\le i\le n}{\text{Max}} \mid (\lambda - \lambda_i)^{-1} \mid] \| \eta \|_2$$

که با (۲۲.۱.۹) هم‌ارز است.

فایدهٔ (۲۲.۱.۹) بعداً در (۱۵.۲.۹) بخش ۲.۹، با استفاده از یک زوج ویژه‌مقدار ـ ویژه‌بردار تقریبی حاصل از روش توانی نشان داده شده است.

پایداری ویژه‌مقدارها در ماتریسهای نامتقارن برای اینکه بتوانیم به امکان ناپایداری در مسألهٔ ویژه‌مقدار ماتریس نامتقارن به‌طور مؤثر بپردازیم، لازم است درک بهتری از ماهیّت آن ناپایداری داشته باشیم. برای مثال، یکی از نتایج تحلیل ناپایداری این است که تبدیلات تشابهی یکانی، شرایط مسأله را بدتر نخواهند کرد.

مانند قبل، فرض می‌کنیم A دارای شکل قطری متعارف ژوردان باشد:

$$P^{-1} AP = \text{diag}[\lambda_1, \ldots, \lambda_n] \equiv D \qquad (۲۳.۱.۹)$$

در این صورت λ_1، ...، λ_n ویژه‌مقداهای A بوده و ستونهای P ویژه‌بردارهای متناظر آنها هستند، که آنها را u_1، ...، u_n می‌نامیم. ماتریس P یکتا نیست. برای مثال، اگر F یک

ماتریس قطری ناتکین باشد، آنگاه

$$(PF)^{-1}A(PF) = F^{-1}DF = D$$

با انتخاب مناسب F، ستونهای PF طول واحد خواهند داشت. بنابراین، بی‌آنکه از کلیت مسأله کاسته شود، فرض می‌کنیم ستونهای P دارای طول واحد باشند:

$$Au_i = \lambda_i u_i \qquad u_i^* u_i = 1 \qquad i = 1, \ldots, n \tag{۲۴.۱.۹}$$

که

$$P = [u_1, \ldots, u_n]$$

ترانهادهٔ مزدوج (۲۳.۱.۹) را می‌گیریم

$$P^* A^* (P^*)^{-1} = D^* = \operatorname{diag}[\bar{\lambda}_1, \ldots, \bar{\lambda}_n]$$

که نشان می‌دهد ویژه‌مقدارهای A^* مزدوجهای مختلط ویژه‌مقدارهای A هستند. با نوشتن

$$(P^*)^{-1} = [w_1, \ldots, w_n] \tag{۲۵.۱.۹}$$

داریم

$$A^* w_i = \bar{\lambda}_i w_i \qquad i = 1, \ldots, n \tag{۲۶.۱.۹}$$

یا هم‌ارز آن، با تشکیل ترانهادهٔ مزدوج داریم:

$$w_i^* \mathbf{A} = \lambda_i w_i^* \tag{۲۷.۱.۹}$$

این رابطه بیان می‌کند که w_i^* **ویژه‌بردار چپ** A برای ویژه‌مقدار λ_i است. چون $P^{-1}P = I$، و از آنجا که

$$P^{-1} = \begin{bmatrix} w_1^* \\ \vdots \\ w_n^* \end{bmatrix}$$

داریم

$$w_i^* u_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{۲۸.۱.۹}$$

این رابطه بیان می‌کند که $\{u_i\}$ یعنی ویژه‌بردارهای A و $\{w_i\}$ یعنی ویژه‌بردارهای A^* یک مجموعهٔ **دو به‌دو متعامد** ایجاد می‌کنند.

با نرمالسازی ویژه‌بردارهای w_i به وسیلهٔ

$$v_i = \frac{w_i}{\| w_i \|_۲} \qquad i = ۱, \dots, n$$

عدد حقیقی و مثبت

$$s_i = v_i^* u_i = \frac{۱}{\| w_i \|_۲} \tag{۲۹.۱.۹}$$

را تعریف می‌کنیم. اکنون ماتریس $(P^*)^{-۱}$ را می‌توانیم به شکل

$$(P^*)^{-۱} = \left[\frac{v_۱}{s_۱}, \dots, \frac{v_n}{s_n}\right]$$

بنویسیم. و

$$A^* v_i = \bar{\lambda}_i v_i \qquad \| v_i \|_۲ = ۱ \qquad i = ۱, \dots, n \tag{۳۰.۱.۹}$$

اکنون پایداری یک ویژه‌مقدار سادهٔ λ_k ی A را امتحان می‌کنیم. منظور از ساده بودن این است که λ_k یک ریشهٔ چندگانهٔ چندجمله‌یی مشخصهٔ A نیست. نتایج را می‌توان برای ویژه‌مقدارهای چندگانه نیز بسط داد ولی ما کاری با آن نداریم. ماتریس اختلال‌یافتهٔ

$$A(\epsilon) = A + \epsilon B \qquad \epsilon > ۰$$

را برای ماتریس دلخواه B ی مستقل از ϵ، درنظر می‌گیریم. ویژه‌مقدارهای $A(\epsilon)$ را با $\lambda_۱(\epsilon)$، ...، $\lambda_n(\epsilon)$ نشان می‌دهیم. پس:

$$P^{-۱} A(\epsilon) P = D + \epsilon C \qquad C = P^{-۱} B P$$

$$c_{ij} = \frac{۱}{s_i} v_i^* B u_j \qquad ۱ \le i, j \le n \tag{۳۱.۱.۹}$$

ثابت می‌کنیم که

$$\lambda_k(\epsilon) = \lambda_k + \frac{\epsilon}{s_k} v_k^* B u_k + O(\epsilon^۲) \tag{۳۲.۱.۹}$$

در به‌دست آوردن این نتیجه از قضیهٔ ۱.۹ گرشگورین، استفاده می‌کنیم. همچنین لازم است توجه کنیم که برای هر ماتریس قطری ناتکین F،

$$F P^{-۱} A(\epsilon) P F^{-۱} = D + \epsilon F C F^{-۱} \tag{۳۳.۱.۹}$$

و این تبدیل ویژه‌مقدارهای $A(\epsilon)$ را عوض نمی‌کند. F را چنین انتخاب می‌کنیم:

$$f_{ii} = \begin{cases} \epsilon\alpha & i = k \\ ۱ & i \neq k \end{cases}$$

که در آن α ثابت مثبتی است که باید بعداً معین شود. بیشتر درایه‌های ماتریس (۳۳.۱.۹) عوض نشده‌اند و فقط آنهایی که روی سطر kام و ستون kام هستند باید بررسی شوند. این درایه‌ها چنین‌اند.

$$[D + \epsilon FCF^{-۱}]_{kj} = \begin{cases} \epsilon^{۲}\alpha c_{kj} & j \neq k \\ \lambda_k + \epsilon c_{kk} & j = k \end{cases}$$

$$[D + \epsilon FCF^{-۱}]_{ik} = \frac{۱}{\alpha} c_{ik} \qquad i \neq k$$

قضیهٔ ۱.۹ را برای ماتریس (۳۳.۱.۹) به‌کار می‌بریم. مراکز دوایر و شعاعهای آنها عبارت‌اند از:

$$\begin{aligned} \text{مرکز} &= \lambda_k + \epsilon c_{kk} \qquad r_k = \epsilon^{۲}\alpha \sum_{j \neq k} | c_{kj} | \\ \text{مرکز} &= \lambda_i + \epsilon c_{ii} \qquad r_i = \epsilon \sum_{j \neq i,k} | c_{ij} | + \frac{۱}{\alpha} | c_{ik} | \qquad i \neq k \end{aligned} \tag{۳۴.۱.۹}$$

می‌خواهیم α را آنقدر بزرگ و ϵ را آن اندازه کوچک بگیریم که دایرهٔ به مرکز $\lambda_k + \epsilon c_{kk}$ را از بقیهٔ دوایر جدا سازیم، و از این راه بدانیم که دقیقاً یک ویژه‌مقدار (۳۳.۱.۹) در داخل دایرهٔ k قرار دارد. فاصلهٔ بین مراکز دوایر k و $i \neq k$ دارای کران پایین

$$| \lambda_i - \lambda_k | - \epsilon | c_{ii} - c_{kk} |$$

است، که برای مقادیر کوچک ϵ، در حدود $| \lambda_i - \lambda_k |$ است. α را طوری می‌گیریم که

$$\frac{۱}{\alpha} | c_{ik} | \leq \frac{۱}{۲} | \lambda_i - \lambda_k | \qquad \text{برای تمام مقادیر } i \neq k \tag{۳۵.۱.۹}$$

سپس $\epsilon_۰$ را طوری می‌گیریم که به‌ازای جمیع مقادیر $۰ < \epsilon < \epsilon_۰$ دایرهٔ k هیچ‌یک از دوایر دیگر را قطع نکند. این کار شدنی است زیرا λ_k از ویژه‌مقدارهای دیگر λ_i متمایز است، و نابرابری (۳۵.۱.۹)، برقرار است.

با توجه به این ساختمان، قضیهٔ ۱.۹ ایجاب می‌کند که دایرهٔ k ام دقیقاً شامل یک ویژه‌مقدار $A(\epsilon)$ باشد، آن را $\lambda_k(\epsilon)$ می‌نامیم. از (۳۴.۱.۹)،

$$| \lambda_k(\epsilon) - \lambda_k - \epsilon c_{kk} | \leq r_k = O(\epsilon^۲)$$

و با به‌کارگرفتن فرمول مربوط به c_{kk} در (۳۱.۱.۹)، نتیجهٔ مطلوب (۳۲.۱.۹) اثبات می‌شود. با کراندارکردن (۳۲.۱.۹)، به‌دست می‌آوریم

$$| \lambda_k(\epsilon) - \lambda_k | \leq \frac{\epsilon}{s_k} \| v_k \|_۲ \| B \|_۲ \| u \|_۲ + O(\epsilon^۲)$$

و با استفاده از (۲۴.۱.۹) و (۳۰.۱.۹)،

$$| \lambda_k(\epsilon) - \lambda_k | \leq \frac{\epsilon}{s_k} \| B \|_۲ + O(\epsilon^۲) \qquad (۳۶.۱.۹)$$

عدد s_k، وقتی ماتریس A به اندازهٔ مقادیرکوچک $E = \epsilon B$ اختلال یابد، به پایداری ویژه‌مقدار λ_k بستگی نزدیکی پیدا می‌کند. اگر A متقارن بود، $u_k = v_k$ و بنابراین $s_k = ۱$، که همان نتیجهٔ کیفی را که قبلاً برای ماتریسهای متقارن به‌دست آوردیم، به‌دست می‌داد. برای ماتریسهای نامتقارن، اگر s_k کاملاً کوچک باشد، آنگاه اختلال کوچک $E = \epsilon B$ می‌تواند به اختلال بزرگ در ویژه‌مقدار λ_k منتهی شود. چنین مسائلی بدوضع خوانده می‌شوند.

مثال مثال (۱۷.۱.۹) را به‌یاد آورید. در این مثال $\epsilon = ۰٫۰۰۱$،

$$P^{-۱}AP = \begin{bmatrix} ۱ & ۰ \\ ۰ & ۲ \end{bmatrix} \qquad B = \begin{bmatrix} -۱ & -۱ \\ ۰ & ۰ \end{bmatrix}$$

$$P = \begin{bmatrix} \frac{۹}{\sqrt{۱۸۱}} & \frac{-۱۰}{\sqrt{۲۲۱}} \\ \frac{۱۰}{\sqrt{۱۸۱}} & \frac{-۱۱}{\sqrt{۲۲۱}} \end{bmatrix} \quad P^{-۱} = \begin{bmatrix} -۱۱\sqrt{۱۸۱} & ۱۰\sqrt{۱۸۱} \\ -۱۰\sqrt{۲۲۱} & ۹\sqrt{۲۲۱} \end{bmatrix} \qquad (۳۷.۱.۹)$$

اگر نرم سطری را برای براورد ضریب وضعیت (۱۶.۱.۹) به‌کار بریم، آنگاه

$$K(A) \leq \| P \|_\infty \| P^{-۱} \|_\infty \doteq ۴۱۹$$

$u_۱$ و $u_۲$ از ستونهای P، و بردارهای $w_۱$ و $w_۲$ از ستونهای $(P^{-۱})^T$ به‌دست می‌آیند، ((۲۵.۱.۹) را ببینید).

برای محاسبهٔ (۳۶.۱.۹)، وقتی $\lambda_1 = 1$،

$$s_1 = v_1^T u_1 = \frac{1}{\| w_1 \|_2} = \frac{1}{\sqrt{(221)(181)}} \doteq 0.005$$

و $\| B \|_2 = \sqrt{2}$. از فرمول (۳۶.۱.۹) نتیجه می‌شود

$$| \lambda_1(\epsilon) - \lambda_1 | \leq \frac{\sqrt{2}\epsilon}{0.005} + O(\epsilon^2) \doteq 283\epsilon + O(\epsilon^2) \qquad (38.1.9)$$

مقدار واقعی خطا $0.298 = 1 - 1.298 \doteq \lambda_1 - \lambda_1(0.001)$، است و براورد قبلی خطا برابر $0.283 \doteq 283\epsilon$ بود. بنابراین (۳۶.۱.۹) یک براورد کران معقول برای خطاست.

در بخشهای بعدی، برخی روشهای عددی، ابتدا توسط بعضی تبدیلات تشابهی، ماتریس A را به شکل ساده‌تری بدل می‌کنند. ما می‌خواهیم تبدیلاتی را به‌کار بریم که اعداد s_k را حتّی کوچکتر نسازند، و موجب نشوند که بدوضعی مسأله بدتر هم بشود. از این دیدگاه، بهترین تبدیلات درکاربرد، تبدیلات متعامد یا یکانی هستند.

گیریم U یکانی باشد و $\hat{A} = U^* A U$. برای یک ویژه‌مقدار سادهٔ λ_k، گیریم s_k و $\hat{s}_k$ معرّف اعداد (۲۹.۱.۹) برای دو ماتریس A و $\hat{A}$، باشند. اگر $\{u_i\}$ و $\{v_i\}$ ویژه‌بردارهای A و A^* باشند، آنگاه $\{U^* u_i\}$ و $\{U^* v_i\}$ ویژه‌بردارهای متناظر برای $\hat{A}$ و $\hat{A}^*$ می‌شوند. برای $\hat{s}_k$،

$$\begin{aligned} \hat{s}_k &= (U^* v_k)^* (U^* u_k) = v_k^* U U^* u_k = v_k^* u_k \\ &= s_k \end{aligned} \qquad (39.1.9)$$

بنابراین پایداری ویژه‌مقدار λ_k نه بهتر می‌شود و نه بدتر. تبدیلات یکانی، طول و زاویهٔ بین بردارها را نیز حفظ می‌کنند (مسألهٔ ۱۳ فصل ۷ را ملاحظه کنید). به‌طورکلی، اعمال ماتریس یکانی روی یک ماتریس داده‌شدهٔ A هیچ خللی در وضعیت مسألهٔ ویژه‌مقدار ایجاد نمی‌کند و همین یکی از دلایل اصلی این است که شکل تبدیلات تشابهی را در حل مسألهٔ ویژه‌مقدار ماتریسی ترجیح می‌دهند.

روشهای مشابه با روش قبلی در (۲۳.۱.۹) ـ (۳۶.۱.۹) را، می‌توان برای به‌دست آوردن قضیّهٔ پایداری برای ویژه‌بردارهای ویژه‌مقدارهای مجزا به‌کار برد. با استفاده از همین مفروضات دربارهٔ $\{\lambda_i\}$، $\{u_i \equiv u_i(0)\}$، و $\{v_i\}$، مسألهٔ ویژه‌بردار زیر را درنظر می‌گیریم

$$(A + \epsilon B) u_k(\epsilon) = \lambda_k(\epsilon) u_k(\epsilon) \qquad \| u_k(\epsilon) \| = 1 \qquad (40.1.9)$$

که λ_k یک ویژه‌مقدار سادهٔ A است. لذا

$$u_k(\epsilon) = u_k + \epsilon a_k u_k + \epsilon \sum_{\substack{j=1 \\ j\neq k}}^{n} \left[\frac{v_j^* B u_k}{(\lambda_k - \lambda_j)s_j} \right] u_j + O(\epsilon^2) \quad (۴۱.۱.۹)$$

برهان این رابطه به مسألهٔ ۶ احاله شده است.

رابطهٔ (۴۱.۱.۹) نشان می‌دهد که پایداری $u_k(\epsilon)$ به ضرایب وضعیت s_j و نزدیکی λ_j به λ_k بستگی دارد. این امر هشداری بر احتمال ناپایداری است، و مثالهای دیگری از این موضوع در مسألهٔ ۷ داده شده‌اند. یک بررسی عمیقتر از رفتار ویژه‌بردارها، وقتی A دستخوش اختلالاتی باشد، نیاز به بررسی زیرفضاهای ویژه‌بردار و ارتباط بین آنها را می‌طلبد، خاصه وقتی که ویژه‌مقدار λ_k ساده نباشد. برای اطّلاعات بیشتر در این زمینه، گلوب و وَن‌لون (۱۹۸۳، صص ۲۰۳ تا ۲۰۷ و ۲۷۱ تا ۲۷۵) را ببینید.

ماتریسهای دارای شکل متعارف ناقطری ژوردان از بحث در مسألهٔ ویژه‌مقدار ماتریسهایی که شکل کانونی ژوردان آنها ناقطری بود، احتراز کرده‌ایم. در مسألهٔ ویژه‌مقدار، مسائلی در مورد ناپایداری وجود دارند که از آنچه در قضیهٔ ۲.۹ داده‌شده بدترند. و مشکلات عمده‌ای در تعیین یک مبنای صحیح برای ویژه‌بردارها وجود دارند.

به جای ارائهٔ یک بحث کلی، مشکلات این رده از ماتریسها را با بررسی یک حالت ساده به تفصیل مطالعه می‌کنیم. گیریم

$$A = \begin{bmatrix} 1 & 1 & 0 & & & \cdots & 0 \\ 0 & 1 & 1 & & 0 & & \vdots \\ \vdots & & & \ddots & & & \\ & & & & & 1 & 1 \\ 0 & & \cdots & & & 0 & 1 \end{bmatrix} \quad (۴۲.۱.۹)$$

ماتریسی از مرتبهٔ n باشد. چندجمله‌یی مشخّصهٔ آن چنین است:

$$f(\lambda) = (1 - \lambda)^n$$

است و $\lambda = 1$ یک ریشهٔ چندگانه از مرتبهٔ n است. تنها یک مجموعهٔ یک ـ بعدی از ویژه‌بردارها وجود دارد که توسط

$$x = [1, 0, \ldots, 0]^T \quad (۴۳.۱.۹)$$

تولید می‌شود.

برای $\epsilon > 0$، یک اختلال به شکل زیر برای A ایجاد می‌کنیم

$$A(\epsilon) = \begin{bmatrix} 1 & 1 & 0 & & & \dots & 0 \\ 0 & 1 & 1 & & 0 & & \\ \vdots & & & \ddots & & & \vdots \\ 0 & & & & & 1 & 1 \\ \epsilon & 0 & \dots & & & 0 & 1 \end{bmatrix}$$

چندجمله‌یی مشخصهٔ آن

$$f_\epsilon(\lambda) = (1-\lambda)^n - (-1)^n\epsilon$$

خواهد بود. در اینجا n ریشهٔ متمایز

$$\lambda_k(\epsilon) = 1 + \omega_k\epsilon^{1/n} \qquad k = 1, \dots, n \qquad (۴۴.۱.۹)$$

وجود دارد که مجموعهٔ $\{\omega_k\}$، ریشه‌های n ام واحد هستند:

$$\omega_k = e^{2\pi ki/n} \qquad k = 1, \dots, n$$

برای اختلالهای حاصل در ویژه‌مقدار A، داریم

$$|\lambda_k(\epsilon) - \lambda_k| = \epsilon^{1/n} \qquad (۴۵.۱.۹)$$

مثلاً، اگر $n = 10$ و $\epsilon = 10^{-10}$، آنگاه

$$|\lambda_k(\epsilon) - \lambda_k| = 0.1 \qquad (۴۶.۱.۹)$$

از قضیّهٔ قبلی (۱۰.۱.۹) کرانی به‌دست آمد که برحسب ϵ خطی بود، و همان‌گونه که از (۴۶.۱.۹) پیداست، (۴۵.۱.۹) از کران قبلی بسیار بدتر است.

چون $A(\epsilon)$ دارای n ویژه‌مقدار متمایز است، یک مجموعهٔ کامل از n ویژه‌بردار مستقل خطی نیز دارد، که آنها را $x_1(\epsilon)$، ...، $x_n(\epsilon)$ می‌نامیم. اولین کاری که باید انجام گیرد، پیداکردن ارتباط بین این ویژه‌بردارها با یک ویژه‌بردار x است که در (۴۳.۱.۹) داده شده است. این مسأله یک مسأله‌ای است مشکل، و همیشه در ماتریسهایی ظاهر می‌شود که شکل ژوردان آنها ناقطری است.

شکل ماتریسهای A و $A(\epsilon)$ به غایت ساده است، و فقط به مشکلاتی اشاره می‌کنند که ممکن است هنگامی پدید آیند که ماتریس با یک ماتریس قطری مشابه نباشد. در کار عملی، خطاهای گردکردن همیشه تضمین می‌کنند که چنین ماتریسی ویژه‌مقدارهای متمایز دارد. این مثال از نقطهٔ‌نظر کیفی، مثال مناسبی برای نشان‌دادن مشکلاتی است که پیش می‌آیند.

۲.۹ روش توانی

این روش، یک روش کلاسیک، برای پیدا کردن ویژه‌مقدار غالب و ویژه‌بردار متناظر با آن، در یک ماتریس است. این، یک روش کلّی نیست ولی در بعضی موارد مفید است. مثلاً برای ماتریسهای بزرگ و تُنک، که روشهای بخشهای اخیر، به علت محدودیت در حافظهٔ رایانه، نمی‌توانند به‌کار برده شوند، اغلب روشی موفق است. به‌علاوه، روش بارست معکوس، که در بخش ۶.۹ توضیح داده‌شده، روش توانی است که برای یک ماتریس معکوس مناسب به‌کار گرفته شده است. و مطالب این بخش، مقدمه‌ای بر مطالب اخیر است. به‌کاربردن روش توانی به‌عنوان یک برنامهٔ همه‌منظورهٔ رایانه‌یی در مورد رده‌های گوناگون ماتریسهای بزرگ، بسیار مشکل است. ولی می‌توان آن را برای بعضی رده‌های خاص به آسانی به‌کار برد.

فرض می‌کنیم A یک ماتریس حقیقی $n \times n$ باشد که شکل متعارف ژوردان آن قطری است. گیریم λ_1، ...، λ_n معرّف ویژه‌مقدارهای A و x_1، ...، x_n ویژه‌بردارهای متناظر با آنها باشند که یک پایه برای $\mathbf{C}^n$ تشکیل می‌دهند. به‌علاوه فرض می‌کنیم

$$| \lambda_1 | > | \lambda_2 | \geq | \lambda_3 | \geq \ldots \geq | \lambda_n | \geq 0 \qquad (۱.۲.۹)$$

گرچه این حالت، حالتی است کاملاً خاص، ولی توجه اصلی ما به‌کاربرد روش توانی در این حالت است. بحث مربوط به آن به سادگی قابل‌تعمیم به حالتی است که یک ویژه‌مقدار غالب چندگانگی هندسی $r > 1$ داشته باشد. (مسألهٔ ۱۰ را ملاحظه کنید). توجه داشته باشید که لازمهٔ این مفروضات حقیقی‌بودن λ_1 و x_1 است.

گیریم $z^{(0)}$ یک حدس اولیهٔ حقیقی برای مضربی از x_1 باشد. اگر هیچ روش منطقی برای انتخاب $z^{(0)}$ در دست نباشد، آنگاه یک مولد عددی تصادفی برای هر مؤلفه انتخاب می‌کنیم. تعریف می‌کنیم

$$w^{(m)} = Az^{(m-1)} \qquad (۲.۲.۹)$$

گیریم β_m یک مؤلفهٔ $w^{(m)}$ باشد که اندازهٔ آن ماکسیمم است. تعریف می‌کنیم

$$z^{(m)} = \frac{w^{(m)}}{\beta_m} \qquad m \geq 1 \qquad (۳.۲.۹)$$

نشان می‌دهیم که بردارهای $\{z^{(m)}\}$ وقتی $m \to \infty$، کمیّت $\hat{\sigma}_m x^{(1)} / \| x^{(1)} \|_\infty$ را با هریک از مقادیر $\hat{\sigma}_m = \pm 1$، تقریب می‌زنند.

ابتدا نشان می‌دهیم که؛

$$z^{(m)} = \sigma_m \frac{A^m z^{(0)}}{\| A^m z^{(0)} \|_\infty} \qquad \sigma_m = \pm 1 \qquad m \geq 1 \qquad (۴.۲.۹)$$

اولاً، $w^{(1)} = Az^{(0)}$.

$$\beta_1 = \sigma_1 \| w^{(1)} \|_\infty = \sigma_1 \| Az^{(0)} \|_\infty \qquad \sigma_1 = \pm 1$$

پس

$$z^{(1)} = \frac{w^{(1)}}{\beta_1} = \sigma_1 \frac{Az^{(0)}}{\| Az^{(0)} \|_\infty}$$

در اثبات (۴.۲.۹) برای حالت کلّیِ $m > 1$ از استقرای ریاضی استفاده می‌کنیم. به‌عنوان مثال برای $m = 2$

$$w^{(2)} = Az^{(1)} = \sigma_1 \frac{A^2 z^{(0)}}{\| Az^{(0)} \|_\infty}$$

$$\beta_2 = \mu \frac{\| A^2 z^{(0)} \|_\infty}{\| Az^{(0)} \|_\infty} \qquad \mu = \pm 1$$

$$z^{(2)} = \frac{w^{(2)}}{\beta_2} = \sigma_1 \frac{A^2 z^{(0)}}{\| Az^{(0)} \|_\infty} \div \frac{\mu \| A^2 z^{(0)} \|_\infty}{\| Az^{(0)} \|_\infty} = \sigma_2 \frac{A^2 z^{(0)}}{\| A^2 z^{(0)} \|_\infty}$$

که در آن $\sigma_2 = \sigma_1 \mu$. حالت کلّی m اساساً به همین شکل به‌دست می‌آید.

برای بررسی همگرایی $\{z^{(m)}\}$، ابتدا $z^{(0)}$ را برحسب ویژه‌بردارهای پایهٔ $\{x_j\}$ بسط می‌دهیم:

$$z^{(0)} = \sum_{j=1}^{n} \alpha_j x_j$$

فرض می‌کنیم $\alpha_1 \neq 0$، که انتخاب تصادفی $z^{(0)}$ معمولاً صحّت آن را تضمین می‌کند. همچنین می‌توان نشان داد که α_1 حقیقی است. پس

$$\begin{aligned} A^m z^{(0)} &= \sum_{j=1}^{n} \alpha_j A_j^m x = \sum_{j=1}^{m} \alpha_j \lambda_j^m x_j \\ &= \lambda_1^m \left[\alpha_1 x_1 + \sum_{j=2}^{n} \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^m x_j \right] \end{aligned} \qquad (۵.۲.۹)$$

با توجّه به (۱.۲.۹) داریم

$$\left(\frac{\lambda_j}{\lambda_1} \right)^m \to 0 \qquad m \to \infty \text{ وقتی} \qquad 2 \leq j \leq m \qquad (۶.۲.۹)$$

با استفاده از این رابطه در (۵.۲.۹) و (۴.۲.۹)، داریم

$$z^{(m)} \doteq \left(\frac{\lambda_1}{|\lambda_1|} \right)^m \frac{\sigma_m \alpha_1 x_1}{|\alpha_1| \| x_1 \|_\infty} \equiv \hat{\sigma}_m \cdot \frac{x_1}{\| x_1 \|_\infty} \qquad (۷.۲.۹)$$

پس $\sigma_m = \pm 1$، و معمولاً مستقل از m است. در حالتی که x_1 یک مؤلفهٔ ماکسیمال یکتا داشته باشد چنین است. در حالتی که x_1 بیش از یک مؤلفهٔ ماکسیمال داشته باشد ممکن است که $\hat{\sigma}_m$

با m تغییر کند (مسألهٔ ۹ را ببینید). نرخ همگرایی در (۷.۲.۹) به $|\lambda_2/\lambda_1|$ بستگی خواهد داشت:

$$\left\| z^{(m)} - \hat{\sigma}_m \frac{x_1}{\| x_1 \|_\infty} \right\|_\infty \leq C \left| \frac{\lambda_2}{\lambda_1} \right|^m \qquad (۸.۲.۹)$$

زیرا همهٔ نسبتهای باقیماندهٔ $|\lambda_j/\lambda_1|$ به کران $|\lambda_2/\lambda_1|$ محدود شده است.

برای به‌دست آوردن یک دنباله از ویژه‌مقدارهای تقریبی، گیریم k اندیس یک مؤلفهٔ ناصفر x_1 باشد. در حالت کلی، k را یک زیرنمایهٔ مؤلفهٔ α_m می‌گیریم و ممکن است با m تغییر کند. تعریف می‌کنیم

$$\lambda_1^{(m)} = \frac{w_k^{(m)}}{z_k^{(m-1)}} \qquad m \geq 1 \qquad (۹.۲.۹)$$

برای بررسی نرخ همگرایی، از (۴.۲.۹) و (۵.۲.۹) استفاده می‌کنیم:

$$\lambda_1^{(m)} = \left[\sigma_m \cdot \frac{A^m z^{(0)}}{\| A^{m-1} z^{(0)} \|} \right]_k \div \left[\sigma_m \cdot \frac{A^{m-1} z^{(0)}}{\| A^{m-1} z^{(0)} \|_\infty} \right]_k$$

$$= \frac{\lambda_1^m \left[\alpha_1 x_1 + \sum_{j=2}^{n} \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^m x_j \right]_k}{\lambda_1^{m-1} \left[\alpha_1 x_1 + \sum_{j=2}^{n} \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^{m-1} x_j \right]_k} \qquad (۱۰.۲.۹)$$

$$\lambda_1^{(m)} = \lambda_1 \left[1 + O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^m \right) \right] \qquad (۱۱.۲.۹)$$

نرخ همگرایی خطی است و به $|\lambda_2/\lambda_1|$ بستگی دارد. اندیس k معمولاً ثابت انتخاب می‌شود. اگر k را α_m، اندیس مؤلفهٔ ماکسیمال $w^{(m)}$ بگیریم، و اگر x_1 یک مؤلفهٔ ماکسیمال تنها داشته باشد، آنگاه وقتی $m \to \infty$، مقدار k ثابت می‌شود. با بیش از یک مؤلفهٔ ماکسیمال، k ممکن است تغییر کند همان‌گونه که در مسألهٔ ۹ نشان داده شده است. یک روش دیگر برای تعریف $\lambda_1^{(m)}$ در (کونت و دبور[1] (۱۹۸۰، ص ۱۹۲) داده شده است، که نیاز به انتخاب یک اندیس مؤلفهٔ خاص ندارد:

$$\lambda_1^{(m)} = \frac{u^T w^{(m)}}{u^T z^{(m-1)}}$$

بردار u باید در $u^T x_1 \neq 0$ صدق کند، و یک انتخاب تصادفی برای u معمولاً کافی است. خطای $\lambda_1^{(m)}$ باز هم در رابطهٔ (۱۱.۲.۹) صدق می‌کند.

1. Conte and de Boor

جدول ۱.۹ مثالی از روش توانی

m	$z_1^{(m)}$	$z_2^{(m)}$	$z_3^{(m)}$	$\lambda_1^{(m)}$	R_m
۱	۰٫۵۰۰۷۷	۰٫۷۵۰۳۸	۱٫۰۰۰۰	۱۱٫۷۶۲۸۱۳۳	
۲	۰٫۵۲۶۲۶	۰٫۷۶۳۱۳	۱٫۰۰۰۰	۹٫۵۰۳۸۴۹۶	
۳	۰٫۵۲۴۵۹	۰٫۷۶۲۳۰	۱٫۰۰۰۰	۹٫۶۳۱۳۲۳۱	−۰٫۰۵۶۴۳
۴	۰٫۵۲۴۷۰	۰٫۷۶۲۳۵	۱٫۰۰۰۰	۹٫۶۲۲۹۶۷۴	−۰٫۰۶۵۵۵
۵	۰٫۵۲۴۶۹	۰٫۷۶۲۳۵	۱٫۰۰۰۰	۹٫۶۲۳۵۰۸۳	۰٫۰۶۴۷۴

مثال گیریم

$$A = \begin{bmatrix} ۱ & ۲ & ۳ \\ ۲ & ۳ & ۴ \\ ۳ & ۴ & ۵ \end{bmatrix} \qquad (۱۲.۲.۹)$$

ویژه‌مقدارهای درست آن عبارت‌اند از:

$$\lambda_۱ = ۹٫۶۲۳۴۷۵۳۸۳ \qquad \lambda_۲ = -۰٫۶۲۳۴۷۵۳۸۳۰ \qquad \lambda_۳ = ۰ \qquad (۱۳.۲.۹)$$

یک حدس اولیهٔ $z^{(۰)}$ با یک مولد عددی تصادفی تولید شده بود. پنج بارست اوّل همراه با نسبتهای

$$R_m = \frac{\lambda_۱^{(m)} - \lambda_۱^{(m-۱)}}{\lambda_۱^{(m-۱)} - \lambda_۱^{(m-۲)}} \qquad (۱۴.۲.۹)$$

در جدول ۱.۹ داده شده‌اند. بارستهای $\lambda_۱^{(m)}$ با استفاده از (۹.۲.۹) و $k = ۳$، تعریف شده بودند. براساس یک بحث بعدی، این نسبتها باید $\lambda_۲/\lambda_۱$ را، وقتی $m \to \infty$، تقریب بزنند.

ما کران خطای قابل‌محاسبهٔ (۲۲.۱.۹) را که در بخش ۱.۹ به‌دست آمده بود به‌کار می‌بریم.

$$\eta = Ax - \lambda x$$

را با

$$x = x^{(۵)} \qquad \lambda = \lambda_۱^{(۵)}$$

محاسبه می‌کنیم. سپس از (۲۲.۱.۹) به‌دست می‌آوریم.

$$\min_i \mid \lambda_i - \lambda_۱^{(۵)} \mid < \frac{\| \eta \|_۲}{\| x \|_۲} = ۳٫۳۰ \times ۱۰^{-۵} \qquad (۱۵.۲.۹)$$

از مقایسهٔ مستقیم با جواب واقعی λ_1 داریم

$$\lambda_1 - \lambda_1^{(5)} = ۰٫۰۰۰۰۳۲۹$$

که نشان می‌دهد (۱۵.۲.۹) در این مورد براورد خیلی دقیق است.

روشهای شتاب چون برای کاهش خطاهای $\lambda_1^{(m)}$ و $z^{(m)}$ یک الگوی منظّم معلومی وجود دا از این الگو می‌توان برای به‌دست آوردن روشهای سریعتر همگرایی استفاده کرد.

حالت ۱. انتقال ویژه‌مقدارها. یک مقدار ثابت b انتخاب می‌کنیم، و به‌جای محاس ویژه‌مقدارهای A محاسبهٔ ویژه‌مقدارهای

$$B = A - bI \qquad (۱۶.۲.۹)$$

را قرار می‌دهیم. ویژه‌مقدارهای B، $\lambda_i - b$، $i = ۱, \ldots, n$، هستند. b را طوری می‌گیریم $\lambda_1 - b$ ویژه‌مقدار غالب B باشد، و b را طوری انتخاب می‌کنیم که نسبت همگرایی را مینیمم ک

به‌عنوان یک حالت خاص، برای وضوح بیشتر، فرض می‌کنیم که همهٔ ویژه‌مقدارهای A حقی و به گونه‌ای مرتب شده باشند که

$$\lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n \qquad \lambda_1 > |\lambda_n|$$

در این صورت ویژه‌مقدار غالب B، بسته به اندازهٔ b، ممکن است $\lambda_1 - b$ یا $\lambda_n - b$ باشد. ا می‌خواهیم که b در

$$|\lambda_1 - b| > |\lambda_n - b|$$

صدق کند. نرخ همگرایی عبارت است از

$$\text{Max}\left\{\frac{|\lambda_2 - b|}{|\lambda_1 - b|}, \frac{|\lambda_n - b|}{|\lambda_1 - b|}\right\} \qquad (۱۷.۲.۹)$$

اگر دقیقاً به رفتار این دو کسر، وقتی b تغییر می‌کند، بنگریم، خواهیم دید که مینیمم (۷.۲.۹ وقتی حاصل می‌شود که

$$(\lambda_1 - b) - (\lambda_2 - b) = (\lambda_n - b) - [-(\lambda_1 - b)]$$

و

$$b^* = \frac{1}{2}(\lambda_2 + \lambda_n) \qquad (۱۸.۲.۹)$$

انتخاب بهینهٔ b خواهد بود. نسبت همگرایی حاصله چنین است:

$$\frac{\lambda_۲ - b^*}{\lambda_۱ - b^*} = -\frac{\lambda_n - b^*}{\lambda_۱ - b^*} = \frac{\lambda_۲ - \lambda_n}{۲\lambda_۱ - \lambda_۲ - \lambda_n} \qquad (۱۹.۲.۹)$$

روشهای تجربی براساس این فرمول و فرمول (۱۱.۲.۹) می‌توانند برای تعیین مقادیر تقریبی b^* به‌کار روند.

تبدیلاتی غیر از (۱۶.۲.۹) را می‌توان برای تبدیل مجموعهٔ ویژه‌مقدارها به گونه‌ای به‌کار برد که همگرایی بازهم سریعتر به‌دست آید. برای بحث بیشتر در مورد این مطالب، ویلکینسن (۱۹۶۵، صص ۵۷۰ ـ ۵۸۰) را ملاحظه کنید.

مثال در مثال قبلی (۱۲.۲.۹)، نسبت همگرایی نظری چنین بود

$$\frac{\lambda_۲}{\lambda_۱} \doteq -۰٫۰۶۴۸$$

با استفاده از مقدار بهینهٔ b که با (۱۸.۲.۹) داده شده است، و استفاده از (۱۳.۲.۹) با تغییر ترتیب،

$$b^* = \frac{۱}{۲}(\lambda_۲ + \lambda_n) \doteq -۰٫۳۱۱۷۴ \qquad (۲۰.۲.۹)$$

ویژه‌مقدارهای $A - bI$ عبارت‌اند از

$$۹٫۹۳۵۲۲ \qquad ۰٫۳۱۱۷۴ \qquad -۰٫۳۱۱۷۴ \qquad (۲۱.۲.۹)$$

نسبت همگرایی روش توانی برای $A - bI$ چنین است

$$\pm\frac{۰٫۳۱۱۷۴}{۹٫۹۳۵۲۲} \doteq \pm۰٫۰۳۱۴$$

که از نصف قدرمطلق نسبت اصلی کمتر است.

حالت ۲. برونیابی ایتکن. شکل همگرایی (۱۱.۲.۹) کاملاً شبیه روشهای ریشه‌یابی به‌طور خطی همگرای بخش ۵.۲ از فصل ۲ است. با دنبال‌کردن همان بحثی که در بخش ۶.۲ داده شد، برونیابی ایتکن را برای تسریع همگرایی $\{\lambda_۱^{(m)}\}$ و $\{z^{(m)}\}$ درنظر می‌گیریم. برای استفاده از بحث زیر، باید در (۱.۲.۹) فرض کنیم که

$$|\lambda_۲| \neq |\lambda_۳| \qquad (۲۲.۲.۹)$$

این فرض ممکن است سست‌ترشده به $|\lambda_۲| = |\lambda_j|$ بینجامد، که درنتیجه $\lambda_۲ = \lambda_j$. ولی دو نسبت همگرایی با اندازهٔ مساوی و مختلف‌العلامه را مجاز نمی‌دانیم (مثال قبلی (۲۱.۲.۹) را برای این مورد، ملاحظه کنید). شیوهٔ ایتکن را نیز می‌توان چنان اصلاح کرد که محدودیت (۲۲.۲.۹) از میان برود.

با (۲۲.۲.۹) و استفاده از (۱۰.۲.۹)،

$$\lambda_1 - \lambda_1^{(m)} \doteq cr^m \qquad (۲۳.۲.۹)$$

که c مقداری است ثابت، r نرخ مجهول همگرایی، و از لحاظ نظری $r = \lambda_2/\lambda_1$. اگر دقیقاً بخش ۶.۲ را دنبال کنیم نتیجه می‌شود که؛

$$R_m = \frac{\lambda_1^{(m)} - \lambda_1^{(m-1)}}{\lambda_1^{(m-1)} - \lambda_1^{(m-2)}} \to r \quad m \to \infty \quad \text{وقتی} \qquad (۲۴.۲.۹)$$

و برونیابی ایتکن مقدار بهبودیافتهٔ $\hat{\lambda_1}$ را چنین می‌دهد:

$$\hat{\lambda_1} = \lambda_1^{(m)} - \frac{[\lambda_1^{(m)} - \lambda_1^{(m-1)}]^2}{[\lambda_1^{(m)} - \lambda_1^{(m-1)}] - [\lambda_1^{(m-1)} - \lambda_1^{(m-2)}]} \qquad m \geq 3 \quad (۲۵.۲.۹)$$

محاسبهٔ مشابهی را برای ویژه‌بردارهای تقریبی می‌توان به‌کار برد تا هر مؤلفهٔ دنبالهٔ $\{z^{(m)}\}$ را شتاب دهد، اگرچه ملاحظاتی باید اعمال شوند.

مثال باز هم مثال (۱۲.۲.۹) را درنظر می‌گیریم. در جدول ۱.۹،

$$R_5 = -۰٫۰۶۴۷۴ \doteq \frac{\lambda_2}{\lambda_1} = -۰٫۰۶۴۷۹$$

به‌عنوان یک مثال از (۲۵.۲.۹)، با مقادیر $\lambda_1^{(2)}$، $\lambda_1^{(3)}$ و $\lambda_1^{(4)}$ از آن جدول برونیابی می‌کنیم. پس

$$\hat{\lambda_1} = ۹٫۶۲۳۴۸۱۴ \qquad \lambda_1 - \hat{\lambda_1} = -۶٫۰۳ \times ۱۰^{-6}$$

در مقایسه با استفاده از مقادیر دقیقتر $\lambda_1^{(5)}$ از همان جدول، $\lambda_1 - \lambda_1^{(5)} = -۳٫۲۹ \times ۱۰^{-5}$. این امر، (هرگاه استفاده از برونیابی از نظر تئوری توجیه‌پذیر باشد)، ارزش این استفاده را مجدداً نشان می‌دهد.

حالت ۳. خارج‌قسمت ریلی ـ ریتس[1]. هرگاه A متقارن باشد، بهتر است که تقریبهای ویژه‌مقدار زیر را به‌کار بریم

$$\lambda_1^{(m+1)} = \frac{(Az^{(m)}, z^{(m)})}{(z^{(m)}, z^{(m)})} = \frac{(w^{(m+1)}, z^{(m)})}{(z^{(m)}, z^{(m)})} \qquad m \geq 0 \quad (۲۶.۲.۹)$$

ما معمولاً قرارداد استانداردِ ضرب داخلی را به‌کار می‌بریم:

1. Rayleigh-Ritz

$$(w, z) = \sum_{1}^{n} w_i z_i \qquad w, z \in \mathbf{R}^n$$

برای تحلیل این دنبالهٔ (۲۶.۲.۹)، توجه کنید که تمام ویژه‌مقدارهای A حقیقی‌اند و ویژه‌بردارهای x_1، ...، x_n می‌توانند یکا متعامد انتخاب شوند. در این صورت (۲.۲.۹)، (۴.۲.۹)، (۵.۲.۹) همراه با (۲۶.۲.۹) به ما می‌دهند

$$\lambda_1^{(m+1)} = \frac{\sum_{j=1}^{n} |\alpha_j|^2 \lambda_j^{2m+1}}{\sum_{j=1}^{n} |\alpha_j|^2 \lambda_j^{2m}}$$

$$\lambda_1^{(m+1)} = \lambda_1 \left[1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2m}\right)\right] \qquad (۲۷.۲.۹)$$

نسبت همگرایی $\lambda_1^{(m)}$ به λ_1 برابر $(\lambda_2/\lambda_1)^2$ است، که بهبودی برای نسبت اصلی λ_2/λ_1 در (۲۷.۲.۹) است.

این رویّه یک رویّهٔ کلاسیک معروف است و جنبه‌های اضافی دیگری در بر دارد که در بعضی مسائل دیگر کاربرد دارند. برای بحث بیشتر ویلکینسن (۱۹۶۵، صص ۱۷۲ تا ۱۷۸) را ببینید.

مثال در مثال (۱۲.۲.۹)، ویژه‌بردار تقریبی $z^{(2)}$ را در (۲۶.۲.۹) به‌کار می‌بریم. پس

$$\lambda_1^{(3)} = \frac{(Az^{(2)}, z^{(2)})}{(z^{(2)}, z^{(2)})} \doteq ۹٫۶۲۳۴۶۴$$

که به همان دقّت $\lambda_1^{(5)}$ است که قبلاً به‌دست آمده بود.

روش توانی را می‌توان وقتی که یک ویژه‌مقدار غالب تنها وجود نداشته باشد نیز به‌کار برد، ولی در آن صورت الگوریتم پیچیده‌تر خواهد شد. روش توانی را می‌توان برای تعیین ویژه‌مقدارهایی غیر از ویژه‌مقدار غالب نیز به‌کار برد. این عمل متضمّن فرایندی است به نام **کاهش** A برای حذف λ به‌عنوان یک ویژه‌مقدار. برای یک بحث کامل در تمام زمینه‌های روش توانی، گلوب و وَن‌لون (۱۹۸۳، صص ۲۰۸-۲۱۸) و ویلکینسن (۱۹۶۵، فصل ۹) را ببینید. گرچه روش توانی یک روش مفید در بعضی شرایط است، ولی باید تأکید کرد که روشهایی که در بخشهای بعدی می‌آیند معمولاً کاراترند. برای یک تغییر همگرایی سریع در روش توانی و خارج قسمت ریلی ـ ریتس، بارست **خارج‌قسمت ریلی** برای ماتریسهای متقارن در پارلت (۱۹۸۰، ص ۷۰) را ببینید.

۳.۹ تبدیلات متعامد با استفاده از ماتریسهای هاؤسهولدر

به‌عنوان یک مرحله در پیدا کردن ویژه‌مقدارهای یک ماتریس، اغلب استفاده از تبدیلات تشابهی، مسأله را به یک صورت ساده‌تری بدل می‌کند. ماتریسهای متعامد ردهٔ ماتریسهایی خواهند بود که ما برای این تبدیلات به‌کار می‌بریم. در (۳۹.۱.۹) نشان داده شده بود که تبدیلات متعامد وضعیت یا پایداری ویژه‌مقدارهای یک ماتریس نامتقارن را بدتر نمی‌کنند. همچنین، ماتریسهای متعامد ویژگیهای مطلوب دیگر انتشار خطا را نیز در بر دارند، که یک نمونه از آن بعداً در این بخش داده شده است؛ و به همین دلایل، ما تبدیلات خود را به تبدیلاتی که در آنها از ماتریسهای متعامد استفاده می‌شود محدود می‌کنیم.

این بخش را با نگاهی به ردهٔ خاص ماتریسهای متعامدی که به **ماتریسهای هاؤسهولدر** معروف‌اند آغاز می‌کنیم. سپس نشان می‌دهیم که چگونه یک ماتریس هاؤسهولدر بسازیم که یک بردار داده‌شده را به شکل ساده‌تری بدل کند. با این ساختمان به‌عنوان یک ابزار، به دو تبدیل یک ماتریس داده‌شدهٔ A توجه می‌کنیم: (۱) به‌دست آوردن تجزیهٔ QR آن، و (۲) ساختن یک ماتریس مشابه سه‌قطری برای حالتی که A متقارن است. این شکلها در بخشهای بعد، در محاسبهٔ ویژه‌مقدارهای A، به‌کار می‌روند. از نظر نمادگذاری، توجه داشته باشید که اصطلاح متعامد را به ماتریسهای حقیقی، محدود کنیم. ولی خیلی متداول شده است که **متعامد** را در این زمینه به‌جای **یکانی**، در حالت کلی مختلط، به‌کار می‌برند، و ما همین نمادگذاری را می‌پذیریم. خوانندگان باید واژهٔ **تعامد** را، وقتی برای ماتریسهای مختلط به‌کار برده می‌شود، به مفهوم **یکانی** منظور کنند.

گیریم $w \in \mathbf{C}^*$ با $\| w \|_۲ = \sqrt{w^* w} = ۱$. تعریف می‌کنیم

$$U = I - ۲ww^* \qquad (۱.۳.۹)$$

این حالت کلّی ماتریس هاؤسهولدر است.

مثال برای $n = ۳$، می‌خواهیم

$$w = [w_۱, w_۲, w_۳]^T \qquad |w_۱|^۲ + |w_۲|^۲ + |w_۳|^۲ = ۱$$

ماتریس U چنین داده می‌شود:

$$U = \begin{bmatrix} ۱ - ۲|w_۱|^۲ & -۲w_۱\overline{w_۲} & -۲w_۱\overline{w_۳} \\ -۲\overline{w_۱}w_۲ & ۱ - ۲|w_۲|^۲ & -۲w_۲\overline{w_۳} \\ -۲\overline{w_۱}w_۳ & -۲\overline{w_۲}w_۳ & ۱ - ۲|w_۳|^۲ \end{bmatrix}$$

برای حالت خاص

$$w = \left[\frac{۱}{۳}, \frac{۲}{۳}, \frac{۲}{۳}\right]^T$$

داریم

$$U = \frac{۱}{۹}\begin{bmatrix} ۷ & -۴ & -۴ \\ -۴ & ۱ & -۸ \\ -۴ & -۸ & ۱ \end{bmatrix}$$

ابتدا ثابت می‌کنیم که U ارمیتی و متعامد است. برای نشان‌دادن ارمیتی‌بودن

$$\begin{aligned} U^* &= (I - ۲ww^*)^* = I^* - ۲(ww^*)^* \\ &= I - ۲(w^*)^*w^* = I - ۲ww^* = U \end{aligned}$$

برای نشان‌دادن تعامد،

$$\begin{aligned} U^*U = U^۲ &= (I - ۲ww^*)^۲ \\ &= I - ۴ww^* + ۴(ww^*)(ww^*) \\ &= I \end{aligned}$$

زیرا استفاده از قانون شرکتپذیری و $w^*w = ۱$ ایجاب می‌کند که داشته باشیم

$$(ww^*)(ww^*) = w(w^*w)w^* = ww^*$$

ماتریسِ U در مثال قبل این ویژگیها را نشان می‌دهد. در مسألهٔ ۱۲، یک تعبیر هندسی از تابع خطی $T(x) = Ux$ برای یک ماتریس هاؤسهولدر U ارائه می‌دهیم.

معمولاً بردارهای w را با چند مؤلفهٔ اول صفر به‌کار می‌بریم،

$$w = [۰, \ldots, ۰, w_r, \ldots, w_n]^T = [O_{r-۱}, \hat{w}^T]^T \qquad (۲.۳.۹)$$

که در آن $\hat{w} \in \mathbf{C}^{n-r+۱}$. پس

$$U = \begin{bmatrix} I_{r-۱} & O \\ O & I_{n-r+۱} - ۲\hat{w}\hat{w}^* \end{bmatrix} \qquad (۳.۳.۹)$$

ضرب قدّامی یک ماتریس A در این U، $r-۱$ سطر اوّل A را تغییر نمی‌دهد، و ضرب خَلْفی A در آن، $r-۱$ ستون اول A را تغییر نخواهد داد. برای بقیّهٔ این بخش ما فرض می‌کنیم تمام ماتریسها و بردارها حقیقی‌اند، تا از مقادیر مختلط احتمالی برای w احتراز کنیم.

ماتریسهای هاوسهولدر برای تبدیل بردار ناصفر به بردار جدیدی که قسمت عمدهٔ مؤلّفه‌های آن صفر است به‌کار می‌برند. گیریم $b \neq 0$ داده شده است، $b \in \mathbf{R}^n$، و بنا به فرض می‌خواهیم ماتریس U به شکل (۱.۳.۹) را به گونه‌ای تولید کنیم که Ub در موضع $r+1$ام تا nام، برای مقادیر داده‌شدهٔ $r \geq 1$، شامل صفر باشد. w را مطابق (۲.۳.۹) انتخاب می‌کنیم. در این صورت $r-1$ عنصر اول b و Ub یکی هستند.

برای ساده‌کردن کار اخیر می‌گیریم، $m = n - r + 1$،

$$w = \begin{bmatrix} O_{r-1} \\ v \end{bmatrix} \qquad b = \begin{bmatrix} c \\ d \end{bmatrix}$$

که $c \in \mathbf{R}^{r-1}$، $v, d \in \mathbf{R}^m$. در این صورت محدودیت ما در شکل Ub ایجاب می‌کند که $r-1$ مؤلفهٔ اول Ub، c باشد، و برای مقداری از α،

$$(I - 2vv^T)d = [\alpha, 0, \ldots, 0]^T \qquad \| v \|_2 = 1 \qquad (۴.۳.۹)$$

چون $I - 2vv^T$ متعامد است، طول d حفظ می‌شود (مسألهٔ ۱۳ فصل ۷)؛ و بنابراین

$$| \alpha | = \| d \|_2 \equiv S$$

$$\alpha = \pm S = \pm\sqrt{d_1^2 + \ldots d_m^2} \qquad (۵.۳.۹)$$

تعریف می‌کنیم

$$p = v^T d$$

با توجّه به (۴.۳.۹)،

$$d - 2pv = [\alpha, 0, \ldots, 0]^T \qquad (۶.۳.۹)$$

از ضرب در v^T و استفاده از $\| v \|_2 = 1$ نتیجه می‌شود که

$$p = -\alpha v_1 \qquad (۷.۳.۹)$$

از گذاشتن این مقدار در اولین مؤلفهٔ (۶.۳.۹) نتیجه می‌شود

$$d_1 + 2\alpha v_1^2 = \alpha$$

$$v_1^2 = \frac{1}{2}[1 - \frac{d_1}{\alpha}] \qquad (۸.۳.۹)$$

علامت α در (۵.۳.۹) را به شکل زیر انتخاب می‌کنیم

$$\text{sign}(\alpha) = -\text{sign}(d_1) \qquad (۹.۳.۹)$$

این انتخاب v_1^2 را ماکسیمم می‌کند، و از هرگونه خطای ممکن از دست‌دادن ارقام بامعنی در محاسبهٔ v_1 جلوگیری می‌کند. علامت v_1 مهّم نیست. با در دست‌داشتن v_1، از (۷.۳.۹)، p به‌دست می‌آید. به (۶.۳.۹) برمی‌گردیم و مؤلفه‌ها را از ۲ تا m می‌گیریم:

$$v_j = \frac{d_j}{2p} \qquad j = 2, 3, \ldots, m \qquad (۱۰.۳.۹)$$

احکام (۵.۳.۹)، (۷.۳.۹) تا (۹.۳.۹)، v را کاملاً مشخص می‌کنند، و بنابراین w و U نیز معین می‌شوند. تعداد محاسبات عبارت است از $2m + 2$ ضرب و تقسیم و دو بار ریشهٔ دوّم. در عمل می‌توان از ریشهٔ دوّم که v_1 را معین می‌کند احتراز کرد، زیرا وقتی ماتریس ww^T تشکیل شد، ریشهٔ دوّم از بین می‌رود. دنباله‌ای از چنین تبدیلات بردار b، مورد استفاده قرار می‌گیرند تا به‌طور منظّم ماتریسها را به شکلهای ساده‌تری تبدیل کنند.

مثال بردار داده‌شدهٔ

$$b = [2, 2, 1]^T$$

را درنظر می‌گیریم. ماتریس U را به گونه‌ای حساب می‌کنیم که Ub در دو موضع آخری خود صفر داشته باشد. برای کمک به ساختمان زیر، بعضی محاسبات میانی نیز آورده شده‌اند. توجه داشته باشید که در این حالت $w = v$ و $b = d$. بنابراین

$$\alpha = -3 \qquad v_1 = \sqrt{\frac{5}{6}} \qquad p = \sqrt{\frac{15}{2}}$$

$$v_2 = \frac{2}{\sqrt{30}} \qquad v_3 = \frac{1}{\sqrt{30}}$$

ماتریس U به شکل زیر داده می‌شود

$$U = \begin{bmatrix} -\frac{2}{3} & -\frac{2}{3} & -\frac{1}{3} \\ -\frac{2}{3} & \frac{11}{15} & -\frac{2}{15} \\ -\frac{1}{3} & -\frac{2}{15} & \frac{14}{15} \end{bmatrix}$$

$$Ub = [-۳, ۰, ۰]^T$$

تجزیهٔ یک ماتریس به QR برای ماتریس مفروض A، نشان می‌دهیم که یک ماتریس متعامد Q و یک ماتریس بالامثلثی R وجود دارند به‌طوری که

$$A = QR \qquad (۱۱.۳.۹)$$

گیریم

$$P_r = I - ۲w^{(r)}w^{(r)T} \qquad r = ۱, \ldots, n-۱ \qquad (۱۲.۳.۹)$$

که در آن $w^{(r)}$ به شکل (۲.۳.۹)، با $r-۱$ صفر، در ابتدای آن است. اگر A را برحسب ستونهای آن، $A_{*۱}$، $A_{*۲}$، ...، A_{*n} بنویسیم، خواهیم داشت

$$P_۱A = [P_۱A_{*۱}, \ldots, P_۱A_{*n}]$$

$P_۱$ و $w^{(۱)}$ را با ساختمان (۵.۳.۹) ـ (۱۰.۳.۹) با $b = A_{*۱}$ انتخاب می‌کنیم. در این صورت زیرقطر $P_۱A$ در اولین ستون همه صفر خواهند بود.

$P_۲$ را به همین شکل انتخاب می‌کنیم به گونه‌ای که زیرقطر $P_۲P_۱A$ در ستون دوّم همه صفر باشند، ابتدا باید توجه کرد که چون $w^{(۲)}$ در موضع اول دارای صفر است، و چون $P_۱A$ در ستون اول در زیرقطر صفر است، حاصلضربهای $P_۲P_۱A$ و $P_۱A$ در سطر و ستون اول، یک مقدار دارند. اکنون $P_۲$ و $w^{(۲)}$ را مانند قبل در (۵.۳.۹) تا (۱۰.۳.۹) انتخاب می‌کنیم، با b برابر ستون دوّم $P_۱A$.

با ادامهٔ این کار برای هر ستون A، یک ماتریس بالامثلثی به شکل زیر خواهیم داشت

$$R \equiv P_{n-۱} \ldots P_۱A \qquad (۱۳.۳.۹)$$

اگر در مرحلهٔ rام ساختمان، تمام عناصر در زیر قطر ستون rام صفر باشند، آنگاه تنها $P_r = I$ را انتخاب کرده و به مرحلهٔ بعدی می‌رویم. برای تکمیل ساختمان، تعریف می‌کنیم

$$Q^T = P_{n-۱} \ldots P_۱ \qquad (۱۴.۳.۹)$$

که متعامد است. پس، همان‌گونه که موردنظر بود، $A = QR$.

مثال ماتریس زیر را درنظر می‌گیریم

$$A = \begin{bmatrix} ۴ & ۱ & ۱ \\ ۱ & ۴ & ۱ \\ ۱ & ۱ & ۴ \end{bmatrix}$$

پس

$$w^{(1)} = [\circ,۹۸۵۵۹۹, \circ,۱۱۹۵۷۳, \circ,۱۱۹۵۷۳]^T$$

$$A_۲ = P_۱A \begin{bmatrix} -۴,۲۴۲۶۴ & -۲,۱۲۱۳۲ & -۲,۱۲۱۳۲ \\ \circ & ۳,۶۲۱۳۲ & \circ,۶۲۱۳۲۱ \\ \circ & \circ,۶۲۱۳۲۱ & ۳,۶۲۱۳۲ \end{bmatrix}$$

$$w^{(۲)} = [\circ, \circ,۹۹۶۳۹۳, \circ,۰۸۴۸۵۷۲]^T$$

$$R = P_۲A_۲ = \begin{bmatrix} -۴,۲۴۲۶۴ & -۲,۱۲۱۳۲ & -۲,۱۲۱۳۲ \\ \circ & -۳,۶۷۴۲۳ & -۱,۲۲۴۷۵ \\ \circ & \circ & ۳,۴۶۴۱۰ \end{bmatrix}$$

برای تجزیهٔ $A = QR$، $Q = P_۱P_۲$ را حساب می‌کنیم. ولی در بیشتر حالات این محاسبه برای تولید صریح Q کارا نخواهد بود. ما در این مورد توضیح مختصری خواهیم داد.

چون متعامدبودن Q مستلزم $\det(Q) = \pm ۱$. داریم

$$|\det(A)| = |\det(Q)\det(R)| = |\det(R)| = ۵۳,۹۹۹۹$$

این عدد با این واقعیت که ویژه‌مقدارهای A برابر $\lambda = ۳, ۳, ۶$ و حاصلضرب آنها $\det(A) = ۵۴$ است، سازگاری دارد.

بحث در تجزیهٔ QR بجاست که بدانیم تا چه اندازه تجزیهٔ $A = QR$ یکتاست. برای A ی ناتکین، فرض می‌کنیم

$$A = Q_۱R_۱ = Q_۲R_۲ \qquad (۱۵.۳.۹)$$

پس $R_۱$ و $R_۲$ نیز باید ناتکین باشند، و

$$Q_۲^TQ_۱ = R_۲R_۱^{-۱}$$

معکوس یک ماتریس بالامثلثی، ماتریسی است بالامثلثی و حاصلضرب دو ماتریس بالامثلثی یک ماتریس بالامثلثی است. بنابراین $R_۲R_۱^{-۱}$ یک ماتریس بالامثلثی است. همچنین، حاصلضرب دو ماتریس متعامد، ماتریسی است متعامد، بنابراین $Q_۲^TQ_۱$ متعامد است. این حقیقت به ما می‌گوید

که $R_۲R_۱^{-۱}$ متعامد است. ولی مشکل نیست که نشان دهیم تنها ماتریسهای بالامثلثی متعامد، ماتریسهای قطری‌اند. برای یک ماتریس قطری D،

$$R_۲R_۱^{-۱} = D$$

چون $R^۲R_۱^{-۱}$ متعامد است،

$$D^۲ = I$$

چون فقط با ماتریسهای حقیقی سروکار داریم، عناصر قطر D، یا $۱+$ هستند یا $۱-$. از ترکیب این نتایج،

$$Q_۲ = Q_۱D \qquad R_۲ = DR_۱ \tag{۱۶.۳.۹}$$

نتیجه می‌گیریم که علامتهای عناصر قطری R در $A = QR$، را می‌توان به دلخواه انتخاب کرد، ولی بقیهٔ تجزیه، به‌طور یکتا تعیین می‌شود.

یک موضوع عملی دیگر این است که معین کنیم چگونه ماتریس R در (۱۳.۳.۹) را حساب کنیم. گیریم

$$A_r = P_rA_{r-۱} = [I - ۲w^{(r)}w^{(r)T}]A_{r-۱} \qquad r = ۱, ۲, \ldots, n-۱ \tag{۱۷.۳.۹}$$

که در آن $A_۰ = A$، $A_{n-۱} = R$. اگر P_r را محاسبه و آن را در $A_{r-۱}$ ضرب کنیم تا A_r به‌دست آید، تعداد ضربها برابر خواهد بود با

$$(n-r+۱)^۲ + \frac{۱}{۲}(n-r+۲)(n-r+۱)$$

روش بسیار کاراتری برای محاسبهٔ A_r وجود دارد. (۱۷.۳.۹) را دوباره به شکل زیر می‌نویسیم

$$A_r = A_{r-۱} - ۲w^{(r)}[w^{(r)T}A_{r-۱}] \tag{۱۸.۳.۹}$$

ابتدا $w^{(r)^T}A_{r-۱}$ را محاسبه می‌کنیم، و سپس $[w^{(r)^T}A_{r-۱}]$ $w^{(r)}$ و A_r را. این محاسبات تقریباً به

$$۲(n-r)(n-r+۱) + (n-r+۱) \tag{۱۹.۳.۹}$$

عمل ضرب نیاز دارد که نشان می‌دهد (۱۹.۳.۹) راه مرجّحی برای محاسبهٔ A_r، و بالاخره $R = A_{n-۱}$ است. این عملیات، هزینهٔ به‌دست آوردن $w^{(r)}$ را در بر ندارد، که به‌دنبال (۱۰.۳.۹) قبلاً مورد بحث قرار گرفته است.

اگر لازم باشد که ماتریسهای P_1، ...، P_{n-1} را برای استفاده‌های بعدی ذخیره نماییم، فقط هر ستون $w^{(r)}$، $r = 1, \ldots, n-1$، را نگه می‌داریم. اولین عنصر ناصفر $w^{(r)}$ را، که در موضع r ام است، در یک جای ویژهٔ حافظه ذخیره می‌کنیم، و بقیهٔ عناصر ناصفر $w^{(r)}$ واقع در موضع $r+1$ ام تا n ام، از ستون r ام ماتریس A_r و R را که در زیرقطر واقع‌اند ذخیره می‌نماییم. ماتریس Q در (۱۴.۳.۹) می‌تواند صریحاً محاسبه شود. ولی همان‌گونه که ساختمان (۱۸.۳.۹) نشان می‌دهد، ما برای ضرب در ماتریسهای دیگر، Q را به‌طور صریح نیاز نداریم.

استفادهٔ اصلی از تجزیهٔ A به QR در تعریف روش QR برای محاسبهٔ ویژه‌مقدارهای A خواهد بود که در بخش ۵.۹ عرضه شده است. این تجزیه را برای حل یک دستگاه خطی $Ax = b$ نیز می‌توان به‌کار برد. در اینجا تجزیه مستقیماً به دستگاه هم‌ارز $Rx = Q^T b$ منجر خواهد شد، و چون Q متعامد است، خطای بسیارکمی دخالت خواهد کرد. دستگاه $Rx = Q^T b$ یک بالامثلثی است و می‌توان آن را با استفاده از جایگذاری پسرو از یک طریق پایداری حل کرد. در حالتی که A بدوضع است، این یک راه برتر در حل دستگاه خطی $Ax = b$ می‌تواند باشد. برای بحث از خطاهای موجود در استفاده از تجزیهٔ QR و مقایسهٔ آن با روش حذف گاوسی در حل $Ax = b$، به ویلکینسن (۱۹۶۵، صص ۲۴۴ تا ۲۴۹) مراجعه کنید. ما این مبحث را باز هم در بخش ۷.۹، هنگام بحث از حل کمترین مربعات دستگاههای خطّی فرامعیّن، دنبال می‌کنیم.

تبدیل یک ماتریس متقارن به شکل سه‌قطری گیریم A یک ماتریس حقیقی متقارن باشد. برای پیداکردن ویژه‌مقدارهای A، معمولاً ابتدا آن را با تبدیلات تشابهی متعامد، به یک شکل سه‌قطری برمی‌گردانند. سپس ویژه‌مقدارهای ماتریس سه‌قطری توسط نظریهٔ دنباله‌های استورم، که در بخش ۴.۹ ارائه‌شده، یا توسط QR که در بخش ۵.۹ آمده است، محاسبه می‌شوند. برای ماتریسهای متعامد، ماتریسهای (۳.۳.۹) هاؤسهولدر را به‌کار می‌بریم.

گیریم

$$P_r = I - 2w^{(r+1)}w^{(r+1)T} \qquad r = 1, \ldots, n-2 \qquad (۲۰.۳.۹)$$

که در آن $w^{(r+1)}$ مانند (۲.۳.۹) تعریف شده است:

$$w^{(r+1)} = [0, \ldots, 0, w_{r+1}, \ldots, w_n]^T$$

[به تغییری که در نمادگذاری P_r در (۱۲.۳.۹) هنگام تعریف تجزیهٔ QR، صورت گرفته توجه کنید.] ماتریس

$$A_2 = P_1^T A P_1 = P_1 A P_1$$

با A متشابه است، عنصر a_{11} تغییر نکرده است، و A_2 متقارن خواهد بود. $w^{(2)}$ و P_1 را می‌سازیم تا شکل زیر به‌ازای مقداری چون $\hat{a}_{21}$ به‌دست آید

$$P_1 A_{*1} = [a_{11}, \hat{a}_{21}, 0, \ldots, 0]^T$$

بردار A_{*1} اولین ستون A است. از (۵.۳.۹) تا (۱۰.۳.۹) با $m = n - 1$، و

$$d = [a_{21}, a_{31}, \ldots, a_{n1}]^T$$

استفاده می‌کنیم. برای مثال، از (۸.۳.۹)،

$$w_2^2 = v_1^2 = \frac{1}{2}\left[1 - \frac{a_{21}}{\alpha}\right]$$
$$\alpha = -\text{sign}(a_{21})\sqrt{a_{21}^2 + \ldots + a_{n1}^2}$$

با داشتن P_1 و $P_1 A$، از ضرب خلفی در P_1 اولین ستون $P_1 A$ تغییر نخواهد یافت. (خواننده باید این موضوع را تحقیق کند). تقارن A_2 از رابطهٔ زیر نتیجه می‌شود

$$A_2^T = (P_1 A P_1)^T = p_1^T A^T P_1^T = P_1 A P_1 = A_2$$

چون A_2 متقارن است، ساختمان در اولین ستون A ایجاب می‌کند که در موضع سوم تا nام اولین سطر و ستون A_2 همه صفر باشند.

این روند را ادامه داده، فرض می‌کنیم

$$A_{r+1} = P_r^T A_r P_r \qquad r = 1, 2, \ldots, n-2 \qquad (۲۱.۳.۹)$$

که در آن $A_1 = A$. سپس P_r را برای ایجاد صفر در موضع $r+2$ تا n از ستون rام انتخاب می‌کنیم. در محاسبهٔ $P_r A_{r-1}$ به علّت شکل خاص P_r، ستونهای ۱ تا $r-1$ بدون تغییر باقی می‌مانند. بردار $w^{(r+1)}$ را همانند توضیح قبلی برای $w^{(2)}$، برمی‌گزینیم.

ماتریس نهایی $T \equiv A_{r-1}$ سه‌قطری و متقارن است.

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & & \ldots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & & & \vdots \\ 0 & \beta_2 & \alpha_3 & & & \\ \vdots & & & \ddots & & \\ & & & & \alpha_{n-1} & \beta_{n-1} \\ 0 & & \ldots & & \beta_{n-1} & \alpha_n \end{bmatrix} \qquad (۲۲.۳.۹)$$

این شکل، شکل بسیار مناسبتری برای محاسبهٔ ویژه‌مقدارهای A خواهد بود و ویژه‌بردارهای A را می‌توان به‌سادگی از ویژه‌بردارهای T به‌دست آورد.

T با رابطهٔ زیر به A وابسته است

$$T = Q^T A Q \qquad Q = P_1 \dots P_{n-2} \tag{۲۳.۳.۹}$$

مانند قبل در تجزیهٔ QR، ما به‌ندرت Q را صریحاً تولید می‌کنیم، و ترجیح می‌دهیم که همانند (۱۸.۳.۹)، با هریک از ماتریسهای P_r جداگانه کار کنیم. برای x، یک ویژه‌بردار A، یعنی $Ax = \lambda x$، داریم

$$Tz = \lambda z \qquad x = Qz \tag{۲۴.۳.۹}$$

اگر یک مجموعه از ویژه‌بردارهای یکا متعامد $\{z_i\}$ را برای T را تولید کنیم، آنگاه $\{Qz_i\}$ یک مجموعهٔ یکا متعامد از ویژه‌بردارهای A خواهد شد، زیرا Q طول و زوایا را حفظ می‌کند (مسألهٔ ۱۳ فصل ۷ را ببینید).

مثال گیریم

$$A = \begin{bmatrix} ۱ & ۳ & ۴ \\ ۳ & ۱ & ۲ \\ ۴ & ۲ & ۱ \end{bmatrix}$$

پس

$$w^{(۲)} = \left[۰, \frac{۲}{\sqrt{۵}}, \frac{۱}{\sqrt{۵}} \right]^T$$

$$P_1 = \begin{bmatrix} ۱ & ۰ & ۰ \\ ۰ & -\frac{۳}{۵} & -\frac{۴}{۵} \\ ۰ & -\frac{۴}{۵} & \frac{۳}{۵} \end{bmatrix}$$

$$T = P_1^T A P_1 = \begin{bmatrix} ۱ & -۵ & ۰ \\ -۵ & \frac{۷۳}{۲۵} & \frac{۱۴}{۲۵} \\ ۰ & \frac{۱۴}{۲۵} & -\frac{۲۳}{۲۵} \end{bmatrix}$$

برای تجزیهٔ خطای این تبدیل به شکل سه‌قطری، قضایایی از ویلکینسن (۱۹۶۵) را در ذیل می‌آوریم. گیریم حساب رایانه، ممیز شناور دودویی و گردکردن با t رقم دودویی در جزء کسری باشد. به‌علاوه فرض می‌کنیم تمام حاصلضربهای داخلی

$$\sum_{i=1}^{m} a_i b_i$$

که در محاسبات پیش می‌آیند با دقت مضاعف انباشته شده و با دقّت ساده هنگام مجموعیابی گرد شده‌اند. این حاصلضربهای داخلی در جاهای گوناگونی در محاسبهٔ T از روی A وجود دارند. گیریم $\hat{T}$ معرّف ماتریس سه‌قطری متقارن است که از A با حساب رایانهٔ قبلی حساب شده است. فرض می‌کنیم $\hat{P}_r$ ماتریسی باشد که از تبدیل A_{r-1} به A_r حاصل شده است. P_r تبدیل دقیق نظری این ماتریس باشد بدون هیچ خطای گردکردن و $Q = P_1 \dots P_{r-1}$ حاصلضرب دقیق این P_r ها باشد که یک ماتریس متعامد است.

قضیهٔ ۴.۹ فرض می‌کنیم A یک ماتریس حقیقی متقارن از مرتبهٔ n باشد. گیریم $\hat{T}$ یک ماتریس حقیقی متقارن سه‌قطری باشد که از اعمال تبدیلات تشابهی (۲۰.۳.۹) هاوسهولدر بر A، مانند (۲۱.۳.۹)، به‌دست آمده است. فرض می‌کنیم حساب ممیّز شناوری که استفاده‌شده دارای همان ویژگیهای مذکور در پاراگراف قبل است. اگر $\{\lambda_i\}$ و $\{\tau_i\}$ به‌ترتیب ویژه‌مقدارهای A و $\hat{T}$ باشند که به‌طور صعودی مرتب شده‌اند، آنگاه

$$\left[\frac{\sum_{i=1}^{n}(\tau_i - \lambda_i)^2}{\sum_{i=1}^{n}\lambda_i^2}\right]^{1/2} \leq c_n 2^{-t} \qquad (۲۵.۳.۹)$$

که در آن

$$c_n = 25(n-1)[1 + (12{,}36)2^{-t}]^{2n-4}$$

برای مقادیر کوچک و متوسط n، $c_n \doteq 25(n-1)$.

برهان از ویلکینسن (۱۹۶۵، ص ۱۶۱)، با استفاده از ماتریس فروبنیوس با نرم F،

$$F(\hat{T} - Q^T A Q) \leq 2x(n-1)(1+x)^{2n-4}F(A) \qquad (۲۶.۳.۹)$$

که در آن $x = (۱۲٫۳۶)۲^{-t}$. از نتیجهٔ (۱۹.۱.۹) قضیهٔ ۳.۹ ویلانت ـ هوفمن داریم

$$\left[\sum_{i=۱}^{n}(\tau_i - \lambda_i)^۲\right]^{۱/۲} \leq F(\hat{T} - Q^T A Q) \qquad (۲۷.۳.۹)$$

زیرا ویژه‌مقدارهای A و $Q^T A Q$ یکی هستند. و از مسألهٔ ۲۸ (ب) فصل ۷،

$$F(A) = \left[\sum_{i=۱}^{n}\lambda_i^۲\right]^{۱/۲}$$

از ترکیب این نتایج، (۲۵.۳.۹) به‌دست می‌آید. ■

برای بحث بیشتر از خطا، از جمله حالتی که حاصلضربهای داخلی با دقت مضاعف ذخیره نشده باشند، ویلکینسن (۱۹۶۵، صص ۲۹۷ـ۲۹۹) را ملاحظه نمایید. قضیهٔ (۲۵.۳.۹) نشان می‌دهد که تبدیل به یک شکل سه‌قطری، عملی است بسیار پایدار که خطای بسیار کم جدیدی در ویژه‌مقدارها به دنبال خواهد داشت.

ماتریسهای متعامد دوران در صفحه رده‌های ماتریسهای متعامد دیگری وجود دارند که می‌توان به‌جای ماتریسهای هاؤسهولدر از آنها استفاده کرد. ردهٔ اصلی، مجموعهٔ دورانها در صفحه‌اند، که دوران دستگاه مختصات در صفحه به زاویهٔ داده‌شدهٔ θ در صفحهٔ مختصات یک تعبیر هندسی آن است. برای اعداد صحیح k، l، $۱ \leq k < l \leq n$، ماتریس متعامد $n \times n$، $R^{(k,l)}$ را با تعویض چهار عنصر ماتریس یکهٔ I_n تعریف می‌کنیم. به‌ازای هر عدد حقیقی θ، عناصر $R^{(k,l)}$ را چنین تعریف می‌کنیم

$$R_{i,j}^{(k,l)} = \begin{cases} \cos\theta & (i,j) = (k,k) \quad \text{یا} \quad (l,l) \\ \sin\theta & (i,j) = (k,l) \\ -\sin\theta & (i,j) = (l,k) \\ (I_n)_{ij} & \text{برای همهٔ مقادیر دیگرِ } (i,j) \end{cases} \qquad (۲۸.۳.۹)$$

به‌ازای $۱ \leq i,j \leq n$.

مثال برای $n = ۳$

$$R^{(۱,۳)} = \begin{bmatrix} \cos\theta & ۰ & \sin\theta \\ ۰ & ۱ & ۰ \\ -\sin\theta & ۰ & \cos\theta \end{bmatrix}$$

به‌عنوان یک حالت خاص، θ را مساوی $\frac{\pi}{۴}$ می‌گیریم. پس

$$R^{(۱,۳)} = \begin{bmatrix} \frac{۱}{\sqrt{۲}} & ۰ & \frac{۱}{\sqrt{۲}} \\ ۰ & ۱ & ۰ \\ -\frac{۱}{\sqrt{۲}} & ۰ & \frac{۱}{\sqrt{۲}} \end{bmatrix}$$

دورانهای $R^{(k,l)}$ صفحه‌یی را می‌توان برای همان تبدیلهایی که با ماتریسهای هاؤسهولدر انجام می‌شوند، مورد استفاده قرار داد. در بسیاری مواقع، ماتریسهای هاؤسهولدر کاراترند، ولی دورانهای صفحه‌یی برای قسمتی از روش QR که در بخش ۵.۹ توضیح داده شده است، کارایی بیشتری دارند. فکر حل مسألهٔ ویژه‌مقدار ماتریسهای متقارن با تبدیل آن در ابتدا به یک شکل سه‌قطری را گیونز[1] در سال ۱۹۵۴ عرضه کرده است. همچنین روشهای محاسبهٔ ویژه‌مقدارهای یک ماتریس سه‌قطری را که در بخش بعد توضیح داده می‌شود، او پیشنهاد نموده است. گیونز دورانهای $R^{(k,l)}$ صفحه را به‌کار برده است و ماتریسهای هاؤسهولدر در ۱۹۵۸ توسط هاؤسهولدر معرفی شده‌اند. برای بحث بیشتر در مورد ماتریسهای دوران و ویژگیهای آنها، گلوب و وَن‌لون (۱۹۸۳، بخش ۴.۳)، پارلت ۱۹۸۰، بخش ۴.۶)، و همچنین مسائل ۱۵ و ۱۷ (ب) را ملاحظه نمایید.

۴.۹ ویژه‌مقدارهای یک ماتریس متقارن سه‌قطری

گیریم T یک ماتریس متقارن سه قطری از مرتبهٔ n، مانند (۲۲.۳.۹) باشد. چندجمله‌یی مشخّصهٔ T را محاسبه و از آن برای پیداکردن ویژه‌مقدارهای T استفاده می‌کنیم. برای محاسبهٔ

$$f_n(\lambda) \equiv \det(T - \lambda I) \qquad (۱.۴.۹)$$

دنبالهٔ

1. W. Givens

$$f_k(\lambda) = \det \begin{bmatrix} \alpha_1 - \lambda & \beta_1 & 0 & \dots & 0 \\ \beta_1 & \alpha_2 - \lambda & \beta_2 & & \vdots \\ 0 & & & \ddots & \\ \vdots & & & & \beta_{k-1} \\ 0 & \dots & & \beta_{k-1} & \alpha_k - \lambda \end{bmatrix} \qquad (۲.۴.۹)$$

را برای $1 \le k \le n$ و $f_0(\lambda) \equiv 1$ معرفی می‌کنیم. با محاسبهٔ مستقیم،

$$\begin{aligned} f_1(\lambda) &= \alpha_1 - \lambda \\ f_2(\lambda) &= (\alpha_2 - \lambda)(\alpha_1 - \lambda) - \beta_1^2 \\ &= (\alpha_2 - \lambda) f_1(\lambda) - \beta_1^2 f_0(\lambda) \end{aligned}$$

فرمول $f_2(\lambda)$ یک رابطهٔ بازگشتی کلّی سه‌گانه را، که دنبالهٔ $\{f_k(\lambda)\}$ در آن صدق می‌کند نشان می‌دهد:

$$f_k(\lambda) = (\alpha_k - \lambda) f_{k-1}(\lambda) - \beta_{k-1}^2 f_{k-2}(\lambda) \qquad 2 \le k \le n \qquad (۳.۴.۹)$$

برای اثبات این رابطه، دترمینان (۲.۴.۹) را برحسب آخرین سطر آن با استفاده از دترمینانهای فرعی بسط می‌دهیم و نتیجه به آسانی به‌دست می‌آید. روش محاسبهٔ $f_n(\lambda)$، پس از آنکه ضرایب $\{\beta_k^2\}$ محاسبه شده باشند، به $3 - 2n$ ضرب نیاز دارد.

مثال گیریم

$$T = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 \end{bmatrix} \qquad (۴.۴.۹)$$

پس

$$\begin{aligned} &f_0(\lambda) = 1 \qquad f_1(\lambda) = 2 - \lambda \\ &f_j(\lambda) = (2 - \lambda) f_{j-1}(\lambda) - f_{j-2}(\lambda) \qquad j = 2, 3, 4, 5, 6 \end{aligned} \qquad (۵.۴.۹)$$

بدون استفاده از رابطهٔ بازگشتی سه‌گانهٔ (۵.۴.۹)، محاسبهٔ $f_6(\lambda)$ بسیار پیچیده‌تر خواهد بود.

در اینجا، می‌توانیم مسأله را حل‌شده تلقی نماییم زیرا $f_n(\lambda)$ یک چندجمله‌یی است و روشهای زیادی برای به‌دست‌آوردن ریشه‌های یک چندجمله‌یی وجود دارند. یا ممکن است روش کلیتری، چون روش خط قاطع یا روش برنت را که هر دو در فصل ۲ تشریح شدند به‌کار بریم. ولی دنبالهٔ $\{f_k(\lambda) \mid 0 \leq k \leq n\}$ دارای ویژگیهای خاصی است که از آن یک **دنبالهٔ استورم** می‌سازد و این ویژگیها جداکردن ویژه‌مقدارهای T را، نسبتاً آسان می‌کند. به محض اینکه ویژه‌مقدارها جدا شدند، یک روش مانند روش برنت [بخش ۸.۲ را ببینید] را می‌توان برای محاسبهٔ سریع ریشه‌ها به‌کار برد. نظریهٔ دنباله‌های استورم در هنریچی (۱۹۷۴، ص ۴۴۴) توضیح داده شده است، ولی ما فقط حالت خاصّ $\{f_k(\lambda)\}$ را مطالعه می‌کنیم.

قبل از بیان نتایج نظریهٔ استورم برای $\{f_k(\lambda)\}$، می‌خواهیم ببینیم که اگر $\beta_l = 0$، چه اتفاقی خواهد افتاد. در این صورت مسألهٔ ویژه‌مقدار را می‌توان به دو مسألهٔ ویژه‌مقدار مجزّای کوچکتر از مرتبه‌های l و $n-l$ تجزیه کرد. به‌عنوان یک مثال

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & 0 & 0 \\ \beta_1 & \alpha_2 & 0 & 0 & 0 \\ 0 & 0 & \alpha_3 & \beta_3 & 0 \\ 0 & 0 & \beta_3 & \alpha_4 & \beta_4 \\ 0 & 0 & 0 & \beta_4 & \alpha_5 \end{bmatrix}$$

را درنظر می‌گیریم. آن را به دو بلوک T_1 و T_2، به‌ترتیب از مرتبه‌های ۲ و ۳، در امتداد قطر می‌شکنیم، و سپس می‌نویسیم

$$T = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}$$

از این رابطه داریم

$$\det[T - \lambda I_5] = \det[T_1 - \lambda I_2] \det[T_2 - \lambda I_3]$$

و می‌توانیم ویژه‌مقدارهای T را با محاسبهٔ ویژه‌مقدارهای T_1 و T_2 به‌دست آوریم. مسألهٔ ویژه‌بردار را نیز می‌توان به شکل مشابهی حل کرد. مثلاً، اگر $T_1\hat{x} = \lambda\hat{x}$ با $\hat{x} \neq 0$ در $\mathbf{R}^2$ باشد، تعریف می‌کنیم

$$x = [\hat{x}^T, 0, 0, 0]^T$$

در این صورت $Tx = \lambda x$. این ساختمان را می‌توان برای محاسبهٔ مجموعهٔ کامل ویژه‌بردارهای T

با استفاده از ویژه‌بردارهای T_1 و T_2 به‌کار برد. در بقیهٔ این بخش فرض می‌کنیم که در ماتریس T، β_l ها مخالف صفرند. با این فرض، تمام ویژه‌مقدارهای T ریشه‌های سادهٔ $f_n(\lambda)$ خواهند بود.

ویژگی دنبالهٔ استورم برای $\{f_k(\lambda)\}$ دنباله‌های $\{f_k(a)\}$ و $\{f_k(b)\}$ را می‌توان برای تعیین تعداد ریشه‌های $f_n(\lambda)$ که در بازهٔ $[a,b]$ واقع‌اند به‌کار برد. برای این کار، تابع صحیح‌المقدار $s(\lambda)$ را به شرح زیر معرفی می‌کنیم. $s(\lambda)$ را تعداد تطابق علامتهای اعضای متوالی دنبالهٔ $\{f_k(\lambda)\}$ تعریف می‌کنیم، و اگر مقدار عضوی نظیر $f_j(\lambda)$ مساوی صفر باشد، علامت آن را مخالف علامت $f_{j-1}(\lambda)$ می‌گیریم. می‌توان نشان داد که $f_j(\lambda) = 0$ ایجاب می‌کند که $f_{j-1}(\lambda) \neq 0$.

مثال دنبالهٔ $f_0(\lambda)$، ...، $f_6(\lambda)$ در (۵.۴.۹) در آخرین مثال را درنظر می‌گیریم. به‌ازای $\lambda = 3$،

$$(f_0(\lambda), \ldots, f_6(\lambda)) = (1, -1, 0, 1, -1, 0, 1)$$

دنبالهٔ علامتهای متناظر آن عبارت است از

$$(+, -, +, +, -, +, +)$$

و $s(3) = 2$.

اکنون قضیّهٔ اساسی را که در محاسبهٔ ریشه‌های $f_n(\lambda)$ و بنابراین در محاسبهٔ ویژه‌مقدارهای T به‌کار رفته است بیان می‌کنیم. برهان از نظریّهٔ کلی که در هنریچی (۱۹۷۴) داده شده است نتیجه می‌شود.

قضیهٔ ۵.۹ گیریم T یک ماتریس متقارن حقیقی سه‌قطری از مرتبهٔ n، مانند (۲۲.۳.۹) باشد. فرض می‌کنیم دنبالهٔ $\{f_k(\lambda) \mid 0 \leq k \leq n\}$ مطابق با (۲.۴.۹) تعریف‌شده، و به‌ازای $l = 1, \ldots, n-1$، و کلیهٔ β_l ها مخالف صفر باشند. در این صورت تعداد ریشه‌های $f_n(\lambda)$ که از $\lambda = a$ بزرگترند برابر است با $s(a)$ که در پاراگراف قبل تعریف شده است. برای $a < b$، تعداد ریشه‌ها در بازهٔ $a < \lambda \leq b$ برابر با $s(a) - s(b)$ خواهد بود.

محاسبهٔ ویژه‌مقدارها قضیهٔ ۵.۹ ابزار اساسی در تعیین جا و جداکردن ریشه‌های $f_n(\lambda)$ است. ابتدا، بازه‌ای را که شامل تمام ریشه‌ها باشد حساب می‌کنیم. با استفاده از قضیهٔ ۱.۹ دایرهٔ گرشگورین،

همهٔ ویژه‌مقدارها در بازهٔ $[a, b]$، با

$$a = \underset{1 \le i \le n}{\text{Min}} \{\alpha_i - | \beta_i | - | \beta_{i-1} |\}$$
$$b = \underset{1 \le i \le n}{\text{Max}} \{\alpha_i + | \beta_i | + | \beta_{i-1} |\}$$

که $\beta_n = \beta_0 = 0$، قرار دارند.

از روش نیمسازی در بازهٔ $[a, b]$ استفاده می‌کنیم تا آن را به زیربازه‌های کوچکتری تقسیم کنیم. با استفاده از قضیهٔ ۵.۹ معلوم می‌کنیم که چند ریشه در هر زیربازه واقع است و زیربازه‌هایی را جستجو می‌کنیم که هرکدام شامل یک ریشه باشد. اگر بعضی از ویژه‌مقدارها تقریباً مساوی باشند، تقسیم زیر بازه‌ها را آنقدر ادامه می‌دهیم تا ریشه با دقت کافی به‌دست آید. وقتی معلوم شد که زیربازه تنها شامل یک ریشه است، می‌توانیم از یک روش همگرای سریعتر استفاده کنیم.

مثال باز هم مثال (۴.۴.۹) را درنظر می‌گیریم. به موجب قضیهٔ گرشگورین، ۱.۹، تمام ویژه‌مقدارها در $[0, 4]$ واقع‌اند. و به آسانی می‌توان تحقیق کرد که نه $\lambda = 0$ یک ویژه‌مقدار است و نه $\lambda = 4$. یک روند منظّم نیمسازی در $[0, 4]$ انجام گرفته است تا شش ویژه‌مقدار $f_6(\lambda)$ در زیربازه‌های مجزا قرارگیرند. نتایج در جدول ۲.۹ به ترتیب به‌دست آمده، نشان داده شده‌اند. ریشه‌ها مانند زیراندیس‌گذاری شده‌اند:

$$0 \le \lambda_6 \le \lambda_5 \le \ldots \le \lambda_1 \le 4$$

ریشه‌ها را می‌توان با ادامهٔ روش نیمسازی پیدا کرد، که در این صورت به قضیهٔ ۵.۹ دیگر نیازی نیست. ولی بهتر است از بعضی روشهای دیگر ریشه‌یابی استفاده کرد.

اگرچه تمام ریشه‌های یک ماتریس سه‌قطری را می‌توان با این راهکار پیدا کرد، ولی معمولاً در چنین حالتی استفاده از الگوریتم QR که در بخش بعد می‌آید، سریعتر جوابها را به ما خواهد

جدول ۲.۹ مثال در استفاده از قضیهٔ ۵.۹

λ	$f_6(\lambda)$	$s(\lambda)$	شرح
۰٫۰	۷٫۰	۶	$\lambda_6 > 0$
۴٫۰	۷٫۰	۰	$\lambda_1 < 4$
۲٫۰	−۱٫۰	۳	$\lambda_4 < 2 < \lambda_3$
۱٫۰	۱٫۰	۴	$\lambda_5 < 1 < \lambda_4 < 2$
۰٫۵	−۱٫۴۲۱۸۷۵	۵	$0 < \lambda_6 < 0.5 < \lambda_5 < 1$
۳٫۰	۱٫۰	۲	$2 < \lambda_3 < 3 < \lambda_2$
۳٫۵	۱٫۴۲۱۸۷۵	۱	$3 < \lambda_2 < 3.5 < \lambda_1 < 4$

داد. برای ماتریسهای بزرگ، معمولاً تمام ریشه‌ها موردنظر نیستند، در چنین حالتی، روشی که در این بخش توضیح داده‌شده ترجیح داده می‌شود. اگر فقط بعضی از ریشه‌های مشخص، مثلاً، پنج ریشهٔ از همه بزرگتر یا همهٔ ریشه‌ها در یک بازهٔ داده‌شده، یا تمام ریشه‌های واقع در $[۱, ۳]$ را بخواهیم، آنگاه با استفاده از قضیهٔ ۵.۹، به‌آسانی می‌توان آنها را جایابی کرد.

۵.۹ روش QR

امروزه، روش QR کاراترین و متداولترین روش برای به‌دست‌آوردن همهٔ ویژه‌مقدارهای یک ماتریس است. این روش ابتدا توسط فرانسیس[1] در سال ۱۹۶۱ انتشار یافت و تاکنون به‌شدت موضوعی تحت بررسی قرار گرفته است. روش QR هم از لحاظ نظری و هم از لحاظ کاربردی، کاملاً پیچیده است، و ما فقط می‌توانیم مقدمه‌ای از نظریه را بیاوریم. برای الگوریتمهای اصلی، هم برای ماتریسهای متقارن و هم نامتقارن، خوانندگان را به آنچه کـه در EISPACK و ولکینسن و راینش (۱۹۷۱) آمده است ارجاع می‌دهیم.

برای ماتریس داده شدهٔ A، یک تجزیهٔ

$$A = QR$$

وجود دارد که R بالامثلثی و Q متعامد است. برای A ی حقیقی، هم Q و هم R را می‌توان حقیقی انتخاب کرد. ساختمان آنها در بخش ۳.۹ داده شده است. ما در سرتاسر این بخش فرض می‌کنیم A حقیقی است. گیریم $A_۱ = A$، و دنبالهٔ ماتریسهای A_m، Q_m و R_m را چنین تعریف می‌کنیم

$$A_m = Q_m R_m \qquad A_{m+۱} = R_m Q_m \qquad m = ۱, ۲, \ldots \qquad (۱.۵.۹)$$

چون $R_m = Q_m^T A_m$، داریم

$$A_{m+۱} = Q_m^T A_m Q_m \qquad (۲.۵.۹)$$

ماتریس $A_{m+۱}$ از نظر تعامد مشابه با A_m، و بنابراین با استقرا، مشابه با $A_۱$ است.

دنبالهٔ $\{A_m\}$ یا به یک ماتریس مثلثی می‌گراید که دارای ویژه‌مقدارهای A بر قطر اصلی خود است، یا به یک ماتریس تقریباً ـ مثلثی که ویژه‌مقدارهای آن به‌سادگی قابل‌محاسبه است، می‌گراید. در این حالت، معمولاً همگرایی کند است و یک تکنیکی برای تسریع همگرایی به نام انتقال به‌کار برده می‌شود. روش انتقال در بخش بعد معرفی و توضیح داده می‌شود.

1. J. G. F. Francis

مثال گیریم

$$A_۱ = \begin{bmatrix} ۲ & ۱ & ۰ \\ ۱ & ۳ & ۱ \\ ۰ & ۱ & ۴ \end{bmatrix} \qquad (۳.۵.۹)$$

ویژه‌مقدارها عبارت‌اند از:

$$\lambda_۱ = ۳ + \sqrt{۳} \doteq ۴٫۷۳۲۱ \quad \lambda_۲ = ۳٫۰ \quad \lambda_۳ = ۳ - \sqrt{۳} \doteq ۱٫۲۶۷۹$$

بارستهای A_m به سرعت همگرا نمی‌شوند، و فقط چند تا از آنها برای نشان‌دادن رفتار کیفی همگرایی آنها داده شده‌اند

$$A_۲ = \begin{bmatrix} ۳٫۰۰۰۰ & ۱٫۰۹۵۴ & ۰ \\ ۱٫۰۹۵۴ & ۳٫۰۰۰۰ & -۱٫۳۴۱۶ \\ ۰ & -۱٫۳۴۱۶ & ۳٫۰۰۰۰ \end{bmatrix} \quad A_۳ = \begin{bmatrix} ۳٫۷۰۵۹ & ۰٫۹۵۵۸ & ۰ \\ ۰٫۹۵۵۸ & ۳٫۵۲۱۴ & ۰٫۹۷۳۸ \\ ۰ & ۰٫۹۷۳۸ & ۱٫۷۷۲۷ \end{bmatrix}$$

$$A_۷ = \begin{bmatrix} ۴٫۶۷۹۲ & ۰٫۲۹۷۹ & ۰ \\ ۰٫۲۹۷۹ & ۳٫۰۵۲۴ & ۰٫۰۲۷۴ \\ ۰ & ۰٫۰۲۷۴ & ۱٫۲۶۸۴ \end{bmatrix} \quad A_۸ = \begin{bmatrix} ۴٫۷۱۰۴ & ۰٫۱۹۲۴ & ۰ \\ ۰٫۱۹۲۴ & ۳٫۰۲۱۶ & -۰٫۰۱۱۵ \\ ۰ & -۰٫۰۱۱۵ & ۱٫۲۶۸۰ \end{bmatrix}$$

$$A_۹ = \begin{bmatrix} ۴٫۷۲۳۳ & ۰٫۱۲۲۹ & ۰ \\ ۰٫۱۲۲۹ & ۳٫۰۰۸۷ & ۰٫۰۰۴۸ \\ ۰ & ۰٫۰۰۴۸ & ۱٫۲۶۸۰ \end{bmatrix} \quad A_{۱۰} = \begin{bmatrix} ۴٫۷۲۸۵ & ۰٫۰۷۸۱ & ۰ \\ ۰٫۰۷۸۱ & ۳٫۰۰۳۵ & -۰٫۰۰۲۰ \\ ۰ & -۰٫۰۰۲۰ & ۱٫۲۶۸۰ \end{bmatrix}$$

عناصر واقع در موضع $(۱, ۲)$ از لحاظ هندسی با نسبتی در حدود ۰٫۶۴ در هر بارست، کاهش می‌یابند و عناصری که در موضع $(۲, ۳)$ هستند با نسبتی در حدود ۰٫۴۲ در هر بارست کاهش می‌یابند. مقدار موضع $(۳, ۳)$ در ماتریس $A_{۱۵}$ برابر ۱٫۲۶۷۹ خواهد بود که تا ۵ رقم درست است.

تبدیل مقدّماتی A به شکل ساده‌تر روش QR نسبتاً گران است، زیرا تجزیهٔ QR وقتی چند بار تکرار شود وقتگیر است. برای کاهش هزینه، ماتریسی را که برای روش QR آماده می‌کنند، نخست به شکل ساده‌تری، که برای تجزیهٔ QR کمتر هزینه دارد، تبدیل می‌کنند.

اگر A متقارن باشد، آن را به یک ماتریس سه‌قطری متقارن متشابه، درست همانگونه که در بخش ۳.۹ تشریح شد، بدل می‌کنند. اگر A نامتقارن باشد، آن را به یک **ماتریس هِسِنبرگ** متشابه،

بدل می‌کنند. ماتریس B هسنبرگ است اگر

$$b_{ij} = ۰ \quad \text{به‌ازای جمیع مقادیر } i > j + ۱ \tag{۴.۵.۹}$$

این ماتریس بالامثلثی است جز اینکه فقط یک زیرقطری ناصفر دارد. ماتریس A با استفاده از همان الگوریتمی که برای تبدیل ماتریسهای متقارن به شکل سه‌قطری به‌کار می‌رفت، به ماتریس هسنبرگ تبدیل می‌شود.

وقتی A سه‌قطری یا هسنبرگ باشد، ماتریسهای هاؤسهولدر بخش ۳.۹، در به‌دست‌آوردن تجزیهٔ QR، شکل ساده‌ای به خود می‌گیرند. ولی معمولاً دورانهای (۲۸.۳.۹) صفحه به‌جای ماتریسهای هاؤسهولدر به‌کار برده می‌شوند زیرا برای محاسبه و کاربرد در چنین وضعیتی، کارایی بیشتری دارند. وقتی $A_۱ = Q_۱R_۱$ و $A_۲ = R_۱Q_۱$ به‌دست آمدند، نیاز داریم بدانیم که شکل $A_۲$ همانند شکل $A_۱$ است تا با استفاده از شکل ارزانتر تجزیهٔ QR، کار را ادامه دهیم.

فرض می‌کنیم $A_۱$ به شکل ماتریس هسنبرگ است. با توجّه به بخش ۳.۹، $Q_۱$ در تجزیهٔ $A_۱ = Q_۱R_۱$ دارای مقدار زیر است:

$$Q_۱ = H_۱ \dots H_{n-۱} \tag{۵.۵.۹}$$

که هریک از H_kها، یک ماتریس هاؤسهولدر (۱۲.۳.۹) است:

$$H_k = I - ۲w^{(k)}w^{(k)T} \qquad ۱ \le k \le n - ۱ \tag{۶.۵.۹}$$

چون ماتریس $A_۱$ به شکل ماتریس هسنبرگ است، می‌توان نشان داد که بردارهای $w^{(k)}$ شکل خاص زیر را دارند.

$$w_i^{(k)} = ۰ \quad \text{برای } i < k \text{ و } i > k + ۱ \tag{۷.۵.۹}$$

این مطلب را می‌توان از روی معادلات مؤلفه‌های $w^{(k)}$، به‌ویژه (۱۰.۳.۹)، ثابت کرد. با توجّه به (۷.۵.۹)، ماتریس H_k فقط در چهار عنصر در مواضع (k,k)، $(k, k+۱)$، $(k+۱, k)$ و $(k+۱, k+۱)$ با ماتریس یکه، تفاوت دارد. و از این‌رو یک، محاسبهٔ نسبتاً ساده نشان می‌دهد که $Q_۱$ باید شکل هسنبرگ داشته باشد. یک لم لازم دیگر این است که ضرب یک ماتریس بالامثلثی و یک ماتریس هسنبرگ باز هم یک ماتریس هسنبرگ است. فقط دو شکل ماتریسها را در هم ضرب و طرز قرارگرفتن صفرهای متناظر را مشاهده کنید تا بتوانید لم را اثبات کنید. از ترکیب این

نتایج، و توجه به این نکته که R_1 یک ماتریس بالامثلثی است، ملاحظه می‌کنیم که $A_2 = R_1 Q_1$ به شکل هسنبرگ است.

اگر A_1 متقارن و سه‌قطری باشد، به‌طور نمایان دیده می‌شود که هسنبرگ است. با توجّه به قضیّهٔ قبلی، A_2 نیز هسنبرگ است. ولی A_2 متقارن است، زیرا

$$A_1^T = (Q_1^T A_1 Q_1)^T = Q_1^T A_1^T Q_1 = Q_1^T A_1 Q_1 = A_2$$

چون هر ماتریس هسنبرگ متقارن سه‌قطری است، نشان داده‌ایم که A_2 سه‌قطری است. ملاحظه می‌کنیم که بارستهای مثال (۳.۵.۹) این نتیجه را نشان می‌دهند.

همگرایی روش QR قضایای همگرایی روش QR را می‌توان در گلوب و وَن‌لون (۱۹۸۳، بخشهای ۵.۷ و ۲.۸)، پارلت (۱۹۶۸) و (۱۹۸۰، فصل ۸) و ویلکینسن (۱۹۶۵، فصل ۸) ملاحظه کرد. قضیه زیر از مرجع آخری گرفته شده است.

قضیهٔ ۶.۹ گیریم A یک ماتریس حقیقی از مرتبهٔ n باشد، و ویژه‌مقدارهای $\{\lambda_i\}$ ی آن، در

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0 \qquad (۸.۵.۹)$$

صدق کنند. در این صورت بارستهای R_m از روش QR، که در (۱.۵.۹) تعریف شدند، به یک ماتریس بالامثلثی D، که ویژه‌مقدارهای $\{\lambda_i\}$ عناصر قطر آن هستند، همگرا می‌شوند. اگر A متقارن باشد، دنبالهٔ $\{A_m\}$ به یک ماتریس قطری همگرا می‌شود. برای سرعت همگرایی، داریم

$$\| D - A_m \| \leq c \operatorname{Max}_i \left| \frac{\lambda_{i+1}}{\lambda_i} \right| \qquad (۹.۵.۹)$$

به‌عنوان یک مثال از این کران خطا، مثال (۳.۵.۹) را درنظر می‌گیریم. نسبتهای ویژه‌مقدارهای متوالی در آن، عبارت‌اند از

$$\frac{\lambda_2}{\lambda_1} \doteq ۰٫۶۳ \qquad \frac{\lambda_3}{\lambda_2} \doteq ۰٫۴۲ \qquad (۱۰.۵.۹)$$

اگر هریک از عناصر ناقطر A_m را در این مثال امتحان کنیم، ملاحظه می‌کنیم که آنها با یکی از دو نسبت (۱۰.۵.۹)، کاهش می‌یابند.

برای ماتریسهایی که ویژه‌مقدارهای آنها، در (۸.۵.۹) صدق نمی‌کنند، بارستهای A_m ممکن است به یک ماتریس مثلثی همگرا نشوند. اگر A متقارن باشد، دنبالهٔ $\{A_m\}$ به یک ماتریس

قطری بلوکی زیر همگرا می‌شود

$$A_m \to D = \begin{bmatrix} B_1 & & & 0 \\ & B_2 & & \\ & & \ddots & \\ 0 & & & B_r \end{bmatrix} \qquad (۱۱.۵.۹)$$

که در آن همهٔ بلوکهای B_i از مرتبه یک یا دو هستند. بنابراین ویژه‌مقدارهای A را، به‌آسانی از روی ویژه‌مقدارهای D می‌توان محاسبه کرد، اگر A حقیقی و نامتقارن باشد، وضعیت پیچیده‌تر ولی قابل قبول است. برای بحث در این مورد، ویلکینسن (۱۹۶۵، فصل ۸) و پارلت (۱۹۶۸) را ملاحظه کنید.

برای اینکه ببینیم $\{A_m\}$ همیشه به یک ماتریس قطری همگرا نمی‌شود، مثال متقارن سادهٔ زیر را درنظر می‌گیریم

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

ویژه‌مقدارهای آن، $\lambda = \pm 1$ اند. چون A متعامد است داریم

$$Q_1 = A \quad R_1 = I \quad \text{که} \quad A = Q_1 R_1$$

و بنابراین

$$A_2 = R_1 Q_1 = A$$

و همهٔ بارستها $A_m = A$. دنبالهٔ $\{A_m\}$ به یک ماتریس قطری میل نخواهد کرد.

روش QR با انتقال الگوریتم QR معمولاً با یک انتقال مبدأ برای ویژه‌مقدارها به‌کار برده می‌شود تا سرعت همگرایی افزایش یابد. برای یک دنباله از مقادیر ثابت $\{c_m\}$، تعریف می‌کنیم $A_1 = A$ و

$$\begin{aligned} A_m - c_m I &= Q_m R_m \\ A_{m+1} &= c_m I + R_m Q_m \qquad m = 1, 2, \ldots \end{aligned} \qquad (۱۲.۵.۹)$$

ماتریسهای A_m با A_1 متشابه‌اند، زیرا

$$\begin{aligned} R_m &= Q_m^T (A_m - c_m I) \\ A_{m+1} &= c_m I + Q_m^T (A_m - c_m I) Q_m \\ &= c_m I + Q_m^T A_m Q_m - c_m I \\ A_{m+1} &= Q_m^T A_m Q_m \qquad m \geq 1 \end{aligned} \qquad (۱۳.۵.۹)$$

ویژه‌مقدارهای A_{m+1} همان ویژه‌مقدارهای A_m اند، و بنابراین با ویژه‌مقدارهای A یکی هستند.

برای اینکه در انتخاب انتقالهای $\{c_m\}$ دقیقتر باشیم، فقط یک ماتریس متقارن سه‌قطری A را درنظر می‌گیریم. برای A_m، گیریم

$$A_m = \begin{bmatrix} \alpha_1^{(m)} & \beta_1^{(m)} & 0 & \cdots & 0 \\ \beta_1^{(m)} & \alpha_2^{(m)} & \beta_2^{(m)} & & \vdots \\ 0 & \ddots & & & \\ \vdots & & & & \beta_{n-1}^{(m)} \\ 0 & & \cdots & \beta_{n-1}^{(m)} & \alpha_n^{(m)} \end{bmatrix} \qquad (۱۴.۵.۹)$$

دو روش برای انتخاب $\{c_m\}$ وجود دارد: (۱) می‌گیریم $c_m = \alpha_n^{(m)}$ و (۲) c_m را آن ویژه‌مقدار ماتریس

$$\begin{bmatrix} \alpha_{n-1}^{(m)} & \beta_{n-1}^{(m)} \\ \beta_{n-1}^{(m)} & \alpha_n^{(m)} \end{bmatrix} \qquad (۱۵.۵.۹)$$

می‌گیریم که به $\alpha_n^{(m)}$ نزدیکتر است. روش دومی بهتر است، ولی در هر دو حالت ماتریسهای A_m به یک ماتریس قطری بلوکی میل می‌کنند که بلوکها، مانند (۱۱.۵.۹)، از مراتب یک یا دو هستند. می‌توان نشان داد که هریک از دو انتخاب $\{c_m\}$ تضمین می‌کند که

$$\beta_{n-1}^{(m)}\beta_{n-2}^{(m)} \to 0 \qquad \text{وقتی } m \to \infty \qquad (۱۶.۵.۹)$$

که معمولاً همگرایی با نرخی سریعتر از روش اوّلیهٔ QR (۱.۵.۹) صورت می‌گیرد.

از (۱۳.۵.۹) و با استفاده از نرم ماتریس عملگر (۱۹.۳.۷) و مسألهٔ ۲۷ (ج) فصل ۷، داریم:

$$\| A_{m+1} \|_2 = \| Q_m^T A_m Q_m \|_2 = \| A_m \|_2$$

ماتریسهای $\{A_m\}$ یکنواخت ـ کراندارند و نتیجتاً همین ویژگی برای هریک از عناصر آنها نیز برقرار است. از (۱۶.۵.۹) و یکنواخت ـ کراندار بودن $\{\beta_{n-1}^{(m)}\}$ و $\{\beta_{n-2}^{(m)}\}$، وقتی $m \to \infty$، یا $\beta_{n-1}^{(m)} \to 0$ یا $\beta_{n-2}^{(m)} \to 0$. در حالت اول، $\alpha_n^{(m)}$ به یک ویژه‌مقدار A همگرا می‌شود. و در حالت دوم، دو ویژه‌مقدار را می‌توان با استفاده از حد زیرماتریس (۱۵.۵.۹) به‌آسانی به‌دست آورد.

وقتی یک یا دو ویژه‌مقدار به این علّت که $\beta_{n-1}^{(m)}$ یا $\beta_{n-2}^{(m)}$ درواقع صفر است، به‌دست آمدند، مرتبهٔ ماتریس A_m ممکن است به‌ترتیب یک یا دو سطر تقلیل یابد. در اینجا، روش QR با انتقال را می‌توان برای ماتریس کاهش یافته به‌کار گرفت. انتخاب انتقالها به گونه‌ای طراحی شده است که

همگرایی به صفر $\beta_{n-1}^{(m)}\beta_{n-2}^{(m)}$، نسبت به بقیهٔ عناصر ناقطر ماتریس، سریعتر انجام پذیرد. بدین طریق، روش QR به یک روش همه‌منظورهٔ سریع تبدیل می‌شود که از هر روش دیگر در حال حاضر سریعتر است. برای اثبات همگرایی روش QR با انتقال، ویلکینسن (۱۹۶۸) را ملاحظه کنید. برای بحث بسیار کاملتر روش QR، از جمله انتخاب انتقال، پارلت (۱۹۸۰، فصل ۸) را ببینید.

مثال مثال قبلی (۳.۵.۹) را به‌کار برده از روش اول انتخاب انتقال $c_m = a_n^{(m)}$ استفاده می‌کنیم. بارستها چنین‌اند:

$$A_1 = \begin{bmatrix} ۲ & ۱ & ۰ \\ ۱ & ۳ & ۱ \\ ۰ & ۱ & ۴ \end{bmatrix} \qquad A_2 = \begin{bmatrix} ۱٫۴۰۰۰ & ۰٫۴۸۹۹ & ۰ \\ ۰٫۴۸۹۹ & ۳٫۲۶۶۷ & ۰٫۷۴۵۴ \\ ۰ & ۰٫۷۴۵۴ & ۴٫۳۳۳۳ \end{bmatrix}$$

$$A_3 = \begin{bmatrix} ۱٫۲۹۱۵ & ۰٫۲۰۱۷ & ۰ \\ ۰٫۲۰۱۷ & ۳٫۰۲۰۲ & ۰٫۲۷۲۴ \\ ۰ & ۰٫۲۷۲۴ & ۴٫۶۸۸۴ \end{bmatrix} \qquad A_4 = \begin{bmatrix} ۱٫۲۷۳۷ & ۰٫۰۹۹۳ & ۰ \\ ۰٫۰۹۹۳ & ۲٫۹۹۴۳ & ۰٫۰۰۷۲ \\ ۰ & ۰٫۰۰۷۲ & ۴٫۷۳۲۰ \end{bmatrix}$$

$$A_5 = \begin{bmatrix} ۱٫۲۶۹۴ & ۰٫۰۴۹۸ & ۰ \\ ۰٫۰۴۹۸ & ۲٫۹۹۸۶ & ۰ \\ ۰ & ۰ & ۴٫۷۳۲۱ \end{bmatrix}$$

عنصر $\beta_2^{(m)}$ خیلی سریع به صفر همگرا می‌شود، ولی عنصر $\beta_1^{(m)}$ به‌طور هندسی و فقط با نسبتی در حدود ۰٫۵ به صفر می‌گراید.

باید به روش پیش از روش QR که خیلی زیاد انگیزهٔ پیدایش آن بوده است اشاره‌ای کنیم. در ۱۹۵۸، روتیس‌هاؤزر[1] یک روش LR براساس تجزیهٔ حذف گاوسی ماتریس به یک ماتریس پایین مثلثی ضرب در یک ماتریس بالامثلثی، معرفی کرد. طبق تعریف

$$A_m = L_m R_m \qquad A_{m+1} = R_m L_m = L_m^{-1} A_m L_m$$

که L_m یک ماتریس پایین مثلثی و R_m یک ماتریس بالامثلثی است. این روش معمولاً از روش QR کاراتر است، اگر بتوانیم آن را به‌کار بریم. ولی تبدیلات تشابهی غیرمتعامد می‌توانند یک انحراف در وضع ویژه‌مقدارهای ماتریسها ایجاد نمایند. و معمولاً پیاده‌کردن آن در برنامه‌نویسی خودکار مشکلتر است. یک بحث کامل از آن در ویلکینسن (۱۹۶۵، فصل ۸) داده شده است.

1. H. Rutishauser

۶.۹ محاسبهٔ ویژه‌بردارها و بارست معکوس

تواناترین ابزار در به‌دست‌آوردن ویژه‌بردارهای یک ماتریس روش بارست معکوس، یعنی روشی است که توسط هـ. ویلَنت[1] (۱۹۴۴) ابداع شده است. ابتدا، بارست معکوس را تعریف کرده توضیح می‌دهیم، و سپس در محاسبهٔ ویژه‌بردارها نظر کلیتری خواهیم داد.

برای ساده‌کردن تحلیل، گیریم A ماتریسی باشد که شکل متعارف ژوردان آن قطری است.

$$P^{-1}AP = \text{diag}[\lambda_1, \dots, \lambda_n] \qquad (۱.۶.۹)$$

گیریم ستونهای P با x_1، ...، x_n نشان داده شده‌اند. پس

$$Ax_i = \lambda_i x_i \qquad i = 1, \dots, n \qquad (۲.۶.۹)$$

بی‌آنکه از کلّیت کاسته شود، می‌توان فرض کرد که به‌ازای جمیع مقادیر i، $\| x_i \|_\infty = 1$.

گیریم λ تقریبی برای یک ویژه‌مقدار سادهٔ λ_k ی A باشد. اگر یک مقدار آغازی $z^{(0)}$ داده شده باشد، $\{w^{(m)}\}$ و $\{z^{(m)}\}$ را با روابط زیر تعریف می‌کنیم

$$(A - \lambda I)w^{(m+1)} = z^{(m)}, \qquad z^{(m+1)} = \frac{w^{(m+1)}}{\| w^{(m+1)} \|_\infty} \qquad m \geq 0 \qquad (۳.۶.۹)$$

دراصل این روش، روش توانی است، که در (۲.۲.۹) ـ (۳.۲.۹) به‌جای A، $(A - \lambda I)^{-1}$ آمده است. از دیدگاه مطالب بخش ۴.۸ فصل ۸، ماتریس $A - \lambda I$ بد ـ وضع است. ولی هرگونه اختلال بزرگ حاصله در جواب، بر ویژه‌بردار x_k از ویژه‌مقدار λ $\lambda_k - \lambda$ در $A - \lambda I$ اثر زیاد می‌گذارد، و این همان برداری است که ما دنبال آن هستیم. برای بحث بیشتر در این منشأ ناپایداری در حل دستگاه خطی، مطالبی را که به دنبال (۸.۴.۸) در بخش ۴.۸ آمده است ملاحظه نمایید. برای آنکه روش (۳.۶.۹) نتیجه‌بخش باشد، ما نمی‌خواهیم $A - \lambda I$ تکین باشد. بنابراین λ نباید دقیقاً λ_k باشد، اگرچه می‌تواند کاملاً به آن نزدیک باشد، همان‌گونه که مثال بعدی نشان خواهد داد.

برای یک تحلیل دقیقتر، گیریم $z^{(0)}$ برحسب پایه‌های ویژه‌بردار (۲.۶.۹) بیان شده باشد:

$$z^{(0)} = \sum_{i=1}^{n} \alpha_i x_i \qquad (۴.۶.۹)$$

و فرض می‌کنیم $\alpha_k \neq 0$. در قیاس با فرمول (۴.۲.۹) برای روش توانی، می‌توانیم نشان دهیم

$$z^{(m)} = \frac{\sigma_m (A - \lambda I)^{-m} z^{(0)}}{\| (A - \lambda I)^{-m} z^{(0)} \|_\infty} \qquad | \sigma_m | = 1 \qquad (۵.۶.۹)$$

1. H. Wielandt

با استفاده از (۴.۶.۹)،

$$(A-\lambda I)^{-m}z^{(0)}=\sum_{i=1}^{n}\alpha_i\left[\frac{1}{\lambda_i-\lambda}\right]^m x_i \qquad (۶.۶.۹)$$

گیریم $\epsilon=\lambda-\lambda_k$، و فرض می‌کنیم

$$|\lambda_i-\lambda|\geq c>0 \qquad i=1,\ldots,n \qquad i\neq k \qquad (۷.۶.۹)$$

با استفاده از (۶.۶.۹) و (۵.۶.۹)،

$$z^{(m)}=\sigma_m\frac{x_k+\epsilon^m\sum_{i\neq k}\frac{\alpha_i}{\alpha_k}\left[\frac{1}{\lambda_i-\lambda}\right]^m x_i}{\left\|x_k+\epsilon^m\sum_{i\neq k}\frac{\alpha_i}{\alpha_k}\left[\frac{1}{\lambda_i-\lambda}\right]^m x_i\right\|_\infty} \qquad (۸.۶.۹)$$

که $|\sigma_m|=1$. اگر $|\epsilon|<c$، آنگاه

$$\left\|\epsilon^m\sum_{i\neq k}\frac{\alpha_i}{\alpha_k}\left[\frac{1}{\lambda_i-\lambda}\right]^m x_i\right\|_\infty\leq\left[\frac{\epsilon}{c}\right]^m\sum_{i\neq k}\left|\frac{\alpha_i}{\alpha_k}\right| \qquad (۹.۶.۹)$$

این کمیت وقتی $m\to\infty$، به صفر میل می‌کند. ترکیب آن با (۸.۶.۹) نشان می‌دهد که وقتی $m\to\infty$، $z^{(m)}$ به مضربی از $x_{(k)}$ میل می‌کند. این همگرایی خطی است، با یک نسبت کاهش خطای $|\epsilon/c|$ در هر بارست. در عمل، $|\epsilon|$ کاملاً کوچک است که همگرایی سریع را تضمین می‌کند.

در پیاده‌کردن (۳.۶.۹)، از تجزیهٔ $A-\lambda I$ با استفاده از تجزیهٔ LU در بخش ۱.۸ از فصل ۸، آغاز می‌کنیم. برای ساده‌کردن نمادگذاری، می‌نویسیم

$$A-\lambda I=LU$$

که در آن از محورگیری استفاده نشده است. در عمل، محورگیری به‌کار خواهد رفت. برای هر بارست $z^{(m+1)}$ مسأله را به شکل زیر حل می‌کنیم:

$$Ly^{(m+1)}=z^{(m)} \qquad Uw^{(m+1)}=y^{(m+1)}$$

$$z^{(m+1)}=\frac{w^{(m+1)}}{\|w^{(m+1)}\|_\infty} \qquad (۱۰.۶.۹)$$

چون $A-\lambda I$ تقریباً تکین است، آخرین عنصر قطری U تقریباً صفر خواهد شد. اگر دقیقاً صفر باشد، آن را به یک عدد کوچک تبدیل می‌کنیم یا λ را اندکی تغییر داده L و U را مجدداً به‌دست می‌آوریم.

برای حدس آغازی $z^{(0)}$، ویلکینسن (۱۹۶۳، ص ۱۴۷) انتخاب زیر را توصیه می‌کند:

$$z^{(0)} = Le \qquad e = [1, 1, \ldots, 1]^T$$

پس در (۱۰.۶.۹)

$$y^{(1)} = e \qquad Uw^{(m+1)} = e \tag{۱۱.۶.۹}$$

این انتخاب بدین منظور صورت گرفته است که تضمین کند α_k در (۴.۶.۹)، نه صفر است و نه کوچک. ولی اگر هم کوچک باشد، این روش معمولاً با سرعت همگراست. برای مثال، فرض می‌کنیم، بعضی یا همهٔ مقادیر α_i/α_k در (۹.۶.۹) در حدود 10^4 باشند. و فرض می‌کنیم $|\epsilon/c| = 10^{-5}$، که یک مقدار واقع‌بینانه در خیلی از حالات است. در این صورت کرانِ (۹.۶.۹) چنین می‌شود

$$(10^{-5})^{(m)} \cdot n \cdot 10^4$$

و این کمیت، وقتی m بزرگ می‌شود، خیلی به سرعت کاهش می‌یابد.

مثال ماتریس قبلی (۳.۵.۹)

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 4 \end{bmatrix} \tag{۱۲.۶.۹}$$

را درنظر می‌گیریم. گیریم $\lambda = 1{,}2679 \doteq \lambda_3 = 3 - \sqrt{3}$، که تا ۵ رقم دقیق است. لذا، به علّت خطاهای گردکردن داریم:

$$L = \begin{bmatrix} 1{,}0 & 0 & 0 \\ 1{,}3659 & 1{,}0 & 0 \\ 0 & 2{,}7310 & 1{,}0 \end{bmatrix} \qquad U = \begin{bmatrix} 0{,}7321 & 1{,}0 & 0 \\ 0 & 0{,}3662 & 1{,}0 \\ 0 & 0 & 0{,}0011 \end{bmatrix}$$

با استفاده از $y^{(۱)} = [۱, ۱, ۱]^T$ داریم

$$
\begin{aligned}
w^{(۱)} &= [۳۳۸۵٫۲, -۲۴۷۷٫۳, ۹۰۸٫۲۰]^T \\
z^{(۱)} &= [۱٫۰۰۰۰, -۰٫۷۳۱۸۰, ۰٫۲۶۸۲۸]^T \\
w^{(۲)} &= [۲۰۳۴۵, -۱۴۸۹۴, ۵۴۵۱٫۹]^T \\
z^{(۲)} &= [۱٫۰۰۰, -۰٫۷۳۲۰۷, ۰٫۲۶۷۹۷]^T
\end{aligned}
\qquad (۱۳.۶.۹)
$$

و بردار $z^{(۲)} = z^{(۳)}$. جواب درست چنین است

$$
\begin{aligned}
x_۳ &= [۱, ۱ - \sqrt{۳}, ۲ - \sqrt{۳}]^T \\
&\doteq [۱٫۰۰۰۰, -۰٫۷۳۲۰۵, ۰٫۲۶۷۹۵]^T
\end{aligned}
\qquad (۱۴.۶.۹)
$$

و $z^{(۲)}$ در حدود انباشتگیهای خطای گردکردن، با $x_۳$ برابر است.

ویژه‌بردارهای ماتریسهای سه‌قطری متقارن گیریم A یک ماتریس سه‌قطری متقارن حقیقی از مرتبهٔ n باشد. مانند قبل، فرض می‌کنیم بعضی یا همه ویژه‌مقدارهای آن با دقت محاسبه شده‌اند. روش بارست معکوس روش ارجح در محاسبهٔ ویژه‌بردارهاست، و انجام آن خیلی آسان است. برای λ به‌عنوان یک ویژه‌مقدار تقریبی A، محاسبهٔ تجزیهٔ LU، حتی وقتی که محورگیری به‌کار رفته باشد، از نظر زمان و از نظر حافظه گران نخواهد بود. برای مثال، مطالب بخش ۳.۸ از فصل ۸ را در باب دستگاههای سه‌قطری ملاحظه کنید. مثال عددی قبلی نیز این روش را برای ماتریسهای سه قطری نشان می‌دهد.

بعضی نتایج دربارهٔ خطا در توجیه بیشتر برای استفاده از بارست معکوس را می‌آوریم. فرض می‌کنیم که حساب رایانه‌یی ممیز شناور دودویی با گردکردن و با t رقم در قسمت اعشاری باشد. در ویلکینسن (۶۳، صص ۱۴۳۱۹-۱۴۷)نشان داده شده است که جواب محاسبه‌شدهٔ $\hat{w}$ ی

$$(A - \lambda I)w^{(m+۱)} = z^{(m)}$$

همان جواب دقیق

$$(A - \lambda I + E)\hat{w} = z^{(m)} \qquad (۱۵.۶.۹)$$

است که در آن

$$\| E \|_۲ \leq K\sqrt{n} \cdot ۲^{-t} \qquad (۱۶.۶.۹)$$

که K مقداری ثابت از مرتبهٔ واحد است. اندازهٔ این کران به همان اندازه‌ای است که از خطاهایی از مرتبهٔ خطای گردکردن انتظار می‌رود.

اگر جواب $\hat{w}$ ی (۱۵.۶.۹) کاملاً بزرگ باشد، این کران یک تقریب خوب برای ویژه‌بردار A خواهد بود. برای اثبات این امر، مطلب را با معرفی نماد زیر آغاز می‌کنیم

$$\hat{z} = \frac{\hat{w}}{\| \hat{w} \|_۲}$$

پس

$$(A - \lambda I + E)\hat{z} = \frac{z^{(m)}}{\| \hat{w} \|_۲}$$

$$\eta \equiv (A - \lambda I)\hat{z} = -E\hat{z} + \frac{\hat{z}^m}{\| \hat{w} \|_۲}$$

برای این ماندهٔ η،

$$\begin{aligned} \| \eta \|_۲ &\leq \| E \|_۲ + \frac{\sqrt{n}}{\| \hat{w} \|_۲} \\ &\leq K\sqrt{n} \cdot ۲^{-t} + \frac{\sqrt{n}}{\| \hat{w} \|_۲} \end{aligned} \qquad (۱۷.۶.۹)$$

که اگر $\| \hat{w} \|_۲$ بزرگ باشد، کمیتی کوچک است. برای اثبات اینکه (۱۷.۶.۹) ایجاب می‌کند که $\hat{z}$ به یک ویژه‌بردار A نزدیک باشد، $\{x_i \mid i = ۱, \ldots, n\}$ را یک مجموعه از ویژه‌بردارهای یکا متعامد می‌گیریم. و فرض می‌کنیم

$$\lambda \doteq \lambda_k \qquad \hat{z} = \sum_{i=۱}^{n} \alpha_i x_i \qquad \| \hat{z} \|_۲^۲ = \sum_{۱}^{n} \alpha_i^۲ = ۱$$

که λ_k یک ویژه‌مقدار تنهای A است. همچنین، فرض می‌کنیم

$$\text{برای هر } i \neq k \qquad |\lambda_i - \lambda_k| \geq c > ۰$$

با

$$c \gg | \lambda_k - \lambda |$$

با این مفروضات، اکنون می‌توانیم یک کران برای خطا در $\hat{z}$ پیدا کنیم. ویژه‌بردارها را به‌عنوان

پایه به‌کاربرده η را برحسب آنها بیان می‌کنیم:

$$\eta = A\hat{z} - \lambda\hat{z} = \sum_{1}^{n} \alpha_i(\lambda_i - \lambda)x_i$$

$$\| \eta \|_2^2 = \sum_{1}^{n} \alpha_i^2(\lambda_i - \lambda)^2$$

$$\geq \sum_{i \neq k} \alpha_i^2(\lambda_i - \lambda)^2 \geq c^2 \sum_{i \neq k} \alpha_i^2$$

$$\sum_{i \neq k} \alpha_i^2 \leq \frac{1}{c^2} \| \eta \|_2^2$$

که با استفاده از (۱۷.۶.۹)، کاملاً کوچک است. از $\| \hat{z} \| = 1$ نتیجه می‌شود که $\alpha_k \doteq 1$ و

$$\| \hat{z} - \alpha_k x_k \|_2 = \sqrt{\sum_{i \neq k} \alpha_i^2} \leq \frac{1}{c} \| \eta \|_2 \qquad (۱۸.۶.۹)$$

که نتیجهٔ مطلوب را نشان می‌دهد. برای بحث بیشتر در مورد خطا، ویلکینسن (۱۹۶۳، صص ۱۴۲ تا ۱۴۶) و (۱۹۶۵، صص ۳۲۱ تا ۳۳۰) را ببینید.

روش دیگر محاسبهٔ ویژه‌بردارها، ظاهراً حل مستقیم

$$(A - \lambda I)x = 0$$

خواهد بود، پس از حذف یک معادله و گرفتن یکی از مؤلفه‌های مجهول برابر یک ثابت ناصفر، مثلاً $x_1 = 1$. این روندی است که اغلب در کتابهای جبرخطی دوره‌های کارشناسی به‌کار می‌رود. ولی به‌عنوان یک روش عددی کلی، ممکن است ناموفّق باشد. یک بحث کامل از این مسأله در ویلکینسن (۱۹۶۵، صص ۳۱۵ـ۳۲۱)، همراه با یک مثال عالی، داده شده است. ما فقط همان مثال قبلی را به‌کار می‌بریم که نشان دهیم این جوابها ممکن است به خوبی جوابهای روش بارست معکوس نباشند.

مثال مثال قبلی (۱۲.۶.۹) را با $\lambda = 1.2679$ درنظر می‌گیریم. معادلهٔ $(A - \lambda I)x = 0$ را انتخاب و آخرین معادله را حذف می‌کنیم تا چنین به‌دست آید:

$$0.7321x_1 + x_2 = 0$$

$$x_1 + 1.7321x_2 + x_3 = 0$$

با گرفتن $x_۱ = ۱ر۰$، ویژه‌بردار تقریبی زیر را به‌دست می‌آوریم

$$x = [۱ر۰۰۰۰, -۰ر۷۳۲۱۰, ۰ر۲۶۸۰۷]^T$$

که در مقایسه با جواب درست (۱۴.۶.۹)، این یک جواب کمی ضعیفتر از جواب (۱۳.۶.۹) است که از بارست معکوس حاصل شده است. به‌طور کلی، نتایج به‌دست‌آمده از این راه ممکن است خیلی ضعیف باشند، و هنگام کاربرد این روش باید خیلی مواظب بود.

روش بارست معکوس در اجرا نیاز به توجه بسیار زیادی دارد. در پرداختن به یک ماتریس خاص، هرگونه مشکلات پیش‌بینی نشده‌ای ممکن است رخ دهد. ولی در یک برنامهٔ رایانه‌یی عمومی زیادی است، با ویژه‌مقدارهایی سروکار خواهیم داشت که یا چندگانه‌اند یا بهم نزدیک‌اند، که اگر به آنها توجه نشود، ممکن است مشکلاتی به بار آورند. برای ماتریسهای نامتقارن، که شکل متعارف ژوردان آنها قطری نیست، مشکلاتی هم درگزینش یک پایهٔ صحیح از ویژه‌مقدارها، وجود دارند. بهترین مرجع برای این موضوع، ویلکینسن (۱۹۶۵) است و همچنین می‌توان گلوب و وَن‌لون (۱۹۸۳، صص ۲۳۸ـ۲۴۰) و پارلت (۱۹۸۰، ۶۲ـ۶۹) را ملاحظه کرد. برای چندین برنامهٔ عالی، ویلکینسن و راینش (۱۹۷۱، صص ۴۱۸ـ۴۳۹) و گاربو[1] و همکاران (۱۹۷۷) را ببینید.

۷.۹ حل دستگاههای خطی به روش کمترین مربعات

اکنون حل دستگاههای خطی فرا معیّن به شکل

$$\sum_{j=۱}^{n} a_{ij}x_j = b_i \qquad i = ۱, \ldots, m \qquad (۱.۷.۹)$$

را مطالعه می‌کنیم، که در آنها $m > n$. این دستگاهها درکاربردهای متنوعی ظاهر می‌شوند، که معروفترین آنها برازش توابع با یک مجموعه از داده‌های $\{(t_i, b_i) \mid i = ۱, \ldots, m\}$ است، که دربارهٔ آنها بعداً بیشتر صحبت خواهیم کرد. شاید به‌نظر آید که منطقیترین جا برای بررسی این دستگاهها، فصل هشتم باشد، ولی بعضی ابزارهای حل (۱.۷.۹) شامل تبدیلات متعامد هستند که در این فصل بررسی شده‌اند. حل عددی (۱.۷.۹) هم از لحاظ نظری و هم از جنبهٔ عملی ممکن است کاملاً پیچیده باشد ولی ما فقط بعضی از قسمتهای مهم آن را بیان می‌کنیم.

1. Garbow

یک دستگاه فرامعیّن (۱.۷.۹) در حالت کلی، جوابی ندارد. بدین دلیل یک بردار $x = (x_1, \ldots, x_n)$ را جستجو می‌کنیم که به تعبیری جواب تقریبی (۱.۷.۹) باشد. نمادهای

$$A = [a_{ij}] \qquad x = [x_1, \ldots, x_n]^T \qquad b = [b_1, \ldots, b_m]^T$$

را که در آن A یک ماتریس $m \times n$ است وارد می‌کنیم. در این صورت معادلهٔ (۱.۷.۹) را می‌توانیم به‌صورت زیر بنویسیم

$$Ax = b \tag{۲.۷.۹}$$

برای سادگی، A و bرا حقیقی فرض می‌کنیم. در بین راههای ممکن پیداکردن یک جواب تقریبی، می‌توانیم بردار x را جویا شویم که

$$\| Ax - b \|_p \tag{۳.۷.۹}$$

را مینیمم سازد، که در آن $1 \leq p \leq \infty$. در این بخش، فقط حالت کلاسیک $p = 2$ را بررسی می‌کنیم، گرچه در سالهای اخیر، کارهای زیادی برای حالتهای $p = 1$ و $p = \infty$ شده است.

جواب x^* ی

$$\underset{x \in \mathbf{R}^n}{\text{Minimize}} \| Ax - b \|_2 \tag{۴.۷.۹}$$

را **جواب کمترین مربعات** دستگاه خطی $Ax = b$ می‌نامند. دلایل متعددی برای حل دستگاه $Ax = b$ بدین طریق وجود دارد. اولاً، تئوری و روشهای عملی مینیمم‌سازی $\| Ax - b \|_2$ آسانتر است، تا حدی به این علت که $\| Ax - b \|_2$ یک تابع پیوسته مشتقپذیر برحسب x_1، ...، x_n است و ثانیاً، مسائل برازش منحنی که به دستگاههای (۱.۷.۹) می‌انجامند، اغلب یک چارچوب آماری دارند که به (۴.۷.۹) منجر می‌شوند، که بر مینیمم‌سازی $\| Ax - b \|_p$، $p \neq 2$، رجحان دارند.

برای درک بهتر ماهیّت حل (۴.۷.۹)، ساختمان نظری زیر را عرضه می‌کنیم. از این ساختمان نیز می‌توان به‌عنوان یک روش عددی عملی نیز استفاده کرد، اگرچه معمولاً روشهای کاراتر دیگری

وجود دارند. نکتهٔ مهم در این نظریه، تجزیهٔ تکین ـ مقدار زیر است،

$$V^T AU = F = \begin{bmatrix} \mu_1 & & & & & \cdots & & 0 \\ & \ddots & & & & & & \\ & & \mu_r & & & & & \\ & & & 0 & & & & \\ & & & & \ddots & & & \\ \vdots & & & & & & 0 & \\ & & & & & & \vdots & \\ 0 & & & & & \cdots & 0 & \end{bmatrix} \qquad (۵.۷.۹)$$

ماتریسهای U و V متعامدند، و مقدارهای تکین μ_i در روابط زیر صدق می‌کنند

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_r > 0$$

برای آگاهی بیشتر، قضیهٔ ۵.۷ در فصل ۷ را ببینید، بعداً در این بخش روش ساختن تجزیهٔ تکین ـ مقدار A را ذکر می‌کنیم.

قضیهٔ ۷.۹ گیریم A ماتریس حقیقی $m \times n$ است، $m \geq n$. تعریف می‌کنیم $z = U^T x$، $c = V^T b$. در این صورت $x^* = Uz^*$، جواب (۴.۷.۹) با رابطهٔ

$$z_i^* = \frac{c_i}{\mu_i} \qquad i = 1, \ldots, r \qquad (۶.۷.۹)$$

داده می‌شود که در آن z_{r+1}، ...، z_n دلخواه‌اند. وقتی $r = n$، x^* یکتاست. وقتی $r < n$، جواب نرم اقلیدسی مینیمال (۴.۷.۹) با قراردادن

$$z_i^* = 0 \qquad i = r+1, \ldots, n \qquad (۷.۷.۹)$$

به‌دست می‌آید. [این نیز جواب کمترین مربعات (۴.۷.۹) خوانده می‌شود، ولو اینکه این یک مینیمم‌ساز یکتا برای $\| Ax - b \|_2$ نیست.] مینیمم (۴.۷.۹) با رابطهٔ زیر داده می‌شود

$$\| Ax^* - b \|_2 = \left[\sum_{j=r+1}^{m} c_j^2 \right]^{1/2} \qquad (۸.۷.۹)$$

برهان مسألهٔ ۱۳ (الف) از فصل ۷ را به یاد می‌آوریم. برای هر $x \in \mathbf{R}^n$ و هر ماتریس متعامد P،

$$\| Px \|_2 = \| x \|_2$$

با بهره‌گیری از این رابطه در $\| Ax - b \|_۲$ و استفاده از (۵.۷.۹)،

$$\begin{aligned} \| Ax - b \|_۲ &= \| V^T Ax - V^T b \|_۲ = \| V^T AUU^T x - c \|_۲ \\ &= \| Fz - c \|_۲ \\ &= \left[\sum_{j=۱}^{r} (\mu_j z_j - c_j)^۲ + \sum_{j=r+۱}^{m} c_j^۲ \right]^{۱/۲} \end{aligned} \tag{۹.۷.۹}$$

این رابطه (۶.۷.۹) و (۸.۷.۹)، بلافاصله نتیجه می‌شوند. برای (۷.۷.۹)، از

$$\| x^* \|_۲ = \| z \|_۲ = \left[\sum_{j=۱}^{r} (z_j^*)^۲ + \sum_{j=r+۱}^{n} z_j^۲ \right]^{۱/۲}$$

استفاده می‌کنیم که در آن به موجب (۹.۷.۹)، $z_{r+۱}$، ...، z_n اختیاری‌اند. انتخاب (۷.۷.۹) به یک مقدار مینیمم یکتا برای $\| x^* \|_۲$ می‌انجامد. ■

ماتریس $n \times m$

$$F^+ = \begin{bmatrix} \mu_۱^{-۱} & ۰ & \cdots & ۰ \\ & \ddots & & \\ \vdots & & \mu_r^{-۱} & \vdots \\ & & ۰ & \\ & & & \ddots \\ ۰ & \cdots & ۰ \cdots ۰ & \end{bmatrix} \tag{۱۰.۷.۹}$$

و

$$A^+ = UF^+V^T \tag{۱۱.۷.۹}$$

را تعریف می‌کنیم. با توجه به (۶.۷.۹) ـ (۸.۷.۹)،

$$\begin{aligned} x^* &= Uz^* = UF^+c = UF^+V^Tb \\ x^* &= A^+b \end{aligned} \tag{۱۲.۷.۹}$$

ماتریس A^+ **معکوس تعمیم‌یافتهٔ** A خوانده می‌شود و جواب کمترین مربعات $Ax = b$ را به‌دست می‌دهد. فرمول (۱۲.۷.۹) نشان می‌دهد که x^* خطی ـ وابسته به b است. این نحوهٔ نمایش x^*

ابزار مهمی در مطالعهٔ حل عددی $Ax = b$ است. بعضی از ویژگیهای دیگر A^+ در مسائل ۲۷ و ۲۸ مطرح شده‌اند.

برای ساده‌کردن مطالعهٔ بقیّهٔ روشها در پیداکردن x^* و تحلیل پایداری آن، A را به حالت ردهٔ کامل $r = n$ محدود می‌سازیم. این حالت مهمترین حالت درکاربردهاست. یادآوری می‌کنیم:برای تکین مقدارهای A،

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n > 0 \qquad (۱۳.۷.۹)$$

مفهوم نرم ماتریسی را، که در بخش ۳.۷ برای ماتریسهای مربعی داده شده بود، می‌توان برای A تعمیم داد. تعریف می‌کنیم

$$\| A \| = \underset{\substack{x \in \mathbf{R}^n \\ x \neq 0}}{\text{Supremum}} \frac{\| Ax \|_2}{\| x \|_2} \qquad (۱۴.۷.۹)$$

با استفاده از تجزیهٔ تکین مقدار ماتریس A، می‌توان نشان داد که،

$$\| A \| = \sqrt{r_\sigma(A^T A)} = \mu_1 \qquad (۱۵.۷.۹)$$

در قیاس با تحلیل خطا در بخش ۴.۸، ضریب وضعیت برای $Ax = b$ را با رابطهٔ زیر تعریف می‌کنیم

$$\text{cond}(A)_2 = \| A \| \| A^+ \| = \frac{\mu_1}{\mu_n} \qquad (۱۶.۷.۹)$$

با استفاده از این نماد، یک قضیّهٔ پایداری از گلوب و وَن‌لون (۱۹۸۳، ص ۱۴۱) را در اینجا می‌آوریم. این قضیّه مشابه قضیهٔ ۴.۸، برای تحلیل اختلال در دستگاههای خطی ناتکین است.

گیریم $b + \delta b$ و $A + \delta A$ به ترتیب شکلهای اختلال‌یافتهٔ b و A باشند. تعریف می‌کنیم

$$x^* = A^+ b \qquad \hat{x}^* = (A + \delta A)^+ (b + \delta b)$$
$$r = b - Ax^* \qquad \hat{r} = (b + \delta b) - (A + \delta A)\hat{x}^* \qquad (۱۷.۷.۹)$$

فرض می‌کنیم

$$\epsilon \equiv \text{Max}\left[\frac{\| \delta A \|}{\| A \|}, \frac{\| \delta b \|_2}{\| b \|_2}\right] < \frac{1}{\text{cond}(A)_2} \qquad (۱۸.۷.۹)$$

و

$$\sin(\theta) \equiv \frac{\| r \|_2}{\| b \|_2} < 1 \qquad (۱۹.۷.۹)$$

که به‌طور ضمنی $0 \leq \theta < \pi/2$ را تعریف می‌کند. در این صورت

$$\frac{\| \hat{x}^* - x^* \|_2}{\| x^* \|_2} \leq \epsilon \left[\frac{2\text{cond}(A)_2}{\cos\theta} + \tan\theta [\text{cond}(A)_2]^2 \right] + O(\epsilon^2) \qquad (20.7.9)$$

$$\frac{\| \hat{r} - r \|_2}{\| b \|_2} \leq \epsilon [1 + 2\text{cond}(A)_2] \text{Min}\{1, m-n\} + O(\epsilon^2) \qquad (21.7.9)$$

برای حالت $m = n$، که n مرتبهٔ (A) است، ماندهٔ r صفر می‌شود و بنابراین (۲.۷.۸) به قضیهٔ قبلی ۴.۸ بَدَل می‌شود.

نتایج قبلی بیان می‌کنند که تغییر در r ممکن است کاملاً کوچک و تغییر در x^* کاملاً بزرگ باشد. توجه کنید که کران در (۲۰.۷.۹)، در مقایسه با وابستگی خطی $\text{cond}(A)$ در حالت ناتکین $m = n$، به مربع $\text{cond}(A)_2$ بستگی دارد [(۱۸.۴.۸) را ببینید]. اگر ستونهای A تقریباً وابسته باشند، آنگاه $\text{cond}(A)_2$ ممکن است خیلی بزرگ باشد، که کران بزرگتری را در (۲۰.۷.۹) در مقایسه با (۲۱.۷.۹)، ایجاد خواهد کرد [مسألهٔ ۳۴ (الف) را ببینید]. قابل‌قبول بودن یا نبودن، به مسأله بستگی دارد، به اینکه آیا مقادیر کوچک r را می‌خواهند یا مقادیر دقیق x^* را.

مسأله برازش منحنی به روش کمترین مربعات منشأ بسیاری از دستگاههای خطی فرامعیّن مسألهٔ برازاندن داده‌ها با یک تابع از یک خانواده از توابع معیّن است. گیریم $\{(t_i, b_i) \mid i = 1, \ldots, m\}$ یک مجموعه از داده‌ها باشد، احتمالاً معرّف یک تابع $b = g(t)$. گیریم $\varphi_1(t)$، ...، $\varphi_n(t)$ توابع داده شده باشند، و گیریم $\mathscr{F}$ خانوادهٔ تمام ترکیبات خطی φ_1، ...، φ_n باشد:

$$\mathscr{F} = \left\{ \sum_{j=1}^{n} x_j \varphi_j(t) \middle| x_j \in \mathbf{R} \right\} \qquad (22.7.9)$$

می‌خواهیم عنصری از $\mathscr{F}$ بیابیم به گونه‌ای که تقریباً در

$$\sum_{j=1}^{n} x_j \varphi_j(t_i) = b_i \qquad i = 1, \ldots, m \qquad (23.7.9)$$

صدق کند. این همان دستگاه (۱.۷.۹)، با $a_{ij} = \varphi_j(t_i)$ است.

به دلایل مدل‌سازی آماری، در پی آن هستیم که

$$E(x) = \left[\frac{1}{m} \sum_{i=1}^{m} \left[b_i - \sum_{j=1}^{n} x_j \varphi_j(t_i) \right]^2 \right]^{1/2} \qquad (24.7.9)$$

را مینیمم سازیم و بنابراین می‌خواهیم برازش داده‌ها را به مفهوم کمترین مربعات بیان کنیم. کمیتِ $E(x^*)$، که برای آن $E(x)$ مینیمم می‌شود، میانگین مربع خطاها، در تقریب داده‌ها با تابع

$$g^*(t) = \sum_{j=1}^{n} x_j^* \varphi_j(t) \qquad (۲۵.۷.۹)$$

نامیده می‌شود. با استفاده از نماد قبلی

$$E(x) = \frac{1}{\sqrt{m}} \| b - Ax \|_2$$

و مینیمم‌کردن $E(x)$ هم‌ارز با پیداکردن جواب کمترین مربعات (۲۳.۷.۹) است.

از گرفتن مشتقات جزیی (۲۴.۷.۹) نسبت به هر x_i، و برابر صفر قراردادن آنها، دستگاه معادلات زیر را به‌دست می‌آوریم

$$A^T Ax = A^T b \qquad (۲۶.۷.۹)$$

این دستگاه شرط لازم برای هر مینیمم‌ساز $E(x)$ است، و می‌توان نشان داد که شرط کافی هم هست. دستگاه (۲۶.۷.۹) معادلهٔ نرمال برای مسألهٔ کمترین مربعات خوانده می‌شود. اگر A دارای رتبهٔ n باشد، آنگاه $A^T A$ یک ماتریس $n \times n$ و ناتکین است و (۲۶.۷.۹) دارای یک جواب یکتاست.

برای اثبات هم‌ارزی (۲۶.۷.۹) با جواب قبلی مسألهٔ کمترین مربعات، تجزیهٔ تکین ـ مقدار A را به‌کار می‌بریم تا (۲۶.۷.۹) را به یک شکل ساده‌تر برگردانیم. از گذاردن $A = VFU^T$ در (۲۶.۷.۹)،

$$UF^T FU^T x = UF^T V^T b$$

از ضرب در U و استفاده از نماد قبلی $z = U^T x$، $c = V^T b$، داریم

$$F^T Fz = F^T c$$

این رابطه یک هم‌ارزی ریاضی کامل بین معادلهٔ نرمال و مینیمم‌سازی قبلی $\| Ax - b \|_2$ در قضیهٔ ۷.۹ برقرار می‌سازد.

فرض می‌کنیم رتبهٔ A برابر n باشد، جواب x^* را می‌توان از حل معادلهٔ نرمال به‌دست آورد. چون $A^T A$ متقارن و معین مثبت است، تجزیهٔ چولسکی را برای پیداکردن جواب می‌توان مورد استفاده قرار داد [(۸.۳.۸) ـ (۱۷.۳.۸) را ببینید]. اثر خطاهای گردکردن روی x^* هم با واحد

خطای گردکردن در رایانه و هم با ضریب وضعیت A^TA، متناسب است. از تجزیهٔ تکین ـ مقدار مربوط به A به سادگی می‌توان دید که

$$\text{cond}(A^TA)_۲ = \frac{\mu_۱^۲}{\mu_n^۲} = [\text{cond}(A)_۲]^۲ \qquad (۲۷.۷.۹)$$

بنابراین حساسیت x^* نسبت به خطاها متناسب است با $[\text{cond}(A)_۲]^۲$، که با کران خطای اختلال (۲۰.۷.۹) سازگار است.

نتیجهٔ (۲۷.۷.۹) معمولاً به‌عنوان دلیل اصلی احتراز از استفاده از معادلهٔ نرمال در حل مسألهٔ کمترین مربعات ذکر شده است. این گوشزد خوبی است، ولی دلایل خیلی ظریفتر از اینها هستند. از (۲۰.۷.۹) دیده می‌شود که اگر $\| r \|_۲$ نزدیک صفر باشد، آنگاه $\sin\theta \doteq ۰$، و کران متناسب با $\text{cond}(A)_۲$ می‌شود. در مقابل، کران خطا در روش چولسکی کمیت $[\text{cond}(A)_۲]^۲$ را نشان می‌دهد که اگر $\text{cond(A)}_۲$ بزرگ باشد، این کمیت بزرگتر می‌شود. یک دلیل دیگر وقتی به‌وجود می‌آید که ستونهای A تقریباً وابسته باشند. در این صورت استفاده از حساب رایانه‌یی محدود ممکن است به یک معادلهٔ نرمال تقریبی منجر شود که اطلاعات ضروری موجود در A را ندارد. در چنین حالتی، A^TA تقریباً تکین است، و جواب معادلهٔ نرمال دقت بسیار کمتری در x^* را، در مقایسه با سایر روشها، که مستقیماً با $Ax = b$ کار می‌کنند، به‌دست خواهد داد. برای یک بحث دقیقتر از این موضوع لاوسن و هنسن[1] (۱۹۷۴، صص ۱۲۶ـ۱۲۹) را ببینید.

مثال داده‌های جدول ۳.۹ و نمودار مربوط به آن را در شکل (۲.۹) درنظر می‌گیریم. یک چندجمله‌یی درجهٔ ۳ برای برازش این داده‌ها به‌کار می‌بریم، و به مینیمم‌کردن عبارت زیر می‌رسیم

$$E(x) = \left[\frac{۱}{m}\sum_{i=۱}^{m}\left[b_i - \sum_{j=۱}^{۴} x_j t_i^{j-۱}\right]^۲\right]^{۱/۲}$$

از اینجا، دستگاه خطی فرامعیّن

$$\sum_{j=۱}^{۴} x_j t_i^{j-۱} = b_i \qquad i = ۱, \ldots, m \qquad (۲۸.۷.۹)$$

نتیجه می‌شود.

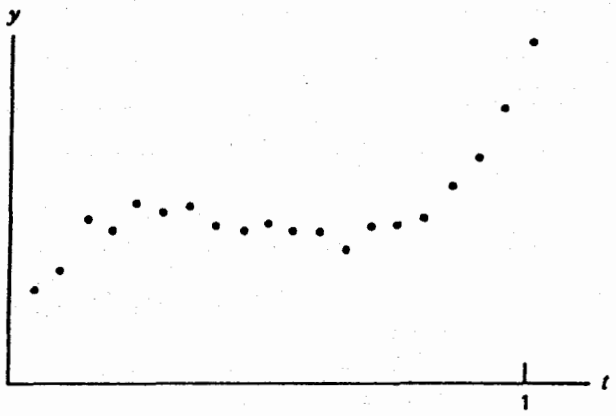
و معادلات نرمال عبارت‌اند از:

$$\sum_{j=۱}^{n} x_j \left[\sum_{i=۱}^{m} t_i^{j+k-۲}\right] = \sum_{i=۱}^{m} t_i^{k-۱} b_i \qquad k = ۱, ۲, ۳, ۴$$

1. Lawson and Hanson

جدول ۳.۹ داده‌ها برای یک برازش کمترین مربعات درجهٔ ۳

x_i	y_i	x_i	y_i
۰٫۰۰	۰٫۴۸۶	۰٫۵۵	۱٫۱۰۲
۰٫۰۵	۰٫۸۶۶	۰٫۶۰	۱٫۰۹۹
۰٫۱۰	۰٫۹۴۴	۰٫۶۵	۱٫۰۱۷
۰٫۱۵	۱٫۱۴۴	۰٫۷۰	۱٫۱۱۱
۰٫۲۰	۱٫۱۰۳	۰٫۷۵	۱٫۱۱۷
۰٫۲۵	۱٫۲۰۲	۰٫۸۰	۱٫۱۵۲
۰٫۳۰	۱٫۱۶۶	۰٫۸۵	۱٫۲۶۵
۰٫۳۵	۱٫۱۹۱	۰٫۹۰	۱٫۳۸۰
۰٫۴۰	۱٫۱۲۴	۰٫۹۵	۱٫۵۷۵
۰٫۴۵	۱٫۰۹۵	۱٫۰۰	۱٫۸۵۷
۰٫۵۰	۱٫۱۲۲		



شکل ۲.۹ نمودار داده‌های جدول ۳.۹.

اگر آن را به شکل (۲۶.۷.۹) بنویسیم، داریم

$$A^TA = \begin{bmatrix} ۲۱ & ۱۰٫۵ & ۷٫۱۷۵ & ۵٫۵۱۲۵ \\ ۱۰٫۵ & ۷٫۱۷۵ & ۵٫۵۱۲۵ & ۴٫۵۱۶۶۶ \\ ۷٫۱۷۵ & ۵٫۵۱۲۵ & ۴٫۵۱۶۶۶ & ۳٫۸۵۴۱۶ \\ ۵٫۵۱۲۵ & ۴٫۵۱۶۶۶ & ۳٫۸۵۴۱۶ & ۳٫۳۸۲۱۲ \end{bmatrix}$$

$$A^Tb = [۲۴٫۱۱۸۰, ۱۳٫۲۳۴۵, ۹٫۴۶۸۳۶۵, ۷٫۵۵۹۴۴۰۵]^T \qquad (۲۹.۷.۹)$$

جواب چنین است

$$x^* = [\circ/۵۷۴۷, ۴/۷۲۵۹, -۱۱/۱۲۸۲, ۷/۶۶۸۷]^T \qquad (۳۰.۷.۹)$$

این جواب نسبت به تغییرات در b خیلی حساس است. به این حساسیت می‌توان از ضریب وضعیت

$$\text{cond}(A^TA) = ۱۲۱۰۵ \qquad (۳۱.۷.۹)$$

پی برد. به‌عنوان یک گواه دیگر، بردار سمت راست A^Tb را با بردار

$$[\circ/۱, -\circ/۱, \circ/۱, -\circ/۱]^T$$

جمع کرده اختلال ایجاد می‌کنیم. این اختلال با اندازهٔ خطاهای موجود در مقادیر داده‌های b_i، سازگار است. با این مقدار جدیدِ سمت راست، معادلهٔ نرمال جواب اختلال‌یافتهٔ زیر را دارد

$$\hat{x}^* = [\circ/۷۴۰۸, ۲/۶۸۲۵, -۶/۱۵۳۸, ۴/۴۵۵۰]^T$$

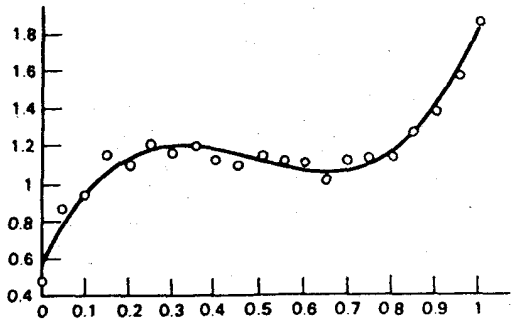
که تفاوت زیادی با x^* دارد. نمودار برازش کمترین مربعات

$$g^*(t) = x_۱^* + x_۲^*t + x_۳^*t^۲ + x_۴^*t^۳$$

همراه با داده‌ها در شکل ۳.۹ نشان داده شده است. خطای ریشهٔ دوم میانگین نمودار چنین است

$$E(x^*) = \circ/۰۴۲۱$$

ماتریس A در این مثال، دارای ستونهایی است که تقریباً خطی ـ وابسته‌اند، و A^TA ضریب وضعیت بزرگی دارد. برای بهبود آن، می‌توانیم یک مجموعهٔ توابع پایهٔ $\{\varphi_i(t)\}$ ی بهتری از



شکل ۳.۹ برازش کمترین مربعات $g^*(t)$

چندجمله‌ییهای درجهٔ نابزرگتر از ۳، برای خانوادهٔ $\mathscr{F}$ انتخاب کنیم. از بررسی ضرایب A^TA، داریم

$$[A^TA]_{jk} = \sum_{i=1}^{m} \varphi_j(t_i)\varphi_k(t_i) \qquad 1 \le j,k \le n \qquad (۳۲.۷.۹)$$

اگر نقاط $\{t_i\}$ در سرتاسر بازهٔ $[a,b]$ خوب توزیع شده باشند، آنگاه مجموع، وقتی در $(b-a)/m$ ضرب شود، تقریبی برای

$$\int_a^b \varphi_k(t)\varphi_j(t)dt$$

خواهد بود. برای به‌دست آوردن یک ماتریس A^TA که ضریب وضعیت کوچکتری داشته باشد، توابع $\varphi_j(t)$ را یکا متعامد انتخاب می‌کنیم. در این صورت A^TA ماتریس تقریباً یکا متعامد است و A ستونهای تقریباً یکا متعامدی خواهد داشت، که به ضریب وضعیت نزدیک به واحد منتهی خواهد شد. درواقع آنچه از همه مهمتر است متعامدبودن خانوادهٔ $\{\varphi_j(t)\}$ است، زیرا در آن صورت ماتریس A^TA تقریباً قطری است، که یک ماتریس خوش ـ وضع است.

مثال مثال قبلی را تکرار می‌کنیم، و از چندجمله‌ییهای لژاندارکه بر بازهٔ $[۰, ۱]$ یکا متعامدند استفاده می‌کنیم. نخستین چهار چندجمله‌یی لژاندارکه در بازهٔ $[۰, ۱]$ یکا متعامدند عبارت‌اند از:

$$\varphi_0(t) = 1 \quad \varphi_1(t) = \sqrt{3}s \quad \varphi_2(t) = \frac{\sqrt{5}}{2}(3s^2 - 1)$$

$$\varphi_3(t) = \frac{\sqrt{7}}{2}(5s^3 - 3s) \qquad (۳۳.۷.۹)$$

که در آنها $s = 2t - 1$، $0 \le t \le 1$. برای معادلهٔ نرمال (۲۶.۷.۹)،

$$A^TA = \begin{bmatrix} ۲۱٫۰۰۰۰ & ۰ & ۲٫۳۴۷۹ & ۰ \\ ۰ & ۲۳٫۱۰۰۰ & ۰ & ۵٫۱۱۶۴ \\ ۲٫۳۴۷۹ & ۰ & ۲۵٫۴۹۹۳ & ۰ \\ ۰ & ۵٫۱۱۶۴ & ۰ & ۲۸٫۳۸۸۹ \end{bmatrix}$$

$$A^Tb = [۲۴٫۱۱۸۰, ۴٫۰۷۲۱, ۳٫۴۰۱۵, ۴٫۸۵۱۹]^T$$

$$x^* = [۱٫۱۴۵۴, ۰٫۱۴۴۲, ۰٫۰۲۷۹, ۰٫۱۴۴۹]^T \qquad (۳۴.۷.۹)$$

ضریب وضعیت A^TA اکنون برابر است با

$$\text{cond}(A^TA) = ۱٫۵۸ \qquad (۳۵.۷.۹)$$

که بسیار کوچکتر از مقدار قبلی در (۳۱.۷.۹) است.

حل با روش QR تجزیه به QR در بخش ۳.۹ و به دنبال آن (۱۱.۳.۹) را به یاد می‌آوریم. مانند آنجا، ماتریسهای هاؤسهولدر از مرتبهٔ $m \times m$ را درنظر می‌گیریم

$$P_j = I - ۲w^{(j)}w^{(j)T} \qquad j = ۱, \ldots, n$$

تا عناصر زیر قطر در A را به صفر بدل کنیم. ماتریسهای متعامد P_j را متوالیاً به‌کار می‌بریم تا عناصر زیرقطر در ستونهای اول تا n به صفر بدل شوند. بردار $w^{(j)}$ در موضع j ام تا m ام، عناصر ناصفر دارد. این فرایند به ماتریس زیر می‌انجامد

$$R = P_n \ldots P_۱ A = Q^T A \qquad (۳۶.۷.۹)$$

اگر این عملیات را در سمت راست دستگاه $Ax = b$ هم اعمال کنیم، دستگاه هم‌ارز زیر را به‌دست می‌آوریم

$$Rx = Q^T b \qquad (۳۷.۷.۹)$$

ماتریس R به شکل زیر است

$$R = \begin{bmatrix} R_۱ \\ ۰ \end{bmatrix} \qquad (۳۸.۷.۹)$$

که $R_۱$ یک ماتریس مربعی بالامثلثی از مرتبه $n \times n$ است. ماتریس $R_۱$ ناتکین است، زیرا A و $R = Q^T A$ یک رتبه دارند که n است. برطبق (۳۸.۷.۹) می‌نویسیم

$$Q^T b = \begin{bmatrix} g_۱ \\ g_۲ \end{bmatrix} \qquad g_۱ \in \mathbf{R}^n \; g_۲ \in \mathbf{R}^{m-n}$$

پس

$$\| Ax - b \|_۲ = \| Q^T Ax - Q^T b \|_۲ = \| Rx - Q^T b \|_۲$$
$$= \left[\| R_۱ x - g_۱ \|_۲^۲ + \| g_۲ \|_۲^۲ \right]^{۱/۲}$$

جواب کمترین مربعات $Ax = b$ از حل دستگاه بالامثلثی ناتکین زیر به‌دست می‌آید

$$R_۱ x = g_۱ \qquad (۳۹.۷.۹)$$

پس مینیمم چنین است

$$\|Ax^* - b\|_۲ = \|g_۲\|_۲ \qquad (۴۰.۷.۹)$$

روش QR برای محاسبهٔ x^* از لحاظ عملیات، کمی از روش چولسکی گرانتر است. تعداد عملیات (ضرب و جمع) روش چولسکی، ازجمله تشکیل A^TA، در حدود

$$\frac{۱}{۲}mn^۲ + \frac{n^۳}{۶}$$

است و تعداد عملیات روش هاؤسهولدر QR در حدود

$$mn^۲ - \frac{n^۳}{۳}$$

است. معذالک معمولاً روش QR برای حل کمترین مربعات توصیه می‌شود. این روش مستقیماً روی A انجام می‌شود و به همین علت و به علّت استفاده از تبدیلات متعامد، اثر خطاهای گردکردن با این روش از نتیجهٔ آن در استفاده از تجزیه به عوامل چولسکی در حل معادلهٔ نرمال بهتر است. برای بحث عمیقتر گلوب و ون‌لون (۱۹۸۳، صص ۱۴۷ـ۱۴۹) و لاوسن و هنسن (۱۹۷۴، فصل ۱۶) را ملاحظه کنید.

مثال مثال قبلی دستگاه خطی (۲۸.۷.۹)

$$A_{ij} = [t_i^{j-۱}] \quad ۱ \le i \le ۲۱, \quad ۱ \le j \le ۴$$

را همراه با داده‌های جدول ۳.۹ درنظر می‌گیریم. در اینجا $\text{cond}(A) = ۱۱۰\text{ر}۰۱$،

$$R_۱ = \begin{bmatrix} -۴\text{ر}۵۸۲۶ & -۲\text{ر}۲۹۱۳ & -۱\text{ر}۵۶۵۷ & -۱\text{ر}۲۰۲۹ \\ ۰ & ۱\text{ر}۳۸۷۴ & ۱\text{ر}۳۸۷۴ & ۱\text{ر}۲۶۸۸ \\ ۰ & ۰ & -۰\text{ر}۳۷۴۴ & -۰\text{ر}۵۶۱۷ \\ ۰ & ۰ & ۰ & -۰\text{ر}۰۹۸۷ \end{bmatrix}$$

$$g_۱ = [-۵\text{ر}۲۶۳۰, ۰\text{ر}۸۴۷۲, -۰\text{ر}۱۴۰۳, -۰\text{ر}۷۵۶۶]^T$$

این جواب x^* با جواب x^* در (۳۰.۷.۹) یکی است، همان‌طورکه خطای ریشهٔ دوّم میانگین یکی است.

تجزیهٔ تکین ـ مقدار تجزیهٔ تکین ـ مقدار ابزار خیلی باارزشی برای تحلیل و حلّ مسائل کمترین مربعات و مسائل دیگر در جبر خطی است. برای مسائل کمترین مربعاتِ در حالتی که رتبه کامل نیست، روش QR که هم‌اکنون توضیح داده شد احتمالاً به ماتریس مثلثی $R_۱$ می‌انجامد که ناتکین است، ولی دارای عناصر قطری بسیار کوچک است. در این صورت تجزیهٔ تکین ـ مقدار A ممکن است در روشنترکردن ساختار A بسیار مفید باشد. اگر بعضی از تکین ـ مقدارهای μ_i

نزدیک به صفر باشند، آنگاه تعیین اثر صفرگذاردن آنها برای یافتن جواب x^* ممکن است ساده‌تر از تعیین نتیجه در سایر روشها برای حل‌کردن x^* باشد. از این‌رو توجیه فراوانی برای پیداکردن روشهای کارا برای محاسبهٔ تجزیهٔ تکین ـ مقدار A وجود دارد.

یکی از بهترین راههای شناخته‌شده برای محاسبهٔ تجزیهٔ تکین ـ مقدار A، راه منسوب به گلوب[1] و راینش[2] و کاهان[3] است و بحث کامل آن در گلوب و ون‌لون (۱۹۸۳، بخش ۵.۶) داده شده است. ما فقط نشان می‌دهیم که چگونه تجزیهٔ تکین ـ مقدار در (۵.۷.۹) را می‌توان از حل مسالهٔ ویژه‌مقدار یک ماتریس متقارن همراه با تجزیهٔ QR به‌دست آورد.

اگر A حقیقی و $m \times n$ با $m \geq n$ باشد، می‌دانیم که A^TA یک ماتریس حقیقی $n \times n$ است. به‌علاوه به‌سادگی می‌توان نشان داد که A^TA متقارن و نیمه معین‌مثبت است [به‌ازای جمیع مقادیر x، $x^TAx \geq 0$]. با استفاده از یک برنامه برای حل مسالهٔ ویژه‌مقدار متقارن، یک ماتریس قطری D و یک ماتریس متعامد U پیدا می‌کنیم که

$$U^T(A^TA)U = D \qquad (۴۱.۷.۹)$$

گیریم $D = \text{diag}[\lambda_1, \ldots, \lambda_n]$، با ویژه‌مقدارهایی که به‌ترتیب نزولی مرتب شده‌اند. اگر یکی از λ_i ها یک عدد کوچک منفی باشد آن را صفر می‌گیریم، زیرا همهٔ ویژه‌مقدارهای A^TA باید نامنفی باشند مگر آنکه اختلالاتی بر اثر خطاهای گردکردن حاصل شده باشد.

در (۴۱.۷.۹)، B از مرتبهٔ $m \times n$ را چنین تعریف می‌کنیم $B = AU$. در این صورت از (۴۱.۷.۹) نتیجه می‌شود

$$B^TB = D$$

پس ستونهای B متعامدند. به‌علاوه اگر یکی از λ_i ها صفر باشد، ستون متناظر با آن در B باید متحد با صفر باشد، زیرا نرم آن صفر است. با به‌کاربردن روش QR، یک ماتریس متعامد V پیدا می‌کنیم که برای آن

$$V^TB = R \qquad (۴۲.۷.۹)$$

زیرقطر اصلی، همهٔ ستونها صفر باشند. ماتریس R در رابطهٔ زیر صدق می‌کند

$$R^TR = B^TV^TVB = B^TB = D$$

1. G. Golub 2. C. Reinsch 3. W. Kahan

باز هم، ستونهای R باید متعامد باشند، و اگر یکی از λ_i ها صفر باشد، ستون متناظر آن در R نیز باید صفر باشد. چون R یک بالامثلثی است، می‌توانیم از تعامد استفاده کرده نشان دهیم که ستونهای R در تمام مواضع بالای قطر دارای عناصر صفرند. بنابراین R به شکل ماتریس F در (۵.۷.۹) است، به جز آنکه بعضی عناصر قطر R ممکن است منفی باشند. در این حالت، علامت این عنصر قطر و ستون متناظر آن در U را عوض می‌کنیم. در این صورت داریم $R = F$، $\mu_i = \sqrt{\lambda_i}$. در (۴۲.۷.۹) قرار می‌دهیم $B = AU$، شکل تجزیهٔ تکین ـ مقدار موردنظر را خواهیم داشت:

$$V^T AU = R$$

یکی از معایب ممکن این شیوه آن است که $A^T A$ را باید تشکیل داد و این کار ممکن است به از دست‌دادن بعضی اطلاعات ناشی از استفاده از حساب با طول متناهی رایانه منجر شود. ولی اگر مسألهٔ ویژه‌مقدار متقارن حلپذیر باشد، اجرای این روش ساده است.

مثال مجدداً ماتریس A در (۲۸.۷.۹) را بر پایهٔ داده‌های جدول ۳.۹ درنظر می‌گیریم. ماتریس $A^T A$ در (۲۹.۷.۹) داده شده است. با به‌کاربردن برنامه‌های EISPACK و LINPACK خواهیم داشت

$$U = \begin{bmatrix} ۰٫۷۸۲۷ & ۰٫۵۹۶۳ & -۰٫۱۷۶۴ & ۰٫۰۲۵۶ \\ ۰٫۴۵۳۳ & -۰٫۳۵۹۶ & ۰٫۷۴۸۹ & -۰٫۳۲۳۱ \\ ۰٫۳۳۲۶ & -۰٫۴۹۹۸ & -۰٫۰۹۸۹ & ۰٫۷۹۳۶ \\ ۰٫۲۶۷۰ & -۰٫۵۱۵۰ & -۰٫۶۳۱۱ & -۰٫۵۱۵۰ \end{bmatrix}$$

مقدارهای تکین چنین‌اند

$$۳۲٫۰۱۰۲, ۳٫۸۹۳۵, ۰٫۱۶۷۴, ۰٫۰۰۲۶$$

ماتریس V متعامد و از مرتبهٔ ۲۱ × ۲۱ است و ما به دلایلی که روشن است آن را نمی‌نویسیم. در عمل، این ماتریس محاسبه نمی‌شود، زیرا حاصلضرب چهار ماتریس هاؤسهولدر است که می‌توان آنها را به شکل ساده‌تری ذخیره کرد.

برای یک بحث مفصلتر در حل مسائل کمترین مربعات گلوب و وَن‌لون (۱۹۸۳، فصل ۶) و کتاب لاوسن و هنسن (۱۹۷۴) را ملاحظه کنید. مسائل عملی بیشتری وجود دارند که باید مورد بحث قرار گیرند، از جمله معین‌کردن رتبهٔ ماتریس وقتی که خطای گردکردن موجب شود که

ماتریس به غلط رتبهٔ کامل پیدا کند. برای برنامه‌ها، پیوست کتاب (۱۹۷۴) و LINPACK را ببینید. برای تجزیهٔ تکین ـ مقدار به LINPACK یا EISPACK مراجعه کنید.

بحث در آثار خواندنی

منبع اصلی اطلاعات این فصل، کتاب دایرةالمعارف گونهٔ معروف ویلکینسن (۱۹۶۵) بوده است. منابع دیگر عبارت بودند از گلوب و وَن‌لون (۱۹۸۳)، گولی و واتسن (۱۹۷۶) استوارت (۱۹۶۴)، نوبل (۱۹۶۹، فصلهای ۹ـ۱۲)، پارلت (۱۹۸۰)، استوارت (۱۹۷۳) و ویلکینسن (۱۹۶۳). برای ماتریسهای با اندازه‌های معمولی، حل عددی مسألهٔ ویژه‌مقدار نسبتاً خوب فهمیده شده است. برای جنبهٔ دیگر روش QR، واتکینز (۱۹۸۲) را ببینید و برای نگاه دقیق به بارست معکوس، پیترز و ویلکینسن (۱۹۷۹) را ملاحظه کنید. الگوریتمهای بسیار عالی برای اغلب مسائل ویژه‌مقدار در ویلکینسن و راینش (۱۹۷۱) و راهنماهای EISPACK توسط اسمیت و همکاران (۱۹۷۶) و گاربو و همکاران (۱۹۷۷) داده شده‌اند. برای تاریخ طرح EISPACK به دونگارا و مولر (۱۹۸۴) مراجعه کنید. یک گزارش کلی عالی از مسائل توسعهٔ نرم‌افزار ریاضی برای مسائل ویژه‌مقداری و سایر مسائل ماتریسی، در رایس (۱۹۸۱) داده شده است. بستهٔ نرم‌افزاری EISPACK پایهٔ بسیاری از برنامه‌های ویژه‌مقدار است که در کتابخانه‌های نرم‌افزاری IMSL و NAG وجود دارند.

بیشتر به علّت محدودیت جا، شماری از مسائل و روشهای عددی در این فصل مورد بحث قرار نگرفته‌اند. برای مسألهٔ ویژه‌مقدار متقارن، روش ژاکوبی ذکر نشده است. این روش یک روش خوب و همگرای سریع برای محاسبهٔ تمام ویژه‌مقدارهای یک ماتریس متقارن است، و برنامه‌ریزی آن نسبتاً ساده است. برای توضیح روش ژاکوبی، گلوب و وَن‌لون (۱۹۸۳، بخش ۴.۸) و پارلت (۱۹۸۰، فصل ۹) و ویلکینسن (۱۹۶۵، صص ۲۶۶ـ۲۸۲) را ببینید. یک برنامه به زبان ALGOL ویلکینسن و راینش (۱۹۷۱، صص ۲۰۲ـ۲۱۱) داده شده است. مسألهٔ تعمیم‌یافتهٔ ویژه‌مقدار $Ax = \lambda Bx$، نیز حذف شده است. در سالهای اخیر این مسأله یک مسألهٔ مهم شده است. متداولترین روش حل آن از مولر و استوارت (۱۹۷۳) است و توضیحات دیگر مسأله و حل آن در گلوب و وَن‌لون (۱۹۸۳، بخشهای ۷.۷ و ۶.۸) و پارلت (۱۹۸۰، فصل ۱۵) داده شده است. برنامه‌های EISPACK برای مسألهٔ ویژه‌مقدار تعمیم‌یافته در گاربو و همکاران (۱۹۷۷) داده شده است.

مسألهٔ پیداکردن ویژه‌مقدارها و ویژه‌بردارهای ماتریسهای بزرگ تنک یک زمینهٔ فعال پژوهشی است. هنگامی که ماتریسها دارای مرتبهٔ بسیار بزرگ باشند (مثلاً $n \geq ۳۰۰$)، استفاده از اغلب روشهای این فصل به‌دلیل ملاحظات حافظه‌ای، دشوارتر است. به‌علاوه، این روشها اغلب توجه خاصی به این موضوع ندارند که در عمل بیشتر این ماتریسها تنک هستند. یک شکل معمول

مسأله شامل ماتریسهای متقارن کراندار است. برنامه‌های مربوط به این مسأله در ویلکینسن و راینش (۱۹۷۱، صص ۲۶۶ـ۲۸۳) و گاربو و همکاران (۱۹۷۷) داده شده‌اند. برای بحثهای کلیتر در مسألهٔ ویژه‌مقدار برای ماتریسهای تنک، جنینگز (۱۹۸۵) و پیسانتسکی[1] (۱۹۸۴، فصل ۶) را ببینید. برای بحث در مورد نرم‌افزار مربوط به مسألهٔ ویژه‌مقدار ماتریسهای تنک، داف[2] (۱۹۸۴، صص ۱۷۹ـ۱۸۲) و هیث[3] (۱۹۸۲) را ملاحظه کنید. یک روش مهم در حل مسألهٔ ویژه‌مقدار برای ماتریسهای تنک متقارن، روش لانتسوش[4] است. برای بحث در این مورد، اسکات[5] (۱۹۸۱) و کتابها و برنامه‌های بسیار جامع کالوم[6] و ویلبی[7] (۱۹۸۴ و ۱۹۸۵) را ببینید.

حل کمترین مربعات دستگاههای خطی فرامعیّن یک وسیلهٔ بسیار باارزش است، وسیله‌ای که در علوم پزشکی، بیولوژیکی و اجتماعی کاربرد گسترده‌ای پیدا کرده است. ما فقط به معرفی برخی از جنبه‌های این موضوع که نقش قاطع تجزیهٔ تکین مقدار را نشان می‌دهد، پرداخته‌ایم. یک مقدمهٔ خیلی جامع برای حل کمترین مربعات دستگاههای خطی در لاوسن و هنسن (۱۹۷۴) داده شده است. این کتاب نحوهٔ به‌کاربردن کلّی این نظریّه و اجرای عملی روشها و طرق استفاده از مجموعه‌های داده‌های بزرگ را به نحو کارایی ارائه می‌دهد. برای مراجع دیگر در حل کمترین مربعات دستگاههای خطی، گلوب و وَن‌لون (۱۹۸۳، فصل ۶) و رایس (۱۹۸۱، فصل ۱۱) را ببینید. برنامه‌هایی برای بعضی از مسائل کمترین مربعات در LINPACK نیز داده شده‌اند.

در بحث حل کمترین مربعات دستگاههای معادله‌های خطی فرامعیّن، از بحث در جنبهٔ آماری موضوع خودداری کرده‌ایم. بخشی به علّت کمبود جا، و بخشی به‌دلیل بی‌اعتمادی در استفاده از توجیه آماری بوده است. زیرا این مسأله اغلب به فرضهایی دربارهٔ توزیع خطا ارتباط پیدا می‌کند که اعتباربخشیدن به آن مشکل است. ما خواننده را به مطالعهٔ یکی ازکتابهای درسی آماری زیادی که یک چارچوب آماری برای روش کمترین مربعات در برازش منحنی داده‌ها ارائه می‌دهند ارجاع می‌دهیم.

مراجع

Chatelin, F. (1987). ***Eigenvalues of Matrices***. **Wiley, London.**

Conte, S., and C. de Boor (1980). ***Elementary Numerical Analysis***, **3rd ed. McGraw-Hill, New York.**

Cullum, J., and R. Willoughby (1984, 1985). ***Lanczos Algorithms for Large Symmetric Eigenvalue Computations***, **Vol. 1,** ***Theory***; **Vol. 2,** ***Programs***. **Birkhäuser, Basel.**

---

1. Pissanetzky 2. Duff 3. Heath 4. Lanczos
5. Scott 6. Culum 7. Willoughby

Dongarra, J., and C. Moler (1984). EISPACK- A package for solving matrix eigenvalue problems. In *Sources and Development of Mathematical Software*, W. Cowell (Ed.), pp. 68–87. Prentice-Hall, Englewood Cliffs, N.J.

Dongarra, J., J. Bunch, C. Moler, and G. Stewart (1979). *LINPACK User's Guide*. SIAM Pub., Philadelphia.

Duff, I. (1984). A survey of sparse matrix software, In *Sources and Development of Mathematical Software*, W. Cowell (Ed.). Prentice-Hall, Englewood Cliffs, N.J.

Garbow, B., J. Boyle, J. Dongarra, and C. Moler (1977). *Matrix Eigensystems Routines—EISPACK Guide Extension*, *Lecture Notes in Computer Science*, Vol. 51. Springer-Verlag, New York.

Golub, G., and C. Van Loan (1983). *Matrix Computations*. Johns Hopkins Press, Baltimore.

Gourlay, A., and G. Watson (1976). *Computational Methods for Matrix Eigenproblems*. Wiley, New York.

Gregory, R., and D. Karney (1969). *A Collection of Matrices for Testing Computational Algorithms*. Wiley, New York.

Heath, M., Ed. (1982). *Sparse Matrix Software Catalog*. Oak Ridge National Laboratory, Mathematics and Statistics Dept., Tech. Rep. Oak Ridge, Tenn.

Henrici, P. (1974). *Applied and Computational Complex Analysis*, Vol. I. Wiley, New York.

Householder, A. (1964). *The Theory of Matrices in Numerical Analysis*. Ginn (Blaisdell), Boston.

Jennings, A. (1985). Solutions of sparse eigenvalue problems. In *Sparsity and Its Applications*, D. Evans (Ed.), pp. 153–184. Cambridge Univ. Press, Cambridge, England.

Lawson, C., and R. Hanson (1974). *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, N.J.

Moler, C., and G. Stewart (1973). An algorithm for generalized matrix eigenvalue problems, *SIAM J. Numer. Anal.* **10**, 241–256.

Noble, B. (1969). *Applied Linear Algebra*. Prentice-Hall, Englewood Cliffs, N.J.

Parlett, B. (1968). Global convergence of the basic *QR* algorithm on Hessenberg matrices, *Math. Comput.* **22**, 803–817.

Parlett, B. (1980). *The Symmetric Eigenvalue Problem*. Prentice-Hall, Englewood Cliffs, N.J.

Peters, G., and J. Wilkinson (1979). Inverse iteration, ill-conditioned equations and Newton's method, *SIAM Rev.* **21**, 339–360.

Pissanetzky, S. (1984). *Sparse Matrix Technology*. Academic Press, New York.

Rice, J. (1981). *Matrix Computations and Mathematical Software*. McGraw-Hill, New York.

Scott, D. (1981). The Lanczos algorithm. In *Sparse Matrices and Their Uses*, I. Duff (Ed.), pp. 139–160. Academic Press, London.

Smith, B. T., J. Boyle, B. Garbow, Y. Ikebe, V. Klema, and C. Moler (1976). *Matrix Eigensystem Routines—EISPACK Guide*, 2nd ed., *Lecture Notes in Computer Science*, Vol. 6. Springer-Verlag, New York.

Stewart, G. (1973). *Introduction to Matrix Computations*. Academic Press, New York.

Watkins, D. (1982). Understanding the *QR* algorithm, *SIAM Rev.* **24**, 427–440.

Wilkinson, J. (1963). *Rounding Errors in Algebraic Processes*. Prentice-Hall, Englewood Cliffs, N.J.

Wilkinson, J. (1965). *The Algebraic Eigenvalue Problem*. Oxford Univ. Press, Oxford, England.

Wilkinson, J. (1968). Global convergence of the tridiagonal *QR* algorithm with origin shifts. *Linear Algebra Its Appl.* **1**, 409–420.

Wilkinson, J., and C. Reinsch, Eds. (1971). *Linear Algebra*. Springer-Verlag, New York.

مسائل

۱. از قضیهٔ ۱.۹، گرشگورین، برای تعیین جای تقریبی ویژه‌مقدارهای ماتریسهای زیر، استفاده کنید.

(الف) $\begin{bmatrix} ۱ & -۱ & ۰ \\ ۱ & ۵ & ۱ \\ -۲ & -۱ & ۹ \end{bmatrix}$ (ب) $\begin{bmatrix} -۲ & ۱ & ۱ \\ ۱ & ۳ & ۱ \\ ۱ & -۱ & ۳ \end{bmatrix}$

هرجا که ممکن است از این نتایج برای تعیین حقیقی یا مختلط‌بودن ویژه‌مقدارها استفاده کنید. برای بررسی این نتایج، ویژه‌مقدارها را مستقیماً از طریق یافتن ریشه‌های چندجمله‌یی مشخصه پیدا کنید.

۲. (الف) چندجمله‌یی زیر داده شده است:

$$p(\lambda) = \lambda^n + a_{n-۱}\lambda^{n-۱} + \ldots + a_۰$$

نشان دهید که برای ماتریس زیر $p(\lambda) = \det[\lambda I - A]$

$$A = \begin{bmatrix} ۰ & ۱ & ۰ & & \cdots & ۰ \\ ۰ & ۰ & ۱ & ۰ & & \vdots \\ \vdots & & & \ddots & & \\ ۰ & & & & ۰ & ۱ \\ -a_۰ & -a_۱ & & \cdots & & -a_{n-۱} \end{bmatrix}$$

ریشه‌های $p(\lambda)$ ویژه‌مقدارهای A هستند. ماتریس A را **ماتریس همراه** چندجمله‌یی $p(\lambda)$ می‌نامند.
(ب) با استفاده از قضیهٔ ۱.۹ گرشگورین، نشان دهید که:

$$|r| \leq ۱ \quad \text{یا} \quad |r + a_{n-۱}| \leq |a_۰| + \ldots |a_{n-۲}|$$

کرانهای ریشه‌های $p(\lambda)$ هستند. اگر این کرانها در صفحهٔ مختلط نواحی مجزا باشند دربارهٔ تعداد ریشه‌ها در هر ناحیه چه می‌توان گفت؟
(ج) قضیهٔ گرشگورین را برای ستونهای A به‌کار برید تا کرانهای دیگری برای ریشه‌های $p(\lambda)$ به‌دست آورید.
(د) از نتایج قسمتهای (ب) و (ج) برای یافتن حدود ریشه‌های معادله‌های چندجمله‌یی زیر استفاده کنید.

(i) $\lambda^{۱۰} + ۸\lambda^{۹} + ۱ = ۰$

(ii) $\lambda^{۶} - ۴\lambda^{۵} + \lambda^{۴} - \lambda^{۳} + \lambda^{۲} - \lambda + ۱ = ۰$

۳. دستگاه خطی (۵.۸.۸) از فصل ۸ را که هنگام حل عددی معادلهٔ پواسون به‌وجود آمده به‌خاطر آورید. اگر معادلات به‌صورت مذکور در ۱۲.۸.۸ و پس از آن مرتب شوند، آنگاه این دستگاه خطی، متقارن با عناصر قطری مثبت است. براساس قضیهٔ ۷.۸ شرط لازم و کافی برای همگرایی روش بارستی گاوس ـ زایدل این است که A معین‌مثبت باشد. از قضیهٔ ۱.۹ گرشگورین برای اثبات معین‌مثبت‌بودن A استفاده کنید. همچنین لازم است قضیهٔ ۸.۸ را برای اینکه $\lambda = ۰$ ویژه‌مقدار ماتریس A نیست ذکر کنید.

۴. مقادیر $\lambda = -۸٫۰۲۸۶۱$ و

$$x = [۱٫۰, ۲٫۵۰۱۴۶, -۰٫۷۵۷۷۳, -۲٫۵۶۴۲۱]^T$$

یک تقریب ویژه‌مقدار و یک تقریب ویژه‌بردار ماتریس زیرند:

$$A = \begin{bmatrix} ۲ & ۱ & ۳ & ۴ \\ ۱ & -۳ & ۱ & ۵ \\ ۳ & ۱ & ۶ & -۲ \\ ۴ & ۵ & -۲ & -۱ \end{bmatrix}$$

نتیجهٔ ۲۲.۱.۹ را برای محاسبهٔ یک کران خطا برای λ به‌کار برید.

۵. برای ماتریس مثال ۱۷.۱.۹ با $\epsilon = ۰٫۰۰۱$ و $\lambda = ۲$ کران خطای اختلال ۳۶.۱.۹ را محاسبه کنید. همین کران در ۳۸.۱.۹ برای ویژه‌مقدار دیگر، $\lambda = ۱$، داده شده بود.

۶. قضیهٔ اختلال ویژه‌بردار (۴۱.۱.۹ را ثابت کنید.

راهنمایی: فرض کنید $\lambda_k(\epsilon)$ و $u_k(\epsilon)$ توابع پیوسته مشتق‌پذیر ϵ باشند. با استفاده از ۳۲.۱.۹ داریم: $\lambda_k'(۰) = u_k^* B u_k / s_k$. بنویسید:

$$u_k(\epsilon) = u_k(۰) + \epsilon u_k'(۰) + O(\epsilon^۲)$$

و آن را نسبت به $u_k'(۰)$ حل کنید. چون $\{u_۱, \ldots, u_n\}$ یک پایه است، بنویسید

$$u_k'(۰) = \sum_{j=۱}^{n} a_j u_j$$

برای یافتن a_j ابتدا از ۴۰.۱.۹ نسبت به ϵ مشتق بگیرید و سپس فرض کنید $\epsilon = ۰$. به‌جای $u_k'(۰)$ نمایش قبلی را بگذارید. از ۲۹.۱.۹ و از دوبه‌دو متعامدبودن ۲۸.۱.۹، یعنی $u_i^* u_j = ۰$ استفاده کنید.

۷. برای ماتریسهای $A(\epsilon)$ در زیر، ویژه‌مقدارها و ویژه‌بردارها را به‌ازای $\epsilon = ۰$ و $\epsilon > ۰$ تعیین کنید. رفتار آنها را هنگامی که $\epsilon \to ۰$ بررسی کنید.

(الف) $\begin{bmatrix} ۱ & ۱ \\ \epsilon & ۱ \end{bmatrix}$ (ب) $\begin{bmatrix} ۱ & ۱ \\ ۰ & ۱+\epsilon \end{bmatrix}$ (ج) $\begin{bmatrix} ۱ & \epsilon \\ ۰ & ۱ \end{bmatrix}$ (د) $\begin{bmatrix} ۱ & ۱ & ۰ \\ ۰ & ۱ & \epsilon \\ ۰ & \epsilon & ۱ \end{bmatrix}$

این مثالها چه چیزی از پایداری زیرفضاهای ویژه‌بردارها را بیان می‌کنند؟

۸. روش توانی را برای محاسبهٔ مقدار غالب و ویژه‌بردار متناظر آن، برای ماتریسهای زیر به‌کار برید.

(الف) $\begin{bmatrix} ۶ & ۴ & ۴ & ۱ \\ ۴ & ۶ & ۱ & ۴ \\ ۴ & ۱ & ۶ & ۴ \\ ۱ & ۴ & ۴ & ۶ \end{bmatrix}$

(ب) $\begin{bmatrix} ۲ & ۱ & ۳ & ۴ \\ ۱ & -۳ & ۱ & ۵ \\ ۳ & ۱ & ۶ & -۲ \\ ۴ & ۵ & -۲ & -۱ \end{bmatrix}$

(ج) $\begin{bmatrix} ۱ & ۳ & -۲ \\ -۱ & -۲ & ۳ \\ ۱ & ۱ & ۲ \end{bmatrix}$

با محاسبهٔ نسبتهای R_m در ۱۴.۲.۹ سرعت همگرایی را بررسی کنید.

هنگامی که نسبتهای R_m تقریباً ثابت‌اند از برونیابی ایتکن برای بهبود سرعت همگرایی ویژه‌مقدار و ویژه‌بردار استفاده کنید. با استفاده از نسبتهای ویژه‌مقدار R_m همگرایی ویژه‌بردارهای $\{z^{(m)}\}$ را تسریع کنید.

۹. روش توانی را برای یافتن ویژه‌مقدار غالب ماتریس

$$A = \begin{bmatrix} ۷ & ۱۳ & -۱۶ \\ ۱۳ & -۱۰ & ۱۳ \\ -۱۶ & ۱۳ & ۷ \end{bmatrix}$$

به‌کار برید. از حدس اولیهٔ $z^{(۰)} = [۱, ۰, ۱]^T$ استفاده کنید. هریک از بارسته‌ای $z^{(m)}$ و $\lambda_۱^{(m)}$ را بنویسید. در باب نتایج توضیح دهید. اگر $\lambda_۱^{(m)}$ با $\lambda_۱^{(m)} = \alpha_m$ تعریف شود چه پیش می‌آید؟

۱۰. برای یک ماتریس A از مرتبهٔ n، فرض کنید شکل متعارف ژوردان آن قطری است و ویژه‌مقدارها با $\lambda_۱$، ...، λ_n نشان داده شده‌اند. فرض کنید $\lambda_۱ = \lambda_۲ = \ldots = \lambda_r$، برای $r > ۱$، و

$$|\lambda_r| > |\lambda_{r+۱}| \geq \ldots \geq |\lambda_n| \geq ۰$$

نشان دهید که روش توانی (۲.۲.۹) ـ (۳.۲.۹) به $\lambda_۱$ و برای اغلب بردارهای اولیهٔ $z^{(۰)}$ به یک ویژه‌بردار متناظر آن همگرا می‌شود.

۱۱. فرض کنید A یک ماتریس متقارن از مرتبهٔ n باشد که ویژه‌مقدارهایش به‌صورت زیر مرتب شده‌اند:

$$\lambda \leq \lambda_۲ \leq \ldots \leq \lambda_n$$

با استفاده از ضرب داخلی استانده، تعریف می‌کنیم:

$$\mathscr{R}(x) = \frac{(Ax, x)}{(x, x)} \qquad x \neq ۰ \quad x \in \mathbf{R}^n$$

نشان دهید هنگامی که $x \neq ۰$ در $\mathscr{R}^n$ است داریم:

$$\mathrm{Max}\mathscr{R}(x) = \lambda_۱ \qquad \mathrm{Min}\mathscr{R}(x) = \lambda_n$$

تابع $\mathscr{R}(x)$ **خارج‌قسمت ریلی** نامیده می‌شود و می‌توان آن را برای مشخص‌کردن ویژه‌مقدارهای باقی‌ماندهٔ A، نیز به‌کار برد. این ماکسیممسازی و مینیمم‌سازی $\mathscr{R}(x)$ پایهٔ بعضی از روشهای عددی کلاسیک برای محاسبهٔ ویژه‌مقدارهای A است.

۱۲. برای به‌دست‌آوردن یک تعبیر هندسی برای ماتریس $n \times n$ هاؤسهولدر، $P = I - ۲ww^T$، فرض کنید $u^{(۲)}, \ldots, u^{(n)}$ یک پایهٔ یکا متعامد برای زیرفضای $(n-۱)$ بعدی عمود بر w باشد. تعریف می‌کنیم

$$T(x) = (I - ۲ww^T)x \qquad x \in \mathbf{R}^n$$

با استفاده از پایهٔ $\{w, u^{(۲)}, \ldots, u^{(n)}\}$ برای $\mathbf{R}^n$ می‌نویسیم:

$$x = a_۱w + a_۲u^{(۲)} + \ldots + a_nu^{(n)}$$

T را برای این نمایش به‌کار برید و نتیجه‌ها را تعبیر کنید.

۱۳. (الف) فرض کنید A یک ماتریس متقارن باشد و λ و x یک جفت ویژه‌مقدار و ویژه‌بردار A با ویژگی $||x||_۲ = ۱$ باشند. فرض می‌کنیم P ماتریس یکا متعامدی باشد که

$$Px = e_۱ = [۱, ۰, \ldots, ۰]^T$$

ماتریس متشابه $B = PAP^T$ را درنظر بگیرید و نشان دهید که اولین سطر و ستون آن صفرند به‌جز عنصر روی قطر که برابر λ است.

راهنمایی: $Be_۱$ را محاسبه و از آن استفاده کنید.

(ب) برای ماتریس

$$A = \begin{bmatrix} ۲ & ۱۰ & ۲ \\ ۱۰ & ۵ & -۸ \\ ۲ & -۸ & ۱۱ \end{bmatrix}$$

$\lambda = ۹$ یک ویژه‌مقدار متناظر با ویژه‌بردار $x = [۲/۳, ۱/۳, ۲/۳]^T$ است. یک ماتریس هاؤسهولدر P بسازید، به‌طوری که $Px = e_۱$ و سپس $B = PAP^T$ را بسازید. مسألهٔ ویژه‌مقدار ماتریس B، به‌آسانی به یک مسأله با ماتریس ۲×۲ بدل می‌شود. از این شیوه برای محاسبهٔ بقیهٔ ویژه‌مقدارها و ویژه‌بردارهای A استفاده کنید. فرایند تبدیل A به B و سپس حل مسألهٔ ویژه‌مقدار یک ماتریس که مرتبه‌اش از مرتبهٔ A یک واحد کمتر است به فرایند **کاهش** معروف است. این فرایند برای تعمیم قابلیت اعمال روش توانی برای ویژه‌مقدارهای دیگر، غیر از ویژه‌مقدار غالب، به‌کار برده می‌شود. برای بحث بیشتر ویلکینسن (۱۹۶۵، صص ۵۸۴ـ۵۹۸) و پارلت (۱۹۸۰، فصل ۵) را ببینید.

۱۴. ماتریسهای هاؤسهولدر را برای تجزیهٔ QR در ماتریسهای زیر به‌کار برید.

(الف) $\begin{bmatrix} ۱ & ۱ & ۱ \\ ۲ & -۱ & -۱ \\ ۲ & -۴ & ۵ \end{bmatrix}$ (ب) $\begin{bmatrix} ۱ & ۳ & -۲ \\ -۱ & -۲ & ۳ \\ ۱ & ۱ & ۲ \end{bmatrix}$

۱۵. ماتریس دوران مرتبهٔ n در زیر را درنظر بگیرید،

$$R^{(k,l)} = \begin{bmatrix} ۱ & ۰ & ۰ & & & \cdots & & & & ۰ \\ ۰ & ۱ & ۰ & & & & & & & \vdots \\ \vdots & & \ddots & & & & & & & \\ ۰ & & & \alpha & ۰ & \cdots & \beta & ۰ & \cdots & ۰ \\ & & & ۰ & ۱ & & ۰ & & & \\ \vdots & & & \vdots & & \ddots & & & & \\ & & & -\beta & ۰ & & \alpha & ۰ & & ۰ \\ & & & & & & & ۱ & & \\ & & & & & & & & \ddots & \vdots \\ ۰ & & & & & \cdots & & & & ۱ \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \text{سطر } k \\ \\ \text{سطر } l \\ \\ \\ \\ \end{matrix}$$

با $\alpha^۲ + \beta^۲ = ۱$. اگر Rb را، برای مقدار داده‌شدهٔ $b \in \mathbf{R}^n$، محاسبه کنیم، تنها عناصری که در مواضع k و l هستند تغییر می‌کنند. با انتخاب مناسب α و β می‌توانیم کاری کنیم که یک صفر در موضع l داشته باشیم. α و β را چنان انتخاب کنید که به‌ازای مقداری از γ،

$$\begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} \begin{bmatrix} b_k \\ b_l \end{bmatrix} = \begin{bmatrix} \gamma \\ ۰ \end{bmatrix}$$

(الف) فرمولهایی برای α و β به‌دست آورید و نشان دهید که $\gamma = \sqrt{b_k^۲ + b_l^۲}$

(ب) با استفاده از دنباله‌ای از ضربهای ماتریسهای دوران:

$$\hat{b} = R^{(۱,۲)} R^{(۱,۳)} R^{(۱,۴)} b$$

$b = [۱, ۱, ۱, ۱]^T$ را به شکل $\hat{b} = [c, ۰, ۰, ۰]$ بدل کنید.

۱۶. نشان دهید که چگونه یک ماتریس دوران $R^{(k,l)}$ را برای تجزیهٔ QR یک ماتریس می‌توان مورداستفاده قرار داد.

۱۷. (الف) مانند بخش ۳.۹، عملها را در به‌دست‌آوردن تجزیهٔ یک ماتریس به QR با استفاده از ماتریسهای هاؤسهولدر شمارش کنید. مطابق معمول ضربها و تقسیمها را یکجا و جذرها را جداگانه بشمارید.

(ب) قسمت (الف) را تکرارکنید، اما این بار برای تبدیل از ماتریسهای دوران $R^{(k,l)}$ استفاده کنید.

۱۸. فرمولهای صریحی برای محاسبهٔ تجزیهٔ QR در یک ماتریس سه‌قطری متقارن به‌دست آورید. تعداد عملها را بشمارید و نتیجه را با آنچه در مسألهٔ ۱۷ به‌دست آمد مقایسه کنید.

۱۹. قضیهٔ ۵.۹ را برای جداکردن ریشه‌های

(الف)
$$\begin{bmatrix} ۰ & ۱ & ۰ & ۰ & ۰ \\ ۱ & ۱ & ۱ & ۰ & ۰ \\ ۰ & ۱ & ۱ & ۱ & ۰ \\ ۰ & ۰ & ۱ & ۱ & ۱ \\ ۰ & ۰ & ۰ & ۱ & ۲ \end{bmatrix}$$

(ب)
$$\begin{bmatrix} ۱ & ۲ & ۰ & ۰ & ۰ \\ ۲ & ۲ & ۳ & ۰ & ۰ \\ ۰ & ۳ & ۳ & ۴ & ۰ \\ ۰ & ۰ & ۴ & ۴ & ۵ \\ ۰ & ۰ & ۰ & ۵ & ۵ \end{bmatrix}$$

به‌کار برید. سپس تقریبهای دقیقی از ریشه‌ها با استفاده از روش تنصیف یا یک تکنیک دیگر ریشه‌یابی به‌دست آورید.

۲۰. (الف) با استفاده از ماتریسهای هاؤسهولدر برای تبدیلات تشابهی، برنامه‌ای برای تبدیل یک ماتریس متقارن به شکل سه‌قطری بنویسید. برای کارایی بیشتر در ضربهای ماتریسی، همانند ضربی که در بخش ۱۸.۳.۹ نشان داده‌شده عمل کنید.

(ب) این برنامه را برای تبدیل ماتریسهای زیر به شکل سه‌قطری به‌کار برید:

(i)
$$\begin{bmatrix} ۱ & ۲ & ۳ \\ ۲ & ۳ & ۵ \\ ۳ & ۵ & ۸ \end{bmatrix}$$

(ii)
$$\begin{bmatrix} ۵ & ۴ & ۱ & ۱ \\ ۴ & ۵ & ۱ & ۱ \\ ۱ & ۱ & ۴ & ۲ \\ ۱ & ۱ & ۲ & ۴ \end{bmatrix}$$

(iii)
$$\begin{bmatrix} ۴ & ۶ & ۲۴۲ & ۱۲ \\ ۶ & ۲۲۵ & ۳ & ۱۸ \\ ۲۴۲ & ۳ & ۲۵ & ۶ \\ ۱۲ & ۱۸ & ۶ & ۰ \end{bmatrix}$$

(ج) ویژه مقدارهای ماتریس سه‌قطری تبدیل‌یافته را با دقت ممکن محاسبه کنید.

۲۱. گیریم $\{p_n(x)|n \geq ۰\}$ یک خانواده از چندجمله‌ییهای متعامد نسبت به یک تابع وزن $\omega(x)$ بر یک بازهٔ $a < x < b$ باشد. به‌علاوه فرض کنید در این چندجمله‌ییها ضریب جملهٔ

پیشرو ۱ باشد:

$$p_n(x) = x^n + \sum_{j=0}^{n-1} a_{n,j}x^j$$

یک ماتریس متقارن سه‌قطری پیدا کنید که $p_n(\lambda)$ چندجمله‌یی مشخصهٔ آن باشد بنابراین محاسبهٔ ریشه‌های یک چندجمله‌یی متعامد (و گرههای فرمول تربیع گاوس) به حل یک مسألهٔ ویژه‌مقدار برای ماتریسهای سه‌قطری متقارن بدل می‌شود.

راهنمایی: رابطهٔ بازگشتی سه‌تایی برای $\{p_n(x)\}$ را به‌خاطر بیاورید و آن را با فرمول (۳.۴.۹) مقایسه کنید.

۲۲. روش QR را (الف) بدون‌انتقال، (ب) با انتقال، برای محاسبهٔ ویژه‌مقدارهای ماتریسهای زیر به‌کار برید

(الف) $\begin{bmatrix} ۳ & ۱ & ۰ \\ ۱ & ۲ & ۱ \\ ۰ & ۱ & ۱ \end{bmatrix}$ (ب) $\begin{bmatrix} ۲ & ۱ & ۰ & ۰ \\ ۱ & ۲ & ۱ & ۰ \\ ۰ & ۱ & ۲ & ۱ \\ ۰ & ۰ & ۱ & ۲ \end{bmatrix}$ (ج) $\begin{bmatrix} ۰ & ۱ & ۰ & ۰ & ۰ \\ ۱ & ۱ & ۱ & ۰ & ۰ \\ ۰ & ۱ & ۱ & ۱ & ۰ \\ ۰ & ۰ & ۱ & ۱ & ۱ \\ ۰ & ۰ & ۰ & ۱ & ۲ \end{bmatrix}$

۲۳. فرض کنید A یک ماتریس هسنبرگ باشد و فرض کنید در تجزیهٔ $A = QR$، Q متعامد و R بالامثلثی باشد.

(الف) با یادآوری بحثی که در پی (۵.۵.۹) آمده؛ نشان دهید که (۷.۵.۹) برقرار است.

(ب) نشان دهید که نتیجهٔ (۷.۵.۹) مستلزم شکلی برای H_k در (۶.۵.۹) است که Q در آن یک ماتریس هسنبرگ است.

(ج) نشان دهید که حاصلضرب یک ماتریس هسنبرگ و یک ماتریس بالامثلثی از هر مرتبه، بازهم یک ماتریس هسنبرگ است. با ترکیب‌کردن این نتایج نشان دهید که RQ باز هم یک ماتریس هسنبرگ است، همان‌گونه که در بند پس از (۷.۵.۹) ادعا شده بود.

۲۴. در ماتریس A در مسألهٔ ۴، $\lambda = ۷٫۹۳۲۹$ و $\lambda = ۵٫۶۶۸۹$ دو ویژه‌مقدار تقریبی دیگر هستند. بارست معکوس را برای محاسبهٔ ویژه‌بردارهای متناظر آنها به‌کار برید.

۲۵. برنامه‌های محاسبهٔ ویژه‌مقدارهای یک ماتریس متقارن حقیقی در مرکز کامپیوترتان را بررسی کنید. از چنین برنامه‌ای برای محاسبهٔ ویژه‌مقدارهای ماتریسهای H_n هیلبرت برای $n = ۳, ۴, ۵, ۶, ۷$ استفاده کنید. برای بررسی صحت جوابهای خودگرگوری[1] وکارنی[2] (۱۹۶۹، صص ۶۶ـ۷۳) را ببینید.

1. Gregory 2. Karney

۲۶. محاسبهٔ ویژه‌مقدارها و ویژه‌تابعهای $x(t)$ ی متناظر با آنها را که در

$$\int_{0}^{1} \frac{x(t)dt}{1+(s-t)^2} = \lambda x(s) \qquad 0 \le s \le 1$$

صدق می‌کنند درنظر بگیرید. یک راه یافتن ویژه‌مقدارهای تقریبی گسسته‌سازی معادله با استفاده از انتگرالگیری عددی است. فرض کنید $h = 1/n$ به‌ازای مقداری از $n \ge 1$ و تعریف کنید $t_j = (j-1/2)h$، $j = 1, \ldots, n$. با گذاردن t_i به‌جای s در این معادله و تقریب‌زدن انتگرال با استفاده از روش انتگرالگیری عددی میانگاهی دستگاه زیر به‌دست می‌آید:

$$h\sum_{j=1}^{n} \frac{\hat{x}(t_j)}{1+(t_i-t_j)^2} = \lambda \hat{x}(t_i) \qquad i = 1, \ldots, n$$

که در آن $\hat{x}(s)$ تابعی است که انتظار داریم $x(s)$ را تقریب بزند. این دستگاه مسألهٔ ویژه‌مقدار برای یک ماتریس متقارن از مرتبهٔ n است. بزرگترین دو ویژه‌مقدار این ماتریس را به‌ازای $n = 2, 4, 8, 16, 32$ پیدا کنید. همگرایی این ویژه‌مقدارها را هنگامی که n افزایش می‌یابد امتحان و سعی کنید خطا را در دقیقترین حالت ($n = 32$)، با مقایسه با ویژه‌مقدارهای واقعی مجهول در معادله انتگرال، پیش‌بینی کنید.

۲۷. نشان دهیدکه معکوس تعمیم‌یافتهٔ A^+ در (۱۱.۷.۹) درشرایط مور- پنروز[1] زیر صدق می‌کند.

۱. $AA^+A = A$ ۲. $A^+AA^+ = A^+$

۳. $(AA^+)^T = AA^+$ ۴. $(A^+A)^T = A^+A$

همچنین نشان دهید:

۵. $(A^+A)^2 = A^+A$ ۶. $(AA^+)^2 = AA^+$

شرایط ۳ تا ۶ نشان می‌دهند که A^+A و AA^+ به‌ترتیب معرف **تصویرهای قائم** بر $\mathbf{R}^n$ و $\mathbf{R}^m$ هستند.

۲۸. برای ماتریس $m \times n$ دلخواه A، نشان دهید که به‌ازای $\alpha > 0$

$$\lim_{\alpha \to 0+} (\alpha I + A^TA)^{-1}A = A^+$$

راهنمایی: از تجزیهٔ تکین ـ مقدار ماتریس A استفاده کنید.

۲۹. برخلاف وضعیت ماتریسهای مربع ناتکین، نیاز نیست وارون تعمیم‌یافتهٔ A^+، با تغییرات A، به‌طور پیوسته تغییر کند. برای تأیید این مطلب یک خانواده $\{A(\epsilon)\}$ بیابید به‌طوری که ماتریس $A(\epsilon)$ به $A(0)$ همگرا باشد ولی $A(\epsilon)^+$ به $A(0)^+$ همگرا نباشد.

1. Moore-Penrose

۳۰. برازش کمترین مربعات چندجمله‌یی خطی را برای داده‌های زیر حساب کنید. نمودار داده‌ها و برازش کمترین مربعات را رسم کنید. همچنین خطای ریشهٔ دوم میانگین مربع را در برازش کمترین مربعات بیابید.

t_i	b_i	t_i	b_i	t_i	b_i
−۱٫۰	۱٫۰۳۲	−۰٫۳	۱٫۱۳۹	۰٫۴	−۰٫۴۱۵
−۰٫۹	۱٫۵۶۳	−۰٫۲	۰٫۶۴۶	۰٫۵	−۰٫۱۱۲
−۰٫۸	۱٫۶۱۴	−۰٫۱	۰٫۴۷۴	۰٫۶	−۰٫۸۱۷
−۰٫۷	۱٫۳۷۷	۰٫۰	۰٫۴۱۸	۰٫۷	−۰٫۲۳۴
−۰٫۶	۱٫۱۷۹	۰٫۱	۰٫۰۶۷	۰٫۸	−۰٫۶۲۳
−۰٫۵	۱٫۱۸۹	۰٫۲	۰٫۳۷۱	۰٫۹	−۰٫۵۳۶
−۰٫۴	۰٫۹۱۰	۰٫۳	۰٫۱۸۳	۱٫۰	−۱٫۱۷۳

۳۱. یک برازش مربعات درجهٔ دوم برای داده‌های زیر درست کنید، از شکل استانداردِ

$$g(t) = x_1 + x_2 t + x_3 t^2$$

و معادلهٔ نرمال ۲۶.۷.۹ استفاده کنید. ضریب وضعیت $A^T A$ چیست؟

t_i	b_i	t_i	b_i	t_i	b_i
−۱٫۰	۷٫۹۰۴	−۰٫۳	۰٫۳۳۵	۰٫۴	−۰٫۷۱۱
−۰٫۹	۷٫۴۵۲	−۰٫۲	−۰٫۲۷۱	۰٫۵	۰٫۲۲۴
−۰٫۸	۵٫۸۲۷	−۰٫۱	−۰٫۹۶۳	۰٫۶	۰٫۶۸۹
−۰٫۷	۴٫۴۰۰	۰٫۰	−۰٫۸۴۷	۰٫۷	۰٫۸۶۱
−۰٫۶	۲٫۹۰۸	۰٫۱	−۱٫۲۷۸	۰٫۸	۱٫۳۵۸
−۰٫۵	۲٫۱۴۴	۰٫۲	−۱٫۳۳۵	۰٫۹	۲٫۶۱۳
−۰٫۴	۰٫۵۸۱	۰٫۳	−۰٫۶۵۶	۱٫۰	۴٫۵۹۹

۳۲. برای ماتریس A که در برازش منحنی کمترین مربعات مسألهٔ ۳۱ به‌دست آمد تجزیهٔ آن به QR و تجزیهٔ تکین ـ مقدار و همچنین وارون تعمیم‌یافتهٔ آن را به‌دست آورید. و از آنها برای حل مسألهٔ کمترین مربعات استفاده کنید.

۳۳. تجزیهٔ QR، تجزیهٔ تکین مقدار و وارون تعمیم‌یافتهٔ ماتریسهای زیر را به‌دست آورید. همچنین $\text{cond}(A)_۲$ را نیز پیدا کنید.

$$A = \begin{bmatrix} ۰٫۹ & ۱٫۱ \\ -۱٫۰ & -۱٫۰ \\ ۱٫۱ & ۰٫۹ \end{bmatrix} \quad \text{(الف)} \qquad A = \begin{bmatrix} ۱ & ۲ & ۳ \\ ۲ & ۳ & ۴ \\ ۳ & ۴ & ۵ \\ ۴ & ۵ & ۶ \end{bmatrix} \quad \text{(ب)}$$

۳۴. (الف) فرض می‌کنیم A ماتریسی $m \times n$، $m \geq n$، باشد و ستونهای b تقریباً وابسته باشند. دقیقتر بگوییم فرض کنید $A = [u_۱, \ldots, u_n]$، $u_j \in \mathbf{R}^m$، و فرض کنید بردار

$$v = \alpha_۱ u_۱ + \ldots + \alpha_n u_n$$

در مقایسه با $||\alpha||_۲$، $\alpha = [\alpha_۱, \ldots, \alpha_n]^T$، بسیار کوچک باشد. نشان دهید که A ضریب وضعیت بزرگی خواهد داشت.

(ب) برخلاف قسمت (الف) فرض می‌کنیم ستونهای A یکا متعامد باشند نشان دهید که $\text{cond}(A)_۲ = ۱$.

پیوستها

نرم‌افزار ریاضی

در آغاز دهۀ ۱۹۷۰، چند برنامۀ پیشرفتۀ رایانه‌یی در موضوعهای اصلی آنالیز عددی نوشته شده‌اند. این برنامه‌ها، برای هر نوع رایانه کارا، قابل اعتماد و قابل اجرا هستند. این موضوعهای آنالیز عددی و علوم رایانه‌یی اکنون **نرم‌افزار ریاضی** نامیده می‌شوند، اصطلاحی که از زبان رایس[1] (۱۹۷۱) متداول شده است. اکنون بسته‌هایی برنامه‌یی با کیفیت عالی برای بیشتر موضوعهای آنالیز عددی وجود دارند. اغلب این برنامه‌ها در دسترس عموم هستند و اغلب جزئی از کتابخانه‌های آنالیز عددی تجاری شده‌اند. برای یک مطالعۀ کلی و گزارش تاریخی این بسته‌های آنالیز عددی کاول[2] (۱۹۸۴) را ببینید و برای سایر کتبی که بحث در این مورد را به طریق قابل توجهی انجام داده‌اند رایس (۱۹۷۷) ، (۱۹۸۱) و (۱۹۸۳) را ملاحظه کنید. در کتاب فورسایت و دیگران (۱۹۷۷) چند برنامه در مسائل مختلف آنالیز عددی دقیقاً انتخاب شده‌اند که کارایی زیادی دارند و نسبتاً به سادگی قابل فهم‌اند.

دو دسته کتابخانه‌های تجاری بزرگی در آنالیز عددی برای رایانه‌های بزرگ، کتابخانه‌های IMSL و کتابخانۀ NAG وجود دارند که هر دو در همۀ خطوط رایانه‌یی، از جمله ریزرایانه‌ها، در دسترس هستند. در این کتابخانه‌ها خیلی از بسته‌های خاصی که در پاراگراف بعدی به آنها اشاره خواهد شد وجود دارند. دو شرکت IMSL و INAG همچنین متنهای خاص دیگری برای استفاده در خط لولۀ رایانه‌های برداری به وجود آورده‌اند. کتابخانه‌های تجاری آنالیز عددی دیگری نیز وجود دارند، ولی معمولاً در انواع کمتری از رایانه‌ها به خدمت گرفته می‌شوند.

ما قویاً به دانشجویان و پژوهشگران توصیه می‌کنیم که ابتدا به این کتابخانه‌های تجارتی برای برنامۀ مورد نیاز خود مراجعه کنند. برنامه‌ها خوب نوشته شده‌اند، توجه خاصی برای سهولت استفاده، دقت و کنترل خطا و کارائی آنها مبذول شده است. این برنامه‌ها بطور مداوم بهتر می‌شوند تا آخرین پژوهشها را در روشهای عددی و تغییرات سخت افزاری رایانه‌ها منعکس کنند. در مقایسه با بسته‌های نرم‌افزاری خاصی که در ذیل به آنها اشاره می‌کنیم، این کتابخانه‌های تجاری بزرگ از نظر کاربرد به مراتب ساده‌ترند و پیاده‌سازی آنها را کارمندان مرکز محاسبات شما می‌توانند انجام دهند بی‌آنکه نیاز به پرداختن به زمینه‌های تخصصی داشته باشند. بسته‌های خاصی که بعداً توضیح داده می‌شوند فقط باید در مواردی به کار روند که کتابخانه‌های تجارتی نیاز کاربر را برآورده ننمایند. اغلب این بسته‌های خاص، ثابت‌اند و برای زمان خاصی نوشته شده‌اند و به روز در نیامده‌اند که تغییرات در الگوریتمها و رایانه‌ها را منعکس کنند.

در صورت امکان و در صورت دسترسی به بسته‌های استاندارد، ما نباید برای مسائل خود

1. Rice 2. Cowell

برنامه‌نویسی کنیم. نوشتن یک برنامهٔ خوب رایانه‌یی بسیار وقت‌گیر و مشکل است، و غیر محتمل است که یک کاربر معمولی بتواند برنامه‌نویسی خود را به پای برنامهٔ کتابخانه‌های تجاری استاندارد برساند.

کتابخانهٔ IMSL به سه مؤلفهٔ اصلی تقسیم شده است. (۱) کتابخانهٔ آنالیز عددی، (۲) کتابخانهٔ آمار و (۳) کتابخانهٔ توابع خاص. هر کتابخانه از زیربرنامه‌هایی به زبان فورترن ۷۷ تشکیل یافته است که از برنامهٔ ارائه شده کاربر فراخوانی می‌شوند. به علاوه، IMSL ، بسته‌های اضافی نیز دارد که استفاده از کتابخانه اصلی را در سطح بسیار بالائی به شکل محاوره‌یی، برای کاربر، ممکن می‌سازد. IMSL همچنین به عنوان یک عامل توزیع چند برنامه برای عموم نیز عمل می‌کند. نشانی IMSL چنین است:

IMSL , Inc.
NBC Building
7500 Bellaire Boulevard
Houston, Texas 77036-5085

از آنجا که کتابخانه‌های آنها به طور مداوم بهنگام می‌شوند، برای اطلاع از موضوعات روز کتابخانه‌ها، می‌توانید با IMSL مکاتبه کنید.

کتابخانهٔ NAG تقریباً همان موضوعات کتابخانهٔ IMSL را دارد و بسته‌های کمکی نیز برای ساده کردن استفاده از کتابخانهٔ اصلی را دارا می‌باشد. برنامه‌های این کتابخانه هم به زبان فورترن و هم به زبان Algol در دسترس‌اند[۱]. برنامه‌های کتابخانه را دانشمندان دانشگاههای مختلف انگلستان نوشته‌اند و بسیاری از بسته‌های نرم‌افزاری خاص که بعداً توضیح داده خواهند شد جزئی از کتابخانهٔ NAG شده‌اند. برای اطلاعات بیشتر با آدرس زیر مکاتبه کنید:

Numerical Algorithm Group, Ltd.
NAG Central office
Mayfield House,
256 Bandburg Road
Oxford ox2 7DE, United Kingdom.

بسته‌های رایانه‌یی زیادی با کیفیت بالا برای مسائل زمینه‌های خاص از ۱۹۷۲ نوشته شده‌اند. اولین، و شاید معروفترین آنها، بستهٔ EISPACK برای حل مسائل ویژه مقادیر ماتریسها است. پس از آن بسته‌هایی در انواع گوناگون، برای بیشتر زمینه‌های اصلی آنالیز عددی، تهیه شده‌اند،

۱. از آنجا که این کتاب در دهه ۱۹۸۰ نوشته شده است به زبانهای پاسکال و C اشاره نشده است.ـم.

و بسیاری از آنها تاکنون در کتابهای درسی گنجانده شده‌اند. بیشتر این نرم‌افزارها در کتاب کاول (۱۹۸۲) به نحو مطلوبی وارد شده‌اند. ما اکنون بسته‌هایی را که از آنها اطلاع داریم ذیلاً می‌آوریم و از اینکه بعضی را از قلم انداخته‌ایم پوزش می‌خواهیم. در بعضی موارد منبع برنامه‌ها را نیز ذکر می‌کنیم. به علاوه، تقریباً تمام این برنامه‌ها، را (بویژه آنهایی را که در آزمایشگاههای ملی ایالات متحده، تولید شده‌اند)، با محدودیتهایی، می‌توان از مؤسسهٔ زیر دریافت کرد:

National Energy Software Center
Argonne National Laboratory
9700 South Cass Avenue
Argonne, Illinois 60439

۱. **EISPACK** این بستهٔ نرم‌افزاری مسألهٔ ویژه مقدار ماتریس را برای انواع گوناگون ماتریسها حل می‌کند، همچنین شامل برنامه‌هایی برای مسألهٔ ویژه مقدار تعمیم‌یافتهٔ $Ax = \lambda Bx$ و برای محاسبهٔ تجزیهٔ ماتریس است. راهنمای کاربران برای این بسته در اسمیت و همکاران و (۱۹۷۶) و در گاربو و همکاران (۱۹۷۷) داده شده است. برای شرح توسعهٔ این بسته، کاول (۱۹۸۴، فصل ۴) را ببینید. IMSL یکی از توزیع‌کنندگان این بستهٔ نرم‌افزاری است.

۲. **LINPACK** این بستهٔ نرم‌افزاری برای حل دستگاههای معادلات خطی است. این برنامه چهار متن دارد: حساب حقیقی و مختلط با دقت ساده و مضاعف. علاوه بر برنامه‌های معمولی برای انواع مختلف دستگاههای ناتکین مربعی، LINPACK حاوی برنامه‌هایی برای حل دستگاههای خطی کمترین مربعات نیز می‌باشد. برای راهنمای کاربران دونگارا و همکاران (۱۹۷۹) را ببینید، و برای شرح توسعه LINPACK ، کاول (۱۹۸۴، فصل ۲) را ملاحظه کنید. IMSL یکی از توزیع کنندگان این بستهٔ نرم‌افزاری است.

۳. **MINPACK** این بستهٔ نرم‌افزاری برای حل دستگاههای معادلات غیرخطی و مسائل بهینه‌سازی است. در حال حاضر فقط متن اول این بسته وجود دارد که برای (الف) حل n معادلهٔ غیرخطی n مجهولی و (ب) برای حل کمترین مربعات دستگاههای معادلات فرا نامعین غیرخطی، به کار می‌رود. متن دیگر قرار است شکلهای مختلف دیگر مسائل بهینه‌سازی مقید و نامقید را در بر گیرد. برای راهنمای کاربران، مورهٔ[1] و همکاران (۱۹۸۰) و برای شرح پروژه کاول (۱۹۸۴، فصل ۵) را ببینید. IMSL یکی از توزیع‌کنندگان این بستهٔ نرم‌افزاری است.

1. Moré

۴. **LLSQ** این یک بستهٔ نرم‌افزاری برای حل کمترین مربعات دستگاههای معادلات خطی است. این برنامه همراه کتاب لاوسن و هنسن آمده و از IMSL می‌توان آن را تهیه نمود.

۵. **QUADPACK** این بستهٔ نرم‌افزاری برای انتگرالگیری عددی توابع یک متغیره است. هم برنامه‌های کلی و هم برنامه‌هایی برای انتگرالهایی با شکل خاص در آن وجود دارند. برای راهنمای کاربران، پیسنز و همکاران (۱۹۸۳) را ببینید.

۶. **DEPAC** این بسته برنامه‌هایی است برای حل دستگاههای معادلات دیفرانسیل معمولی. در حال حاضر شامل سه برنامهٔ اصلی است: یک برنامهٔ متغیر ـ مرتبهٔ ادامز، یک برنامهٔ ثابت مرتبهٔ رونگه ـ فلبرک، و یک برنامهٔ متغیر مرتبه برای مسائل سرسخت بر پایهٔ فرمولهای مشتقگیری پسرو. این برنامه‌ها در آزمایشگاه ملی سندیا[1] در شهر آلبوکرک[2] در نیومکزیکو[3] موجودند. برای بحث در مورد تولید نرم‌افزاری در معادلات دیفرانسیل معمولی کاول (۱۹۸۴، فصل ۶) را ملاحظه کنید.

۷. **GRDI** این یک بسته برنامه‌هایی برای حل معادلات دیفرانسیل معمولی و جزئی است که این آخری براساس روش خطوط پایه‌گذاری شده است. برنامه‌هایی برای معادلات دیفرانسیل معمولی بر پایهٔ روشهایی است که هایندمارش و همکاران او در آزمایشگاه لاورنس لیورمور[4] ایجاد کرده‌اند. برای کسب اطلاعات یا دریافت این بستهٔ نرم‌افزاری با آدرس زیر مکاتبه کنید:

Dr. W. E. Schiesser
Whitaker Laboartory No.5
Lehigh University
Bethlehen, Pennsylvania 18015

۸. **FISHPAK** این یک بستهٔ نرم‌افزاری برای حل تقریبهای تفاضلات متناهی برای بعضی از دسته‌های معادلات دیفرانسیل جزئی تفکیکپذیر معروف (مثلاً معادلهٔ پواسن و معادلهٔ هلمهولتس) در نواحی خاص، مانند مستطیلها یا دوایر، است. این بسته برای محاسبات مکانیک سیالات در NCAR (مرکز ملی برای تحقیقات جوی) تولید شده است. برای بحث در مورد این بسته اسوارتس تراوبر[5] و سوییت[6] (۱۹۷۹) را ببینید. برای یک آشنایی کلیتر با نرم‌افزار حل معادلات دیفرانسیل جزئی کاول (۱۹۸۴، فصل ۹) را ملاحظه کنید. این کتاب شامل بحثی است در ELLPACK [برای معادلات دیفرانسیل جزئی بیضوی، رایس (۱۹۷۷، صص ۳۱۹ـ۳۴۱) را ببینید] و ITPACK

1. Sandia 2. Albuquerque 3. New Mexico
4. Lawronce Livormore 5. Swarztrauber 6. Sweet

[برای روشهای بارستی حل معادلات تفاضلات متناهی، هاگمن و یانگ (۱۹۸۱) را ببینید]، می‌باشد.

علاوه بر بسته‌های نرم‌افزاری فوق‌الذکر، بسیاری از برنامه‌ها و الگوریتمها در ACM, Transactions on Mathematical Software انتشار یافته‌اند. برای لیست بعضی از الگوریتمهای موجود به یک شمارهٔ این مجله مراجعه کنید. برنامه‌های اصلی را می‌توان از IMSL تهیه نمود.

یک مبادلهٔ نرم‌افزاری الکترونیکی یک وسیلهٔ به دست آوردن بسیاری از نرم‌افزارهای فوق‌الذکر توسط پست الکترونیکی مستقر در آزمایشگاه ملی آرگون در شیکاگو، ایلینویز، است. این سیستم NETLIB خوانده شده و می‌توان از طریق آدرس پست الکترونیکی زیر با آن تماس برقرار نمود

NETLIB@AML-MCS on ARPANET

پیام زیر را به این آدرس بفرستید تا راههای استفاده از NETLIB را دریافت دارید

send index

این سیستم در زمان نوشتن این کتاب وجود داشته است ولی هیچ تضمینی برای ادامهٔ موجودیت آن وجود ندارد. یک ابزار بسیار مفیدی است برای به دست آوردن نرم‌افزار ریاضی با کیفیت بالا.

نرم‌افزار برای ریزرایانه‌ها. کتابخانه‌های IMSL و NAG ، زیر مجموعه‌های خاصی از کتابخانه‌های خود را برای تعدادی از ریزرایانه‌ها از جمله ماشینهای سازگار با IBM شخصی، در دسترس دارند. به علاوه بسته‌های دیگری نیز وجود دارند که پیوسته توسعه می‌یابند. توجه خاص ما به بسته‌هایی است که دسترسی آسانتری به ابزارهای پیچیدهٔ آنالیز عددی را به‌وجود می‌آورند. دو بستهٔ متداول از این نوع، PC-MATLAB و GAUSS هستند، که هر دو دسترسی به تعدادی از برنامه‌های موجود در بسته‌های EISPACK و LINPACK را همراه با تسهیلات آنالیز عددی و گرافیک دیگر مهیا می‌سازند. این نرم‌افزارها در اغلب ریزرایانه‌ها و ایستگاههای علمی در دسترس‌اند. برای اطلاعات در مورد این بسته‌ها با آدرس زیر مکاتبه کنید.

برای GAUSS:

APTECH System, Inc.

P. O. Box 6487

Kent, Washington 98064

برای PC-MATLAB:

The Math Works, Inc.
24 Prime Park Way
Natick, MA01760-1500
Fax: (508) 653-2997
E-mail: info@Mathworks.com

مراجع

Cowell, W., Ed. (1984). *Sources and Development of Mathematical Software*. Prentice-Hall, Englewood Cliffs, N.J.

Dongarra, J., J. Bunch, C. Moler, and G. Stewart (1979). *LINPACK User's Guide*. SIAM Pub., Philadelphia.

Forsythe, G., M. Malcolm, and C. Moler (1977). *Computer Methods for Mathematical Computations*. Prentice-Hall, Englewood Cliffs, N.J.

Garbow, B., J. Boyle, J. Dongarra, and C. Moler (1977). *Matrix Eigensystem Routines—EISPACK Guide Extension*. *Lecture Notes in Computer Science 51*. Springer-Verlag, New York.

Hageman, L., and D. Young (1981). *Applied Iterative Methods*. Academic Press, New York.

Lawson, C., and R. Hanson (1974). *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, N.J.

Moré, J., B. Garbow, and K. Hillstrom (1980). *User Guide for MINPACK-1*. Argonne National Laboratory Rep. ANL-80-74, Chicago, Ill.

Piessens, R., E. deDoncker-Kapenga, C. Überhuber, and D. Kahaner (1983). *QUADPACK: A Subroutine Package for Automatic Integration*. Springer-Verlag, New York.

Rice, J., Ed. (1971). *Mathematical Software*. Academic Press, New York.

Rice, J., Ed. (1977). *Mathematical Software III*. Academic Press, New York.

Rice, J. (1981). *Matrix Computations and Mathematical Software*. McGraw-Hill, New York.

Rice, J. (1983). *Numerical Methods, Software, and Analysis*. McGraw-Hill, New York.

Smith, B., J. Boyle, B. Garbow, Y. Ikebe, V. Klema, and C. Moler (1976). *Matrix Eigensystem Routines—EISPACK Guide*, 2nd ed., *Lecture Notes in Computer Science 6*. Springer-Verlag, New York.

Swarztrauber, P., and R. Sweet (1979). Algorithm 541: Efficient Fortran subprograms for the solution of separable elliptic partial differential equations. *ACM Trans. Math. Softw.* 5, 352–364.

پاسخهای تمرینهای انتخابی

فصل ۱

۵. (الف)

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

$$|\text{خطا}| \leq \frac{|x|^{۲n}}{(۲n+۱)(n!)}$$

۷. (الف) $p_۱(x,y) = ۱ + x - \frac{۱}{۲}y$، $p_۲(x,y) = p_۱(x+y) - \frac{۱}{۲}x^۲ + \frac{۱}{۲}xy - \frac{۱}{۸}y^۲$

۱۰. (الف) ۲۱٫۶۲۵ (ج) $\frac{۲}{۳}$ (د) $\frac{۲}{۳}$

۱۱. (ب) ۱۱۱۱۱۱۱

۱۲. (الف) ۰٫۱۱۰۱ (ج) ۰٫۰۰۰۱۱۰۰۱۱۰۰ ...

۲۱. (الف) ۴ (ب) ۲ (ج) ۳

۲۲. (الف) [۲٫۰۵۲۶۵, ۲٫۰۵۳۷۵]

(د) $[۱۳۳٫۴۴۰۹, ۱۳۱٫۳۵۶۶] \doteq \left[\frac{۸٫۴۷۳۵}{۰٫۰۶۳۵}, \frac{۸٫۴۷۲۵}{۰٫۰۶۴۵}\right]$

۲۶. (ب) برای $x \doteq ۰$ تقریب سری درجه دوم تیلر عبارت است از

$$f(x) \doteq \frac{-۵}{۲۴} - \frac{۱۱}{۴۸}x - \frac{۳۷۹}{۱۹۲۰}x^۲$$

همچنین $f(۰)$ دقیقاً برابر است با $\frac{-۵}{۲۴}$.

۲۸. (الف) ۰٫۰۰۳۰− (ج) ۰٫۰۰۰۶۸

۳۰. (الف) با $x_A = ۳٫۱۴$ و $y_A = ۲٫۶۸۵$ خطا تقریباً برابر است با $-۱٫۴۶(x_T - x_A) - ۱٫۷۱(y_T - y_A)$ ≒ خطا x_T و y_T اعداد واقعی گردنشده مربوط به x_A و y_A هستند همچنین ۰٫۰۰۹۸ ≥| خطای نسبی |

فصل ۲

۱. $\prod_{j=۰}^{\infty}[۱ + r^{۲^j}] = ۱/(۱ - r)$ و حاصلضرب نامتناهی همگراست اگر و تنها اگر $|r| < ۱$.

۳. (الف) ریشه برابر است با ۴٫۴۹۳۴۰۹۴۵۸

۴. (ب) ۱٫۸۳۹۲۸۶۸ (د) ۱٫۱۲۸۴۲۵۱

۵. (الف) ۴٫۴۹۳۴۰۹۴۵۸ (ب) ۹۸٫۹۵۰۰۶۲۸۲

۶.

B	ریشه
۱	−۰٫۵۸۸۴۰۱۷۷۶۵
۵	−۰٫۴۰۴۹۱۱۵۴۸۲
۱۰	−۰٫۳۲۶۵۰۲۰۱۰۱
۲۵	−۰٫۲۳۷۴۳۶۲۴۳۹
۵۰	−۰٫۱۸۳۲۹۱۳۳۳۳

۱۲. (ه) $|\mathrm{Rel}(x_۴)| \le ۲^{-۱۵}۱۰^{-۱۶} \doteq ۳٫۰۵ \times ۱۰^{-۲۱}$

۱۳. بارست $x_۴$ به اندازهٔ کافی دقیق خواهد بود.

۱۵. (الف) همگرا نیست. (ب) همگراست.

۱۹. $\alpha = ۲٫۱۳۲۲۶۷۷$، $[a, b] = [۱٫۱, ۱ + \frac{\pi}{۲}]$

۲۴. (الف) واگراست. (ج) همگراست.

۲۵. $\lim_{n\to\infty} \frac{\sqrt{a} - x_{n+۱}}{(\sqrt{a} - x_n)^۳} = \frac{۱}{۴a}$

۲۹. (الف) نرخ ≒ ۰٫۶۲۵

۳۹. (ب) $۲٫۴۷۰۶۳۸۹۷۰ + ۴٫۶۴۰۵۳۳۱۶۲i$

۴۰. (الف) این چندجمله‌یی، چندجمله‌یی لژاندار از درجهٔ ۱۲ بر بازهٔ $-۱ \le x \le ۱$ است. ریشه‌های آن عبارت‌اند از:

$\pm ۰٫۷۶۹۹۰۲۶۷۴۲$ $\pm ۰٫۱۲۵۲۳۳۴۰۸۵$

$\pm$۰٫۹۰۴۱۱۷۲۵۶۴ $\pm$۰٫۳۶۷۸۳۱۴۹۹۰

$\pm$۰٫۹۸۱۵۶۰۶۳۴۲ $\pm$۰٫۵۸۷۳۱۷۹۵۴۳

.۵۱

(x_0, y_0)	روش همگراست به
(۱٫۲, ۲٫۵)	(۱٫۳۳۶۳۵۵۳۷۷, ۱٫۷۵۴۲۳۵۱۹۸)
(−۲, ۲٫۵)	(−۰٫۹۰۱۲۶۶۱۹۰۸, −۲٫۰۸۶۵۸۷۵۹۵)
(−۱٫۲, −۲٫۵)	(−۰٫۹۰۱۲۶۶۱۹۰۸, −۲٫۰۸۶۵۸۷۵۹۵)
(۲٫۰, −۲٫۵)	(−۳٫۰۰۱۶۲۴۸۸۷, ۰٫۱۴۸۱۰۷۹۹۵۰)

در آخرین حالت، این روش ظاهراً به‌طور تصادفی بالا و پایین پرش دارد. و بعد از ۱۶ بارست به ریشه‌ای که قبلاً داده‌شده همگرا می‌شود. معادلات حتماً ریشهٔ دیگری دارند.

.۵۲ (الف) دو ریشه وجود دارد (a, b) و (b, a) با $a =$ ۰٫۶۷۳۰۰۷۱۷۰، $b =$ ۱٫۹۴۵۰۲۶۸۲

فصل ۳

.۴

$$|\text{خطای درونیابی}| \leq (\frac{h^2}{8})e^2 \doteq ۹٫۲ \times ۱۰^{-۵}$$

$$|\text{خطای گردکردن}| \leq ۵ \times ۱۰^{-۵}$$

$$|\text{خطای کل}| \leq ۱٫۴۲ \times ۱۰^{-۴}$$

.۷ h را برابر ۰٫۰۰۱ انتخاب و فرض کنید درایه‌های جدول هفت رقم بامعنی دارند. خطای کل با $۸٫۴ \times ۱۰^{-۷}$ کراندار می‌شود.

.۱۰ درایه‌های جدول را با ۶ رقم بامعنی انتخاب کنید. یک افراز ممکن $۱ \leq x \leq ۱۰$ با اندازهٔ شبکهٔ پیشنهادی و خطای درونیابی کلی حاصل در جدول زیر نشان داده شده است.

بازه	h	خطای کلی
$۱ \leq x \leq ۲$	۰٫۰۱	$۶٫۸۱ \times ۱۰^{-۷}$
$۲ \leq x \leq ۳$	۰٫۰۲۵	$۷٫۳۵ \times ۱۰^{-۷}$
$۳ \leq x \leq ۶$	۰٫۰۵	$۸٫۸۵ \times ۱۰^{-۷}$
$۶ \leq x \leq ۱۰$	۰٫۱	$۸٫۸۵ \times ۱۰^{-۷}$

.۲۰ درجهٔ ۳

۲۲. ششمین ورودی ۴۱۹۳۲۷ باید به ۴۱۹۲۳۷ یا ۴۱۹۲۳۸ تغییر یابد. دو خطای دیگر وجود دارند و اثر آنها در تفاضلات مرتبهٔ بالاتر روی یکدیگر می‌افتند.

۲۹. شرط لازم و کافی برای اینکه چندجمله‌یی درونیاب با درجهٔ نابیشتر از ۲ یکتا باشد

$$x_1 \neq \frac{1}{2}(x_0 + x_2)$$

۳۴. $s(x)$ بر بازهٔ $[x_0, x_2]$ با فرمول زیر داده شده است:

$$s(x) = \frac{(x_2 - x)^3 M_0 + (x - x_0)^3 M_2}{6(h_0 + h_1)} + \frac{(x_2 - x)y_0 + (x - x_0)y_2}{h_0 + h_1} - \frac{h_0 + h_1}{6}[(x_2 - x)M_0 + (x - x_0)M_2]$$

$s(x_1) = y_1$ را تشکیل دهید تا یک شرط درونیاب در x_1 برقرار کنید، که یک معادله برای M_0 و M_2 خواهد داد. شکل مشابهی برای $s(x)$ بر بازهٔ $[x_{n-2}, x_n]$ با رابطهٔ $s(x_{n-1}) = y_{n-1}$ به‌کار برده می‌شود.

۳۶. (الف) برای $n = 96$ زیربازه، خطای $s(x)$ برای شرایط مرزی مختلف عبارت‌اند از:

(الف) $1{,}35\text{E} - 10$ (ب) $8{,}32\text{E} - 11$ (ج) $8{,}25\text{E} - 10$

۳۹. $$B_i^{(m)}(x) = \frac{x_{i+m} - x}{x_{i+m} - x_{i+1}} B_{i+1}^{(m-1)}(x) - \frac{x - x_i}{x_{i+m-1} - x_i} B_i^{(m-1)}(x)$$

۴۳. (الف) $d_0 = 1$؛ $d_k = 0$ به‌ازای $1 \leq k \leq m - 1$

فصل ۴

۱. چندجمله‌یی درجهٔ ۴ برنشتاین عبارت است از:

$$p_4(x) = 2\sqrt{2}x(1 - x)^3 + 6x^2(1 - x)^2 + 2\sqrt{2}x^3(1 - x)$$

چندجمله‌یی درجهٔ ۴ تیلر عبارت است از:

$$f_4(x) = 1 - \frac{\pi^2}{2}\left(x - \frac{1}{2}\right)^2 + \frac{\pi^4}{24}\left(x - \frac{1}{2}\right)^4$$

جدول زیر مقادیر $p_4(x)$ و $f_4(x)$ را در چندین نقطه همراه با خطاها هنگامی که با $\sin(\pi x)$

مقایسه می‌شوند نشان می‌دهد.

x	$p_۴(x)$	خطا	$f_۴(x)$	خطا
۰٫۰	۰٫۰	۰٫۰	۰٫۲۰۰	$-۲٫۰\text{E}-۲$
۰٫۱	۰٫۲۵۷۳	۰٫۰۵۱۷	۰٫۳۱۴۳	$-۵٫۳\text{E}-۳$
۰٫۲	۰٫۴۶۱۳	۰٫۱۲۶۵	۰٫۵۸۸۷	$-۹٫۶\text{E}-۴$
۰٫۳	۰٫۶۰۹۱	۰٫۱۹۹۹	۰٫۸۰۹۱	$-۸٫۵\text{E}-۵$
۰٫۴	۰٫۶۹۸۶	۰٫۲۵۰۵	۰٫۹۵۱۱	$-۱٫۳\text{E}-۶$
۰٫۵	۰٫۷۲۸۶	۰٫۲۷۱۴	۱٫۰۰۰۰	۰٫۰

۳. (ب) فرض کنید

$$S=\sum_{۱}^{\infty}\frac{(-۱)^j x^{۲j}}{j^۲}$$

بازهٔ همگرایی این سری $|x|\le ۱$. برای خطا

$$\left|S-\sum_{j=۱}^{n}(-۱)^j\cdot\frac{x^{۲j}}{j^۲}\right|\le\frac{|x|^{۲n+۲}}{(n+۱)^۲}$$

اگر $n\ge ۳۱۶$، خطا برای همهٔ مقادیر $|x|\le ۱$ از $۱۰^{-۵}$ کمتر است. اما برای $|x|<۱$، احتمالاً بسیار بزرگ است، زیرا عامل $|x|^{۲n+۲}$ برای تخمین خطا به حساب نیامده است.

۶. تقریب پاد[1] عبارت است از $R(x)=(۱+\frac{۱}{۲}x)/(۱-\frac{۱}{۲}x)$ در خطای

$$e^x-R(x)=-[\frac{۱}{۱۲}x^۳+\frac{۱}{۱۲}x^۴+\frac{۱۳}{۲۴۰}x^۵+\dots]$$

و خطای سری تیلر $e^x-p_۲(x)=\frac{۱}{۶}x^۳+\frac{۱}{۲۴}x^۴+\dots$. برای مقدار کوچک $|x|$ خطای $R(x)$ در حدود نصف خطای $p_۲(x)$است. اما برای $|x|=۱$ خطای $R(x)$ بزرگتر است.

۱۰. (الف) $p_۱(x)=\ln\frac{۳}{۲}-۱+\frac{۲}{۳}x;\ \|\ln(x)-p_۱\|_\infty=$ ۰٫۰۷۲

(ب) $q_۱^*(x)=-۰٫۶۶۳۳+۰٫۶۹۳۱x;\ \|\ln(x)-q_۱^*\|_\infty=$ ۰٫۰۳۰

۱۳. مقدار مینیمم‌ساز α عبارت است از $\alpha=\sqrt{e}$. مقدار مینیمم برابر است با $e+۱-۲\sqrt{e}\doteq$ ۰٫۴۲.

۲۰. (الف) $\psi_۰(x)\equiv ۱$، $\psi_۱(x)=x-\frac{۱}{۴}$، $\psi_۲(x)=x^۲-\frac{۵}{۷}x+\frac{۱۷}{۲۵۲}$

۲۸. (الف) $\rho_n(\sin x)\le\dfrac{\pi^{n+۱}}{(n+۱)!۲^{۳n+۲}}$

۳۱. (ج) $I_۲(x)=$ ۰٫۹۶۸۷۰۶x−۰٫۱۸۷۱۳۰$x^۳$; $\|\tan^{-۱}x-I_۲\|_\infty=$ ۰٫۰۰۵۹۰

1. pade

(د)

n	$\| f - I_n \|_\infty$	n	$\| f - I_n \|_\infty$
۱	۳٫۷۲E − ۲	۵	۱٫۱۴E − ۵
۲	۴٫۳۷E − ۳	۶	۱٫۶۹E − ۶
۳	۵٫۷۲E − ۴	۷	۲٫۵۵E − ۷
۴	۷٫۹۴E − ۵	۸	۳٫۹۱E − ۸

۳۲. (ج)

$$F_۲(x) = ۰٫۹۷۲۴۲۰x - ۰٫۱۹۱۸۹۸x^۳; \| \tan^{-۱} x - F_۲ \|_\infty = ۰٫۰۰۵۰۰$$

۳۶. $$q^*_{n-۱}(x) = p(x) - \frac{a_n}{۲^{n-۱}} T_n(x), \rho_{n-۱}(p) = \frac{|a_n|}{۲^{n-۱}}$$

۴۰. (د)

n	$\rho_n(f)$	n	$\rho_n(f)$
۱	۲٫۹۸E − ۲	۵	۸٫۶۹E − ۶
۲	۳٫۴۲E − ۳	۶	۱٫۲۸E − ۶
۳	۴٫۴۲E − ۴	۷	۱٫۹۲E − ۷
۴	۶٫۰۷E − ۵	۸	۲٫۹۳E − ۸

فصل ۵

۱.

$\int_۰^۱ e^{-x^۲} dx$ (الف)			$\int_{-۴}^{۴} \frac{dx}{۱+x^۲}$ (ج)		
n	I_n	R_n	n	I_n	R_n
۲	۰٫۷۳۱۳۷۰۲۵		۲	۴٫۲۳۵۲۹۴	
۴	۰٫۷۴۲۹۸۴۱۰		۴	۲٫۹۱۷۶۴۷	
۸	۰٫۷۴۸۶۶۵۶۱	۴٫۰۳	۸	۲٫۶۵۸۸۲۴	۵٫۰۹
۱۶	۰٫۷۴۶۵۸۴۶۰	۴٫۰۱	۱۶	۲٫۶۵۰۵۰۷	۳۱٫۱
۳۲	۰٫۷۴۶۷۶۴۲۵	۴٫۰۰	۳۲	۲٫۶۵۱۳۴۷	−۹٫۹۰
۶۴	۰٫۷۴۶۸۰۹۱۶	۴٫۰۰	۶۴	۲٫۶۵۱۵۶۳	۳٫۸۹
۱۲۸	۰٫۷۴۶۸۲۰۳۹	۴٫۰۰	۱۲۸	۲٫۶۵۱۶۱۷	۴٫۰۰
۲۵۶	۰٫۷۴۶۸۲۳۲۰	۴٫۰۰	۲۵۶	۲٫۶۵۱۶۳۱	۴٫۰۰

۲.

(الف) $\int_0^1 e^{-x^2}dx$			(ج) $\int_{-۴}^{۴} \frac{dx}{۱+x^۲}$		
n	I_n	R_n	n	I_n	R_n
۲	۰٫۷۴۷۱۸۰۴۲۸۹۰۹۵		۲	۵٫۴۹۰۱۹۶۰۷۸۴	
۴	۰٫۷۴۶۸۵۵۳۷۹۷۹۱۰		۴	۲٫۴۷۸۴۳۱۳۷۲۵	
۸	۰٫۷۴۶۸۲۶۱۲۰۵۲۷۵	۱۱٫۱	۸	۲٫۵۷۲۵۴۹۰۱۹۶	−۳۲٫۰
۱۶	۰٫۷۴۶۸۲۴۲۵۷۴۳۵۷	۱۵٫۷	۱۶	۲٫۶۴۷۷۳۴۵۶۳۵	۱٫۲۵
۳۲	۰٫۷۴۶۸۲۴۱۴۰۶۰۷۰	۱۵٫۹	۳۲	۲٫۶۵۱۶۲۷۲۸۳۰	۱۹٫۳
۶۴	۰٫۷۴۶۸۲۴۱۳۳۲۹۹۷	۱۶٫۰	۶۴	۲٫۶۵۱۶۳۵۲۸۰۷	۴۸۷
۱۲۸	۰٫۷۴۶۸۲۴۱۳۲۸۴۲۹	۱۶٫۰	۱۲۸	۲٫۶۵۱۶۳۵۳۲۴۴	۱۸۳
۲۵۶	۰٫۷۴۶۸۲۴۱۳۸۱۴۳	۱۶٫۰	۲۵۶	۲٫۶۵۱۶۳۵۳۲۷۲	۱۶٫۰

۶. با افزایش n، نسبتهای R_n به مقادیر ثابت میل می‌کند. مقادیر برای $n = ۱۶$ به صورت زیرند:

α	۰٫۲۵	۰٫۵	۰٫۷۵	۱٫۰
R_n	۲٫۱۱	۲٫۵۲	۳٫۰۳	۳٫۶۴

۹. $I_h(f) = \frac{۳}{۴}h[f(۰) + ۳f(۲h)]$

۱۵.

n	I_n	خطا
۲	۱٫۲۶۳۱۶	۱٫۳۸۸
۴	۲٫۰۴۷۲۹	۰٫۶۰۴۴
۶	۲٫۴۱۱۶۹	۰٫۲۳۹۹
۸	۲٫۵۶۰۰۸	۰٫۰۹۱۵۶
۱۰	۲٫۶۱۷۲۵	۰٫۰۳۴۳۸

۱۷. $\int_0^1 f(x)\ln\left(\frac{۱}{x}\right)dx \doteq w_۱f(x_۱) + w_۲f(x_۲)$

$$x_۱ = \frac{۱۵ - \sqrt{۱۰۶}}{۴۲} \doteq ۰٫۱۱۲۰۰۸۸۰۶۲ \qquad x_۲ = \frac{۱۵ + \sqrt{۱۰۶}}{۴۲} \doteq ۰٫۶۰۲۲۷۶۹۰۸۱$$

$$w_۱ = \frac{۲۱}{\sqrt{۱۰۶}}\left[x_۲ - \frac{۱}{۴}\right] \doteq ۰٫۷۱۸۵۳۹۳۱۹ \qquad w_۲ = ۱ - w_۱ = ۰٫۲۸۱۴۶۰۶۸۱$$

۲۰. برای انتگرالگیری گاوسی، $|I-I_n| \le (۰٫۰۸۰/n^۴) \|f^{(۴)}\|_\infty$. برای قاعدهٔ سیمپسون بر بازهٔ $[-۱,۱]$، $|I-I_n| \le (۰٫۱۸/n^۴) \|f^{(۴)}\|_\infty$

۲۴. $\sum_{۱}^{\infty} \frac{۱}{n^{۵/۲}} = ۴٫۵۹۵۱۱۲۵۴ + E \quad ۰ < -E < ۱٫۳۴ \times ۱۰^{-۶}$

۲۶. با فرض $I - I_n \doteq c/n^p$ و با استفاده از $I_{۱۶}$ و $I_{۳۲}$ و $I_{۶۴}$ به دست می‌آوریم $p \doteq ۳٫۴۴$، $c \doteq ۰٫۰۱۵۴$. برای برونیاب ایتکن، $I_{۶۴} = ۰٫۲۸۵۷۱۴۲۸۵۸۶$، و $I - I_n| \le ۱۰^{-۱۱}$ به $۹٫۴۳ \times ۱۰^{-۹} = I_{۶۴} - \tilde{I}_{۶۴} \doteq I - I_{۶۴}$. برای اینکه داشته باشیم $n \ge ۴۶۹$ احتیاج خواهیم داشت. انتخاب منطقی استفاده از $n = ۵۱۲$ است.

۲۸. $\tilde{I}_n = \frac{۱}{۳}[I_n^{(T)} + ۲I_n^{(M)}]$ که قاعدهٔ سیمپسون خواهد بود.

۳۸. (الف) انتگرال $I = \left(\int_۰^۱ + \int_۱^{۴n}\right) \cos(x)\ln(x)dx$ را بشکنید. با استفاده از سری تیلر

$$\int_۰^۱ \cos(x)\ln(x)dx \doteq -\left(۱ - \frac{۱}{۲!۹} + \frac{۱}{۴!۲۵} - \frac{۱}{۶!۴۹} + \frac{۱}{۸!۸۱}\right)$$
$$\doteq -۰٫۹۴۶۰۸۳۰۷۲۷$$

که در آن خطا کمتر از $۲٫۲۸ \times ۱۰^{-۹}$ است. برای محاسبهٔ انتگرال باقی‌مانده می‌توان از یک روش استانده استفاده کرد. به‌عنوان مثال، انتگرالگیری رامبرگ با $n = ۱۲۹$ گره به مقدار $-۰٫۵۴۶۰۷۸۱۵۸۰$ منجر می‌شود. که در آن خطا کمتر از $۲٫۸ \times ۱۰^{-۹}$ است. بنابراین انتگرال برابر است با $I = ۱٫۴۹۲۱۶۱۲۳۰۷$ که در آن خطا از $۵٫۱ \times ۱۰^{-۹}$ کمتر است.

(ب) از آنجایی که $|x^۲ \sin(\frac{۱}{x})| \le x^۲$، $\varepsilon > ۰$ را چنان انتخاب کنید که

$$\left|\int_۰^\varepsilon x^۲ \sin\left(\frac{۱}{x}\right)dx\right| \le \int_۰^\varepsilon x^۲dx \le ۰٫۰۰۰۵$$

یک انتخاب مناسب $\varepsilon = ۰٫۱$. سپس $\int_{۰٫۱}^{۲/\pi} (x^۲)\sin(۱/x)dx$ را با دقت حداقل ۰٫۰۰۰۵ محاسبه کنید.

۴۱. (الف) $$D_h^{(۲)} f(x) = \frac{f(x) - ۲f(x+h) + f(x+۲h)}{h^۲}$$
$$f''(x) - D_h^{(۲)} f(x) \doteq hf^{(۳)}(x)$$

فصل ۶

۲. (الف) ۲ (ب) ۱ (ج) ۰ (د) ۲۰

۳. (ب)

$$y_۱(۰) = ۱ \quad y_۱' = y_۲$$
$$y_۲(۰) = ۰ \quad y_۲' = ۰٫۱(۱ - y_۱^۲)y_۲ - y_۱$$

۵. (ج) جواب واقعی $Y(x) = ۲e^x + \sin(x) - \cos(x)$. بارستهای پیکار عبارتند از:

$$Y_۰(x) = ۱,\ Y_۱(x) = ۱ + x + ۲\sin(x)$$

$$Y_۲(x) = ۳ + x + \frac{x^۲}{۲} + ۲\sin(x) - ۲\cos(x)$$

$$Y_۳(x) = ۳ + ۳x + \frac{۱}{۲}x^۲ + \frac{۱}{۶}x^۳ - ۲\cos(x)$$

برای $x \doteq ۰$ آنها را با استفاده از بسط سری تیلر و $Y(x)$ با هم مقایسه کنید.

۸. فرض کنید خطای اولیه $e_۰ = ۰$. برای کران از (۱۳.۲.۶)

$$\underset{۰\le x\le b}{\text{Max}} \mid Y(x_n) - y_n \mid \le \frac{h}{۲}(e^{۲b} - ۱)$$

خطای مجانبی به وسیلهٔ فرمول زیر داده شده است

$$Y(x) - y_h(x) = h\frac{\ln(۱+x)}{(۱+x)^۲} + O(h^۲) \qquad x \ge ۰$$

این عبارت نشان می‌دهد که با افزایش x خطا کاهش می‌یابد، در حالی که کران قبلی پیش‌بینی می‌کند که وقتی بازهٔ $[۰, b]$ افزایش می‌یابد خطا نیز افزایش می‌یابد.

۱۰. (ب) $y_{۱,n+۱} = y_{۱,n} + hy_{۲,n}$

$$y_{۲,n+۱} = y_{۲,n} + h[۰.۱(۱ - y_{۱,n}^۲)y_{۲,n} - y_{۱,n}] \qquad n \ge ۰$$

۱۴. $T_n(Y) = -\frac{۵}{۸}h^۲Y^{(۴)}(x_n) + O(h^۴)$

۲۶. برای $c = \frac{۱}{۴}$، $u_n = c_۱ + c_۲(\frac{۱}{۲})^n + c_۳(-\frac{۱}{۲})^n$ با $c_۱$، $c_۲$، و $c_۳$ دلخواه. هنگامی که $n \to \infty$، $u_n \to c_۱$، که $c_۱$ یک عدد ثابت است.

۳۱. (الف) $۰ \le a_۰ < ۲$ (ب) $a_۰ = ۰$

(ج) در $a_۰ = ۰$ هیچ ناحیه‌ای برای پایداری مطلق وجود ندارد. هنگامی که $a_۰$ از ۰ به ۲ می‌رود ناحیهٔ پایداری مطلق در حد به $-۲ < h\lambda < ۰$ افزایش می‌یابد. این رابطه فقط برای مقادیر حقیقی در ناحیهٔ پایداری مطلق برقرار است.

۳۳. (الف) $a_۰ = -۹ - ۳b_۲ \qquad a_۱ = ۹ \qquad a_۲ = ۱ + ۳b_۲$

$$b_۰ = ۶ + b_۲ \qquad b_۱ = ۶ + ۴b_۲$$

(ب) معادلهٔ مشخصه برای $h = ۰$ چنین است

$$\rho(r) \equiv r^۳ - a_۰r^۲ - ۹r + a_۰ + ۸ = ۰$$

با تقسیم بر $(r-1)$ دو ریشهٔ باقیمانده در معادلهٔ زیر صادق‌اند.

$$r^2 + (1 - a_0)r - (a_0 + 8) = 0 \qquad (*)$$

به ازای همهٔ مقادیر حقیقی a_0، عبارت مبیّن

$$(1 - a_0)^2 + 4(a_0 + 8) = (1 + a_0)^2 + 32 \geq 32 > 0$$

مثبت است. بنابراین برای هر مقدار a_0 هر دو ریشهٔ r_1 و r_2ی معادلهٔ (*)حقیقی‌اند. یک امتحان مستقیم این ریشه‌ها نشان خواهد داد که این ریشه‌هاهمزمان به ازای یک مقدار a_0 در $-1 < r < 1$ صدق نمی‌کنند.

۳۶. $y_{n+1} = y_{n-1} + 2hf(x_{n-1}, y_{n-1})$ از مرتبهٔ اول است و پایدار نسبی است اما در شرط ریشهٔ قوی صدق نمی‌کند.

۴۰. از $Y(x) = \frac{1}{4}x^2$ استفاده کنید بنابراین $Y''(x) = 1$. در این صورت خطای $e_n = Y(x_n) - y_n$ در رابطهٔ زیر صدق می‌کند.

$$e_n = \frac{h}{2\lambda}[(1 + h\lambda)^n - 1] \qquad n \geq 0$$

۴۸.

$$\gamma_1 + \gamma_2 + \gamma_3 = 1$$
$$\gamma_2\alpha_2 + \gamma_3\alpha_3 = \frac{1}{2}$$
$$\gamma_1\beta_{21} + \gamma_3(\beta_{31} + \beta_{32}) = \frac{1}{2}$$
$$\gamma_2\alpha_2^2 + \gamma_3\alpha_3^2 = \frac{1}{3}$$
$$\gamma_2\alpha_2\beta_{21} + \gamma_3\alpha_3(\beta_{31} + \beta_{32}) = \frac{1}{3}$$
$$\gamma_2\beta_{21}^2 + \gamma_3(\beta_{31} + \beta_{32})^2 = \frac{1}{3}$$
$$\gamma_3\alpha_3\beta_{32} = \frac{1}{6}$$
$$\gamma_3\beta_{32}\beta_{21} = \frac{1}{6}$$

این معادلات وابسته‌اند و می‌توان آنها را به ۶ معادلهٔ مستقل بدل کرد. یک جواب خاص عبارت است از

$$Y_{n+1} = Y_n + \frac{h}{6}(V_1 + 4V_2 + V_3)$$
$$V_1 = f(x_n, y_n) \qquad V_2 = f(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hV_1)$$
$$V_3 = f[x_n + h, y_n + h(2V_2 - V_1)]$$

اگر $f(x,y)$ به y وابسته نباشد این قاعده همان قاعدهٔ سیمپسن است.

۵۱. (الف) ریشه‌های $h\lambda$ی $| 1 + h\lambda + \frac{1}{2}(h\lambda)^2 | < 1$. برای λ ی حقیقی، این بازه $-2 < h\lambda < 0$، است.

فصل ۷

۱. (الف) وابسته (ب) مستقل

۷. (الف) $(Ax, x) = x^T(Ax)$. اما همچنین

$$(Ax, x) = (x, Ax) = (Ax)^T x = x^T A^T x = -x^T Ax$$

از آنجایی که $x^T Ax = -x^T Ax$ ، باید داشته باشیم $x^T Ax = 0$ که نتیجهٔ مطلوب است.

۸. $u^{(3)} = (-7, 4, 1)$. برای نرمال کردن، هر بردار را بر طولش تقسیم می‌کنیم.

$$\| u^{(1)} \|_2 = \sqrt{6} \qquad \| u^{(2)} \|_2 = \sqrt{11} \qquad \| u^{(3)} \|_2 = \sqrt{66}$$

۹. (الف)
$$A = \frac{1}{9}\begin{bmatrix} 7 & -4 & -4 \\ -4 & 1 & -8 \\ -4 & -8 & 1 \end{bmatrix}$$

۱۱. (الف) $\lambda_1 = -1$ ، $x = [2, -1]^T$

$\lambda_2 = 3$ ، $x = [2, 1]^T$

۱۳. (الف) $\| Ux \|_2^2 = (Ux, Ux) = (x, U^*Ux) = (x, Ix) = (x, x) = \| x \|_2^2$، نشان می‌دهد که برای هر x، $\| Ux \|_2 = \| x \|_2$. برای فاصلهٔ بین Ux و Uy ، از نتیجهٔ قبل،

$$\| Ux - Uy \|_2 = \| U(x - y) \|_2 = \| x - y \|_2$$

۱۴. چون A ارمیتی است، فرض کنید $x^{(n)},\dots,x^{(1)}$ یک پایهٔ یکامتعامد برای C^n مربوط به ویژه مقدارهای حقیقی λ_1 ، ... ، λ_n باشد. برای هر x در C^n می‌توانیم بنویسیم

$$x=\sum_1^n \alpha_i x^{(i)} \qquad \text{با} \quad \alpha_i=(x,x^{(i)}) \qquad i=1,\dots,n$$

از یکا متعامد بودن داریم $\| x \|_2^2=\sum_1^n |\alpha_i|^2$. با استفاده از این شکل برای x نشان دهید

$$(Ax,x)=\sum_1^n \lambda_i |\alpha_i|^2$$

چون α_1 ، ... ، α_n برای بدست آوردن عناصر مختلف x از C^n می‌توانند به طور دلخواه تغییر کنند این فرمول ثابت می‌کند که A معین مثبت است اگر و تنها اگر همهٔ λ_i ها مثبت باشند.

۲۳. ابتدا ثابت کنید

$$\| x \|_\infty \le \| x \|_p \le n^{1/p} \| x \|_\infty \qquad \text{به ازای همهٔ مقادیر} \quad x\in C^n$$

از این رابطه به آسانی دیده می‌شود که هرگاه $p\to\infty$ ، $\| x \|_p \to \| x \|_\infty$.

۲۷. (الف) فرض کنید λ و x یک جفت ویژه مقدار ـ ویژه بردار ماتریس A باشند. آنگاه جفت ویژه‌های مربوط به تغییرات مختلف A عبارت‌اند از: (۱) λ^m و x برای A^m ؛ (۲) $1/\lambda$ و x برای A^{-1} ؛ و (۳) $\lambda+c$ و x برای $A+cI$.

۲۸. با استفاده از ستونهای A_{*j}ی ماتریس A ، بنویسید $A=[A_{*1},\dots,A_{*n}]$. بنابراین

$$F(A)=\sqrt{\| A_{*1} \|_2^2+\cdots+\| A_{*n} \|_2^2}$$

(الف) $F(UA)=\sqrt{\| UA_{*1} \|_2^2+\cdots+\| UA_{*n} \|_2^2}$. با استفاده از مسألهٔ ۱۳ (الف) برای جمیع مقادیر j ، $\| UA_{*j} \|_2=\| A_{*j} \|_2$. بنابراین $F(UA)=F(A)$

(ب) فرض کنید $U^*AU=D=\operatorname{diag}[\lambda_1,\dots,\lambda_n]$. برای به دست آوردن

$$F(A)=F(U^*AU)=F(D)=\sqrt{\lambda_1^2+\cdots+\lambda_n^2}$$

قسمت (الف) را بکار برید.

۳۲. $\| A \|_\infty \le 10, \| A^{-1} \|_\infty < \frac{1}{2}$

فصل ۸

۱. (الف)

$$L=\begin{bmatrix}۱ & ۰ & ۰\\ ۱ & ۱ & ۰\\ -۲ & ۳ & ۱\end{bmatrix}\quad U=\begin{bmatrix}۱ & ۱ & -۱\\ ۰ & ۱ & -۱\\ ۰ & ۰ & ۲\end{bmatrix}\quad x=\begin{bmatrix}۲\\ ۲\\ ۳\end{bmatrix}$$

۲. (الف) بدون محورگیری، ماتریس افزودهٔ $[A \mid b]$ به

$$\left[\begin{array}{ccc|c} ۶٫۰۰۰ & ۲٫۰۰۰ & ۲٫۰۰۰ & -۲٫۰۰۰ \\ ۰ & ۰٫۰۰۰۱۰۰۰ & -۰٫۳۳۳۳ & ۱٫۶۶۷ \\ ۰ & ۰ & ۵۵۵۵ & -۲۷۷۹۰ \end{array}\right]\begin{array}{l} m_{۲۱}=۰٫۳۳۳۳ \\ m_{۳۱}=۰٫۱۶۶۷ \\ m_{۳۲}=۱۶۶۷۰ \end{array}$$

بدل می‌شود. جواب با استفاده از جایگذاری پسرو عبارت است از

$$x_۱=۱٫۳۳۵ \qquad x_۲=۰ \qquad x_۳=-۵٫۰۰۳$$

همهٔ اعمال حساب انجام شده تا ۴ رقم اعشاری با ممیز شناورگردشده‌اند. این برای قسمت (ب) نیز درست است.

(ب) با محورگیری $[A \mid b]$ به

$$\left[\begin{array}{ccc|c} ۶٫۰۰۰ & ۲٫۰۰۰ & ۲٫۰۰۰ & -۲٫۰۰۰ \\ ۰ & ۱٫۶۶۷ & -۱٫۳۳۳ & ۰٫۳۳۳۴ \\ ۰ & ۰ & ۰٫۳۳۳۲ & ۱٫۶۶۷ \end{array}\right]\begin{array}{l} m_{۲۱}=۰٫۳۳۳۳ \\ m_{۳۱}=۱٫۶۶۷ \\ m_{۳۲}=۰٫۰۰۰۰۵۹۹۹ \end{array}$$

بدل می‌شود. جواب با استفاده از جایگذاری پسرو عبارت است از

$$x_۱=۲٫۶۰۲ \qquad x_۲=-۳٫۸۰۱ \qquad x_۳=-۵٫۰۰۳$$

۵. (الف) در شکل افرازشده،

$$\begin{bmatrix}A_۱ & -A_۲\\ A_۲ & A_۱\end{bmatrix}\begin{bmatrix}x_۱\\ x_۲\end{bmatrix}=\begin{bmatrix}b_۱\\ b_۲\end{bmatrix}$$

(ب) فرض کنید دستگاه ۱ معرّف دستگاه حقیقی قسمت (الف) باشد، و فرض کنید دستگاه ۲ دستگاه مختلط اوّلیهٔ $Ax=b$ را نشان دهد. برای ذخیره‌سازی ماتریس، دستگاه ۱ به $۴n^۲$

حافظه نیاز دارد و دستگاه ۲ به $2n^2$ حافظه. (هر عدد مختلط به دو حافظهٔ ذخیره‌سازی نیاز دارد). برای حل، دستگاه ۱ تقریباً به $\frac{1}{3}(2n)^3 = \frac{8}{3}n^3$ عملیات ضرب و تقسیم نیاز دارد. دستگاه ۲؛ به $\frac{1}{3}n^3$ عملیات ضرب و تقسیم مختلط نیاز دارد. چون هر ضرب مختلط به ۴ ضرب حقیقی نیاز دارد تعداد عملیات واقعی $\frac{4}{3}n^3$ است. بنابراین دستگاه ۲ به نصف حافظهٔ ذخیره‌سازی و در حدود نصف زمان عملیات نسبت به دستگاه ۱ نیاز دارد.

۹. چون L^T ناتکین است $\| L^T x \|_2^2 > 0$ برای هر $x \neq 0$. همچنین $(Ax, x) = (LL^T, x) = (L^T x, L^T x) =$ $\det(A) = \det(L)^2$.

۱۰. (الف)

$$L = \begin{bmatrix} ۱٫۵ & ۰ & ۰ \\ -۲٫۰ & ۱٫۰ & ۰ \\ ۳٫۰ & -۴٫۰ & ۳٫۰ \end{bmatrix}$$

۱۴.

$$L = \begin{bmatrix} ۲ & ۰ & ۰ & ۰ & ۰ \\ ۱ & \frac{۵}{۲} & ۰ & ۰ & ۰ \\ ۰ & ۱ & \frac{۱۲}{۵} & ۰ & ۰ \\ ۰ & ۰ & ۱ & \frac{۲۹}{۱۲} & ۰ \\ ۰ & ۰ & ۰ & ۱ & \frac{۷۰}{۲۹} \end{bmatrix} \qquad U = \begin{bmatrix} ۱ & -\frac{۱}{۲} & ۰ & ۰ & ۰ \\ ۰ & ۱ & -\frac{۲}{۵} & ۰ & ۰ \\ ۰ & ۰ & ۱ & -\frac{۵}{۱۲} & ۰ \\ ۰ & ۰ & ۰ & ۱ & -\frac{۱۲}{۲۹} \\ ۰ & ۰ & ۰ & ۰ & ۱ \end{bmatrix}$$

$Lz = b$ دارای جواب $z = [\frac{۳}{۲}, \frac{-۷}{۵}, \frac{۱۷}{۱۲}, \frac{-۲۱}{۲۹}, ۱]^T$ است

$Ux = z$ دارای جواب $x = [۱, -۱, ۱, -۱, ۱]^T$ است.

۱۸. $\text{cond}(A)_۱ = \text{cond}(A)_\infty = ۳۹۶۰۱, \text{cond}(A)_۲ = ۳۹۲۰۶$

۲۶. برای نرخهای همگرایی μ از (۸-۶-۵) و η از (۸-۶-۱۷)، در این حالت داریم: $\mu = \frac{۲}{۴}$ و $\eta = \frac{۲}{۳}$. نرخهای واقعی کاهش مشاهده شده، ۶۱ر۰ برای روش گاوس ـ ژاکوبی، و ۳۷ر۰ برای روش گاوس ـ زایدل هستند.

۲۷. (الف) همگراست اگر و تنها اگر $r_\sigma(A^{-1}B) < ۱$، و نرخ همگرایی اساساً $r_\sigma(A^{-1}B)$ یا $\| A^{-1}B \|$ است بسته به نرمی که استفاده‌شده باشد.

(ب) باز هم همگراست اگر و تنها اگر $r_\sigma(A^{-1}B) < ۱$. اما نرخ همگرایی در حدود $r_\sigma(A^{-1}B)^2$ یا $\| A^{-1}B \|^2$ است که سریعتر از قسمت (الف) است.

۳۲. (الف) $A^{-1} - C_m = A^{-1}R_m = A^{-1}R_0^{2^m}$ ، $m \geq 0$. جملهٔ C_m به A^{-1} همگراست اگر و تنها اگر $r_\sigma(R_0) < 1$.

(ب) با نوشتن $(I - R_0)^{-1}$ به صورت یک حاصلضرب نامتناهی،

$$A^{-1} = C_0(I + R_0)(I + R_0^2)(I + R_0^4)(I + R_0^8)\dots$$
$$C_m = C_0(I + R_0)(I + R_0^2)(I + R_0^4)\dots(I + R_0^{2^{m-1}})$$

فصل ۹

۱. (الف) با به‌کار بردن قضیهٔ گرشگورین برای سطرهای ماتریس A دایره‌های زیر به دست می‌آیند $|\lambda - 1| \leq 1$ ، $|\lambda - 5| \leq 2$ ، و $|\lambda - 9| \leq 3$. دایره‌های دوم و سوم متقاطع‌اند و بنابراین باید شامل دو ویژه مقدار در قسمت اشتراکشان باشند. اولین دایره متمایز از دو دایرهٔ دیگر است. چون چندجمله‌یی مشخّصهٔ A ضرایب حقیقی دارد،ویژه مقدارها اگر مختلط باشند، باید مزدوج همدیگر باشند. بنابراین $|\lambda - 1| \leq 1$ فقط شامل یک ویژه مقدار حقیقی دربازهٔ $[0, 2]$ است.

با به کار بردن قضیه برای ستونهای A دایره‌ها عبارت‌اند از $|\lambda - 1| \leq 3$ ، $|\lambda - 5| \leq 2$ ، و $|\lambda - 9| \leq 1$. با استدلالهای مشابه و استفاده از نتایج قبلی، ویژه مقدارها در هر یک از بازه‌های $[0, 2]$ ، $[3, 7]$ و $[8, 10]$ قرار دارند. ویژه مقدارهای واقعی عبارت‌اند از:

۱٫۳۳۱۹۲۷۶۸۹ ۴٫۸۵۶۹۳۳۶۹۲ ۸٫۸۱۱۱۳۸۶۱۶

۲. (ج) n دایره عبارت‌اند از:

$$|r| \leq |a_0| \quad |r| \leq 1 + |a_j| \quad j = 1, \dots, n-2 \quad \text{و} \quad |r + a_{n-1}| \leq 1$$

د (i)، ۹ ریشه وجود دارند که در $|r - 1| \leq 1$ صدق می‌کنند و یک ریشهٔ حقیقی در $-9 \leq r \leq -7$ صدق می‌کند.

۴. $\min_{1 \leq i \leq 4} |\lambda - \lambda_i| \leq 0.0000356$

۸. (الف) $\lambda = 15$ ، $x = [1, 1, 1, 1]^T$. در هر بارست خطا با ضریب $\frac{1}{2}$ کاهش می‌یابد. بنابراین $\lambda = 5$ احتمالاً بزرگترین ویژه مقدار دوم است.

۱۳. (ب) با استفاده از $w = [1/\sqrt{6}, -1/\sqrt{6}, -2/\sqrt{6}]^T$ برای ساختن

$P = I - ۲ww^T$ ، به دست می‌آوریم $Px = e_۱$. بنابراین

$$B = PAP^T = \begin{bmatrix} ۹ & ۰ & ۰ \\ ۰ & ۱۸ & ۰ \\ ۰ & ۰ & -۹ \end{bmatrix}$$

در این حالت ماتریس B قطری است اما معمولاً بدین سادگی نخواهد بود. با استفاده از $w = (۱/\sqrt{۳۰})[۵, ۱, ۲]^T$ به یک P برای هر $Px = -e_۱$ می‌رسیم. اما $B = PAP^T$ باز همان شکل را برای سطر و ستون ۱ خواهد داشت. در این حالت خاص

$$B = \begin{bmatrix} ۹ & ۰ & ۰ \\ ۰ & \frac{۱۸}{۲۵} & -\frac{۳۲۴}{۲۵} \\ ۰ & -\frac{۳۲۵}{۲۵} & \frac{۲۰۷}{۲۵} \end{bmatrix}$$

اولین شکل P با علامت مخالف با (۹-۳-۹) ساخته شد. در حالی که دومین انتخاب P براساس (۹-۳-۹) ساخته می‌شود. در حالت عادی شکل P که با (۹ـ۳ـ۹) مطابقت می‌کند به دلیل دقّت در ملاحظاتی که پس از (۹ـ۳ـ۹) آورده‌ایم، شکل ارجح خواهد بود.

۱۴.

$$Q = \frac{۱}{۳}\begin{bmatrix} -۱ & ۲ & ۲ \\ -۲ & ۱ & -۲ \\ -۲ & -۲ & ۱ \end{bmatrix} \qquad R = ۳\begin{bmatrix} -۱ & ۱ & -۱ \\ ۰ & ۱ & -۱ \\ ۰ & ۰ & ۱ \end{bmatrix}$$

۱۵. (الف) $\alpha = \frac{b_k}{\gamma}$ ، $\beta = \frac{b_l}{\gamma}$

(ب)

$$R^{(۱,۳)} = \begin{bmatrix} \frac{\sqrt{۳}}{۲} & ۰ & \frac{۱}{۲} & ۰ \\ ۰ & ۱ & ۰ & ۰ \\ -\frac{۱}{۲} & ۰ & \frac{\sqrt{۳}}{۲} & ۰ \\ ۰ & ۰ & ۰ & ۱ \end{bmatrix} \qquad R^{(۱,۴)} = \begin{bmatrix} \frac{۱}{\sqrt{۲}} & ۰ & ۰ & \frac{۱}{\sqrt{۲}} \\ ۰ & ۱ & ۰ & ۰ \\ ۰ & ۰ & ۱ & ۰ \\ -\frac{۱}{\sqrt{۲}} & ۰ & ۰ & \frac{۱}{\sqrt{۲}} \end{bmatrix}$$

$$\hat{b} = Ub = [۲, ۰, ۰, ۰]^T \qquad R^{(۱,۲)} = \begin{bmatrix} \frac{\sqrt{۳}}{۲} & \frac{۱}{۲} & ۰ & ۰ \\ -\left(\frac{۱}{۲}\right) & \frac{\sqrt{۳}}{۲} & ۰ & ۰ \\ ۰ & ۰ & ۱ & ۰ \\ ۰ & ۰ & ۰ & ۱ \end{bmatrix}$$

$$U = R^{(1,2)}R^{(1,3)}R^{(1,4)} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{-1}{2\sqrt{3}} & \frac{\sqrt{3}}{2} & \frac{-1}{2\sqrt{3}} & \frac{-1}{2\sqrt{3}} \\ \frac{-1}{\sqrt{6}} & 0 & \frac{\sqrt{2}}{\sqrt{3}} & \frac{-1}{\sqrt{6}} \\ \frac{-1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

۱۹. (الف) $f_0(\lambda) = 1$ ، $f_1(\lambda) = -\lambda$ ، $f_j(\lambda) = (1-\lambda)f_{j-1}(\lambda) - f_{j-2}(\lambda)$ به ازای $j = 2, 3, 4$ ، $f_5(\lambda) = (2-\lambda)f_4(\lambda) - f_3(\lambda)$. فرض کنید $\lambda_5 < \lambda_4 < \lambda_3 < \lambda_2 < \lambda_1$. به موجب قضیهٔ گرشگورین همهٔ λ_iها در بازهٔ $[-1, 3]$ قرار دارند. بنابراین همان گونه که در جدول زیر نشان داده شده است، می‌توان ریشه‌ها را با استفاده از قضیهٔ ۹ـ۵ جدا کرد. ریشه‌های واقعی را می‌توان توسط روش دیگری پیدا کرد. به عنوان مثال روش خط قاطع، سریعاً به جواب $\lambda_5 \doteq 2{,}902113003$ می‌رسد.

λ	$f_5(\lambda)$	$s(\lambda)$	ملاحظات
$-1{,}0$	$2{,}0$	5	$\lambda_1 > -1$
$3{,}0$	$-2{,}0$	0	$\lambda_5 < 3$
$1{,}0$	$0{,}0$	2	$\lambda_3 = 1{,}0, \lambda_4 > 1$
$0{,}0$	$1{,}0$	3	$\lambda_2 < 0{,}0$
$-0{,}5$	$-1{,}78$	4	$-1 < \lambda_1 < 0{,}5 < \lambda_2 < 0$
$2{,}0$	$-1{,}0$	2	$\lambda_4 > 2$
$2{,}5$	$1{,}78$	1	$2 < \lambda_4 < 2{,}5 < \lambda_5 < 3$

۲۰. (ب) (۲) با استفاده از تبدیلات هاوسهولدر مانند (۲۱.۳.۹) تا (۲۳.۳.۹) به دست می‌آوریم $T = Q^T A Q$

$$T = \begin{bmatrix} 5{,}0 & -4{,}242640687\,1 & 0{,}0 & 0{,}0 \\ -4{,}242640687\,1 & 6{,}0 & 1{,}414213562\,4 & 0{,}0 \\ 0{,}0 & 1{,}414213562\,4 & 5{,}0 & \\ 0{,}0 & 0{,}0 & 0{,}0 & 2{,}0 \end{bmatrix}$$

$$Q = \begin{bmatrix} 1{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}0 & -0{,}942809041\,6 & 0{,}333333333\,3 & 0{,}0 \\ 0{,}0 & -0{,}235702260\,4 & -0{,}666666666\,7 & -0{,}707106781\,2 \\ 0{,}0 & -0{,}235702260\,4 & -0{,}666666666\,7 & 0{,}707106781\,2 \end{bmatrix}$$

در اینجا Q صریحاً محاسبه شده است، معمولاً در عمل ساختن Q به طور صریح اتلاف وقت خواهد بود.

۲۲. (۱) ویژه مقدارها عبارت‌اند از

$$\lambda_1 = 2 - \sqrt{3} \quad \lambda_2 = 2 \quad \lambda_3 = 2 + \sqrt{3}$$

(الف) برای روش QR بدون انتقال عناصری که خارج از قطر قرار دارند بطور خطی به صفر همگرا می‌شوند. در هر بارست عنصرهای موضع (۱,۲) با ضریب ۰٫۵۳۶ کاهش می‌یابند و عناصر موضع (۲,۳) با ضریب ۰٫۱۳۴- در هر بارست کاهش می‌یابند.

(ب) برای روش QR با انتقال، با استفاده از انتخاب

$$c_m = a_{3,3}^{(m)}$$

همگرایی سریع خواهیم داشت. بعد از پنج بارست اندازهٔ عنصر موضع (۲,۳) کمتر از 5×10^{-11} خواهد بود. با نشان دادن ماتریس اولیهٔ A با A_0 نتیجهٔ زیر را خواهیم داشت:

$$A_5 = \begin{bmatrix} 3{,}7316925974 & 0{,}0249060210 & 0{,}0 \\ 0{,}0249060210 & 2{,}0003582102 & \varepsilon \\ 0{,}0 & \varepsilon & 0{,}2679491924 \end{bmatrix}$$

با $|\varepsilon| < 5 \times 10^{-11}$. وقتی روش QR با انتقال، برای ماتریس تبدیل شدهٔ حاصل از حذف سطر و ستون سوم A_5 به کار رود، ماتریس بسیار سریع همگرا می‌شود.

۲۴. برای $\lambda = 7{,}9329$ ویژه بردار عبارت است از

$$x = [0{,}7211774817, 0{,}2724739430, 1{,}0, 0{,}2515508351]^T$$

برای $\lambda = 5{,}6689$،

$$x = [0{,}5736191640, 0{,}5489544105, -0{,}8148078321, 1{,}0]^T$$

۲۸. مقادیر به دست آمده در جدول زیر داده شده‌اند

n	$\lambda_1^{(n)}$	$\lambda_2^{(n)}$
۲	۰٫۹	۰٫۱
۴	۰٫۸۸۴۵۶۶	۰٫۱۰۳۹۵۷
۸	۰٫۸۸۰۹۷۱	۰٫۱۰۵۳۶۰
۱۶	۰٫۸۸۰۰۸۵	۰٫۱۰۵۷۱۵
۳۲	۰٫۸۷۹۸۶۵	۰٫۱۰۵۸۰۵

با بررسی نسبتهای تفاضلات پیاپی به طور تجربی می‌توان نشان داد که

$$\lambda_i - \lambda_i^{(n)} = O\left(\frac{1}{n^2}\right)$$

این نرخ همگرایی به صورت نظری قابل توجیه است. برای براورد خطاهای ایجاد شده برونیابی ریچاردسون را می‌توان به‌کار برد.

$$\lambda_1 - \lambda_1^{(32)} \doteq 7{,}3 \times 10^{-5} \qquad \lambda_2 - \lambda_2^{(32)} \doteq 3{,}0 \times 10^{-5}$$

۲۷. (۱) A^+, A را با استفاده از شکل تجزیهٔ تکین ـ مقداری‌شان بیان کنید: $A = VFU^T$ و $A^+ = UF^+V^T$. بنابراین

$$\begin{aligned} AA^+A &= (VFU^T)(UF^+V^T)(VFU^T) \\ &= VFF^+FU^T \\ &= VFU^T \qquad : FF^+F = F \text{ چون با محاسبهٔ مستقیم خواهیم داشت} \\ &= A \end{aligned}$$

۳۰. $p(x) = -1{,}269091x + 0{,}392952$ ، خطای ریشه دوم میانگین برابر است با $0{,}243$

۳۳. (الف)

$$QR = \begin{bmatrix} -0{,}5179 & -0{,}7517 & 0{,}4082 \\ 0{,}5754 & 0{,}0470 & 0{,}8165 \\ -0{,}6330 & 0{,}6578 & 0{,}4082 \end{bmatrix} \begin{bmatrix} -1{,}7378 & -1{,}7148 \\ 0 & -0{,}2819 \\ 0 & 0 \end{bmatrix}$$

$$\text{SVD} = \begin{bmatrix} 0{,}5774 & 0{,}7071 & 0{,}4082 \\ -0{,}5774 & 0 & 0{,}8165 \\ 0{,}5774 & -0{,}7071 & 0{,}4082 \end{bmatrix} \begin{bmatrix} 2{,}4495 & 0 \\ 0 & 0{,}2000 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0{,}707 & 0{,}7071 \\ -0{,}7071 & 0{,}7071 \end{bmatrix}$$

$$A^+ = \begin{bmatrix} -\frac{7}{3} & -\frac{1}{6} & \frac{8}{3} \\ \frac{8}{3} & -\frac{1}{6} & -\frac{7}{3} \end{bmatrix} \qquad \text{cond}(A) = 12{,}2474$$

نمایه