به نام خداوند

بخشنده ی مهربان

مبانی
آنالیز عددی

زمستان ۱۳۹۴

فهرست مطالب

فصل ۱

آنالیز عددی؟

به موضوعی زیبا در ریاضیات کاربردی خوش آمدید؛ *آنالیز عددی*. از آنجا که این اولین برخورد شما با این موضوع است بهتر است پیش از هر چیز تعریفی از آن ارائه دهیم:

آنالیز عددی علم طراحی و تحلیل الگوریتم‌های حلِ تقریبیِ مسائلِ پیوسته در ریاضیات است.

بگذارید ابتدا در مورد "مسائل پیوسته" صحبت کنیم. لفظ پیوسته به این معناست که مسئله‌ی ما دارای متغیرهایی از جنس حقیقی یا مختلط است. اما این مسائل از کجا می‌آیند و معمولاً به چه صورت هستند؟ برای بررسی بسیاری از پدیده‌های طبیعی، غالباً قوانین حاکم بر آن‌ها را به زبان ریاضی فرمول‌بندی می‌کنیم که به آن مدل ریاضی می‌گوییم. مطالعه و پیش‌بینی رفتار یک پدیده با بررسی و حل مدل ریاضی آن صورت می‌گیرد.

برای مثال یک مدل ساده رشد جمعیت را در اینجا توضیح می‌دهیم. یک گونه خاص حیوان در یک محیط بسته را در نظر بگیرید که جمعیت اولیه آن‌ها برابر p_0 است. فرض کنیم در یک بازه زمانی متناهی هیچ مرگ و میری اتفاق نیفتد، هیچ ورود و خروجی در محیط صورت نگیرد، و نرخ رشد جمعیت در هر زمان متناسب با جمعیت فعلی باشد. یعنی هرچه جمعیت زیادتر باشد، نرخ رشد هم بیشتر باشد. این یک فرض معقول است زیرا با افزایش جمعیت، تعداد زاد و ولد هم افزایش می‌یابد. اگر جمعیت در زمان t را با $p(t)$ نشان دهیم، آنگاه معادله‌ی ساده‌ی زیر مدلی برای این نوع رشد جمعیت خواهد بود

$$p'(t) = rp(t), \quad 0 < t \leqslant b,$$
$$p(0) = p_0,$$

که در آن ثابت $r > 0$ نرخ ذاتی رشد جمعیت است که ضریب تناسب نرخ رشد جمعیت با خود جمعیت در زمان t است. سمت چپ معادله یعنی $p'(t)$، نرخ رشد جمعیت در زمان t است که برابر با ضریبی از خود جمعیت در همان زمان t است. مدل بالا که از یک معادله دیفرانسیل معمولی مرتبه اول خطی و یک مقدار اولیه (که بیان‌کننده جمعیت اولیه است) تشکیل

شده است، یک مدل ساده برای رشد جمعیت برای بازه‌های زمانی کوتاه تحت فرضیات محدودکننده‌ی مشخصی است. این مدل، یک مدل پیوسته ریاضی با متغیرهای پیوسته‌ی r، b، t، p و $p_۰$ است. اگر چه در واقعیت جمعیت گونه‌های حیوانات همواره یک عدد طبیعی است، اما در مدل‌سازی معمولاً آن را عددی حقیقی فرض می‌کنیم. در جمعیت‌های چگال مانند حجم یک ماده رادیواکتیو، یا غلظت مواد سمی در یک گالن آب لازم است متغیرها حتماً حقیقی (پیوسته) منظور شوند.

جواب مدل بالا به صورت زیر بدست می‌آید. ابتدا توجه داریم جمعیت همواره مثبت است یعنی $p(t) \geqslant ۰$ برای $t \geqslant ۰$. واضح است که تابع $p(t) \equiv ۰$ یک جواب مسئله است به شرطی که $p_۰ = ۰$. در حالتی که $p_۰ \neq ۰$، با تقسیم طرفین معادله بر $p(t)$ و انتگرال‌گیری از ۰ تا $\tau > ۰$ داریم

$$\int_۰^\tau \frac{p'(t)}{p(t)} dt = \int_۰^\tau r\, dt,$$

که این هم نتیجه می‌دهد

$$\ln \frac{p(\tau)}{p(۰)} = r\tau,$$

یا

$$p(\tau) = p_۰ \exp(r\tau), \quad \forall \tau > ۰.$$

روشن است که جواب $p = ۰$ نیز با فرض $p_۰ = ۰$ با تابع بالا داده می‌شود. روشی که به کمک آن جواب دقیق مسئله را بدست آوردیم یک روش *حلِ تحلیلی* می‌نامیم. جواب بدست آمده در بالا، رشد نمایی جمعیت را با گذشت زمان نشان می‌دهد. قطعاً این مدل نمی‌تواند در واقعیت و برای یک بازه زمانی بلند مدت درست باشد زیرا عواملی مانند مرگ و میر، رقابت بر سر منابع غذایی و مهاجرت باعث تغییر در رشد جمعیت می‌شوند. اگر به دنبال مدل واقعی‌تر باشیم لازم است فرضیات حاکم بر مسئله را کمتر کنیم. مثلا نرخ ذاتی رشد می‌تواند (برای گونه‌هایی از حیوانات) تابعی از زمان و حتی جمعیت به صورت $r = r(t, p)$ باشد. در یک حالت کلی‌تر (و نه لزوماً کلی‌ترین حالت ممکن) یک مدل معادله دیفرانسیل مقدار اولیه به صورت

$$p'(t) = f(t, p), \quad ۰ < t \leqslant b,$$
$$p(۰) = p_۰,$$

خواهیم داشت که f تابعی غیرخطی از t و p است. بدست آوردن جواب تحلیلی مسئله‌ی غیرخطی اخیر به سادگی مسئله‌ی اول نیست و برای برخی f های حتی ساده، معمولاً روشی تحلیلی برای یافتن جواب (در صورت وجود) در دست نیست، اگر هم راهی وجود داشته باشد سرراست نیست. همچنین در مدل می‌توان وجود یک شکارچی را در محیط فرض کرد که اگر جمعیت شکارچی هم مد نظر باشد و آن را با $q(t)$ نشان دهیم، به یک دستگاه از معادلات دیفرانسیل با دو تابع مجهول p و q خواهیم رسید که دارای پیچیدگی‌های بیشتری خواهد بود و یافتن جوابهای دقیق آن به عنوان توابعی با فرم بسته یا حتی فرم سری در اکثر مواقع غیر ممکن است.

مثالی که در بالا آورده شد یک مثال ساده از مدل‌سازی ریاضیِ پدیده‌های طبیعی است. به عنوان مثال‌های دیگر می‌توان به مدل‌سازی ریاضی مسائل فیزیکی، شیمیایی، اقتصادی، زیستی و غیره اشاره کرد که بررسی وجود و یکتایی جواب و همچنین راه بدست آوردن جواب‌های آن‌ها بخشی از ریاضیات امروزی است. همانگونه که در بالا گفته شد در اکثر حالت‌ها علیرغم وجود جواب برای این مدل‌ها، راهی برای بدست آوردن جواب واقعی (روش تحلیلی) وجود ندارد. در اینجا است که باید دست به دامان *روش‌های عددی* برای تولید یک "جواب تقریبی" شد. مثلاً برای حل معادله دیفرانسیل مقدار اولیه بالا می‌توان یک روش عددی ساده به صورت زیر طراحی کرد: ابتدا بازه‌ی $[۰, b]$ را به n زیربازه هر کدام به طول h تقسیم می‌کنیم و برای $k = ۰, ۱, \ldots, n$ قرار می‌دهیم $t_k := kh$. سپس معادله‌ی دیفرانسیل $p'(t) = f(t, p(t))$ را در نقطه‌ی $t = t_k$ می‌نویسیم. حال عبارت مشتق را به صورت زیر *تقریب* می‌زنیم

$$p'(t_k) \approx \frac{p(t_{k+۱}) - p(t_k)}{t_{k+۱} - t_k}.$$

این فرمول یک فرمول مشتق‌گیری عددی بسیار ساده است. فصلی از این درس در مورد انواع فرمول‌های *مشتق‌گیری و انتگرال‌گیری عددی* است که جایگاه ویژه‌ای در طراحی روش‌های عددی دارند. با جایگذاری تقریب مشتق بالا در معادله‌ی دیفرانسیل و با توجه به اینکه $t_{k+۱} - t_k = h$ به معادله‌ی گسسته‌ی زیر می‌رسیم

$$p_{k+۱} = p_k + hf(t_k, p_k), \quad k = ۰, ۱, \ldots, n - ۱,$$

که در آن p_k تقریبی از مقدار واقعیِ $p(t_k)$ است و $p_۰$ مقدار اولیه‌ی داده شده است. روش ساده‌ی بالا به *روش اویلر* مشهور است. در حقیقت ما مدل معادله دیفرانسیل پیوسته را *گسسته‌سازی* کردیم. گسسته‌سازی واژه‌ای آشنا در آنالیز عددی است.

روش عددی در نهایت منجر به یک "الگوریتم" شد که با پیاده‌سازی آن بر روی رایانه به هدف نهایی خواهیم رسید. کد الگوریتم روش اویلر در محیط متلب به صورت زیر است (تمام برنامه‌های این درس در محیط متلب نوشته و اجرا خواهند شد. اگر با این نرم‌افزار ریاضی آشنایی کامل ندارید به پیوست ... مراجعه کنید)

```
function p = euler(b,n,f,p0)
h=b/n;
t=0:h:b;
p(1) = p0;
for k = 1:n
    p(k+1) = p(k)+h*f(t(k),p(k));
end
```

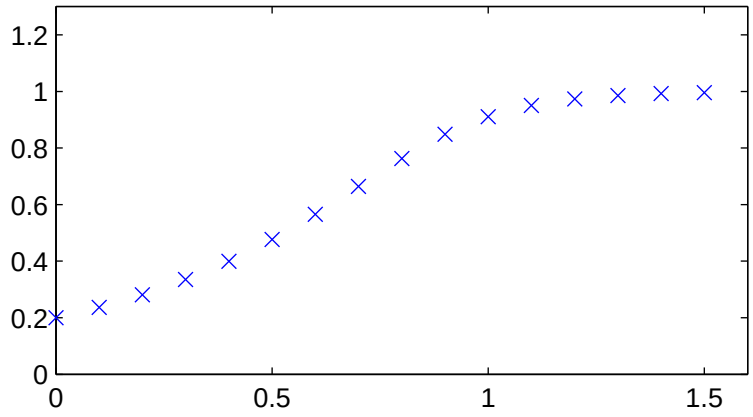
در این برنامه اعداد b (طول بازه)، n (تعداد زیربازه‌ها)، p0 (مقدار اولیه)، و تابع سمت راست f به عنوان ورودی داده می‌شوند و جواب مسئله در نقاط t_k در بردار p بازگردانده می‌شود. برای مثال فرض کنید می‌خواهیم تقریب جواب مسئله‌ی مقدار اولیه

$$p' = \sin(\pi(p^2 + p)/2)$$

با شرط اولیه مشخص p_0 را بدست آوریم. این نوع معادلات که سمت راست آن‌ها یعنی f فقط تابعی از جمعیت p است و به زمان t بستگی ندارد (یعنی $f = f(p)$) به معادلات رشد لجستیک معروفند و برای برخی از گونه‌های حیوانات تحت شرایط خاصی درست هستند. جواب این معادله‌ی جدا شدنی با انتگرال‌گیری از تابع $1/\sin(\pi(p^2 + p)/2)$ نسبت به p حاصل می‌شود که تابع اولیه‌ای برای آن در دست نیست. بنابراین اجازه دهید آن را با روش اویلر حل کنیم. برای این منظور دستورات زیر را می‌نویسیم که در آن تابع euler بالا فراخوانی شده است:

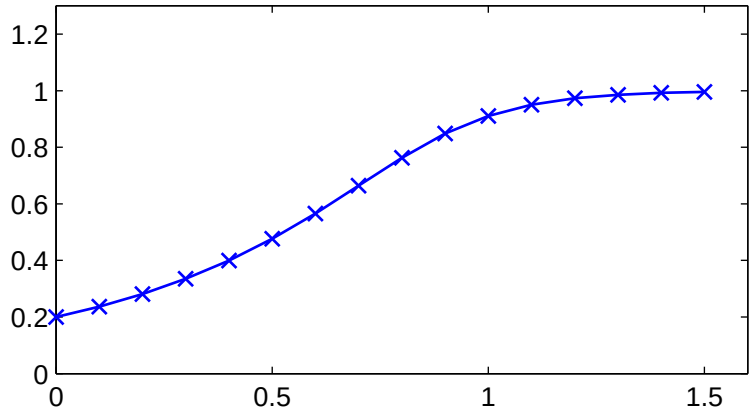
```
f = @(t,p) sin(pi/2*(p+p^2));
n = 15; b=1.5;
p = euler(b,n,f,0.2);
t = 0:b/n:b;
plot(t,p,'x')
```

اولین دستور برنامه بالا، تابع سمت راست را به کمک عملگر @ تعریف می‌کند. در انتها هم دستور plot نمودار شکل ۱.۱ را برایمان ترسیم می‌کند که در آن مقادیر تقریبی بدست آمده با روش اویلر در نقاط t_k با علامت $\times$ مشخص شده‌اند. این جواب با تقسیم بازه‌ی $[0, 1/5]$ به ۱۵ زیربازه و با مقدار اولیه $p_0 = 0/2$ حاصل شده است. همانطور که مشاهده



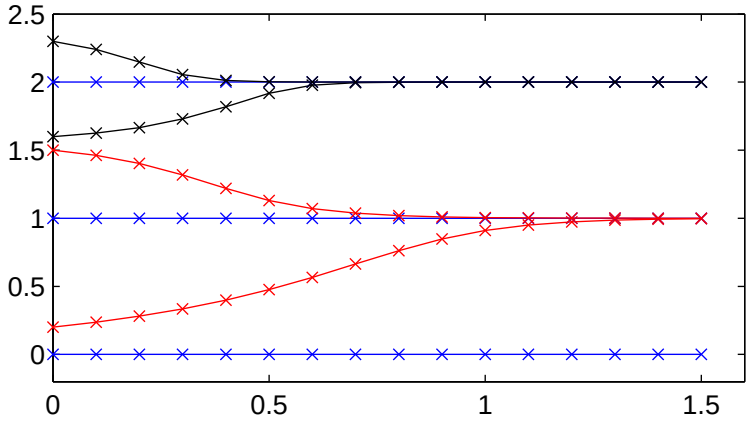
شکل ۱.۱: جواب تقریبی با روش اویلر در نقاط t_k

می‌کنید جواب تقریبی در تعداد متناهی نقطه بدست آمده است. اگر بخواهیم با در دست داشتن همین تعداد محدود مقادیر، جواب را در نقاط دیگر بازه هم بدست آوریم باز هم آنالیز عددی راهی به ما پیشنهاد می‌دهد؛ *تقریب و درونیابی* که موضوع یکی از فصل‌های این درس است. آسان‌ترین راه برای این کار وصل کردن نقاط با خط مستقیم به همدیگر و تولید یک تابع قطعه‌ای خطی پیوسته است. این تابع تقریب، پیوسته است اما مشتق‌پذیر نیست و برای اهدافی که به مشتق تقریب نیاز داریم قابل استفاده نیست. این تقریب قطعه‌ای خطیِ پیوسته در واقع ساده‌ترین نوع تقریب با *اسپلاین‌ها* است که بعداً در مورد آن‌ها و انواع دیگرشان، که همواری لازم را ارائه می‌دهند، صحبت خواهیم کرد. نمودار این تابع در شکل ۲.۱ رسم شده است.



شکل ۲.۱: جواب تقریبی با روش اویلر در نقاط t_k و تقریب جواب در نقاط دیگر

اگر با نظریه معادلات دیفرانسیل معمولی قدری آشنا باشیم می‌دانیم که رفتار جواب معادله لجستیک را می‌توان به صورت کیفی از روی تابع سمت راست معادله، یعنی f، و شرط اولیه p_0 حدس زد. برای مثال واضح است که یک تابع ثابت $p = c$ جواب معادله است اگر داشته باشیم $f(c) = 0$ زیرا مشتق تابع ثابت برابر صفر است. این جواب در صورت وجود، جواب تعادلی نامیده می‌شود. اگر $p = c$ یک جواب تعادلی باشد و مقدار اولیه p_0 هم c انتخاب شود روش ما نیز این تابع را نتیجه می‌دهد و جواب عددی با جواب دقیق یکسان است. با انتخاب مقادیر اولیه متفاوت جواب‌های دیگری حاصل می‌شوند که لزوماً توابع ثابت نیستند ولی این جواب‌ها با افزایش زمان یا به جواب تعادلی میل می‌کنند یا از آن دور می‌شوند. اگر جواب‌های دیگر به جواب تعادلی میل کنند، جواب تعادلی را پایدار و در غیر این صورت آن را ناپایدار می‌گوییم. در شکل ۳.۱ روش اویلر برای مقادیر اولیه مختلف $p_0 = 0, 0/2, 1, 1/5, 1/6, 2, 2/3$ اجرا شده و نمودار جواب‌ها (با تقریب قطعه‌ای خطی) رسم شده است. در اینجا توابع $p = 0$، $p = 1$ و $p = 2$ جواب‌های تعادلی هستند زیرا ریشه‌های $f(p) = \sin(\pi(p^2 + p)/2)$ می‌باشند. واضح است که این تابع ریشه‌های دیگری نیز دارد که با تغییر مقدار اولیه می‌توان آن‌ها را نیز بدست آورد و رفتار جواب‌های دیگر نسبت به آن‌ها را مشاهده کرد. بنابراین داشتن اطلاعاتی پیرامون مدل ریاضی برای موفقیت در حل عددی آن بسیار کارساز است.



شکل ۳.۱: جواب‌های تقریبی معادله لجستیک با روش اویلر با مقادیر اولیه مختلف

اکنون پرسشی که مطرح است این است که آیا روشی که طراحی کرده‌ایم همواره کار می‌کند؟ برای چه دسته از مسائل مطمئن هستیم که به جواب خوبی خواهیم رسید؟ جواب عددیِ بدست آمده تا چه میزان به جواب واقعی مسئله (که احیاناً در دست نیست) نزدیک است؟ اگر بخواهیم جوابی با دقت بهتر بدست آوریم چه باید بکنیم؟ آیا اگر n را افزایش دهیم (طول گام h را کاهش دهیم) به جواب دقیق‌تری خواهیم رسید؟ اگر جواب پرسش آخر مثبت است، با مثلاً نصف کردن h جواب چقدر بهتر می‌شود؟ این سؤالات و سؤالات احتمالی دیگر بر این واقعیت تأکید دارند که این الگوریتم و هر الگوریتم عددی دیگر باید مورد "تجزیه و تحلیل" قرار گیرد و از دیدگاه‌های متفاوتی، که برخی از آن‌ها را در این درس خواهیم خواند، مورد واکاوی واقع شود. برای مثال در آنالیز عددی ثابت می‌شود اگر سمت راست معادله یعنی $f(t,p)$ نسبت به هر دو متغیرش به اندازه‌ی کافی مشتق‌پذیر باشد و اگر طول گام h به اندازه‌ی کافی کوچک اختیار شود آنگاه اختلاف مقدار تقریبی p_n بدست آمده از روش اویلر و مقدار واقعی $p(t_n) = p(b)$ در کران زیر صدق می‌کند

$$|p_n - p(t_n)| \leqslant Cbh,$$

که در آن C یک ثابت مثبت است. کران بالا نشان می‌دهد با کاهش h، خطا به صورت خطی کاهش می‌یابد. برای مثال اگر h را نصف کنیم، خطا هم نصف می‌شود. همچنین این کران نشان می‌دهد هرچه b بزرگتر باشد خطای بیشتری ایجاد خواهد شد. این کران به نوعی یک *آنالیز خطای* ساده برای روش اویلر تحت فرضیات خاصی است، که البته ما اثبات آن را در اینجا نیاوردیم.

الگوریتم اویلر دارای همگرایی کند است و همچنین برای حل دسته‌ای از مسائل ناکارامد است. از این رو هدف بعدی ارائه الگوریتمی بهتر و کاراتر است. در آنالیز عددی همواره به سمت ساختن الگوریتم‌های بهتر که برای مسائل بیشتر و پیچیده‌تر کارایی دارند پیش می‌رویم. از آنجا که اینگونه مسائل روز به روز در علوم و مهندسی کشف و ظاهر می‌شوند و نیاز به الگوریتم‌های جدید دارند، آنالیز عددی همواره علمی پویا و روبه رشد است.

خلاصه‌ی روند طی شده برای یافتن جواب تقریبیِ یک پدیده‌ی طبیعی با الهام از مثال ساده‌ی بالا به صورت زیر است:

جواب تقریبی ⟵ الگوریتم ⟵ گسسته‌سازی ⟵ مدل پیوسته ریاضی ⟵ پدیده طبیعی

گذر از هر مرحله نیازمند بررسی‌های مربوط به خود است و لازم است معیارهایی برای اعتبارسنجی در نظر گرفته شود و تا مادامی که آن معیارها برآورده نشده‌اند آن مرحله مرتباً اصلاح شود. مدل‌سازیِ پدیده‌های طبیعی موضوعی بین رشته‌ای و تخصصی است که در این درس در مورد آن صحبت نخواهیم کرد. فرآیند رسیدن از معادله‌ی پیوسته به معادلات گسسته بخشی اعظمی از آنالیز عددی را شامل می‌شود که در موضوعی با عنوان *نظریه تقریب* موشکافی می‌شود. از سوی دیگر، اکثر مسائل گسسته منجر به حل دستگاههای معادلات جبری می‌شوند که معمولاً در قالب مباحث *جبرخطی عددی* باید به آنها پرداخت. در این درس مقداری در مورد مسائل جبرخطی عددی صحبت خواهیم کرد اما قسمت عمده‌ی آن را به درس بعدی که با همین عنوان است موکول می‌کنیم. در پایان، لازم به ذکر است که الگوریتم‌های حل مسائل گسسته باید بر اساس منطق، توانایی و پیشرفت‌های علوم کامپیوتر طراحی و تحلیل شوند و لذا وجه دیگر آنالیز عددی ارتباط تنگاتنگ و فهم صحیح از *محاسبات علمی* است تا بتوان مدل‌های گسسته را به شکل بهینه و قابل اعتماد پیاده سازی نمود. در یک حالت ساده، با توجه به اینکه متغیرهای حقیقی (یا مختلط) به صورت دقیق در کامپیوتر قابل ذخیره سازی نیستند و با سیستم ممیز شناور نمایش داده می‌شوند، بخشی از آنالیز عددی معطوف به مطالعه و بررسی نحوه کارکرد کامپیوتر و آنالیز خطاهای ایجاد شده در ذخیره سازی و نمایش این متغیرها است. بخش‌هایی از فصل دوم این درس به همین منظور نگارش شده است.

فصل ۲

خطاها و مفهوم پایداری

همانگونه که در فصل ۱ مشاهده کردیم در عمل به جای یافتن جواب تحلیلی یک مدل پیوسته یک جواب تقریبی برای آن بدست می‌آوریم. در مراحل مختلفِ فرآیند یافتن چنین جوابی لازم است از معادلات، کمیت‌ها و مقادیر تقریبی بجای نوع دقیقشان استفاده کنیم. مثلاً وقتی مدل پیوسته را گسسته‌سازی می‌کنیم در حقیقت بجای معادله‌ی پیوسته از مجموعه‌ای متناهی از معادلات گسسته استفاده می‌کنیم. همچنین همه‌ی اعداد و توابع مورد استفاده در الگوریتم در کامپیوتر به صورت دقیق ذخیره نمی‌شوند. لازم است ارتباط و اختلاف بین این تقریب‌ها و نوع دقیقشان در هر مرحله کنترل شود. در اینجا با تقریب اعداد شروع می‌کنیم و با نحوه‌ی ذخیره‌سازی اعداد در کامپیوتر آشنا می‌شویم. قبل از آن در مورد خطای مطلق و خطای نسبی توضیحاتی ارائه می‌دهیم.

۱.۲ خطای مطلق و نسبی

اختلاف مقدار واقعی یک کمیت و مقدار تقریبی آن را "خطا" می‌نامیم. معمولاً دو متر مختلف برای اندازه‌گیری خطا وجود دارد: فرض کنیم $x \in \mathbb{R}$ و $\widehat{x}$ تقریبی از آن باشد، *خطای مطلق* این تقریب عبارتست از

$$|\widehat{x} - x|$$

و اگر $x \neq 0$ *خطای نسبی* این تقریب عبارتست از

$$\frac{|\widehat{x} - x|}{|x|}.$$

اگر $x \in \mathbb{C}$ آنگاه $|x|$ اندازه‌ی x است که با $\sqrt{\mathrm{Re}(x)^2 + \mathrm{Im}(x)^2}$ تعریف می‌شود. در محاسبات معمولاً خطای نسبی مورد توجه است چراکه برخلاف خطای مطلق وابسته به مقیاس نیست به این معنی که اگر $x \mapsto \alpha x$ و $\widehat{x} \mapsto \alpha\widehat{x}$ خطای نسبی تغییر نخواهد کرد. در بسیاری از موارد مطلوب است که یک *کران بالا* برای خطای مطلق یا نسبی بدست آوریم. به

عنوان مثال می‌دانیم که عدد گنگ π دارای یک نمایش اعشاری با بینهایت رقم است که قسمتی از نمایش آن به صورت

$$۳/۱۴۱۵۹۲۶۵۳۵۸۹۷۹\cdots$$

است. عدد $\widehat{x} = ۳/۱۴۲$ تقریبی از π با کران خطای مطلق

$$|\pi - ۳/۱۴۲| \leqslant ۰/۵ \times ۱۰^{-۳}$$

است. اگر x برداری در $\mathbb{R}^n$ باشد به‌جای قدرمطلق از نرم برداری استفاده می‌کنیم و خطای مطلق را با

$$\|\widehat{x} - x\|$$

و خطای نسبی را با

$$\frac{\|\widehat{x} - x\|}{\|x\|}, \quad x \neq ۰,$$

تعریف می‌کنیم که در آن‌ها $\|\cdot\|$ یک نرم برداری است. برای مثال فرض کنیم برای $x = (x_۱, \ldots, x_n) \in \mathbb{R}^n$ نرم به صورت زیر تعریف شده باشد

$$\|x\|_\infty := \max_{۱\leqslant k\leqslant n} |x_k|, \tag{۱.۲}$$

که به آن نرم ماکزیمم یا نرم بینهایت می‌گوییم. فرض کنیم $x = (۱, ۲, -۳)$ و $\widehat{x} = (۰/۹۹۹۷, ۲/۰۰۳۲, -۲/۹۸۴۰)$. در این صورت داریم

$$\|x - \widehat{x}\|_\infty = \|(۰/۰۰۰۳, -۰/۰۰۳۲, -۰/۰۱۶۰)\|_\infty = ۰/۰۱۶۰.$$

نرم بینهایت تنها نرمی نیست که می‌توان روی $\mathbb{R}^n$ تعریف کرد. برای مثال می‌توان نرم اقلیدسی (یا نرم دو) را به صورت

$$\|x\|_۲ := \left(\sum_{k=۱}^{n} |x_k|^۲\right)^{۱/۲}, \tag{۲.۲}$$

تعریف کرد. به طور کلی یک p-نرم برداری به صورت زیر تعریف می‌شود

$$\|x\|_p := \left(\sum_{k=۱}^{n} |x_k|^p\right)^{۱/p}, \quad p \geqslant ۱. \tag{۳.۲}$$

این تعاریف برای $x \in \mathbb{C}^n$ نیز برقرارند و در این صورت $|\cdot|$ همان اندازه‌ی عدد مختلط است. در مثال بالا داریم

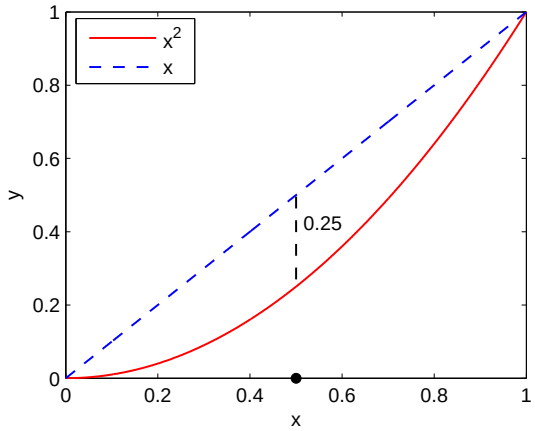
$$\|x - \widehat{x}\|_۲ = \sqrt{(۰/۰۰۰۳)^۲ + (۰/۰۰۳۲)^۲ + (۰/۰۱۶۰)^۲} \doteq ۰/۰۱۶۳.$$

از این پس هرگاه از نماد $\doteq$ استفاده کنیم منظورمان تساوی تا تعداد ارقام اعشار ذکر شده است. مثلاً در بالا مقدار عبارت رادیکال ۰٫۰۱۶۳۱۹۶۲۰۰۹۳۶۲۰۰۰ است که تا چهار رقم با ۰٫۰۱۶۳ یکی است.

عبارت خطا تنها در مورد اعداد (و بردارها) به کار نمی‌رود. در حالت کلی‌تر در هر فضای نرم دار این مفهوم قابل تعریف است. به عنوان مثال فرض کنید $C[a,b]$ فضای همه توابع پیوسته حقیقی-مقدار روی بازه‌ی $[a,b]$ باشد. یک نرم روی این فضا به صورت

$$\|f\|_\infty := \max_{x\in[a,b]} |f(x)| \tag{۴.۲}$$

تعریف می‌شود که به آن نرم بینهایت (یا نرم چبیشف) می‌گویند. برای مثال تابع $f(x)=x^2$ در $C[0,1]$ است و روشن است که $\|f\|_\infty = 1$. حال فرض کنید این تابع را با تابع $\widehat{f}(x)=x$ روی $[0,1]$ تقریب بزنیم. (این تقریب یک درونیاب خطی f است که در فصل بعد در مورد آن صحبت خواهیم کرد). خطای تقریب عبارتست از $e(x) := x^2 - x$ که ماکزیمم قدر مطلق آن روی $[0,1]$ در نقطه‌ی $x=0.5$ با مقدار $|e(0.5)| = 0.25$ رخ می‌دهد. بنابراین $\|e\|_\infty = 0.25$. در شکل ۱.۲ نمودار این دو تابع و نقطه‌ای که خطای ماکزیمم در آن رخ داده و مقدار این خطا مشخص شده‌اند.



شکل ۱.۲: تقریب تابع x^2 با تابع خطی x و خطای ماکزیمم این تقریب.

همانند آنچه در مورد فضای $\mathbb{R}^n$ گفتیم، برای فضای $C[a,b]$ هم می‌توان نرم‌های مختلفی تعریف کرد. به عنوان مثال نرم یک به صورت زیر تعریف می‌شود

$$\|f\|_1 := \int_a^b |f(x)|dx. \tag{۵.۲}$$

با توجه به اینکه $|f|$ پیوسته و $[a,b]$ فشرده است، انتگرال بالا قابل تعریف است. به عنوان مثال برای $f(x)=x^2$ روی $[0,1]$ داریم

$$\|f\|_1 = \int_0^1 x^2 dx = \frac{1}{3},$$

و برای خطای تقریب خطی آن در نرم یک داریم

$$\|e\|_۱ = \int_۰^۱ |x - x^۲| dx = \frac{۱}{۶} \doteq ۰/۱۶۶۷.$$

در حقیقت نرم یک مساحت بین دو نمودار $x^۲$ و x است. هر چه این مساحت کمتر باشد یعنی تقریب ما بهتر است. بر خلاف نرم بینهایت که خطا را در بدترین نقطه‌ی بازه در نظر می‌گیرد، نرم یک، خطا روی کل بازه (مساحت بین دو نمودار) را محاسبه می‌کند. یکی دیگر از پر استفاده‌ترین نرم‌های توابع، نرم دو است که به صورت

$$\|f\|_۲ := \left(\int_a^b f(x)^۲ dx \right)^{۱/۲}, \tag{۶.۲}$$

تعریف می‌شود. باز هم چون $f^۲$ پیوسته و $[a, b]$ فشرده است، انتگرال قابل تعریف است. برای خطای تقریب خطی مذکور در نرم دو داریم

$$\|e\|_۲ = \left(\int_۰^۱ (x - x^۲)^۲ dx \right)^{۱/۲} = \frac{۱}{\sqrt{۳۰}} \doteq ۰/۱۸۲۶.$$

نرم دوی خطا در واقع جذر مجموع مربعات خطا روی کل بازه است.

۲.۲ نمایش ماشینی اعداد

ما در زندگی روزمره‌ی خود برای نمایش اعداد از *دستگاه ارزش مکانی* در مبنای ۱۰ (دستگاه دهدهی) استفاده می‌کنیم. بنابراین برای نشان دادن یک عدد از کاراکترهای ۰، ۱، . . .، ۹ استفاده می‌کنیم. مقدار عدد به ارزش کاراکترهایش و مکان آنها وابسته است. برای مثال عدد ۳۴۶/۴۹۷۵ یعنی

$$۳ \times ۱۰^۲ + ۴ \times ۱۰^۱ + ۶ \times ۱۰^۰ + ۴ \times ۱۰^{-۱} + ۹ \times ۱۰^{-۲} + ۷ \times ۱۰^{-۳} + ۵ \times ۱۰^{-۴}.$$

هر عدد حقیقی دارای نمایشی یکتا به صورت بالاست مگر اینکه بینهایت ۹ بعد از اعشار تکرار شود. مثلاً ۳/۱۴۹۹۹۹۹۰۰۰ و ۳/۱۵ هر دو نمایش یک عدد حقیقی هستند. همچنین می‌توان دستگاه ارزش مکانی را در مبناهای غیر از ۱۰ نیز به‌کار برد. هر عدد صحیح $\beta \geq ۲$ را می‌توان به عنوان مبنا در نظر گرفت. هر عدد حقیقی مثبت a را می‌توان به‌صورت

$$a = d_n \beta^n + d_{n-۱} \beta^{n-۱} + \cdots + d_۱ \beta^۱ + d_۰ \beta^۰ + d_{-۱} \beta^{-۱} + d_{-۲} \beta^{-۲} + \cdots,$$

یا فشرده‌تر

$$a = (d_n d_{n-۱} \cdots d_۱ d_۰ / d_{-۱} d_{-۲} \cdots)_\beta,$$

نشان داد که $d_i \in \{۰, ۱, \ldots, \beta - ۱\}$. نمایش بالا منحصر بفرد است مگر زمانی که بینهایت $\beta - ۱$ بعد از اعشار وجود داشته باشد. هر چه مبنا کوچکتر باشد محاسبات ساده‌تر است. بنابراین مبنای ۲ ساده‌ترین خواهد بود و همچنین

چون اجزای مکانیکی و الکتریکی همواره دو حالت (صفر و یک) را به راحتی نمایش می‌دهند (مثلاً لامپ یا روشن است یا خاموش، جهت میدان یا ساعتگرد است یا پادساعتگرد و ...)، از این رو غالباً ۲ مبنای نمایش اعداد در رایانه است و به آن *دستگاه دودویی* می‌گویند. هر رقم در دستگاه دودویی بیت گفته می‌شود.

برای ذخیره‌ی اعداد حقیقی در کامپیوتر می‌توان از دو نوع نمایش استفاده کرد که در ادامه به آن‌ها می‌پردازیم و در آخر خواهیم دید که در ماشین‌های امروزی از نمایش دوم استفاده می‌شود.

۱.۲.۲ نمایش ممیز ثابت و ممیز شناور

در دستگاه نمایش ممیز ثابت به خاطر محدودیت حافظه، t رقم برای ارقام بعد از ممیز و n رقم برای ارقام قبل از ممیز در نظر گرفته می‌شود. و همین‌طور یك بیت نیز برای علامت در نظر گرفته می‌شود. بنابراین در این دستگاه فرم کلی اعداد به صورت

$$a = \pm(d_{n-1}d_{n-2}\cdots d_1 d_0 / d_{-1}d_{-2}\cdots d_{-t})_\beta,$$

خواهد بود. مجموعه‌ی اعداد این دستگاه را با $\mathbb{F}_0(\beta, n, t)$ نمایش می دهیم. این مجموعه دارای خواص زیر است:

- $\mathbb{F}_0(\beta, n, t)$ یک زیرمجموعه از اعداد حقیقی است و تعداد اعضای آن $2 \times \beta^n \times \beta^t - 1$ است. دقت کنید که با این دستگاه عدد صفر دو نمایش $0+$ و $0-$ دارد.

- با فرض اینکه $\gamma = \beta - 1$، بزرگترین عدد مثبت در این دستگاه عبارتست از

$$x_{\max} = \big(\underbrace{\gamma\gamma\cdots\gamma}_{n \text{ تا}} / \underbrace{\gamma\gamma\cdots\gamma}_{t \text{ تا}}\big)_\beta = \gamma \sum_{i=-t}^{n-1} \beta^i = \beta^n - \beta^{-t} \approx \beta^n.$$

- کوچکترین عدد مثبت از لحاظ قدرمطلق در این دستگاه عبارتست از

$$x_{\min} = \big(\underbrace{00\cdots0}_{n \text{ تا}} / \underbrace{00\cdots01}_{t-1 \text{ تا}}\big)_\beta = \beta^{-t}.$$

- توزیع اعداد در این دستگاه یکنواخت است. یعنی اعداد به‌صورت هم‌فاصله در بازه‌ی $[-x_{\max}, x_{\max}]$ قرار دارند. فاصله‌ی دو عدد متوالی β^{-t} است.

در اولین کامپیوترها محاسبات با دستگاه ممیز ثابت در مبنای ۲ انجام می‌شد. در این دستگاه $n + t + 1$ بیت برای نمایش یک عدد لازم است که یکی از بیتها، بیت علامت است. ضعف این دستگاه در این است که بازه‌ی $[-x_{\max}, x_{\max}] \approx [-2^n, 2^n]$ محدوده‌ی آنچنان بزرگی از اعداد حقیقی نیست مگر آنکه n بسیار بزرگ انتخاب شود که حافظه‌ی محدود کامپیوتر چنین اجازه‌ای را نمی‌دهد. این ضعف از آن‌جا ناشی می‌شود که توزیع اعداد در این دستگاه یکنواخت است، یعنی

حساسیت آن روی اعداد کوچک و اعداد بزرگ یکی است. از این رو در کامپیوترهای امروزی از دستگاه ممیز شناور استفاده می‌شود. در دستگاه ممیز شناور هر عدد حقیقی a به‌فرم

$$a = \pm m \times \beta^e, \quad \beta^{-1} \leqslant m < 1, \quad e \in \mathbb{Z},$$

نمایش داده می‌شود. برای هر عدد حقیقی غیر صفر چنین نمایشی یکتاست. قسمت اعشاری m را *مانتیس*، e را *نما* و طبق معمول β را مبنا می‌گوییم. در عمل تعداد ارقام مانتیس و نما محدود انتخاب می‌شوند. اگر تعداد t رقم برای مانتیس استفاده شود و نما در یک بازه متناهی محدود شود، تنها می‌توان اعداد ممیز شناور به شکل زیر را نمایش داد

$$\overline{a} = \pm \overline{m} \times \beta^e = \pm(0/d_1 d_2 \cdots d_t)_\beta \times \beta^e, \quad d_i \in \{0, 1, \ldots, \beta - 1\}, \tag{۷.۲}$$

طوری‌که $\overline{m}$ گرد شده‌ی m تا t رقم بعد از اعشار است و محدوده‌ی نما

$$L \leqslant e \leqslant U,$$

است. برای اینکه نمایش بالا منحصربفرد باشد همواره فرض می‌کنیم $d_1 \neq 0$ و به مجموعه اعدادی که با این دستگاه قابل نمایش هستند اعداد ممیز شناور نرمال می‌گوییم و آن را با $\mathbb{F}(\beta, t, L, U)$ نمایش می‌دهیم. این دستگاه دارای خواص زیر است:

- مجموعه‌ی $\mathbb{F}(\beta, t, L, U)$ زیرمجموعه‌ای از اعداد حقیقی است و به راحتی می‌توان نشان داد که تعداد اعضای آن $2 \times (\beta - 1) \times \beta^{t-1} \times (U - L + 1)$ است.

- با فرض اینکه $\gamma = \beta - 1$، بزرگترین عدد مثبت در این دستگاه عبارتست از

$$x_{\max} = \big(0/\underbrace{\gamma\gamma\cdots\gamma}_{t \text{ تا}}\big)_\beta \times \beta^U = (\beta - 1)\beta^U \sum_{i=1}^{t} \beta^{-i} = \beta^U(1 - \beta^{-t}) \approx \beta^U.$$

- کوچکترین عدد مثبت از لحاظ قدرمطلق در این دستگاه عبارتست از

$$x_{\min} = \big(0/1\underbrace{00\cdots0}_{t-1 \text{ تا}}\big)_\beta \times \beta^L = \beta^{L-1}.$$

- توزیع اعداد در این دستگاه یکنواخت نیست. در حقیقت اگر $x \in \mathbb{F}(\beta, t, L, U)$ یک عدد مثبت و دارای نمایش $x = (0/d_1 \cdots d_t)_\beta \times \beta^e$ باشد، عدد ماشینی بلافاصله بعد از آن که با x_+ نشان داده می‌شود دارای نمایش $x_+ = \big(0/d_1 \cdots d_t + 0/\underbrace{00\cdots0}_{t-1 \text{ تا}}1\big)_\beta \times \beta^e$ است. از این رو فاصله‌ی آنها عبارتست از $\big(0/\underbrace{00\cdots0}_{t-1 \text{ تا}}1\big)_\beta \times \beta^e = \beta^{e-t}$ که بوضوح وابسته به نما است، یعنی هرچه نما بزرگتر باشد (اعداد بزرگتر باشند) فاصله‌ی آن‌ها بیشتر خواهد بود.

بازه‌ی $\Omega = [-x_{\max}, x_{\max}]$ را *دامنه‌ی اعداد ممیز شناور* می‌گوییم.

مثال ۱.۲. دستگاه ممیز شناور $\mathbb{F}(۲, ۳, -۱, ۲)$ را در نظر بگیرید. نیمه‌ی مثبت این اعداد در شکل ۲.۲ نمایش داده شده است. مجموعه‌ی $\mathbb{F}$ دقیقاً **۳۲** عضو دارد که طبق فرمول‌های بالا بزرگترین عضو آن $(۰/۱۱۱)_۲ \times ۲^۲ = ۳/۵$ و کوچکترین عدد مثبت عضو آن $(۰/۱۰۰)_۲ \times ۲^{-۱} = ۰/۲۵$ است. مشاهده می‌کنید که توزیع نقاط یکنواخت نیست، همچنین یک فضای خالی نسبتاً بزرگ اطراف صفر وجود دارد.

شکل ۲.۲: اعداد مثبت دستگاه نرمال $\mathbb{F}(۲, ۳, -۱, ۲)$

تعریف ۱.۲. در دستگاه ممیز شناور نرمال $\mathbb{F}(\beta, t, L, U)$ به فاصله‌ی عدد ۱ و عدد بلافاصله بعد از آن *اپسیلون ماشین* می‌گویند و با ϵ_M نمایش می‌دهند.
با توجه به اینکه $(۰/۱۰۰\cdots۰)_\beta \times \beta^۱ = ۱$ و عدد بلافاصله بعد از آن $\big(۰/۱\underbrace{۰۰\cdots۰}_{t-۲\ \text{تا}}۱\big)_\beta \times \beta^۱ = ۱_+$، است لذا $\epsilon_M = \beta^{۱-t}$.

واضح است $\mathbb{F}(\beta, t, L, U) \subset \Omega$، اگر عدد حقیقی a در Ω بوده اما عضو $\mathbb{F}$ نباشد، می‌توان با نگاشت

$$fl \ : \ \Omega \longrightarrow \mathbb{F}(\beta, t, L, U),$$

آن را با یک عدد ممیز شناور $\overline{a}$ نمایش داد که این عدد می‌تواند یکی از اعداد ممیز شناور بلافاصله بعد یا قبل a باشد. می‌نویسیم $\overline{a} = fl(a)$. این نگاشت را *گردکردن* می‌گوییم. از جمله می‌توان به گردکردن به سمت صفر (بریدن) و گردکردن به سمت نزدیکترین عدد ممیز شناور اشاره کرد. همچنین گردکردن به سمت $+\infty$ (گردکردن به بالا) و گردکردن به سمت $-\infty$ (گردکردن به پایین) را نیز می‌توان درنظر گرفت. بریدن و گردکردن به نزدیکترین در حساب ممیز شناور عادی و گردکردن به بالا و پایین در حساب ممیزشناور بازه‌ای کاربرد دارند.

بریدن: اگر $a \in \Omega$ به‌صورت $a = \pm(۰/d_۱ \cdots d_t d_{t+۱} \cdots)_\beta \times \beta^e$ باشد، آنگاه بریده شده یا گرد شده‌ی a به سمت صفر اولین عضو مجموعه‌ی $\mathbb{F}(\beta, t, L, U)$ است که سر راه حرکت a به سمت صفر قرار دارد. از این‌رو اگر نگاشت بریدن را با fl_{c} نشان دهیم، داریم:

$$\overline{a} = fl_{\mathrm{c}}(a) = \pm(۰/d_۱ \cdots d_t)_\beta \times \beta^e.$$

بسادگی می‌توان دید که از لحاظ خطای مطلق

$$\begin{aligned}\big|a - fl_{\rm c}(a)\big| &= (\text{۰/}\underbrace{\text{۰۰}\cdots\text{۰}}_{\text{تا } t} d_{t+\text{۱}}\cdots)_\beta \times \beta^e \\ &\leqslant (\text{۰/}\underbrace{\text{۰۰}\cdots\text{۰}}_{\text{تا } t-\text{۱}}\text{۱})_\beta \times \beta^e \\ &= \beta^{e-t},\end{aligned} \tag{۸.۲}$$

و از لحاظ خطای نسبی

$$\begin{aligned}\frac{\big|a - fl_{\rm c}(a)\big|}{|a|} &\leqslant \frac{\beta^{e-t}}{(\text{۰/}d_\text{۱}\cdots d_t d_{t+\text{۱}}\cdots)_\beta \times \beta^e} \\ &= \frac{\beta^{-t}}{(\text{۰/}d_\text{۱}\cdots d_t d_{t+\text{۱}}\cdots)_\beta} \\ &\leqslant \frac{\beta^{-t}}{(\text{۰/۱۰}\cdots\text{۰})_\beta} \\ &= \beta^{\text{۱}-t}.\end{aligned} \tag{۹.۲}$$

همانگونه که ملاحظه می‌شود، کران بالای خطای نسبی به اندازه عدد (نما) بستگی ندارد و فقط به تعداد ارقام مانتیس وابسته می‌باشد. بنابراین می‌بینیم که اگر چه فاصله‌ی دو عدد متوالی بزرگ زیاد است اما تأثیری در خطای نسبی ندارد و این مزیت اعداد ممیز شناور است.

گرد کردن به نزدیکترین: اگر $a = \pm(\text{۰/}d_\text{۱}\cdots d_t d_{t+\text{۱}}\cdots)_\beta \times \beta^e$، آنگاه گرد شده‌ی a به نزدیکترین، عددی عضو $\mathbb{F}(\beta, t, L, U)$ است که نزدکترین فاصله را با a دارد. نگاشت گردکردن به نزدیکترین را با $fl_{\rm rn}$ نشان می‌دهیم. با توجه به اینکه اعداد ممیز شناور بلافاصله قبل و بعد a در $\mathbb{F}$ عبارتند از $a_- = \pm(\text{۰/}d_\text{۱}\cdots d_t)_\beta \times \beta^e$ و $a_+ = \pm(\text{۰/}d_\text{۱}\cdots d_t + \text{۰/}\underbrace{\text{۰۰}\cdots\text{۰}}_{\text{تا } t-\text{۱}}\text{۱})_\beta \times \beta^e$، پس در این حالت گرد شده‌ی a یکی از این دو عدد است. اگر a از عدد وسط این دو کوچکتر باشد $fl_{\rm rn}(a) = a_-$ در غیر اینصورت $fl_{\rm rn}(a) = a_+$. به‌سادگی می‌توان نشان داد

$$\big|a - fl_{\rm rn}(a)\big| \leqslant \frac{\text{۱}}{\text{۲}}\beta^{e-t}, \qquad \frac{\big|a - fl_{\rm rn}(a)\big|}{|a|} \leqslant \frac{\text{۱}}{\text{۲}}\beta^{\text{۱}-t}. \tag{۱۰.۲}$$

پرسش ۳ را ببینید. پس خطای مطلق و نسبی گردکردن به نزدیکترین نصف بریدن است، از این‌رو در اکثر ماشین‌های محاسباتی و بخصوص کامپیوترها به صورت پیش‌فرض از نگاشت گردکردن به نزدیکترین استفاده می‌شود. در ادامه برای رعایت اختصار منظور از "گردکردن" همان "گردکردن به نزدیکترین" است مگر اینکه خلاف آن ذکر شود.

گردکردن به بالا و گردکردن به پایین: اگر $a = \pm(\text{۰/}d_\text{۱}\cdots d_t d_{t+\text{۱}}\cdots)_\beta \times \beta^e$ گرد شده‌ی a به بالا (پایین) اولین عدد عضو $\mathbb{F}(\beta, t, L, U)$ است که سر راه حرکت از a به سمت $+\infty$ ($-\infty$) است. بنابراین اگر a عددی منفی (مثبت) باشد، گرد شده‌ی آن به بالا (پایین) و بریده شده‌ی آن هر دو یکی هستند. به راحتی می‌توان نشان داد کران‌هایی که برای خطای مطلق و نسبی در حالت بریدن بدست آمدند، برای گردکردن به سمت بالا و پایین نیز برقرارند. پرسش ۴ را ببینید.

تعریف ۲.۲. در دستگاه $\mathbb{F}(\beta,t,L,U)$ همراه با نگاشت گردکردن fl، عدد حقیقی و مثبت u به طوری‌که

$$fl(1+\delta)=1, \quad \forall \delta \leqslant u,$$

را *واحد گردکردن* می‌گویند.

مقدار واحد گردکردن هم به مقدار t و هم به نوع نگاشت fl بستگی دارد، در حالی که اپسیلون ماشین فقط به مقدار t وابسته است. با توجه به مقدار اپسیلون ماشین به‌راحتی می‌توان نشان داد

$$u=\begin{cases} \frac{1}{2}\beta^{-t+1}, & \text{اگر از گردکردن استفاده شود} \\ \beta^{-t+1}, & \text{اگر از بریدن استفاده شود} \end{cases}. \tag{۱۱.۲}$$

در یک دستگاه $\mathbb{F}(\beta,t,L,U)$، هر عدد حقیقی که در دامنه‌ی $\mathbb{F}$ قرار داشته باشد را می‌توان با خطای نسبی نابیشتر از واحد گردکردن u نمایش داد. با توجه به روابط (۹.۲)، (۱۰.۲) و (۱۱.۲) اگر x یک عدد حقیقی باشد طوری‌که $x_{\min} \leqslant |x| \leqslant x_{\max}$، داریم

$$fl(x)=x(1+\varepsilon), \quad |\varepsilon| \leqslant u. \tag{۱۲.۲}$$

در یک دستگاه ممیز شناور اعداد بزرگ و اعداد کوچک با یک دقت نسبی تقریباً برابر نمایش داده می‌شوند. مقدار واحد گردکردن u همواره به عنوان یک معیار نسبی در تغییرات نسبی و خطاهای نسبی مورد استفاده قرار می‌گیرد. به عنوان مثال معیار توقف در الگوریتم‌های تکراری معمولاً به واحد گردکردن وابسته است.

دیدیم که در دستگاه اعداد ممیز شناور نرمال عدد صفر نمایش داده نمی‌شود و همچنین یک فضای خالی نسبتاً بزرگ اطراف صفر وجود دارد. برای تعریف صفر و همچنین پوشش این فضا *اعداد زیرنرمال* را تعریف می‌کنیم. یک عدد زیرنرمال به‌صورت

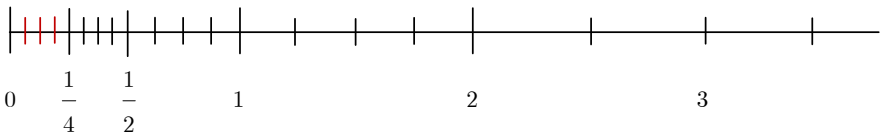
$$\overline{a}=\pm(0/0 d_2 \cdots d_t)_\beta \times \beta^L, \quad d_i \in \{0,1,\ldots,\beta-1\}, \tag{۱۳.۲}$$

تعریف می‌شود. توجه داریم که رقم بلافاصله بعد از ممیز در این عدد برابر صفر و نما برابر کمترین کران یعنی L است. با توجه به اینکه نما ثابت است این نمایش منحصر بفرد است. به عنوان نمونه در دستگاه $\mathbb{F}(2,3,-1,2)$ که در مثال ۱.۲ بررسی شد، اعداد زیرنرمال عبارتند از

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

که نیمه‌ی مثبت آن‌ها به همراه اعداد نرمال در شکل ۳.۲ ترسیم شده است.

ملاحظه ۱.۲. اگر $x \in (-\infty,-x_{\max}) \cup (x_{\max},\infty)$ آنگاه مقدار $fl(x)$ تعریف نمی‌شود، در حالی که اگر $x \in (-x_{\min},x_{\min})$ نگاشت گردکردن قابل تعریف است (حتی اگر اعداد زیرنرمال تعریف نشده باشند). در حالت اول گوییم



شکل ۳.۲: اعداد مثبت نرمال و زیرنرمال در $\mathbb{F}(۲, ۳, -۱, ۲)$

سرریز و در حالت دوم گوییم *پی‌ریز* رخ داده است. اگر اعداد زیرنرمال تعریف شده باشند در حالت دوم گوییم *پی‌ریز تدریجی* رخ داده است. در حقیقت وقتی نمایش یک عدد حقیقی یک عدد زیرنرمال باشد گوییم پی‌ریز تدریجی رخ داده است. با توجه به فضای خالی نسبتاً بزرگ اطراف صفر در اعداد نرمال، اضافه کردن اعداد زیرنرمال این ناحیه را منظم‌تر و یکنواخت‌تر می‌کند. برای لمس بهتر این موضوع شکل‌های ۲.۲ و ۳.۲ را دوباره مقایسه کنید.

۲.۲.۲ استاندارد IEEE

اگرچه در برخی از ماشین‌حساب‌ها از دستگاه ممیز شناور دهدهی استفاده می‌شود، اما تقریباً در همه کامپیوترهای جدید دستگاه دودویی مورد استفاده قرار می‌گیرد. در اوایل عصر توسعه‌ی کامپیوترها بیشتر آن‌ها از این دستگاه استفاده می‌کردند اما تفاوت‌هایی در مبنا و اندازه های مانتیس و محدوده‌ی نما و حتی نحوه‌ی نرمال‌سازی وجود داشت. این تفاوت‌ها منجر به انتقال ناپذیری نرم افزارها روی ماشین‌های متفاوت می‌گردید . در این راستا نیاز به وجود یک استاندارد برای حساب ممیز شناور باعث شد تا در نتیجه همکاری کارخانجات تولید سخت افزار و دانشمندان علوم کامپیوتر به سرپرستی ویلیام کاهان [۱] از دانشگاه کالیفرنیا در برکلی، در سال ۱۹۸۵ میلادی استانداردی برای نمایش اعداد در انجمن مهندسان برق و الکترونیک [۲] (IEEE) وضع شود که در سال ۱۹۸۹ مورد تأیید کمسیون بین‌المللی الکترونیک [۳] (IEC) قرار گرفت [۴]. نام این استاندارد IEEE 754 است و امروزه اکثر کامپیوترها برای محاسبات ممیز شناور دودویی از آن پیروی می‌کنند.

استاندارد IEEE 754 دو فرمت عمده‌ی *دقت معمولی* و *دقت دوبرابر* را پشتیبانی می‌کند که در اولی ۳۲ و در دومی ۶۴ بیت برای نمایش هر عدد ممیز شناور مورد استفاده قرار می‌گیرد. دقت معمولی دودویی با دستگاه $\mathbb{F}(۲, ۲۴, -۱۲۵, ۱۲۸)$ و دقت دوبرابر دودویی با دستگاه $\mathbb{F}(۲, ۵۳, -۱۰۲۱, ۱۰۲۴)$ ساخته می‌شوند و هر دو هم شامل اعداد نرمال هستند و هم اعداد زیرنرمال. در دقت معمولی برای نمایش یک عدد ممیز شناور a، ۱ بیت برای نمایش علامت (برای علامت منفی مقدار آن یک و برای علامت مثبت مقدار آن صفر است)، ۸ بیت برای نمایش نما و ۲۳ بیت برای نمایش مانتیس در نظر گرفته می‌شود. در دقت دوبرابر ۱ بیت برای نمایش علامت، ۱۱ بیت برای نما و ۵۲ بیت برای مانتیس استفاده می‌شود.

[۱] وی که به‌خاطر تلاش‌هایش در زمینه‌ی ابداع استاندارد IEEE 754، جایزه تورینگ را دریافت کرده، از اولین محققین آنالیز بازه‌ای نیز محسوب می‌شود

[۲] Institute of Electrical and Electronics Engineers (IEEE).

[۳] International Electronical Commission (IEC).

[۴] قابل ذکر است که IEC همانند سازمان استاندارد بین‌المللی (ISO) ، یک سازمان استاندارد جهانی در زمینه‌ی الکترونیک می‌باشد.

اگر a یک عدد ممیز شناور نرمال باشد، تقریب $\overline{a}$ از a به‌صورت

$$\overline{a} = \pm(۱/m)_۲ \times ۲^e, \quad e_{\min} \leqslant e \leqslant e_{\max}, \tag{۱۴.۲}$$

در نظر گرفته می‌شود. نکته اینکه رقم قبل از ممیز برای اعداد نرمال همیشه ۱ است، بنابراین نرمال‌سازی مانتیس در این معادله متفاوت از معادله‌ی (۷.۲) است. این بیت (که همواره برابر ۱ است) ذخیره نمی‌شود و به آن *بیت پنهان* می‌گویند. به همین علت است که برای نمایش مانتیس یک بیت کمتر استفاده می‌شود و آن بیت برای نمایش نما مورد استفاده قرار می‌گیرد. برای نمایش نما از روش مکمل دو، که برای ذخیره اعداد صحیح علامتدار استفاده می‌شود، استفاده نمی‌شود بلکه یک *توان اُریبی* برای نمایش نما در نظر گرفته می‌شود و نما با این توان جمع شده و به‌صورت یک عدد صحیح بدون علامت ذخیره می‌شود. با مقایسه‌ی فرمول‌های (۷.۲) و (۱۴.۲) در دقت معمولی $e_{\min} = -۱۲۶$ و $e_{\max} = ۱۲۷$ و همچنین توان اریبی برابر $۱۲۷ = ۱ - ۲^{۸-۱}$ در نظر گرفته می‌شود و از این رو به‌جای نمای e، عدد صحیح بدون علامت $e + ۱۲۷$ ذخیره می‌شود. به عنوان مثال اگر نما برابر -۲۰ باشد به‌جای آن عدد ۱۰۷ در ۸ بیت (در نظر گرفته شده برای نما) به‌صورت دودویی ذخیره می‌شود. در دقت دوبرابر $e_{\min} = -۱۰۲۲$ و $e_{\max} = ۱۰۲۳$ و توان اریبی برابر $۱۰۲۳ = ۱ - ۲^{۱۱-۱}$ می‌باشد. بزرگترین عدد قابل نمایش در دستگاه با دقت معمولی $۲/۰ \times ۲^{۱۲۷} \doteq ۳/۴۰۲۸ \times ۱۰^{۳۸}$ و بزرگترین عدد قابل نمایش در دستگاه با دقت دوبرابر $۲/۰ \times ۲^{۱۰۲۳} \doteq ۱/۷۹۷۷ \times ۱۰^{۳۰۸}$ است. همچنین کوچکترین اعداد مثبت نرمال قابل نمایش $۱/۱۷۵۵ \times ۱۰^{-۳۸} \doteq ۱/۰ \times ۲^{-۱۲۶}$ و $۲/۲۲۵۱ \times ۱۰^{-۳۰۸} \doteq ۱/۰ \times ۲^{-۱۰۲۲}$ به ترتیب در دقت‌های معمولی و دوبرابر هستند. این دستگاه‌ها برای نمایش اعداد تمامی نگاشت‌های گردکردن را پشتیبانی می‌کنند اما پیش‌فرض آن‌ها گردکردن به نزدیکترین است، از این‌رو واحد گردکردن برابر

$$u = \begin{cases} ۲^{-۲۴} \approx ۵/۹۶ \times ۱۰^{-۸}, & \text{در دقت معمولی} \\ ۲^{-۵۳} \approx ۱/۱۱ \times ۱۰^{-۱۶}, & \text{در دقت دوبرابر} \end{cases}.$$

است که مقدار آن در هر مورد نصف اپسیلون ماشین است.

یک عدد زیرنرمال با نمایش

$$\overline{a} = \pm(۰/m)_۲ \times ۲^{e_{\min}}, \tag{۱۵.۲}$$

به صورت زیر ذخیره می‌شود: بجای $e_{\min}$ نمای $e = e_{\min} - ۱$ ذخیره می‌شود، اما مانتیس $m \neq ۰$ بدون تغییر ذخیره می‌شود. برای مثال در دقت معمولی بجای -۱۲۶ در نما عدد -۱۲۷ ذخیره می‌شود. به عبارت دیگر می‌توان گفت در دقت معمولی اگر نمای ذخیره شده -۱۲۷ و مانتیس غیر صفر باشد، آن عدد یک عدد زیرنرمال است. کوچکترین عدد مثبت زیرنرمال قابل نمایش با دقت معمولی عبارتست از $(۰/۰۰\cdots۰۱)_۲ ۲^{-۱۲۶} = ۲^{-۲۳-۱۲۶} \doteq ۱/۴۰۱۳ \times ۱۰^{-۴۵}$ و با دقت دوبرابر عبارتست از $۲^{-۵۲-۱۰۲۲} \doteq ۴/۹۴۰۷ \times ۱۰^{-۳۲۴}$.

برای نمایش ±۰، نما برابر $e_{\min} - ۱$ و مانتیس برابر صفر است. بیت علامت نیز صفر مثبت و صفر منفی را از هم متمایز می‌کند. یک استفاده مهم از صفر علامتدار تشخیص پی‌ریز مثبت و پی‌ریز منفی است و استفاده‌ی دیگر آن در انجام محاسبات با توابع مختلط ظاهر می‌شود.

جدول ۱.۲: نحوه‌ی نمایش اعداد در استاندارد IEEE 754

نما	مانتیس	نمایش
$e = e_{\min} - ۱$	$m = ۰$	± ۰
$e = e_{\min} - ۱$	$m \neq ۰$	$\pm(۰/m)_۲ \times ۲^{e_{\min}}$
$e_{\min} \leqslant e \leqslant e_{\max}$		$\pm(۱/m)_۲ \times ۲^{e}$
$e_{\max} + ۱$	$m = ۰$	$\pm\infty$
$e_{\max} + ۱$	$m \neq ۰$	NaN

جدول ۲.۲: فرمت‌های مختلف در استاندارد IEEE

انواع فرمت‌ها	بیت مورد نیاز	t	e	$e_{\min}$	$e_{\max}$
معمولی	۳۲ بیت	$۲۳ + ۱$	۸ بیت	-۱۲۶	۱۲۷
دوبرابر	۶۴ بیت	$۵۲ + ۱$	۱۱ بیت	-۱۰۲۲	۱۰۲۳

برای نمایش $\pm\infty$، نما برابر $e_{\max} + ۱$ و مانتیس برابر صفر است. این مقدار در حالتی که $\frac{a}{۰}$ با $a \neq ۰$ داشته باشیم برگشت داده می‌شود. مثلاً در نرم‌افزار متلب با علامت Inf نمایش داده می‌شود. همچنین این نماد از قراردادهای ریاضی نظیر $\infty + \infty = \infty$ و $-\infty \times (-۱) = \infty$ و $\frac{a}{\infty} = ۰$ نیز تبعیت می‌کند. استاندارد IEEE بسیار به واقعیت ریاضی نزدیک است تا آنجا که حتی محاسبات خاص از قبیل $\frac{۰}{۰}$، $(-\infty) + \infty$+، $\infty \times ۰$ و $\sqrt{-۱}$ را نیز پشتیبانی می‌کند. هر یک از این حالات منجر به نمایش یک NaN یا به‌صورت مفصل "Not a Number" می‌شوند که برای ذخیره‌ی آن $e = e_{\max} + ۱$ و $m \neq ۰$ در نظر گرفته می‌شود. نکته اینکه به‌ازای هر $m \neq ۰$ یک NaN داریم. اگر یک NaN و یک عدد ممیز شناور معمولی با چهار عمل اصلی ترکیب شوند حاصل حتماً NaN خواهد بود. همچنین برای متغیری که مقداردهی اولیه نشده یا گاهی برای داده‌های تعریف نشده نیز یک NaN به‌صورت پیش‌فرض در نظر گرفته می‌شود.

در جدول ۱.۲ مقادیر نما و مانتیس برای نمایش انواع حالات در استاندارد IEEE آمده است. استاندارد IEEE همچنین دو فرمت با *دقت توسعه‌یافته* را نیز پشتیبانی می‌کند که در اینجا به آن نمی‌پردازیم. در جدول ۲.۲ مقایسه‌ای بین دو فرمت معمولی و دوبرابر نشان شده است. بیت پنهان به‌صورت $+۱$ در هر حالت بیان شده است.

مثال ۲.۲. می‌خواهیم عدد $+۸۹/۷۵$ را در استاندارد IEEE با دقت معمولی ذخیره کنیم. داریم $(۱۰۱۱۰۰۱)_۲ = ۸۹$ و نیز برای قسمت اعشار این عدد داریم $(۰/۱۱)_۲ = ۰/۷۵$. بنابراین می‌توان نوشت

$$+۸۹/۷۵ = (۱۰۱۱۰۰۱/۱۱)_۲ = (۱/۰۱۱۰۰۱۱۱۱)_۲ \times ۲^{۶},$$

از این‌رو نما $e = ۶$ و مانتیس $m = ۰/۰۱۱۰۰۱۱۱۱$ هستند. مانتیس به همین صورت ذخیره می‌شود ولی به‌جای نما $(۱۰۰۰۰۱۰۱)_۲ = ۱۳۳ = e + ۱۲۷$ در هشت بیت ذخیره می‌شود. چون علامت عدد مثبت است مقدار بیت علامت ۰

خواهد بود. پس داریم:

$$0\big|10000101\big|01100111000000000000000$$

توجه کنید که ۱۵ بیت سمت راست مانتیس صفر هستند. گاهی قسمت اعشار عدد دهدهی نامختوم است، یا حتی مختوم است ولی نمایش آن در مبنای دو نامختوم است یا اینکه بعد از مرتب کردن به‌صورت بالا طول مانتیس بیشتر از ۲۳ بیت است، در این صورت بایستی ۲۳ بیت ابتدایی آن را ذخیره کرد، که قطعاً منجر به خطای گرد کردن می‌شود.

حال می‌خواهیم ببینیم ۳۲ بیت زیر نمایش چه عددی هستند.

$$1\big|01011001\big|01110100000000000000000$$

بیت‌های بالا یک عدد را در دقت معمولی نمایش می‌دهند. با توجه به اینکه بیت علامت ۱ می‌باشد پس عدد مورد نظر منفی است. هشت بیت نما نمایش دهنده‌ی عدد $۸۹ = ۲(۰۱۰۱۱۰۰۱)_۲$ می‌باشد. با توجه به مقدار توان اریبی، نمای عدد مورد نظر $e = ۸۹ - ۱۲۷ = -۳۸$ است. با توجه به مقدار e عدد بالا یک عدد ممیز شناور نرمال خواهد بود. مقدار مانتیس برابر است با $۲(۰/۰۱۱۱۰۱۰۰...۰)_۲ = ۲(۱/۰۱۱۱۰۱)_۲ + ۱$. بنابراین عدد مورد نظر

$$\begin{aligned}
-(۱/۰۱۱۱۰۱)_۲ \times ۲^{-۳۸} &= -(۰/\underbrace{۰۰\cdots۰}_{\text{۳۷تا}}۱۰۱۱۱۰۱)_۲ \\
&= -\left(۲^{-۳۸} + ۲^{-۴۰} + ۲^{-۴۱} + ۲^{-۴۲} + ۲^{-۴۴}\right) \\
&= -۲^{-۳۸}\left(۱ + ۲^{-۲} + ۲^{-۳} + ۲^{-۴} + ۲^{-۶}\right) \\
&= -۱/۴۵۳۱۲۵ \times ۲^{-۳۸} \\
&\doteq -۰/۵۲۸۶۴۳۷۹ \times ۱۰^{-۱۱}.
\end{aligned}$$

مثال بعد مربوط به یک عدد زیرنرمال است. می‌خواهیم ببینیم جدول زیر نمایش چه‌عددی است.

$$1\big|00000000\big|01110100000000000000000$$

علامت عدد منفی است. با توجه به مقدار نمای ذخیره شده که صفر است در می‌یابیم نمای عدد مورد نظر $e = ۰ - ۱۲۷ = -۱۲۷$ است که برابر $e_{\min} - ۱$ است. چون مانتیس غیرصفر است عدد مورد نظر عددی زیرنرمال است. (اگر مانتیس صفر بود عدد مورد نظر -۰ می‌بود). با توجه به رابطه‌ی (۱۵.۲) این عدد برابر است با

$$\begin{aligned}
-(۰/۰۱۱۱۰۱)_۲ \times ۲^{-۱۲۶} &= -(۰/\underbrace{۰۰\cdots۰}_{\text{۱۲۷تا}}۱۱۱۰۱)_۲ \\
&= -۲^{-۱۲۸} + ۲^{-۱۲۹} + ۲^{-۱۳۰} + ۲^{-۱۳۲} \\
&= -۲^{-۱۲۸}\left(۱ + ۲^{-۱} + ۲^{-۲} + ۲^{-۴}\right) \\
&= -۱/۸۱۲۵ \times ۲^{-۱۲۸} \\
&\doteq -۰/۵۳۲۶۴۵۸۸ \times ۱۰^{-۴۰}.
\end{aligned}$$

۳.۲ ارقام بامعنا

در اینجا دوباره به خطای اعداد بر می‌گردیم. خطای نسبی در اعداد رابطه‌ی نزدیکی با آنچه *ارقام بامعنای درست* گفته می‌شود، دارد که به شرح آن خواهیم پرداخت. *ارقام بامعنای* یک عدد اولین رقم غیر صفر (از سمت چپ) و ارقام بعد از آن (صفر و غیرصفر) می‌باشند. با این تعریف ۱٫۲۰۳۰ دارای پنج و ۰٫۰۰۱۲۳ دارای سه رقم بامعناست. گیریم $\widehat{x}$ تقریبی از x باشد، گوییم $\widehat{x}$ تا n رقم بامعنا با x یکی است اگر n رقم بامعنای اولیه‌ی آن‌ها یکی باشد. برای مثال دو دسته اعداد زیر را درنظر بگیرید:

$$\text{(الف)} \quad x = ۱٫۰۰۰۰۰, \quad \widehat{x} = ۱٫۰۰۴۹۹, \quad \frac{|x-\widehat{x}|}{|x|} \doteq ۴٫۹۹ \times ۱۰^{-۳},$$

$$\text{(ب)} \quad x = ۹٫۰۰۰۰۰, \quad \widehat{x} = ۸٫۹۹۸۹۹, \quad \frac{|x-\widehat{x}|}{|x|} \doteq ۱٫۱۲ \times ۱۰^{-۴}.$$

در دسته‌ی اول $\widehat{x}$ و x تا سه رقم بامعنا یکی هستند و دسته‌ی دوم هیچ رقم بامعنای یکسانی ندارند، در حالی که خطای نسبی تقریب دوم حدود ۴۴ بار بهتر از تقریب اول است. پس با همین مثال ساده نتیجه می‌گیریم تعداد ارقام بامعنا معیار مناسبی برای سنجش دقت یک تقریب نیست. به‌جای ارقام بامعنا، معمولاً "ارقام بامعنای درست" یک تقریب را می‌توان بکار برد. در کتاب‌ها تعریف یکسانی از این مفهوم وجود ندارد. می‌توان آن را به صورت زیر بیان کرد: تقریب $\widehat{x}$ از x دارای t رقم بامعنای درست است اگر $\widehat{x}$ و x به یک عدد سوم با t رقم بامعنا گرد (به نزدیکترین) شوند. این تعریف ارقام بامعنای درست اغلب مفید و از لحاظ شهودی درست است اما دو عدد زیر را در نظر بگیرید:

$$\text{(پ)} \quad x = ۰٫۹۹۴۹, \quad \widehat{x} = ۰٫۹۹۵۱,$$

با توجه به تعریف بالا $\widehat{x}$ دارای دو رقم بامعنای درست نیست چراکه تا دو رقم $x \to ۰٫۹۹$ درحالی‌که $\widehat{x} \to ۱٫۰$. اما $\widehat{x}$ دارای یک رقم بامعنای درست و نیز سه رقم بامعنای درست است! بنابراین همواره بزرگترین عدد طبیعی t که خاصیت بالا را برقرار سازد مد نظر خواهد بود. یک تعریف و یک فرمول برای محاسبه تعداد ارقام بامعنای درست به صورت زیر است.

تعریف ۳.۲. گیریم عدد ناصفر x دارای نمایش ممیز شناور زیر در مبنای β باشد:

$$x = \pm(۰/d_۱d_۲\cdots)_\beta \times \beta^e, \quad d_k \in \{۰, ۱, \ldots, \beta-۱\}, \quad d_۱ \neq ۰, \quad e \in \mathbb{Z}, \qquad (۱۶.۲)$$

و $\widehat{x}$ تقریبی از x باشد. در این صورت بزرگترین عدد صحیح نامنفی t که بازای آن داشته باشیم

$$\frac{|\widehat{x}-x|}{\beta^{e-۱}} \leqslant \frac{۱}{۲} \times \beta^{۱-t}, \qquad (۱۷.۲)$$

را تعداد ارقام بامعنای درست $\widehat{x}$ تعریف می‌کنیم.

تقسیم بر $\beta^{e-۱}$ در حقیقت خطای تقریب را به نوعی به شکل نسبی در می‌آورد. در گزاره‌ی زیر ارتباط بین خطای نسبی یک تقریب و تعداد ارقام با معنای درست آمده است.

گزاره ۱.۲. *فرض کنیم عدد ناصفر x دارای نمایش (۱۶.۲) در مبنای β باشد. گیریم یک تقریب $\widehat{x}$ از x دارای t رقم بامعنای درست باشد. در این صورت کران خطای نسبی*

$$\frac{|\widehat{x}-x|}{|x|} \leqslant \frac{۱}{۲} \times \beta^{۱-t}$$

برقرار خواهد بود.

برهان. با توجه به اینکه $|x| \geqslant \beta^{e-۱}$ داریم

$$\frac{|\widehat{x}-x|}{|x|} \leqslant \frac{|\widehat{x}-x|}{\beta^{e-۱}}$$

و با توجه به (۱۷.۲) کران خطای نسبی برقرار خواهد بود. □

گزاره بالا ارتباط خطای نسبی و تعداد ارقام با معنای درست را بیان می‌کند. حال چند مثال ارائه می‌دهیم.

مثال ۳.۲. فرض کنیم $\beta = ۱۰$. در این مبنا کران موجود در (۱۷.۲) به صورت ۵×۱۰^{-t} در می‌آید. بنابراین در دسته‌ی (پ) در بالا، $\widehat{x}$ دارای سه رقم بامعنای درست است، چراکه برای عدد ۰/۹۹۴۹ داریم $e = ۰$ و

$$\frac{|\widehat{x}-x|}{۱۰^{-۱}} = ۰/۰۰۲ \leqslant ۵ \times ۱۰^{-۳}.$$

به عنوان یک مثال دیگر فرض کنید

$$(ت) \quad x = ۴۵/۷۲۴۶۳, \quad \widehat{x} = ۴۵/۷۲۸۱۳.$$

داریم $e = ۲$ و

$$\frac{|\widehat{x}-x|}{۱۰^{۱}} = ۰/۰۰۰۳۵ \leqslant ۵ \times ۱۰^{-۴},$$

که نشان می‌دهد این دو عدد دارای ۴ رقم بامعنای درست یکسان هستند. به عنوان یک مثال دیگر دو عدد

$$(ث) \quad x = ۰/۰۰۰۴۰۰, \quad \widehat{x} = ۰/۰۰۰۳۹۸$$

را در نظر بگیرید. داریم $e = -۳$ و بنابراین

$$\frac{|x-\widehat{x}|}{۱۰^{-۴}} = \frac{۰/۰۰۰۰۰۲}{۰/۰۰۰۱} = ۰/۰۲ \leqslant ۵ \times ۱۰^{-۲}$$

که نشان می‌دهد $\widehat{x}$ و x دارای دو رقم بامعنای درست یکسان هستند. برای مقایسه، این‌بار دو عدد

$$(ج) \quad x = ۰/۴۰۰, \quad \widehat{x} = ۰/۳۹۸$$

را در نظر بگیرید. در این حالت $e = ۰$ و خواهیم داشت

$$\frac{|x - \widehat{x}|}{۱۰^{-۱}} = \frac{۰/۰۰۲}{۰/۱} = ۰/۰۲ \leqslant ۵ \times ۱۰^{-۲}$$

که همانند قبل نشان می‌دهد $\widehat{x}$ و x دارای دو رقم بامعنای درست یکسان هستند. در حقیقت صفرهای سمت چپ به عنوان ارقام بامعنای درست شمرده نمی‌شوند. ♢

همانطور که گفته شد تعداد ارقام بامعنا معیار مناسبی برای سنجش دقت یک تقریب نیست. تعداد ارقام بامعنای درست معیار بسیار بهتری است اما گاهی به خوبیِ خطای نسبی نیست. برای مثال همانطور که گفته شد خطای نسبی در دسته‌ی (ب) در بالا حدود ۴۴ بار بهتر از خطای نسبی دسته‌ی (الف) است، با این حال تعداد ارقام با معنای درست برای هر دو دسته برابر ۳ است (محاسبه کنید).

از این پس هر گاه می‌گوییم "ارقام بامعنا" منظورمان "ارقام بامعنای درست" است.

۴.۲ آنالیز خطاهای گردکردن

برای یک روش گاهی چندین الگوریتم برای رسیدن به جواب تقریبی وجود دارد که همگی از دید ریاضی معادل یکدیگرند. اما این الگوریتم‌ها لزوماً از دید عددی با هم معادل نیستند و وقتی روی کامپیوتر اجرا می‌شوند گاهی جواب‌های کاملاً متفاوتی ارائه می‌دهند. به مثال زیر توجه کنید:

مثال ۴.۲. روش ارشمیدس برای محاسبه‌ی عدد اصم π دنباله‌ی تکرار زیر را با استفاده از تقریب محیط یک دایره با محیط چندضلعی‌های منتظم، تولید می‌کند:

$$p_{k+۱} = ۲^{k+۱}\sqrt{\frac{۱}{۲}\big(۱ - \sqrt{۱ - [۲^{-k}p_k]^۲}\big)}, \quad k = ۱, ۲, \ldots, \quad p_۱ = ۲. \tag{۱۸.۲}$$

از دید نظری با افزایش k مقادیر p_k به عدد π میل می‌کنند. یک برنامه ساده با متلب به صورت زیر می‌نویسیم و نتایج حاصل را در ستون (I) جدول ۳.۲ ارائه می‌دهیم.

```
p(1) = 2;
for k=1:28
    p(k+1) = 2^(k+1)*sqrt(1/2*(1-sqrt(1-(2^(-k)*p(k))^2)));
end
```

جدول ۳.۲: محاسبه‌ی π

k	(I)	(II)
۱	۲/۰۰۰۰۰۰۰۰۰۰۰۰۰۰	۲/۰۰۰۰۰۰۰۰۰۰۰۰۰۰
۲	۲/۸۲۸۴۲۷۱۲۴۷۴۶۱۹۰	۲/۸۲۸۴۲۷۱۲۴۷۴۶۱۹۰
۳	۳/۰۶۱۴۶۷۴۵۸۹۲۰۷۱۹	۳/۰۶۱۴۶۷۴۵۸۹۲۰۷۱۹
⋮	⋮	⋮
۱۰	۳/۱۴۱۵۸۷۷۲۵۲۷۹۹۶۱	۳/۱۴۱۵۸۷۷۲۵۲۷۷۱۶۰
۱۱	۳/۱۴۱۵۹۱۴۲۱۵۰۴۶۳۵	۳/۱۴۱۵۹۱۴۲۱۵۱۱۲۰۰
⋮	⋮	⋮
۲۷	۳/۱۶۲۲۷۷۶۶۰۱۶۸۳۸۰	۳/۱۴۱۵۹۲۶۵۳۵۸۹۷۹۳
۲۸	۳/۴۶۴۱۰۱۶۱۵۱۳۷۷۵۴	۳/۱۴۱۵۹۲۶۵۳۵۸۹۷۹۳
۲۹	۴/۰۰۰۰۰۰۰۰۰۰۰۰۰۰	۳/۱۴۱۵۹۲۶۵۳۵۸۹۷۹۳

آنچه مشاهده می‌کنیم خطای بسیار زیاد در محاسبات است تا آنجا که $p_{۲۹}$ برابر ۴/۰ شده است. عدد π تا پانزده رقم اعشار عبارتست از $\pi = ۳/۱۴۱۵۹۲۶۵۳۵۸۹۷۹۳\ldots$. حال یک تغییر کوچک در فرمول (۱۸.۲) اعمال می‌کنیم. با توجه به اتحاد $۱ - x = \frac{۱-x^۲}{۱+x}$ و با فرض اینکه $x := \sqrt{۱ - [۲^{-k}p_k]^۲}$ داریم

$$p_{k+۱} = p_k \sqrt{\frac{۲}{۱ + \sqrt{۱ - [۲^{-k}p_k]^۲}}}, \quad k = ۱, ۲, \ldots, \quad p_۱ = ۲. \tag{۱۹.۲}$$

برنامه متلب آن نیز به صورت زیر است

```
p(1) = 2;
for k=1:28
    p(k+1) = p(k)*sqrt(2/(1+sqrt(1-(2^(-k)*p(k))^2)));
end
```

نتایج در ستون (II) جدول ۳.۲ ارائه شده‌اند. مشاهده می‌کنیم فرمول بازگشتی (۱۹.۲) عدد π را به خوبی تقریب می‌زند. به این نکته توجه داریم که هر دو الگوریتم از دید نظری با هم معادلند، اما نتایج عددی متفاوتی ارائه می‌دهند. در واقع می‌توان گفت این دو الگوریتم از دید عددی با هم معادل نیستند، چرا که با ورودی یکسان، خروجی‌های متفاوتی تولید می‌کنند. آنچه در مورد الگوریتم اول در این مثال رخ داده است، خطای حذف ارقام بامعناست که در ادامه این بخش به آن خواهیم رسید.

مثال بالا نشان می‌دهد در روند هر الگوریتم اتفاقات مهمی رخ می‌دهد که باید به درستی مورد واکاوی قرار گیرند. برای این کار باید بتوانیم تک تک محاسبات در کامپیوتر از چهار عمل اصلی گرفته تا محاسبات پیچیده‌تر مبتنی بر محاسبات ماتریسی را آنالیز کنیم.

۱.۴.۲ مدل‌ها و مقدمات آنالیز خطا

در این بخش می‌خواهیم ببینیم چهار عمل اصلی ریاضی (جمع، تفریق، ضرب و تقسیم) به چه سبکی انجام می‌شوند و خطای تولید شده از عمل این عملگرها روی دو یا چند عدد چگونه است. اگر x و y دو عدد ممیز شناور باشند ممکن است به عنوان مثال حاصل‌ضرب آن‌ها یک عدد ممیز شناور نبوده و لازم باشد دوباره گرد شود. بنابراین برای دو عدد ممیز شناور x و y

$$fl(x+y), \quad fl(x-y), \quad fl(x\times y), \quad fl(x\div y),$$

مقادیر ذخیره شده در حافظه‌ی کامپیوتر به عنوانِ حاصل‌جمع، حاصل‌تفریق، حاصل‌ضرب و حاصل تقسیم با نگاشت گردکردن یا بریدن (یا هر نگاشت دیگر) در نظر گرفته می‌شوند. مقادیر بالا را به ترتیب جمع، تفریق، ضرب و تقسیم ماشینی دو عدد x و y می‌گوییم. بنابراین عملگرهایی که در ماشین تعریف می‌شوند با عملگرهای ریاضی متفاوت هستند، یعنی در مقابل جمع ریاضی جمع ماشینی و در مقابل ضرب ریاضی ضرب ماشینی و غیره داریم. واضح است اگر x و y اعداد ممیز شناور نباشند، قبل از اعمال جمع ماشینی بایستی گرد شوند. به عنوان مثال، عددی که به عنوان حاصل‌جمع در حافظه ذخیره می شود $fl(fl(x)+fl(y))$ خواهد بود. در اینجا فرض می‌کنیم x و y عضو مجموعه اعداد ممیز شناور $\mathbb{F}$ باشند و نیز فرض را بر این می‌گذاریم که فعلاً خطاهای سرریز و پی‌ریز رخ ندهند (البته اگر از استاندارد IEEE 754 استفاده شود عملاً هیچگاه پی‌ریز رخ نمی‌دهد و می‌توان این فرض را حذف کرد). اگر $*$ نماینده‌ی یکی از عملگرهای ریاضی باشد، با توجه به رابطه‌ی (۱۲.۲) داریم:

$$fl(x*y)=(x*y)(1+\varepsilon), \quad |\varepsilon|\leqslant u, \quad *\in\{+,-,\times,\div\}. \tag{۲۰.۲}$$

قابل ذکر است که در استاندارد IEEE این امکان فراهم شده است که تابع جذر دوم نیز با دقت ماشین محاسبه شود. در واقع داریم

$$fl(\sqrt{x})=\sqrt{x}(1+\varepsilon), \quad |\varepsilon|\leqslant u. \tag{۲۱.۲}$$

گاهی حاصل عملگر ماشینی روی دو عدد دقیق‌تر از آنچه ما از قبل پیش‌بینی کرده‌ایم است، برای مثال اگر حاصل $x*y$ خود نیز یک عدد ممیز شناور باشد خطای محاسباتی نخواهیم داشت و در حقیقت عملگر ماشینی با عملگر ریاضی یکی خواهد بود. عملگرهای ماشینی دارای خواص دیگری نیز هستند که آن‌ها را از عملگرهای ریاضی جدا می‌کنند، مثلاً برخلاف عملگرهای ریاضی برای عملگرهای ماشینی خواص شرکت‌پذیری و توزیع‌پذیری همواره برقرار نیستند. در زیر مثالی خواهیم

آورد که در آن برای سه عدد ماشینیِ x، y و z ، رابطه‌ی

$$fl(x + fl(y + z)) = fl(fl(x + y) + z)$$

(خاصیت شرکت‌پذیری) برقرار نخواهد بود.

مثال ۵.۲. در دستگاه دهدهی ممیز شناور با طول مانتیس $t = ۷$ سه عدد زیر را درنظر بگیرید

$$x = ۰/۱۲۳۴۵۶۷ \times ۱۰^{۰}, \quad y = ۰/۴۷۱۱۳۲۵ \times ۱۰^{۴}, \quad z = -y.$$

از این‌رو داریم

$$fl(y + z) = ۰, \quad fl(x + fl(y + z)) = x = ۰/۱۲۳۴۵۶۷ \times ۱۰^{۰},$$

از طرف دیگر جدول جمع‌بندی زیر را در نظر بگیرید

$x = \quad 0.0000123$	4567×10^4
$y = \quad 0.4711325$	$\times 10^4$
$fl(x+y) = \quad 0.4711448$	$\times 10^4$
$z = -0.4711325$	$\times 10^4$

که در سطر اول آن به خاطر اینکه بایستی ممیزها زیر هم نوشته شده و نماها یکسان شوند (این عمل را مقیاس کردن گوییم)، چهار رقم آخر بعد از اعشار به بیرون انتقال یافته و خاصیت خود را از دست داده‌اند. با توجه به جدول داریم

$$fl(fl(x + y) + z) = ۰/۰۰۰۰۱۲۳ \times ۱۰^{۴} = ۰/۱۲۳۰۰۰۰ \times ۱۰^{۰} \neq fl(x + fl(y + z)).$$

در مورد تفاوت عملگرهای ماشینی و ریاضی آوردن مثال‌هایی شبیه مثال ۵.۲ برای نشان دادن عدم برقراری خاصیت توزیع‌پذیری جمع ماشینی روی ضرب ماشینی و برعکس، و همچنین عدم وجود عضو خنثای یکتای جمع ماشینی و عضو خنثای یکتای ضرب ماشینی مشکل نخواهد بود. البته واضح است که خاصیت جابجایی برای جمع و ضرب ماشینی برقرار است.

مدل (۲۰.۲) برای حالت تفریق زمانی برقرار است که محاسبات به کمک بیتِ *پشتیبان* انجام شود. با یک مثال ساده نقش بیتِ پشتیبان را در عمل تفریق نشان می‌دهیم: یک دستگاه ممیز شناور با $\beta = ۲$ و $t = ۳$ را درنظر بگیرید. می‌خواهیم عدد ممیز شناور بلافاصله قبل از ۱/۰ را از ۱/۰ کم کنیم. داریم

$$\begin{array}{rcl} 0.100 \times 2^1 & & 0.100 \quad \times 2^1 \\ -0.111 \times 2^0 & \longrightarrow & -0.011\underline{1} \times 2^1 \\ \hline & & 0.0001 \times 2^1 = 0.100 \times 2^{-2} \end{array}$$

نکته اینکه برای انجام عمل تفریق بایستی ابتدا مقیاس‌سازی کنیم، بنابراین عدد دوم یک رقم چهارم در مانتیس نیاز دارد. به این رقم بیت پشتیبان می‌گوییم. ماشین‌های قدیمی از بیت پشتیبان بهره نمی‌بردند. بدون بیت پشتیبان تفریق بالا به‌صورت زیر در ماشین انجام می‌شود

$$\begin{array}{rcl} 0.100 \times 2^1 & & 0.100 \times 2^1 \\ -0.111 \times 2^0 & \longrightarrow & -0.011 \times 2^1 \\ \hline & & 0.001 \times 2^1 = 0.100 \times 2^{-1} \end{array}$$

جواب محاسبه شده در بالا دوبرابر جواب قبل است. یعنی دارای خطای نسبی ۱۰۰٪ است. قابل ذکر است که برای ماشینی که بیت پشتیبان استفاده نکند، مدل (۲۰.۲) در حالت جمع و تفریق برقرار نخواهد بود. کمبود بیت پشتیبان یک مشکل کاملاً جدی است، خوشبختانه ماشین‌های امروزی همگی از بیت پشتیبان برای عمل تفریق استفاده می‌کنند و از این رو در محاسبات امروزی مدل (۲۰.۲) همواره برقرار است.

مثال ۶.۲. گیریم $x, y, z \in \mathbb{F}$ و فرض کنیم ماشین از استاندارد IEEE با دقت معمولی و نگاشت گردکردن استفاده می‌کند. برای معادل ماشینیِ عبارت $x \times (y+z)$ طبق (۲۰.۲) داریم:

$$\begin{aligned} fl(x \times fl(y+z)) &= fl(x \times [(y+z)(۱+\varepsilon_۱)]) \\ &= [x \times (y+z)(۱+\varepsilon_۱)](۱+\varepsilon_۲) \\ &= x \times (y+z)(۱+\varepsilon_۱+\varepsilon_۲+\varepsilon_۱\varepsilon_۲) \end{aligned}$$

که در آن $|\varepsilon_k| \leqslant ۲^{-۲۴},\ k = ۱, ۲$. در عمل با توجه به اینکه اندازه‌ی همه‌ی ε_k ها با یک مقدار کران‌دار شده است می‌توان تفاوت آن‌ها را فقط در علامت در نظر گرفت و برای هر k قرار داد $۱ + \varepsilon_k \equiv ۱ \pm \varepsilon$. از این‌رو رابطه‌ی بالا در بدبینانه‌ترین حالت به‌صورت

$$fl(x \times fl(y+z)) = x \times (y+z)(۱ \pm ۲\varepsilon \pm \varepsilon^۲), \quad |\varepsilon| \leqslant ۲^{-۲۴}$$

خواهد بود. در رابطه‌ی بالا می‌توان از $\varepsilon^۲$ که کمتر از $۲^{-۴۸}$ است چشم‌پوشی کرد و نوشت

$$fl(x \times fl(y+z)) \approx x \times (y+z)(۱+\eta), \quad |\eta| \leqslant ۲^{-۲۳} \doteq ۱/۱۹۲ \times ۱۰^{-۷}.$$

اگر اعداد x، y یا z خود نیز ممیز شناور نباشند محاسبه بالا دارای پیچیدگی بیشتر و در نهایت دقت کمتری خواهد بود.

برای ساده سازی آنالیزهای خطای گردکردن، لم زیر یک روش بسیار زیبا و کاربردی ارائه می‌دهد که در آن نماد جدید γ_n معرفی شده است [۲، ۴].

لم ۲.۲. *اگر برای* $k = ۱, \dots, n$ *داشته باشیم* $|\varepsilon_k| \leqslant u$ *و* $\rho_k = \pm ۱$ *و همچنین* $nu < ۱$ *آنگاه*

$$\prod_{k=۱}^{n} (۱ + \varepsilon_k)^{\rho_k} = ۱ + \theta_n,$$

به‌طوریکه

$$|\theta_n| \leqslant \frac{nu}{۱ - nu} =: \gamma_n. \tag{۲۲.۲}$$

برهان. اثبات با استقرا صورت می‌گیرد. برای $n = ۱$ حکم بسادگی برقرار است. فرض کنیم $\rho_n = +۱$. داریم

$$\begin{aligned}
\prod_{k=۱}^{n} (۱ + \varepsilon_k)^{\rho_k} &= (۱ + \varepsilon_n)(۱ + \theta_{n-۱}) = ۱ + \theta_n, \\
\theta_n &= \varepsilon_n + (۱ + \varepsilon_n)\theta_{n-۱}, \\
|\theta_n| &\leqslant u + (۱ + u)\frac{(n-۱)u}{۱ - (n-۱)u} \\
&= \frac{u(۱ - (n-۱)u) + (۱ + u)(n-۱)u}{۱ - (n-۱)u} \\
&= \frac{nu}{۱ - (n-۱)u} \leqslant \gamma_n,
\end{aligned}$$

و برای $\rho_n = -۱$ بطور مشابه خواهیم داشت

$$\begin{aligned}
\prod_{k=۱}^{n} (۱ + \varepsilon_k)^{\rho_k} &= (۱ + \varepsilon_n)^{-۱}(۱ + \theta_{n-۱}) = ۱ + \theta_n, \\
\theta_n &= \frac{\theta_{n-۱} - \varepsilon_n}{۱ + \varepsilon_n}, \\
|\theta_n| &\leqslant \frac{|\theta_{n-۱}| + u}{۱ - u} \leqslant \frac{nu - (n-۱)u^۲}{۱ - nu + (n-۱)u^۲} \leqslant \gamma_n.
\end{aligned}$$

□

ملاحظه ۲.۲. برای حالتی که $\rho_k = ۱, \forall k$ و با فرض اینکه $nu < ۲$ کران قوی‌تر

$$|\theta_n| \leqslant \frac{nu}{۱ - nu/۲}$$

برقرار است (پرسش ۱۱ را ببینید). اما برای یک‌دست شدن محاسبات و با توجه به اختلاف ناچیز این دو کران همواره (۲۲.۲) را مبنای آنالیزهای خود قرار می‌دهیم.

مثال ۷.۲. با توجه به تعریف نماد جدید γ_n برای مثال ۶.۲ داریم:

$$fl(x \times fl(y+z)) = x \times (y+z)(۱+\varepsilon_۱)(۱+\varepsilon_۲) = x \times (y+z)(۱+\theta_۲),$$
$$|\theta_۲| \leqslant \gamma_۲ = \frac{۲^{-۲۳}}{۱ - ۲^{-۲۳}} \doteq ۱/۱۹۲ \times ۱۰^{-۷}.$$

از این پس برای رعایت اختصار در نوشتن، عملگر $\times$ را حذف و به‌جای عباراتی به‌فرم $x \times y$ از xy استفاده کنیم. در ادامه سعی می‌کنیم برای چند الگوریتم ساده، آنالیز خطا ارائه دهیم.

۲.۴.۲ الگوریتم ضرب کردن

در اینجا با یک مثال ساده شروع می‌کنیم. فرض کنیم حاصلضرب دو عدد ممیز شناور غیر صفر x و y مدنظر است که آن را با s نمایش می‌دهیم، یعنی قرار می‌دهیم $s = xy$. فرض کنیم آنچه ماشین محاسبه می‌کند $\widehat{s}$ باشد که با توجه به رابطه‌ی (۲۰.۲) برابر است با

$$\widehat{s} = fl(xy) = xy(۱+\varepsilon), \quad |\varepsilon| \leqslant u. \tag{۲۳.۲}$$

این رابطه بیانگر این است که مقدار محاسبه شده‌ی $\widehat{s}$ دقیقاً برابر حاصل‌ضرب دو عدد x و $y(۱+\varepsilon)$ است که $|\varepsilon| \leqslant u$. به عبارت دیگر اگر x و y را داده‌های ورودی و $\widehat{s}$ را جواب تقریبی درنظر بگیریم، می‌توان گفت حاصل‌ضرب داده‌های اختلال‌یافته‌ی

$$\widehat{x} = x, \quad \widehat{y} = y(۱+\varepsilon), \quad |\varepsilon| \leqslant u,$$

منجر به جواب تقریبی مورد نظر شده است و می‌توان نوشت

$$\widehat{s} = \widehat{x}\widehat{y}, \quad \frac{|x - \widehat{x}|}{|x|} = ۰, \quad \frac{|y - \widehat{y}|}{|y|} \leqslant u.$$

در واقع جواب تقریبی، حاصلضرب دو داده‌ی نزدیک به داده‌های واقعی است. به این آنالیز، آنالیز خطای پسرو می‌گوییم. البته در این مثال به متغیر x اختلالی وارد نشده است، در اینجا می‌توان اختلال وارده را بر x در نظر گرفت و y را تنها نوشت و یا برای هر دو داده، اختلال $\sqrt{۱+\varepsilon}$ را در نظر گرفت. از طرف دیگر با توجه به رابطه‌ی (۲۳.۲) به‌سادگی نتیجه می‌گیریم

$$|\widehat{s} - s| \leqslant u|s|, \quad \text{یا} \quad \frac{|\widehat{s} - s|}{|s|} \leqslant u.$$

در رابطه‌ی بالا کرانی برای اختلاف جواب واقعی s و جواب محاسبه شده $\widehat{s}$ بر حسب میزان اختلال در داده‌ها (u) داده شده است، که به آن آنالیز خطای پیشرو می‌گوییم. در واقع تعریف می‌کنیم:

تعریف ۴.۲. به یافتن کرانی برای اختلالات وارده به داده‌ها (ورودی‌ها) به گونه‌ای که داده‌های اختلال یافته منجر به جواب تقریبی شوند، *آنالیز خطای پسرو* می‌گوییم. اگر این کران‌ها در حد واحد گرد کردن باشند (یعنی ضریب کوچکی از

u باشند) گوییم روش *پسرو پایدار* است. از طرف دیگر به یافتن کران خطای اختلاف جواب تقریبی و جواب واقعی، *آنالیز خطای پیشرو* می‌گوییم. اگر این کران در حد واحد گرد کردن باشد گوییم روش *پیشرو پایدار* است.

مثال ۸.۲. برای حاصلضرب سه عدد ممیز شناور غیر صفر $x_۱$، $x_۲$ و $x_۳$ یعنی $s_۳ = x_۱x_۲x_۳$ داریم

$$\begin{aligned}\widehat{s}_۳ &= fl(x_۱x_۲x_۳)\\ &= fl(fl(x_۱x_۲)x_۳)\\ &= x_۱x_۲(۱+\varepsilon_۲)x_۳(۱+\varepsilon_۳)\\ &= x_۱x_۲x_۳(۱+\theta_۲)\\ &= s_۳(۱+\theta_۲), \quad |\theta_۲| \leqslant \gamma_۲.\end{aligned}$$

و بطور کلی برای حاصلضرب n عدد ممیز شناور غیر صفر $s_n = x_۱x_۲\cdots x_n$ داریم

$$\begin{aligned}\widehat{s}_n = fl(x_۱x_۲\cdots x_n) &= x_۱x_۲(۱+\varepsilon_۲)x_۳(۱+\varepsilon_۳)\cdots x_n(۱+\varepsilon_n)\\ &= x_۱x_۲\cdots x_n(۱+\theta_{n-۱}) \qquad (۲۴.۲)\\ &= s_n(۱+\theta_{n-۱}), \quad |\theta_{n-۱}| \leqslant \gamma_{n-۱}.\end{aligned}$$

با توجه به سطر اول رابطه‌ی (۲۴.۲) می‌توان گفت مقدار تقریبی $\widehat{s}_n$ دقیقاً برابر حاصل‌ضرب مقادیر اختلال‌یافته‌ی

$$\widehat{x}_۱ = x_۱, \quad \widehat{x}_k = x_k(۱+\varepsilon_k), \quad |\varepsilon_k| \leqslant u, \quad k = ۲, \ldots, n,$$

است، یعنی

$$\widehat{s}_n = \widehat{x}_۱\widehat{x}_۲\cdots\widehat{x}_n, \quad \frac{|\widehat{x}_k - x_k|}{|x_k|} \leqslant u, \quad k = ۱, ۲, \ldots, n.$$

آنالیز ارائه شده در بالا یک آنالیز خطای پسرو می‌باشد. چون کران اختلالات در داده‌ها کوچک است (برابر u است)، پس این الگوریتم پسرو پایدار است. همچنین با توجه به رابطه‌ی (۲۴.۲) کران خطای پیشرو به صورت زیر بسادگی نتیجه می‌شود

$$\frac{|\widehat{s}_n - s_n|}{|s_n|} \leqslant \gamma_{n-۱}. \qquad (۲۵.۲)$$

با توجه به تعریف γ_n در لم ۲.۲، داریم

$$\gamma_n = nu\frac{۱}{۱-nu} = nu\left(۱ + nu + (nu)^۲ + \cdots\right) = nu + \mathcal{O}(u^۲).$$

بنابراین کران بالای (۲۵.۲) را می‌توان با $nu + \mathcal{O}(u^۲)$ جایگزین کرد. طبق کران بدست آمده، اگر n یعنی تعداد عملوندها بسیار بزرگ باشد، این الگوریتم پیشرو پایدار نیست.

۳.۴.۲ الگوریتم جمع زدن

در این بخش آنالیز خطای حاصل‌جمع n عدد ممیز شناور را بدست می‌آوریم که کمی از آنالیز خطای حاصل‌ضرب پیچیده‌تر است. فرض کنیم n عدد ممیز شناور $x_۱$، $x_۲$ ، ... و x_n در دست است و گیریم $s_n = x_۱ + x_۲ + \cdots + x_n$. مجموع‌های جزئی را به‌صورت $s_j = x_۱ + \cdots + x_j,\ j = ۱, \ldots, n$ درنظر می‌گیریم. داریم

$$\begin{aligned}
\widehat{s}_۱ &= x_۱, \\
\widehat{s}_۲ &= fl(\widehat{s}_۱ + x_۲) = (x_۱ + x_۲)(۱ + \varepsilon_۱), \\
\widehat{s}_۳ &= fl(\widehat{s}_۲ + x_۳) = [(x_۱ + x_۲)(۱ + \varepsilon_۱) + x_۳]\,(۱ + \varepsilon_۲) \\
&= (x_۱ + x_۲)(۱ + \varepsilon_۱)(۱ + \varepsilon_۲) + x_۳(۱ + \varepsilon_۲),
\end{aligned}$$

که در آن $|\varepsilon_k| \leqslant u$ برای $k = ۱, ۲$. همانند مثال ۶.۲ لازم نیست بین ε_k ها تفاوتی قائل شویم (مگر در علامت آن‌ها) بنابراین اندیس آن‌ها را درنظر نمی‌گیریم و قرار می‌دهیم $۱ + \varepsilon_k \equiv ۱ \pm \varepsilon$. داریم

$$\widehat{s}_۳ = x_۱(۱ \pm \varepsilon)^۲ + x_۲(۱ \pm \varepsilon)^۲ + x_۳(۱ \pm \varepsilon), \quad |\varepsilon| \leqslant u,$$

و به طور مشابه با الهام از الگوی بالا داریم

$$\widehat{s}_n = x_۱(۱ \pm \varepsilon)^{n-۱} + x_۲(۱ \pm \varepsilon)^{n-۱} + x_۳(۱ \pm \varepsilon)^{n-۲} + \cdots + x_n(۱ \pm \varepsilon).$$

اگر شرط $nu < ۱$ برقرار باشد (که غالباً برقرار است مگر اینکه n بسیار بزرگ باشد)، با توجه به نمادهای لم ۲.۲ داریم

$$\widehat{s}_n = x_۱(۱ + \theta'_{n-۱}) + x_۲(۱ + \theta_{n-۱}) + x_۳(۱ + \theta_{n-۲}) + \cdots + x_n(۱ + \theta_۱), \tag{۲۶.۲}$$

که $|\theta'_{n-۱}| \leqslant \gamma_{n-۱}$ و $|\theta_j| \leqslant \gamma_j$ برای $j = ۱, \ldots, n-۱$. بنابراین اگر داده‌های ورودی اختلال یافته به صورت

$$\widehat{x}_۱ = x_۱(۱ + \theta'_{n-۱}), \quad \widehat{x}_k = x_k(۱ + \theta_{n-k+۱}), \quad k = ۲, \ldots, n,$$

تعریف شوند، آنالیز خطای پسرو به صورت زیر است

$$\widehat{s}_n = \widehat{x}_۱ + \widehat{x}_۲ + \cdots + \widehat{x}_n, \quad \frac{|\widehat{x}_k - x_k|}{|x_k|} \leqslant \gamma_{n-k+۱}, \quad k = ۱, ۲, \ldots, n.$$

این روابط نشان می‌دهند برای n های نه چندان بزرگ، الگوریتم معمولی جمع کردن پسرو پایدار است. از طرفی با توجه به (۲۶.۲) داریم

$$\widehat{s}_n - s_n = x_۱\theta'_{n-۱} + \sum_{k=۲}^{n} x_k\theta_{n-k+۱},$$

با توجه به اینکه بزرگترین کران اختلالات در تمام داده‌ها γ_{n-1} است، می‌توان نوشت

$$|\widehat{s}_n - s_n| \leqslant \gamma_{n-1} \sum_{k=1}^{n} |x_k|,$$

و با فرض اینکه $s_n \neq 0$، به صورت نسبی داریم

$$\frac{|\widehat{s}_n - s_n|}{|s_n|} \leqslant \gamma_{n-1} \frac{\sum_{k=1}^{n} |x_k|}{|\sum_{k=1}^{n} x_k|},$$

که یک آنالیز خطای پیشرو است. اگر همه‌ی x_k ها هم علامت باشند، ضریب کسری در کران بالا برابر ۱ است، در غیر این صورت این ضریب می‌تواند بزرگ باشد و روش پیشرو ناپایدار می‌شود.

از معادله‌ی (۲۶.۲) یک نکته‌ی مهم دیگر نیز برداشت می‌شود. می‌دانیم هرچه اندیس k کوچکتر باشد کران بالای γ_{n-k+1} بزرگتر خواهد بود. برای اینکه کمترین خطای محاسباتی را داشته باشیم بایستی ترتیب جمع شدن جملات طوری باشد که بزرگترین جمله در کوچکترین فاکتور ضرب شود. با توجه به اینکه کوچکترین فاکتور، $(1 + \theta_1)$، در x_n ضرب شده است، پس بایستی x_n بزرگترین جمله و به همین ترتیب x_1 کوچکترین جمله باشد، یعنی اگر جملات از کوچک به بزرگ (از لحاظ قدرمطلق) با هم جمع شوند خطای محاسباتی کمتری در جواب نهایی رخ خواهد داد و خطای پسرو بهینه می‌شود.

۴.۴.۲ الگوریتم ضرب داخلی

ضرب داخلی دو بردار از اهمیت ویژه‌ای برخوردار است زیرا اکثر عملگرهای جبر خطی عددی مبتنی بر ضرب داخلی‌اند. بنابراین لازم است الگوریتم‌های پایدار برای آن طراحی شود. ضرب داخلی $s_n = x^T y$ را درنظر بگیرید که در آن $x, y \in \mathbb{R}^n$. با توجه به اینکه $s_n = x_1 y_1 + \cdots + x_n y_n$. یک الگوریتم برای آن به صورت زیر است

```
s = 0;
for k = 1 : n
   s = s + x(k)*y(k);
end
```

برای آنالیز کردن این الگوریتم، گیریم $s_k = x_۱y_۱ + \cdots + x_ky_k$ بیانگرِ k امین مجموع جزئی باشد، داریم

$$\begin{aligned}
\widehat{s}_۱ &= fl(x_۱y_۱) = x_۱y_۱(۱ + \varepsilon_۱),\\
\widehat{s}_۲ &= fl(\widehat{s}_۱ + x_۲y_۲)\\
&= [\widehat{s}_۱ + x_۲y_۲(۱ + \varepsilon_۲)]\,(۱ + \varepsilon_۳)\\
&= [x_۱y_۱(۱ + \varepsilon_۱) + x_۲y_۲(۱ + \varepsilon_۲)]\,(۱ + \varepsilon_۳)\\
&= x_۱y_۱(۱ \pm \varepsilon)^۲ + x_۲y_۲(۱ \pm \varepsilon)^۲,
\end{aligned}$$

که در آن $|\varepsilon| \leqslant u$. با محاسباتی مشابه داریم

$$\begin{aligned}
\widehat{s}_۳ &= fl(\widehat{s}_۲ + x_۳y_۳)\\
&= [\widehat{s}_۲ + x_۳y_۳(۱ \pm \varepsilon)](۱ \pm \varepsilon)\\
&= [x_۱y_۱(۱ \pm \varepsilon)^۲ + x_۲y_۲(۱ \pm \varepsilon)^۲ + x_۳y_۳(۱ \pm \varepsilon)](۱ \pm \varepsilon)\\
&= x_۱y_۱(۱ \pm \varepsilon)^۳ + x_۲y_۲(۱ \pm \varepsilon)^۳ + x_۳y_۳(۱ \pm \varepsilon)^۲.
\end{aligned}$$

و با الهام از الگوی بالا در حالت کلی می‌توان نوشت

$$\widehat{s}_n = x_۱y_۱(۱ \pm \varepsilon)^n + x_۲y_۲(۱ \pm \varepsilon)^n + x_۳y_۳(۱ \pm \varepsilon)^{n-۱} + \cdots + x_ny_n(۱ \pm \varepsilon)^۲.$$

اگر شرط $nu < ۱$ برقرار باشد، با توجه به لم ۲.۲ داریم

$$\widehat{s}_n = x_۱y_۱(۱ + \theta'_n) + x_۲y_۲(۱ + \theta_n) + x_۳y_۳(۱ + \theta_{n-۱}) + \cdots + x_ny_n(۱ + \theta_۲), \qquad (۲۷.۲)$$

که در آن $|\theta'_n| \leqslant \gamma_n$ و $|\theta_j| \leqslant \gamma_j$ برای $j = ۲, \ldots, n$. در حقیقت داده‌های اختلال‌یافته‌ی

$$\begin{cases}
\widehat{x}_k = x_k, \quad k = ۱, \ldots, n,\\
\widehat{y}_۱ = y_۱(۱ + \theta'_n), \quad \widetilde{y}_k = y_k(۱ + \theta_{n-k+۲}), \quad k = ۲, \ldots, n,
\end{cases}$$

(یا برعکس می‌توان اختلالات را با x_k ها در نظر گرفت و y_k ها را تنها نوشت) منجر به جواب محاسباتیِ

$$\widehat{s}_n = \widehat{x}_۱\widehat{y}_۱ + \cdots + \widehat{x}_n\widehat{y}_n$$

می‌شوند بطوریکه

$$\frac{|\widehat{x}_k - x_k|}{|x_k|} = ۰, \quad \frac{|\widehat{y}_k - y_k|}{|y_k|} \leqslant \gamma_k, \quad k = ۱, ۲, \ldots, n.$$

این یک آنالیز خطای پسرو می‌باشد. آنالیز خطای (۲۷.۲) را می‌توان بصورت دیگری نیز بیان کرد. گیریم $|x|$ نشان‌دهنده‌ی برداری با درایه‌های $|x_k|$ باشد یعنی $|x| = (|x_1|, \ldots, |x_n|)$. به کمک (۲۷.۲) بسادگی می‌توان نشان داد

$$fl(x^T y) = (x + \delta x)^T y = x^T (y + \delta y), \quad |\delta x| \leqslant \gamma_n |x|,\ |\delta y| \leqslant \gamma_n |y|, \tag{۲۸.۲}$$

در اینجا علامت نامساوی بین بردارها به‌صورت درایه به درایه تفسیر می‌شود. یعنی برای بردارهای x و y در $\mathbb{R}^n$ داریم $|x| < |y|$ اگر و تنها اگر $x_k < y_k$ بازای $k = 1, \ldots, n$. بلافاصله از (۲۸.۲) نتیجه می‌گیریم

$$\left| fl(x^T y) - x^T y \right| \leqslant \gamma_n \sum_{k=1}^{n} |x_k y_k| = \gamma_n |x|^T |y|,$$

که یک آنالیز خطای پیشرو است. اگر $x = y$، کران بالا نشان می‌دهد در محاسبه‌ی $x^T x$ خطای نسبی γ_n بدست می‌آید یعنی

$$\frac{\left| fl(x^T x) - x^T x \right|}{|x^T x|} \leqslant \gamma_n.$$

اما اگر $|x^T y| \ll |x|^T |y|$ دقت نسبی بالا حاصل نخواهد شد. در واقع این الگوریتم در حالت کلی پیشرو پایدار نیست.

۵.۴.۲ عملگر ضرب-جمع ترکیبی

امروزه بسیاری از کامپیوترها عملگر ضرب-جمعِ ترکیبی را پشتیبانی می‌کنند. بدین معنا که عبارتی به فرم $x \times y + z$ (یا $x \times y - z$) تنها با یک دستور محاسبه می‌شود و فقط یک‌بار خطای گردکردن در آن رخ می‌دهد، یعنی داریم

$$fl(xy \pm z) = (xy \pm z)(1 + \varepsilon), \quad |\varepsilon| \leqslant u.$$

عملگر ضرب-جمع ترکیبی در بسیاری از الگوریتم‌ها سودمند است و خطای گردکردن را در آن‌ها تقریباً نصف می‌کند. برای نمونه با استفاده از عملگر ضرب-جمع ترکیبی ضرب داخلیِ $x^T y$ بین دو بردار n تایی x و y می‌تواند تنها با n خطای محاسبات بجای $2n - 1$ در حالت جمع و ضرب معمولی انجام شود. (توجه داریم که برای انجام یک ضرب داخلی در یک حلقه هربار یک ضرب و یک جمع با مجموع قبلی داریم.)

مثال دیگر الگوریتم هورنر برای محاسبه مقدار چندجمله‌ای $p(x) = a_n x^n + \cdots + a_1 x + a_0$ است که با رابطه‌ی بازگشتیِ $b_n = a_n$ و $b_k = x b_{k+1} + a_k,\ k = n-1 : -1 : 0$ داده شده است، که تنها به n عملگر ضرب-جمع ترکیبی نیاز دارد.

به عنوان یک مثال دیگر روش نیوتن برای حل $f(x) = a - 1/x = 0$ را در نظر بگیرید. بعداً خواهیم دید، رابطه‌ی بازگشتی متناظر عبارتست از

$$\begin{aligned} x_{k+1} &= x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{a - 1/x_k}{x_k^{-2}} \\ &= x_k + (1 - x_k a) x_k. \end{aligned}$$

در کامپیوترهای امروزی این روش برای محاسبه‌ی معکوس یک عدد و از آنجا برای انجام عمل تقسیم به‌صورت $a/b = a \times (1/b)$ استفاده می‌شود. همانگونه که مشاهده می‌کنیم در فرمول بالا در هر گام، محاسبه‌ی x_{k+1} بر حسب x_k نیاز به دو عمل ضرب-جمع ترکیبی دارد که منجر به کاهش خطاهای گردکردن به حدود $\frac{1}{4}$ حالت عادی می‌شود.

۶.۴.۲ جلوگیری از سرریزی

با توجه به اینکه حوزه‌ی نمایش اعداد در ماشین محدود است همواره باید مراقب خطاهای سرریز و پی‌ریز بود. اگر اعداد زیرنرمال تعریف شده باشند نگران خطای پی‌ریز نیستیم، اما احتمال خطای سرریز همیشه وجود دارد. گرچه به‌نظر می‌رسد استاندارد IEEE حوزه‌ی وسیعی از اعداد را پوشش دهد، اما حتی در الگوریتم‌های خیلی ساده ممکن است به سرعت از حوزه‌ی نمایش خارج شده و گرفتار خطای سرریز شویم. برای مثال اگر $x_0 = 2$ و $x_{n+1} = x_n^2$ بلافاصله داریم $x_{10} = 2^{1024}$ که بزرگتر از حوزه‌ی استاندارد IEEE با دقت دوبرابر است. در مورد محاسباتی که شامل فاکتوریل هستند نیز بایستی مواظب خطای سرریز بود، برای مثال $10^{309} \times 1.24 \doteq 171!$ که از بزرگترین عدد با دقت دوبرابر بزرگتر است.

برای جلوگیری از خطای سرریز می‌توان از دقت‌های بالاتر در کامپیوتر استفاده کرد یا الگوریتمی که منجر به این خطا می‌شوند را بازنویسی نمود. مثلاً یک راه ممکن گاهی تغییر متغیر لگاریتمی است. همچنین بسته به نوع مسئله تغییر متغیرهای دیگر می‌توانند مورد استفاده قرار گیرند. در روند الگوریتم بایستی مراقب باشیم در محاسبات میانی دچار اینگونه خطاها نشویم. یک مثال ساده محاسبه‌ی مجموع فیثاغورثی $c = \sqrt{a^2 + b^2}$ است که در عمل بسیار با آن مواجه هستیم مثلاً در محاسبه‌ی اندازه در مختصات قطبی یا اعداد مختلط. حتی اگر a و b در حوزه‌ی نمایش باشند، c نیز در حوزه خواهد بود اما ممکن است مجموع مربعات آن‌ها (محاسبه‌ی میانی) سرریز شود. یک راه حل ساده الگوریتم زیر است: اگر $a = b = 0$ آنگاه $c = 0$، در غیر این‌صورت قرار دهید $p = \max\{|a|, |b|\}$ و $q = \min\{|a|, |b|\}$ و از این‌رو

$$\rho = q/p, \quad c = p\sqrt{1 + \rho^2}.$$

اگر a و b در حوزه‌ی نمایش باشند، این روند هیچ‌گاه دچار سرریز نخواهد شد.

مثالِ مشابه دیگر، محاسبه اندازه‌ی اقلیدسی یک بردار غیر صفر به‌صورت $\|x\|_2 = (\sum_{k=1}^n x_k^2)^{1/2}$ است. برای جلوگیری از خطای سرریز ابتدا بزرگترین المان از نظر قدرمطلق یعنی $x_{\max} = \max_k |x_k|$ را بدست آورده و قرار می‌دهیم

$$s = \sum_{k=1}^{n} (x_k/x_{\max})^2, \quad \|x\|_2 = x_{\max}\sqrt{s}.$$

۷.۴.۲ جلوگیری از خطای حذف

آنچه منجر به نتیجه نادرست در الگوریتم اولِ مثال ۴.۲ برای محاسبه π شد، خطای حذف ارقام با معنا بود که از این پس آن را *خطای حذف* می‌نامیم. خطای حذف در عمل تفریق دو عدد نزدیک به هم رخ می‌دهد. برای مثال دو عدد نزدیک به هم

$$x = ۰/۴۷۱۱۴۴۸۴۵۶۷ \times ۱۰^{۴}, \quad y = ۰/۴۷۱۱۳۲۵ \times ۱۰^{۴}$$

را در نظر بگیرید و فرض کنیم طول مانتیس درنظرگرفته شده در ماشین $t = ۷$ است. داریم

$$fl(x - y) = ۰/۰۰۰۰۱۲۳ \times ۱۰^{۴} = ۰/۱۲۳۰۰۰۰ \times ۱۰^{۰}.$$

چهار رقم انتهایی مانتیس که برابر صفر هستند از بین رفته‌اند و در حقیقت چهار رقم بامعنا در این تفریق از دست داده‌ایم. واضح است که اگر محاسبات با تعداد ارقام بیشتری انجام می‌شد، به نسبت تعداد ارقام بامعنای کمتری را از دست می‌دادیم. هرچه اعداد به هم نزدیک‌تر باشند تعداد ارقام بامعنای بیشتری در تفریق از دست خواهند رفت. برای جلوگیری از خطای حذف می‌توان از دقت بالاتر در ماشین استفاده کرد یا الگوریتم را طوری بازنویسی نمود که از تفریق اعداد نزدیک به هم اجتناب شود.

در فرمول بازگشتی (۱۸.۲) برای محاسبه عدد π، اختلاف عدد ۱ و عبارت $\sqrt{۱ - [۲^{-k}p_k]^{۲}}$ برای k بزرگ ناچیز است و منجر به خطای حذف می‌شود. یک بازنویسی ساده از الگوریتم منجر به فرمول بازگشتی (۱۹.۲) شد، که در آن از این تفریق اجتناب شده است.

مثال دیگر، محاسبه‌ی ریشه‌های یک چندجمله‌ای درجه دوم $ax^{۲} + bx + c = ۰$، برای $a \neq ۰$ با دستور

$$r_{۱,۲} = \frac{-b \pm \sqrt{b^{۲} - ۴ac}}{۲a}$$

است. این فرمول برای محاسبه‌ی ریشه‌ی کوچکتر (از لحاظ قدرمطلق) در ماشین دقیق نخواهد بود. برای جلوگیری از این پدیده می‌توان ریشه‌ی کوچکتر را با توجه به رابطه‌ی $r_۱r_۲ = c/a$ محاسبه کرد.

مثال بعد، نحوه یافتن ریشه‌ی دوم عدد مختلط $x = a + ib$ است. فرض کنیم $u + iv = \sqrt{a + ib}$. در این صورت می‌توان نشان داد

$$u = \left(\frac{r + a}{۲}\right)^{۱/۲}, \quad v = \left(\frac{r - a}{۲}\right)^{۱/۲}, \quad r = \sqrt{a^{۲} + b^{۲}}.$$

واضح است هرگاه $a > ۰$ و $|a| \gg |b|$ در محاسبه‌ی v خطای حذف رخ می‌دهد، چراکه در این صورت a و r خیلی به هم نزدیکند. برای جلوگیری از این خطا، با توجه به این واقعیت که $uv = \sqrt{r^{۲} - a^{۲}}/۲ = b/۲$ خواهیم داشت $v = b/(۲u)$. همچنین اگر $a < ۰$، در محاسبه‌ی u به خطای حذف برخواهیم خورد که بایستی از فرمول $u = b/(۲v)$ استفاده کنیم.

آنچه باعث حصول نتیجه‌ی نامعتبر در محاسبات می‌شود، همیشه خطای حذف نیست. می‌توان الگوریتم‌هایی را مثال زد که هیچگونه عمل تفریق در آن‌ها وجود ندارد با این‌حال جوابهایی بسیار دور از واقعیت ارائه می‌دهند.

۵.۲ هزینه‌های محاسباتی

هزینه محاسباتی یا پیچیدگی یک روش، تعداد کل اعمال ریاضی جمع، ضرب، تقسیم و تفریق و همچنین ارزیابی‌های توابع در سرتاسر الگوریتم است. اگر برای حل یک مسئله دو روش مناسب (سازگار، پایدار و همگرا) موجود باشد، روشی که هزینه محاسباتی کمتری داشته باشد ارجحیت خواهد داشت. شاید تصور شود با وجود کامپیوترهای سریع که اعمال ریاضی را در کسر ناچیزی از ثانیه انجام می‌دهند پرداختن به این موضوع بی اهمیت باشد. برای دریافتن اهمیت این موضوع حل یک دستگاه معادلات خطی از بعد ۱۱ × ۱۱ را در نظر بگیرید. اگر چه روش کرامر برای حل چنین دستگاهی پایدار نیست، با این حال اگر از این روش استفاده شود نیاز به محاسبه ۱۲ دترمینان است. اگر دترمینان‌ها با روش بسط حول یک سطر یا ستون به صورت بازگشتی محاسبه شوند، حل این دستگاه به بیش از ۱۳ میلیارد عمل محاسباتی جمع، تفریق، ضرب و تقسیم نیاز دارد که مدت زمان قابل توجهی برای پردازش این تعداد عمل محاسباتی مورد نیاز است. اما اگر همین دستگاه با روش حذفی گاوس حل شود فقط حدود ۹۰۰ عمل محاسباتی مورد نیاز است که کامپیوترهای امروزی در کسری از ثانیه آن را انجام می‌دهند. اگر ابعاد مسئله بزرگتر شود، اهمیت استفاده از روش‌هایی با پیچیدگی محاسباتی کم آشکارتر می‌شود. در مثال‌های زیر هزینه محاسباتی چند الگوریتم را بدست می‌آوریم.

مثال ۹.۲. فرض کنید به دنبال ارزیابی چندجمله‌ای

$$p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$$

در یک نقطه‌ی دلخواه مانند x_0 هستیم. ابتدا از الگوریتم جایگذاری مستقیم استفاده کنیم. بخش اصلی برنامه آن در متلب به صورت زیر است. توجه کنید چون در این نرم‌افزار اندیس آرایه‌ها به صورت پیش‌فرض از ۱ شروع می‌شود، اندیس ضرایب یک واحد افزایش یافته است.

```
p = a(1);
for k = 1:n
   p = p + a(k+1)*x0^k;
end
```

این حلقه n بار تکرار می‌شود و در تکرار k-ام یک جمع و k ضرب مورد نیاز است (توان نیز حالت خاصی از ضرب است). بنابراین هزینه محاسباتی این الگوریتم عبارتست از

$$\sum_{k=1}^{n}(k+1) = \frac{n^2+3n}{2} = \mathcal{O}(n^2).$$

معمولاً ضریب بزرگترین توان n نیز از اهمیت ویژه‌ای برخوردار است (برای مثال هر دوی $۱۰۰۰n^۲$ و $۰/۱n^۲$ از $\mathcal{O}(n^۲)$ هستند، اما تفاوت چشم‌گیری دارند)، به همین علت گاهی ضریب بزرگترین توان قید می‌شود. برای مثال در الگوریتم بالا می‌گوییم هزینه محاسباتی تقریباً $\frac{۱}{۲}n^۲$ است. الگوریتم بالا را می‌توان با حفظ توان‌های $x_۰$ در گام‌های قبل، اصلاح کرد. الگوریتم اصلاح شده به صورت زیر است

```
p = a(1); s = 1;
for k = 1:n
   p = p + a(k+1)*s*x0;
   s = s*x0;
end
```

هزینه محاسباتی این الگوریتم $۴n$ است.

یک الگوریتم دیگر برای ارزیابی $p(x)$ در نقطه‌ی $x_۰$، الگوریتم مشهور هورنر است که به صورت زیر طراحی می‌شود: چندجمله‌ای $p_n(x)$ را به شکل زیر بازنویسی می‌کنیم

$$p_n(x) = \Big(\big(\cdots\big((a_nx + a_{n-۱})x + a_{n-۲}\big)x + \cdots\big)x_۰ + a_۰\Big).$$

با یک انتقال اندیس برای همراه شدن با متلب الگوریتم زیر را خواهیم داشت که در نهایت مقدار b همان $p(x_۰)$ خواهد بود.

```
b=a(n+1);
for k = n:-1:1
   b = b*x0+a(k);
end
```

حلقه بالا نیز n بار تکرار می‌شود و در هر تکرار تنها یک جمع و یک ضرب مورد نیاز است و از این رو هزینه این روش $۲n$ خواهد بود که نصف هزینه‌ی الگوریتم قبل است. الگوریتم هورنر از نظر خطاهای محاسباتی نیز بهینه است.

پیچیدگی یا هزینه محاسباتی، یکی از عامل‌های تعیین کننده‌ی برتری یک روش نسبت به روش دیگر است. برای همین منظور در بسیاری از الگوریتم‌های عددی ارائه شده در این کتاب هزینه‌های محاسباتی محاسبه خواهند شد.

۶.۲ وضعیت و پایداری

در حل عددیِ یک مدل ریاضی با یک روش عددی، هم وضعیت خود مدل و هم نحوه رفتار روش عددی دارای اهمیت ویژه‌اند و بایستی به طور دقیق بررسی شوند. وضعیت مدل یا مسئله‌ی ریاضی متفاوت از وضعیت روش عددی است و این دو موضوع باید به طور جداگانه بررسی شوند. برای ایجاد انگیزه، بحث خود را با ارائه چند مثال ساده آغاز می‌کنیم.

مثال ۱۰.۲. یکی از مسائلی که همواره با آن مواجهیم حل یک دستگاه معادلات خطی است. تقریباً تمام مدل‌های خطی (و حتی مدل‌های غیر خطی در صورت خطی‌سازی) منجر به حل یک دستگاه معادلات خطی می‌شوند. حل چنین دستگاهی، بر خلاف آنچه در نگاه اول نظر می‌رسد، می‌تواند بسیار مشکل‌ساز باشد. فرض کنید حل دستگاه $Ax = b$ مد نظر است که در آن A *ماتریس هیلبرت* است. ماتریس هیلبرت به صورت زیر تعریف می‌شود

$$H_n := \begin{bmatrix} ۱ & \frac{۱}{۲} & \cdots & \frac{۱}{n} \\ \frac{۱}{۲} & \frac{۱}{۳} & \cdots & \frac{۱}{n+۱} \\ \vdots & \vdots & & \vdots \\ \frac{۱}{n} & \frac{۱}{n+۱} & \cdots & \frac{۱}{۲n-۱} \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

در اینجا فرض می‌کنیم $n = ۱۱$ و بردار سمت راست یعنی b نیز داده شده است. به دنبال بردار جواب x هستیم. برای اینکه جواب دقیق مسئله را داشته باشیم فرض می‌کنیم $x = [۱, ۱, \ldots, ۱]^T \in \mathbb{R}^{۱۱}$ و با ضرب کردن A در x بردار سمت راست را بدست می‌آوریم. حال یک روش برای حل دستگاه با A و b داده شده به کار می‌گیریم و جواب تقریبی $\widehat{x}$ را با جواب دقیق x مقایسه می‌کنیم. برنامه را در محیط متلب می‌نویسیم. برای تولید ماتریس هیلبرت از دستور `A = hilb(11)` و برای تولید برداری ستونی از اندازه‌ی ۱۱ با درایه‌های ۱ از دستور `ones(11,1)` استفاده می‌کنیم. به کمک دستور `linsolve` جواب عددی دستگاه را بدست می‌آوریم:

```
>> A = hilb(11); b = A*ones(11,1); xhat = linsolve(A,b)
xhat =
   0.999999993350113
   1.000000703429746
   0.999981603925342
   1.000207015815544
   0.998760068660420
   1.004379228959193
   0.990427925814003
```

```
1.013093309178271
0.989092115526001
1.005059843203712
0.998998187609104
```

اگر ماتریس A مربعی باشد، دستور linsolve از روش تجزیه‌ی LU با محورگیری استفاده می‌کند. دقت پیش‌فرض در متلب، دقت دوبرابر است و دیدیم در این حالت $u \doteq ۱/۱ \times ۱۰^{-۱۶}$، اما اگر خطای ماکزیمم نسبی در حل این دستگاه را حساب کنیم داریم

$$\frac{\|x - \widehat{x}\|_\infty}{\|x\|_\infty} \doteq ۰/۰۱۳۰۹۳۳۰۹۱۷۸۲۷۱,$$

که نشان می‌دهد برای این ماتریس با این ابعاد کوچک تقریباً خطای نسبی $۱۰^{۱۵}$ بار بدتر از انتظار ماست. شاید فکر کنیم مشکل از روش تجزیه‌ی LU است. اما در کتاب‌های جبرخطی عددی ثابت می‌شود که این روش در صورت محورگیری روش پایداری است و باید مشکل را در جای دیگری جستجو کرد. در این بخش خواهیم دید علت این مشکل مربوط به خود ماتریس A است. این ماتریس یک ماتریس *بیمار* است! و باید از روش‌های عددیی که منجر به چنین ماتریسی می‌شوند اجتناب کرد.

مثال ۱۱.۲. فرض کنید $p(x) = (x - ۱)(x - ۲) \cdots (x - ۸)$ به شکل گسترده به صورت زیر نوشته شود

$$p(x) = x^۸ - ۳۶x^۷ + ۵۴۶x^۶ - ۴۵۳۶x^۵ + ۲۲۴۴۹x^۴ - ۶۷۲۸۴x^۳ + ۱۱۸۱۲۴x^۲ - ۱۰۹۵۸۴x + ۴۰۳۲۰$$

واضح است که ریشه‌های آن اعداد طبیعی متمایز ۱، ۲، ...، ۸ می‌باشند. در متلب دستور roots برای یافتن ریشه‌های یک چندجمله‌ای به کار می‌رود. ورودی آن بردار ضرایب چندجمله‌ای است. با این دستور می‌توان ریشه‌های یک چندجمله‌ای (از درجه پایین مانند همین مثال) را با دقت بسیار خوب بدست آورد. اگر از این دستور استفاده کنیم نتایج زیر را خواهیم داشت

```
>> c = [1 -36 546 -4536 22449 -67284 118124 -109584 40320];
>> r = roots(c)
r =
```

```
    8.0000
    7.0000
    6.0000
    5.0000
    4.0000
    3.0000
    2.0000
    1.0000
```

که حداقل تا ۱۱ رقم اعشار دقیق‌اند و در اینجا تا چهار رقم اعشار ارائه شده‌اند. حال ضریب دوم یعنی **۳۶**− را به **۳۶٫۰۰۱**− تغییر می‌دهیم و باز به کمک همین دستور ریشه‌ها را می‌یابیم. داریم

```
>> c = [1 -36.001 546 -4536 22449 -67284 118124 -109584 40320];
>> r = roots(c)
r =
   8.2726
   6.4999 + 0.7293i
   6.4999 - 0.7293i
   4.5748
   4.1625
   2.9911
   2.0002
   1.0000
```

برخی از ریشه‌ها خیلی تغییر نکرده‌اند، اما به عنوان مثال در بازه‌ی [**۵٫۵, ۶٫۵**] دیگر ریشه‌ای وجود ندارد و در عوض یک جفت ریشه مختلط مزدوج داریم. خاطر نشان می‌کنیم که این بدوضعی مربوط به خود مسئله است نه الگوریتم خاصی که برای حل آن به کار گرفته شد. همانطور که گفته شد دستور `roots` برای یافتن ریشه‌های چندجمله‌ایها با درجه پایین، پایدار است. اما مشاهده کردیم که بدوضعی خود مسئله باعث شد که اندکی اختلال در ورودی ساختار ریشه‌ها را عوض کند تا آنجا که ریشه‌های مختلط نیز بدست آمدند.

مثال ۱۲.۲. این مثال متفاوت از مثال‌های قبل است به این معنی که مسئله‌یی که قرار است حل کنیم مشکلی ندارد، اما الگوریتم عددیی که برای حل آن به کار می‌گیریم روش مناسبی نیست. می‌خواهیم $I_n = \int_۰^۱ \frac{t^n}{t+۱۰}\,dt$ را برای عدد ثابت $n \geqslant ۱$ محاسبه کنیم. یک روش بازگشتی با برقراری ارتباط بین I_k و $I_{k-۱}$ بدست می‌آوریم. ابتدا داریم $I_۰ = \int_۰^۱ \frac{dt}{t+۱۰}\,dt = \ln ۱/۱$. با توجه به اینکه $\frac{t}{t+۱۰} = ۱ - \frac{۱۰}{t+۱۰}$، رابطه بازگشتی زیر برای بدست آوردن I_n بدست می‌آید:

$$\begin{cases} I_k = -۱۰I_{k-۱} + \dfrac{۱}{k}, \quad k = ۱, ۲, \ldots, n \\ I_۰ = \ln ۱/۱. \end{cases} \tag{۲۹.۲}$$

دستورات زیر در متلب برای محاسبه $I_{۱۶}$ نوشته می‌شوند

```
I = log(1.1);
for k = 1 : 16
   I = -10*I + 1/k;
end
```

نتیجه، مقدار ۰/۰۶۹۱۲۲۱۲۳۱۰۷۳۰۵− است که قطعاً نادرست است، زیرا حداقل مطمئن هستیم جواب واقعی مثبت است زیرا از یک تابع مثبت روی $[۰, ۱]$ انتگرال گرفته‌ایم. در اینجا مسئله مشکلی ندارد اما روش ارائه شده نیاز به درمان دارد.

در این مثال‌ها، دو مفهوم متفاوت اما مرتبط با هم را مشاهده کردیم. یکی بدوضعی مسئله تحت بررسی، و دیگری ناپایداریِ روش یا الگوریتم عددی. مفهوم پایداری هم در مورد مسئله (مدل) ریاضی و هم در مورد روش و الگوریتم عددیِ حل مسئله به کار می‌رود. در واقع باید پایداری مسئله و پایداری روش و الگوریتم به طور جداگانه بررسی شوند. معمولاً به پایداری در حالت اول *وضعیت مسئله* می‌گویند و مسئله پایدار را *خوش‌وضع* می‌نامند. در مورد روش و الگوریتم معمولاً فقط از واژه پایداری استفاده می‌شود. در ادامه‌ی این بخش فقط به وضعیت مسئله (مدل) می‌پردازیم. برای بررسی پایداری روش‌های عددی ابتدا لازم است خود روش‌ها را مطالعه کنیم که تا اینجای درس هنوز به آن نرسیده‌ایم. اما در بخش قبل اندکی در مورد پایداری الگوریتم‌های محاسباتی مانند جمع، ضرب و ضرب داخلی صحبت کردیم. توضیحات بیشتر در این زمینه در کتاب آنالیز عددی پیشرفته [۶] آمده است.

۱.۶.۲ وضعیت یک مسئله

یک مسأله معمولاً دارای ورودی و خروجی است. ورودی شامل مجموعه‌ای از داده‌ها است و خروجی مجموعه‌ی دیگری است که به صورت یکتا توسط ورودی تعیین می‌شود. فرض کنیم ورودی $x \in D \subseteq \mathbb{R}^m$ و خروجی $y \in \mathbb{R}^n$ باشد.

معمولاً x و y توسط یک نگاشت مانند F به صورت ضمنیِ

$$F(x,y)=0 \tag{۳۰.۲}$$

با هم در ارتباطند. اگر y بطور یکتا و صریح بر حسب x تعیین شود، نگاشت f وجود دارد که

$$\begin{cases} f:\ D\to\mathbb{R}^n \\ y=f(x) \end{cases}. \tag{۳۱.۲}$$

همچنین می‌توان آن را به میدان اعداد مختلط نیز تعمیم داد. آنچه مدنظر ماست بررسی حساسیت نگاشت f نسبت به تغییرات جزئی در داده‌ی ورودی x است. می‌خواهیم بدانیم تغییرات کوچک در بردار x چه اثری بر y می‌گذارد. مسئله‌ی (۳۰.۲) یا (۳۱.۲) را *خوش وضع* گوییم اگر تغییرات (اختلالات) جزئی در ورودی x باعث ایجاد تغییرات زیاد در جواب y نشود. میزان کوچک و بزرگ بودن این تغییرات به نوع مسئله و دقتی که از آن انتظار داریم بستگی دارد و با نرم فضاهایی که x و y به آن‌ها تعلق دارند اندازه‌گیری می‌شود. بنابراین تعریف می‌کنیم

تعریف ۵.۲ (وضعیت مسئله‌ی ریاضی)**.** گیریم $x\in D$ و δx یک اختلال باشد بگونه‌ای که $x+\delta x\in D$ و δy تغییرات به وجود آمده در y باشد به گونه‌ای که

$$F(x+\delta x, y+\delta y)=0, \tag{۳۲.۲}$$

و یا در فرم صریح

$$y+\delta y=f(x+\delta x). \tag{۳۳.۲}$$

آنگاه مسئله‌ی (۳۰.۲) خوش‌وضع است اگر دارای جواب یکتا باشد و ثابت مثبت C موجود باشد بطوریکه

$$\|\delta y\|\leqslant C\|\delta x\|,\quad \text{یا}\quad \frac{\|\delta y\|}{\|y\|}\leqslant C\frac{\|\delta x\|}{\|x\|}, \tag{۳۴.۲}$$

که یکی به صورت مطلق و دیگری به صورت نسبی بیان شده است. نرم های تعریف شده برای داده و جواب ممکن است یکی نباشند، چرا که احیاناً داده‌ی ورودی و جواب به دو فضای متفاوت متعلق‌اند. رابطه (۳۴.۲) می‌گوید، برای اینکه مسئله (۳۰.۲) خوش‌وضع باشد، لازم است اختلال کوچک در داده منجر به اختلالی از همان مرتبه در جواب شود.

اگر مسئله‌ای به مفهوم بالا خوش‌وضع نباشد، *بدوضع* گفته می‌شود. اگر یک مسئله ریاضی بدوضع باشد، باید برای آن راه حل عددی خاصی اندیشیده شود زیرا روش‌های عددی معمولی برای حل آن کارایی نخواهند داشت. یک راه این است که مسئله بدوضع را به یک مسئله‌ی معادل خوش‌وضع تبدیل و سپس روش عددی را روی آن اعمال کنیم.

مثال ۱۳.۲. معادله دیفرانسیل مقدار اولیه‌ی زیر را در نظر بگیرید:

$$y' = ۱۰۰y - ۱۰۱e^{-t}, \quad y(۰) = y_۰, \quad t > ۰, \tag{۳۵.۲}$$

که به ازای $y_۰ = ۱$ دارای جواب دقیق $y(t) = e^{-t}$ است. یکی از ورودی‌های این مسئله مقدار شرط اولیه آن یعنی $y_۰$ است و جواب مسئله تابع $y(t)$ است. حال مسئله‌ی اختلال یافته‌ی زیر را درنظر بگیرید

$$y' = ۱۰۰y - ۱۰۱e^{-t}, \quad y(۰) = y_۰ + \delta,$$

طوری‌که δ یک اختلال کوچک در مقدار شرط اولیه است. به‌سادگی می‌توان نشان داد جواب این مسئله‌ی اختلال یافته به ازای $y_۰ = ۱$ عبارتست از $y_\delta(t) = e^{-t} + \delta e^{۱۰۰t}$. از این‌رو

$$y_\delta(t) - y(t) = \delta e^{۱۰۰t},$$

که واضح است با افزایش t به‌سرعت رشد می‌کند. بنابراین مسئله‌ی (۳۵.۲) نسبت به اختلال در مقدار شرط اولیه خود با افزایش t بدوضع است.

برای بررسی میزان حساسیت جواب نسبت به تغییرات جزئی در ورودی نیازمند یک متر و معیار مناسب هستیم. این متر را *ضریب وضعیت* می‌گوییم. در اینجا ضریب وضعیت خاصیتی از نگاشت f خواهد بود. برای بدست آوردن ضریب وضعیتِ نگاشت f در نقطه‌ی x، ابتدا فرض می‌کنیم وقتی x را تغییر می‌دهیم نگاشت f آن را با بی‌نهایت رقم (دقیق) محاسبه می‌کند. در حقیقت عدد وضعیت در این حالت در ذات f است و هیچ ربطی به الگوریتمی که قرار است آن را به صورت عددی محاسبه کند ندارد. اگر بدانیم نگاشت f نسبت به تغییر x به $x + \delta x$ چقدر حساسیت نشان می‌دهد آنگاه می‌توانیم تغییرات y را نیز حساب کنیم.

۲.۶.۲ ضریب وضعیت

با حالت ساده‌ی یک متغیره شروع می‌کنیم. فرض کنیم $m = n = ۱$. ابتدا فرض می‌کنیم $x \neq ۰, y \neq ۰$ و δx تغییرات جزئی در x و δy تغییرات متناظر در y باشد. در فرم صریح (۳۱.۲) با فرض مشتق‌پذیری f در x از بسط تیلر داریم

$$\delta y = f(x + \delta x) - f(x) = f'(x)\delta x + \mathcal{O}(|\delta x|^۲).$$

با چشم‌پوشی از جمله‌ی خطا، از دید نسبی داریم

$$\frac{\delta y}{y} \approx \frac{x f'(x)}{f(x)} \cdot \frac{\delta x}{x}, \quad |\delta x| \ll ۱.$$

تعریف می‌کنیم

$$(\text{cond } f)\,(x) := \left| \frac{x f'(x)}{f(x)} \right| \tag{۳۶.۲}$$

که عبارت سمت چپ ضریب وضعیت نگاشت f در نقطه‌ی x است. بنابراین می‌توان نوشت

$$\left|\frac{\delta y}{y}\right| \approx (\text{cond } f)(x) \left|\frac{\delta x}{x}\right|.$$

این ضریب به ما می‌گوید اختلال نسبی در y (جواب)، چقدر نسبت به اختلال نسبی در x (داده ورودی) بزرگ است. اگر $x = ۰$ و $y \neq ۰$ بهتر است δx را مطلق و δy را نسبی در نظر بگیریم که در این صورت داریم

$$(\text{cond } f)(x) = \left|\frac{f'(x)}{f(x)}\right|, \qquad \left|\frac{\delta y}{y}\right| \approx (\text{cond } f)(x) \ |\delta x|.$$

اگر $x \neq ۰$ و $y = ۰$ باز به همین صورت عمل می‌کنیم و δx را نسبی و δy را مطلق در نظر می‌گیریم. اما اگر $y = x = ۰$ آنگاه

$$(\text{cond } f)(x) = |f'(x)|, \qquad |\delta y| \approx (\text{cond } f)(x) \ |\delta x|,$$

که هر دوی ورودی و جواب بطور مطلق بررسی شده‌اند. گاهی در دو حالت $x = ۰, y \neq ۰$ و $x \neq ۰, y = ۰$ همانند حالت آخر عمل می‌شود و هر دوی δx و δy به صورت مطلق لحاظ می‌شوند.

اکنون حالتی که m و n دلخواه‌اند را بررسی می‌کنیم. قرار می‌دهیم

$$x = [x_۱, x_۲, \ldots, x_m]^T \in D \subseteq \mathbb{R}^m, \quad y = [y_۱, y_۲, \ldots, y_n]^T \in \mathbb{R}^n$$

و نگاشت f را درایه به درایه به صورت

$$y_j = f_j(x_۱, x_۲, \ldots, x_m), \quad j = ۱, ۲, \ldots, n$$

در نظر می‌گیریم. فرض می‌کنیم f_j ها دارای مشتقات جزئی نسبت به تمام m متغیر x_i هستند. اختلال نسبی در x را با

$$\frac{\|\delta x\|_{\mathbb{R}^m}}{\|x\|_{\mathbb{R}^m}}, \quad \delta x = [\delta x_۱, \ldots, \delta x_m]^T$$

و اختلال نسبی در y را با

$$\frac{\|\delta y\|_{\mathbb{R}^n}}{\|y\|_{\mathbb{R}^n}}, \quad \delta y = [\delta y_۱, \ldots, \delta y_n]^T$$

اندازه‌گیری می‌کنیم و سعی‌مان این است که اختلال در جواب را با اختلال در ورودی مرتبط کنیم. از بسط تیلر چند متغیره داریم

$$\delta y_j = f_j(x + \delta x) - f_j(x) = \sum_{i=۱}^{m} \frac{\partial f_j}{\partial x_i}(x)\delta x_i + \mathcal{O}(\|\delta x\|^۲).$$

با چشم‌پوشی از جمله‌ی خطا داریم

$$|\delta y_j| \leqslant \sum_{i=۱}^{m} \left|\frac{\partial f_j}{\partial x_i}\right| \ |\delta x_i| \leqslant \max_i |\delta x_i| \times \sum_{i=۱}^{m} \left|\frac{\partial f_j}{\partial x_i}\right|$$
$$\leqslant \max_i |\delta x_i| \times \max_j \sum_{i=۱}^{m} \left|\frac{\partial f_j}{\partial x_i}\right|.$$

از آنجا که این رابطه برای هر $j = ۱, ۲, \dots, n$ برقرار است، برای $\max_j |\delta y_j|$ نیز برقرار است. پس

$$\|\delta y\|_\infty \leqslant \|\delta x\|_\infty \left\| \frac{\partial f}{\partial x} \right\|_\infty . \tag{۳۷.۲}$$

در اینجا

$$\frac{\partial f}{\partial x} = J_f := \begin{bmatrix} \frac{\partial f_۱}{\partial x_۱} & \frac{\partial f_۱}{\partial x_۲} & \dots & \frac{\partial f_۱}{\partial x_m} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_۱} & \frac{\partial f_n}{\partial x_۲} & \dots & \frac{\partial f_n}{\partial x_m} \end{bmatrix} \in \mathbb{R}^{n \times m}$$

به *ماتریس ژاکوبین* معروف است. (ماتریس ژاکوبین متناظر با مشتق مرتبه اول برای توابع چند متغیره است.) با توجه به (۳۷.۲) داریم:

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leqslant \frac{\|x\|_\infty \|J_f(x)\|_\infty}{\|f(x)\|_\infty} \cdot \frac{\|\delta x\|_\infty}{\|x\|_\infty}.$$

اگر چه این رابطه به صورت نامعادله برقرار است اما تساوی برای δx های مناسب تقریباً برقرار خواهد بود، بنابراین تعریف می‌کنیم

$$(\operatorname{cond} f)(x) := \frac{\|x\|_\infty \|J_f(x)\|_\infty}{\|f(x)\|_\infty}. \tag{۳۸.۲}$$

به وضوح وقتی $m = n = ۱$، رابطه‌ی (۳۶.۲) حاصل می‌شود.

مثال ۱۴.۲. برای $g \in C^۱$ حل معادله‌ی غیرخطی $g(y) = x$ برای ورودی x و جواب y مد نظر است. در حالتی که $x = ۰$ این مسأله همان مسأله‌ی ریشه‌یابی است. اگر g در همسایگی y معکوس‌پذیر باشد این مسأله خوش تعریف است. در این حالت

$$y = g^{-۱}(x) := f(x), \qquad f'(x) = \frac{۱}{g'(y)}$$

و داریم

$$(\operatorname{cond} f)(x) = \left| \frac{۱}{g'(y)} \right| \frac{|x|}{|y|} \quad x, y \neq ۰$$

اگر $x = ۰$ (مسئله‌ی ریشه‌یابی) یا $y = ۰$ آنگاه

$$(\operatorname{cond} f)(x) = \left| \frac{۱}{g'(y)} \right|.$$

بنابراین در مسئله ریشه یابی اگر ریشه‌ی y تکراری باشد ($g'(y) = ۰$) مسئله بدوضع است.

مثال ۱۵.۲. جواب‌های معادله جبری درجه دوم $y^۲ - ۲xy + ۱ = ۰$ که در آن $x \geqslant ۱$ است را در نظر بگیرید. فرض کنید x ورودی و y جواب باشد. در فرم ضمنی می‌توان نوشت $F(x, y) = y^۲ - ۲xy + ۱ = ۰$ و با حل y بر حسب x به فرم صریح داریم $y_\pm = x \pm \sqrt{x^۲ - ۱} := f_\pm(x)$. داریم $f'_\pm(x) = ۱ \pm \frac{x}{\sqrt{x^۲ - ۱}}$ و

$$(\operatorname{cond} f)(x) = \frac{|x|}{\sqrt{x^۲ - ۱}}, \quad x > ۱.$$

بنابراین وقتی x از ۱ دور باشد (ریشه‌ها متمایز باشند) مسئله خوش‌وضع است. در حالتی که $x = ۱$ یعنی ریشه‌ها مضاعف باشند، مسأله بدوضع است. با یک تغییر متغیر می‌توان این مسئله را به یک مسئله خوش‌وضع تبدیل کرد. فرض کنیم $t = x + \sqrt{x^۲ - ۱}$. داریم

$$F(t, y) = y^۲ - \frac{۱ + t^۲}{t} y + ۱ = ۰,$$

و ریشه‌ها به‌صورت $y_- = t$ و $y_+ = ۱/t$ خواهند بود که برای $t = ۱$ برابر هستند. این تغییر پارامتر تکینگی مسئله را – که در فرم قبل برای حالت $x = ۱$ به وجود آمد – از بین می‌برد. در فرم جدید هر دو ریشه $y_- = y_-(t)$ و $y_+ = y_+(t)$ توابعی هموار نسبت به t در همسایگی $t = ۱$ هستند. با محاسبه‌ی ضریب وضعیت، برای هر مقدار t داریم $(\text{cond } f)(t) = ۱$. بنابراین فرم جدید مسئله خوش‌وضع خواهد بود.

مثال ۱۶.۲. معادله غیرخطی زیر را در نظر بگیرید

$$x^n = ae^{-x}, \quad a > ۰, \quad n \geq ۱.$$

این معادله برای هر n دقیقاً یک ریشه مثبت $\xi(a)$ دارد. چراکه اگر دو ریشه‌ی مثبت ξ و ζ برای این معادله مفروض باشند و بدون اینکه از کلیت کاسته شود فرض کنیم $\xi > \zeta$ آنگاه $\xi^n > \zeta^n$ و $-ae^{-\xi} > -ae^{-\zeta}$ و بنابراین $\xi^n - ae^{-\xi} > \zeta^n - ae^{-\zeta}$ که با فرض ریشه بودن ξ و ζ در تناقض است. اگر قرار دهیم $f(x) = x^n - ae^{-x}$ آنگاه $f(۰) = -a < ۰$ و $f(\sqrt[n]{a}) = a - ae^{-\sqrt[n]{a}} > ۰$، پس این ریشه‌ی یکتا، در $(۰, \sqrt[n]{a})$ قرار دارد. حال نشان می‌دهیم $\xi(a)$ به عنوان تابعی از a خوش‌وضع است. داریم $[\xi(a)]^n - ae^{-\xi(a)} = ۰$. با مشتق‌گیری ضمنی از ξ نسبت به a داریم

$$n\xi^{n-۱}\frac{d\xi}{da} - e^{-\xi} + ae^{-\xi}\frac{d\xi}{da} = ۰ \quad \text{یا} \quad \xi'(a) = \frac{d\xi}{da} = \frac{e^{-\xi}}{n\xi^{n-۱} + ae^{-\xi}}.$$

حال ضریب وضعیت به صورت زیر قابل محاسبه است:

$$(\text{cond } \xi)(a) = \frac{\xi'(a)}{\xi(a)} a = \frac{ae^{-\xi}}{n\xi^n + a\xi e^{-\xi}} = \frac{ae^{-\xi}}{nae^{-\xi} + a\xi e^{-\xi}} = \frac{۱}{n + \xi} < \frac{۱}{n}.$$

نامساوی بالا نشان می‌دهد ξ بعنوان تابعی از a خوش‌وضع است، یعنی تغییرات جزئی در a، ریشه‌ی ξ را دچار اختلال زیاد نخواهد کرد.

مثال ۱۷.۲ (ضریب وضعیت دستگاه معادلات خطی). دستگاه معادلات خطی

$$Ax = b, \quad A \in \mathbb{R}^{n \times n}, \quad b \in \mathbb{R}^n$$

را در نظر می‌گیریم که در آن ماتریس معکوس‌پذیر A و بردار $b \neq ۰$ ورودی و x جواب است. برای بررسی وضعیت این مسئله ابتدا فرض کنیم به درایه‌های ماتریس A اختلالی وارد نشود و فقط بردار b دچار تغییر شود. بنابراین

$$\begin{cases} f : \mathbb{R}^n \to \mathbb{R}^n \\ x = f(b) := A^{-۱}b, \end{cases}$$

با توجه به اینکه $\partial f/\partial b = A^{-1}$ طبق تعریف (۳۸.۲) داریم

$$(\text{cond } f)(b) = \frac{\|b\|\|A^{-1}\|}{\|A^{-1}b\|} = \frac{\|Ax\|\|A^{-1}\|}{\|x\|}, \qquad (Ax = b \quad \text{که}),$$

که در آن از p-نرم‌های برداری و ماتریسی استفاده شده است. با توجه به اینکه یک رابطه‌ی یک به یک بین x و b برقرار است می‌توان بدترین (بزرگترین) عدد وضعیت را به صورت زیر یافت

$$\max_{0\neq b\in\mathbb{R}^n} (\text{cond } f)(b) = \max_{0\neq x\in\mathbb{R}^n} \frac{\|Ax\|}{\|x\|} \cdot \|A^{-1}\| = \|A\| \cdot \|A^{-1}\|.$$

آخرین عبارت سمت راست به b وابسته نیست و به آن ضریب *وضعیت ماتریس* A می‌گویند. بنابراین تعریف می‌کنیم

$$\text{cond } (A) := \|A\| \cdot \|A^{-1}\|. \tag{۳۹.۲}$$

چون از p-نرم ماتریسی استفاده کرده‌ایم، ضریب وضعیت را با $\text{cond}_p(A)$ نشان می‌دهیم که به آن ضریب وضعیت ماتریس A در نرم p گفته می‌شود. لازم به ذکر است که این عدد، وضعیت یک دستگاه معادلات خطی با ماتریس ضرایب A را اندازه‌گیری می‌کند، نه وضعیت کمیت‌های دیگر وابسته به A مانند مقادیر ویژه و غیره را. اگر اختلالی به اندازه δb به b وارد شود، متناظر با آن اختلال δx در جواب x ایجاد می‌شود به طوریکه $A(x+\delta x) = b + \delta b$. از این رو داریم $A\delta x = \delta b$ یا $\delta x = A^{-1}\delta b$. گرفتن نرم از طرفین نتیجه می‌دهد $\|\delta x\| \leqslant \|A^{-1}\|\,\|\delta b\|$. از طرفی با گرفتن نرم از طرفین $Ax = b$ داریم $\frac{1}{\|x\|} \leqslant \frac{\|A\|}{\|b\|}$، که این هر دو نتیجه می‌دهند

$$\frac{\|\delta x\|}{\|x\|} \leqslant \text{cond } (A)\frac{\|\delta b\|}{\|b\|}. \tag{۴۰.۲}$$

رابطه بالا نشان می‌دهد اگر ضریب وضعیت ماتریس A بزرگ باشد، اختلال جزئی در ورودی باعث اختلال زیاد در جواب x می‌شود. در این حالت گوییم دستگاه بدوضع است.

اگر بجای بردار سمت راست b، اختلال به ماتریس A وارد شود و یا در حالت کلی اگر به هر دوی A و b اختلال وارد شود باز هم می‌توان نشان داد اختلال به وجود آمده در جواب x به ضریب وضعیت ماتریس A وابسته است.

در متلب دستور

```
cond(A,p)
```

ضریب وضعیت ماتریس A در نرم p را ارائه می‌دهد که p می‌تواند 1، 2، `inf`، و `'fro'` را به ترتیب برای نرم‌های ماتریسی یک، دو، بینهایت و فروبنیوس اختیار کند.

حال وضعیت چند ماتریس که به وفور در آنالیز و محاسبات عددی ظاهر می‌شوند را بررسی می‌کنیم. اولی ماتریس هیلبرت است که در مثال ۱۰.۲ ارائه شد. در جدول زیر ضریب وضعیت ماتریس هیلبرت به ازای چند مقدار n آمده است.

n	۱۰	۲۰	۴۰
$\text{cond}_2 H_n$	$۱/۶۰ \times ۱۰^{۱۳}$	$۲/۴۵ \times ۱۰^{۲۸}$	$۷/۶۵ \times ۱۰^{۵۸}$

یک دستگاه معادلات خطی مرتبه ۱۰ با ماتریس هیلبرت در "دقت معمولی" قابل حل نخواهد بود. همچنین "دقت دوبرابر" برای یک ماتریس هیلبرت مرتبه $n = ۲۰$ کافی نخواهد بود. اثبات شده است

$$\text{cond}_۲(H_n) \sim \frac{(\sqrt{۲}+۱)^{۴n+۴}}{۲^{\frac{۱۵}{۴}}\sqrt{\pi n}}, \quad n \to \infty \text{ وقتی} .$$

بنابراین ماتریس هیلبرت به عنوان یک ماتریس بدوضع شناخته می‌شود.

حال به مثال ۱۰.۲ بر می‌گردیم. درایه‌های ماتریس هیلبرت ۱۱×۱۱ و بردار سمت راست کسری‌اند و به صورت دقیق در کامپیوتر ذخیره نمی‌شوند. در واقع در دقت پیش‌فرض متلب (دقت دوبرابر) داریم

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} \approx ۱۰^{-۱۶}, \quad \frac{\|\delta b\|_\infty}{\|b\|_\infty} \approx ۱۰^{-۱۶}.$$

حتی اگر روش عددی (در اینجا روش تجزیه LU با محورگیری) هیچ خطایی تولید نکند و حتی اگر از خطای درایه‌های A صرف نظر کنیم، طبق (۴۰.۲) و با علم به اینکه $\text{cond}_\infty(A) \approx ۱/۲۵ \times ۱۰^{۱۵}$، داریم

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \lesssim ۱۰^{-۱۶} \times ۱/۲۵ \times ۱۰^{۱۵} = ۰/۱۲۵$$

که با نتیجه بدست آمده در آن مثال همخوانی دارد.

ماتریس مشهور دیگر *ماتریس واندرموند* است که برای مقادیر حقیقی و متمایز $t_۱$، $t_۲$، ...، t_n، به صورت زیر تعریف می‌شود:

$$V_n = \begin{bmatrix} ۱ & ۱ & \cdots & ۱ \\ t_۱ & t_۲ & \cdots & t_n \\ \vdots & \vdots & & \vdots \\ t_۱^{n-۱} & t_۲^{n-۱} & \cdots & t_n^{n-۱} \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

حال چند مثال که از انتخاب‌های مختلف مقادیر t_k بدست می‌آیند را بررسی می‌کنیم. اگر این مقادیر در بازه‌ی $[-۱, ۱]$ به صورت هم‌فاصله با

$$t_k := ۱ - \frac{۲(k-۱)}{n-۱}, \quad k = ۱, ۲, \ldots, n$$

تعریف شوند، آنگاه ثابت شده است

$$\text{cond}_\infty(V_n) \sim \frac{۱}{\pi} e^{-\frac{\pi}{۴}} e^{n(\frac{\pi}{۴}+\frac{۱}{۲}\ln ۲)}, \quad n \to \infty.$$

این حالت به عنوان مثال در مسئله‌ی درونیابی به کمک چندجمله‌ایها ظاهر می‌شود که در فصل‌های بعدی به آن خواهیم رسید. در جدول زیر چند مقدار از ضریب وضعیت‌ها در نرم بینهایت آمده است.

n	۱۰	۲۰	۴۰	۸۰
$\text{cond}_\infty V_n$	$۱/۳۶ \times ۱۰^{۴}$	$۱/۰۵ \times ۱۰^{۹}$	$۶/۹۳ \times ۱۰^{۱۸}$	$۳/۱۵ \times ۱۰^{۳۸}$

با اینکه رشد ضریب وضعیت ماتریس واندرموند در این حالت به اندازه‌ی ماتریس هیلبرت نیست، اما باز رشد آن نمایی است و بنابراین یک ماتریس بدوضع است. شرایط بدتر خواهد شد اگر مقادیر t_k اعداد هارمونیک

$$t_k = \frac{۱}{k}, \quad k = ۱, ۲, \ldots, n$$

انتخاب شوند. در این صورت $\mathrm{cond}_\infty(V_n) > n^{n+۱}$ که رشد بسیار بیشتر از حالت نمایی دارد.

۷.۲ پرسش‌ها

۱. دستگاه نمایش ممیز شناور نرمال در مبنای ۲ را در نظر بگیرید که هر عدد حقیقی در آن به صورت $\pm ۰/d_۱d_۲d_۳ \times ۲^e$ نمایش داده می‌شود که $d_j \in \{۰, ۱\}$ و $-۱ \leqslant e \leqslant ۱$. فرض کنید برای نمایش دیگر اعداد از گردکردن استفاده شود. اپسیلون دستگاه و واحد گردکردن و کوچکترین و بزرگترین اعداد مثبت و منفی و تعداد اعداد قابل نمایش با این دستگاه را محاسبه کنید. اگر دستگاه $\pm ۰/d_۱d_۲d_۳d_۴ \times ۲^e$ باشد و $-۳ \leqslant e \leqslant ۳$، موارد بالا را یکبار دیگر محاسبه کنید.

۲. با مثال نشان دهید در $\mathbb{F}(\beta, t, L, U)$ قاعده شرکت پذیری جمع برقرار نیست و همچنین عضو خنثی عمل جمع منحصربفرد نمی‌باشد.

۳. روابط (۱۰.۲) را اثبات کنید.

۴. نشان دهید اگر $a \in \mathbb{R}$ و $fl_{\mathrm{up}}(a)$ گرد شده‌ی a به بالا در $\mathbb{F}(\beta, t, L, U)$ باشد، آنگاه

$$\left|a - fl_{\mathrm{up}}(a)\right| \leqslant \beta^{e-t}, \qquad \frac{\left|a - fl_{\mathrm{up}}(a)\right|}{|a|} \leqslant \beta^{۱-t}.$$

این نامساوی‌ها را برای گردکردن به پایین نیز اثبات کنید.

۵. ثابت کنید اگر $x \in \mathbb{R}$ در دامنه اعداد ممیز شناور باشد، داریم

$$fl(x) = \frac{x}{۱ + \varepsilon}, \quad |\varepsilon| \leqslant u.$$

۶. استاندارد IEEE با دقت دوبرابر را در نظر بگیرید. تعداد اعضای بین ۱ و ۲ چقدر است؟ اعداد در این دستگاه حداکثر چند رقم در مبنای ۱۰ دارند؟ بزرگترین و کوچکترین اعداد قابل نمایش از لحاظ قدر مطلق چقدرند؟ اپسیلون ماشین را بر حسب توان ۱۰ بدست آورید.

۷. در استاندارد IEEE با دقت معمولی، بازه‌ای را تعیین کنید در آن فاصله‌ی اعداد ممیز شناور برابر ۱ است.

۸. در استاندارد IEEE چند عدد ممیز شناور با دقت دوبرابر بین دو عدد متوالی غیرصفر با دقت معمولی وجود دارد؟

۹. تعداد ارقام بامعنای درست دهدهی تقریب‌های زیر را بدست آوردید.

$$x = ۳۱/۹۹۸۲۳, \quad \widehat{x} = ۳۲/۰۰۱۲۹,$$
$$x = ۲/۵۳۰۱۲۳۲, \quad \widehat{x} = ۲/۵۳۰۲۳۹۴,$$
$$x = ۰/۰۰۵۲۱۰, \quad \widehat{x} = ۰/۰۰۵۹۱.$$

۱۰. دو عدد دهدهی $x = ۲/۴۳۱$ و $\widehat{x} = ۲/۴۲۵$ را در مبنای دو بنویسید و تعداد ارقام بامعنای درست دودویی یکسان آنها را پیدا کنید.

۱۱. نشان دهید اگر در لم ۲.۲ داشته باشیم $\rho_k = ۱, \forall k$ آنگاه کران بهتر $|\theta_n| \leqslant nu/(۱ - nu/۲)$ برای $nu < ۲$ برقرار است.

۱۲. با فرض اینکه ماشین عمل ضرب-جمع ترکیبی را پشتیبانی کند، آنالیزهای خطای پسرو و پیشرو جدید برای الگوریتم ضرب داخلی ارائه دهید.

۱۳. شکل دیگری برای عبارات زیر جهت جلوگیری از خطای حذف بنویسید.

$$۱ - \cos x, \quad |x| \ll ۱,$$
$$\sin x - \cos x, \quad |x| \approx \pi/۴,$$
$$\ln(\sqrt{x^۲ + ۱} - x), \quad |x| \gg ۱$$
$$\sin x - \sin y, \quad x \approx y.$$

۱۴. ضریب وضعیت توابع زیر را تعیین کنید و امکان وقوع بدوضعی را مشخص نمایید.

$$(a)\ \ f(x) = \ln x, \quad x > ۰$$
$$(b)\ \ f(x) = \sin^{-۱} x, \quad |x| \leqslant ۱$$
$$(c)\ \ f(x) = \sin^{-۱} \frac{x}{\sqrt{۱ + x^۲}}$$
$$(d)\ \ f(x) = x^{۱/n}, \quad x > ۰, n \in \mathbb{N}$$

۱۵. وضعیت مسئله‌ی $x_{\pm} = -p \pm \sqrt{p^۲ + q}$ برای معادله درجه دوم $x^۲ + ۲px - q = ۰$ را نسبت به اختلال در p و q به صورت مجزا بررسی کنید.

۱۶. برای تابع ترکیبی $h(x) = g(f(x))$ ضریب وضعیت h را بر حسب ضرایب وضعیت g و f بنویسید. دقت کنید که ضرایب وضعیت هر یک در چه نقطه‌ای محاسبه می‌شوند. نتایج را برای تابع $h(x) = \frac{۱+\sin x}{۱-\sin x}$ در $x = \frac{\pi}{۴}$ بکار گیرید.

۱۷. قرار دهید $f(x, y) = x + y$ و ضریب وضعیت این نگاشت در نرم بینهایت را بدست آورید. نشان دهید اگر x و y هم‌علامت باشند عمل جمع آن‌ها خوش‌وضع است، اما اگر x و y علامتشان عکس هم و اندازه‌ی آنها تقریباً برابر باشد، عمل جمع بدوضع خواهد بود.

۱۸. معادله جبری زیر را در نظر بگیرید:

$$x^n + ax - ۱ = ۰, \quad a > ۰, \quad n \geq ۲.$$

الف. نشان دهید این معادله دقیقاً یک ریشه مثبت $\xi(a)$ دارد.

ب. فرمولی برای $(\text{cond } \xi)(a)$ بیابید.

ج. کرانهای پایین و بالای مناسبی برای $(\text{cond } \xi)(a)$ بدست آورید.

فصل ۳

درونیابی و تقریب

فرض کنیم مقادیر یک تابع در تعداد متناهی نقطه از دامنه‌اش در دست باشد. هدف از مسئله‌ی تقریب، یافتن الگویی (غالباً) پیوسته برای این مجموعه از نقاط است بطوریکه این الگو تا آنجا که ممکن است به تابع اصلی نزدیک باشد. بعلاوه همواره علاقه‌مندیم الگوی ساخته شده دارای ساختار ساده تری نسبت به ساختار تابع اصلی باشد. به فضایی از توابع که الگوی ما به آن تعلق دارد *فضای تقریب* می‌گوییم. یکی از فضاهای تقریبِ مناسب، که مبنای اصلی این فصل کتاب (و البته فصل‌های بعدی) نیز می‌باشد، فضای چندجمله‌ایهای حداکثر از درجه‌ی n روی $\mathbb{R}$ است. از این پس این فضا را با نماد $\mathbb{P}_n$ نمایش می‌دهیم. همانگونه که می‌دانیم این فضا متناهی‌البعد با بعد $n+1$ و یک پایه برای آن مجموعه‌ی تک‌جمله‌ایهای

$$\{1, x, x^2, \ldots, x^n\}$$

است، یعنی هر چندجمله‌ای درجه n مانند p_n را می‌توان به صورت ترکیب خطی زیر نوشت

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n, \quad x \in \mathbb{R}. \tag{۱.۳}$$

در ادامه‌ی این فصل برای یافتن روش‌های عددی مناسب، پایه‌های دیگری نیز برای $\mathbb{P}_n$ معرفی خواهیم کرد. یکی از روش‌های تقریب روش *درونیابی* است. مسئله‌ی درونیابی به صورت زیر معرفی می‌شود.

مسئله (درونیابی چندجمله‌ای)**.** فرض کنیم $[a,b]$ بازه‌ای در $\mathbb{R}$ باشد و نقاط متمایز x_0، x_1، ...، x_n همگی در $[a,b]$ واقع شده باشند. همچنین فرض کنیم f_0، f_1، ...، f_n مقادیری دلخواه در $\mathbb{R}$ باشند. چندجمله‌ای $p_n \in \mathbb{P}_n$ را بیابید بگونه‌ای که

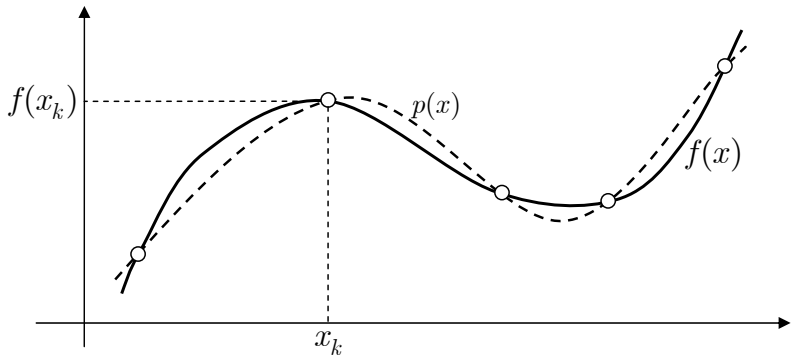
$$p_n(x_k) = f_k, \quad k = 0, 1, \ldots, n. \tag{۲.۳}$$

به x_k نقاط درونیابی و به $n+1$ شرط (۲.۳) *شرایط درونیابی* می‌گوییم. نقاط درونیابی را گاهی با مجموعه‌ی

$$X = \{x_0, x_1, \ldots, x_n\}$$

نمایش می‌دهیم که نمایش آنها در یک مجموعه، متمایز بودن آنها را نیز نشان می‌دهد.

در شکل ۱.۳ شمایی از مسئله‌ی درونیابی روی پنج نقطه نشان داده شده است. همانگونه که مشاهده می‌کنیم تابع f و درونیاب آن در نقاط درونیابی یکدیگر را قطع می‌کنند.



شکل ۱.۳: شمایی از مسئله‌ی درونیابی

به نظر می‌رسد شرایط درونیابی برای یافتن p_n کافی باشند، زیرا کافی است $n+1$ ضریب a_k، $k = 0, 1, \ldots, n$ تعیین شوند. اگر مقادیر f_k از یک تابع پیوسته حقیقی-مقدار مانند $f \in C[a,b]$، بدست آمده باشند، یعنی بازای هر k داشته باشیم $f_k = f(x_k)$، آنگاه با توجه به مسئله‌ی درونیابی، خطای

$$f(x) - p_n(x)$$

در نقاط درونیابی x_k برابر صفر است. انتظار داریم برای دیگر نقاط $x \in [a,b]$ نیز این خطا (اگرچه صفر نیست اما) حداقل ناچیز باشد. این انتظارِ ماست، اما خواهیم دید حداقل در مورد فضای چندجمله‌ایها این انتظار گاهی برآورده نمی‌شود.

یک روش سرراست برای تعیین چندجمله‌ای درونیاب وجود دارد که به صورت زیر بیان می‌شود: چندجمله‌ای p_n را به صورت (۱.۳) بسط می‌دهیم و برای تعیین ضرایب a_k، شرایط درونیابی (۲.۳) را اِعمال می‌کنیم. در این صورت داریم

$$\begin{aligned}
a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n &= f_0 \\
a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n &= f_1 \\
\vdots \qquad\qquad & \qquad \vdots \\
a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_n x_n^n &= f_n
\end{aligned}$$

که می‌توان آن را به شکل ماتریسی زیر

$$\begin{bmatrix}
1 & x_0 & x_0^2 & \cdots & x_0^n \\
1 & x_1 & x_1^2 & \cdots & x_1^n \\
\vdots & \vdots & \vdots & & \vdots \\
1 & x_n & x_n^2 & \cdots & x_n^n
\end{bmatrix}
\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}
=
\begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{bmatrix}$$

یا به صورت فشرده‌ی

$$V\boldsymbol{a} = \boldsymbol{f}$$

نمایش داد، که در آن $\boldsymbol{a} \in \mathbb{R}^{n+1}$ بردار ضرایب مجهول، $\boldsymbol{f} \in \mathbb{R}^{n+1}$ بردار مقادیر معلوم و $V \in \mathbb{R}^{(n+1)\times(n+1)}$ ماتریس ضرایب است. همانگونه که مشاهده می‌کنیم، ماتریس درونیابیِ V، ماتریس واندرموند است که در بخش ۶.۲ فصل قبل معرفی شد. با توجه به اینکه ماتریس واندرموند برای نقاط متمایز x_k معکوس‌پذیر است (پرسش ۴ را ببینید)، بردار ضرایب $\boldsymbol{a}$ به صورت یکتا تعیین خواهد شد و بنابراین چندجمله‌ای p_n به صورت یکتا بدست می‌آید، که این وجود و یکتایی چندجمله‌ای درونیاب را اثبات می‌کند. بنابراین قضیه زیر را داریم:

قضیه ۱.۳. *یک و تنها یک چندجمله‌ای $p_n \in \mathbb{P}_n$ وجود دارد که در مسئله‌ی درونیابی چندجمله‌ای با $n+1$ نقطه‌ی متمایز صدق می‌کند.* □

مثال ۱.۳. چندجمله‌ای درونیاب درجه دو روی نقاط

x_k	-1	0	1
f_k	-1	0	3

با حل دستگاه

$$\begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 3 \end{bmatrix}$$

بدست می‌آید. جواب این دستگاه عبارت است از $[a_0, a_1, a_2] = [0, 2, 1]$، بنابراین $p_2(x) = x^2 + 2x$. ◇

اگر چه روش ارائه شده در بالا از دیدگاه نظری دارای اهمیت است و وجود و یکتایی p_n را اثبات می‌کند، اما از دیدگاه عددی روشی ناکارامد است. زیرا همانگونه که در بخش ۶.۲ مشاهده کردیم، ماتریس واندرموند بدوضع است و ضریب وضعیت آن با افزایش n به سرعت افزایش می‌یابد و بنابراین بردار جواب $\boldsymbol{a}$ با دقت مناسبی در کامپیوتر محاسبه نخواهد شد. در واقع این روش *ناپایدار* است. حتی اگر این ماتریس خوش‌وضع می‌بود، چون ماتریسی پُر است هزینه‌ی محاسباتی حل دستگاه با روش‌های تجزیه حدود $\frac{2}{3}n^3$ است. در بخش‌های بعدی روش‌هایی با پایداری مناسب‌تر و هزینه‌ی کمتر ارائه خواهیم داد.

ملاحظه ۱.۳. اگر درونیابی در یک زیرفضای متناهی‌البعد دلخواه مانند $\mathcal{U}$ با بعد $n+1$ مد نظر باشد، و به دنبال تابع درونیاب $u \in \mathcal{U}$ روی نقاط (x_k, f_k)، $k = 0, 1, \ldots, n$، باشیم، می‌توان روشی مشابه بالا برای تعیین درونیاب به کار برد. در واقع یک پایه مانند

$$\{u_0(x), u_1(x), \ldots, u_n(x)\}$$

برای $\mathcal{U}$ در نظر می‌گیریم، تابع u را به صورت $u(x) = a_۰u_۰(x) + a_۱u_۱(x) + \cdots + a_nu_n(x)$ بسط می‌دهیم و با اِعمال شرایط درونیابی

$$u(x_k) = f_k, \quad k = ۰, ۱, \ldots, n,$$

به دستگاه

$$\begin{bmatrix} u_۰(x_۰) & u_۱(x_۰) & \cdots & u_n(x_۰) \\ u_۰(x_۱) & u_۱(x_۱) & \cdots & u_n(x_۱) \\ \vdots & \vdots & & \vdots \\ u_۰(x_n) & u_۱(x_n) & \cdots & u_n(x_n) \end{bmatrix} \begin{bmatrix} a_۰ \\ a_۱ \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f_۰ \\ f_۱ \\ \vdots \\ f_n. \end{bmatrix} \qquad \text{یا} \qquad U\boldsymbol{a} = \boldsymbol{f},$$

می‌رسیم. به U ماتریس درونیابی می‌گوییم. اگر U معکوس‌پذیر باشد، مسئله‌ی درونیابی خوش‌تعریف است یعنی دارای جواب یکتاست. اما همانگونه که در درونیابی چندجمله‌ای گفتیم، این روش در عمل هزینه‌ی بالایی دارد و گاهی ناپایدار است، و معمولاً بایستی روش‌های دیگری طراحی شوند. از جمله $\mathcal{U}$ می‌تواند فضای چندجمله‌ایهای مثلثاتی، فضای اسپلاین‌ها و غیره باشد. ♡

در ادامه توجه خود را معطوف به درونیابی چندجمله‌ای می‌کنیم و در ابتدا فرمولی برای خطای درونیابی چندجمله‌ای ارائه می‌دهیم. در این فرمول شرایط هموارای قوی روی f نیاز است و در واقع باید $f \in C^{n+۱}[a, b]$. در بخش‌های بعد کران خطای دیگری نیز ارائه خواهیم داد که این فرض محدود کننده را ندارد. ابتدا لازم است قضیه‌ی رول تعمیم یافته را یادآوری کنیم.

لم ۲.۳. *فرض کنیم تابع $f \in C^\ell[a, b]$ حداقل $\ell + ۱$ ریشه متمایز در $[a, b]$ داشته باشد. آنگاه $f^{(\ell)}$ حداقل یک ریشه در $[a, b]$ دارد.*

برهان. فرض کنیم f دارای ریشه‌های متمایز $\alpha_۰, \ldots, \alpha_\ell$ در $[a, b]$ باشد. روی هر بازه‌ی $[\alpha_{j-۱}, \alpha_j]$، $j = ۱, \ldots, \ell$، چون $f(\alpha_{j-۱}) = f(\alpha_j) = ۰$ طبق قضیه رول f' دارای حداقل یک ریشه مانند β_j است. پس f' دارای حداقل ℓ ریشه روی $[a, b]$ است. به همین ترتیب f'' دارای حداقل $\ell - ۱$ ریشه و در آخر $f^{(\ell)}$ دارای حداقل یک ریشه در $[a, b]$ است. □

قضیه ۳.۳. *گیریم $X = \{x_۰, \ldots, x_n\} \subset [a, b]$ و $f \in C^{n+۱}[a, b]$ داده شده‌اند. فرض کنیم p_n چندجمله‌ای درونیاب f مبتنی بر X باشد، آنگاه برای هر $x \in [a, b]$ وجود دارد $\xi(x) \in [a, b]$ بطوریکه*

$$R_n(f; x) := f(x) - p_n(x) = \frac{(x - x_۰) \cdots (x - x_n)}{(n + ۱)!} f^{(n+۱)}\big(\xi(x)\big). \tag{۳.۳}$$

برهان. اگر $x \in X$، معادله‌ی (۳.۳) بوضوح برقرار است چراکه طرفین برابر صفر خواهند بود. فرض می‌کنیم $x \notin X$، قرار می‌دهیم $\pi_{n+۱}(x) = (x - x_۰) \cdots (x - x_n)$، و تعریف می‌کنیم

$$g(t) = f(t) - p_n(t) - \frac{f(x) - p_n(x)}{\pi_{n+۱}(x)} \pi_{n+۱}(t), \quad t \in [a, b].$$

به روشنی g در $C^{(n+1)}[a,b]$ است و دارای حداقل $n+2$ ریشه شامل تمام اعضای X و خود x است. طبق قضیه رول تعمیم‌یافته $g^{(n+1)}(t)$ دارای حداقل یک ریشه در $[a,b]$، مثلاً ξ، است. به عبارت دیگر $g^{(n+1)}(\xi)=0$. با مشتق‌گیری از تابع $g(t)$ داریم

$$g^{(n+1)}(t)=f^{(n+1)}(t)-0-\frac{f(x)-p_n(x)}{\pi_{n+1}(x)}(n+1)!,$$

که در آن از این واقعیت که $p_n^{(n+1)}(x)=0$ و $\pi_{n+1}^{(n+1)}(x)=(n+1)!$ استفاده کرده‌ایم. با توجه به اینکه $g^{(n+1)}(\xi)=0$، معادله‌ی (۳.۳) به آسانی نتیجه می‌شود. وابستگی ξ به x نیز آشکار است و بهمین علت آن را به عنوان تابعی از x به صورت $\xi(x)$ نشان می‌دهیم. □

در بخش بعد روش دیگری برای تعیین چندجمله‌ای درونیاب معرفی می‌کنیم که به روش درونیابی لاگرانژ مشهور است.

۱.۳ روش لاگرانژ

در روش لاگرانژ پایه‌ی فضای $\mathbb{P}_n$ را طوری می‌سازیم که ماتریس درونیابی یک ماتریس قطری باشد. اگر این پایه را با

$$\{\ell_0(x),\ell_1(x),\dots,\ell_n(x)\}$$

نشان دهیم بایستی اعضای آن وابسته به نقاط درونیابی باشند. برای این منظور تعریف می‌کنیم

$$\ell_k(x)=\prod_{\substack{j=0\\ j\neq k}}^{n}\frac{x-x_j}{x_k-x_j},\quad k=0,1,\dots,n \tag{۴.۳}$$

که $\ell_k(x)$ چندجمله‌ایهایی از درجه n هستند و در شرط دلتای کرونکر زیر صدق می‌کنند

$$\ell_k(x_i)=\delta_{k,i}=\begin{cases}1, & i=k\\ 0, & i\neq k\end{cases},\quad i,k=0,1,\dots,n. \tag{۵.۳}$$

اگر قرار می‌دهیم

$$p_n(x)=\sum_{k=0}^{n}f_k\ell_k(x) \tag{۶.۳}$$

آنگاه به وضوح داریم $p_n(x_k)=f_k$. به چندجمله‌ایهای $\ell_k(x)$ *چندجمله‌ایهای لاگرانژ* می‌گوییم. فرمول درونیابی لاگرانژ به نوعی دیگر وجودِ چندجمله‌ای درونیاب را نیز اثبات می‌کند.

اگر چه هزینه‌ای برای حل دستگاه نخواهیم داشت (چون دستگاه قطری است)، اما برای ساختن چندجمله‌ایهای $\ell_k(x)$ هزینه‌ی محاسباتی داریم. به سادگی می‌توان دید برای نقطه ثابت x، هزینه محاسبه هر $\ell_k(x)$ با فرمول (۴.۳) برابر $3n$

و هزینه‌ی محاسبه همه‌ی آن‌ها $۳n(n+۱)$ است. در آخر محاسبه‌ی p_n در نقطه‌ی x با فرمول (۶.۳) به $۲n+۱$ عمل نیاز دارد. بنابراین هزینه کلی روش لاگرانژ برای محاسبه درونیاب در یک نقطه $۳n^۲+۵n+۱ \approx ۳n^۲$ است. اگر نقطه‌ی جدیدی، مثلاً $x_{n+۱}$، به نقاط درونیابی اضافه شود تمام محاسبات باید از ابتدا انجام شود و محاسبات قبلی استفاده نخواهند شد. همچنین محاسبه مستقیم چندجمله‌ایهای لاگرانژ از دید محاسباتی به ناپایداری منجر خواهد شد، بویژه وقتی برخی از نقاط درونیابی به هم نزدیک باشند که باعث بروز خطای حذف خواهد شد. فرمول درونیابی لاگرانژ را می‌توان اصلاح نمود و فرم دیگری از آن، که به فرمول *گرانیگاهی* معروف است، را بدست آورد که از دید محاسباتی پایدار و کم هزینه‌تر است. در بخش ۴.۳ این روش را تشریح خواهیم کرد. در مثال زیر درونیاب‌های درجه یک و درجه دو را برای یک تابع مفروض بدست می‌آوریم.

مثال ۲.۳. می‌خواهیم تابع $f(x)=\frac{۱}{۱+x}$ را روی نقاط $x_۰=۰$ و $x_۱=۱$ درونیابی کنیم و کران خطای درونیابی را بدست آوریم. ابتدا چندجمله‌ایهای لاگرانژ را محاسبه می‌کنیم. طبق (۴.۳) داریم

$$\ell_۰(x)=\frac{x-x_۱}{x_۰-x_۱}=۱-x, \quad \ell_۱(x)=\frac{x-x_۰}{x_۱-x_۰}=x,$$

و از آنجا $f_۰=f(۰)=۱$ و $f_۱=f(۱)=\frac{۱}{۲}$، درونیاب درجه یک به صورت زیر نوشته می‌شود

$$p_۱(x)=f_۰\ell_۰(x)+f_۱\ell_۱(x)=۱-\frac{۱}{۲}x.$$

طبق قضیه‌ی ۳.۳ کران خطای درونیابی خطی عبارتست از

$$|f(x)-p_۱(x)| \leqslant \frac{۱}{۲!}\max_{x\in[۰,۱]}|x(x-۱)| \times \max_{x\in[۰,۱]}|f''(x)|.$$

اکسترمم تابع $x(x-۱)$ در $۱/۲$ رخ می‌دهد و مقدار آن $-۱/۴$ است. از طرفی $f''(x)=۲/(۱+x)^۳$ و اکسترمم آن در نقطه‌ی ۰ با مقدار ۲ اتفاق می‌افتد. پس داریم

$$|f(x)-p_۱(x)| \leqslant \frac{۱}{۲}\times\frac{۱}{۴}\times ۲ = ۰/۲۵, \quad \forall\, x\in[۰,۱].$$

حال نقطه‌ی $x_۲=۲$ را به نقاط درونیابی اضافه می‌کنیم و درونیاب درجه دو را تعیین می‌کنیم. بایستی چندجمله‌ایهای لاگرانژ را از نو محاسبه کنیم. داریم

$$\begin{aligned}
\ell_۰(x)&=\frac{(x-x_۱)(x-x_۱)}{(x_۰-x_۱)(x_۰-x_۲)}=\frac{۱}{۲}(x-۱)(x-۲),\\
\ell_۱(x)&=\frac{(x-x_۰)(x-x_۲)}{(x_۱-x_۰)(x_۱-x_۲)}=-x(x-۲),\\
\ell_۲(x)&=\frac{(x-x_۰)(x-x_۱)}{(x_۲-x_۰)(x_۲-x_۱)}=\frac{۱}{۲}x(x-۱),
\end{aligned}$$

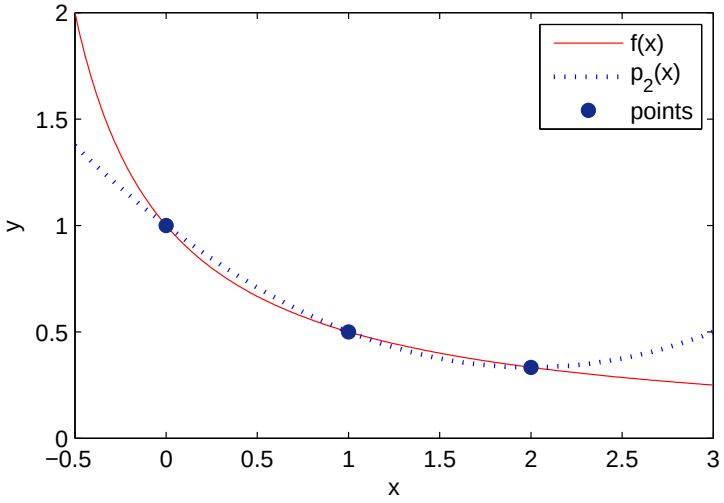
و با توجه به اینکه $f_۲ = f(۲) = \frac{۱}{۳}$ داریم

$$p_۲(x) = f_۰\ell_۰(x) + f_۱\ell_۱(x) + f_۲\ell_۲(x) = \frac{۱}{۶}x^۲ - \frac{۲}{۳}x + ۱.$$

کران خطای درونیابی با محاسبه‌ی اکسترمم‌های تابع $x(x-۱)(x-۲)$ و اکسترمم‌های $f'''(x) = -۶/(۱+x)^۴$ روی $[۰, ۲]$ بدست می‌آید. در مورد اول با محاسبه‌ی ریشه‌های مشتق می‌توان نشان داد اکسترمم‌ها در نقاط $۱ \pm \frac{\sqrt{۳}}{۳}$ با مقادیر $\pm\frac{۲\sqrt{۳}}{۹}$ رخ می‌دهند. اکسترمم تابع $f'''(x)$ هم به روشنی در صفر و با مقدار -۶ قرار دارد. بنابراین می‌توان نوشت

$$\begin{aligned}|f(x) - p_۲(x)| &\leqslant \frac{۱}{۳!}\max_{x\in[۰,۲]}|x(x-۱)(x-۲)| \times \max_{x\in[۰,۲]}|f'''(x)| \\ &= \frac{۱}{۶} \times \frac{۲\sqrt{۳}}{۹} \times ۶ = \frac{۲\sqrt{۳}}{۹}, \quad \forall x \in [۰, ۲].\end{aligned}$$

نمودار تابع f و چندجمله‌ای درونیاب درجه دوم آن بهمراه نقاط درونیابی در شکل ۲.۳ رسم شده است. ♢



شکل ۲.۳: نمودار تابع و درونیاب درجه دوم مثال ۲.۳

در حالت کلی کران خطای درونیابی خطی عبارتست از

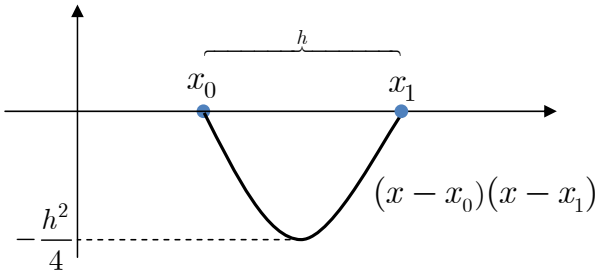
$$\begin{aligned}|f(x) - p_۱(x)| &= \frac{۱}{۲!}|(x-x_۰)(x-x_۱)||f''(\xi)| \\ &\leqslant \frac{۱}{۸}h^۲ \max_{x\in[x_۰,x_۱]}|f''(x)|, \qquad \forall x \in [x_۰, x_۱],\end{aligned} \tag{۷.۳}$$

که در آن $h = x_۱ - x_۰$ و از این واقعیت که

$$\max_{x\in[x_۰,x_۱]}|(x-x_۰)(x-x_۱)| = \frac{(x_۱-x_۰)^۲}{۴} = \frac{h^۲}{۴}$$

استفاده کرده‌ایم. نمودار نوعیِ تابع $(x-x_۰)(x-x_۱)$ در شکل ۳.۳ رسم شده است.

یافتن کران خطای درونیابی درجه دوم روی نقاط هم‌فاصله به عنوان تمرین در پرسش ۱ به عهده‌ی خواننده واگذار شده است.



شکل ۳.۳: نمودار تابع $(x-x_0)(x-x_1)$ و مقدار اکسترمم آن

۲.۳ روش نیوتن

در روش زیبا و کارایی که ایساک نیوتن برای محاسبه‌ی چندجمله‌ای درونیاب یافت، به جای استفاده از پایه‌ی متشکل از تک‌جمله‌ایها، یعنی $\{1, x, \ldots, x^n\}$ یا پایه‌ی متشکل از چندجمله‌ایهای لاگرانژ، یعنی $\{\ell_0(x), \ell_1(x), \ldots, \ell_n(x)\}$ برای فضای $\mathbb{P}_n$، از چندجمله‌ایهای π_0، π_1، ...، π_n، که به صورت

$$\pi_k(x) = \begin{cases} 1, & k = 0 \\ (x-x_0)\cdots(x-x_{k-1}), & 1 \leqslant k \leqslant n \end{cases} \tag{۸.۳}$$

تعریف می‌شوند، استفاده می‌شود. در این صورت چندجمله‌ای درونیاب $p_n(x)$، که در نقاط x_k با $f(x)$ یکی است، به شکل ترکیب خطی زیر نوشته می‌شود

$$p_n(x) = \alpha_0\pi_0(x) + \alpha_1\pi_1(x) + \cdots + \alpha_n\pi_n(x). \tag{۹.۳}$$

ضرایب a_k با اعمال شرایط درونیابی $p_n(x_k) = f(x_k)$، $0 \leqslant k \leqslant n$، تعیین می‌شوند. اعمال این شرایط منجر به دستگاه مثلثی

$$\begin{bmatrix} \pi_0(x_0) & 0 & 0 & \cdots & 0 \\ \pi_0(x_1) & \pi_1(x_1) & 0 & \cdots & 0 \\ \pi_0(x_2) & \pi_1(x_2) & \pi_2(x_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \pi_0(x_n) & \pi_1(x_n) & \pi_2(x_n) & \cdots & \pi_n(x_n) \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix}$$

برای تعیین بردار ضرایب می‌شود. دستگاه بالا را می‌توان به شکل فشرده

$$U\boldsymbol{\alpha} = \boldsymbol{f}$$

نوشت. این دستگاه دارای جواب یکتاست زیرا متمایز بودن x_k ها نتیجه می‌دهد

$$\det(U) = \pi_0(x_0)\pi_1(x_1)\cdots\pi_n(x_n) \neq 0.$$

جواب دستگاه پایین مثلثی بالا را می‌توان با روش جایگذاری پیشرو بدست آورد. در این صورت

$$\alpha_۰ = f(x_۰), \quad \alpha_۱ = \frac{f(x_۱) - \alpha_۰\pi_۰(x_۱)}{\pi_۱(x_۱)} = \frac{f(x_۱) - f(x_۰)}{x_۱ - x_۰}, \quad \dots$$

با توجه به مثلثی بودن دستگاه، ضریب $\alpha_۰$ فقط به f و $x_۰$، و ضریب $\alpha_۱$ به f، $x_۰$، $x_۱$، و بطور کلی ضریب α_k به f و $x_۰$، ...، x_k بستگی دارد. بنابراین ضرایب را با نماد

$$\alpha_k = f[x_۰, x_۱, \dots, x_k], \quad k = ۰, ۱, \dots, n$$

که به *تفاضلات تقسیم‌شده نیوتن* معروف‌اند، نمایش می‌دهیم و قرار می‌دهیم

$$\begin{aligned} p_n(x) &= f[x_۰]\pi_۰(x) + f[x_۰, x_۱]\pi_۱(x) + \cdots + f[x_۰, \dots, x_n]\pi_n(x) \\ &= f[x_۰] + f[x_۰, x_۱](x - x_۰) + \cdots + f[x_۰, \dots, x_n](x - x_۰)\cdots(x - x_{n-۱}), \end{aligned} \tag{۱۰.۳}$$

که به آن فرمول درونیابی نیوتن می‌گویند. اگر $p_{n-۱}(x)$ چندجمله‌ای درونیاب روی نقاط $x_۰$، ...، $x_{n-۱}$ باشد آنگاه

$$p_n(x) = p_{n-۱}(x) + f[x_۰, \dots, x_n](x - x_۰)\cdots(x - x_{n-۱}),$$

که خاصیت بازگشتی فرمول درونیاب نیوتن را نشان می‌دهد. استفاده از عنوان تفاضلات تقسیم شده بعداً مشخص خواهد شد. از آنجا که چندجمله‌ای درونیاب یکتاست، ضریب جمله‌ی x^n در هر دوی فرمول‌های لاگرانژ و نیوتن بایستی یکسان باشد، پس داریم

$$f[x_۰, x_۱, \dots, x_n] = \sum_{k=۰}^{n} f(x_k) \prod_{j=۰, j\neq k}^{n} \frac{۱}{x_k - x_j}. \tag{۱۱.۳}$$

سمت راست معادله بالا یک فرمول متقارن است، بنابراین تفاضل تقسیم شده $f[x_۰, \dots, x_n]$ نسبت به آرگومان‌های خود متقارن است یعنی با تغییر ترتیب x_j ها مقدار آن ثابت می‌ماند. به عنوان مثال

$$f[x_۰, x_۱, x_۲] = f[x_۲, x_۰, x_۱] = f[x_۲, x_۱, x_۰].$$

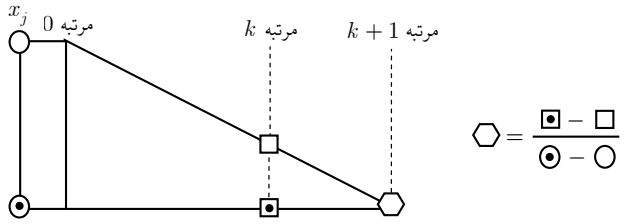
اغلب به جای حل دستگاه مثلثی $U\boldsymbol{\alpha} = \boldsymbol{f}$ از جدول تفاضلات تقسیم‌شده نیوتن استفاده می‌شود که مبتنی بر رابطه‌ی بازگشتی زیر برای $k > i$ است

$$f[x_i, \dots, x_k] = \frac{f[x_{i+۱}, \dots, x_k] - f[x_i, \dots, x_{k-۱}]}{x_k - x_i}, \quad f[x_i] = f(x_i). \tag{۱۲.۳}$$

این رابطه را می‌توان با تغییر اندیس آرگومان‌ها و با اعمال شرایط درونیابی در فرمول درونیابی نیوتن و استفاده از استقرا و یا به کمک فرمول متقارن (۱۱.۳) به سادگی اثبات کرد. چون در این رابطه بازگشتی یک تفاضل و یک تقسیم وجود دارد، عنوان تفاضلات تقسیم شده برای آن‌ها انتخاب شده است.

x_j	مرتبه صفر	مرتبه یک	مرتبه دو	مرتبه سه
x_0	$f[x_0]$			
x_1	$f[x_1]$	$f[x_0, x_1]$		
x_2	$f[x_2]$	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$	
x_3	$f[x_3]$	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	$f[x_0, x_1, x_2, x_3]$

هر عنصر جدول با تفاضل دو عنصر ستون قبل خود تقسیم بر تفاضل دو عنصر ستون اول طبق الگوی ترسیم شده در شکل ۴.۳ ساخته می‌شود. ضرایب بسط نیوتن عناصر روی قطر بالایی جدول‌اند، با این حال بایستی تمام $\frac{n(n+1)}{2}$ عنصر



شکل ۴.۳: الگوی محاسبه‌ی تفاضلات تقسیم‌شده‌ی نیوتن

جدول محاسبه شوند. برای محاسبه‌ی هر یک سه عملگر (دو جمع و یک تقسیم) نیاز است، بنابراین هزینه محاسبه جدول $\frac{3}{2}n^2 + \frac{3}{2}n$ است. در آخر باید چندجمله‌ای p_n را با فرمول (۱۰.۳) برای یک نقطه ثابت x محاسبه کنیم. برای این کار از الگوریتمی شبیه روش هرنر استفاده می‌کنیم. فرمول درونیابی نیوتن را می‌توان به شکل

$$p_n(x) = \alpha_0 + (x - x_0)\Big(\alpha_1 + (x - x_1)\big(\alpha_2 + (x - x_2)(\alpha_3 + \cdots + (x - x_{n-1})\alpha_n)\big)\Big)$$

نوشت. بنابراین الگوریتم هرنر برای محاسبه $b_0 = p_n(x)$ به صورت زیر خواهد بود

$$b_n := \alpha_n, \quad b_j = b_{j+1}(x - x_j) + \alpha_j, \quad j = n-1 : -1 : 0, \tag{۱۳.۳}$$

که هزینه محاسباتی آن $3n$ است. از این رو هزینه محاسباتی روش نیوتن تقریباً نصف روش لاگرانژ است. اما اگر نقطه درونیابی جدیدی به انتهای جدول اضافه شود، محاسبات قبل همگی معتبر بوده و یک قطر به پایین جدول با هزینه $3n$ اضافه می‌شود. این خصوصیت بازگشتی حسن برجسته روش نیوتن نسبت به روش لاگرانژ است.

از دید نظری هر ترتیبی از نقاط درونیابی x_j منجر به چندجمله‌ای درونیاب یکتایی خواهد شد. اما از منظر عددی، ضریب وضعیتِ محاسبه ضرایب α_j در بسط نیوتن به شدت وابسته به ترتیب نقاط درونیابی است و اغلب اگر نقطه x که ارزیابی چندجمله‌ای درونیاب در آن مد نظر است مشخص باشد، ترتیب نقاط به گونه‌ای که

$$|x - x_0| \leqslant |x - x_1| \leqslant \cdots \leqslant |x - x_n| \tag{۱۴.۳}$$

پیشنهاد می‌شود.

برای خطای درونیابی می‌توان فرمولی برحسب تفاضلات تقسیم شده بدست آورد. فرض کنیم خطا در نقطه‌ی دلخواه $x \neq x_j$، $0 \leqslant j \leqslant n$، مد نظر است. گیریم $q(t)$ چندجمله‌ای درونیاب از درجه $n+1$ تابع $f(t)$ روی نقاط x_0، ...، x_n و x باشد. پس طبق خاصیت بازگشتی فرمول نیوتن داریم

$$q(t) = p_n(t) + f[x_0, \ldots, x_n, x](t - x_0) \cdots (t - x_n).$$

از طرفی طبق فرض درونیابی $q(x) = f(x)$ و از این‌رو

$$f(x) - p_n(x) = f[x_0, \ldots, x_n, x](x - x_0) \cdots (x - x_n), \qquad (۱۵.۳)$$

که در مقایسه با خطای لاگرانژ (۳.۳) کاربردی‌تر است چراکه نیاز به هموار بودن f ندارد و تنها لازم است $f(x)$ در نقاط x_k تعریف شده باشد. از طرفی این فرمول خطا شامل پارامتر مجهول نیز نیست.

مثال ۳.۳. درونیاب مبتنی بر نقاط $(-1, 1)$، $(3, 2)$ و $(-1, 1)$ را به کمک روش نیوتن بدست می‌آوریم. جدول تفاضلات تقسیم شده به صورت زیر است

x_j	مرتبه صفر	مرتبه یک	مرتبه دو
-1	1		
2	3	$\frac{3-1}{2+1} = \frac{2}{3}$	
1	-1	$\frac{-1-3}{1-2} = 4$	$\frac{4-\frac{2}{3}}{1+1} = \frac{5}{3}$

و با توجه به عناصر روی قطر جدول چندجمله‌ای درونیاب به کمک فرمول (۱۰.۳) به صورت زیر بدست می‌آید

$$p_2(x) = 1 + \frac{2}{3}(x+1) + \frac{5}{3}(x+1)(x-2).$$

اگر نقطه‌ی جدید $(0, 2)$ به نقاط درونیابی اضافه شود، محاسبات قبل در جدول کماکان معتبرند و کافی است یک سطر دیگر به انتهای جدول برای نقطه‌ی جدید اضافه شود.

x_j	مرتبه صفر	مرتبه یک	مرتبه دو	مرتبه سه
-1	1			
2	3	$\frac{3-1}{2+1} = \frac{2}{3}$		
1	-1	$\frac{-1-3}{1-2} = 4$	$\frac{4-\frac{2}{3}}{1+1} = \frac{5}{3}$	
0	2	$\frac{2+1}{0-1} = -3$	$\frac{-3-4}{0-2} = \frac{7}{2}$	$\frac{\frac{7}{2}-\frac{5}{3}}{0+1} = \frac{11}{6}$

و چندجمله‌ای درونیاب درجه سه با اضافه کردن یک جمله به p_2 به صورت زیر محاسبه می‌شود

$$p_3(x) = 1 + \frac{2}{3}(x+1) + \frac{5}{3}(x+1)(x-2) + \frac{11}{6}(x+1)(x-2)(x-1).$$

$\diamondsuit$

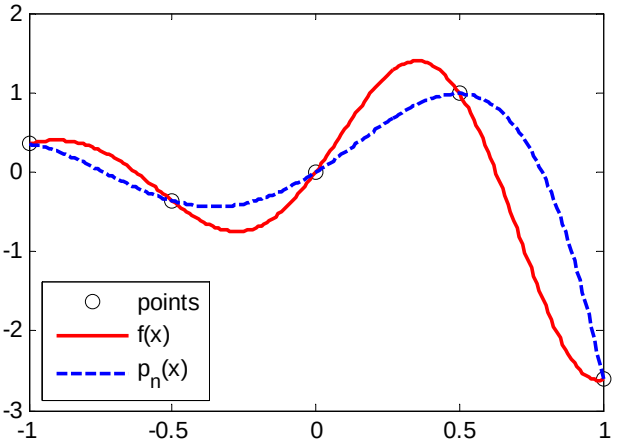
برنامه متلب روش نیوتن برای تعیین چندجمله‌ای درونیاب به صورت زیر نوشته می‌شود.

```
1 function [p, a, D] = NewtonInterp(x, f, s)
2  n = length(x);
3  D(:,1)=f';
4  for j=2:n
5      for i=j:n
6          D(i,j) = (D(i-1,j-1)-D(i,j-1))/(x(i-j+1)-x(i));
7      end
8  end
9  a = diag(D); p = a(n);
10 for j=n-1:-1:1
11     p = p.*(s'-x(j))+a(j);
12 end
```

در برنامه‌ی بالا بردارهای n تایی x (نقاط درونیابی) و f (مقادیر تابع) و بردار با بعد متفاوت s که قرار است درونیاب در آن محاسبه شود به عنوان ورودی دریافت می‌شوند و خروجی‌های برنامه بردار p که مقادیر چندجمله‌ای درونیاب در بردار s است، بردار a متشکل از ضرایب چندجمله‌ای درونیاب، و ماتریس D متشکل از تفاضلات تقسیم شده‌ی نیوتن می‌باشند. در حلقه‌ی آخر درون برنامه مقدار چندجمله‌ای درونیاب به کمک روش هرنر محاسبه شده است. لازم به ذکر است که در این برنامه، درونیاب از درجه‌ی $n-1$ است زیرا تعداد نقاط درونیابی n تاست. به عنوان یک مثال برای درونیابی روی نقاط هم‌فاصله روی بازه‌ی $[0, 1]$ برای تابع $f(x) = e^x \sin 5x$ تابع متلب بالا را به صورت زیر فراخوانی می‌کنیم

```
1 h = 0.5; x = -1:h:1; s = -1:0.01:1;
2 f = exp(x).*sin(5*x);
3 [p, a, D] = NewtonInterp(x, f, s);
4 plot(x,f,'ok',s,exp(s).*sin(5*s),'-r',s,p,'--b')
```

که در آن دستور plot برای رسم نمودار نقاط درونیابی به شکل گوی‌های مشکی، تابع f با خط پر و چندجمله‌ای درونیاب روی شبکه‌ی ریز s با خط چین، نوشته شده است. این نمودار در شکل ۵.۳ رسم شده است.



شکل ۵.۳: نقاط درونیابی، تابع اصلی و درونیاب درجه چهارم آن

۱.۲.۳ اصلاح فرمول درونیابی روی نقاط هم‌فاصله

اگر نقاط درونیابی با فاصله‌ی یکسان روی $[a,b]$ پخش شده باشند، می‌توان فرمولی با هزینه‌ی محاسباتی کمتر برای بدست آوردن درونیاب طراحی کرد. فرض کنیم بازای یک h حقیقی معین داشته باشیم

$$x_k = x_۰ + kh, \quad k = ۰, ۱, \dots, n,$$

نقاط درونیابی در بازه‌ی $[a,b]$ باشند. در این صورت در فرمول (۱۰.۳) می‌توان ضرایب $f[x_۰, \dots, x_k]$ و پایه‌های $\pi_k(x)$ را به طریق بر حسب h دیگری بازنویسی کرد. برای این کار ابتدا عملگرهای Δ^m، که روی مقادیر f_k اثر می‌کنند را به صورت زیر تعریف می‌کنیم:

$$\begin{aligned}\Delta^۰ f_k :=& f_k,\\ \Delta^m f_k :=& \Delta^{m-۱} f_{k+۱} - \Delta^{m-۱} f_k, \quad m \geqslant ۱.\end{aligned}$$

به Δ^m عملگر *تفاضلات پیشرو* مرتبه m می‌گوییم که با توجه به تعریف از روی دو تفاضل پیشرو مرتبه‌ی $m-۱$ بدست می‌آید. علت استفاده از لفظ پیشرو این است که در رابطه‌ی بازگشتی عملگر روی f_k از عملگر روی $f_{k+۱}$ (مقدار بعدی) کم شده است. اگر نقاط درونیابی هم‌فاصله باشند، تفاضلات تقسیم‌شده را می‌توان بر حسب تفاضلات پیشرو نوشت:

$$f[x_i, x_{i+۱}, \dots, x_{i+k}] = \frac{\Delta^k f_i}{h^k k!}. \tag{۱۶.۳}$$

اثبات این ادعا بسیار ساده است. در واقع حکم برای تفاضلات مرتبه صفر یعنی بازای $k = ۰$ بوضوح برقرار است زیرا

$$f[x_i] = f_i = \Delta^۰ f_i.$$

با استقرا روی k فرض کنید حکم برای $k = m$ درست است. بنابراین برای $k = m + ۱$ داریم

$$\begin{aligned} f[x_i, \dots, x_{i+m}, x_{i+m+۱}] &= \frac{f[x_{i+۱}, \dots, x_{i+m+۱}] - f[x_i, \dots, x_{i+m}]}{x_{i+m+۱} - x_i} \\ &= \frac{۱}{(m+۱)h} \left(\frac{\Delta^m f_{i+۱}}{h^m m!} - \frac{\Delta^m f_i}{h^m m!} \right) \\ &= \frac{\Delta^{m+۱} f_i}{(m+۱)! h^{m+۱}}. \end{aligned}$$

برای بازنویسی توابع $\pi_k(x)$ بر حسب h، ابتدا تغییر متغیر زیر را در نظر می‌گیریم

$$t = \frac{x - x_۰}{h}.$$

اگر $x_۰ \leqslant x \leqslant x_n$ آنگاه $۰ \leqslant t \leqslant n$. در واقع این تغییر متغیر بازه‌ی $[x_۰, x_n]$ را به بازه‌ی $[۰, n]$ می‌نگارد. حال داریم

$$\begin{aligned} &(x - x_۰) = th, \\ &(x - x_k) = (\underbrace{x - x_۰}_{} + \underbrace{x_۰ - x_k}_{}) = (th - kh) = (t - k)h, \quad k = ۱, ۲, \dots. \end{aligned}$$

بنابراین می‌توان نوشت

$$\begin{aligned} &\pi_۱(x) = (x - x_۰) = th, \\ &\pi_k(x) = \pi_{k-۱}(x)(x - x_{k-۱}) = t(t-۱)\cdots(t-k+۱)h^k, \quad k = ۱, ۲, \dots, \end{aligned}$$

و چندجمله‌ای درونیاب p_n به صورت زیر بازنویسی می‌شود

$$\begin{aligned} p_n(x) &= f[x_۰]\pi_۰(x) + f[x_۰, x_۱]\pi_۱(x) + \cdots + f[x_۰, \dots, x_n]\pi_n(x) \\ &= \Delta^۰ f_۰ + \frac{\Delta^۱ f_۰}{h} th + \frac{\Delta^۲ f_۰}{h^۲ ۲!} t(t-۱)h^۲ + \cdots + \frac{\Delta^n f_۰}{h^n n!} t(t-۱)\cdots(t-n+۱)h^n \\ &= f_۰ + \frac{t}{۱!}\Delta^۱ f_۰ + \frac{t(t-۱)}{۲!}\Delta^۲ f_۰ + \cdots + \frac{t(t-۱)\cdots(t-n+۱)}{n!}\Delta^n f_۰ \\ &= \sum_{k=۰}^{n} \binom{t}{k} \Delta^k f_۰, \end{aligned} \tag{۱۷.۳}$$

که در آن از نماد ترکیبیاتی $\binom{t}{k} = \frac{t(t-۱)\cdots(t-k+۱)}{k!}$ استفاده کرده‌ایم. برای محاسبه‌ی تفاضلات پیشرو می‌توان از جدول تفاضلات پیشرو استفاده کرد که در زیر حالت $n = ۳$ به تصویر کشیده شده است.

x_j	مرتبه صفر	مرتبه یک	مرتبه دو	مرتبه سه
$x_۰$	$f_۰$			
		$\Delta^۱ f_۰$		
$x_۱$	$f_۱$		$\Delta^۲ f_۰$	
		$\Delta^۱ f_۱$		$\Delta^۳ f_۰$
$x_۲$	$f_۲$		$\Delta^۲ f_۱$	
		$\Delta^۱ f_۲$		
$x_۳$	$f_۳$			

هر عنصر جدول، تفاضل عنصرهای بالا و پایین ستون قبل خود است. بنابراین هزینه تولید این جدول برای درونیابی درجه n برابر $\frac{۱}{۲}n(n+۱)$ است زیرا برای هر عضو جدول فقط یک عمل تفریق نیاز است. در فرمول (۱۷.۳) از عناصر قطر بالایی این جدول استفاده شده است. نوشتن برنامه کامپیوتری این روش به خواننده واگذار می‌شود و در اینجا تنها به ارائه‌ی یک مثال بسنده می‌کنیم.

مثال ۴.۳. می‌خواهیم چندجمله‌ای درونیاب درجه سه روی نقاط $(۰,-۱)$، $(۰/۵, ۲)$، $(۱, ۱)$ و $(۱/۵, -۴)$ را با روش نیوتن بدست آوریم. چون نقاط درونیابی هم‌فاصله هستند، روش تفاضلات پیشرو را به کار می‌گیریم. جدول تفاضلات پیشرو به صورت زیر است

x_j	مرتبه صفر	مرتبه یک	مرتبه دو	مرتبه سه
۰	-۱			
		$۲+۱=۳$		
۰/۵	۲		$-۱-۳=-۴$	
		$۱-۲=-۱$		$-۴+۴=۰$
۱	۱		$-۵+۱=-۴$	
		$-۴-۱=-۵$		
۱/۵	-۴			

با تغییر متغیر $t=(x-x_۰)/h=۲x$ درونیاب به صورت زیر نوشته میشود

$$\begin{aligned} p_۳(x) &= f_۰ + t\Delta f_۰ + \frac{t(t-۱)}{۲!}\Delta^۲ f_۰ + \frac{t(t-۱)(t-۲)}{۳!}\Delta^۳ f_۰ \\ &= -۱ + ۳t - ۲t(t-۱) + ۰ \\ &= -۱ + ۶x - ۴x(۲x-۱) = -۸x^۲ + ۱۰x - ۱. \end{aligned}$$

در این مثال چندجمله‌ای درونیاب از درجه‌ی دو است و یک درجه کمتر از انتظار ما است. در واقع این چهار نقطه روی یک سهمی واقعند و اگر یکی از آن‌ها حذف شود باز هم درونیاب همان است که در بالا محاسبه شد. در جدول هم مشاهده می‌کنیم که تفاضلات پیشرو مرتبه سوم صفر است. پس یادمان باشد "درجه‌ی چندجمله‌ای درونیاب گذرنده از $n+۱$ نقطه‌ی متمایز، **حداکثر** برابر n است." $\diamondsuit$

فرمول مشابهی را می‌توان به کمک *تفاضلات پسرو* ∇^m نیز بدست آورد. تفاضلات پسرو به صورت زیر تعریف می‌شوند

$$\begin{aligned}\nabla^۰ f_k :=& f_k,\\ \nabla^m f_k :=& \nabla^{m-۱} f_k - \nabla^{m-۱} f_{k-۱}, \quad m \geqslant ۱.\end{aligned}$$

برای بازنویسی درونیاب بر حسب تفاضلات پسرو، نقطه‌ی x_n را محور قرار داده و تغییر متغیر

$$t = \frac{x - x_n}{h}$$

که بازه‌ی $[x_۰, x_n]$ را به بازه‌ی $[-n, ۰]$ می‌نگارد، را در نظر می‌گیریم. در این صورت مشابه آنچه در مورد تفاضلات پیشرو گفته شد می‌توان نشان داد

$$\begin{aligned} p_n(x) &= f_n + \frac{t}{۱!}\nabla^۱ f_n + \frac{t(t+۱)}{۲!}\nabla^۲ f_n + \cdots + \frac{t(t+۱)\cdots(t+n-۱)}{n!}\nabla^n f_n \\ &= \sum_{k=۰}^{n} \binom{t+k-۱}{k} \nabla^k f_n. \end{aligned} \tag{۱۸.۳}$$

جدول تفاضلات پسرو را می‌توان همانند جدول تفاضلات پیشرو تشکیل داد. درایه‌های این دو جدول یکسانند و می‌توان نشان داد $\Delta^k f_{n-k} = \nabla^k f_n$ که هر دو در یک مکان جدول قرار دارند. در تفاضلات پیشرو درایه‌های قطر بالا و در تفاضلات پسرو درایه‌های قطر پایین استفاده می‌شوند.

هر دو روش پیشرو و پسرو چندجمله‌ای درونیاب یکسان ارائه می‌دهند، زیرا هر دو بازنویسی جدیدی از فرمول تفاضلات تقسیم شده‌ی نیوتن هستند. اما با توجه به آنچه در (۱۴.۳) گفته شد، از دید پایداری عددی، اگر نقطه‌ی x که قرار است درونیاب در آن محاسبه شود به ابتدای جدول (یعنی به $x_۰$) نزدیک باشد، بهتر است از روش تفاضلات پیشرو استفاده کرد و اگر x به انتهای جدول (یعنی به x_n) نزدیک باشد روش تفاضلات پسرو پیشنهاد می‌شود.

لازم به ذکر است که روش‌های دیگری نیز وجود دارند که از دیگر درایه‌های جدول تفاضلاتی نیوتن استفاده می‌کنند. مثلاً فرمول استرلینگ از درایه‌های میانی جدول استفاده می‌کند و برای وقتی که x در میانه‌های بازه‌ی $[x_۰, x_n]$ واقع است، پایدارتر است. در اینجا بیش از این به آن‌ها نمی‌پردازیم و در عوض روش‌های درونیابی دیگری را در بخش‌های بعد معرفی می‌کنیم.

۳.۳ روش نویل-ایتکن*

این روش یک روش تکراری برای محاسبه چندجمله‌ای درونیاب است که بعد از ای. اچ. نویل (۱۸۸۹-۱۹۶۱) و اِی. سی. ایتکن (۱۸۹۵-۱۹۶۷) به این نام مشهور شده است. برای تشریح این روش ابتدا یک مثال ساده ارائه می‌دهیم. فرض کنید $p_1(x)$ چندجمله‌ای درونیاب خطی تابع f روی دو نقطه‌ی $\{x_0, x_1\}$ و $q_1(x)$ چندجمله‌ای درونیاب خطی تابع f روی دو نقطه‌ی $\{x_1, x_2\}$ باشد که $x_2 \neq x_0$. در واقع x_1 نقطه‌ی مشترک برای هر دو درونیاب است. در این صورت می‌توان درونیاب درجه دو $p_2(x)$ روی نقاط $\{x_0, x_1, x_2\}$ را بر حسب $p_1(x)$ و $q_1(x)$ نوشت. در واقع داریم

$$p_2(x) = \frac{(x - x_0)q_1(x) - (x - x_2)p_1(x)}{x_2 - x_0}.$$

علت این امر واضح است، زیرا اولاً با توجه به اینکه $p_1, q_1 \in \mathbb{P}_1$ روشن است که $p_2 \in \mathbb{P}_2$، و از طرفی داریم

$$\begin{aligned}
p_2(x_0) &= \frac{0 \times q_1(x_0) + (x_2 - x_0)p_1(x_0)}{x_2 - x_0} = p_1(x_0) = f(x_0),\\
p_2(x_1) &= \frac{(x_1 - x_0)q_1(x_1) + (x_2 - x_1)p_1(x_1)}{x_2 - x_0} = \frac{(x_1 - x_0)f(x_1) + (x_2 - x_1)f(x_1)}{x_2 - x_0} = f(x_1),\\
p_2(x_2) &= \frac{(x_2 - x_0)q_1(x_2) + 0 \times p_1(x_2)}{x_2 - x_0} = q_1(x_2) = f(x_2),
\end{aligned}$$

که نشان می‌دهد p_2 درونیاب درجه دوم تابع f روی نقاط $\{x_0, x_1, x_2\}$ است. در حقیقت چندجمله‌ای درونیاب درجه دو را به صورت بازگشتی برحسب درونیاب‌های درجه اول نوشتیم. در قضیه زیر حالت کلی بررسی شده است که پایه‌ی روش نویل-ایتکن است.

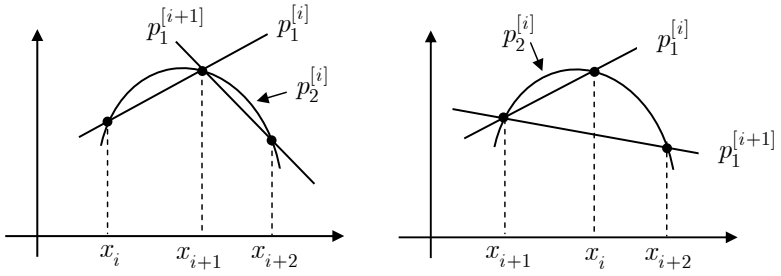
قضیه ۴.۳. *گیریم $p_0^{[i]} := f(x_i)$ برای $0 \leqslant i \leqslant n$. آنگاه برای هر $0 \leqslant k \leqslant n - 1$ داریم*

$$p_{k+1}^{[i]}(x) = \frac{(x - x_i)p_k^{[i+1]}(x) - (x - x_{i+k+1})p_k^{[i]}(x)}{x_{i+k+1} - x_i}, \quad 0 \leqslant i \leqslant n - k - 1, \tag{۱۹.۳}$$

که در آن $p_k^{[i]}(x)$ چندجمله‌ای درونیاب درجه k تابع f روی نقاط $\{x_i, x_{i+1}, \ldots, x_{i+k}\}$ است. و بخصوص $p_n^{[0]}(x)$ چندجمله‌ای درونیاب درجه n روی نقاط $\{x_0, x_1, \ldots, x_n\}$ است.

برهان. با توجه به اینکه $p_k^{[i]}, p_k^{[i+1]} \in \mathbb{P}_k$، طبق فرمول واضح است که $p_{k+1}^{[i]} \in \mathbb{P}_{k+1}$. کافی است نشان دهیم بازای $i \leqslant j \leqslant i + k + 1$ داریم $p_{k+1}^{[i]}(x_j) = f(x_j)$، که به سادگی و با استفاده از خصوصیات درونیابی $p_k^{[i]}$ و $p_k^{[i+1]}$ برقرار است. □

در واقع، درونیاب درجه $k + 1$ به صورت بازگشتی از روی دو درونیاب درجه k بدست می‌آید. در این سه درونیاب، k نقطه مشترکند. در دو نمودار شکل ۶.۳، درونیاب درجه دو مبتنی بر نقاط $\{x_i, x_{i+1}, x_{i+2}\}$ که به کمک درونیاب‌های درجه یک مبتنی بر $\{x_i, x_{i+1}\}$ و $\{x_{i+1}, x_{i+2}\}$ بدست آمده است، رسم شده است.



شکل ۶.۳: شمایی از درونیابی تکراری به روش نویل-ایتکن

با تغییر متغیر $t_i := x - x_i$ می‌توان شکل دیگری از فرمول (۱۹.۳) بدست آورد که در الگوریتم این روش آمده است. در محاسبات می‌توان از یک جدول مثلثی که برای $n = ۳$ در زیر طراحی شده است استفاده کرد.

$x - x_۰$	$p_۰^{[۰]}(x)$			
		$p_۱^{[۰]}(x)$		
$x - x_۱$	$p_۰^{[۱]}(x)$		$p_۲^{[۰]}(x)$	
		$p_۱^{[۱]}(x)$		$p_۳^{[۰]}(x)$
$x - x_۲$	$p_۰^{[۲]}(x)$		$p_۲^{[۱]}(x)$	
		$p_۱^{[۲]}(x)$		
$x - x_۳$	$p_۰^{[۳]}(x)$			

هر عنصر جدول از روی دو عنصر بالا و پایین ستون قبل طبق فرمول (۱۹.۳) بدست می‌آید. در این روش برای هر x بایستی یک جدول مجزا تولید کرد. بنابراین اگر لازم است درونیاب در تعداد نقاط زیاد x بدست آید بهتر است از روش نیوتن استفاده کرد زیرا جدول آن مستقل از x است و کافی است یک بار تهیه شود. برنامه‌ی روش نویل-ایتکن در زیر آمده است.

```
1  function p = NevilleAitken(x,f,s)
2  n = length(x); P = zeros(n,n);
3  P(:,1) = f'; t = x-s;
4  for k=1:n
5      for i=1:n-k
6          P(i,k+1) = (t(i)*P(i+1,k)-t(i+k)*P(i,k))/(t(i)-t(i+k));
7      end
```

```
8  end
9  p = P(1,n);
10 end
```

در این برنامه آرگومان s یک عدد (اسکالر) است که درونیاب در آن محاسبه می‌شود و بر خلاف روش نیوتن نمی‌توان آن را به صورت یک بردار وارد کرد. برای اینکه مقادیر درونیاب را در یک بردار بدست آوریم باید این تابع را در یک حلقه فراخوانی کرد. به عنوان مثال برای درونیابی تابع $f(x) = e^x \sin ۵x$ روی نقاط هم‌فاصله با $h = ۰/۵$ روی بازه‌ی $[-۱, ۱]$ می‌نویسیم

```
1 h = 0.5; x = -1:h:1; s = -1:0.01:1;
2 f = exp(x).*sin(5*x);
3 for k=1:length(s)
4     p(k) = NevilleAitken(x, f, s(k));
5 end
6 plot(x,f,'ok',s,exp(s).*sin(5*s),'-r',s,p,'--b')
```

هزینه محاسباتی روش نویل-ایتکن برای محاسبه p_n در نقطه‌ی ثابت x با الگوریتم بالا تقریباً برابر $\frac{۹}{۲}n^۲$ است که اثبات آن به عنوان تمرین در پرسش ۱۷ از شما خواسته شده است.

۴.۳ فرم گرانیگاهی درونیابی لاگرانژ*

یک بازنویسی ساده از فرمول لاگرانژ منجر به فرمول دیگری برای درونیابی خواهد شد که از جنبه‌های مختلف برتری قابل توجهی بر فرمول قبلی دارد. این فرمول به *درونیابی گرانیگاهی* مشهور است که در این بخش قدری در مورد آن صحبت خواهیم کرد.

با توجه به فرمول درونیابی لاگرانژ، اگر قرار دهیم

$$\pi_{n+۱}(x) = \prod_{j=۰}^{n}(x - x_j), \tag{۲۰.۳}$$

و

$$\beta_j^{-1} = \prod_{\substack{k=0 \\ k\neq j}}^{n}(x_j - x_k), \tag{۲۱.۳}$$

آنگاه چندجمله‌ایهای لاگرانژ بفرم زیر بازنویسی خواهند شد

$$\ell_j(x) = \pi_{n+1}(x)\frac{\beta_j}{(x - x_j)}, \quad j = 0, 1, \ldots, n. \tag{۲۲.۳}$$

ضرایب β_j را *ضرایب تکیه‌گاه* می‌نامیم. در این صورت چندجمله‌ای درونیاب را می‌توان به‌صورت زیر نوشت

$$p_n(x) = \sum_{j=0}^{n}\ell_j(x)f_j = \pi_{n+1}(x)\sum_{j=0}^{n}\frac{\beta_j}{(x - x_j)}f_j, \tag{۲۳.۳}$$

که فرمول *لاگرانژ اصلاح شده* نام دارد. با توجه به اینکه چندجمله‌ای درونیاب از هر درجه برای تابع ثابت $f(x) \equiv 1$ دقیق است، با جایگذاری در (۲۳.۳) داریم

$$\pi_{n+1}(x) = \frac{1}{\sum_{j=0}^{n}\frac{\beta_j}{(x-x_j)}}, \quad x \neq x_j, \quad j = 0, 1, \ldots, n. \tag{۲۴.۳}$$

بنابراین از (۲۳.۳) و (۲۴.۳) داریم

$$p_n(x) = \frac{\sum_{j=0}^{n}\frac{\beta_j}{(x-x_j)}f_j}{\sum_{j=0}^{n}\frac{\beta_j}{(x-x_j)}}, \quad x \neq x_j, \quad j = 0, 1, \ldots, n. \tag{۲۵.۳}$$

فرمول (۲۵.۳) فرمول *گرانیگاهی* نام دارد که دارای تقارنی خاص است از این جهت که هم در صورت و هم در مخرج عوامل $\frac{\beta_j}{(x-x_j)}$ وجود دارند.

محاسبه‌ی هر عامل β_j به $2n$ عمل ریاضی نیاز دارند و چون تعداد آن‌ها $n+1$ تا است، پس محاسبه‌ی کل این عوامل $2n^2 + 2n$ هزینه در بر دارد. بعد از این محاسبات، محاسبه‌ی $p_n(x)$ از رابطه‌ی (۲۳.۳) به $\mathcal{O}(n)$ عمل محاسباتی دیگر نیاز دارد. بنابراین هزینه محاسباتی روش لاگرانژ اصلاح شده تقریباً $2n^2$ است. از طرفی محاسبه‌ی $p_n(x)$ رابطه‌ی گرانیگاهی (۲۵.۳) نیز به $\mathcal{O}(n)$ عملگر نیازمند است، از این‌رو هزینه‌ی محاسباتی روش گرانیگاهی نیز تقریباً $2n^2$ است. اما اگر نقطه‌ی جدید (x_{n+1}, f_{n+1}) اضافه شود، نیازی نیست که β_j ها از نو محاسبه شوند بلکه بصورت زیر عمل می‌کنیم: فرض کنیم β_j های قدیم که مبتنی بر نقاط $\{x_0, x_1, \ldots, x_n\}$ بودند را با $\beta_j^{(n)}$ برای $j = 0, \ldots, n$ نمایش دهیم، همچنین β_j های جدید مبتنی بر نقاط $\{x_0, x_1, \ldots, x_n, x_{n+1}\}$ را با $\beta_j^{(n+1)}$ برای $j = 0, \ldots, n+1$ نشان دهیم. داریم

$$\beta_j^{(n+1)} = \frac{\beta_j^{(n)}}{(x_j - x_{n+1})}, \quad j = 0, \ldots, n, \tag{۲۶.۳}$$

و آخرین عامل نیز مستقیماً بصورت

$$\beta_{n+1}^{(n+1)} = \frac{1}{\prod_{k=0}^{n}(x_{n+1} - x_k)}, \tag{۲۷.۳}$$

محاسبه می‌شود. با توجه به اینکه عوامل $\beta_j^{(n)}$ را برای $j = ۰, \dots, n$ داریم، واضح است که محاسبه‌ی $\beta_j^{(n+۱)}$ برای $j = ۰, \dots, n+۱$ به $۴n + ۴$ عمل نیاز دارد.

همچنین ضرایب تکیه‌گاه به تابعی که درونیابی می‌شود وابسته نیستند، از این رو برای درونیابی یک تابع جدید کافی است فقط فرمول (۲۵.۳) برای f_j های جدید با هزینه $\mathcal{O}(n)$ استفاده شود. می‌توان این نکته را مزیت دیگر این روش نسبت به روش نیوتن دانست که برای هر تابع نیازمند تولید یک جدول مجزا با مرتبه محاسباتی $\mathcal{O}(n^۲)$ است.

در فرمول گرانیگاهی به نظر می‌رسد در نزدیکی نقاط گره‌ای که مخرج کسر $\frac{\beta_j}{(x-x_j)}$ به صفر می‌گراید (وقتی x به یکی از x_j ها نزدیک باشد و خطای حذف رخ دهد) مشکل ناپایداری پیش آید. اما چون این عامل هم در صورت و هم در مخرج وجود دارد خطای تولید شده خنثی می‌شود. بعلاوه در حالتی که $fl(x - x_j) = ۰$ قرار می‌دهیم $fl(x - x_j) = u$ که u واحد گردکردن ماشین است. پس از آن عامل u در صورت و مخرج حذف خواهد شد. همچنین هرگاه با افزایش n ضرایب تکیه‌گاه رشد زیاد داشته باشند، گاهی می‌توان در فرمول گرانیگاهی با مقیاس کردن (حذف عوامل مشترک در صورت و مخرج) از خطای سرریز جلوگیری کرد.

در برخی حالتهای خاص ضرایب تکیه‌گاه را می‌توان به طور صریح بدست آورد و در این موارد هزینه محاسباتی روش $\mathcal{O}(n)$ خواهد بود. از جمله می‌توان به حالتی که توزیع نقاط درونیابی در بازه مورد نظر یکنواخت (هم‌فاصله) باشد اشاره کرد. در این حالت اگر فرض کنیم $x_j - x_{j-۱} = h$، به کمک فرمول‌های بازگشتی (۲۶.۳) و (۲۷.۳) و با استقرا روی n، می‌توان نشان داد (پرسش ۱۸ را ببینید)

$$\beta_j = \frac{(-۱)^{n-j}}{h^n j!(n-j)!} = \frac{(-۱)^n}{h^n n!}(-۱)^j \binom{n}{j}, \quad j = ۰, ۱, \dots, n.$$

باتوجه به اینکه β_j هم در صورت و هم در مخرج (۲۵.۳) ظاهر شده است، عامل $\frac{(-۱)^n}{h^n n!}$ که به اندیس j وابسته نیست را از سیگما بیرون آورده و در صورت و مخرج ساده می‌کنیم. بنابراین تعریف می‌کنیم

$$\beta_j^* = (-۱)^j \binom{n}{j}, \quad j = ۰, ۱, \dots, n. \tag{۲۸.۳}$$

و خواهیم داشت

$$p_n(x) = \frac{\sum_{j=۰}^n \frac{\beta_j^*}{(x-x_j)} f_j}{\sum_{j=۰}^n \frac{\beta_j^*}{(x-x_j)}}, \quad x \neq x_j, \quad j = ۰, ۱, \dots, n. \tag{۲۹.۳}$$

با توجه به فرمول

$$\binom{n}{j} = \frac{n-j+۱}{j}\binom{n}{j-۱},$$

تکیه‌گاه‌های β_j^* با رابطه‌ی بازگشتی زیر بسادگی محاسبه می‌شوند

$$\beta_۰^* = ۱, \quad \beta_j^* = -\beta_{j-۱}^* \frac{n-j+۱}{j}, \quad j = ۱, ۲, \dots, n. \tag{۳۰.۳}$$

واضح است که محاسبه‌ی تمامی β_j^* ها با رابطه‌ی بازگشتی بالا به $4n$ عملگر محاسباتی نیاز دارد. یک برنامه برای روش گرانیگاهی روی نقاط هم‌فاصله به صورت زیر است.

```
1  function p = BaryEquidis (f,a,b,n,x)
2  h = (b-a)/n;   x0 = a;
3  bstr = 1;  % the first support
4  s = 0; t = 0;   % initial values of numerator and denominator
5  for j = 1:n+1
6     xx0 = x-x0;
7     ind = find (xx0 == 0);
8     xx0(ind) = eps;
9     t = t + bstr./xx0;
10    s = s + bstr./xx0*f(x0);
11    bstr = -(n-j+1)/j*bstr;
12    x0 = x0 + h;
13 end
14 p = s./t;
```

در این برنامه f تابع تحت درونیابی است که می‌توان آن را با دستور @(x) تولید کرد. مقادیر چندجمله‌ای درونیاب در بردار x محاسبه می‌شوند. ضرایب تکیه‌گاه در متغیر bstr به طور بازگشتی ذخیره می‌شوند. دستور

```
ind = find (xx0 == 0)
```

اندیس j هایی را می‌یابد که $x - x_j = 0$ و بعد از آن، دستور xx0(ind) = eps مقادیر صفر را با اپسیلون ماشین جایگزین می‌کند. صورت کسر (۲۹.۳) در s و مخرج آن در t ذخیره می‌شوند. اجرای این برنامه برای یافتن درونیاب درجه چهار تابع $f(x) = e^x \sin 5x$ روی $[-1, 1]$ به صورت زیر است

```
1 f = @(x) exp(x).*sin(5*x);
2 x = -1:0.01:1;
3 p = BaryEquidis (f,-1,1,4,x);
4 plot(x,f(x),'-b', x,p,'--k');
```

همچنین اگر نقاط درونیابی ریشه‌های چندجمله‌ای چبیشف نوع اول و نوع دوم در بازه‌ی $[-۱, ۱]$ باشند، باز هم می‌توان برای ضرایب تکیه‌گاه فرمول صریح بدست آورد. خوانندگان علاقه‌مند را به [۶] ارجاع می‌دهیم.

۵.۳ همگرایی و پایداری

در بخش‌های قبل کران‌های خطایی برای درونیابی چندجمله‌ای بدست آوردیم. در یک مورد ثابت کردیم اگر $f \in C^{n+۱}[a,b]$ آنگاه وجود دارد $\xi = \xi(x) \in [a,b]$ به طوریکه

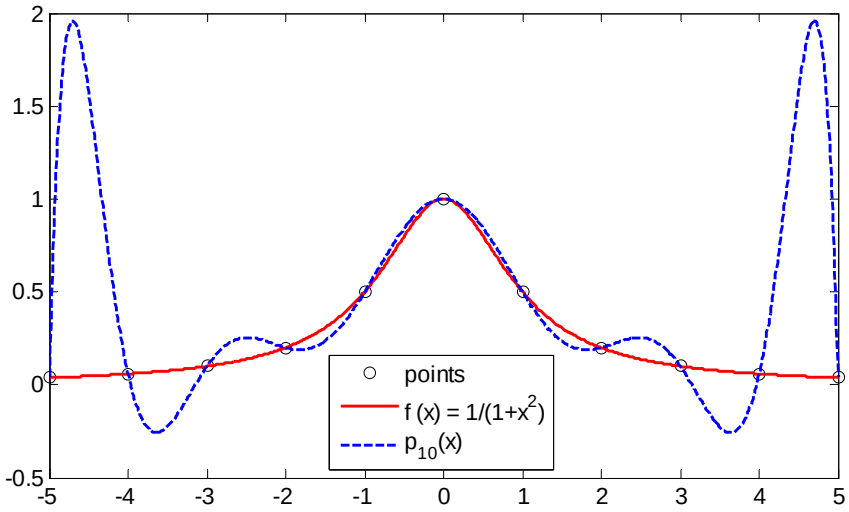
$$R_n(f;x) := f(x) - p_n(x) = \frac{f^{(n+۱)}(\xi(x))}{(n+۱)!}\pi_{n+۱}(x), \quad \pi_{n+۱}(x) = \prod_{k=۰}^{n}(x - x_k). \tag{۳۱.۳}$$

سؤالی که تا به اینجای فصل هنوز به آن پاسخ نداده‌ایم این است که آیا با افزایش تعداد نقاط درونیابی، چندجمله‌ای $p_n(x)$ به صورت یکنواخت به $f(x)$ همگرا می‌شود؟ بهتر است قبل از بحث نظری با یک مثال به این پرسش جواب منفی بدهیم.

مثال ۵.۳ (مثال نقض رونگه). در سال ۱۹۰۱ رونگه نشان داد [۷] دنباله‌ی چندجمله‌ایهای درونیاب تابع

$$f(x) = \frac{۱}{۱ + x^۲}, \quad x \in [-۵, ۵],$$

روی نقاط هم‌فاصله برای $|x| > c \approx ۳/۶۳$ واگرا است. در شکل ۷.۳ نمودار تابع f و چندجمله‌ای درونیاب درجه ۱۰ آن روی نقاط هم‌فاصله ترسیم و مقایسه شده‌اند. این شکل با فراخوانی تابع درونیابی نیوتن به کمک دستورات زیر بدست



شکل ۷.۳: چندجمله‌ای درونیاب درجه ۱۰ تابع $\frac{۱}{۱+x^۲}$ روی بازه‌ی $[-۵, ۵]$ با نقاط هم‌فاصله

آمده است

```
1 x = -5:1:5; f = 1./(1+x.^2); s = -5:0.01:5;
2 [p, a, D] = NewtonInterp(x, f, s);
3 plot(x,f,'ok',s,1./(1+s.^2),'-b', s,p,'--k');
```

اگر درجه‌ی درونیاب را بالاتر ببریم، وضع از این هم بدتر می‌شود و اختلاف بین درونیاب و تابع اصلی به بینهایت میل می‌کند. نکته قابل توجه این است که این تابع در $C^\infty[-۵,۵]$ قرار دارد یعنی تا هر مرتبه‌ای مشتق‌پذیر است و مشتقات آن پیوسته‌اند، اما دنباله‌ی درونیاب‌های آن واگراست. همچنین مشاهده می‌شود که خطای درونیابی در نزدیکی دو نقطه‌ی انتهایی بازه بیشتر است. نکته‌ای که باید در نظر بگیریم این است که این نتیجه‌ی غیر قابل قبول به خاطر بدوضعی خودِ مسئله‌ی درونیابی روی نقاط هم‌فاصله است، نه به خاطر روش درونیابی نیوتن که به کمک آن درونیاب را محاسبه کرده‌ایم. ♢

از جمله مثال‌های دیگری که شما می‌توانید به صورت عددی واگرایی آن‌ها را چک کنید توابع $f(x)=|x|$ روی $[-۱,۱]$ و $f(x)=e^{-x^۲}$ روی $[-۵,۵]$ است. تابع دوم شکلی شبیه تابع مثال رونگه دارد و همانند آن بینهایت بار مشتق‌پذیر است. اما درونیابی برای توابعی مانند e^x و $\sin x$ که دارای دنباله‌ی مشتقات کراندار یکنواخت هستند، همگراست. در واقع اگر برای تابع $f\in C^\infty[a,b]$ داشته باشیم

$$|f^{(k)}(x)|\leqslant M<\infty,\quad \forall\, k\in\mathbb{N},\quad \forall\, x\in[a,b]$$

که در آن M به k وابسته نیست، گوییم دنباله‌ی مشتقات f کراندار یکنواخت است. در این صورت با توجه به اینکه بازای هر j داریم $|x-x_j|\leqslant b-a$، از (۳۱.۳) نتیجه می‌گیریم

$$|f(x)-p_n(x)|\leqslant M\frac{(b-a)^{n+۱}}{(n+۱)!},\quad \forall\, x\in[a,b].$$

از آنجا که رشد تابع فاکتوریل از تابع توانی بیشتر است نتیجه می‌گیریم بازای هر $x\in[a,b]$ چندجمله‌ای $p_n(x)$ به $f(x)$ همگرا خواهد شد. باز هم تأکید می‌کنیم که این همگرایی به علت کراندار یکنواخت بودن دنباله‌ی مشتقات f بدست آمد. برای مثال درونیاب‌های توابع $\sin x$ و e^x همگرایند زیرا برای اولی کران یکنواخت $M=۱$ و برای دومی $M=e^b$ وجود دارند. اما تابع مثال رونگه چنین خاصیتی ندارد و تنها می‌توان گفت مشتقات آن روی بازه‌ی $[a,b]$ کراندارند و در واقع بازای هر $k\in\mathbb{N}$ وجود دارد $M_k<\infty$ بطوریکه

$$|f^{(k)}(x)|\leqslant M_k,\quad \forall\, x\in[a,b],$$

که در آن دنباله‌ی $\{M_k\}_{k\in\mathbb{N}}$ به سرعت رشد می‌کند و

$$\max_{x\in[a,b]}|\pi_{n+۱}(x)|\frac{M_{n+۱}}{(n+۱)!}\to\infty.$$

این همان اتفاقی است که در مورد مثال $e^{-x^۲}$ روی $[-۵, ۵]$ نیز رخ می‌دهد. اما گاهی می‌توان با تغییر جایگاه نقاط درونیابی مقدار

$$\|\pi_{n+۱}\|_\infty := \max_{x\in[a,b]} |\pi_{n+۱}(x)|,$$

را تا آنجا که ممکن است کوچک کرد بطوریکه کران درونیابی به صفر میل کند. در دروس پیشرفته تر ثابت می‌شود (برای مثال فصل سوم [۶] را ببینید)، $\|\pi_{n+۱}\|_\infty$ کمترین مقدار خود را وقتی اختیار می‌کند که نقاط درونیابی x_k ریشه‌های چندجمله‌ای چبیشف درجه‌ی $n+۱$ روی $[a, b]$ باشند. این چندجمله‌ایها به صورت زیر تعریف می‌شوند:

تعریف ۱.۳. چندجمله‌ای چبیشف درجه n-ام $T_n(x)$ برای $x \in [-۱, ۱]$ به‌صورت

$$T_n(x) := \cos\left(n \cos^{-۱} x\right), \quad n = ۰, ۱, \ldots, \tag{۳۲.۳}$$

تعریف می‌شود.

اگرچه ظاهر آنها شبیه چندجمله‌ای نیست اما به روشنی $T_۰(x) = ۱$ و $T_۱(x) = x$. می‌توان ثابت کرد چندجمله‌ایهای درجه بالاتر با رابطه‌ی بازگشتی

$$T_{n+۱}(x) = ۲xT_n(x) - T_{n-۱}(x), \quad n = ۱, ۲, \ldots, \quad x \in [-۱, ۱], \tag{۳۳.۳}$$

بدست می‌آیند (پرسش ۲۰ را ببینید). همچنین اگر قرار دهیم $T_{n+۱}(x) = ۰$ طبق تعریف (۳۲.۳) ریشه‌های $T_{n+۱}$ به صورت زیر بدست می‌آیند

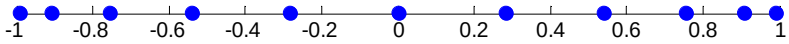
$$x_j = \cos\frac{(۲j+۱)\pi}{۲n+۲}, \quad j = ۰, \ldots, n \tag{۳۴.۳}$$

که نشان می‌دهد همگی حقیقی، متمایز و در بازه‌ی $(-۱, ۱)$ قرار دارند. در زیر چندجمله‌ایهای چبیشف تا درجه‌ی پنج به کمک رابطه‌ی بازگشتی (۳۳.۳) بدست آمده‌اند

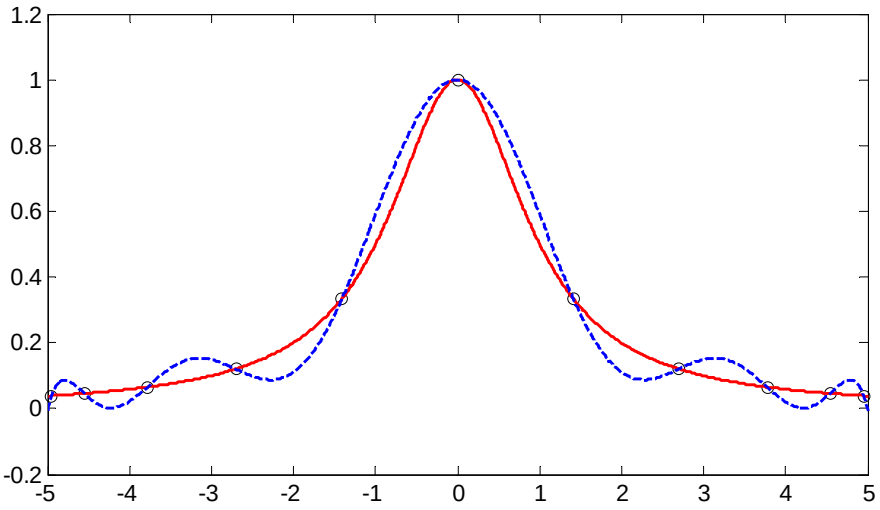
$$\begin{aligned}
T_۰(x) &= ۱,\\
T_۱(x) &= x,\\
T_۲(x) &= ۲x^۲ - ۱,\\
T_۳(x) &= ۴x^۳ - ۳x,\\
T_۴(x) &= ۸x^۴ - ۸x^۲ + ۱,\\
T_۵(x) &= ۱۶x^۵ - ۲۰x^۳ + ۵x.
\end{aligned}$$

همانطور که می‌بینید و می‌توان به کمک رابطه بازگشتی (۳۳.۳) نیز اثبات کرد، ضریب پیشرو در چندجمله‌ای $T_{n+۱}$ (یعنی ضریب $x^{n+۱}$) برابر ۲^n است. لازم به ذکر است که با تغییر متغیر خطی $t = \frac{b-a}{۲}x + \frac{b+a}{۲}$ می‌توان این چندجمله‌ایها

را از بازه‌ی $[-۱, ۱]$ به بازه‌ی $[a, b]$ منتقل کرد. ریشه‌ها نیز با همین تغییر متغیر منتقل خواهند شد. اگر $n+۱$ نقطه‌ی درونیابی ریشه‌های چندجمله‌ای $T_{n+۱}$ باشند، داریم $\pi_{n+۱} = ۲^{-n}T_{n+۱}$ و همانگونه که قبلاً اشاره شد می‌توان ثابت کرد نرم $\|\pi_{n+۱}\|_\infty$ مینیمم می‌شود. در واقع چندجمله‌ایهای چبیشف (بعد از تقسیم بر ضریب پیشرو) کمترین نرم بینهایت را در بین هم‌نوعان خود (تمام اعضای $\mathbb{P}_{n+۱}$ با ضریب پیشرو یک) دارند. شمایی از نقاط چبیشف (ریشه‌های $T_{۱۱}(x)$) در بازه‌ی $[-۱, ۱]$ در شکل زیر رسم شده است.



با انتخاب ریشه‌های چبیشف به عنوان نقاط درونیابی، درونیابِ مثال رونگه و بسیاری دیگر از توابع که درونیابشان روی نقاط هم‌فاصله واگراست، همگرا می‌شوند. برای مثال درونیاب درجه ۱۰ مثال رونگه روی نقاط چبیشف در شکل ۸.۳ رسم شده است. این شکل با فراخوانی تابع درونیابی نیوتن به کمک دستورات زیر بدست آمده است



شکل ۸.۳: چندجمله‌ای درونیاب درجه ۱۰ تابع $\frac{۱}{۱+x^۲}$ روی بازه‌ی $[-۵, ۵]$ با نقاط چبیشف

```
1 n = 10;
2 x = cos((2*(0:n)+1)*pi/(2*n+2));
3 t = 5*x; s = -5:0.01:5;
4 [p, a, D] = NewtonInterp(t, f, s);
5 plot(t,f,'ok',s,1./(1+s.^2),'-b', s,p,'--k');
```

اگر درجه‌ی درونیاب را بالاتر ببریم، نمودار تابع و درونیاب آن بر هم منطبق خواهند شد. بنابراین نقاط چبیشف جزء بهترین نقاط برای درونیابی هستند. اما نکته‌ی بسیار مهمی که باید به آن توجه کرد این است که "توابعی وجود دارند که

درونیاب‌هایشان حتی روی نقاط چبیشف نیز واگراست.“ این یک قضیه مشهور در نظریه تقریب به نام قضیه‌ی فابرِ است. بحث همگرایی درونیاب‌ها نکات دیگری نیز دارد که در اینجا بیش از این به آن نمی‌پردازیم.

حال می‌خواهیم اندکی هم در مورد پایداری مسئله‌ی درونیابی صحبت کنیم که رابطه‌ی تنگاتنگی با همگرایی دارد. برای بررسی پایداری باید به ورودی‌های مسئله‌ی درونیابی اختلال وارد کنید و ببینیم در خروجی (یعنی چندجمله‌ای درونیاب) چه اتفاقی می‌افتد. فرض کنیم فقط در مقادیر $f_k = f(x_k)$ اختلالاتی به اندازه‌ی ϵ_k وارد شده است. این اختلالات می‌تواند ناشی از خطای محاسبه‌ی تابع در ماشین یا در ذات خود این مقادیر باشد. تابع درونیاب برای مسئله‌ی مختل شده را با $p_{n,\epsilon}$ نشان می‌دهیم که طبق روش درونیابی لاگرانژ به صورت زیر بدست می‌آید

$$p_{n,\epsilon}(x) = \sum_{k=0}^{n} (f_k + \epsilon_k)\ell_k(x), \quad x \in [a,b].$$

بنابراین طبق فرمول (۶.۳) می‌توان نوشت

$$p_{n,\epsilon}(x) - p_n(x) = \sum_{k=0}^{n} \epsilon_k \ell_k(x), \quad x \in [a,b].$$

از این رابطه نتیجه می‌گیریم

$$|p_{n,\epsilon}(x) - p_n(x)| \leqslant \max_{0\leqslant k\leqslant n} |\epsilon_k| \sum_{k=0}^{n} |\ell_k(x)| = \lambda_n(x) \max_{0\leqslant k\leqslant n} |\epsilon_k|,$$

که در آن $\lambda_n(x) := \sum_{k=0}^{n} |\ell_k(x)|$ به *تابع لبگ* مشهور است. اگر قرار دهیم $\Lambda_n := \max_{x\in[a,b]} \lambda_n(x)$ آنگاه داریم

$$\|p_{n,\epsilon} - p_n\|_\infty \leqslant \Lambda_n \|\boldsymbol{\epsilon}\|_\infty,$$

که نشان می‌دهد اگر Λ_n عدد بزرگی باشد، اختلال اندک در ورودی باعث اختلال زیاد در درونیاب می‌شود. به Λ_n *ثابت لبگ* می‌گویند که به وضوح مقدار آن به تابع f وابسته نیست و فقط به چیدمان نقاط درونیابی بستگی دارد. می‌توان ثابت کرد اگر نقاط درونیابی هم‌فاصله باشند ثابت لبگ با افزایش n به صورت نمایی رشد می‌کند. بنابراین مسئله‌ی درونیابی روی نقاط هم‌فاصله بدوضع است. اما در عوض می‌توان نشان داد ثابت لبگ نقاط چبیشف دارای رشد کم است و ثابت می‌شود که رشد آن از مرتبه‌ی لگاریتمی است. برای همین درونیابی روی نقاط چبیشف بسیار پایدارتر از درونیابی روی نقاط هم‌فاصله است. همانطور که دیدیم چگالی نقاط چبیشف در نزدیکی دو انتهای بازه بیشتر از مرکز بازه است. می‌توان گفت مجموعه نقاط درونیابی که دارای این خاصیت هستند دارای ثابت لبگ کوچکتری می‌باشند.

۶.۳ درونیابی ارمیت*

هرگاه در تقریب و درونیابی نام ارمیت ظاهر می‌شود، احتمالاً ردپایی از مشتقات تابع وجود خواهد داشت. در درونیابی ارمیت به غیر از مقادیر تابع یعنی $f(x_k)$ ها، مقادیر مشتقات تابع نیز در دست است، یعنی اطلاعات بیشتری نسبت به

درونیابی لاگرانژ در اختیار داریم و همین باعث می‌شود تقریب دقیق‌تری بدست آوریم. در این بخش توجه خود را معطوف به یک حالت خاص می‌کنیم که در آن تنها مقادیر تابع و مشتق مرتبه اولش در تمامی نقاط درونیابی مشخص است، یعنی مقادیر

$$f(x_0), f'(x_0), f(x_1), f'(x_1), \ldots, f(x_n), f'(x_n),$$

را در نقاط متمایز x_0، x_1، ...، x_n داریم و بدنبال یک چندجمله‌ای حداکثر درجه $m = 2n+1$ مانند $p_m(x)$ هستیم به طوریکه شرایط درونیابی زیر برقرار باشند

$$\begin{aligned} p_m(x_k) &= f_k, \quad k = 0, 1, \ldots, n, \quad (i) \\ p'_m(x_k) &= f'_k, \quad k = 0, 1, \ldots, n, \quad (ii) \end{aligned} \tag{۳۵.۳}$$

که در آن $f_k = f(x_k)$ و $f'_k = f'(x_k)$. شاید بپرسیم اگر چنین چندجمله‌ای وجود دارد چرا درجه‌ی آن باید $2n+1$ باشد. در درونیابی لاگرانژ تعداد اطلاعات ما $n+1$ شرط درونیابی بود و درجه‌ی درونیاب یکی کمتر از آن یعنی n بدست آمد. در اینجا نیز چون $2n+2$ معادله‌ی معلوم طبق شرایط درونیابی داریم پس انتظار است که درجه‌ی درونیاب $2n+1$ باشد. برای اینکه نشان دهیم چنین چندجمله‌ای وجود دارد، کافی است یک چندجمله‌ای مانند p_m از درجه‌ی $2n+1$ بسازیم که در شرایط درونیابی (۳۵.۳) صدق کند. این چندجمله‌ای را به شکل زیر در نظر می‌گیریم

$$p_m(x) = \sum_{k=0}^{n} h_{k0}(x) f_k + \sum_{k=0}^{n} h_{k1}(x) f'_k, \tag{۳۶.۳}$$

که $h_{k0}(x)$ و $h_{k1}(x)$ چندجمله‌ایهایی از درجه‌ی $2n+1$ هستند و طوری آن‌ها را می‌سازیم که شرایط درونیابی (۳۵.۳) برقرار باشند. برای اینکه شرایط (i) برقرار باشند کافی است

$$h_{k0}(x_j) = \delta_{jk}, \quad h_{k1}(x_j) = 0, \quad j, k = 0, \ldots, m, \qquad (iii)$$

که $\delta_{jk} = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}$ تابع دلتای کرونکر است. با مشتق‌گیری از چندجمله‌ای درونیاب داریم

$$p'_m(x) = \sum_{k=0}^{m} h'_{k0}(x) f_k + \sum_{k=0}^{m} h'_{k1}(x) f'_k,$$

و برای اینکه شرایط (ii) برقرار باشد کافی است

$$h'_{k1}(x_j) = \delta_{jk}, \quad h'_{k0}(x_j) = 0, \quad j, k = 0, \ldots, m. \qquad (iv)$$

ابتدا $h_{k1}(x)$ را می‌سازیم. طبق (iii) چندجمله‌ای h_{k1} دارای ریشه‌های x_0، x_1، ...، x_n است و طبق (iv) بغیر از x_k بقیه‌ی ریشه‌ها تکراری هستند زیرا مشتق در آنها صفر است. بنابراین می‌توان نوشت

$$h_{k1}(x) = c(x - x_0)^2 \cdots (x - x_{k-1})^2 (x - x_k)(x - x_{k+1})^2 \cdots (x - x_n)^2, \tag{۳۷.۳}$$

که در آن c یک ضریب ثابت است. واضح است که $h_{k1} \in \mathbb{P}_{2n+1}$ پس کافی است ضریب c تعیین شود. برای این امر از شرط $h'_{k1}(x_k) = 1$ که هنوز از آن استفاده نکرده‌ایم، استفاده می‌کنیم. اگر از (۳۹.۳) مشتق گرفته و $h'_{k1}(x_k) = 1$ را به کار بریم، خواهیم داشت

$$\frac{1}{c} = (x_k - x_0)^2 \cdots (x_k - x_{k-1})^2 (x_k - x_{k+1})^2 \cdots (x_k - x_n)^2.$$

در آخر با استفاده از تعریف چندجمله‌ایهای لاگرانژ $\ell_k(x)$ می‌توان نوشت

$$h_{k1}(x) = (x - x_k)\ell_k^2(x). \tag{۳۸.۳}$$

به طور مشابه چندجمله‌ای h_{k0} طبق (iii) دارای ریشه‌های x_0، ...، x_{k-1}، x_{k+1}، ...، x_n است و طبق (iv) همه‌ی آن‌ها تکراری هستند. یعنی داریم

$$h_{k1}(x) = c(x - x_0)^2 \cdots (x - x_{k-1})^2 (x - x_{k+1})^2 \cdots (x - x_n)^2.$$

اما در اینجا c ثابت نیست و برای اینکه h_{k0} از درجه‌ی $2n+1$ باشد، بایستی c یک چندجمله‌ای درجه یک باشند. می‌توان آن را به صورت $ax + b$ در نظر گرفت، اما برای اینکه محاسبات ساده‌تر شود آن را به صورت $a(x - x_k) + b$ در نظر می‌گیریم و داریم

$$h_{k1}(x) = [a(x - x_k) + b](x - x_0)^2 \cdots (x - x_{k-1})^2 (x - x_{k+1})^2 \cdots (x - x_n)^2. \tag{۳۹.۳}$$

کافی است ضرایب a و b تعیین شوند. هنوز از دو شرط $h_{k0}(x_k) = 1$ و $h'_{k0}(x_k) = 0$ استفاده نکرده‌ایم. با اعمال این دو شرط a و b تعیین می‌شوند (به عنوان تمرین انجام دهید) و خواهیم داشت

$$h_{k0}(x) = [1 - 2(x - x_k)\ell'_k(x_k)]\ell_k^2(x). \tag{۴۰.۳}$$

با روندی که در بالا اشاره شد، حداقل یک چندجمله‌ای در $p_m \in \mathbb{P}_m$ ساختیم که در شرایط درونیابی (۳۵.۳) صدق می‌کند. حال نشان می‌دهیم این چندجمله‌ای یکتاست. فرض کنیم چندجمله‌ای دیگری مانند $q_m \in \mathbb{P}_m$ وجود داشته باشد که در شرایط درونیابی (۳۵.۳) صدق کند. اگر قرار دهیم $e_m := p_m - q_m$، واضح است که e_m از درجه‌ی حداکثر $m = 2n + 1$ است و از طرفی چون p_m و q_m در شرایط درونیابی (۳۵.۳) صدق می‌کنند، پس e_m در شرایط زیر صدق می‌کند

$$e_m(x_k) = 0, \quad e'_m(x_k) = 0, \quad k = 0, 1, \ldots, n,$$

که نشان می‌دهد هر x_k ریشه‌ی تکراری e_m است، یعنی e_m حداقل $2n + 2$ ریشه دارد. طبق قضیه‌ی اساسی جبر e_m چندجمله‌ای صفر است، پس $q_m = p_m$ که یکتایی درونیاب را نتیجه می‌دهد. بنابراین قضیه‌ی زیر را داریم:

قضیه ۵.۳. *یک و تنها یک چندجمله‌ای از درجه‌ی $2n + 1$ وجود دارد که در شرایط درونیابی (۳۵.۳) صدق می‌کند.* □

مثال ۶.۳. چندجمله‌ای درونیاب ارمیت مبتنی بر داده‌های جدول زیر را بدست آورید.

	$x_۰ = ۰$	$x_۱ = ۱$
f_k	۱	۲
f'_k	۲	۳

کافی است چندجمله‌ایهای $h_{۰۰}$، $h_{۰۱}$، $h_{۱۰}$، و $h_{۱۱}$، را طبق فرمول‌های (۳۸.۳) و (۴۰.۳) بدست آوریم. به سادگی داریم

$$\ell_۰(x) = ۱ - x, \quad \ell_۱(x) = x,$$

و از آن بدست می‌آوریم

$$h_{۰۰}(x) = (۱ - ۲x)(۱ - x)^۲, \quad h_{۱۰}(x) = (۳ - ۲x)x^۲, \quad h_{۰۱}(x) = x(۱ - x)^۲, \quad h_{۱۱}(x) = (x - ۱)x^۲.$$

و در آخر طبق رابطه‌ی (۳۶.۳) داریم

$$p_۳(x) = (۱ - ۲x)(۱ - x)^۲ + ۲(۳ - ۲x)x^۲ + ۲x(۱ - x)^۲ + ۳(x - ۱)x^۲.$$

$\diamondsuit$

در روش بالا، مجموعه‌ی مستقل خطی

$$\{h_{۰۰}, h_{۱۰}, \ldots, h_{n۰}, h_{۰۱}, \ldots, h_{nn}\} \tag{۴۱.۳}$$

را به عنوان پایه‌ای برای $\mathbb{P}_{۲n+۱}$ در نظر گرفتیم و چون این پایه در شرایط (iii) و (iv) صدق می‌کند، چندجمله‌ای p_m به فرم (۳۶.۳) در شرایط درونیابی (۳۵.۳) صدق می‌کند. همانند درونیابی لاگرانژ، اگرچه نیازی به حل دستگاه برای تعیین ضرایب درونیابی نیست، اما پایه‌ی (۴۱.۳) را باید برای هر مجموعه از نقاط بسازیم. چون این پایه از روی چندجمله‌ایهای لاگرانژ $\ell_k(x)$ ساخته می‌شود، همه‌ی کاستی‌های روش لاگرانژ را به ارث می‌برد. مثلاً امکان بروز خطای حذف ارقام با معنا در آن زیاد است، و اگر نقاط جدیدی برای درونیابی اضافه شوند تمام محاسبات باید از نو تکرار شوند.

اگر بخواهیم روش نیوتن را برای درونیابی ارمیت تعمیم دهیم باید پایه‌ی جدیدی برای فضای $\mathbb{P}_{۲n+۱}$ انتخاب کنیم که منجر به یک فرمول بازگشتی برای تعیین چندجمله‌ای درونیاب شود. با الهام از (۸.۳) و با توجه به اینکه در هر نقطه‌ی x_k هم مقدار خود تابع معلوم است و هم مقدار مشتق آن، پایه‌ی زیر را برای درونیابی ارمیت روی فضای $\mathbb{P}_{۲n+۱}$ انتخاب می‌کنیم

$$\{\pi_۰(x), \pi_۱(x), \ldots, \pi_{۲n+۱}(x)\}$$

که در آن چندجمله‌ایهای $\pi_k(x)$ متفاوت از تعریف (۸.۳) هستند و به صورت زیر تعریف می‌شوند

$$\begin{aligned}
\pi_۰(x) &= ۱,\\
\pi_۱(x) &= (x-x_۰),\\
\pi_۲(x) &= (x-x_۰)^۲,\\
\pi_۳(x) &= (x-x_۰)^۲(x-x_۱),\\
&\vdots\\
\pi_{۲n}(x) &= (x-x_۰)^۲(x-x_۱)^۲\cdots(x-x_{n-۱})^۲,\\
\pi_{۲n+۱}(x) &= (x-x_۰)^۲(x-x_۱)^۲\cdots(x-x_{n-۱})^۲(x-x_n).
\end{aligned}$$

در واقع اگر هر نقطه‌ی درونیابی را دو بار در نظر بگیریم و نقاط t_k را به صورت زیر تعریف کنیم

$$t_۰ = t_۱ := x_۰, \quad t_۲ = t_۳ := x_۱, \quad \dots \quad t_{۲n} = t_{۲n+۱} := x_n,$$

آنگاه توابع پایه‌ی $\pi_k(x)$ به شکل زیر تعریف می‌شوند

$$\pi_۰(x) = ۱, \quad \pi_k(x) = \prod_{j=۰}^{k-۱}(x-t_j), \quad k = ۱, ۲, \dots, ۲n+۱.$$

اگر چندجمله‌ای درونیاب را بر حسب این پایه به صورت

$$p_{۲n+۱}(x) = \alpha_۰\pi_۰(x) + \alpha_۱\pi_۱(x) + \cdots + \alpha_{۲n+۱}\pi_{۲n+۱}(x) \qquad (۴۲.۳)$$

بسط دهیم و شرایط درونیابی (۳۵.۳) را اعمال کنیم و از خاصیت

$$\pi_k(t_j) = ۰, \quad \text{for } k > j, \qquad \pi_k'(t_j) = ۰, \quad \text{for } k > j$$

استفاده کنیم به دستگاه معادلات خطی پایین مثلثی زیر می‌رسیم

$$\begin{bmatrix}
\pi_۰(x_۰) & & & & & & \\
۰ & \pi_۱'(x_۰) & & & & & \\
\pi_۰(x_۱) & \pi_۱(x_۱) & \pi_۲(x_۱) & & & & \\
۰ & \pi_۱'(x_۱) & \pi_۲'(x_۱) & \pi_۳'(x_۱) & & & \\
\vdots & \vdots & \vdots & & \ddots & & \\
\pi_۰(x_n) & \pi_۱(x_n) & \pi_۲(x_n) & \cdots & \pi_{۲n}(x_n) & & \\
۰ & \pi_۱'(x_n) & \pi_۲'(x_n) & \cdots & \pi_{۲n}'(x_n) & \pi_{۲n+۱}'(x_n)
\end{bmatrix}
\begin{bmatrix} \alpha_۰ \\ \alpha_۱ \\ \alpha_۲ \\ \alpha_۳ \\ \vdots \\ \alpha_{۲n} \\ \alpha_{۲n+۱} \end{bmatrix}
=
\begin{bmatrix} f_۰ \\ f_۰' \\ f_۱ \\ f_۱' \\ \vdots \\ f_n \\ f_n' \end{bmatrix}$$

با توجه به اینکه عناصر روی قطر این ماتریس همگی غیر صفرند (نشان دهید!)، این دستگاه معکوس‌پذیر است و دارای جواب یکتاست که اثبات دیگری برای وجود و یکتایی درونیاب ارمیت ارائه می‌دهد. می‌توان این دستگاه را با روش جایگذاری پیشرو با هزینه‌ی محاسباتی تقریباً $\mathcal{O}(n^۲)$ حل کرد و ضرایب α_k را بدست آورد و با جایگذاری در (۴۲.۳) چندجمله‌ای p_m را تعیین کرد. اما همانند قبل بهتر است از جدول تفاضلات تقسیم شده‌ی نیوتن بجای روش جایگذاری پیشرو استفاده کرد. برای ساختن این جدول ابتدا لم زیر را اثبات می‌کنیم.

لم ۶.۳. *اگر f تابعی مشتق‌پذیر در نقطه‌ی x باشد، آنگاه $f[x,x] = f'(x)$.*

برهان. برای اثبات می‌نویسیم

$$f[x,x] = \lim_{h\to ۰} f[x, x+h] = \lim_{h\to ۰} \frac{f(x+h)-f(x)}{h} = f'(x).$$

تساوی آخر به این علت برقرار است که f در x مشتق‌پذیر است. □

اکنون جدول تفاضلات نیوتن را برای مقادیر t_k تشکیل می‌دهیم و محاسبه‌ی $f[t_k, t_{k+۱}]$ برای وقتی که $t_k = t_{k+۱}$ از مقدار مشتق $f'(t_k)$ که جزء داده‌های معلوم مسئله است استفاده می‌کنیم. این فرآیند را با یک مثال توضیح می‌دهیم.

مثال ۷.۳. درونیاب ارمیت مبتنی بر داده‌های جدول مثال ۶.۳ را به کمک روش نیوتن بدست می‌آوریم. جدول تفاضلاتی به صورت زیر نوشته می‌شود

$t_۰ = x_۰ = ۰$	$f_۰ = ۱$			
$t_۱ = x_۰ = ۰$	$f_۰ = ۱$	$f'_۰ = ۲$		
$t_۲ = x_۱ = ۱$	$f_۱ = ۲$	$\frac{۲-۱}{۱-۰} = ۱$	$\frac{۱-۲}{۱-۰} = -۱$	
$t_۳ = x_۱ = ۱$	$f_۱ = ۲$	$f'_۱ = ۳$	$\frac{۳-۱}{۱-۰} = ۲$	$\frac{۲-(-۱)}{۱-۰} = ۳$

یا

۰	۱			
۰	۱	۲		
۱	۲	۱	−۱	
۱	۲	۳	۲	۳

از مقادیر روی قطر جدول داریم $\alpha_۰ = f[x_۰] = ۱$، $\alpha_۱ = f[x_۰, x_۰] = ۲$، $\alpha_۲ = f[x_۰, x_۰, x_۱] = -۱$ و $\alpha_۳ = f[x_۰, x_۰, x_۱, x_۱] = ۳$. بنابراین درونیاب درجه سه ارمیت به صورت زیر نوشته می‌شود

$$\begin{aligned} p_۳(x) &= ۱ + ۲(x-t_۰) - (x-t_۰)(x-t_۱) + ۳(x-t_۰)(x-t_۱)(x-t_۲) \\ &= ۱ + ۲x - x^۲ + ۳x^۲(x-۱). \end{aligned}$$

به طور کلی می‌توان گفت اگر درونیاب ارمیت به صورت (۴۲.۳) بسط داده شود، ضرایب بسط عبارتند از

$$\alpha_k = f[t_۰, t_۱, \ldots, t_k], \quad k = ۰, ۱, \ldots, ۲n+۱.$$

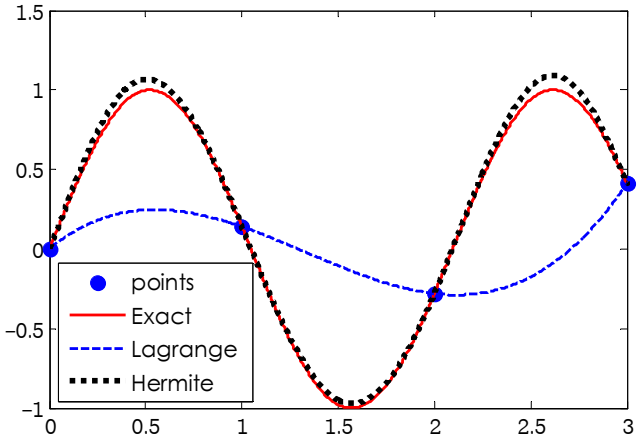
♢

برنامه‌ی متلب درونیابی ارمیت با روش نیوتن در زیر آمده است. در انتهای برنامه درونیاب به کمک روش هرنر همانند بخش (۲.۳) محاسبه شده است.

```
1  function [p, a, D] = HermiteInterp(x, f, f1, s)
2  n = length(x);
3  t(1:2:2*n-1)=x; t(2:2:2*n)=x;
4  D(1:2:2*n-1,1)=f'; D(2:2:2*n,1)=f'; D(2:2:2*n,2)=f1';
5  for j=2:2*n
6      for i=j:2*n
7          if t(i-j+1)-t(i)~=0
8              D(i,j)= (D(i-1,j-1)-D(i,j-1))/(t(i-j+1)-t(i));
9          end
10     end
11 end
12 a = diag(D); p = a(2*n);
13 for j=2*n-1:-1:1
14     p = p.*(s-t(j))+a(j);
15 end
```

نقاط درونیابی در بردار x، مقادیر تابع در بردار f و مقادیر مشتق تابع در بردار f1 به برنامه داده می‌شوند. در آخر درونیاب ارمیت در بردار s محاسبه و توسط بردار p بازگردانده می‌شوند. همچنین جدول تفاضلاتی در ماتریس D و ضرایب درونیابی یعنی قطر ماتریس در بردار a جهت اطلاع کاربر بازگردانده می‌شوند.

مثال ۸.۳. تابع $f(x) = \sin(\mathbf{۳}x)$ را در نظر بگیرید. در شکل ۹.۳ این تابع به همراه درونیاب لاگرانژ آن روی نقاط $X = \{\mathbf{۰}, \mathbf{۱}, \mathbf{۲}, \mathbf{۳}\}$ و همچنین درونیاب ارمیت آن بازای مقادیر خود تابع و مشتق مرتبه اول آن در همین نقاط رسم شده است. از روی شکل مشخص است که درونیاب لاگرانژ دقت مناسبی ندارد اما درونیاب ارمیت همخوانی بسیار مناسب‌تری با تابع f دارد. علت این است که درونیاب لاگرانژ صرفاً در صدد عبور کردن از نقاط درونیابی است اما درونیاب ارمیت به غیر از عبور از نقاط، مراقب یکسان بودن شیب تابع f و درونیابش در نقاط درونیابی نیز هست. البته مقایسه‌ای که در این مثال صورت گرفته است عادلانه نیست زیرا درونیاب لاگرانژ از درجه ۳ است در حالی که درونیاب ارمیت از درجه ۷ است. ♢



شکل ۹.۳: مقایسه درونیابی ارمیت و لاگرانژ

خطای درونیابی ارمیت را می‌توان هم در فرم لاگرانژی و هم در فرم نیوتنی بدست آورد. اثبات قضیه‌ی زیر به عنوان تمرین واگذار می‌شود زیرا مشابه اثبات خطای درونیابی معمولی است. پرسش ۲۳ را ببینید.

قضیه ۷.۳. *فرض کنید $x_0, \ldots, x_n$ نقاط متمایز در بازه‌ی $[a,b]$ باشند و تابع $f \in C^{2n+2}[a,b]$ و چندجمله‌ای $p_{2n+1} \in \mathbb{P}_{2n+1}$ در شرایط درونیابی*

$$f^{(j)}(x_k) = p_{2n+1}^{(j)}(x_k), \quad k = 0, 1, \ldots, n, \quad j = 0, 1,$$

صدق کنند. آنگاه برای هر $x \in [a,b]$ وجود دارد یک $\xi(x) \in [a,b]$ بطوریکه

$$f(x) - p_{2n+1}(x) = \frac{(x-x_0)^2 \cdots (x-x_n)^2}{(2n+2)!} f^{(2n+2)}\big(\xi(x)\big), \tag{۴۳.۳}$$

که فرم لاگرانژی فرمول خطاست. همچنین با شرط ضعیف‌تر $f \in C^1[a,b]$ داریم

$$f(x) - p_{2n+1}(x) = (x-x_0)^2 \cdots (x-x_n)^2 f[x_0, x_0, x_1, x_1, \ldots, x_n, x_n, x],$$

که فرمول خطای فرم نیوتن درونیابی ارمیت است. □

مثال ۹.۳. خطای درونیابی ارمیت روی دو نقطه به فاصله‌ی x_0 و x_1 که $h := x_1 - x_0$ با فرض اینکه $f \in C^4[x_0, x_1]$ به صورت زیر است

$$f(x) - p_3(x) = \frac{(x-x_0)^2(x-x_1)^2}{4!} \max_{t \in [x_0,x_1]} |f^{(4)}(t)|, \quad x \in [x_0, x_1].$$

با توجه به اینکه $\max_{x \in [x_0,x_1]} |(x-x_0)^2(x-x_1)^2| \leqslant (h/2)^4$ (اثبات کنید!) ، می‌توان نوشت

$$\|f - p_3\|_\infty = \frac{h^4}{384} \|f^{(4)}\|_\infty,$$

که نشان می‌دهد اگر f شرایط همواری لازم را داشته باشد خطا از $\mathcal{O}(h^4)$ است. ♢

در انتهای این بخش به این نکته اشاره می‌کنیم که درونیابی ارمیت را می‌توان حالتی که مقادیر مشتقات مراتب بالاتر تابع نیز در دست باشند تعمیم داد، که در اینجا به آن نمی‌پردازیم.

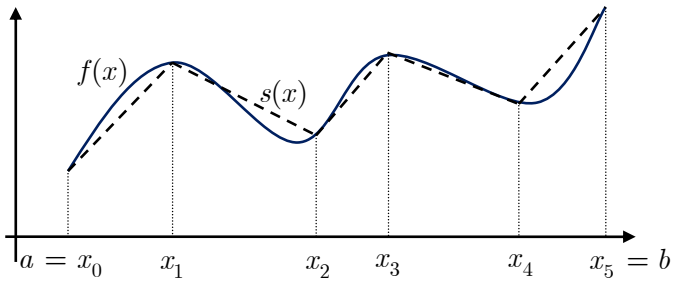
۷.۳ اسپلاین‌ها

در بخش‌های قبل مشاهده کردیم که درونیابی چندجمله‌ای در حالت کلی پایدار نیست و دنباله‌ی چندجمله‌ایهای درونیاب گاهی واگراست. در یک نگاه کلی می‌توان گفت چون با افزایش درجه، نوسانات چندجمله‌ایها افزایش می‌یابد در بسیاری از مواقع درونیاب درجه‌ی بالا اگرچه از تمام نقاط درونیابی می‌گذرد اما تقریب مناسبی برای تابع روی کل بازه نیست. یک ایده برای اجتناب از این معضل این است که بازه‌ی درونیابی را به چندین قسمت تقسیم کنیم و روی هر قسمت یک تقریب چندجمله‌ای با درجه‌ی پایین بکار بریم. البته گاهی لازم است این تقریب طوری ساخته شود که روی کل بازه هموار باشد، یعنی در محل اتصال زیربازه‌های همسایه شرایط پیوستگی تابع و مشتقاتش را فراهم کنیم. به این روش *تقریب موضعی* می‌گوییم. با این نگاه تقریب‌های قبلی همگی از نوع *تقریب سراسری* بودند زیرا روی کل بازه عمل و از تمام اطلاعات مسئله به صورت یکجا استفاده می‌کردند. *اسپلاین‌ها* توابعی برای تقریب موضعی هستند که در این بخش تا حدودی به آن‌ها می‌پردازیم و در آخر برخی از مزایای آن‌ها نسبت به تقریب‌های سراسری را بیان می‌کنیم.

یک حالت ساده این است که بازه‌ی $[a,b]$ را به n زیر بازه به صورت $[x_0,x_1]$، $[x_1,x_2]$، ...، $[x_{n-1},x_n]$ که در آن

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b,$$

تقسیم کنیم و روی هر زیربازه تابع $f(x)$ را با یک چندجمله‌ای خطی طوری تقریب بزنیم که تابع قطعه‌ای حاصل در نقاط x_k با f یکی باشد. به شکل (۱۰.۳) نگاه کنید. در این درونیاب قطعه‌ای خطی که ما آن را اسپلاین خطی می‌نامیم، تابع



شکل ۱۰.۳: درونیاب اسپلاین خطی

اسپلاین در هر زیر بازه یک چندجمله‌ای خطی است و روی کل بازه‌ی $[a,b]$ پیوسته است. به طور کلی تعریف یک اسپلاین درجه‌ی ℓ به صورت زیر است.

تعریف ۲.۳. گیریم ℓ عدد صحیح نامنفی باشد. یک تابع $s : [a,b] \longrightarrow \mathbb{R}$ را یک اسپلاین چندجمله‌ای از درجه‌ی ℓ گوییم اگر

۱. $s \in C^{\ell-1}[a,b]$

۲. $s \in \mathbb{P}_\ell$ برای $x \in [x_k, x_{k+1}]$ که $0 \leqslant k \leqslant n-1$،

که در آن $X = \{a = x_0 < x_1 < \cdots < x_n = b\}$ یک افراز نقطه‌ای از بازه‌ی $[a,b]$ است. در اینجا $C^{-1}[a,b]$ فضای توابع قطعه‌ای پیوسته روی $[a,b]$ است. فضای اسپلاین‌های درجه‌ی ℓ روی X را با $\mathbb{S}_X^\ell$ نمایش می‌دهیم.

با این تعریف تابع s در شکل ۱۰.۳ (یعنی نمودار خط چین) یک اسپلاین درجه یک (خطی) است. به همین ترتیب یک اسپلاین درجه دوم در هر زیربازه یک چندجمله‌ای درجه دو است و روی کل بازه یک تابع C^1 است. واضح است که در نقاط غیر x_k، $k = 1, \ldots, n-1$، تابع اسپلاین از هر مرتبه مشتق‌پذیر است. پس برای اسپلاین درجه دو، کافی است اسپلاین و مشتق آن در نقاط x_k، $k = 1, \ldots, n-1$، دارای حد چپ و حد راست برابر باشند.

دقت کنید که در تعریف اسپلاین، تابع f ی در کار نیست. در ادامه اسپلاین را به گونه‌ای تعیین می‌کنیم که درونیابِ یک تابعِ مفروضِ f باشد. در واقع درونیابِ f را در فضای $\mathbb{S}_X^\ell$ بدست می‌آوریم، همانگونه که قبلاً این کار را در فضای $\mathbb{P}_n$ انجام دادیم. اگر بخواهیم همانند قبل عمل کنیم، باید پایه‌یی برای فضای $\mathbb{S}_X^\ell$ معرفی کنیم و تابع اسپلاین s را بر حسب پایه بسط دهیم و با اعمال شرایط درونیابی، ضرایب بسط را تعیین کنیم. اما معرفی پایه‌ی فضای اسپلاین‌ها قدری از حوصله‌ی این درس خارج است و در درس‌های پیشرفته‌تر مطرح می‌شود. شما می‌توانید برای مثال به [۶] مراجعه کنید. اما در اینجا به طریق دیگری که نیازی به معرفی پایه نباشد اسپلاین درونیاب را تعیین می‌کنیم.

فرض کنید مقادیر f_0، f_1، ...، f_n از یک تابع مفروض f در دست باشند. اسپلاین درونیاب خطی به سادگی با درونیابی خطی تابع f در هر زیربازه همانند شکل (۱۰.۳) تعیین می‌شود. در واقع داریم

$$s(x) = \begin{cases} f(x_0) + (x - x_0) f[x_0, x_1], & x \in [x_0, x_1] \\ f(x_1) + (x - x_1) f[x_1, x_2], & x \in [x_1, x_2] \\ \qquad \vdots & \\ f(x_{n-1}) + (x - x_{n-1}) f[x_{n-1}, x_n], & x \in [x_{n-1}, x_n] \end{cases}$$

که در آن در هر زیر بازه، فرمول درونیاب خطی نیوتن را نوشته‌ایم. اسپلاین درونیاب خطی به همین سادگی تعیین می‌شود زیرا فرضی بر پیوستگی مشتقات نداریم و تنها کافی است s پیوسته باشد که آن هم در ذات ضابطه‌های s لحاظ شده است. یافتن خطای اسپلاین درونیاب خطی نیز سرراست است. فرض کنیم

$$h_k = x_{k+1} - x_k, \quad k = 0, 1, \ldots, n-1.$$

و همچنین فرض کنیم $f \in C^2[a,b]$. در این صورت طبق فرمولِ خطای درونیابی خطی در (۷.۳)، برای هر زیر بازه‌ی $[x_k, x_{k+1}]$ می‌توان نوشت

$$|f(x) - s(x)| \leqslant \frac{h_k^2}{8} \max_{t \in [x_{k-1}, x_k]} |f''(t)|, \quad x \in [x_k, x_{k+1}], \quad k = 0, 1, \ldots, n-1.$$

و اگر قرار دهیم $h = \max_{۰\leqslant k\leqslant n-۱} h_k$ داریم

$$|f(x) - s(x)| \leqslant \frac{h^۲}{۸} \max_{t\in[a,b]} |f''(t)|, \quad x \in [a,b].$$

چون نابرابری بالا بازای هر $x \in [a,b]$، بخصوص برای x ی که سمت چپ را ماکزیمم می‌کند، برقرار است، داریم

$$\|f - s\|_\infty \leqslant \frac{h^۲}{۸}\|f''\|_\infty. \tag{۴۴.۳}$$

کران بالا نشان می‌دهد اگر $f \in C^۲[a,b]$ آنگاه خطای اسپلاین درونیاب خطی $\mathcal{O}(h^۲)$ است. یعنی با کاهش فاصله‌ی بین نقاط، f به s همگرا می‌شود. این اولین حسن اسپلاین‌ها نسبت به درونیاب‌های چندجمله‌ای است که حتی برای توابع C^∞ نیز همگرایی تضمین شده نداشتند.

مثال ۱۰.۳. می‌خواهیم اسپلاین درونیاب خطی برای تابع $f(x) = ۱ - x^۲$ روی نقاط $X = \{-۱, -\frac{۱}{۲}, ۰, \frac{۱}{۲}, ۱\}$ بنویسیم و کران خطای درونیابی را بدست آوریم. با توجه به ضابطه‌ی تابع داریم

$$f_۰ = ۰, \;\; f_۱ = \frac{۳}{۴}, \;\; f_۲ = ۱, \;\; f_۳ = \frac{۳}{۴}, \;\; f_۴ = ۰.$$

همچنین مقادیر تفاضلات تقسیم شده مرتبه اول عبارتند از

$$f[x_۰, x_۱] = \frac{۳}{۲}, \;\; f[x_۱, x_۲] = \frac{۱}{۲}, \;\; f[x_۲, x_۳] = -\frac{۱}{۲}, \;\; f[x_۳, x_۴] = -\frac{۳}{۲}.$$

بنابراین ضابطه‌ی اسپلاین به صورت زیر تعیین می‌شود

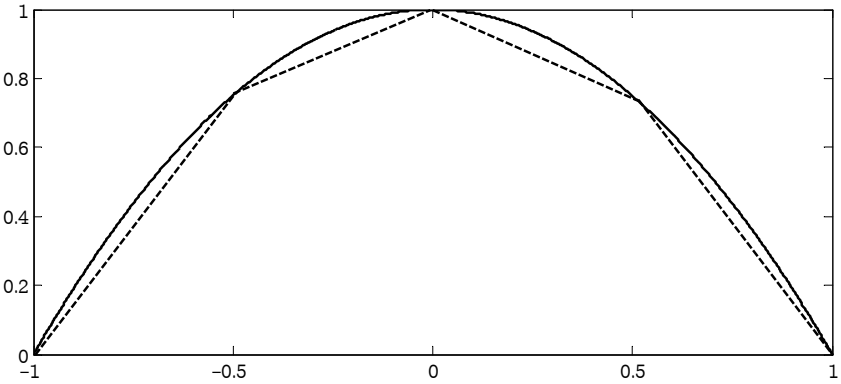
$$s(x) = \begin{cases} \frac{۳}{۲}(x+۱), & x \in [-۱, -\frac{۱}{۲}] \\ \frac{۳}{۴} + \frac{۱}{۲}(x+\frac{۱}{۲}), & x \in [-\frac{۱}{۲}, ۰] \\ ۱ - \frac{۱}{۲}x, & x \in [۰, \frac{۱}{۲}] \\ -\frac{۳}{۲}(x-۱), & x \in [\frac{۱}{۲}, ۱] \end{cases}$$

نمودار تابع $۱-x^۲$ و اسپلاین s در شکل ۱۱.۳ رسم شده است. طبق کران خطای (۴۴.۳) و با توجه به اینکه $f''(x) = -۲$ و $h = h_k = \frac{۱}{۲}$، داریم

$$\|s - f\|_\infty \leqslant \frac{(۱/۲)^۲}{۸} \times ۲ = \frac{۱}{۱۶} = ۰/۰۶۲۵.$$

اگر فاصله‌ی h را کوچکتر کنیم، کران خطا با نسبت ۲ کاهش می‌یابد. ♢

برنامه‌ی متلب اسپلاین درونیاب خطی به صورت زیر نوشته می‌شود.



شکل ۱۱.۳: نمودار $f(x) = ۱ - x^۲$ و یک اسپلاین درونیاب خطی آن

```
1 function s = LinearSpline(x, f, t)
2 n = length(x); s = [];
3 for k=1:n-1
4     ind = find( t>=x(k) & t<x(k+1));
5     tk = t(ind);
6     sk = f(k)+(tk-x(k))*(f(k+1)-f(k))/(x(k+1)-x(k));
7     s = [s sk];
8 end
9 s = [s f(end)];
```

در این برنامه x بردار نقاط، f بردار مقادیر، t برداری که اسپلاین در آن محاسبه می‌شود و s بردار مقادیر اسپلاین در t است. در سطر دوم درون حلقه دستور find اندیس نقاطی را می‌یابد که آن نقاط بین x_k و $x_{k+۱}$ قرار دارند و tk بخشی از t است که در این زیربازه قرار دارد. به عنوان نمونه، این برنامه برای مثال قبل به صورت زیر فراخوانی شده است.

```
1 x = [-1 -.5 0 .5 1]; f = 1-x.^2; t = -1:0.01:1;
2 s = LinearSpline(x, f, t);
3 plot(t,1-t.^2,'-b', t,s,'--r')
```

بنابر نوبت هم که باشد، باید در اینجا به اسپلاین‌های درونیاب درجه دو بپردازیم. اما بنابر دلایلی که در بعداً توضیح

خواهیم داد، اسپلاین‌های درجه دوم را کنار گذاشته و درونیابی با اسپلاین‌های درجه سه را مطرح می‌کنیم. شاید بتوان گفت اسپلاین‌های درجه سه که به آنها *اسپلاین‌های مکعبی* نیز می‌گویند بیش از بقیه‌ی اسپلاین‌ها در تقریب و درونیابی مورد استفاده قرار می‌گیرند، زیرا طبق تعریف، این نوع اسپلاین‌ها دارایِ همواری $C^۲[a,b]$ هستند که در بسیاری از کاربردهای فیزیکی و مهندسی کافی است و نیازی به اسپلاین‌های درجه بالاتر، که تولید و کار با آنها هزینه‌ی محاسباتی بیشتری می‌طلبد، نیست.

اگر باز هم به تعریف بازگردیم، یک اسپلاین مکعبی مانند s روی هر زیر بازه‌ی $[x_k, x_{k+۱}]$ یک چندجمله‌ای درجه سه است و در نقاط گره‌ای داریم

$$s(x_k^-) = s(x_k^+), \quad s'(x_k^-) = s'(x_k^+), \quad s''(x_k^-) = s''(x_k^+), \quad k = ۱, ۲, \ldots, n-۱, \qquad (۴۵.۳)$$

که در آن منظور از بالا اندیس‌های $+$ و $-$ مقادیر حد سمت راست و حد سمت چپ در نقاط x_k است. در حقیقت معادلات بالا نشان‌دهنده‌ی پیوستگیِ s، s' و s'' در نقاط گره‌ای می‌باشند. لازم به ذکر است که در دو نقطه‌ی ابتدایی و انتهایی بازه، یعنی $x_۰$ و x_n نیازی به اعمال شرایط پیوستگی s و مشتقاتش نیست. معادلات (۴۵.۳) تعداد $۳(n-۱)$ شرط برای تعیین اسپلاین به ما ارائه می‌دهند. اگر بخواهیم اسپلاین s تابع f را در نقاط $X = \{x_۰, x_۱, \ldots, x_n\}$ درونیابی کند، علاوه بر شرایط بالا $n+۱$ شرط درونیابی

$$s(x_k) = f_k, \quad k = ۰, ۱, ۲, \ldots, n, \qquad (۴۶.۳)$$

هم اضافه می‌شوند. بنابراین برای تعیین اسپلاین درونیاب مکعبی، $۳(n-۱)+n+۱ = ۴n-۲$ معادله‌ی معلوم داریم. اما مجهولات ما چقدر است؟ برای تعیین اسپلاین کافی است ضابطه‌های اسپلاین در هر زیربازه را تعیین کنیم. هر ضابطه شامل ۴ مجهول است زیرا هر ضابطه به صورت یک $\alpha_۰ + \alpha_۱x + \alpha_۲x^۲ + \alpha_۳x^۳$ می‌باشد. چون n ضابطه داریم پس تعداد کل مجهولات $۴n$ است. بنابراین دستگاه معادلات دو درجه‌ی آزادی دارد و می‌تواند بینهایت جواب داشته باشد. در واقع بینهایت اسپلاین مکعبی وجود دارند که در شرایط درونیابی (۴۶.۳) صدق می‌کنند. در اسپلاین درونیاب خطی چنین چیزی نداشتیم و اگر به همین منوال تعداد مجهولات و معلومات آن را حساب کنیم دو عدد یکسان بدست می‌آوریم (انجام دهید!). برای اینکه یک اسپلاین درونیاب مکعبیِ یکتا بدست آوریم، می‌توان دو شرط اختیاری دیگر اضافه کرد. این دو شرط با توجه به فیزیک مسئله تعیین می‌شود و معمولاً (نه همیشه) در دو نقطه‌ی ابتدا و انتهای بازه یعنی a و b اضافه می‌شوند که به آن‌ها شرایط مرزی یا شرایط انتهایی می‌گوییم. غالباً برای اسپلاین مکعبی این شرایط اضافی به سه شکل زیر تحمیل می‌شوند:

الف: اگر اسپلاین درونیاب مکعبی به گونه‌ای تعیین شود که

$$s'(a) = f'_۰, \quad s'(b) = f'_n \qquad (۴۷.۳)$$

که در آن $f'_۰$ و f'_n مقادیر داده شده‌ی معلوم هستند (مشتقات f در ابتدا و انتهای بازه هستند)، به آن *اسپلاین درونیاب مکعبی ارمیتی* یا *اسپلاین درونیاب مکعبی مقید* می‌گوییم.

ب: اگر اسپلاین درونیاب مکعبی به گونه‌ای تعیین شود که

$$s''(a) = ۰, \quad s''(b) = ۰ \tag{۴۸.۳}$$

به آن *اسپلاین درونیاب مکعبی طبیعی* می‌گوییم.

ج: اگر تابع f و مشتق آن هر دو متناوب با دوره‌ی تناوب $b - a$ باشند و اسپلاین درونیاب مکعبی به گونه‌ای تعیین شود که

$$s'(a) = s'(b), \quad s''(a) = s''(b) \tag{۴۹.۳}$$

به آن *اسپلاین درونیاب مکعبی متناوب* می‌گوییم. در این صورت لزوماً $s(a) = s(b)$، و اسپلاین بگونه‌ای تعیین می‌شود که خودش، مشتق مرتبه اولش و مشتق مرتبه دومش متناوب هستند.

در اینجا می‌توانیم علت مطرح نکردن درونیابی با اسپلاین‌های درجه دو را بیان کنیم. اختلاف معلومات و مجهولات در درونیابی با اسپلاین درجه دوم، یک واحد است. یعنی تنها کافی است یک شرط به مسئله تحمیل کنیم که در این صورت نمی‌توان یک اسپلاین متقارن بدست آورد. روش‌های دیگری برای تعیین اسپلاین‌های درونیاب درجه دو و به طور کلی اسپلاین‌های درجه زوج وجود دارند که می‌توانید آن‌ها را در [۶] مشاهده کنید.

اکنون روشی برای تعیین اسپلاین درونیاب مکعبی ارائه می‌دهیم و تحت هر یک از شرایط سه گانه‌ی بالا یک اسپلاین یکتا بدست می‌آوریم. ضابطه‌ی s در زیربازه‌ی $[x_k, x_{k+۱}]$ را با s_k نمایش می‌دهیم، یعنی $s_k := s\big|_{[x_k, x_{k+۱}]}$. طبق شرایط درونیابی داریم

$$s_k(x_k) = f_k, \quad s_k(x_{k+۱}) = f_{k+۱}, \quad k = ۰, ۱, \ldots, n - ۱, \tag{۵۰.۳}$$

که به خودی خود پیوسته بودن s روی کل بازه‌ی $[x_۰, x_n]$ را نشان می‌دهد. از سوی دیگر فرض کنیم

$$s'_k(x_k) =: m_k, \quad s'_k(x_{k+۱}) =: m_{k+۱}, \quad k = ۰, ۱, \ldots, n - ۱. \tag{۵۱.۳}$$

که این فرض نیز پیوستگی s' روی کل بازه را تضمین می‌کند زیرا

$$s'(x_k^-) = s'_{k-۱}(x_k) = m_k = s'_k(x_k) = s'(x_k^+), \quad k = ۱, ۲, \ldots, n - ۱.$$

مقادیر m_k فعلاً مجهولند اما بعداً تعیین می‌شوند. اکنون چندجمله‌ای درجه‌ی سه s_k در بازه‌ی $[x_k, x_{k+۱}]$ را طوری تعیین می‌کنیم که در شرایط درونیابی (۵۰.۳) و (۵۱.۳) صدق کند. این یک مسئله‌ی درونیابی ارمیت است. جدول تفاضلات تقسیم شده‌ی نیوتن با فرض اینکه $h_k = x_{k+۱} - x_k$ به صورت زیر است:

x_k	f_k			
x_k	f_k	m_k		
x_{k+1}	f_{k+1}	$f[x_k, x_{k+1}]$	$\frac{f[x_k,x_{k+1}]-m_k}{h_k}$	
x_{k+1}	f_{k+1}	m_{k+1}	$\frac{m_{k+1}-f[x_k,x_{k+1}]}{h_k}$	$\frac{m_{k+1}+m_k-2f[x_k,x_{k+1}]}{h_k^2}$

همانگونه که در درونیابی ارمیت دیدیم، چندجمله‌ای درونیاب عبارتست از

$$s_k(x) = f_k + m_k(x - x_k) + \frac{f[x_k, x_{k+1}] - m_k}{h_k}(x - x_k)^2 + \frac{m_{k+1} + m_k - 2f[x_k, x_{k+1}]}{h_k^2}(x - x_k)^2(x - x_{k+1}),$$

که می‌توان آن را به شکل ساده‌ی زیر نوشت

$$s_k(x) = \alpha_{k,0} + \alpha_{k,1}(x - x_k) + \alpha_{k,2}(x - x_k)^2 + \alpha_{k,3}(x - x_k)^3, \quad x \in [x_k, x_{k+1}], \tag{۵۲.۳}$$

که در آن ضرایب $\alpha_{k,j}$ با توجه به اینکه $(x - x_{k+1}) = (x - x_k) - h_k$، به صورت زیر بدست می‌آیند

$$\begin{aligned} \alpha_{k,0} &= f_k, \\ \alpha_{k,1} &= m_k, \\ \alpha_{k,2} &= \frac{f[x_k, x_{k+1}] - m_k}{h_k} - h_k\alpha_{k,3}, \\ \alpha_{k,3} &= \frac{m_{k+1} + m_k - 2f[x_k, x_{k+1}]}{h_k^2}. \end{aligned} \tag{۵۳.۳}$$

بنابراین برای تعیین s_k کافی است ضرایب m_k تعیین شوند و با جایگذاری آنها در (۵۳.۳) ضرایب $\alpha_{k,j}$ بدست آیند. برای محاسبه‌ی m_k ها از شرط پیوستگی مشتق دوم s استفاده می‌کنیم. داریم

$$s''_{k-1}(x_k) = s''_k(x_k), \quad k = 1, 2, \ldots, n - 1. \tag{۵۴.۳}$$

با دو بار مشتق‌گیری از فرمول (۵۲.۳) و بازنویسی آن برای زیر بازه‌ی $[x_{k-1}, x_k]$ و سپس اعمال شرایط پیوستگی (۵۴.۳) به معادله‌ی زیر می‌رسیم

$$2\alpha_{k-1,2} + 6h_{k-1}\alpha_{k-1,3} = 2\alpha_{k,2}, \quad k = 1, 2, \ldots, n - 1.$$

اگر ضرایب $\alpha_{k,j}$ را از فرمول‌های (۵۳.۳) جایگزین کنیم و در آخر طرفین را بر $(h_{k-1} + h_k)$ تقسیم کنیم به معادلات زیر می‌رسیم

$$\lambda_k m_{k-1} + 2m_k + \beta_k m_{k+1} = d_k, \quad k = 1, 2, \ldots, n - 1, \tag{۵۵.۳}$$

که در آن مقادیر λ_k، β_k و d_k برای $1 \leqslant k \leqslant n-1$ به صورت زیرند

$$\lambda_k = \frac{h_k}{h_{k-1}+h_k}, \quad \beta_k = \frac{h_{k-1}}{h_{k-1}+h_k}, \quad d_k = 3\lambda_k f[x_{k-1}, x_k] + 3\beta_k f[x_k, x_{k+1}]. \tag{۵۶.۳}$$

معادلات (۵۵.۳) یک دستگاه با $n-1$ معادله و $n+1$ مجهول m_k، $0 \leqslant k \leqslant n$، تشکیل می‌دهند. با اعمال هر یک از شرایط مرزی سه گانه، تعداد معادلات و مجهولات را برابر کرده و ضرایب m_k را به صورت یکتا تعیین می‌کنیم.

ابتدا شرایط مرزی ارمیتی را بررسی می‌کنیم. طبق دو شرط (۴۷.۳) به سادگی داریم

$$m_0 = f'_0, \quad m_n = f'_n.$$

اگر این مقادیر را در معادلات (۵۵.۳) جایگزین کنیم به دستگاه معادلات سه قطری زیر برای تعیین مقادیر m_1، ...، m_{n-1} می‌رسیم،

$$\begin{bmatrix} 2 & \beta_1 & & & \\ \lambda_2 & 2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda_{n-2} & 2 & \beta_{n-2} \\ & & & \lambda_{n-1} & 2 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{n-2} \\ m_{n-1} \end{bmatrix} = \begin{bmatrix} \widehat{d}_1 \\ d_2 \\ \vdots \\ d_{n-2} \\ \widehat{d}_{n-1} \end{bmatrix}, \tag{۵۷.۳}$$

این دستگاه از بعد $(n-1) \times (n-1)$ است و در آن

$$\widehat{d}_1 = d_1 - \lambda_1 m_0 = d_1 - \lambda_1 f'_0, \quad \widehat{d}_{n-1} = d_{n-1} - \beta_{n-1} m_n = d_{n-1} - \beta_{n-1} f'_n, \quad n \geqslant 2.$$

با توجه به اینکه $\lambda_k > 0$ و $\beta_k > 0$ و $\lambda_k + \beta_k = 1$، نتیجه می‌گیریم ماتریس سه قطری بالا اکیداً غالب قطر سطری است و بنابراین معکوس پذیر است. پس ضرایب m_k به صورت یکتا با حل این دستگاه تعیین می‌شوند. سپس ضرایب $\alpha_{k,j}$ طبق فرمول‌های (۵۳.۳) محاسبه می‌شوند و اسپلاین روی هر زیربازه توسط (۵۲.۳) نوشته می‌شود.

برای بدست آوردن اسپلاین درونیاب طبیعی باید از دو شرط (۴۸.۳) استفاده کنیم. ضابطه‌ی اسپلاین در زیر بازه‌ی اول s_0 است و طبق (۵۲.۳) داریم

$$s''_0(x) = 2\alpha_{0,2} + 6\alpha_{0,3}(x-a).$$

با اعمال شرط مرزی $s''(a) = 0$ و استفاده از فرمول‌های (۵۳.۳) به معادله‌ی زیر می‌رسیم

$$2m_0 + m_1 = 3f[x_0, x_1]. \tag{۵۸.۳}$$

به همین ترتیب با استفاده از ضابطه‌ی زیر بازه‌ی آخر و اعمال شرط طبیعی $s''(b) = 0$ به معادله‌ی

$$m_{n-1} + 2m_n = 3f[x_{n-1}, x_n] \tag{۵۹.۳}$$

خواهیم رسید. اگر معادله‌ی (۵۸.۳) را به ابتدا و معادله‌ی (۵۹.۳) را به انتهای معادلات (۵۵.۳) اضافه کنیم، به دستگاه معادلات سه قطری زیر برای تعیین مقادیر $m_۰$، $m_۱$، ...، m_n می‌رسیم

$$\begin{bmatrix} ۲ & ۱ & & & & \\ \lambda_۱ & ۲ & \beta_۱ & & & \\ & \ddots & \ddots & \ddots & & \\ & & \lambda_{n-۱} & ۲ & \beta_{n-۱} \\ & & & ۱ & ۲ \end{bmatrix} \begin{bmatrix} m_۰ \\ m_۱ \\ \vdots \\ m_{n-۱} \\ m_n \end{bmatrix} = \begin{bmatrix} d_۰ \\ d_۱ \\ \vdots \\ d_{n-۱} \\ d_n \end{bmatrix}, \tag{۶۰.۳}$$

که از بعد $(n+۱) \times (n+۱)$ است و در آن مقادیر $d_۱$، ...، $d_{n-۱}$ از (۵۶.۳) بدست می‌آیند و $d_۰ = ۳f[x_۰, x_۱]$ و $d_n = ۳f[x_{n-۱}, x_n]$. واضح است که این دستگاه نیز اکیداً غالب قطر سطری و لذا معکوس‌پذیر است.

اسپلاین درونیاب متناوب نیز با اضافه کردن شرایط (۴۹.۳) تعیین می‌شود. از شرط $s'(a) = s'(b)$ نتیجه می‌گیریم

$$m_۰ = m_n,$$

و همانند آنچه در قبل گفته شد با نوشتن ضابطه‌ی اسپلاین در بازه‌ی اول و آخر و مشتق‌گیری از آنها داریم

$$s''(a) = ۲\alpha_{۰,۲} = -\frac{۲}{h_۰}\left(۲m_۰ + m_۱ - ۳f[x_۰, x_۱]\right),$$
$$s''(b) = ۲\alpha_{n-۱,۲} + ۶h_{n-۱}\alpha_{n-۱,۳} = \frac{۲}{h_{n-۱}}\left(m_{n-۱} + ۲m_n - ۳f[x_{n-۱}, x_n]\right).$$

طبق شرط $s''(a) = s''(b)$ و با استفاده از $m_۰ = m_n$ به معادله‌ی زیر می‌رسیم

$$۲m_۰ + \beta_۰ m_۱ + \lambda_n m_{n-۱} = d_۰, \tag{۶۱.۳}$$

که در آن

$$\beta_۰ = \frac{h_{n-۱}}{h_۰ + h_{n-۱}}, \quad \lambda_n = \frac{h_۰}{h_۰ + h_{n-۱}}, \quad d_۰ = ۳\beta_۰ f[x_۰, x_۱] + ۳\lambda_n f[x_{n-۱}, x_n].$$

با توجه به متناوب بودن اسپلاین، با فرض اینکه یک بازه مجازی مانند $[x_n, x_{n+۱}]$ به طول $h_۰$ به انتهای بازه‌ی اصلی و یک بازه‌ی مجازی مانند $[x_{-۱}, x_۰]$ به طول $h_{n-۱}$ به ابتدای بازه‌ی اصلی اضافه کنیم، این تعاریف با (۵۶.۳) هم‌خوانی خواهند داشت. اگر معادله‌ی (۶۱.۳) را به ابتدای معادلات (۵۵.۳) اضافه کنیم (با توجه به اینکه $m_n = m_۰$ نیازی به اضافه کردن یک معادله‌ی دیگر به انتهای معادلات (۵۵.۳) نیست)، به دستگاه زیر می‌رسیم

$$\begin{bmatrix} ۲ & \beta_۰ & & & \lambda_n \\ \lambda_۱ & ۲ & \beta_۱ & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda_{n-۲} & ۲ & \beta_{n-۲} \\ & & & \lambda_{n-۱} & ۲ \end{bmatrix} \begin{bmatrix} m_۰ \\ m_۱ \\ \vdots \\ m_{n-۲} \\ m_{n-۱} \end{bmatrix} = \begin{bmatrix} d_۰ \\ d_۱ \\ \vdots \\ d_{n-۲} \\ d_{n-۱} \end{bmatrix}, \tag{۶۲.۳}$$

این دستگاه اگرچه دیگر سه قطری نیست اما اکیداً غالب قطر سطری و معکوس‌پذیر است. با حل آن مقادیر $m_۰$، . . .، $m_{n-۱}$ تعیین می‌شوند و در آخر قرار می‌دهیم $m_n = m_۰$.

مثال ۱۱.۳. می‌خواهیم اسپلاین‌های مکعبی که داده‌های $(۰, ۲)$، $(-۱, ۱)$ و $(۲, ۳)$ را درونیابی می‌کنند تحت هر یک از شرایط مرزی بدست آوریم. ابتدا اسپلاین ارمیتی را بررسی می‌کنیم که به دو شرط دیگر روی مشتقات تابع در دو انتها نیاز دارد. فرض کنیم این دو شرط با $s'(۰) = -۲$ و $s'(۳) = ۱$ داده شده باشند. بنابراین $m_۰ = -۲$ و $m_۲ = ۱$. با توجه به اینکه $h_۰ = ۱$، $h_۱ = ۲$، داریم $\lambda_۱ = \frac{۲}{۳}$، $\beta_۱ = \frac{۱}{۳}$ و $\widehat{d}_۱ = -\frac{۱۹}{۶}$. دستگاه (۵۷.۳) یک دستگاه ۱×۱ به صورت $۲m_۱ = -\frac{۱۹}{۶}$ خواهد بود که می‌دهد $m_۱ = -\frac{۱۹}{۱۲}$. با توجه به مقادیر $m_۰$، $m_۱$ و $m_۲$ ضرایب $\alpha_{k,j}$ بدست می‌آیند. با محاسبه‌ی آن‌ها تا چهار رقم اعشار داریم

$$s(x) \doteq \begin{cases} ۲ - ۲x - ۳/۴۱۶۷x^۲ + ۲/۴۱۶۷x^۳, & x \in [۰, ۱], \\ -۱ - ۱/۵۸۳۳(x-۱) + ۳/۳۳۳۳(x-۱)^۲ - ۰/۸۹۵۸(x-۱)^۳, & x \in [۱, ۳]. \end{cases}$$

در اسپلاین درونیاب طبیعی، بدست می‌آوریم $d_۰ = -۹$، $d_۱ = -۴/۵$ و $d_۲ = ۴/۵$. دستگاه معادلات (۶۰.۳) به شکل زیر است

$$\begin{bmatrix} ۲ & ۱ & ۰ \\ \frac{۲}{۳} & ۲ & \frac{۱}{۳} \\ ۰ & ۱ & ۲ \end{bmatrix} \begin{bmatrix} m_۰ \\ m_۱ \\ m_۲ \end{bmatrix} = \begin{bmatrix} -۹ \\ -۴/۵ \\ ۴/۵ \end{bmatrix},$$

که حل آن منجر به جواب $[m_۰, m_۱, m_۲] = [-۳/۷۵, -۱/۵, ۳]$ می‌شود. به کمک این مقادیر، ضرایب $\alpha_{k,j}$ را با فرمول‌های (۵۳.۳) محاسبه می‌کنیم و ضابطه‌های اسپلاین را به صورت زیر می‌نویسیم

$$s(x) = \begin{cases} ۲ - ۳/۷۵x + ۰/۷۵x^۳, & x \in [۰, ۱], \\ -۱ - ۱/۵(x-۱) + ۲/۲۵(x-۱)^۲ - ۰/۳۷۵(x-۱)^۳, & x \in [۱, ۳]. \end{cases}$$

با توجه به اینکه در داده‌های این مسئله داریم $f_۰ = f_۲$، پس اسپلاین درونیاب متناوب نیز قابل تعریف است. داریم $\beta_۰ = \frac{۲}{۳}$ و $\lambda_۲ = \frac{۱}{۳}$ و $d_۰ = -\frac{۹}{۲}$. پس دستگاه معادلات (۶۲.۳) برای این مثال به صورت زیر است

$$\begin{bmatrix} ۲ & ۱ \\ \frac{۲}{۳} & ۲ \end{bmatrix} \begin{bmatrix} m_۰ \\ m_۱ \end{bmatrix} = \begin{bmatrix} -۴/۵ \\ -۴/۵ \end{bmatrix}.$$

درایه‌ی سطر اول و ستون دوم ماتریس بالا (درایه ۱) مجموع $\beta_۰ + \lambda_۲$ است. زیرا طبق (۶۱.۳) برای $n = ۲$ خواهیم داشت $۲m_۰ + (\beta_۰ + \lambda_۲)m_۱ = d_۰$. با حل این دستگاه $m_۰ = -۱/۳۵$ و $m_۱ = -۱/۸$ بدست می‌آیند و $m_۲ = m_۰$. ادامه‌ی محاسبات که شامل نوشتن ضابطه‌ی اسپلاین است به خواننده واگذار می‌شود. ◇

در اینجا برنامه‌ای به زبان مت‌لب می‌نویسیم که اسپلاین درونیاب مکعبی با هر یک از شرایط مرزی سه گانه را تولید می‌کند.

```
1  function s = CubicSpline(char, x, f, t, f0, fn)
2  n = length(x); h = x(2:n)-x(1:n-1);
3  if n<3 error('length of x should be =>3'); end
4  lam = h(2:n-1)./(h(1:n-2)+h(2:n-1)); bet = 1-lam;
5  d = 3*lam.*(f(2:n-1)-f(1:n-2))./h(1:n-2)+ ...
6      3*bet.*(f(3:n)-f(2:n-1))./h(2:n-1);
7  switch (char)
8      case ('hermite')
9          A = diag(2*ones(1,n-2))+diag(lam(2:n-2),-1)+diag(bet(1:n-3),1);
10         d(1) = d(1)- lam(1)*f0;
11         if n>3 d(n-2)= d(n-2)-bet(n-2)*fn; end
12         m = A\d'; m = [f0;m;fn];
13     case ('natural')
14         lam = [lam 1]; bet = [1 bet];
15         A = diag(2*ones(1,n))+diag(lam,-1)+diag(bet,1);
16         d0 = 3*(f(2)-f(1))/h(1); dn = 3*(f(n)-f(n-1))/h(n-1);
17         d = [d0 d dn];
18         m = A\d';
19     case ('periodic')
20         bet0 = h(n-1)/(h(n-1)+h(1)); lamn = 1-bet0;
21         bet = [bet0 bet];
22         A = diag(2*ones(1,n-1))+diag(lam,-1)+diag(bet(1:n-2),1);
23         if n>2 A(1,n-1)=lamn; else A(1,2)=A(1,2)+lamn; end
24         d0 = 3*bet0*(f(2)-f(1))/h(1)+3*lamn*(f(n)-f(n-1))/h(n-1);
25         m = A\[d0 d]'; m = [m;m(1)];
26 end
```

```
27 a0 = f(1:n-1);
28 a1 = m(1:n-1)';
29 a3 = (m(2:n)'+m(1:n-1)'-2*(f(2:n)-f(1:n-1))./h)./h.^2;
30 a2 = ((f(2:n)-f(1:n-1))./h-m(1:n-1)')./h-h.*a3;
31 s = [];
32 for k=1:n-1
33     ind = find( t>=x(k) & t<x(k+1));
34     tk = t(ind);
35     sk = a0(k)+a1(k)*(tk-x(k))+a2(k)*(tk-x(k)).^2+a3(k)*(tk-x(k)).^3;
36     s = [s sk];
37 end
38 s = [s f(n)];
```

در ورودی‌های این برنامه رشته‌ی char نوع اسپلاین را مشخص می‌کند که یکی از مقادیر 'hermite'، 'natural' یا 'periodic' را توسط کاربر اختیار می‌کند. متغیرهای x، f، t و s همان هایی هستند که در برنامه‌ی قبل (اسپلاین خطی) معرفی شدند. متغیرهای f0 و fn مقادیر مشتق در دو نقطه‌ی انتهایی را در بر دارند و فقط برای اسپلاین ارمیتی به عنوان ورودی وارد می‌شوند. در واقع اسپلاین طبیعی و متناوب چهار ورودی و اسپلاین ارمیتی شش ورودی دارد.

نکته‌ی دیگری که باید به آن اشاره کنیم نحوه‌ی حل دستگاه‌های معادلات خطی سه قطری است. در برنامه‌ی بالا، این دستگاه‌ها با دستور "بک‌اسلش" متلب که با روش حذفی گاوس با محورگیری (در صورت مربعی بودن دستگاه) عمل می‌کند، حل کردیم. اما اگر n بزرگ باشد بهتر است روش از حذفی گاوس مخصوص دستگاه‌های سه قطری که هزینه‌ی محاسباتی آن $\mathcal{O}(n)$ است، استفاده کنیم.

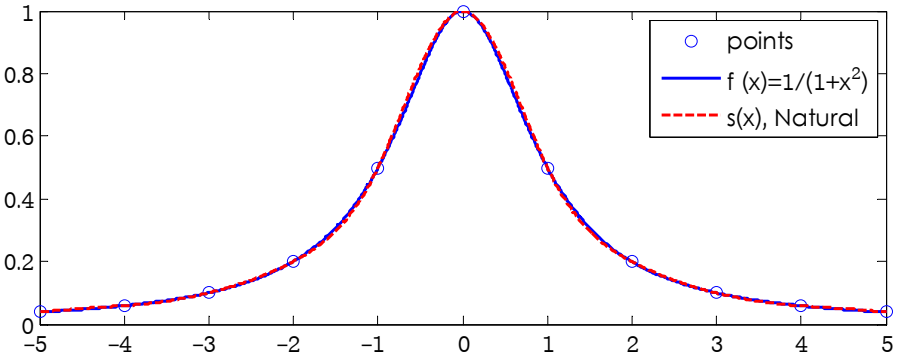
مثال ۱۲.۳. در این مثال درونیاب‌های اسپلاین مکعبی تابع رونگه یعنی $f(x) = ۱/(۱ + x^۲)$ روی $[-۵, ۵]$ را بررسی می‌کنیم. قبلاً دیده‌ایم که درونیاب‌های چندجمله‌ای روی نقاط هم‌فاصله برای این تابع واگرا هستند. فرض کنیم نقاط درونیابی به صورت هم‌فاصله با فاصله‌ی $h = ۱$ در بازه قرار گرفته باشند، یعنی

$$x_k = -۵ + k, \quad k = ۰, ۱, \ldots, ۱۰.$$

اسپلاین درونیاب طبیعی با فراخوانی برنامه به صورت زیر محاسبه می‌شود.

```
1 g = @(x) 1./(1+x.^2);
2 x= -5:1:5; f = g(x); t=-5:0.01:5;
3 s = CubicSpline('natural', x, f, t);
4 plot(x,f,'bo',t,g(t),'-b',t,s,'--r')
5 err = norm(s-g(t),inf)
```

نمودارهای تابع و درونیاب آن به همراه نقاط درونیابی در شکل ۱۲.۳ رسم شده‌اند. نمودارها تقریباً بر هم منطبق هستند. خطای درونیابی روی بردار t که در متغیر err ذخیره شده است تقریباً ۰٫۰۲۲۰ است. با کوچکتر کردن h دقت درونیابی



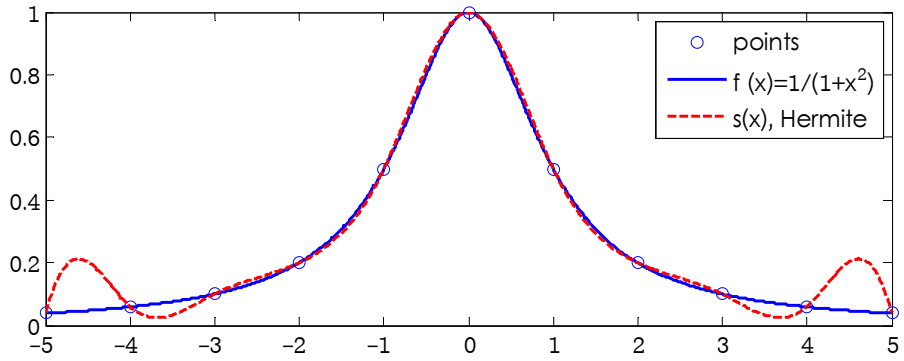
شکل ۱۲.۳: نمودار تابع $f(x) = \frac{1}{1+x^2}$ و اسپلاین درونیاب مکعبی طبیعی آن

نیز بهتر می‌شود. اسپلاین درونیاب متناوب را می‌توانید با جایگزینی 'natural' با 'periodic' در سطر سوم برنامه بالا، محاسبه کنید. انجام آن به شما واگذار می‌شود. در اینجا اسپلاین ارمیتی را بررسی می‌کنیم. ابتدا فرض کنیم مقادیر مشتق در نقاط انتهایی به صورت $s'(-5) = 1$ و $s'(5) = -1$ داده شده‌اند. برای فراخوانی برنامه‌ی اصلی، به جای سطر سوم برنامه بالا می‌نویسیم

```
1 s = CubicSpline('hermite', x, f, t, 1, -1);
```

و بقیه دستورات را تغییر نمی‌دهیم. نمودار ۱۳.۳ حاصل اجرای این برنامه است، که در نزدیکی دو انتها با تابع اصلی تفاوت زیادی دارد. علت این امر انتخاب مقادیر نامناسب برای مشتقات اسپلاین در نقاط ابتدا و انتها است. اما همانگونه که در این شکل می‌بینیم اختلاف تابع و درونیاب فقط در بازه‌های نزدیک به دو انتها زیاد است و اگر قدری از کناره‌ها فاصله

بگیریم تقریب مناسبی خواهیم داشت. این یک ویژگی مثبتِ اسپلاین‌ها است که یک خصوصیت بدِ موضعی کل دامنه را تحت تأثیر قرار نمی‌دهد. تقریب‌های سراسری معمولاً چنین ویژگی‌ای ندارند.

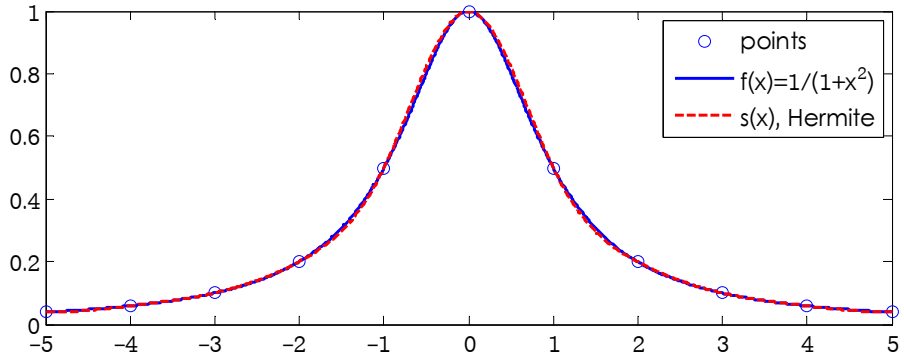


شکل ۱۳.۳: نمودار تابع $f(x) = \frac{۱}{۱+x^۲}$ و اسپلاین درونیاب مکعبی ارمیتی آن با فرض $s'(-۵) = ۱$ و $s'(۵) = -۱$.

حال اگر شرایط مرزی ارمیتی را با $s'(-۵) = ۰$ و $s'(۵) = ۰$ جایگزین کنیم، که با شیب تابع رونگه در نقاط ابتدا و انتها تناسب بیشتری دارند، به نمودار شکل ۱۴.۳ می‌رسیم که در آن نمودار تابع و درونیاب آن تقریباً بر هم منطبق هستند. این حالت که با جایگزینی سطر سوم با

```
1 s = CubicSpline('hermite', x, f, t, 0, 0);
```

اجرا شده است دارای خطای درونیابی در حدود ۰٫۰۲۲۷ است. ♢



شکل ۱۴.۳: نمودار تابع $f(x) = \frac{۱}{۱+x^۲}$ و اسپلاین درونیاب مکعبی ارمیتی آن با فرض $s'(-۵) = s'(۵) = ۰$.

روشِ مشابه دیگر برای تعیین اسپلاین‌های درونیاب مکعبی، روش انتگرال‌گیری پی در پی است که در آن از خاصیت قطعه‌ای خطی بودن s'' استفاده می‌کنیم و فرمولی برای آن بر حسب مقادیر مجهول $q_k := s''(x_k)$ می‌نویسیم. سپس

با یک بار انتگرال‌گیری، فرمولی برای s' بدست می‌آوریم و با انتگرال‌گیری مجدد s را تعیین می‌کنیم. مقادیر q_k با حل دستگاه‌هایی مشابه دستگاه‌های ما برای m_k بدست می‌آیند. جزئیات این روش را می‌توانید در فصل دوم مرجع [۸] ببینید.

تا به اینجا حتماً به این نکته توجه کرده‌اید که بر خلاف درونیابی چندجمله‌ای که با افزایش درجه‌ی درونیاب به دنبال همگرایی بودیم (که گاهی حاصل نمی‌شد)، در درونیابی با اسپلاین‌ها درجه‌ی اسپلاین از ابتدا ثابت است (مثلاً در اسپلاین مکعبی، سه است) و با کاهش فاصله‌ی h به دنبال همگرایی هستیم. با اجرای برنامه متلب بالا و با کاهش h ملاحظه خواهید کرد که خطای درونیابی کاهش می‌یابد. جالب است بدانید که این اتفاق نه تنها برای خود تابع می‌افتد بلکه مشتقات s نیز به مشتقات f همگرا خواهند بود. قضیه‌ی زیر که تنها یک حالت خاص (اسپلاین مکعبی با شرایط مرزی ارمیتی) را در بر دارد، بیانگر همین نکته است. ما این قضیه را بدون اثبات مطرح می‌کنیم. علاقه‌مندان می‌توانند جزئیات اثبات آن را در مراجع [۸] مشاهده کنند.

قضیه ۸.۳. *فرض کنیم $f \in C^۴[a,b]$ و $X = \{a = x_۰ < x_۱ < \cdots < x_n = b\}$ یک افراز از $[a,b]$ باشد. قرار می‌دهیم*

$$h = \max_{۰ \leqslant j \leqslant n-۱} |x_{j+۱} - x_j|, \quad \eta = \min_{۰ \leqslant j \leqslant n-۱} |x_{j+۱} - x_j|,$$

و فرض می‌کنیم L ثابتی باشد که $\frac{h}{\eta} \leqslant L$. آنگاه برای اسپلاین مکعبی $s \in \mathbb{S}^۳_X$ که تابع f را با شرایط مرزی ارمیتی در نقاط x_k، $۰ \leqslant k \leqslant n$، درونیابی می‌کند، داریم

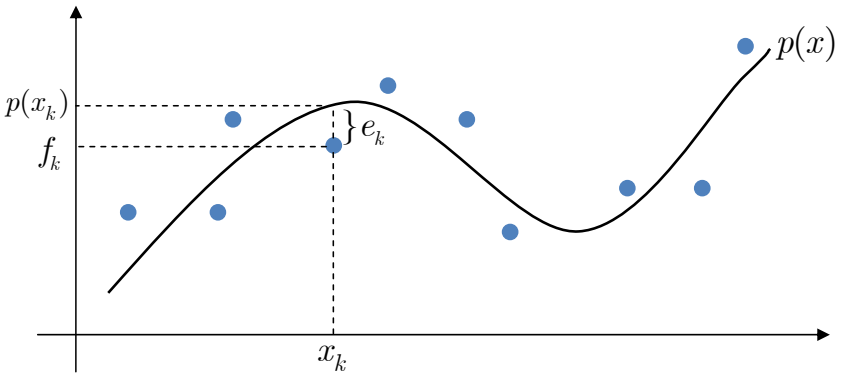
$$\|f^{(j)} - s^{(j)}\|_\infty \leqslant L c_j h^{۴-j} \|f^{(۴)}\|_\infty, \quad j = ۰, ۱, ۲, ۳, \tag{۶۳.۳}$$

که در آن $c_j \leqslant ۲$. □

در قضیه بالا ثابت $L \geqslant ۱$ میزان انحرافِ توزیع نقاط از یکنواختی را نشان می‌دهد. اگر L متناهی و نسبتاً کوچک باشد می‌گوییم توزیع نقاط شبه-یکنواخت است. این کران‌های خطا نشان می‌دهند، نه تنها خود درونیاب s بطور یکنواخت به f همگراست، بلکه مشتقات آن تا مرتبه سه نیز به طور یکنواخت به مشتقات متناظر f همگرایند. بخصوص s''' که یک تابع ناپیوسته قطعه‌ای ثابت است نیز به f''' همگرای یکنواخت است.

۸.۳ برازش منحنی

یکی دیگر از روش‌های تقریب، روش "برازش منحنی" است که در آن الزامی نیست که منحنی تقریب از نقاط عبور کند، بلکه کافی است با فاصله‌ی کمی از نزدیکی آن‌ها بگذرد. در این بخش برازش منحنی به کمک چندجمله‌ایها را بررسی می‌کنیم. چند حالت خاص غیر چندجمله‌ای را هم در تمرینات مطرح می‌کنیم. در برازش منحنیِ چندجمله‌ای، درجه‌ی چندجمله‌ای از ابتدا ثابت است و ضرایب آن را طوری بدست می‌آوریم که چندجمله‌ای، تقریب مناسبی برای نقاط باشد. شکل ۱۵.۳ را ببینید. وقتی تعداد نقاط بسیار زیاد باشد، درونیابی منجر به یک چندجمله‌ای با درجه‌ی بالا می‌شود که غالباً نوسانات زیادی



شکل ۱۵.۳: شمایی از برازش منحنی

دارد. در این حالت برازش منحنی ایده‌ی بهتری است. فرض کنیم همانند قبل نقاط $X = \{x_0, x_1, \ldots, x_n\}$ با مقادیر متناظر f_0، f_1، ...، f_n داده شده‌اند. به دنبال یافتن تقریب در فضای $\mathbb{P}_m$ برای $m \leqslant n$ هستیم. بنابراین چندجمله‌ای تقریب را به صورت

$$p_m(x) = a_0 + a_1 x + \cdots + a_m x^m,$$

بسط می‌دهیم و ضرایب a_j را طوری بدست می‌آوریم که خطاهای

$$e_k := p_m(x_k) - f_k, \quad k = 0, 1, \ldots, n,$$

تا آنجا که امکان دارد کوچک باشند. برای این کار فرض می‌کنیم $\boldsymbol{e} = [e_0, e_1, \ldots, e_n]$ بردار خطا باشد و برای اینکه بهترین تقریب را بدست آوریم یک نرم از این بردار را مینیمم می‌کنیم. ساده‌ترین حالت وقتی اتفاق می‌افتد که نرم ۲ آن یعنی $\|\boldsymbol{e}\|_2 = \sum_{k=0}^{n} e_k^2$ را مینیمم کنیم. پس مسئله به صورت زیر مطرح می‌شود:

مسئله (برازش منحنیِ چندجمله‌ای)**.** ضرایب a_j، $0 \leqslant j \leqslant n$، از چندجمله‌ای $p_m \in \mathbb{P}_m$ را طوری بدست آورید که خطای

$$E(a_0, a_1, \ldots, a_m) := \|\boldsymbol{e}\|_2 = \sum_{k=0}^{n} e_k^2, \tag{۶۴.۳}$$

مینیمم شود. با توجه به اینکه در این مسئله مجموع مربعات خطا مینیمم می‌شود، گاهی به آن "تقریب کمترین مربعات" نیز می‌گویند.

اولین سؤالی که پیش می‌آید این است که آیا چنین مسئله‌ای خوش‌تعریف است؟ یعنی آیا این مسئله جواب دارد و آیا جواب آن یکتاست؟ به عبارت دیگر آیا بردار یکتای $\boldsymbol{a} = [a_0, a_1, \ldots, a_m] \in \mathbb{R}^{m+1}$ یافت می‌شود که تابع E را مینیمم کند؟ جواب این سؤال مثبت است و در دروس پیشرفته‌تر اثبات می‌شود که در اینجا به آن نمی‌پردازیم و فقط روشی برای بدست آوردن بردار ضرایب $\boldsymbol{a}$ ارائه می‌دهیم.

ابتدا حالت خاص $m = ۱$ را بررسی می‌کنیم. در این حالت

$$E(a_۰, a_۱) = \sum_{k=۰}^{n} [(a_۰ + a_۱ x_k) - f_k]^۲.$$

واضح است که E نسبت به دو متغیر $a_۰$ و $a_۱$ یک چندجمله‌ای درجه دو (سهمی‌گون) و مشتق‌پذیر روی $\mathbb{R}^۲$ است. طبق آنچه در دروس حسابان دانشگاه خوانده‌اید، اکسترمم‌های E جایی رخ می‌دهند که

$$\frac{\partial E}{\partial a_۰} = ۰, \quad \frac{\partial E}{\partial a_۱} = ۰.$$

یک اکسترمم، یک نقطه‌ی مینیمم سراسری است اگر ماتریس هِسین، که با

$$H(a_۰, a_۱) := \begin{bmatrix} \frac{\partial^۲ E}{\partial a_۰^۲} & \frac{\partial^۲ E}{\partial a_۰ \partial a_۱} \\ \frac{\partial^۲ E}{\partial a_۱ \partial a_۰} & \frac{\partial^۲ E}{\partial a_۱^۲} \end{bmatrix}$$

تعریف می‌شود، بازای هر $(a_۰, a_۱) \in \mathbb{R}^۲$ معین مثبت باشد. با مشتق‌گیری از E و برابر صفر قرار دادن مشتقات، داریم

$$\begin{aligned} \frac{\partial E}{\partial a_۰} &= ۲ \sum_{k=۰}^{n} [(a_۰ + a_۱ x_k) - f_k] = ۰, \\ \frac{\partial E}{\partial a_۱} &= ۲ \sum_{k=۰}^{n} x_k [(a_۰ + a_۱ x_k) - f_k] = ۰ \end{aligned}$$

که به صورت ماتریسی می‌توان آن را به شکل زیر بازنویسی کرد:

$$\begin{bmatrix} \sum_{k=۰}^{n} ۱ & \sum_{k=۰}^{n} x_k \\ \sum_{k=۰}^{n} x_k & \sum_{k=۰}^{n} x_k^۲ \end{bmatrix} \begin{bmatrix} a_۰ \\ a_۱ \end{bmatrix} = \begin{bmatrix} \sum_{k=۰}^{n} f_k \\ \sum_{k=۰}^{n} x_k f_k \end{bmatrix}, \qquad \text{یا} \qquad A\boldsymbol{a} = \boldsymbol{b}. \tag{۶۵.۳}$$

ماتریس ضرایب A متقارن است و می‌توان آن را به صورت زیر تجزیه کرد: اگر فرض کنیم

$$V = \begin{bmatrix} ۱ & x_۰ \\ ۱ & x_۱ \\ \vdots & \vdots \\ ۱ & x_n \end{bmatrix} \in \mathbb{R}^{(n+۱)\times ۲}$$

آنگاه به روشنی می‌توان دید

$$A = V^T V, \quad \boldsymbol{b} = V^T \boldsymbol{f}.$$

که در آن $\boldsymbol{f} = [f_0, f_1, \ldots, f_n]^T$. چون x_k ها متمایزند، ستون‌های V روی $\mathbb{R}^{n+1}$ مستقل خطی هستند و بنابراین V از رتبه‌ی کامل است. برای پاسخ دادن به این پرسش که آیا دستگاه $A\boldsymbol{a} = \boldsymbol{b}$ دارای جواب یکتاست، ثابت می‌کنیم ماتریس A معین مثبت است و لذا معکوس‌پذیر است. فرض کنیم $\boldsymbol{x} \in \mathbb{R}^2$ بردار دلخواه غیر صفر باشد. داریم

$$\boldsymbol{x}^T A\boldsymbol{x} = \boldsymbol{x}^T V^T V\boldsymbol{x} = (V\boldsymbol{x})^T(V\boldsymbol{x}) = \|V\boldsymbol{x}\|_2^2 \geqslant 0.$$

حال برای اثبات معین مثبت بودن A کافی است نشان دهیم $\|V\boldsymbol{x}\|_2 \neq 0$، که این هم واضح است زیرا V از رتبه‌ی کامل است و چون $\boldsymbol{x} \neq 0$ پس $V\boldsymbol{x} \neq 0$. پس ثابت کردیم دستگاه (۶۵.۳) دارای جواب یکتاست. اما آیا این جواب یک مینیمم است یا یک ماکزیمم؟ به سادگی و با مشتق‌گیری می‌توانید نشان دهید

$$H(a_0, a_1) := 2\begin{bmatrix} \sum_{k=0}^{n+1} 1 & \sum_{k=0}^{n+1} x_k \\ \sum_{k=0}^{n+1} x_k & \sum_{k=0}^{n+1} x_k^2 \end{bmatrix}$$

که ضریب مثبتی از (دو برابرِ) ماتریس A است که در بالا نشان دادیم معین مثبت است. پس جواب دستگاه (۶۵.۳) یک مینیمم سراسری روی $\mathbb{R}^2$ است. بنابراین $p_1(x) = a_0 + a_1 x$ خطِ کمترین مربعات (برازشِ خطی) داده‌های (x_0, f_0)، (x_n, f_n)، ... است. در دروس آمار گاهی به p_1 "رگرسیون" این داده‌ها نیز می‌گویند.

مثال ۱۳.۳. خط کمترین مربعات برای نقاط جدول زیر را تعیین کنید

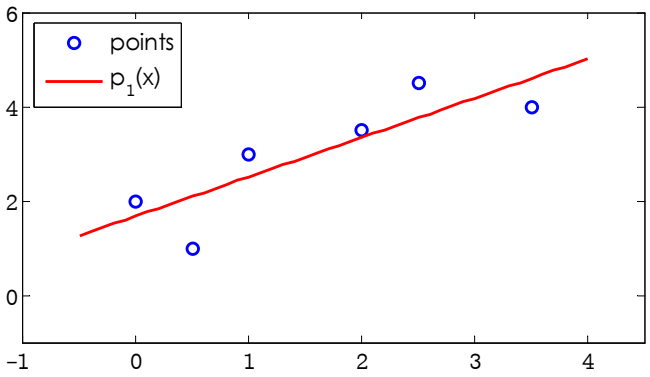
x_k	۰	۰/۵	۱	۲	۲/۵	۳/۵
f_k	۲	۱	۳	۳/۵	۴/۵	۴

می‌توانیم از دستورات زیر برای محاسبه و رسم نمودار جواب استفاده کنیم

```
1 x = [0 0.5 1 2 2.5 3.5]; f = [2 1 3 3.5 4.5 4];
2 A = [length(x) sum(x); sum(x) sum(x.^2)]; b = [sum(f);sum(f.*x)];
3 a = A\b;
4 t = -0.5:0.1:4; p = a(1)+a(2)*t;
5 plot(x,f,'bo', t,p,'-r')
```

نمودار حاصل به صورت زیر است

♢



شکل ۱۶.۳: خط کمترین مربعات

اکنون حالت کلی برازش منحنی در $\mathbb{P}_m$ را در نظر می‌گیریم. تابع $E : \mathbb{R}^{m+1} \to \mathbb{R}$ به صورت زیر تعریف می‌شود

$$E(a_0, a_1, \ldots, a_m) = \sum_{k=0}^{n} [(a_0 + a_1 x_k + \cdots + a_m x_k^m) - f_k]^2,$$

که یک چندجمله‌ای درجه دوی $m+1$ بعدی است و همانند آنچه در حالت خاص $m = 1$ گفته شد، اکسترمم‌های خود را در جایی می‌گیرد که

$$\frac{\partial E}{\partial a_j} = 0, \quad j = 0, 1, \ldots, m.$$

این معادلات به دستگاه معادلات خطی زیر منجر می‌شوند

$$\begin{bmatrix} \sum_{k=0}^{n} 1 & \sum_{k=0}^{n} x_k & \cdots & \sum_{k=0}^{n} x_k^m \\ \sum_{k=0}^{n} x_k & \sum_{k=0}^{n} x_k^2 & \cdots & \sum_{k=0}^{n} x_k^{m+1} \\ \vdots & \vdots & & \vdots \\ \sum_{k=0}^{n} x_k^m & \sum_{k=0}^{n} x_k^{m+1} & \cdots & \sum_{k=0}^{n} x_k^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{k=0}^{n} f_k \\ \sum_{k=0}^{n} x_k f_k \\ \vdots \\ \sum_{k=0}^{n} x_k^m f_k \end{bmatrix}, \quad \text{یا} \quad A\boldsymbol{a} = \boldsymbol{b}. \qquad (۶۶.۳)$$

ماتریس متقارن $A \in \mathbb{R}^{(m+1)\times(m+1)}$ و بردار سمت راست $\boldsymbol{b}$ را می‌توان همانند قبل به کمک ماتریس

$$V = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^m \\ 1 & x_1 & x_1^2 & \cdots & x_1^m \\ \vdots & \vdots & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix} \in \mathbb{R}^{(n+1)\times(m+1)}$$

به صورت $A = V^T V$ و $\boldsymbol{b} = V^T \boldsymbol{f}$ تجزیه کرد. ماتریس V یک ماتریس واندرموند مستطیلی است و چون نقاط x_k متمایز هستند، این ماتریس از رتبه‌ی کامل است. مانند قبل می‌توان معین مثبت بودن A روی $\mathbb{R}^{m+1}$ را اثبات کرد. بنابراین

بردار ضرایب $\boldsymbol{a}$ به صورت یکتا با حل دستگاه (۶۶.۳) تعیین می‌شود و جواب بدست آمده تابع $E = \|\boldsymbol{e}\|_۲$ را مینیمم می‌کند.

ملاحظه ۲.۳. دستگاه معادلات (۶۶.۳) را دستگاه معادلات نرمال می‌گویند. با توجه به اینکه ماتریس A حاصلضرب دو ماتریس واندرموند است، با افزایش m سریعاً بدوضع می‌شود. بهتر است بجای حل دستگاه نرمال مربعی $V^TV\boldsymbol{a} = V^T\boldsymbol{f}$، دستگاه مستطیلی $V\boldsymbol{a} = \boldsymbol{f}$ را با روش تجزیه‌ی QR حل کنیم. نکته‌ی دیگر اینکه اگر نقاط x_k طوری انتخاب شده باشند که ستون‌های V یکامتعامد باشند، آنگاه $A = V^TV = I$ که I ماتریس همانی است. در این صورت نیازی به حل دستگاه نیست و ضرایب $\boldsymbol{a}$ به صورت صریح بدست می‌آیند. اما در عمل معمولاً انتخاب نقاط دست ما نیست. ♡

در نرم‌افزار مت‌لب از دستور آماده‌ی `polyfit` برای بدست آوردن چندجمله‌ای برازش منحنی استفاده می‌شود. این دستور به صورت

```
a = polyfit(x,f,m)
```

وارد می‌شود و بردار ضرایب `a` را برمی‌گرداند. برای محاسبه‌ی چندجمله‌ای می‌توان از دستور `polyval` استفاده کرد. این دستورات را در مثال زیر امتحان می‌کنیم.

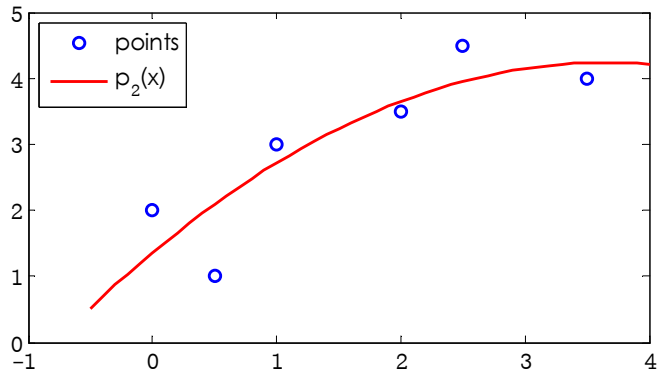
مثال ۱۴.۳. منحنی کمترین مربعات درجه دو که داده‌های مثال ۱۳.۳ را برازش کند به صورت زیر در مت‌لب محاسبه و نمودار آن رسم می‌شود (شکل ۱۷.۳ را ببینید):

```
1 x = [0 0.5 1 2 2.5 3.5]; f = [2 1 3 3.5 4.5 4];
2 a = polyfit(x, f, 2);
3 t = -0.5:0.1:4;
4 p = polyval(a, t);
5 plot(x,f,'bo', t,p,'-r')
```

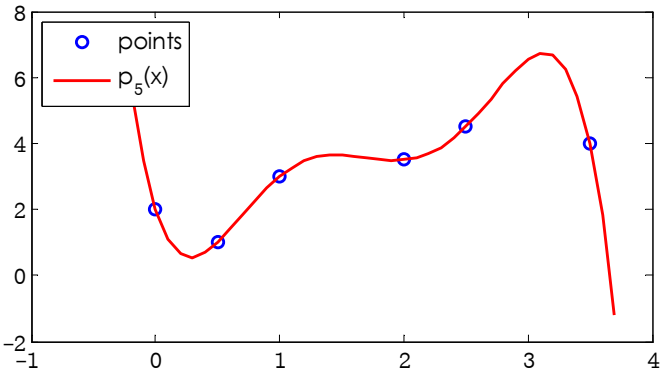
در دستور `polyfit` اگر قرار دهیم $m = n$، تقریب کمترین مربعات، همان چندجمله‌ای درونیاب خواهد شد. در پرسش ۳۰ اثبات این ادعا از شما خواسته شده است. پس اگر در برنامه‌ی بالا بجای سطر سوم دستور

```
a = polyfit(x, f, 5);
```

را بنویسیم، به شکل ۱۸.۳ می‌رسیم. ♢



شکل ۱۷.۳: منحنی کمترین مربعات درجه دو



شکل ۱۸.۳: منحنی کمترین مربعات درجه پنج (چندجمله‌ای درونیاب)

۹.۳ درونیابی چندمتغیره

روش‌هایی که تا به اینجای فصل بررسی کردیم همگی برای درونیابی و تقریب یک تابع یک متغیره به کار می‌روند. در این بخش می‌خواهیم اندکی در مورد حالت‌های چندمتغیره صحبت کنیم. مهمترین نکته‌ای که باید در همین ابتدای کار به آن اشاره کنیم این است که "مسئله‌ی درونیابی چندمتغیره همواره روی نقاط متمایز جواب یکتا ندارد." برای مثال یک چندجمله‌ای خطی در $\mathbb{R}^2$ (یعنی معادله‌ی یک صفحه) با پایه‌ی $\{1, x, y\}$ تعیین می‌شود که $(x, y) \in \mathbb{R}^2$. مشابه حالت یک بعدی، تعداد نقاط برای درونیابی برابر با تعداد اعضای پایه است. پس انتظار داریم از سه نقطه‌ی متمایز یک صفحه‌ی یکتا بگذرد، اما اگر سه نقطه روی یک خط واقع باشند بینهایت صفحه از آن‌ها می‌گذرد. توضیح بیشتر در این مورد از حوصله‌ی این درس خارج است. دانشجویان علاقه‌مند می‌توانند به فصل هفتم [۶] مراجعه کنند. در اینجا مسائل خاص روی دامنه‌های خاص و توزیع نقاط خاص را در نظر می‌گیریم که مشکل بالا را نداشته باشند. یک مثال در حالت دو بعدی، درونیابی روی مستطیل با یک شبکه‌ی منظم از نقاط است.

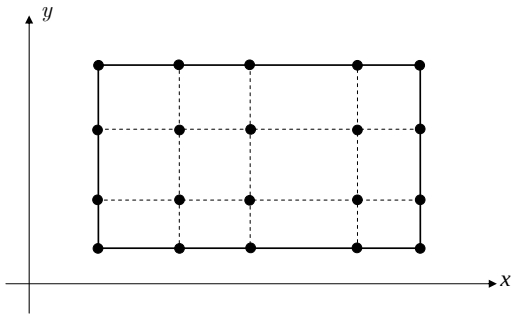
گیریم $R = [a, b] \times [c, d]$ و فرض کنیم $f \in C(R)$. فرض کنیم $[a, b]$ و $[c, d]$ بصورت زیر افراز شده‌اند:

$$\begin{aligned} a \leqslant x_0 < x_1 < \cdots < x_m \leqslant b \\ c \leqslant y_0 < y_1 < \cdots < y_n \leqslant d \end{aligned} \quad . \tag{۶۷.۳}$$

نقاط

$$(x_i, y_j), \quad i = ۰, ۱, \ldots, m, \quad j = ۰, ۱, \ldots, n \tag{۶۸.۳}$$

یک شبکه مستطیلی را تشکیل می‌دهند. در شکل زیر یک شبکه مستطیلی از نقاط برای $m = ۴$ و $n = ۳$ ترسیم شده است. حال تعریف می‌کنیم



شکل ۱۹.۳: نقاط درونیابی روی یک مستطیل

$$p_{m,n}(x, y) := \sum_{i=۰}^{m} \sum_{j=۰}^{n} f(x_i, y_j) \ell_i^x(x) \ell_j^y(y) \tag{۶۹.۳}$$

که در آن

$$\ell_i^x(x) = \prod_{\substack{k=۰ \\ k \neq i}}^{m} \frac{(x - x_k)}{x_i - x_k}, \quad \ell_j^y(y) = \prod_{\substack{k=۰ \\ k \neq j}}^{n} \frac{(y - y_k)}{y_i - y_k},$$

به ترتیب چندجمله‌ایهای لاگرانژ مبتنی بر نقاط $\{x_i\}_{i=۰}^{m}$ و $\{y_j\}_{j=۰}^{n}$ هستند. چندجمله‌ای $p_{m,n}(x, y)$ تابع $f(x, y)$ را در نقاط (۶۸.۳) درونیابی می‌کند. این چندجمله‌ای از درجه‌ی $m + n$ است اما شامل تمام تک‌جمله‌ایها از درجه‌ی $m + n$ نیست. آنچه در بالا گفته شد وجود چندجمله‌ای درونیاب روی یک شبکه مستطیلی از نقاط را اثبات می‌کند. اکنون ثابت می‌کنیم این درونیاب یکتا است. یکتایی این مسئله از یکتایی درونیاب در حالت یک بعدی نتیجه می‌شود. فرض کنیم

$$q(x, y) := \sum_{i=۰}^{m} \sum_{j=۰}^{n} c_{ij}\, x^i y^j$$

یک چندجمله‌ای درونیاب برای تابع $f(x, y)$ روی شبکه مستطیلی (۶۸.۳) باشد. برای هر $k = ۰, ۱, \ldots, m$،

$$q_k(y) := q(x_k, y) = \sum_{j=۰}^{n} \Big(\sum_{i=۰}^{m} c_{ij}\, x_k^i \Big) y^j = \sum_{j=۰}^{n} b_{kj} y^j$$

یک چندجمله‌ای یک متغیره از y است که مقادیر $f(x_k, y_j)$, $j = ۰, \ldots, n$ را درونیابی می‌کند. بنابراین طبق قضیه یکتاییِ یک متغیره، ضرایب b_{kj} بصورت یکتا تعیین می‌شوند. اکنون برای هر $۰ \leqslant j \leqslant n$ ضرایب c_{ij} از حل دستگاه

معادلات

$$\sum_{i=۰}^{m} c_{ij}\, x_k^i = b_{kj}, \quad ۰ \leqslant k \leqslant m$$

بدست می‌آیند. هر یک از $n+۱$ دستگاه بالا، با توجه به اینکه ماتریس متناظر آنها یعنی $\left[x_i^k\right]$ یک ماتریس واندرموند است، دارای جواب یکتا هستند. بنابراین در مجموع ضرایب c_{ij} بصورت یکتا تعیین می‌شوند. روی هم رفته قضیه زیر را داریم:

قضیه ۹.۳. *یک و تنها یک چندجمله‌ای $p_{m,n} \in \mathbb{P}_{m,n}$ وجود دارد که در $(m+۱)(n+۱)$ شرط درونیابی $p(x_i, y_j) = f(x_i, y_j)$ صدق می‌کند، که در آن $x_i,\ i = ۰, \ldots, m$ و $y_j,\ j = ۰, \ldots, n$ نقاط متمایز به ترتیب روی محور x و محور y هستند.* □

یک شکل ساده از (۶۹.۳)، حالت $m = n = ۱$ است که منجر به چندجمله‌ای درونیاب دوخطی زیر می‌شود

$$\begin{aligned} p_{۱,۱}(x,y) =& \frac{(b-x)(d-y)}{(b-a)(d-c)} f(a,c) + \frac{(b-x)(y-c)}{(b-a)(d-c)} f(a,d) \\ &+ \frac{(x-a)(d-y)}{(b-a)(d-c)} f(b,c) + \frac{(x-a)(y-c)}{(b-a)(d-c)} f(b,d). \end{aligned} \tag{۷۰.۳}$$

چندجمله‌ای $p_{۱,۱}(x,y)$ تابع $f(x,y)$ را در نقاط $\big\{(a,c),(a,d),(b,c),(b,d)\big\}$ درونیابی می‌کند. این چندجمله‌ای از درجه‌ی دو است اما شامل جمله‌های $x^۲$ و $y^۲$ نیست.

کران خطای درونیابی دوبعدی روی مستطیل از روی کران خطای (۳.۳) حاصل می‌شود. جزئیات در قضیه زیر آمده است.

قضیه ۱۰.۳. *فرض کنیم $R = [a,b] \times [c,d]$ و نقاط درونیابی $\{(x_i, y_j)\}$ در شرایط (۶۷.۳) صدق کنند. گیریم $\partial^{n+۱} f/\partial y^{n+۱}$ و $\partial^{m+۱} f/\partial x^{m+۱}$ برای تمام $(x,y) \in R$ موجود و پیوسته باشند. آنگاه برای هر $(x,y) \in R$ داریم*

$$\begin{aligned} \big|f(x,y) - p_{m,n}(x,y)\big| \leqslant& \frac{|\pi_{m+۱}^x(x)|}{(m+۱)!} \max_{a \leqslant \xi \leqslant b} \left|\frac{\partial^{m+۱} f(\xi, y)}{\partial x^{m+۱}}\right| \\ &+ \lambda_m(x) \frac{|\pi_{n+۱}^y(y)|}{(n+۱)!} \max_{c \leqslant \eta \leqslant d} \left|\frac{\partial^{n+۱} f(x, \eta)}{\partial y^{n+۱}}\right|, \end{aligned} \tag{۷۱.۳}$$

که در آن $\pi_{m+۱}^x(x) = (x-x_۰)(x-x_۱)\cdots(x-x_m)$ و $\pi_{n+۱}^y(y) = (y-y_۰)(y-y_۱)\cdots(y-y_n)$، و همچنین

$$\lambda_m(x) = \sum_{j=۰}^{m} |\ell_j^x(x)|$$

تابع لبگ متناظر با متغیر x است.

برهان. اختلاف تابع f و درونیابش را بصورت زیر می‌نویسیم

$$\begin{aligned} f(x,y) - p_{m,n}(x,y) =& \Big[f(x,y) - \sum_{i=۰}^{m} f(x_i, y)\ell_i^x(x)\Big] \\ &+ \sum_{i=۰}^{m} \ell_i^x(x) \Big[f(x_i, y) - \sum_{j=۰}^{n} f(x_i, y_j)\ell_j^y(y)\Big]. \end{aligned}$$

از رابطه‌ی بالا کران می‌گیریم و دو بار از کران خطای درونیابی یک متغیره (۳.۳) استفاده می‌کنیم. □

مثال ۱۵.۳. ابتدا چندجمله‌ای درونیاب دو بعدی مبتنی بر مقادیر $f_{ij} = f(x_i, y_j)$ در جدول زیر را می‌یابیم

	$x_۰ = ۰$	$x_۱ = ۱$	$x_۲ = ۳$
$y_۰ = ۰$	۱	۲	-۱
$y_۱ = ۳$	۲	۳	۰

چندجمله‌ایهای لاگرانژ را هم برای متغیر x و هم برای متغیر y محاسبه می‌کنیم

$$\ell_۰^x(x) = \frac{۱}{۳}(x-۱)(x-۳), \quad \ell_۱^x(x) = -\frac{۱}{۲}x(x-۳), \quad \ell_۲^x(x) = \frac{۱}{۶}x(x-۱),$$
$$\ell_۰^y(y) = \frac{۱}{۳}(۳-y), \quad \ell_۱^y(y) = \frac{۱}{۳}y,$$

سپس طبق فرمول درونیابی (۶۹.۳) داریم

$$\begin{aligned} p_{۲,۱}(x,y) &= f_{۰۰}\ell_۰^x(x)\ell_۰^y(y) + f_{۱۰}\ell_۱^x(x)\ell_۰^y(y) + f_{۲۰}\ell_۲^x(x)\ell_۰^y(y) \\ &\quad + f_{۰۱}\ell_۰^x(x)\ell_۱^y(y) + f_{۱۱}\ell_۱^x(x)\ell_۱^y(y) + f_{۲۱}\ell_۲^x(x)\ell_۱^y(y) \\ &= \frac{۱}{۳}y - \frac{۵}{۶}x^۲ + \frac{۱۱}{۶}x + ۱. \end{aligned}$$

که یک چندجمله‌ای درجه دو نسبت به x و درجه یک نسبت به y است.

بعنوان یک مثال دیگر، می‌خواهیم درونیاب درجه $(m,n) = (۲,۳)$ روی نقاط $\boldsymbol{x} = (۰,۲,۳)$ و $\boldsymbol{y} = (-۲,۰,۱,۳)$ با مقادیری که آن‌ها را بجای جدول، در ماتریس

$$F = \begin{bmatrix} ۳ & ۲ & ۰ \\ ۱ & ۱ & ۴ \\ ۸ & ۳ & ۲ \\ ۴ & ۵ & ۱ \end{bmatrix}$$

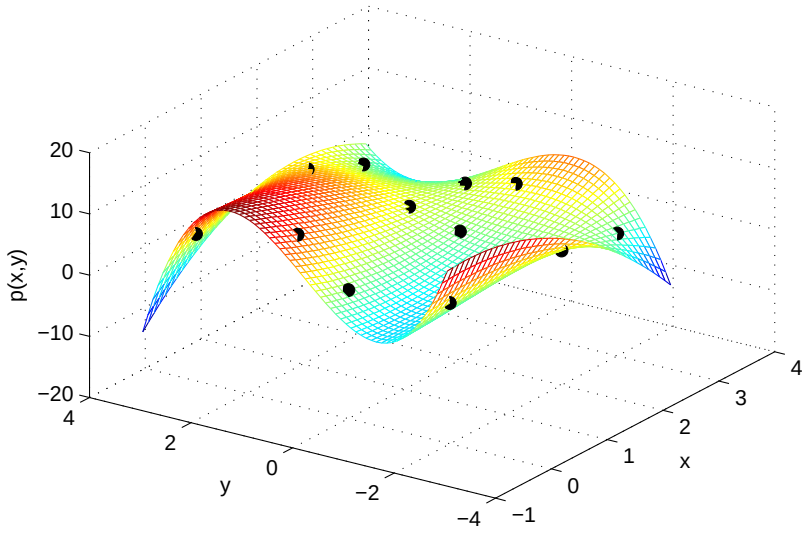
قرار داده‌ایم، محاسبه کنیم. این‌بار برنامه‌ای در متلب می‌نویسیم و نمودار چندجمله‌ای درونیاب را هم رسم می‌کنیم.

```
1  x = [0 2 3]; y = [-2 0 1 3];
2  F = [3 2 0; 1 1 4; 8 3 2; 4 5 1];
3  m = length(x); n = length(y);
4  s = x(1)-0.5:0.1:x(end)+0.5; t = y(1)-0.5:0.1:y(end)+0.5;
```

```
5  for k=1:m
6     L = 1;
7     for j=1:m
8         if j~=k L = L.*(s-x(j))/(x(k)-x(j)); end
9     end
10    Lx(k,:)=L;
11 end
12 for k=1:n
13    L = 1;
14    for j=1:n
15        if j~=k L = L.*(t-y(j))/(y(k)-y(j)); end
16    end
17    Ly(k,:)=L;
18 end
19 p = Ly'*F*Lx;
20 [S,T]=meshgrid(s,t);
21 mesh(S,T,p);
22 [X,Y]=meshgrid(x,y);
23 hold on
24 plot3(X(:),Y(:),F(:),'o')
```

در برنامه بالا در دو حلقه چندجمله‌ایهای لاگرانژ نسبت به متغیر x و y به طور مجزا محاسبه شده‌اند. چون هدف نهایی این بوده که چندجمله‌ای درونیاب $p_{۲,۳}(x,y)$ را بازای مقادیر برداری $x = \boldsymbol{s}$ و $y = \boldsymbol{t}$ محاسبه کنیم، از ابتدا چندجمله‌ایهای لاگرانژ ℓ_k^x را در بردار $\boldsymbol{s}$ و ℓ_k^y را در بردار $\boldsymbol{t}$ محاسبه کرده‌ایم. چندجمله‌ای درونیاب در معادله‌ی (۶۹.۳) که دارای دو سیگما است هم به کمک ضرب سه ماتریس در سطر 19 محاسبه شده است. دستورات meshgrid، mesh و plot3 برای رسم نمودارهای توابع دو متغیره به کار رفته‌اند. با اجرای برنامه بالا شکل ۲۰.۳ حاصل شده است. این شکل نمودار چندجمله‌ای درونیاب و همچنین مقادیر نقاط درونیابی را نشان می‌دهد. ♢

این روش درونیابی را می‌توان به حالت سه بعدی روی مکعب یا مکعب مستطیل تعمیم داد. اما درونیابی روی ناحیه‌های دیگر دارای پیچیدگی‌های خاص و نیازمند روش‌های دیگر است که در اینجا به آن‌ها نمی‌پردازیم.



شکل ۲۰.۳: درونیاب دو بعدی $p_{۲,۳}$ و مقادیر نقاط درونیابی

۱۰.۳ پرسش‌ها

۱. فرض کنیم نقاط $x_۰$، $x_۱$ و $x_۲$ به صورت هم‌فاصله با فاصله‌ی h روی محور x واقع شده‌اند. کران خطای درونیابیِ تابع مفروض $f \in C^۳[x_۰, x_۲]$ روی این نقاط را بدست آورید و آن را برای تابع $f(x) = \frac{۱}{۱+x}$ امتحان کنید.

۲. برای چندجمله‌ایهای لاگرانژ مبتنی بر نقاط متمایز $x_۰$، $x_۱$، ... و x_n نشان دهید

$$\sum_{k=۰}^{n} \ell_k(x) = ۱.$$

۳. فرض کنید $x_۰$، $x_۱$، ... و x_n نقاط متمایز و ℓ_j چندجمله‌ایهای لاگرانژ مبتنی بر این نقاط باشند. ثابت کنید

$$\ell_k'(x_k) = \sum_{\substack{i=۰ \\ i \neq k}}^{n} \frac{۱}{x_k - x_i}.$$

۴. ماتریس واندرموند را در نظر بگیرید

$$V_n = \begin{bmatrix} ۱ & x_۰ & x_۰^۲ & \dots & x_۰^n \\ ۱ & x_۱ & x_۱^۲ & \dots & x_۱^n \\ \vdots & \vdots & \vdots & \vdots & \\ ۱ & x_n & x_n^۲ & \dots & x_n^n \end{bmatrix}.$$

نشان دهید

$$\det(V_n) = \prod_{۰ \leqslant j < i \leqslant n} (x_i - x_j).$$

۵. گیریم x_0، x_1، ... و x_n نقاط متمایز و ℓ_j چندجمله‌ایهای لاگرانژ مبتنی بر این نقاط باشند. فرض کنید $\pi(x) = \prod_{j=0}^{n}(x - x_j)$. نشان دهید

$$\ell_j(x) = \frac{\pi(x)}{(x - x_j)\pi'(x_j)}, \quad x \neq x_j, \quad j = 0, 1, \ldots, n.$$

۶. گیریم x_0، x_1، ... و x_n نقاط حقیقی متمایز باشند، و مسئله درونیابی زیر را در نظر بگیرید. یک تابع

$$p_n(x) = \sum_{j=0}^{n} c_j e^{jx}$$

انتخاب می کنیم به طوری که $p_n(x_i) = f_i$ که f_i ها داده‌های معلومند. نشان دهید یک انتخاب یکتا برای مقادیر c_0، c_1، ... و c_n وجود دارد.

۷. فرض کنید $n, j = 0, 1, \ldots, n$ ، ℓ_j چندجمله‌ایهای لاگرانژ مبتنی بر نقاط x_j باشند. و نیز فرض کنید $c_i = \ell_i(0)$. نشان دهید

$$\sum_{i=0}^{n} c_i x_i^j = \begin{cases} 1, & j = 0, \\ 0, & j = 1, 2, \ldots, n, \\ (-1)^n x_0 x_1 \cdots x_n, & j = n + 1. \end{cases}$$

۸. فرض کنید چندجمله‌ای درجه n ام p_n تابع f را در نقاط متمایز x_0، x_1، ...، و x_n درونیابی کند. نشان دهید

$$\det \begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n & f(x_0) \\ 1 & x_1 & x_1^2 & \ldots & x_1^n & f(x_1) \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n & f(x_n) \\ 1 & x & x^2 & \ldots & x^n & p_n(x) \end{bmatrix} = 0.$$

۹. درجه‌ی چندجمله‌ای درونیاب مبتنی بر نقاط $(-1, 1)$، $(0, 1)$، $(1, 3)$ و $(2, 6)$ چند است؟

۱۰. نشان دهید اگر f یک چندجمله‌ای درجه k باشد، به ازای هر $n > k$ داریم $f[x_0, x_1, \ldots, x_n] = 0$.

۱۱. فرض کنید f یک چندجمله‌ای درجه ۳ باشد. نشان دهید

$$f[x_0, x_1, x_2] = \frac{1}{2} f''\left(\frac{x_0 + x_1 + x_2}{3}\right),$$

که x_0، x_1 و x_2 نقاط متمایزند.

۱۲. فرض کنید $f(x)$ یک تابع مفروض و $x_۰$، $x_۱$، ... و x_n نقاط متمایز باشند. تعریف می کنیم

$$g(x) = f[x_۰, x_۱, \ldots, x_n, x],$$

نشان دهید $g'(x) = f[x_۰, x_۱, \ldots, x_n, x, x]$.

۱۳. برای تابع f تعریف شده روی نقاط متمایز $x_۰$، $x_۱$، ... و x_n نشان دهید

$$f[x_۰, \ldots, x_n] = \sum_{j=۰}^{n} f(x_j) \left(\prod_{i=۰, i\neq j}^{n} (x_j - x_i) \right)^{-۱}.$$

۱۴. فرض کنید $f(x) = \frac{۱}{x}$ ثابت کنید

$$f[x_۰, x_۱, \ldots, x_n] = (-۱)^n \prod_{j=۰}^{n} \frac{۱}{x_j}, \quad x_j \neq ۰.$$

۱۵. ثابت کنید

$$f[x_۰, x_۱, \ldots, x_n] = \frac{\det \begin{bmatrix} ۱ & x_۰ & x_۰^۲ & \ldots & x_۰^{n-۱} & f(x_۰) \\ ۱ & x_۱ & x_۱^۲ & \ldots & x_۱^{n-۱} & f(x_۱) \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ ۱ & x_n & x_n^۲ & \ldots & x_n^{n-۱} & f(x_n) \end{bmatrix}}{\det \begin{bmatrix} ۱ & x_۰ & x_۰^۲ & \ldots & x_۰^{n-۱} & x_۰^n \\ ۱ & x_۱ & x_۱^۲ & \ldots & x_۱^{n-۱} & x_۱^n \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ ۱ & x_n & x_n^۲ & \ldots & x_n^{n-۱} & x_n^n \end{bmatrix}}$$

۱۶. نشان دهید اگر $f \in C^m[a, b]$، $m \geqslant ۰$، و $x_j \in [a, b]$ آنگاه وجود دارد $\xi \in [a, b]$ بطوریکه

$$f[x_۰, x_۱, \ldots, x_m] = \frac{۱}{m!} f^{(m)}(\xi).$$

برای $m = ۱$ این همان قضیه مقدار میانگین است. بنابراین رابطه بالا تعمیمی از قضیه مقدار میانگین به کمک تفاضلات نیوتن است. *راهنمایی:* جملات خطای درونیابی لاگرانژ و نیوتن را مقایسه کنید.

۱۷. هزینه‌ی محاسباتی روش نویل-ایتکن برای تولید چندجمله‌ای درونیاب را محاسبه کنید و آن را با روش نیوتن مقایسه کنید.

۱۸. نشان دهید در حالتی که نقاط $n+1$ نقطه درونیابی بصورت هم‌فاصله با $h = x_j - x_{j-1}$ قرار دارند، ضرایب تکیه‌گاه در فرمول گرانیگاهی به صورت زیر بدست می‌آیند

$$\beta_j = \frac{(-1)^n}{h^n n!}(-1)^j \binom{n}{j}, \quad j = 0, 1, \ldots, n.$$

۱۹. یک برنامه‌ی متلب برای روش درونیابی گرانیگایی روی نقاط دلخواه بنویسید.

۲۰. به کمک اتحاد مثلثاتی $\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta$ رابطه‌ی بازگشتی

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n = 1, 2, \ldots, \quad x \in [-1, 1],$$

برای چندجمله‌ایهای چبیشف را اثبات کنید.

۲۱. الف: نشان دهید اگر $\pi_{n+1}(x) = \prod_{k=0}^{n}(x - x_k)$ آنگاه $\pi'_{n+1}(x_j) = \prod_{k=0, k\neq j}^{n}(x_j - x_k)$ و به کمک آن فرمول جدیدی برای ضرایب تکیه‌گاه β_j در درونیابی گرانیگاهی بدست آورید.

ب: فرض کنید نقاط درونیابی، ریشه‌های چندجمله‌ای چبیشف T_{n+1} روی $[-1, 1]$ باشند (تعریف ۱.۳ و نقاط (۳۴.۳) را ببینید). با توجه به قسمت الف ثابت کنید ضرایب تکیه‌گاه β_j در درونیابی گرانیگاهی روی این نقاط با فرمول صریح

$$\beta_j = \frac{2^n}{n+1}(-1)^j \sin\frac{(2j+1)\pi}{2(n+1)}, \quad j = 0, 1, \ldots, n$$

بدست می‌آیند. ضرایب β_j^* را هم بنویسید.

ج: یک برنامه‌ی متلب برای روش گرانیگاهی روی نقاط چبیشف بنویسید و تابع مثال رونگه را روی این نقاط بازای $n = 10$ درونیابی کنید و شکل ۸.۳ این فصل را دوباره تولید کنید.

۲۲. درونیاب ارمیت روی نقاط جدول زیر را تعیین کنید.

x_k	0	1	3
f_k	0	2	1
f'_k	1	0	0

۲۳. قضیه ۷.۳ (کران خطای درونیابی ارمیت) را اثبات کنید.

۲۴. مقادیر a و b و c را به گونه ای بیابید که تابع

$$s(x) = \begin{cases} x^3, & x \in [0, 1], \\ \frac{1}{2}(x-1)^3 + a(x-1)^2 + b(x-1) + c, & x \in [1, 3], \end{cases}$$

یک اسپلاین مکعبی روی $X = \{۰, ۱, ۳\}$ باشد. آیا این تابع یک اسپلاین مکعبی طبیعی نیز هست؟

۲۵. فرض کنید $x \in [-۱, ۰]$ و $s_۱(x) = ۱ + c(x+۱)^۳$ که c یک پارامتر حقیقی است. چندجمله‌ای $s_۲$ روی $[۰, ۱]$ را بگونه‌ای بیابید که $s(x) = \begin{cases} s_۱(x), & -۱ \leqslant x \leqslant ۰, \\ s_۲(x), & ۰ \leqslant x \leqslant ۱ \end{cases}$ یک اسپلاین مکعبی طبیعی روی $[-۱, ۱]$ با $X = \{-۱, ۰, ۱\}$ باشد.

۲۶. چندجمله‌ای $p \in \mathbb{P}_۳$ را بگونه‌ای بیابید که برای $s(x) = \begin{cases} p(x), & ۰ \leqslant x \leqslant ۱, \\ (۲-x)^۳, & ۱ \leqslant x \leqslant ۲ \end{cases}$ داشته باشیم $s(۰) = ۰$ و s یک اسپلاین مکعبی باشد. آیا s یک اسپلاین مکعبی طبیعی نیز هست؟

۲۷. برای تابع $f(x) = \sin(۲\pi x)$ اسپلاین درونیاب مکعبی روی $[۰, ۱]$ با شرایط مرزی متناوب در حالت‌های $n = ۴, ۸$ را بدست آورید.

۲۸. اسپلاین مکعبی متناوب بر بازه‌ی $[-۱, ۱]$ و متناظر داده‌های جدول زیر را بدست آورید.

x_k	-۱	۰	۱
f_k	۱	۲	۱

۲۹. الف: برای دو تابع f و s در $C^۲[a,b]$، نشان دهید

$$\int_a^b [f''(x)]^۲ dx - \int_a^b [s''(x)]^۲ dx = \int_a^b [f''(x) - s''(x)]^۲ dx + ۲\int_a^b s''(x)[f''(x) - s''(x)]dx.$$

ب: اگر s اسپلاین مکعبی درونیاب تابع f روی $X = \{a = x_۰ < x_۱ < \cdots < x_n = b\}$ باشد، نشان دهید

$$\int_a^b s''(x)[f''(x) - s''(x)]dx = s''(b)[f'(b) - s'(b)] - s''(a)[f'(a) - s'(a)].$$

نشان دهید با هریک از شرایط مرزی سه گانه طبیعی، ارمیتی یا متناوب سمت راست معادلعه‌ی بالا برابر صفر است و طبق قسمت الف نتیجه بگیرید

$$\int_a^b [s''(x)]^۲ dx \leqslant \int_a^b [f''(x)]^۲ dx.$$

ج: فرض کنید $\mathcal{I}$ مجموعه‌ی همه‌ی $g \in C^۲[a,b]$ است که تابع f را با شرایط مشابه s درونیابی می‌کنند. با توجه به نتیجه نهایی قسمت ب نشان دهید برای هر $g \in \mathcal{I}$ داریم

$$\int_a^b [s''(x)]^۲ dx \leqslant \int_a^b [g''(x)]^۲ dx.$$

به عبارت دیگر در بین تمامی اعضای $\mathcal{I}$، (در حالتی که شرط مرزی متناوب است، در بین تمام اعضای متناوب $\mathcal{I}$) تابع درونیاب اسپلاین مکعبی دارای کمترین انحنا در روی بازه $[a,b]$ است. راهنمایی: فرمول انحنای f در نقطه‌ی x عبارتست از

$$\kappa(x) = \frac{f''(x)}{\left(۱ + [f'(x)]^۲\right)^{۳/۲}},$$

و انحنای کل f برابر انتگرال مربع $\kappa(x)$ است. با چشم‌پوشی از رفتار $f'(x)$، انحنای کل را با انتگرال مربع f'' تخمین بزنید.

۳۰. ثابت کنید اگر در مسئله‌ی برازش منحنی درجه‌ی m روی نقاط $(x_۰, f_۰)$، $(x_۱, f_۱)$، ...، (x_n, f_n) قرار دهیم $m = n$ آنگاه تقریب چندجمله‌ای همان چندجمله‌ای درونیاب خواهد بود.

۳۱. فرض کنید به دنبال یک برازش منحنی به فرم نمایی $y = ae^{bx}$ برای داده‌های $(x_۰, f_۰)$، $(x_۱, f_۱)$، ...، (x_n, f_n) باشیم. در این صورت باید ضرایب a و b را طوری بدست آوریم که نرم دوم خطا مینیمم شود. اگر همان روند برازش چندجمله‌ای را تکرار کنیم به یک دستگاه معادلات غیر خطی می‌رسیم. اما در این پرسش به طریق دیگری عمل کنید و با تغییر متغیر $Y = \ln y$ این مسئله را به یک مسئله‌ی کمترین مربعات درجه یک تبدیل و حل کنید. کد مت‌لب برای محاسبه‌ی برازش نمایی را بنویسید.

۳۲. برازش نمایی برای داده‌های جدول زیر را بدست آورید

x_k	۰	۱	۲	۳	۴
f_k	۱	۳	۶	۱۴	۲۰

۳۳. با تغییر متغیر $Y = ۱/y$ و $X = x^۲$ روشی برای برازش منحنی با الگوی $y = ۱/(a + bx^۲)$ برای داده‌های $(x_۰, f_۰)$، $(x_۱, f_۱)$، ...، (x_n, f_n) ارائه دهید.

فصل ۴

مشتق‌گیری عددی

مشتق‌گیری عددی یکی از ابزارهای لازم در طراحی برخی از روش‌های عددی مانند روش‌های حل عددی معادلات دیفرانسیل است. برای راضی کردن شما برای مطالعه‌ی دقیق این فصل چند مورد از دلایلی که ما را مجبور به استفاده از مشتق‌گیری عددی می‌کنند، ذکر می‌کنیم. نخست اینکه گاهی ضابطه‌ی تابع بسیار پیچیده و مشتق‌گیری از آن مشکل است، در این صورت می‌توان مشتق را به صورت عددی محاسبه کرد. دوم و مهمتر اینکه در عمل معمولاً فرم بسته‌ای از تابع در دست نیست و تنها مقادیری از آن در تعداد متناهی نقطه در اختیار است که در این صورت مشتق‌گیری تحلیلی غیرممکن است. در معادلات دیفرانسیل ضابطه‌ی تابع به صورت صریح در دست نیست و یک راه برای یافتن تقریبی از آن استفاده از فرمول‌های مشتق‌گیری عددی و تبدیل معادله دیفرانسیل به یک معادله جبری است.

فرمول‌های مشتق‌گیری عددی معمولاً به شکل زیر هستند

$$f^{(\ell)}(x_j) = \sum_k \lambda_k f(x_k) + E(f,h),$$

که در آن مقدار مشتق مرتبه‌ی ℓ-ام تابع بر حسب مقادیر تابع در نقاط x_k نوشته شده است. این نقاط در همسایگی x_j قرار دارند و اگر فرض کنیم h متری برای اندازه‌گیری چگالی این نقاط باشد، $E_n(f,h)$ خطای مشتق‌گیری عددی است.

۱.۴ استخراج فرمول‌ها به کمک بسط تیلر

پیش از هر چیز یادآوری می‌کنیم، اگر f در یک همسایگی از نقطه‌ی $x_۰$ تابعی $C^{m+۱}$، باشد، در آن همسایگی یک بسط تیلر مرتبه m حول $x_۰$ برای f وجود دارد. برای ثابت $h > ۰$ داریم

$$f(x_۰ \pm h) = f(x_۰) \pm hf'(x_۰) + \frac{h^۲}{۲!}f''(x_۰) + \cdots + (\pm ۱)^m \frac{h^m}{m!}f^{(m)}(x_۰) + (\pm ۱)^{m+۱}\frac{h^{m+۱}}{(m+۱)!}f^{(m+۱)}(\xi),$$

که در آن ξ مجهولی بین x_0 و $x_0 \pm h$ است.

با فرض $m = 1$ و علامت $+$ برای h داریم

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(\xi),$$

که نتیجه می‌دهد

$$f'(x_0) = \frac{f(x_1) - f(x_0)}{h} - \frac{h}{2!}f''(\xi) =: F_h^1 + E(f,h), \quad \xi \in [x_0, x_1], \tag{۱.۴}$$

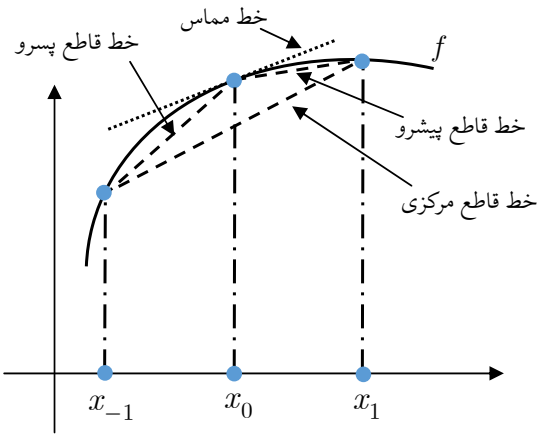
که در آن $x_1 = x_0 + h$. در اینجا جمله‌ی خطا با

$$E(f,h) = -\frac{h}{2!}f''(\xi),$$

نشان داده شده است که با فرض $f \in C^2[x_0, x_1]$، نشان می‌دهد این فرمول از $\mathcal{O}(h)$ است. بالا اندیس ۱ در F_h^1 نشان‌دهنده‌ی مرتبه‌ی فرمول است. در این فرمول (از دید نظری) اگر h را مثلاً نصف کنیم، خطا هم نصف مقدار قبل می‌شود، اگر h را یک‌سوم کنیم خطا هم یک‌سوم مقدار قبل می‌شود. به $E(f,h)$ خطای برشی هم می‌گوییم زیرا با برش جمله‌های بسط تیلر حاصل شده است. به فرمول (۱.۴) فرمول *تفاضلی پیشرو* برای مشتق مرتبه اول می‌گوییم. به همین ترتیب می‌توان فرمول *تفاضلی پسرو* را به کمک بسط تیلر به صورت زیر بدست آورد

$$f'(x_0) = \frac{f(x_0) - f(x_{-1})}{h} + \frac{h}{2!}f''(\eta) =: B_h^1 + E(f,h), \quad \eta \in [x_{-1}, x_0], \tag{۲.۴}$$

که در آن $x_{-1} = x_0 - h$. شکل ۱.۴ تصویری نمادین از این دو فرمول و فرمولی که بعداً ارائه می‌دهیم، را نشان می‌دهد. مقدار $f'(x_0)$ شیب خط مماس در نقطه‌ی x_0 است و تقریب تفاضلی پیشرو، شیب خط قاطع از مقادیر $f(x_0)$ و $f(x_1)$



شکل ۱.۴: تصویر نمادین فرمول‌های پیشرو، پسرو و مرکزی

است، در حالیکه تقریب تفاضلی پسرو، شیب خط قاطع از مقادیر $f(x_0)$ و $f(x_{-1})$ است. هرچه h کوچکتر باشد، شیب خط مماس و خطوط قاطع به هم نزدیک‌تر است و فرمول خطای $E(f,h)$ سرعت میل کردن این شیب‌ها به هم را بر حسب سرعت میل کردن h به صفر تعیین می‌کند.

مثال ۱.۴. مشتق تابع $f(x) = e^x$ در نقطه‌ی صفر را به کمک فرمول تفاضلی پیشرو (۱.۴) با مقادیر مختلف h محاسبه می‌کنیم. می‌دانیم جواب دقیق ۱ است. فرض کنیم $h = ۰/۱$، داریم

$$f'(۰) \approx F^{۱}_{۰/۱} = \frac{۱}{۰/۱}[e^{۰/۱} - e^{۰}] \doteq ۱/۰۵۱۷$$

که در آن $\doteq$ نشان‌دهنده‌ی تساوی طرفین تا چهار رقم اعشار است. بنابراین خطای این تقریب در حدود ۰/۰۵۱۷ است. اگر طول گام h را نصف کنیم تقریب بهتری طبق زیر خواهیم داشت

$$f'(۰) \approx F^{۱}_{۰/۰۵} = \frac{۱}{۰/۰۵}[e^{۰/۰۵} - e^{۰}] \doteq ۱/۰۲۵۴.$$

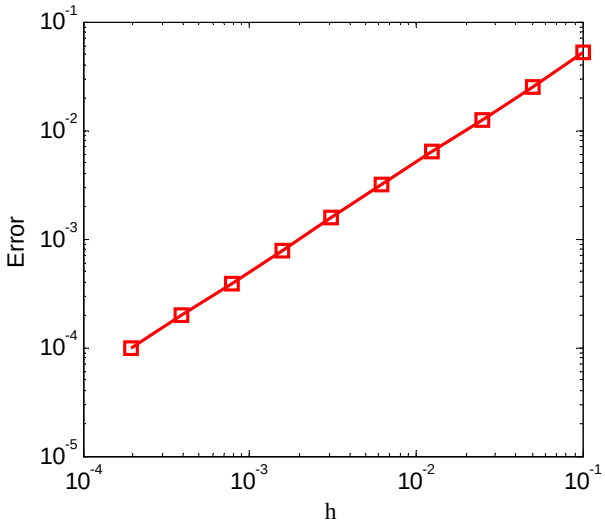
تقریب بالا دارای خطای ۰/۰۲۵۴ است که تقریباً نصف خطای اول است. یعنی با نصف شدن h خطا هم تقریباً نصف شده است. این معنیِ $\mathcal{O}(h)$ است. ادامه‌ی محاسبات به صورت دستی آزار دهنده است و بهتر است برنامه‌ای در متلب بنویسیم که این کار را برایمان انجام دهد.

```
f = @(x) exp(x);
h(1)=0.1; K=10;
for k=1:K
   F =(f(h(k))-f(0))/h(k);
   err(k)=abs(F-f(0));
   h(k+1)=h(k)/2;
end
loglog(h(1:end-1),err)
xlabel('h'); ylabel('Error');
```

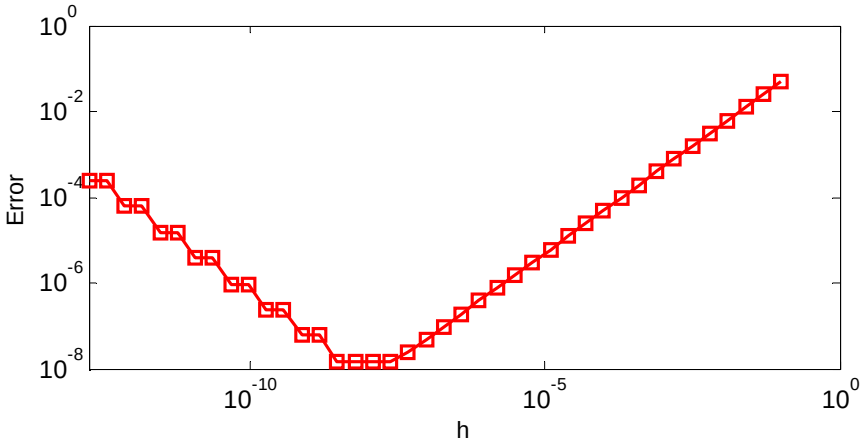
در این برنامه h ابتدا با مقدار ۰/۱ شروع و هر بار در درون حلقه نصف می‌شود تا تقریب بهتری بدست آید. حاصل اجرای این برنامه نمودار ۲.۴ است.

این نمودار در مختصات log-log ترسیم شده است، یعنی h روی محور افقی و خطاها روی محور عمودی به صورت لگاریتمی مقیاس شده‌اند. شیب این نمودار مرتبه‌ی خطا را (که در اینجا ۱ است)، تعیین می‌کند. (علت آن چیست؟)

در عمل بخاطر خطاهای محاسباتی، با کاهش h خطا به صفر میل نمی‌کند. در ادامه علت آن را دقیقاً تشریح خواهیم کرد. اما در اینجا می‌خواهیم این پدیده را به صورت عددی مشاهده کنیم. برای همین منظور در برنامه‌ی بالا متغیر K را برابر ۴۰ قرار می‌دهیم و برنامه را دوباره اجرا می‌کنیم. حاصل، شکل ۳.۴ است. می‌بینیم که برای مقادیر h کوچکتر از حدود



شکل ۲.۴: خطای مشتق‌گیری عددی در فرمول پیشرو مرتبه اول



شکل ۳.۴: ناپایداری در مشتق‌گیری عددی در فرمول پیشرو مرتبه اول

$۱۰^{-۸}$ ، نمودار خطا بجای اینکه کاهش یابد، افزایش می‌یابد و به یک نمودار V شکل تبدیل می‌شود. در روش‌های عددی گاهی با این نمودارها که نشان‌دهنده‌ی ناپایداری عددی هستند، مواجهیم و بایستی برای رفع آن‌ها چاره‌ای باندیشیم. ♢

در اینجا علت بروز ناپایداری عددی برای h های خیلی کوچک را توضیح می‌دهیم. همانگونه که گفته شد این ناپایداری به دلیل خطاهای گردکردن است. در محاسبه‌ی مشتق تفاضلی پیشرو، خطاهای محاسباتی هم در محاسبه‌ی مقادیر h، $f(x_۰)$ و $f(x_۱)$ و هم در عمل تفریق و تقسیم رخ می‌دهند. فرض کنیم مقادیر تابع با خطای نسبیِ u (واحد گرد کردن) محاسبه شوند، یعنی

$$\widehat{f_k} = fl(f_k) = f_k(۱ + \varepsilon_k), \quad |\varepsilon_k| \leqslant u, \quad k = ۰, ۱,$$

که در آن فرض کرده‌ایم $f_k = f(x_k)$. پس می‌توان مقدار محاسبه شده‌ی $\widehat{F}_h^{۱}$ را به صورت زیر نوشت

$$\begin{aligned}\widehat{F}_h^{۱} &= fl\left(\frac{fl(\widehat{f}_{۱} - \widehat{f}_{۰})}{fl(h)}\right)\\ &= \frac{\big(f_{۱}(۱+\varepsilon_{۱}) - f_{۰}(۱+\varepsilon_{۰})\big)(۱+\varepsilon_{۲})}{h/(۱+\varepsilon_{۴})}(۱+\varepsilon_{۳})\\ &= \frac{f_{۱}(۱+\theta_{۴}) - f_{۰}(۱+\theta'_{۴})}{h}\\ &= F_h^{۱} + \frac{f_{۱}\theta_{۴} - f_{۰}\theta'_{۴}}{h},\end{aligned}$$

که در آن $|\varepsilon_k| \leqslant u$ برای $k = ۰, ۱, ۲, ۳, ۴$ و از این رو $|\theta_{۴}| \leqslant \gamma_{۴} = ۴u/(۱-۴u)$ و $|\theta'_{۴}| \leqslant \gamma_{۴}$. در مخرج سطر دوم از رابطه‌ی $fl(h) = h/(۱+\varepsilon_{۴})$ استفاده کرده‌ایم که در پرسش ۵ فصل ۲ آن را اثبات کرده‌اید. حال می‌توان نوشت

$$|\widehat{F}_h^{۱} - F_h^{۱}| \leqslant \frac{۲\gamma_{۴}}{h}\max_{x\in[x_{۰},x_{۱}]}|f(x)| = \frac{۲\gamma_{۴}M_{۰}}{h},$$

که $M_{۰} = \max\limits_{[x_{۰},x_{۱}]}|f(x)|$. با توجه به اینکه $E(f,h) = -\frac{h}{۲}f''(\xi)$، اختلاف بین مقدار دقیق $f'(x_{۰})$ و مقدار محاسبه شده‌ی $\widehat{F}_h^{۱}$ به صورت زیر بدست می‌آید

$$|f'(x_{۰}) - \widehat{F}_h^{۱}| \leqslant |f'(x_{۰}) - F_h^{۱}| + |F_h^{۱} - \widehat{F}_h^{۱}| \leqslant \frac{hM_{۲}}{۲} + \frac{۲M_{۰}\gamma_{۴}}{h} =: e(h)$$

که در آن $M_{۲} = \max\limits_{[x_{۰},x_{۱}]}|f''(x)|$. خطای $e(h)$ شامل دو جمله است. جمله‌ی اول که کران خطای برشی است با کاهش h کاهش می‌یابد، اما جمله‌ی دوم که از خطای محاسبات ناشی می‌شود، با کاهش h افزایش می‌یابد. سؤال این است که در چه طول گامی e مینیمم خود را اختیار می‌کند. با مشتق‌گیری از e می‌توان نشان داد مینیمم آن در

$$h = h^* = ۲\sqrt{\frac{\gamma_{۴}M_{۰}}{M_{۲}}}$$

رخ می‌دهد. به h^* *طول گام بهینه* می‌گوییم. برای مقادیر بیشتر از مقدار بهینه، تابع e نزولی و برای مقادیر کمتر از آن صعودی است. در مثال ۱.۴، محاسبات در دقت دو برابر یعنی با $u \approx ۱۰^{-۱۶}$ انجام شده‌اند، بنابراین $\gamma_{۴} \approx ۴ \times ۱۰^{-۱۶}$. همچنین برای همسایگی‌های نزدیک صفر مقدار $M_{۰}$ و $M_{۲}$ تقریباً برابر ۱ است. بنابراین طول گام بهینه عبارت است از

$$h^* \approx ۲\sqrt{۴ \times ۱۰^{-۱۶}} = ۴ \times ۱۰^{-۸}$$

که نتایج عددی در شکل ۳.۴ نیز آن را تصدیق می‌کنند.

طبق شکل ۳.۴، با دقت دوبرابر نمی‌توان خطایی بهتر از $۱۰^{-۸}$ با فرمول پیشرو مرتبه اول بدست آورد. بنابراین باید به دنبال ساختن فرمول‌های با مرتبه بالاتر بود. یکی از این فرمول‌ها، فرمول *تفاضل مرکزی* است که به صورت زیر بدست

می‌آید. فرض $f \in C^۳[x_۰ - h, x_۰ + h]$ و بسط‌های تیلر زیر را در نظر بگیرید

$$f(x_۰ + h) = f(x_۰) + hf'(x_۰) + \frac{h^۲}{۲}f''(x_۰) + \frac{h^۳}{۳!}f'''(\xi), \quad \xi \in [x_۰, x_۰ + h],$$
$$f(x_۰ - h) = f(x_۰) - hf'(x_۰) + \frac{h^۲}{۲}f''(x_۰) - \frac{h^۳}{۳!}f'''(\eta), \quad \eta \in [x_۰, x_۰ + h].$$

با کم کردن دو معادله‌ی بالا و تقسیم کردن بر $۲h$ به فرمول زیر می‌رسیم

$$f'(x_۰) = \frac{f(x_۰ + h) - f(x_۰ - h)}{۲h} + E(f, h), \tag{۳.۴}$$

که در آن $E(f, h)$ به صورت زیر است

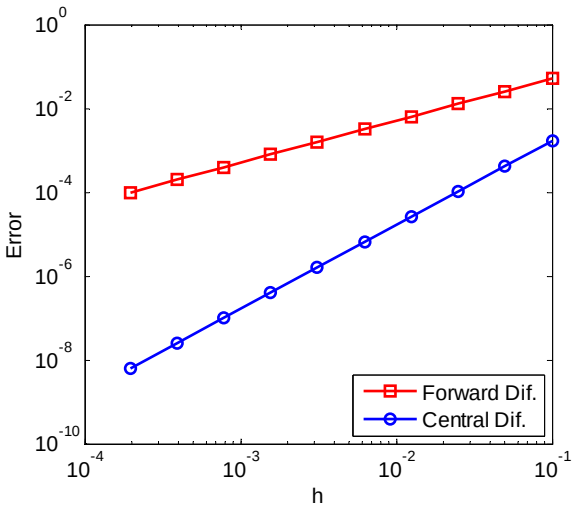
$$E(f, h) = -\frac{h^۲}{۶}\frac{f'''(\xi) + f'''(\eta)}{۲} = -\frac{h^۲}{۶}f'''(\zeta), \quad \zeta \in [x_۰ - h, x_۰ + h].$$

تساوی دوم به خاطر پیوستگی f''' و استفاده از قضیه‌ی مقدار میانی برقرار است (اثبات کنید!). بنابراین فرمول تفاضل مرکزی به شرط هموار f، از مرتبه‌ی $h^۲$ است. فرمول تفاضل مرکزی را با C_h نشان می‌دهیم. این فرمول شیب خط قاطعِ مقادیر $f(x_۱)$ و $f(x_{-۱})$ است که در شکل ۱.۴ نشان داده شده است. همانطور که در این شکل می‌بینید شیب خط قاطع مرکزی، تقریب بهتری برای شیب خط مماس ارائه می‌دهد. اگر برنامه‌ی متلب این فرمول را نوشته و با فرمول تفاضل پیشرو مقایسه کنیم، نمودار خطای شکل ۴.۴ بدست می‌آید که در آن h را K=10 بار نصف کرده‌ایم. مشاهده می‌کنیم که شیب نمودار خطای فرمول مرکزی دو است که همان مرتبه‌ی خطاست. اگر قرار دهیم K=40، شکل ۵.۴ بدست می‌آید که نشان می‌دهد هر دو فرمول از ناپایداری عددی در h های خیلی کوچک رنج می‌برند. مقدار طول گام بهینه برای فرمول پیشرو را بدست آوردیم، شما هم به عنوان تمرین طول گام بهینه برای فرمول تفاضل مرکزی را بدست آورید و با نتایج عددی شکل ۵.۴ مقایسه کنید (پرسش ۱ را ببینید). طبق این شکل، با فرمول مرکزی خطایی بهتر از حدود $۱۰^{-۱۲}$ در دقت دوبرابر قابل حصول نیست.

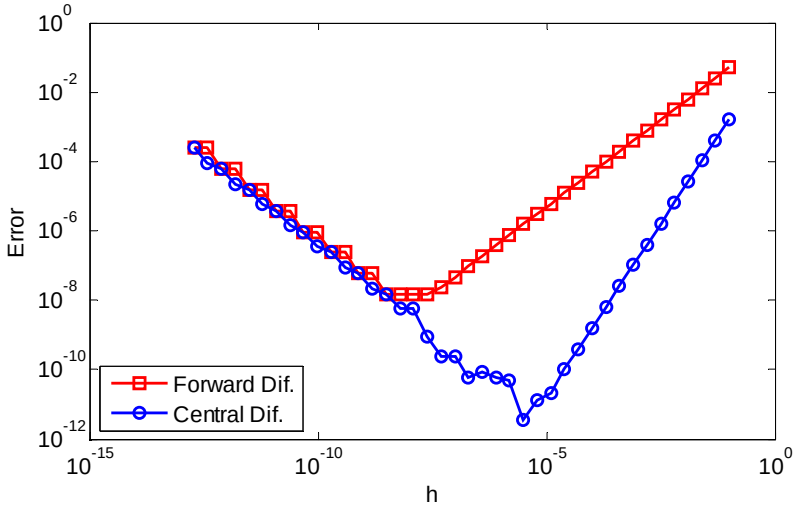
فرمول‌های با مراتب بالاتر را می‌توان با نوشتن بسط‌های تیلر از درجه بالاتر بدست آورد. مثلاً اگر بخواهیم فرمولی برای $f'(x_۰)$ بنویسیم که از مقادیر $f(x_۰)$، $f(x_۰ + h)$ و $f(x_۰ + ۲h)$ استفاده کند، باید برای هر یک از این مقادیر یک بسط تیلر حول $x_۰$ بنویسیم و با کم و زیاد کردن آن‌ها به فرمولی برای $f'(x_۰)$ برسیم. اگر این کار را انجام دهیم، فرمول زیر حاصل می‌شود

$$f'(x_۰) = \frac{-۳f(x_۰) + ۴f(x_۱) - f(x_۲)}{۲h} + \frac{h^۲}{۳}f'''(\xi), \quad \xi \in [x_۰, x_۰ + ۲h], \tag{۴.۴}$$

که در آن $x_۱ = x_۰ + h$ و $x_۲ = x_۰ + ۲h$. این یک فرمولِ تفاضل پیشروی سه نقطه‌ای است.



شکل ۴.۴: مشتق‌گیری عددی با فرمول‌های تفاضل پیشرو و تفاضل مرکزی



شکل ۵.۴: ناپایداری در مشتق‌گیری عددی با فرمول‌های تفاضل پیشرو و تفاضل مرکزی

برونیابی ریچاردسون

فرمول (۴.۴) را می‌توان به صورت دیگری نیز بدست آورد. فرض کنیم f به اندازه‌ی کافی در همسایگی x_0 مشتق‌پذیر باشد. اگر جملات بیشتری از بسط تیلر بنویسیم، فرمول تفاضل پیشرو به صورت زیر است

$$f'(x_0) = \underbrace{\frac{f(x_0+h)-f(x_0)}{h}}_{F_h^1} + \alpha_1 h + \alpha_2 h^2 + \alpha_3 h^3 + \cdots, \qquad (۵.۴)$$

که در آن $\alpha_k = -\frac{h^k}{(k+1)!} f^{(k+1)}(x_0)$. در فرمول بالا یک *بسط مجانبی* برای خطا نوشته‌ایم. همانگونه که قبلاً دیده‌ایم و در بسط مجانبی بالا هم مشخص است، فرمول پیشرو از $\mathcal{O}(h)$ است. اگر در (۵.۴) بجای h مقدار $2h$ را جایگزین کنیم

به فرمول زیر می‌رسیم

$$f'(x_۰) = \underbrace{\frac{f(x_۰ + ۲h) - f(x_۰)}{۲h}}_{F^{۱}_{۲h}} + ۲\alpha_۱ h + ۴\alpha_۲ h^۲ + ۸\alpha_۳ h^۳ + \cdots, \tag{۶.۴}$$

که هنوز از $\mathcal{O}(h)$ است. برای بدست آوردن فرمول دقیق‌تر می‌توان بین دو فرمول بالا ضریب h را در بسط مجانبی خطا حذف کرد. کافی است معادله‌ی (۶.۴) را از ۲ برابر معادله‌ی (۵.۴) کم کنیم، که نتیجه می‌دهد

$$f'(x_۰) = \underbrace{\frac{-۳f(x_۰) + ۴f(x_۰ + h) - f(x_۰ + ۲h)}{۲h}}_{F^{۲}_{h} = ۲F^{۱}_{h} - F^{۱}_{۲h}} - ۲\alpha_۲ h^۲ - ۶\alpha_۳ h^۳ - \cdots,$$

که نشان می‌دهد فرمول جدید، یعنی $F^{۲}_{h}$، از $\mathcal{O}(h^۲)$ است. می‌توان این روش را برای بدست آوردن فرمول‌های دقیق‌تر باز هم ادامه داد. به این روش *برونیابی ریچاردسون* می‌گویند. در مثال ۱.۴ دیدیم که برای محاسبه‌ی مشتق تابع e^x در صفر، $F^{۱}_{۰/۱} = ۱/۰۵۱۷$ دارای خطای ۰/۰۵۱۷ و $F^{۱}_{۰/۰۵} = ۱/۰۲۵۴$ دارای خطای ۰/۰۲۵۴ هستند. تقریب جدید به کمک فرمول ریچاردسون به صورت زیر محاسبه می‌شود

$$F^{۲}_{۰/۰۵} = ۲F^{۱}_{۰/۰۵} - F^{۱}_{۰/۱} = ۰/۹۹۹۱,$$

که دارای خطای ۰/۰۰۰۹ است.

در پرسش ۲ از شما خواسته شده است که فرمول پسروی مرتبه دوم

$$f'(x_۰) = \frac{۳f(x_۰) - ۴f(x_{-۱}) + f(x_{-۲})}{۲h} + \mathcal{O}(h^۲), \tag{۷.۴}$$

را به کمک برونیابی ریچاردسون و فرمول پسرو مرتبه اول (۲.۴) بدست آورید. همچنین در پرسش ۳، استخراج فرمول مرکزی

$$f'(x_۰) = \frac{-f(x_۲) + ۸f(x_۱) - ۸f(x_{-۱}) + f(x_{-۲})}{۱۲h} + \mathcal{O}(h^۴) \tag{۸.۴}$$

به کمک برونیابی ریچاردسون و فرمول مرکزی (۳.۴) خواسته شده است.

لیست برخی از فرمول‌های مشتق‌گیری برای مشتق مرتبه اول به همراه جمله‌ای خطای هر یک در جدول ؟؟ آمده است. (بعداً اضافه شود ...)

مشتقات مراتب بالاتر

فرمول‌های تفاضلی برای مشتقات مراتب بالاتر نیز به طریق مشابه بدست می‌آیند. برای مثال یک فرمول تفاضل مرکزی برای $f''(x_۰)$ به صورت زیر بر حسب مقادیر $f(x_۰)$، $f(x_۰ + h)$ و $f(x_۰ - h)$ بدست می‌آید. ابتدا بسط‌های تیلر زیر را

می‌نویسیم

$$f(x_۰+h)=f(x_۰)+hf'(x_۰)+\frac{h^۲}{۲}f''(x_۰)+\frac{h^۳}{۳!}f'''(x_۰)+\frac{h^۴}{۴!}f^{(۴)}(\xi),\quad \xi\in[x_۰,x_۰+h],$$
$$f(x_۰-h)=f(x_۰)-hf'(x_۰)+\frac{h^۲}{۲}f''(x_۰)-\frac{h^۳}{۳!}f'''(x_۰)+\frac{h^۴}{۴!}f^{(۴)}(\eta),\quad \eta\in[x_۰,x_۰+h].$$

اگر دو معادله‌ی بالا را با هم جمع کنیم و $۲f(x_۰)$ را از آن کم و بر $h^۲$ تقسیم کنیم به فرمول تفاضل مرکزی زیر می‌رسیم

$$\begin{aligned}f''(x_۰)&=\frac{f(x_۰-h)-۲f(x_۰)+f(x_۰+h)}{h^۲}-\frac{h^۲}{۱۲}f^{(۴)}(\zeta)\\&=D_h^۲+E(f,h),\quad \zeta\in[x_۰-h,x_۰+h].\end{aligned}$$

یک راه دیگر برای تقریبِ مشتق مرتبه دوم، اِعمال دو عملگر مرتبه اول روی هم است. با توجه به اینکه f'' مشتق مرتبه اول f' است، می‌توان نوشت

$$\begin{aligned}f''(x_۰)&\approx F_h^۱[F_h^۱f(x_۰)]=F_h^۱[\frac{f(x_۱)-f(x_۰)}{h}]\\&=\frac{۱}{h}[F_h^۱f(x_۱)-F_h^۱f(x_۰)]=\frac{f(x_۲)-f(x_۱)}{h^۲}-\frac{f(x_۱)-f(x_۰)}{h^۲}\\&=\frac{f(x_۲)-۲f(x_۱)+f(x_۰)}{h^۲}\end{aligned}$$

لیستی از فرمول‌های مشتق‌گیری برای مشتق دوم در جدول ؟؟ آمده است. (بعداً اضافه شود ...)

۲.۴ استخراج فرمول‌ها به کمک چندجمله‌ای درونیاب

یکی دیگر از روش‌های بدست آوردن فرمول‌های مشتق‌گیری عددی، استفاده از فرمول درونیابی چندجمله‌ای است. حسن این روش در این است که می‌توان فرمول‌هایی روی نقاط غیر هم‌فاصله نیز بدست آورد. در فصل ۳ دیدیم که می‌توان تابع f را با درونیاب یکتای p_n روی $n+۱$ نقطه‌ی متمایز تقریب زد،

$$f(x)=p_n(x)+R_n(f,x).$$

می‌توان از این فرمول برای تقریب مشتق تابع نیز استفاده کرد و نوشت

$$f'(x)=p_n'(x)+R_n'(f,x).$$

در این صورت $p_n'(x)$ تقریبی برای $f'(x)$ خواهد بود و $E_n(f,x)=R_n'(f,x)$ نیز خطای مشتق‌گیری است.

فرض کنیم x_0، x_1، ...، x_n نقاط متمایز روی خط حقیقی باشند. فرم نیوتن درونیاب تابع مفروض f به صورت زیر نوشته می‌شود

$$p_n(x) = f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] + \\ \cdots + (x - x_0)\cdots(x - x_{n-1})f[x_0, \ldots, x_n].$$

و خطای درونیابی هم عبارت است از

$$R_n(f, x) = (x - x_0)(x - x_1)\cdots(x - x_n)f[x_0, \ldots, x_n, x].$$

با مشتق‌گیری از p_n و محاسبه‌ی آن در x_0 داریم

$$p_n'(x_0) = f[x_0, x_1] + (x_0 - x_1)f[x_0, x_1, x_2] + \\ \cdots + (x_0 - x_1)(x_0 - x_2)\cdots(x_0 - x_{n-1})f[x_0, \ldots, x_n]. \tag{۹.۴}$$

همچنین با مشتق‌گیری از جمله‌ی خطا و محاسبه‌ی آن در x_0 داریم

$$E(f) = R_n'(f, x_0) = (x_0 - x_1)(x_0 - x_2)\cdots(x_0 - x_n)f[x_0, \ldots, x_n, x_0]. \tag{۱۰.۴}$$

با انتخاب ترتیب مناسب برای نقاط درونیابی، فرمول‌های پیشرو، پسرو و مرکزی بدست می‌آیند.

فرمول‌های پیشرو

اگر نقاط به صورت $x_0 < x_1 < \cdots < x_n$ مرتب شده باشد، آنگاه (۹.۴) فرمول‌های تفاضل پیشرو برای تقریب $f'(x_0)$ ارائه می‌دهد. برای مثال در درونیابی خطی، یعنی برای $n = 1$، با فرض اینکه $x_1 - x_0 = h$ داریم

$$f'(x_0) \approx p_1'(x_0) = f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{h},$$
$$E(f) = (x_0 - x_1)f[x_0, x_1, x_0] = -h\frac{f''(\xi)}{2},$$

که همان فرمول تفاضل پیشرو (۱.۴) است که در بخش قبل به کمک بسط تیلر بدست آوردیم. جمله‌ی خطا با فرض اینکه $f \in C^2[x_0, x_1]$ و با توجه به پرسش ۱۶ فصل ۳ برای قضیه مقدار میانگین تعمیم‌یافته بدست آمده است.

اگر درونیابی درجه دو را در نظر بگیریم، یک فرمول پیشرو سه نقطه‌ای بدست خواهیم آورد. برای این منظور فرض می‌کنیم $x_1 - x_0 = h_0$ و $x_2 - x_1 = h_1$. بنابراین داریم

$$\begin{aligned} f'(x_0) &\approx p_2'(x_0) = f[x_0, x_1] + (x_0 - x_1)f[x_0, x_1, x_2] \\ &= \frac{f(x_1) - f(x_0)}{h_0} - h_0\frac{\big[f(x_2) - f(x_1)\big]/h_1 - \big[f(x_1) - f(x_0)\big]/h_0}{h_0 + h_1} \\ &= \frac{-(2h_0 + h_1)f(x_0) + (h_0 + h_1)(1 + \frac{h_0}{h_1})f(x_1) - \frac{h_0^2}{h_1}f(x_2)}{h_0(h_0 + h_1)}. \end{aligned}$$

اگر در فرمول بالا داشته باشیم $h_۰ = h_۱ = h$، به فرمول (۴.۴) می‌رسیم. بنابراین فرمول بالا تعمیمی از فرمول تفاضل پیشروی سه نقطه‌ای برای نقاط غیر هم‌فاصله است. برای تخمین خطای این فرمول داریم

$$E(f) = (x_۰ - x_۱)(x_۰ - x_۲)f[x_۰, x_۱, x_۲, x_۰] = h_۰(h_۰ + h_۱)\frac{f'''(\xi)}{۳!}, \qquad (۱۱.۴)$$

که این هم تعمیمی از خطای فرمول (۴.۴) برای نقاط غیر هم‌فاصله است.

فرمول‌های مرتبه بالاتر از درونیاب‌های درجه بالاتر نتیجه می‌شوند که حتماً به معادلات پیچیده‌تری منجر می‌شوند. با توجه به اینکه از برکردن این فرمول‌ها مشکل است، بهتر است فرمول (۹.۴) را به خاطر داشته باشیم و به کمک جدول تفاضلات تقسیم‌شده‌ی نیوتن مشتق را تقریب بزنیم.

مثال ۲.۴. به کمک مقادیر تابع $f(x) = \sin x$ در نقاط ۰، ۰٫۲ و ۰٫۵ تقریبی از مشتق آن در نقطه‌ی صفر بدست می‌آوریم. برای این منظور از جدول تفاضلات نیوتن استفاده می‌کنیم. در اینجا محاسبات را تا چهار رقم اعشار انجام می‌دهیم.

$x_۰ = ۰$	$f[x_۰] = ۰$
$x_۱ = ۰٫۲$	$f[x_۱] \doteq ۰٫۱۹۸۷ \quad f[x_۰, x_۱] \doteq ۰٫۹۹۳۵$
$x_۲ = ۰٫۵$	$f[x_۲] \doteq ۰٫۴۷۹۴ \quad f[x_۱, x_۲] \doteq ۰٫۹۳۵۷ \quad f[x_۰, x_۱, x_۲] \doteq -۰٫۱۱۵۶$

طبق فرمول (۹.۴) داریم

$$f'(۰) \approx f[x_۰, x_۱] + (x_۰ - x_۱)f[x_۰, x_۱, x_۲] \doteq ۰٫۹۹۳۵ + ۰٫۲ \times ۰٫۱۱۵۶ \doteq ۱٫۰۱۶۶.$$

مقدار واقعیِ مشتق $\cos ۰ = ۱$ است. بنابراین مقدار عددی دارای خطای ۰٫۰۱۶۶ است. برای مقایسه، کران خطای تقریب را هم طبق (۱۱.۴) محاسبه می‌کنیم،

$$|E(f)| \leqslant \frac{۱}{۶}h_۰(h_۰ + h_۱) \max_{t\in[۰,۰٫۵]} |\cos(t)| = \frac{۰٫۲(۰٫۲ + ۰٫۳)}{۶} = \frac{۱}{۶۰} = ۰٫۰۱۶۷,$$

که با خطای عددی هم‌خوانی دارد. ♢

لازم به ذکر است که فرمول‌های پیشرو روی نقاط هم‌فاصله را می‌توان به کمک جدول تفاضلات متناهی Δ^m (روش نیوتن روی نقاط هم‌فاصله) و مشتق‌گیری از فرمول درونیابی (۱۷.۳) نیز بدست آورد. در پرسش ۵ انجام آن از شما خواسته شده است.

فرمول‌های پسرو

فرمول‌های پسرو برای تقریب $f'(x_۰)$، با رعایت ترتیبِ $x_n < x_{n-۱} < \cdots x_۱ < x_۰$ در فرمول (۹.۴) بدست می‌آیند.

مثلاً بازای $n = ۱$ و با فرض اینکه $x_۱ - x_۰ = -h$ داریم

$$\begin{aligned} f'(x_۰) &= f[x_۰, x_۱] + (x_۰ - x_۱)f[x_۰, x_۱, x_۰] \\ &= \frac{f(x_۰) - f(x_۱)}{h} + h\frac{f''(\xi)}{۲!} \end{aligned}$$

که همان فرمول تفاضل پسرو (۲.۴) به همراه جمله‌ی خطا است. معمولاً برای اینکه ترتیب نقاط x_j با ترتیب اندیس آن‌ها هم‌خوانی داشته باشد، بجای $x_۱$، ...، x_n به ترتیب از نمادهای $x_{-۱}$، ...، x_{-n} استفاده می‌شود. با این نمادگذاری فرمول بالا به صورت

$$f'(x_۰) = \frac{f(x_۰) - f(x_{-۱})}{h} + h\frac{f''(\xi)}{۲!},$$

نوشته می‌شود.

فرمول‌های مرتبه بالاتر (روی نقاط ناهم‌فاصله) را می‌توانید به طور مشابه بدست آورید. پرسش ۴ یک مورد را از شما خواسته است.

همانند آنچه در مورد فرمول‌های پیشرو گفته شد، فرمول‌های پسرو روی نقاط هم‌فاصله را نیز می‌توان با مشتق‌گیری از فرمول تفاضل متناهی پسرو (۱۸.۳) و استفاده از جدول تفاضلات پسرو ∇^m، که در واقع همان جدول تفاضلات پیشرو است، بدست آورد. پرسش ۶ را ببینید.

فرمول‌های مرکزی

برای بدست آوردن فرمول‌های مرکزی با روش درونیابی، کافی است در ترتیب نقاط تغییراتی ایجاد کنیم. مثلاً فرض کنید $n = ۲$ و $x_۱ = x_۰ + h$ و $x_۲ = x_۰ - h$. همچنین در اینجا از نماد $x_{-۱}$ بجای $x_۲$ استفاده می‌کنیم. طبق فرمول (۹.۴) و با توجه به اینکه $x_۱ - x_۰ = h$ و $x_۰ - x_{-۱} = h$ داریم

$$\begin{aligned} f'(x_۰) \approx p'_۲(x_۰) &= f[x_۰, x_۱] + (x_۰ - x_۱)f[x_۰, x_۱, x_{-۱}] \\ &= \frac{f(x_۱) - f(x_۰)}{h} - h\frac{\frac{f(x_{-۱})-f(x_۱)}{-۲h} - \frac{f(x_۱)-f(x_۰)}{h}}{-h} \\ &= \frac{f(x_۱) - f(x_{-۱})}{۲h}, \end{aligned}$$

که همان فرمول مرکزی است که قبلاً هم آن را بدست آورده‌ایم. برای تعیین جمله‌ی خطا با فرض $f \in C^۳[x_{-۱}, x_۱]$، طبق (۱۰.۴) داریم

$$E(f) = (x_۰ - x_۱)(x_۰ - x_{-۱})f[x_۰, x_۱, x_{-۱}, x_۰] = -h^۲\frac{f'''(\xi)}{۶}, \quad \xi \in [x_{-۱}, x_۱].$$

فرمول بالا را به سادگی می‌توان به حالتی که این سه نقطه هم‌فاصله نیستند، تعمیم داد. پرسش ۷ را ببینید. فرمول‌های مرتبه بالاتر هم به همین ترتیب بدست می‌آیند. به مثال زیر توجه کنید.

مثال ۳.۴. تقریبی مرکزی از $f'(۱)$ به کمک مقادیر جدول زیر بدست می‌آوریم.

x_k	۰/۲۵	۰/۷۵	۱	۱/۵	۲
$f(x_k)$	۱	۰/۵	۰	۰/۲۵	−۰/۲۵

برای این منظور جدول تفاضلات تقسیم شده را به صورت زیر تشکیل می‌دهیم. در این جدول برای رعایت اختصار از نماد $f_{ij\dots k}$ بجای $f[x_i, x_j, \dots, x_k]$ استفاده می‌کنیم.

$x_۴ = ۰/۲۵$	$f_۴ = ۱$				
$x_۳ = ۰/۷۵$	$f_۳ = ۰/۵$	$f_{۴۳} = -۱$			
$x_۰ = ۱$	$f_۰ = ۰$	$f_{۳۰} = -۲$	$f_{۴۳۰} \doteq -۱/۳۳۳۳$		
$x_۱ = ۱/۵$	$f_۱ = ۰/۲۵$	$\underline{f_{۰۱} = ۰/۵}$	$f_{۳۰۱} \doteq -۳/۳۳۳۳$	$f_{۴۳۰۱} \doteq -۱/۸۶۶۷$	
$x_۲ = ۲$	$f_۲ = -۰/۲۵$	$f_{۱۲} = -۱$	$\underline{f_{۰۱۲} = -۱/۵}$	$\underline{f_{۳۰۱۲} \doteq ۱/۴۶۶۷}$	$\underline{f_{۴۳۰۱۲} \doteq ۱/۹۰۴۸}$

طبق فرمول (۹.۴)، تقریب $f'(۱)$ به صورت زیر نوشته می‌شود

$$\begin{aligned} f'(۱) &\approx f_{۰۱} + (x_۰ - x_۱)f_{۰۱۲} + (x_۰ - x_۱)(x_۰ - x_۲)(x_۰ - x_۳)f_{۰۱۲۳} \\ &\quad + (x_۰ - x_۱)(x_۰ - x_۲)(x_۰ - x_۳)(x_۰ - x_۴)f_{۰۱۲۳۴} \\ &\doteq ۲/۱۶۱۹ \end{aligned}$$

تفاضلات تقسیم‌شده که در این تقریب استفاده می‌شوند، در جدول با خط زیر مشخص شده‌اند. توجه کنید که از خاصیت متقارن بودن تفاضلات تقسیم‌شده هم استفاده کرده‌ایم. مثلاً $f_{۰۱۲۳} = f_{۳۰۱۲}$ یا $f_{۰۱۲۳۴} = f_{۴۳۰۱۲}$. ◇

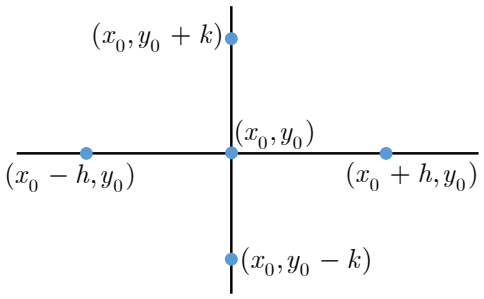
۳.۴ مشتقات جزئی

فرمول‌های مشتق‌گیری که در بخش‌های قبل توضیح داده شدند، برای مشتق‌گیری روی خط حقیقی، یعنی در حالت یک بعدی، استفاده می‌شوند. اگر تابع f چندمتغیره باشد، باید فرمول‌هایی برای مشتقات جزئی f طراحی کنیم. این فرمول‌ها از روی فرمول‌های یک متغیره ساخته می‌شوند.

در اینجا حالتِ دو متغیره را توضیح می‌دهیم که تعمیم آن به ابعاد بالاتر سرراست است. فرض می‌کنیم $f(x, y)$ تابعی مشتق‌پذیر نسبت به هر دو متغیرش در نقطه‌ی $(x_۰, y_۰) \in \mathbb{R}^۲$ باشد. در این صورت فرمول‌های تفاضل پیشرو برای مشتقات جزئی مرتبه اول به همراه جملات خطا به صورت زیر نوشته می‌شوند

$$\begin{aligned} \frac{\partial f}{\partial x}(x_۰, y_۰) &= \frac{f(x_۰ + h, y_۰) - f(x_۰, y_۰)}{h} - \frac{h}{۲}\frac{\partial^۲ f}{\partial x^۲}(\xi, y_۰), \quad \xi \in [x_۰, x_۰ + h], \\ \frac{\partial f}{\partial y}(x_۰, y_۰) &= \frac{f(x_۰, y_۰ + k) - f(x_۰, y_۰)}{k} - \frac{h}{۲}\frac{\partial^۲ f}{\partial y^۲}(x_۰, \eta), \quad \eta \in [y_۰, y_۰ + k]. \end{aligned}$$

همانطور که مشاهده می‌کنید، در مشتق جزئی نسبت به x، نمو h فقط روی متغیر x تغییرات ایجاد می‌کند و به طور مشابه در مشتق جزئی نسبت به y، تغییرات روی متغیر y است. شکل ۶.۴ را ببینید. فرمول‌های تفاضل پسرو و تفاضل مرکزی



شکل ۶.۴: مشتق‌گیری عددی در جهت محورهای x و y

برای مشتقات مرتبه اول به طور مشابه بدست می‌آیند. فرمول تفاضل مرکزی برای مشتق مرتبه دوم به کمک فرمول (۸.۴) به صورت زیر بدست می‌آیند

$$\frac{\partial^۲ f}{\partial x^۲}(x_۰, y_۰) = \frac{f(x_۰ - h, y_۰) - ۲f(x_۰, y_۰) + f(x_۰ + h, y_۰)}{h^۲} - \frac{h^۲}{۱۲}\frac{\partial^۴ f}{\partial x^۴}(\xi, y_۰), \quad \xi \in [x_۰ - h, x_۰ + h],$$
$$\frac{\partial^۲ f}{\partial y^۲}(x_۰, y_۰) = \frac{f(x_۰, y_۰ - k) - ۲f(x_۰, y_۰) + f(x_۰, y_۰ + k)}{k^۲} - \frac{k^۲}{۱۲}\frac{\partial^۴ f}{\partial y^۴}(x_۰, \eta), \quad \eta \in [x_۰ - h, x_۰ + h].$$

پس در مورد مشتقات جزئی نکته‌ی جدیدی اضافه نمی‌شود و همگی با توجه به مشتقات معمولی تعیین می‌شوند. بنابراین بیش از این به آن نمی‌پردازیم.

۴.۴ پرسش‌ها

۱. طول گام بهینه‌ی فرمول تفاضل مرکزی در مشتق مرتبه اول و مشتق مرتبه دوم را بدست آورید.

۲. با روش برونیابی ریچاردسون و استفاده از فرمول تفاضل پسروی مرتبه اول (۲.۴)، فرمول پسروی مرتبه دوم (۷.۴) را بدست آورید.

۳. با روش برونیابی ریچاردسون و استفاده از فرمول مرکزی مرتبه دوم (۳.۴)، فرمول مرتبه چهارم (۸.۴) را بدست آورید.

۴. یک فرمول مشتق‌گیری پسرو سه نقطه‌ای برای تقریب $f'(x_۰)$ روی نقاط غیر هم‌فاصله به کمک درونیابی روی نقاط $x_۰$، $x_{-۱} = x_۰ - h_۱$ و $x_{-۲} = x_{-۱} - h_۲$ بدست آورید و خطای آن را با اعمال شرایط همواری لازم روی تابع f تعیین کنید.

۵. با مشتق‌گیری از فرمول درونیابی (۱۷.۳)، روشی برای تولید فرمول‌های مشتق‌گیری پیشرو روی نقاط هم‌فاصله ارائه دهید و کران خطای هر فرمول را بدست آورید.

۶. با مشتق‌گیری از فرمول درونیابی (۱۸.۳)، روشی برای تولید فرمول‌های مشتق‌گیری پسرو روی نقاط هم‌فاصله ارائه دهید و کران خطای هر فرمول را بدست آورید.

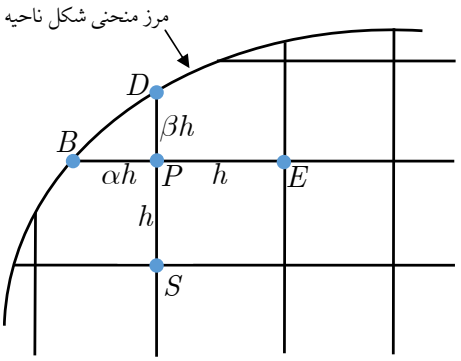
۷. فرض کنید $x_۱ = x_۰ + h_۰$ و $x_{-۱} = x_۰ - h_۱$، که $h_۰, h_۱ > ۰$. به کمک درونیابی، فرمولی مرکزی برای تقریب $f'(x_۰)$ بدست آورید.

۸. در بخش ۱.۴ دیدیم که به علت خطای حذف ارقام بامعنا در عمل تفریق، خطای فرمول‌های مشتق‌گیری برای طول گام‌های کوچکتر از طول گام بهینه، افزایش می‌یابد. سؤالی که ممکن است مطرح شود این است که آیا فرمول مشتق‌گیری‌ای وجود دارد که در آن عمل تقریق نداشته باشیم؟ در این پرسش یک فرمول از این نوع برای تقریب مشتق مرتبه اول ارائه می‌دهیم. این فرمول از طول گام مختلط $\mathrm{i}h$ بجای طول گام حقیقی h استفاده می‌کند. فرض کنید f تابعی تحلیلی در صفحه مختلط و تحدید آن روی محور حقیقی، حقیقی مقدار باشد. برای $x \in \mathbb{R}$ و $h \in \mathbb{R}$ به کمک فرمول تیلر بسط $f(x+\mathrm{i}h)$ را حول x بنویسید و با در نظر گرفتن قسمت موهومی در هر دو طرف ثابت کنید

$$f'(x) = \frac{\mathrm{Im}(f(x+\mathrm{i}h))}{h} + \mathcal{O}(h^۲),$$

که در آن $\mathrm{Im}(z)$ بیانگر قسمت موهومی عدد مختلط z است. یک برنامه متلب برای محاسبه مشتق به کمک این فرمول بنویسید و مرتبه خطا را به صورت تقریبی بدست آورید و نمودارهای لازم را رسم کنید. آیا برای این فرمول نمودار V شکل در خطا ظاهر می‌شود؟ چرا؟

۹. در روش تفاضلات متناهی برای حل معادلات دیفرانسیلِ دو بعدی، معمولاً از فرمول‌های مشتق‌گیری عددی روی یک شبکه‌ی منظم استفاده می‌شود. اگر ناحیه دارای مرز منحنی شکل باشد، برای نقاط نزدیکِ مرز باید از تقریب‌های ساخته شده برای نقاط غیر هم‌فاصله استفاده کنیم. شکل زیر را ببینید. تقریب مشتقات در نقطه‌ی $P = (x_p, y_p)$



مدنظر است. فرض کنید $x_e - x_p = h$، $x_p - x_b = \alpha h$، $y_p - y_s = h$ و $y_d - y_p = \beta h$ که $0 < \alpha \leqslant 1$ و $0 < \beta \leqslant 1$.

الف: تقریب‌های مرکزی برای $\frac{\partial f}{\partial x}(P)$ و $\frac{\partial f}{\partial y}(P)$ بدست آورید. راهنمایی: از پرسش ۷ استفاده کنید.

ب: تقریب‌های مرکزی برای $\frac{\partial^2 f}{\partial x^2}(P)$ و $\frac{\partial^2 f}{\partial y^2}(P)$ بدست آورید.

فصل ۵

انتگرال‌گیری عددی

در فصل پیش، مباحثی در مشتق‌گیری عددی مطرح شد و در این فصل روش‌هایی برای انتگرال‌گیری عددی ارائه خواهیم داد. اگر برای تابع تحت انتگرال، تابع اولیه وجود نداشته باشد یا محاسبه‌ی تابع اولیه مشکل باشد، انتگرال‌گیری عددی به کار خواهد آمد. گاهی نیز انتگرالده فرم بسته ندارد و تنها مقادیری از آن در تعداد متناهی نقطه قابل محاسبه یا در دست است، که در این صورت نیز از انتگرال‌گیری عددی برای تقریب انتگرال استفاده می‌کنیم. تکنیک‌های انتگرال‌گیری عددی به خصوص در حل عددی معادلات دیفرانسیل و معادلات انتگرال کاربرد فراوان دارند.

یک فرمول انتگرال‌گیری برای تابع انتگرال‌پذیر f، غالباً به شکل زیر است

$$\int_a^b f(x)dx = \sum_{k=0}^{n} \omega_k f(x_k) + E_n(f), \tag{۱.۵}$$

که در آن a و b کران‌های انتگرال‌گیری هستند که می‌توانند $\pm\infty$ را نیز اختیار کنند. ضرایب ω_k *وزن‌های انتگرال‌گیری* نامیده می‌شوند و $E_n(f)$ کران خطای انتگرال‌گیری است. روشن است که f باید در نقاط x_k که *گره‌های انتگرال‌گیری* نامیده می‌شوند، قابل تعریف باشد. در این فرمول انتگرال بر حسب مقادیر تابع در گره‌ها نوشته شده است. فرمول‌هایی هم وجود دارند که از مقادیر مشتقات تابع نیز استفاده می‌کنند. در این فصل معمولاً فرمول‌هایی به شکل (۱.۵) بدست می‌آوریم، اما مختصری هم در مورد فرمول‌های مبتنی بر مشتقات صحبت خواهیم کرد.

دو دسته‌ی مهم فرمول‌های انتگرال‌گیری، فرمول‌های نیوتن-کاتس و فرمول‌های گاوسی هستند که در این فصل به آن‌ها می‌پردازیم.

۱.۵ فرمول‌های نیوتن-کاتس

یک روش سرراست برای بدست آوردن فرمول‌های انتگرال‌گیری، تقریب تابع f با یک تابع با ساختار ساده و سپس انتگرال‌گیری از تابع تقریب است. اگر چندجمله‌ای درونیاب روی نقاط x_0، x_1، ...، x_n در بازه‌ی متناهی $[a,b]$ را به

عنوان تقریب f در نظر بگیریم، داریم

$$f(x) = \sum_{k=0}^{n} \ell_k(x) f(x_k) + R_n(f, x),$$

که در آن ℓ_k چندجمله‌ایهای لاگرانژ هستند که در فصل ۳ به صورت

$$\ell_k(x) = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{x - x_i}{x_k - x_i},$$

معرفی شدند. اگر از طرفین تقریب بالا انتگرال‌گیری کنیم، داریم

$$\int_a^b f(x)dx = \sum_{k=0}^{n} \int_a^b \ell_k(x)dx\, f(x_k) + \int_a^b R_n(f, x)dx.$$

با مقایسه‌ی این رابطه با (۱.۵) می‌بینیم که گره‌های انتگرال‌گیری همان نقاط درونیابی هستند و

$$\omega_k = \int_a^b \ell_k(x)dx, \quad k = 0, 1, \ldots, n. \tag{۲.۵}$$

خطای انتگرال‌گیری نیز انتگرال خطای درونیابی روی $[a, b]$ است، یعنی

$$E_n(f) = \int_a^b R_n(f, x)dx.$$

چنین فرمول‌هایی در اوایل قرن هجدهم توسط نیوتن و کاتس توسعه یافتند و به همین علت امروزه به فرمول‌های نیوتن-کاتس مشهورند. گاهی به آن‌ها "درونیاب-فرمول" هم می‌گویند، زیرا همانطور که در بالا دیدیم به کمک چندجمله‌ای درونیاب بدست می‌آیند. این فرمول‌ها گاهی به دو دسته‌ی *فرمول‌های بسته* و *فرمول‌های باز* تقسیم می‌شوند. در فرمول‌های بسته دو نقطه‌ی انتهایی a و b هم جزء نقاط انتگرال‌گیری هستند، در حالی که در فرمول‌های باز نقاط انتگرال‌گیری همگی در (a, b) قرار دارند. این دسته‌بندی نه تنها برای فرمول‌های نیوتن-کاتس، بلکه برای همه‌ی فرمول‌های انتگرال‌گیری استفاده می‌شود. در ادامه‌ی این بخش چند نوع از فرمول‌های نیوتن-کاتس را بدست می‌آوریم، و در بخش‌های بعد برخی فرمول‌های دیگر را هم بررسی می‌کنیم.

فرمول ذوزنقه‌ای

فرض کنیم انتگرال‌گیری روی بازه‌ی متناهی $[x_0, x_1]$ مد نظر است. تابع تحت انتگرال، f، را با چندجمله‌ای درونیاب خطی ($n = 1$) روی نقاط x_0 و x_1 تقریب می‌زنیم. داریم

$$\int_{x_0}^{x_1} f(x)dx = \int_{x_0}^{x_1} p_1(x)dx + \int_{x_0}^{x_1} R_1(x, f)dx,$$

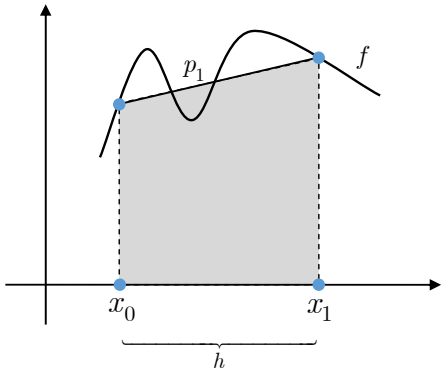
که در آن برای $f \in C^{۲}[x_۰, x_۱]$ و با قرار دادن $f_k = f(x_k)$ داریم

$$p_۱(x) = f_۰ + (x - x_۰)f[x_۰, x_۱], \quad R_۱(f, x) = (x - x_۰)(x - x_۱)\frac{f''\big(\xi(x)\big)}{۲!}, \quad \xi(x) \in [x_۰, x_۱].$$

در اینجا برای سادگی از فرم درونیابی نیوتن به جای درونیابی لاگرانژ استفاده کرده‌ایم ولی خطا در فرم لاگرانژ نوشته شده است. محاسبه‌ی انتگرال $p_۱$ کار ساده‌ای است. با فرض اینکه $h = x_۱ - x_۰$ داریم

$$\int_{x_۰}^{x_۱} p_۱(x)dx = \int_{x_۰}^{x_۱} \big(f_۰ + (x - x_۰)f[x_۰, x_۱]\big)dx = \frac{h}{۲}\big[f_۰ + f_۱\big],$$

که مساحت ذوزنقه‌ای با ارتفاع‌های $f_۰$ و $f_۱$ و عرض h است. شکل ۱.۵ را ببینید. برای محاسبه‌ی خطا باید انتگرال



شکل ۱.۵: انتگرال‌گیری با روش ذوزنقه‌ای

$$E(f) = \frac{۱}{۲}\int_{x_۰}^{x_۱} (x - x_۰)(x - x_۱)f''\big(\xi(x)\big)dx$$

را محاسبه کنیم. با توجه به اینکه f'' پیوسته است و تابع $(x - x_۰)(x - x_۱)$ همواره منفی است (شکل ۳.۳ در فصل ۳ را ببینید)، وجود دارد $\xi_۰ \in [x_۰, x_۱]$ بطوریکه (پرسش ۱ را ببینید)

$$E(f) = \frac{۱}{۲}f''(\xi_۰)\int_{x_۰}^{x_۱} (x - x_۰)(x - x_۱)dx = \frac{-h^{۳}}{۱۲}f''(\xi_۰).$$

این خطای انتگرال‌گیری روی بازه‌ای به طول h است. پس فرمول ذوزنقه‌ای به همراه جمله‌ی خطا به صورت زیر است

$$\int_{x_۰}^{x_۱} f(x)dx = \frac{h}{۲}\big[f_۰ + f_۱\big] - \frac{h^{۳}}{۱۲}f''(\xi_۰), \quad \xi_۰ \in [x_۰, x_۱]. \tag{۳.۵}$$

روشن است که این فرمول از نوع بسته است، زیرا کران‌های انتگرال جزء نقاط انتگرال‌گیری هستند.

با توجه به اینکه درونیاب یک تابع خطی با خودش یکی است، فرمول ذوزنقه‌ای برای توابع خطی، یعنی برای هر $f \in \mathbb{P}_۱$ دقیق است. به طور خلاصه می‌توان نوشت

$$E(f) = ۰, \quad \forall\, f \in \mathbb{P}_۱.$$

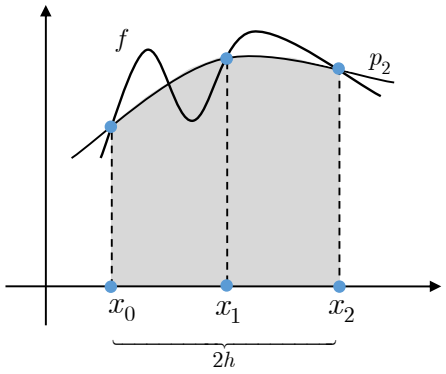
در این صورت می‌گوییم فرمول ذوزنقه‌ای دارای "درجه دقت چندجمله‌ای" یک است. جمله‌ی خطا هم که شامل f'' است مؤید این امر است. به طور کلی تعریف زیر را داریم

تعریف ۱.۵. گوییم فرمول انتگرال‌گیری (۱.۵) دارای درجه دقت چندجمله‌ای m است، اگر بازای هر $f \in \mathbb{P}_m$ داشته باشیم $E_n(f) = ۰$.

در بخش‌های بعد فرمول‌هایی با درجه دقت بالاتر بدست خواهیم آورد.

فرمول سیمسُن

در فرمول سیمسن، تابع f با چندجمله‌ای درونیاب درجه دو ($n = ۲$) تقریب زده می‌شود. در اینجا فرمول سیمسن روی زیربازه‌ی دلخواه $[x_۰, x_۲]$ به طول $۲h$ با نقطه‌ی مرکزی $x_۱ = (x_۰ + x_۲)/۲$ را بدست می‌آوریم. شکل ۲.۵ را ببینید. فرض



شکل ۲.۵: انتگرال‌گیری با روش سیمسُن

کنیم مقادیر تابع در این نقاط $f_۰$، $f_۱$ و $f_۲$ باشند. درونیابِ تابع f روی نقاط $\{x_۰, x_۱, x_۲\}$ را با $p_۲(x)$ نشان می‌دهیم و داریم

$$p_۲(x) = f_۰ + (x - x_۰)f[x_۰, x_۱] + (x - x_۰)(x - x_۱)f[x_۰, x_۱, x_۲],$$
$$R_۲(f, x) = (x - x_۰)(x - x_۱)(x - x_۲)f[x_۰, x_۱, x_۲, x].$$

فرمول سیمسن با انتگرال‌گیری از $p_۲$ و فرمول خطا با انتگرال‌گیری از $R_۲$ روی $[x_۰, x_۲]$ بدست می‌آیند. با قدری محاسبات داریم

$$\int_{x_۰}^{x_۲} p_۲(x)dx = \frac{h}{۳}\big[f_۰ + ۴f_۱ + f_۲\big].$$

واضح است که فرمول سیمسن از درجه دقت دو است، اما خوشبختانه می‌توان نشان داد این فرمول حتی برای چندجمله‌ایهای درجه سه نیز دقیق است. فرض کنیم $p_۳$ یک چندجمله‌ای درجه سه باشد که از مقادیرِ نقاط $x_۰$، $x_۱$ و $x_۲$ می‌گذرد. گیریم x^* نقطه‌ای دیگری در بازه‌ی $[x_۰, x_۲]$ به غیر از سه نقطه‌ی بالا باشد. طبق درونیابی نیوتن داریم

$$p_۳(x) = p_۲(x) + (x - x_۰)(x - x_۱)(x - x_۲)f[x_۰, x_۱, x_۲, x^*].$$

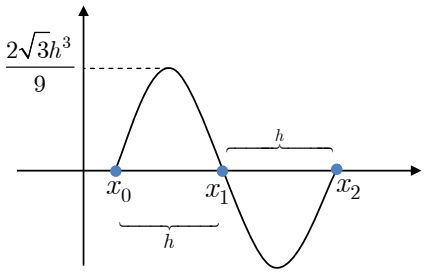
با انتگرال‌گیری از طرفین معادله‌ی بالا داریم

$$\int_{x_0}^{x_2} p_3(x)dx = \int_{x_0}^{x_2} p_2(x)dx + f[x_0, x_1, x_2, x^*] \int_{x_0}^{x_2} (x-x_0)(x-x_1)(x-x_2)dx.$$

به سادگی می‌توان نشان داد $\int_{x_0}^{x_2} (x-x_0)(x-x_1)(x-x_2)dx = 0$. شکل ۳.۵ را هم ببینید. بنابراین داریم

$$\int_{x_0}^{x_2} p_3(x)dx = \int_{x_0}^{x_2} p_2(x)dx.$$

از آنجا که فرمول سیمسن برای انتگرال p_2 دقیق است، برای انتگرال p_3 نیز دقیق است.



شکل ۳.۵: نمودار تابع $(x-x_0)(x-x_1)(x-x_2)$

بدست آوردن فرمول خطا با انتگرال‌گیری از R_2 به آسانیِ فرمولِ ذوزنقه‌ای نیست زیرا تابع $(x-x_0)(x-x_1)(x-x_2)$ تغییر علامت می‌دهد و نمی‌توان از نتیجه‌ی پرسش ۱ استفاده کرد. اما در اینجا به صورت دیگری عمل می‌کنیم. فرض کنیم $f \in C^4[x_0, x_2]$. با توجه به اینکه درجه دقت فرمول سیمسن سه است، جمله‌ی خطا شامل عبارت $f^{(4)}(\xi)$ برای $\xi \in [x_0, x_2]$ است و باید به شکل

$$E(f) = c \times f^{(4)}(\xi)$$

باشد که در آن c یک ضریب ثابت و مستقل از f است. ضریب c را با جایگذاریِ یک تابع ساده بجای f تعیین می‌کنیم. بهترین انتخاب $f(x) = x^4$ است. از آنجا که $f^{(4)}(x) = 4!$، می‌توان نوشت

$$\int_{x_0}^{x_0+2h} x^4 dx = \frac{h}{3}[x_0^4 + 4(x_0+h)^4 + (x_0+2h)^4] + c4!,$$

و با قدری محاسبات، ضریب c با مقدار $-h^5/90$ مستقل از x_0 بدست می‌آید. بنابراین داریم

$$E(f) = -\frac{h^5}{90} f^{(4)}(\xi), \quad \xi \in [x_0, x_2].$$

پس فرمول سیمسن به همراه جمله‌ی خطا برای انتگرال‌گیری روی بازه‌ی $[x_0, x_2]$ و با فرض $h = x_2 - x_1 = x_1 - x_0$ به صورت زیر است

$$\int_{x_0}^{x_2} f(x)dx = \frac{h}{3}[f_0 + 4f_1 + f_2] - \frac{h^5}{90} f^{(4)}(\xi), \quad \xi \in [x_0, x_2], \tag{۴.۵}$$

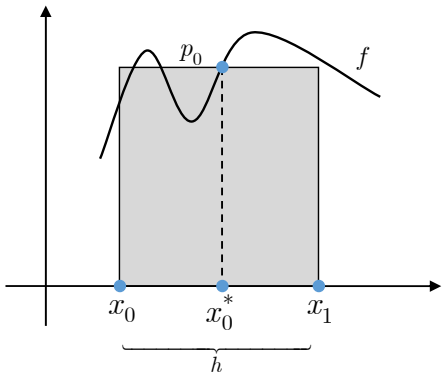
که در آن جمله‌ی خطا با فرض $f \in C^4[x_0, x_2]$ نوشته شده است. این فرمول هم از نوع فرمول‌های بسته است.

فرمول نقطه میانی

در این بخش یک فرمول باز معرفی می‌کنیم. فرض کنیم تقریب انتگرال f روی بازه‌ی متناهی $[x_0, x_1]$ مد نظر است. اگر درونیاب درجه صفر در نقطه‌ی مرکزی $x_0^* = (x_0 + x_1)/2$ را به عنوان تقریبی از f روی $[x_0, x_1]$ در نظر بگیریم، داریم $p_0(x) = f(x_0^*)$ و

$$\int_{x_0}^{x_1} f(x)dx = \int_{x_0}^{x_1} p_0(x)dx + E(f) = hf(x_0^*) + E(f).$$

به این فرمول، فرمول نقطه میانی می‌گوییم که مساحت مستطیلی به عرض $h = x_1 - x_0$ و طول $f(x_0^*)$ است. شکل ۴.۵ را هم ببینید. با توجه به اینکه این فرمول از روی درونیاب درجه صفر حاصل شده است، درجه دقت آن حداقل صفر است.



شکل ۴.۵: انتگرال‌گیری با روش نقطه میانی

اما این فرمول برای چندجمله‌ای‌های خطی نیز دقیق است. زیرا اگر قرار دهیم $f(x) = x$ آنگاه داریم

$$\int_{x_0}^{x_1} x\,dx = \frac{1}{2}(x_1^2 - x_0^2) = \frac{1}{2}(x_1 + x_0)(x_1 - x_0) = hx_0^* = hf(x_0^*).$$

به طریق دیگری هم می‌توان نشان داد که این فرمول دارای درجه دقت دو است. فرض کنیم $\widehat{x}$ نقطه‌ی دیگری غیر از x_0^* در (x_0, x_1) باشد. اگر p_1 درونیاب خطی f روی نقاط x_0^* و $\widehat{x}$ باشد، طبق فرمول درونیابی نیوتن داریم

$$p_1(x) = p_0(x) + (x - x_0^*)f[x_0, \widehat{x}].$$

با توجه به اینکه $\int_{x_0}^{x_1}(x - x_0^*)dx = 0$، داریم

$$\int_{x_0}^{x_1} p_1(x)dx = \int_{x_0}^{x_1} p_0(x)dx = hf(x_0^*),$$

که نشان می‌دهد فرمول نقطه میانی برای چندجمله‌ایهای درجه یک هم دقیق است.

برای بدست آوردن فرمول $E(f)$ بر حسب توان‌های h می‌توانیم همانند روش سیمسن عمل کنیم، که انجام آن در پرسش ۴ از شما خواسته شده است. اما در اینجا با روش دیگری فرمول خطا را بدست می‌آوریم تا شما هم با تکنیک‌های مختلف آشنا شوید. فرض کنیم $f \in C^2[x_0, x_1]$. با توجه به اینکه خطای درونیابی درجه صفر روی نقطه‌ی x_0^* به صورت

$$R_0(x, f) = (x - x_0^*)f[x_0^*, x],$$

است، داریم

$$E(f) = \int_{x_0}^{x_1} (x - x_0^*) f[x_0^*, x] dx.$$

حال تعریف می‌کنیم

$$v(x) = \int_{x_0}^{x} (t - x_0^*) dt.$$

روشن است که $v(x_0) = v(x_1) = 0$. از طرفی با یک انتگرال‌گیری ساده می‌توان دید $v(x) \leqslant 0$ برای هر $x \in [x_0, x_1]$. همچنین طبق قضیه اساسی حساب دیفرانسیل جمله‌ی خطا به صورت زیر نوشته می‌شود

$$E(f) = \int_{x_0}^{x_1} v'(x) f[x_0^*, x] dx.$$

با یک انتگرال‌گیری جزء به جزء و با توجه به اینکه $\frac{d}{dx} f[x_0^*, x] = f[x_0^*, x, x]$، داریم

$$E(f) = v(x) f[x_0^*, x]\Big|_{x_0}^{x_1} - \int_{x_0}^{x_1} v(x) f[x_0^*, x, x] dx.$$

با توجه به $v(x_0) = v(x_1) = 0$، جمله اول سمت راست صفر است. در جمله‌ی دوم سمت راست، طبق پرسش ۱۶ فصل ۳ داریم $f[x_0^*, x, x] = \frac{1}{2} f''\big(\xi(x)\big)$ که $\xi(x) \in [x_0, x_1]$. از آنجا که v تغییر علامت نمی‌دهد و f'' طبق فرض پیوسته است، داریم

$$E(f) = -\frac{1}{2} f''(\xi_0) \int_{x_0}^{x_1} v(x) dx, \quad \xi_0 \in [x_0, x_1].$$

انتگرال v به سادگی قابل محاسبه است و در آخر خواهیم داشت

$$E(f) = \frac{h^3}{24} f''(\xi_0), \quad \xi_0 \in [x_0, x_1].$$

بنابراین فرمول نقطه میانی به همراه جمله‌ی خطا عبارت است از

$$\int_{x_0}^{x_1} f(x) dx = h f(x_0^*) + \frac{h^3}{24} f''(\xi_0), \quad \xi_0 \in [x_0, x_1], \tag{۵.۵}$$

که در آن جمله‌ی خطا با فرضِ $f \in C^2[x_0, x_1]$ بدست آمده است. مشاهده می‌کنیم که اندازه‌ی خطای این فرمول نصف اندازه‌ی خطای فرمول ذوزنقه‌ای است، اگر چه ممکن است مجهول ξ_0 برای این دو فرمول یکسان نباشد.

فرمول‌های از درجه‌ی بالاتر

اگر تابع f را با یک چندجمله‌ای درجه سه ($n = 3$) روی نقاط هم‌فاصله تقریب بزنیم فرمول جدیدی بدست می‌آید که روی زیربازه‌ی نوعیِ $[x_0, x_3]$ با نقاط انتگرال‌گیری $\{x_0, x_1, x_2, x_3\}$ به صورت زیر است

$$\int_{x_0}^{x_3} f(x) dx = \frac{3}{8} h\big[f_0 + 3f_1 + 3f_2 + f_3\big] + E(f).$$

جدول ۱.۵: ضرایب انتگرال‌گیری فرمول‌های نیوتن-کاتس بسته

$n \rightarrow$	۱	۲	۳	۴	۵	۶	۷	۸
ω_0	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{3}{8}$	$\frac{14}{45}$	$\frac{95}{288}$	$\frac{41}{140}$	$\frac{5257}{17280}$	$\frac{3956}{14175}$
ω_1		$\frac{4}{3}$	$\frac{9}{8}$	$\frac{64}{45}$	$\frac{375}{288}$	$\frac{216}{140}$	$\frac{25039}{17280}$	$\frac{23552}{14175}$
ω_2				$\frac{24}{45}$	$\frac{250}{288}$	$\frac{27}{140}$	$\frac{9261}{17280}$	$-\frac{3712}{14175}$
ω_3						$\frac{272}{140}$	$\frac{20923}{17280}$	$\frac{41984}{14175}$
ω_4								$-\frac{18160}{14175}$

این فرمولِ بسته، به خاطر ضریب پشت براکت به فرمول سیمسن $\frac{3}{8}$ مشهور است و درجه دقت آن سه است. بنابراین هیچ مزیتی بر روش سیمسن ندارد در حالی که ارزیابی‌های بیشتری از تابع نیاز دارد. به همین سبب این فرمول معمولاً استفاده نمی‌شود.

ملاحظه ۱.۵. به همان طریقی که نشان دادیم درجه دقت فرمول سیمسن برابر سه است، می‌توان نشان داد درجه دقت درونیاب-فرمول‌ها (فرمول‌های نیوتن-کاتس مبتنی بر p_n یا $n+1$ نقطه‌ای) برابر $n+1$ است، اگر n زوج باشد. در حالیکه اگر n فرد باشد، درجه دقت برابر n است. فرمول ذوزنقه‌ای و فرمول سیمسن $\frac{3}{8}$ جزء دسته‌ی دوم هستند. ♡

اگرچه بدست آوردن فرمول‌های درجه بالاتر سرراست است، اما نیاز به انجام محاسبات زیادی دارد. برای سادگی می‌توان از یک نرم افزار محاسبات نمادین مانند میپل استفاده کرد. در جدول ۱.۵ ضرایب فرمول‌های نیوتن-کاتس بسته با فاکتورگیری از h برای چند مقدار n (درجه‌ی درونیاب) داده شده است. با توجه به اینکه ضرایب متقارن هستند، فقط نیمی از آن‌ها در جدول آورده شده است. فرمول نیوتن-کاتس بسته که بازای $n=4$ بدست می‌آید، فرمول *میلنِ* نام دارد، که به صورت زیر است

$$\int_{x_0}^{x_4} f(x)dx = \frac{h}{45}\left[14f_0 + 64f_1 + 24f_2 + 64f_3 + 14f(x_4)\right] + E(f),$$

و درجه دقت آن پنج است. همچنین اگر برای انتگرال‌گیری در بازه‌ی متناهی $[x_0, x_n]$ نقاط انتگرال‌گیری را به صورت زیر تعریف کنیم

$$x_0 = a+h, \quad x_n = b-h, \quad h = \frac{b-a}{n+1}, \quad n \geqslant 0, \tag{۶.۵}$$

می‌توان فرمول‌های نیوتن-کاتس باز را بدست آورد، که فرمول نقطه میانی، اولینِ آن‌هاست. اما توجه کنید که ما فرمول نقطه میانی را روی بازه‌ای به طول h بدست آوردیم، در حالیکه با انتخاب نقاط به صورت بالا، این فرمول به صورت زیر است

$$\int_{x_0}^{x_2} f(x)dx = 2hf(x_1) + E(f),$$

جدول ۲.۵: ضرایب انتگرال‌گیری فرمول‌های نیوتن-کاتس باز

$n \to$	۰	۱	۲	۳	۴	۵
ω_0	۲	$\frac{۳}{۲}$	$\frac{۸}{۳}$	$\frac{۵۵}{۲۴}$	$\frac{۳۳}{۱۰}$	$\frac{۴۲۷۷}{۱۴۴۰}$
ω_1			$-\frac{۴}{۳}$	$\frac{۵}{۲۴}$	$-\frac{۴۲}{۱۰}$	$-\frac{۳۱۷۱}{۱۴۴۰}$
ω_2					$\frac{۷۸}{۱۰}$	$\frac{۳۹۳۴}{۱۴۴۰}$

که تنها در طول بازه با فرمول (۵.۵) متفاوت است. با این حال اگر نقاط طبق الگوی (۶.۵) انتخاب شوند، ضرایب فرمول‌های نیوتن-کاتس باز برای چند مقدار n (با فاکتورگیری از h) به صورتی است که در جدول ۲.۵ آورده شده است.

انتظار داریم با افزایش n، خطای انتگرال‌گیری به صفر میل کند، اما همانگونه که در فصل ۳ دیدیم خطای درونیابی با افزایش درجه‌ی درونیاب به صفر میل نمی‌کند. به تبع آن خطای انتگرال‌گیری $E_n(f)$ نیز به صفر میل نمی‌کند. این اساسی‌ترین مشکل فرمول‌های نیوتن-کاتس است. همانطور که در جدول مشاهده می‌کنید، برخی از ضرایب فرمول ۹ نقطه‌ای ($n = ۸$) بسته منفی هستند. فرمول‌های بزرگتر نیز که در این جدول ارائه نشده‌اند، دارای ضرایب منفی هستند. ضرایب منفی در فرمول‌های باز از فرمول سه نقطه‌ای به بعد شروع می‌شوند. وجودِ ضرایب منفی، بر پایداری یک فرمول تأثیرگذار است.

برای بررسی پایداری یک فرمول انتگرال‌گیری، به داده‌های ورودی اختلالات کوچکی وارد می‌کنیم و تأثیر آن‌ها را در جواب نهایی اندازه‌گیری می‌کنیم. فرض کنیم اختلالات ε_k به مقادیر $f(x_k)$ برای $۰ \leqslant k \leqslant n$ وارد شده‌اند. این اختلالات می‌توانند به خاطر خطاهای محاسباتی در کامپیوتر یا خطاهای وسایل اندازه‌گیری، به وجود آمده باشند. اگر جوابِ اصلی و جواب مسئله‌ی مختل شده را به ترتیب با

$$Q_n = \sum_{k=۰}^{n} \omega_k f(x_k),$$
$$Q_{n,\varepsilon} = \sum_{k=۰}^{n} \omega_k [f(x_k) + \varepsilon_k],$$

نشان دهیم، آنگاه اختلاف دو جواب به صورت زیر است

$$|Q_n - Q_{n,\varepsilon}| \leqslant \sum_{k=۰}^{n} |\omega_k|\,|\varepsilon_k| \leqslant \|\boldsymbol{\varepsilon}\|_\infty \sum_{k=۰}^{n} |\omega_k|,$$

که در آن $\boldsymbol{\varepsilon} = [\varepsilon_۰, \ldots, \varepsilon_n]$. رابطه‌ی بالا نشان می‌دهد، در صورتی اختلالات کوچک در ورودی منجر به اختلال کوچک در جواب می‌شوند که

$$\sum_{k=۰}^{n} |\omega_k|,$$

مقدار کوچکی باشد. انتظار داریم با افزایش n جواب عددی به مقدار دقیق انتگرال میل کند. در این صورت، برای اینکه

شرط پایداری حفظ شود، لازم است دنباله‌ی

$$\Big\{\sum_{k=0}^{n}|\omega_k|\Big\}_{n\in\mathbb{N}} \tag{۷.۵}$$

کراندار باشد. در فرمول‌های با ضرایب مثبت که حداقل برای توابع ثابت دقیق هستند، این فرض برقرار است زیرا بازای هر n داریم

$$\sum_{k=0}^{n}|\omega_k| = \sum_{k=0}^{n}\omega_k = b-a.$$

برای اثبات، تابع ثابت $f(x)\equiv 1$ را در فرمول انتگرال‌گیری قرار دهید (پرسش ۳ را ببینید). بنابراین فرمول‌هایی که همه‌ی ضرایب آن‌ها مثبت هستند، پایدارند. اما می‌توان ثابت کرد دنباله‌ی (۷.۵) برای فرمول‌های نیوتن-کاتس واگرا است. به همین سبب هیچگاه فرمول‌های نیوتن-کاتس درجه بالا استفاده نمی‌شوند.

برای رفع این مشکل، می‌توان بازه‌ی انتگرال‌گیری را به زیربازه‌هایی با طول کوچک تقسیم کرد و در هر زیربازه یک فرمول‌های نیوتن-کاتس با درجه‌ی پایین به کار برد و همگرایی را با ثابت نگاه داشتن درجه‌ی درونیاب و کاهش طول زیربازه‌ها بدست آورد. به این فرمول‌ها، فرمول‌های نیوتن-کاتس مرکب می‌گوییم.

فرمول ذوزنقه‌ای مرکب

فرض کنیم انتگرال‌گیری روی بازه‌ی متناهی $[a,b]$ مد نظر است. ابتدا بازه‌ی $[a,b]$ را به n زیربازه به طول مساوی h تقسیم می‌کنیم. یعنی

$$h=\frac{b-a}{n},\quad x_k=a+kh,\quad k=0,1,\dots,n.$$

روی هر زیربازه فرمول ذوزنقه‌ای ساده را به کار می‌بریم و در آخر تمام مقادیر بدست آمده را با هم جمع می‌کنیم. در زیربازه‌ی $[x_k,x_{k+1}]$ طبق فرمول ذورنقه‌ای (۳.۵) داریم

$$\int_{x_k}^{x_{k+1}} f(x)dx=\frac{h}{2}\big[f_k+f_{k+1}\big]-\frac{h^3}{12}f''(\xi_k),\quad \xi_k\in[x_k,x_{k+1}],$$

که در آن جمله‌ی خطا با فرض $f\in C^2[x_k,x_{k+1}]$ نوشته شده است و $f_k=f(x_k)$. برای بدست آوردن فرمول انتگرال‌گیری روی کل بازه‌ی $[a,b]$ از خاصیت جمعی عملگر انتگرال استفاده می‌کنیم. داریم

$$\int_a^b f(x)dx=\sum_{k=0}^{n-1}\int_{x_k}^{x_{k+1}} f(x)dx,$$

که این نتیجه می‌دهد

$$\begin{aligned}\int_a^b f(x)dx &= h\Big[\frac{f_0}{2}+f_1+f_2+\cdots+f_{n-1}+\frac{f_n}{2}\Big]-\frac{h^3}{12}\sum_{k=0}^{n-1}f''(\xi_k)\\ &=: T_n(f)+E_n^T(f),\end{aligned}$$

که در آن $T_n(f)$ فرمول انتگرال‌گیری ذوزنقه‌ای مرکب و $E_n^T(f)$ خطای انتگرال‌گیری روی کل بازه‌ی $[a,b]$ است. می‌توان آن را به صورت زیر ساده‌سازی کرد

$$E_n^T(f) = -\frac{۱}{۱۲}h^۲(b-a)\left[\frac{۱}{n}\sum_{k=۰}^{n-۱} f''(\xi_k)\right].$$

با توجه به پیوستگی f''، طبق پرسش ۲ وجود دارد $\xi \in [a,b]$ بطوریکه

$$E_n^T(f) = -\frac{۱}{۱۲}(b-a)h^۲ f''(\xi) \quad a \leqslant \xi \leqslant b. \tag{۸.۵}$$

چون f'' روی $[a,b]$ کران‌دار است، $E_n^T(f) = \mathcal{O}(h^۲)$ وقتی $h \to ۰$. در حقیقت فرمول ذوزنقه‌ای مرکب وقتی $h \to ۰$ با مرتبه $h^۲$ همگراست به شرطی که $f \in C^۲[a,b]$.

برنامه روش ذوزنقه‌ای مرکب به صورت زیر است.

```
function int = trapez(a,b,n,f)
h=(b-a)/n; x=a:h:b; y=f(x);
int=h*(0.5*y(1)+sum(y(2:n))+0.5*y(n+1));
```

در این برنامه a و b کران‌های متناهی انتگرال، n تعداد زیربازه‌ها و f تابع تحت انتگرال است که می‌توان آن را با دستور @(x) تولید کرد.

مثال ۱.۵. می‌خواهیم به کمک فرمول ذوزنقه‌ای مرکب مقادیر تقریبی برای

$$I = \int_{\frac{\pi}{۴}}^{\frac{\pi}{۲}} \cos(\ln x)dx$$

بدست آوریم. مقدار دقیق انتگرال بالا را می‌توان به صورت تحلیلی با توجه به

$$\int \cos(\ln x)dx = \frac{x}{۲}\big[\cos(\ln x) + \sin(\ln x)\big] + c$$

بدست آورد، که تا ۱۶ رقم اعشار عبارت است از $I \doteq ۰/۷۶۲۰۵۳۱۵۷۳۲۲۰۷۸۰$. در اینجا می‌خواهیم به کمک این جواب دقیق، رفتار خطا را بررسی کنیم. بازای $n = ۱$ داریم $h = b - a = \frac{\pi}{۴}$ و

$$T_۱ = \frac{\pi}{۴}\Big[\frac{۱}{۲}f\big(\frac{\pi}{۴}\big) + \frac{۱}{۲}f\big(\frac{\pi}{۲}\big)\Big] \doteq ۰/۷۳۴۶.$$

اگر بازه را به دو زیربازه تقسیم کنیم، داریم $h = \frac{\pi}{۸}$ و

$$T_۲ = \frac{\pi}{۸}\Big[\frac{۱}{۲}f(\frac{\pi}{۴}) + f(\frac{۳\pi}{۸}) + \frac{۱}{۲}f(\frac{\pi}{۲})\Big] \doteq ۰/۷۵۴۸.$$

برای ادامه‌ی کار، از الگوریتم روش کمک می‌گیریم. دستورات زیر را می‌نویسیم:

```
f=@(x) cos(log(x)); a=pi/4; b=pi/2; I=0.762053157322078;
disp(sprintf('  h   numer  error  order\n-----------------------------'));
for k=1:6
    int=trapez(a,b,2^(k-1),f);
    e(k)=abs(I-int);
    h = (b-a)/2^(k-1);
    if k==1
        disp(sprintf('%3.4f  %3.4f  %3.4f',h,int,e(k)));
    else
        order = log2(e(k-1)/e(k));
        disp(sprintf('%3.4f  %3.4f  %3.4f  %3.4f',h,int,e(k),order));
    end
end
```

در درون حلقه هر بار تعداد نقاط دو برابر می‌شوند، یعنی هر بار فاصله‌ی h نصف می‌شود. دستورات `disp` و `sprintf` برای تزیین خروجی برنامه استفاده شده‌اند. شما می‌توانید این دستورات را حذف کرده و خروجی را بر حسب سلیقه‌ی خود ارائه دهید. خروجی برنامه‌ی بالا به صورت زیر است:

```
   h    numer   error   order
-----------------------------
0.7854  0.7346  0.0274
0.3927  0.7548  0.0073  1.9091
0.1963  0.7602  0.0019  1.9736
0.0982  0.7616  0.0005  1.9931
0.0491  0.7619  0.0001  1.9982
```

```
0.0245   0.7620   0.0000   1.9996
```

در ستون اول مقادیر h، در ستون دوم جواب‌های عددی، در ستون سوم خطای انتگرال‌گیری و در ستون آخر مرتبه‌های خطا داده شده‌اند که به نظر می‌رسد مرتبه با کاهش h به عدد ۲ میل می‌کند. در مورد نحوه‌ی محاسبه‌ی ستون آخر باید کمی توضیح دهیم. فرض کنیم خطای یک روش عددی از $\mathcal{O}(h^p)$ است. این بدان معنی است که اگر به عنوان مثال h را نصف کنیم (یعنی h را $\frac{1}{2}$ برابر کنیم)، خطای روش $(\frac{1}{2})^p$ برابر می‌شود. برای محاسبه‌ی مرتبه‌ی p به صورت عددی، فرض می‌کنیم در گام h خطا به صورت

$$e(h) = ch^p$$

برای یک ثابت c باشد. در این صورت با نصف کردن h جمله‌ی خطا عبارت است از

$$e(h/2) = c\Big(\frac{h}{2}\Big)^p.$$

با تقسیم کردن $e(h)$ بر $e(h/2)$ نتیجه می‌گیریم

$$\frac{e(h)}{e(h/2)} = 2^p,$$

که این هم نتیجه می‌دهد

$$p = \log_2 \left(\frac{e(h)}{e(h/2)} \right). \tag{۹.۵}$$

ستون آخر خروجی، `order`، طبق معادله‌ی بالا ساخته شده است، یعنی هر سطر را بر سطر بعدی تقسیم و از آن لگاریتم در مبنای دو گرفته‌ایم (سطر دهم برنامه را ببینید). واضح است که برای سطر اول (یعنی $h = 1$)، نمی‌توان مرتبه‌ی خطا را محاسبه کرد.

داده‌های ستون آخر، مرتبه‌ی $p = 2$ را برای فرمول ذوزنقه‌ای مرکب تصدیق می‌کنند. البته این مرتبه با فرضِ $f \in C^2[a,b]$ بدست آمده است. تأیید اینکه تابع $\cos(\ln x)$ در $C^2[\frac{\pi}{4}, \frac{\pi}{2}]$ است، کار دشواری نیست. ♢

مثال ۲.۵. سؤال این است که بازه‌ی $[0, 1]$ را حداقل به چند زیربازه باید تقسیم کنیم تا خطای محاسبه‌ی

$$I = \int_0^1 \sin x \, dx$$

با روش ذوزنقه‌ای مرکب کمتر از $\varepsilon = 10^{-4}$ شود؟ طبق فرمول خطا، E_n^T، داریم

$$|I - T_n| \leqslant \frac{h^2(b-a)}{12} \max_{t\in[a,b]} |f''(t)|, \quad h = \frac{b-a}{n}.$$

برای اینکه خطای روش از ε کمتر باشد، کافی است

$$\frac{h^2(b-a)}{12} \max_{t\in[a,b]} |f''(t)| \leqslant \varepsilon.$$

در این مثال داریم

$$\frac{h^۲}{۱۲} \leqslant ۱۰^{-۴},$$

که این هم نتیجه می‌دهد

$$h \leqslant \sqrt{۱۲ \times ۱۰^{-۴}},$$

و از آن داریم

$$n \geqslant \frac{۱}{\sqrt{۱۲ \times ۱۰^{-۴}}} \doteq ۸/۳۳۳۳.$$

چون n عددی طبیعی است، باید داشته باشیم $n \geqslant ۹$. ♢

با توجه به جمله‌ی $E_n^T(f)$، فرمول ذوزنقه‌ای مرکب برای توابع $f \in C^۲[a,b]$ با کاهش h همگراست. سؤالی که ممکن است ذهن شما را مشغول کرده باشد، این است که اگر تابع f درجه‌ی همواری کمتری داشته باشد، آیا باز هم فرمول همگراست؟ پاسخ مثبت است به شرطی که f حداقل پیوسته باشد، اما در این صورت مرتبه‌ی همگرایی کمتر است. مثلاً اگر $f \in C^۱[a,b]$، مرتبه‌ی همگرایی فرمول ذوزنقه‌ای مرکب $\mathcal{O}(h)$ است. اثبات این ادعا را می‌توانید در فصل ششم [۶] ببینید.

فرمول سیمسن مرکب

فرمول سیمسن مرکب روی بازه‌ی متناهی $[a,b]$ را می‌سازیم. قرار می‌دهیم $h = (b-a)/n$ و $x_k = a + kh$، برای $k = ۰, ۱, \ldots, n$. با توجه به اینکه در هر زیر بازه به سه نقطه‌ی درونیابی نیاز داریم، لازم است n زوج باشد. برای انتگرال‌گیری در هر زیربازه‌ی $[x_{۲k-۲}, x_{۲k}]$ بازای $k = ۱, ۲, \ldots, \frac{n}{۲}$ طبق فرمول سیمسن ساده‌ی (۴.۵) داریم

$$\int_{x_{۲k-۲}}^{x_{۲k}} f(x)dx = \frac{h}{۳}\left[f_{۲k-۲} + ۴f_{۲k-۱} + f_{۲k}\right] - \frac{h^۵}{۹۰} f^{(۴)}(\xi_k), \quad \xi_k \in [x_{۲k-۲}, x_{۲k}].$$

با جمع بستن روی همه‌ی زیربازه‌ها به فرمول زیر می‌رسیم

$$\begin{aligned}\int_a^b f(x)dx &= \frac{h}{۳}\left[f_۰ + ۴f_۱ + ۲f_۲ + ۴f_۳ + \cdots + ۲f_{n-۲} + ۴f_{n-۱} + f_n\right] - \frac{h^۵}{۹۰}\sum_{k=۱}^{n/۲} f^{(۴)}(\xi_k) \\ &=: S_n(f) + E_n^S(f),\end{aligned}$$

که در آن $S_n(f)$ فرمول سیمسن مرکب و $E_n^S(f)$ جمله‌ی خطای آن است. در این فرمول ضریب اولین و آخرین مقدار داخل براکت برابر ۱ است، ضریب مقادیر با اندیس زوج ۲ و ضریب مقادیر با اندیس فرد ۴ است. همانند روش ذوزنقه‌ای، می‌توان جمله‌ی خطا را بگونه‌ای دیگری بدون علامت سیگما نمایش داد. برای این منظور، در جمله‌ی خطا قرار می‌دهیم

$۲h = \frac{b-a}{n/۲}$ و از فرض $f \in C^۴[a,b]$ و پرسش ۲ استفاده می‌کنیم. داریم

$$E_n^S(f) = -\frac{h^۵}{۹۰}\sum_{k=۱}^{n/۲} f^{(۴)}(\xi_k) = -\frac{h^۴(b-a)}{۱۸۰}\left[\frac{۱}{n/۲}\sum_{k=۱}^{n/۲} f^{(۴)}(\xi_k)\right]$$
$$= -\frac{h^۴(b-a)}{۱۸۰} f^{(۴)}(\xi), \quad \xi \in [a,b].$$

بنابراین اگر $f \in C^۴[a,b]$، آنگاه همگرایی فرمول سیمسن مرکب از $\mathcal{O}(h^۴)$ است.

برنامه روش سیمسن مرکب به صورت زیر است.

```
function int = simpson(a,b,n,f)
if floor(n/2) ~= n/2
   erorr('n should be an even number')
end
h=(b-a)/n; x=a:h:b; y=f(x);
int=h/3*(y(1)+4*sum(y(2:2:n))+2*sum(y(3:2:n-1))+y(n+1));
```

مثال ۳.۵. تقریب‌های سیمسن مرکب برای انتگرال مثال (۱.۵) را بدست می‌آوریم. بازای $n = ۲$ داریم $h = \frac{\pi}{۸}$ و

$$S_۲ = \frac{\pi}{۲۴}\left[f(\frac{\pi}{۴}) + ۴f(\frac{۳\pi}{۸}) + f(\frac{\pi}{۲})\right] \doteq ۰/۷۶۱۵.$$

بازای $n = ۴$ داریم $h = \frac{\pi}{۱۶}$ و مقدار تقریبی انتگرال برابر است با

$$S_۴ = \frac{\pi}{۴۸}\left[f(\frac{\pi}{۴}) + ۴f(\frac{۵\pi}{۱۶}) + ۲f(\frac{۶\pi}{۱۶}) + ۴f(\frac{۷\pi}{۱۶}) + f(\frac{\pi}{۲})\right] \doteq ۰/۷۶۲۰.$$

با تغییرات اندکی در برنامه‌ی مثال ۱.۵ می‌توان آن را برای روش سیمسن بازنویسی کرد و نرخ همگرایی $p = ۴$ را در ستون آخرِ خروجی مشاهده کرد. انجام این کار به شما واگذار می‌شود. ♢

ملاحظه ۲.۵. با توجه به خطای فرمول سیمسن مرکب، اگر $f \in C^۴[a,b]$ آنگاه این فرمول با مرتبه‌ی $h^۴$ به مقدار دقیق انتگرال همگرا است. این بدان معنا نیست که فرمول برای توابعی که هموارای کمتری دارند، همگرا نیست بلکه مرتبه همگرایی پایین‌تر است. در [۶] ثابت شده است، اگر $f \in C^\ell[a,b]$ و $۱ \leqslant \ell \leqslant ۴$ آنگاه فرمول سیمسن مرکب با مرتبه‌ی h^ℓ همگرا است. ♡

فرمول نقطه میانی مرکب

همانند فرمول‌های مرکب قبلی، این فرمول به با استفاده از (۵.۵) به صورت زیر نوشته می‌شود

$$\begin{aligned}\int_a^b f(x)dx &= h\big[f(x_0^*)+f(x_1^*)+\cdots+f(x_{n-1}^*)\big]-\frac{h^2(b-a)}{24}f''(\xi)\\ &=: M_n(f)+E_n^M(f),\quad \xi\in[a,b],\quad x_k^*=\frac{x_k+x_{k+1}}{2}.\end{aligned}\tag{۱۰.۵}$$

جمله‌ی خطا با فرضِ $f\in C^2[a,b]$ بدست آمده است. اگر انتگرالده در نقاط a و b تکینگی ضعیف داشته باشد، فرمول‌های بسته قابل استفاده نیستند. در این صورت می‌توان از فرمول نقطه میانی مرکب استفاده کرد. برنامه‌ی این روش به صورت زیر است:

```
function int = midpoint(a,b,n,f)
h=(b-a)/n; x=a:h:b-h; y=f(x+h/2);
int = h*sum(y);
```

مثال ۴.۵. باز هم انتگرال مثال ۱.۵ را در نظر بگیرید. تقریب‌های نقطه میانی مرکب برای این انتگرال به صورت زیر است

$$\begin{aligned}M_1 &= \frac{\pi}{4}f\big(\frac{3\pi}{8}\big)\doteq 0/7749,\\ M_2 &= \frac{\pi}{8}\Big[f\big(\frac{5\pi}{16}\big)+f\big(\frac{7\pi}{16}\big)\Big]\doteq 0/7656.\end{aligned}$$

برای ادامه‌ی محاسبات از برنامه‌ی روش استفاده می‌کنیم. برای فراخوانی تابع روش نقطه میانی، همان برنامه‌ی روش ذوزنقه‌ای در مثال ۱.۵ را می‌نویسیم و تنها سطر چهارم آن را با

```
int = midpoint(a,b,2^(k-1),f);
```

جایگزین می‌کنیم. خروجی به صورت زیر خواهد بود:

```
   h     numer    error    order
------------------------------
0.7854   0.7749   0.0128
0.3927   0.7656   0.0036   1.8390
```

```
0.1963   0.7630   0.0009   1.9536
0.0982   0.7623   0.0002   1.9878
0.0491   0.7621   0.0001   1.9969
0.0245   0.7621   0.0000   1.9992
```

اگر مقادیر خطا در ستون سوم را با مقادیر خطای روش ذوزنقه‌ای مقایسه کنیم، مشاهده می‌کنیم که خطاها در روش نقطه میانی تقریباً نصف خطاها در روش ذوزنقه‌ای هستند. اما همانند روش ذوزنقه‌ای، نسبت‌ها در ستون آخر به عدد ۲ میل می‌کنند. نتایج عددیِ بالا، کرانِ خطای روش نقطه میانی مرکب را تصدیق می‌کنند، زیرا در (۱۰.۵) دیدیم که مرتبه‌ی خطای روش نقطه میانی مرکب $p = ۲$ است. اما باید به این نکته هم توجه کنیم که این مرتبه با فرضِ $f \in C^۲[۰, ۱]$ بدست آمده است. تابع $\cos(x)$ در این مثال این فرض را برآورده می‌کند. حال می‌خواهیم ببینیم رفتارِ روش برای تابعی که در این شرط همواری صدق نکند، چگونه است. مثال زیر را در نظر می‌گیریم

$$I = \int_۰^۱ \frac{۱}{۲\sqrt{x}} dx.$$

این تابع در نقطه‌ی ابتدای بازه تعریف نشده است. بنابراین روی $[۰, ۱]$ در شرط همواری مورد نیاز در (۱۰.۵) صدق نمی‌کند. فرمول‌های بسته مانند ذوزنقه‌ای و سیمسن را هم نمی‌توان برای محاسبه‌ی این انتگرال به کار گرفت. اما می‌توان از فرمول نقطه میانی مرکب با جایگزینیِ سطر اول برنامه با

```
f = @(x) 1./(2*sqrt(x)); a=0; b=1; I=1;
```

استفاده کرد. نتیجه به صورت زیر است:

```
   h     numer    error    order
-----------------------------
1.0000   0.7071   0.2929
0.5000   0.7887   0.2113   0.4709
0.2500   0.8494   0.1506   0.4890
0.1250   0.8932   0.1068   0.4960
0.0625   0.9244   0.0756   0.4986
0.0313   0.9465   0.0535   0.4995
```

ستون سوم نشان می‌دهد با کاهش h، خطا هم در حال کاهش است اما کاهش خطا سرعت کمتری نسبت به مثال قبل دارد. ستون آخر نشان می‌دهد مرتبه همگرایی تقریباً به $p = ۰/۵$ میل می‌کند. پس اگر تابع تحت انتگرال مرتبه‌ی همواری کمتری داشته باشد، مرتبه‌ی همگرایی روش نقطه میانی (و روش‌های دیگر) کاهش می‌یابد. اثبات دقیقِ این ادعا و تعیین دقیق مرتبه‌ی خطا به صورت تحلیلی، از حوصله‌ی این درس خارج است. ♢

۲.۵ روش ضرایب نامعین

یک روش دیگر برای بدست آوردن فرمول‌های انتگرال‌گیری، روش ضرایب نامعین است. فرض کنیم بدنبال یافتن یک فرمول انتگرال‌گیری به شکل

$$\int_a^b w(x)f(x)dx = \sum_{k=0}^{n} \omega_k f(x_k) + E_n(f), \qquad (۱۱.۵)$$

هستیم که در آن ضرایب ω_k (و حتی گاهی نقاط x_k) مجهول هستند. مجهولات را طوری تعیین می‌کنیم که فرمول انتگرال‌گیری برای تمام توابع در یک زیرفضا از $C[a,b]$ دقیق باشد. مثلاً اگر مجهولات طوری بدست آیند که برای هر $f \in \mathbb{P}_m$ داشته باشیم $E_n(f) = 0$، آنگاه یک فرمول با درجه دقت چندجمله‌ای m خواهیم داشت. با توجه به اینکه انتگرال یک عملگر خطی است، کافی است فرمول برای اعضای پایه‌ی زیرفضا دقیق باشد.

در فرمول (۱۱.۵) یک تابع وزن $w(x)$ که تابعی مثبت و پیوسته روی (a,b) است، در انتگرالده ظاهر شده است. فرمول (۱.۵) حالت خاصی از (۱۱.۵) است که در آن $w(x) \equiv 1$. ابتدا فرض می‌کنیم نقاط انتگرال‌گیری متمایز x_0، x_1، ...، x_n داده شده باشند و فقط وزن‌های ω_k مجهول باشند. همچنین فعلاً فرض می‌کنیم به دنبال فرمول‌هایی با درجه دقت چندجمله‌ای هستیم. با توجه به اینکه تعداد مجهولات $n+1$ است، برای تعیین آن‌ها به صورت یکتا باید $n+1$ معادله‌ی مستقل داشته باشیم. پس زیر فضا باید از بُعد $n+1$ یعنی $\mathbb{P}_n$ باشد. یک پایه برای آن

$$\{1, x, x^2, \ldots, x^n\}$$

است. برای تعیین ضرایب مجهول ω_k کافی است فرمول (۱۱.۵) برای تمام اعضای پایه دقیق باشد. یعنی

$$\begin{array}{lcccccccc}
f(x)=1 & : & \int_a^b w(x)\,dx & = & \omega_0 & + & \omega_1 & + \cdots + & \omega_n \\
f(x)=x & : & \int_a^b xw(x)\,dx & = & \omega_0 x_0 & + & \omega_1 x_1 & + \cdots + & \omega_n x_n \\
\vdots & & \vdots & & \vdots & & & & \\
f(x)=x^n & : & \int_a^b x^n w(x)\,dx & = & \omega_0 x_0^n & + & \omega_1 x_1^n & + \cdots + & \omega_n x_n^n
\end{array}$$

که منجر به حل دستگاه معادلات خطی زیر با یک ماتریس واندرموند می‌شود

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_n \end{bmatrix} = \begin{bmatrix} \int_a^b w(x)\,dx \\ \int_a^b xw(x)\,dx \\ \vdots \\ \int_a^b x^n w(x)\,dx \end{bmatrix}.$$

چون ماتریس واندرموند معکوس‌پذیر است (پرسش ۴ فصل ۳ را ببینید)، ضرایب ω_k به صورت یکتا بدست می‌آیند. با افزایش n ماتریس واندرموند بدوضع می‌شود، اما در اینجا نگران این موضوع نیستیم زیرا همانگونه که قبلاً گفته شد، لزومی به تولید فرمول‌های نیوتن-کاتس درجه بالا نیست.

ملاحظه ۳.۵. اگر نقاط انتگرال‌گیری با فاصله‌ی یکسان به صورت

$$x_k = a + kh, \quad k = ۰, ۱, \dots, n, \quad h = \frac{b-a}{n},$$

در بازه‌ی $[a, b]$ پخش شده باشند، بهتر است با تغییر متغیر

$$t = \frac{x - x_۰}{h}$$

بازه‌ی انتگرال‌گیری را به $[۰, n]$ منتقل کنیم و سپس روش ضرایب نامعین را بکار بریم. در این صورت داریم

$$\int_a^b w(x)f(x)dx = h\int_۰^n \lambda(t)g(t)dt = \omega_۰ g(۰) + \omega_۱ g(۱) + \cdots + \omega_n g(n) + E_n(g), \qquad (۱۲.۵)$$

که در آن $g(t) = f(x_۰ + ht)$ و $\lambda(t) = w(x_۰ + ht)$. حال در فرمول (۱۲.۵) توابع پایه‌ی $\{۱, t, \dots, t^n\}$ را بجای $g(t)$ قرار داده و با فرض $E_n(g) = ۰$، وزن‌های ω_k را با حل دستگاه تعیین می‌کنیم. ♡

مثال ۵.۵. یک فرمول سه نقطه‌ای با درجه دقت حداقل دو به کمک روش ضرایب نامعین بدست می‌آوریم. فرض می‌کنیم $w(x) \equiv ۱$. برای این منظور قرار می‌دهیم

$$\int_{x_۰}^{x_۲} f(x)dx = h\int_۰^۲ g(t)dt = \omega_۰ g(۰) + \omega_۱ g(۱) + \omega_۲ g(۲) + E(g).$$

با فرض اینکه این فرمول برای زیرفضای $\mathbb{P}_۲$ یعنی برای اعضای $\{۱, t, t^۲\}$ دقیق باشد، داریم

$$\begin{array}{lcccccccc} g(t)=۱ & : & \omega_۰ & + & \omega_۱ & + & \omega_۲ & = & h\int_۰^۲ dt = ۲h \\ g(t)=t & : & & & \omega_۱ & + & ۲\omega_۲ & = & h\int_۰^۲ t dt = ۲h \\ g(t)=t^۲ & : & & & \omega_۱ & + & ۴\omega_۲ & = & h\int_۰^۲ t^۲ dt = \frac{۸}{۳}h \end{array}$$

که می‌توان آن را به شکل ماتریسی زیر هم نمایش داد

$$\begin{bmatrix} ۱ & ۱ & ۱ \\ ۰ & ۱ & ۲ \\ ۰ & ۱ & ۴ \end{bmatrix} \begin{bmatrix} \omega_۰ \\ \omega_۱ \\ \omega_۲ \end{bmatrix} = \begin{bmatrix} ۲h \\ ۲h \\ \frac{۸}{۳}h \end{bmatrix}.$$

با حل این دستگاه، $\omega_۰ = \omega_۲ = h/۳$ و $\omega_۱ = ۴h/۳$ بدست می‌آیند، که همان ضرایب فرمول سیمسن ساده هستند. ♢

مثال ۶.۵. می‌خواهیم فرمولی دو نقطه‌ای به شکل

$$\int_{۰}^{۱} \ln \frac{۱}{x} f(x)dx = \omega_{۰} f(۰) + \omega_{۱} f(۱) + E(f) \qquad (۱۳.۵)$$

بدست آوریم که از ماکزیمم درجه دقت چندجمله‌ای باشد. در این مثال تابع وزن $w(x) = \ln \frac{۱}{x}$ هم در انتگرالده ظاهر شده است. با توجه به اینکه نقاط انتگرال‌گیری داده شده‌اند و فقط دو ضریب $\omega_{۰}$ و $\omega_{۱}$ مجهول هستند، پس درجه دقت مورد نظر حداقل یک است و برای تعیین این ضرایب مجهول (نامعین) کافی است فرمول برای اعضای $\{۱, x\}$ دقیق باشد. بنابراین داریم

$$\begin{array}{rcl} f(x) = ۱ \ : \ \omega_{۰} + \omega_{۱} & = & \int_{۰}^{۱} \ln \frac{۱}{x} dx = ۱ \\ f(x) = x \ : \ \omega_{۱} & = & \int_{۰}^{۱} x \ln \frac{۱}{x} dx = \frac{۱}{۴} \end{array}$$

با حل این دستگاه $\omega_{۱} = ۱/۴$ و $\omega_{۰} = ۳/۴$ بدست می‌آیند. این فرمول برای چندجمله‌ایهای درجه دو دقیق نیست زیرا با یک محاسبه‌ی ساده داریم

$$\int_{۰}^{۱} x^{۲} \ln \frac{۱}{x} dx = \frac{۱}{۹} \neq ۰ \times \omega_{۰} + ۱ \times \omega_{۱} = \frac{۱}{۴}.$$

بنابراین ماکزیمم درجه دقت، همان یک است. می‌توان به همان طریقی که خطای روش سیمسن را بدست آوردیم، جمله‌ی خطای این فرمول را هم تعیین کنیم. فرض کنیم $f \in C^{۲}[۰, ۱]$. از آنجا که درجه دقت این فرمول یک است، $E(f)$ شامل عبارت $f''(\xi)$ برای یک $\xi \in [۰, ۱]$ است، یعنی $E(f) = cf''(\xi)$، که c یک ثابت مستقل از f است. برای تعیین c، تابع $f(x) = x^{۲}$ را در (۱۳.۵) جایگزین می‌کنیم. داریم

$$\int_{۰}^{۱} x^{۲} \ln \frac{۱}{x} dx - \left[\frac{۳}{۴} \times ۰ + \frac{۱}{۴} \times ۱ \right] = c \times ۲,$$

که این هم می‌دهد $c = -۵/۷۲$. بنابراین

$$E(f) = -\frac{۵}{۷۲} f''(\xi), \quad \xi \in [۰, ۱].$$

نکته‌ی مهم در مورد فرمول وزن‌دار (۱۳.۵) این است که تابع وزن $\ln \frac{۱}{x}$ در نقطه‌ی ابتدایی $x = ۰$ تعریف نشده است. در حقیقت انتگرال

$$\int_{۰}^{۱} \ln \frac{۱}{x} f(x)dx$$

یک انتگرال تکین (منفرد) است و اگر برای مقادیر x نزدیک صفر داشته باشیم $f(x) = \mathcal{O}(x^{\ell})$ که $\ell > -۱$، این تکینگی ضعیف است یعنی انتگرال همگراست. فرمول‌های ذوزنقه‌ای و سیمسن که مقادیر انتگرالده در نقطه‌ی صفر را نیاز دارند، برای محاسبه‌ی این انتگرال کارایی ندارند، مگر اینکه مقدار انتگرالده در صفر به صورت حدی حساب شود. اما بسیار کاراتر است که از یک فرمول وزن‌دار به شکل (۱۳.۵) (یا فرمول‌های مشابه با تعداد نقاط بیشتر) استفاده کنیم، زیرا فرمول فقط بر حسب مقادیر f نوشته شده است و اثر تکینگی به وزن‌های ω_k منتقل شده است. به این نوع فرمول‌ها، فرمول‌های *انتگرال ضربی* هم می‌گویند. ♢

در مثال بعد یک فرمول انتگرال‌گیری بدست می آوریم که در آن علاوه بر مقادیر تابع، از مقادیر مشتق آن نیز استفاده شده است.

مثال ۷.۵. فرض کنید می‌خواهیم فرمولی به شکل

$$\int_{x_0}^{x_1} f(x)dx = \omega_0 f(x_0) + \omega_1 f(x_1) + \omega_0' f'(x_0) + \omega_1' f'(x_1) + E(f),$$

بدست آوریم که دارای حداکثر درجه دقت چندجمله‌ای باشد. از آنجا که چهار ضریب مجهول داریم، کافی است فرمول برای $\{1, x, x^2, x^3\}$ دقیق باشد. با جایگزینی f با هر یک از اعضای این پایه به دستگاه معادلات زیر می‌رسیم

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ x_0 & x_1 & 1 & 1 \\ x_0^2 & x_1^2 & 2x_0 & 2x_1 \\ x_0^3 & x_1^3 & 3x_0^2 & 3x_1^2 \end{bmatrix} \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_0' \\ \omega_1' \end{bmatrix} = \begin{bmatrix} x_1 - x_0 \\ \frac{1}{2}(x_1^2 - x_0^2) \\ \frac{1}{3}(x_1^3 - x_0^3) \\ \frac{1}{4}(x_1^4 - x_0^4) \end{bmatrix}.$$

اگر قرار دهیم $x_1 - x_0 = h$، با حل دستگاه بالا ضرایب انتگرال‌گیری به صورتی که در فرمول زیر آمده‌اند بدست می‌آیند

$$\begin{aligned} \int_{x_0}^{x_1} f(x)dx &= \frac{h}{2} f(x_0) + \frac{h}{2} f(x_1) - \frac{h^2}{12} f'(x_0) + \frac{h^2}{12} f'(x_1) + E(f) \\ &= \frac{h}{2}\big[f(x_0) + f(x_1)\big] - \frac{h^2}{12}\big[f'(x_0) - f'(x_1)\big] + E(f) \end{aligned} \tag{۱۴.۵}$$

که به آن فرمول ذوزنقه‌ای اصلاح‌شده هم می‌گویند. واضح است که $E(\mathbb{P}_3) = 0$ یعنی درجه دقت آن ۳ است. جمله‌ی خطا را می‌توان مشابه مثال قبل بدست آورد، اما در ادامه آن را به طریق دیگری بدست می‌آوریم.

این فرمول را می‌توان با انتگرال‌گیری از چندجمله‌ای درونیاب ارمیت p_3 که f و مشتق آن را در x_0 و x_1 درونیابی می‌کند، بدست آورد. ابتدا چندجمله‌ای درونیاب ارمیت مبتنی بر داده‌های جدول

x	x_0	x_1
$f(x)$	$f(x_0)$	$f(x_1)$
$f'(x)$	$f'(x_0)$	$f'(x_0)$

را با روش نیوتن بدست می‌آوریم. جدول تفاضلات تقسیم‌شده به صورت زیر است.

x_0	$f(x_0)$			
x_0	$f(x_0)$	$f'(x_0)$		
x_1	$f(x_1)$	$f[x_0, x_1]$	$\frac{f[x_0,x_1]-f'(x_0)}{h}$	
x_1	$f(x_1)$	$f'(x_1)$	$\frac{f'(x_1)-f[x_0,x_1]}{h}$	$\frac{f'(x_1)+f'(x_0)-2f[x_0,x_1]}{h^2}$

چندجمله‌ای درونیاب عبارتست از

$$p_۳(x) = f(x_۰) + (x - x_۰)f'(x_۰) + \frac{f[x_۰, x_۱] - f'(x_۰)}{h}(x - x_۰)^۲ + \frac{f'(x_۱) + f'(x_۰) - ۲f[x_۰, x_۱]}{h^۲}(x - x_۰)^۲(x - x_۱).$$

با فرض اینکه $f \in C^۴[x_۰, x_۱]$، جمله‌ی خطا طبق فرمول (۴۳.۳) به صورت زیر نوشته می‌شود

$$R_۳(f, x) = \frac{۱}{۴!}(x - x_۰)^۲(x - x_۱)^۲ f^{(۴)}\big(\xi(x)\big), \quad \xi(x) \in [x_۰, x_۱].$$

با انتگرال‌گیری از $p_۳$ روی بازه‌ی $[x_۰, x_۱]$ و با قدری محاسبات داریم

$$\int_{x_۰}^{x_۱} p_۳(x)dx = \frac{h}{۲}f(x_۰) + \frac{h}{۲}f(x_۱) - \frac{h^۲}{۱۲}f'(x_۰) + \frac{h^۲}{۱۲}f'(x_۱),$$

که همان فرمول انتگرال‌گیری (۱۴.۵) است. حسن روش درونیابی نسبت به روش ضرایب نامعین این است که می‌توان فرمول خطای $E(f)$ را مستقیماً با انتگرال‌گیری از $R_۳(f, x)$ بدست آورد. داریم

$$\begin{aligned} E(f) &= \int_{x_۰}^{x_۱} R_۳(f, x)dx = \frac{۱}{۴!}\int_{x_۰}^{x_۱} (x - x_۰)^۲(x - x_۱)^۲ f^{(۴)}\big(\xi(x)\big)dx \\ &= \frac{f^{(۴)}(\eta)}{۴!}\int_{x_۰}^{x_۱} (x - x_۰)^۲(x - x_۱)^۲ dx \\ &= \frac{h^۵}{۷۲۰}f^{(۴)}(\eta), \quad \eta \in [x_۰, x_۱]. \end{aligned}$$

تساوی در سطر دوم به دلیل پیوستگی $f^{(۴)}$ و تغییر علامت ندادن تابع $(x - x_۰)^۲(x - x_۱)^۲$ روی $[x_۰, x_۱]$ و طبق پرسش ۱ برقرار است. برای انتگرال‌گیری روی یک بازه‌ی بزرگ، می‌توان صورت مرکب این فرمول را بکار برد. پرسش ۵ را ببینید.

به این فرمول و فرمول‌های مشابه آن که از مقادیر مشتقات تابع استفاده می‌کنند، گاهی "ارمیت-درونیاب-فرمول" می‌گویند. ♢

در مثال بعد فرض می‌کنیم علاوه بر وزن‌های ω_k، نقاط انتگرال‌گیری x_k هم مجهول باشند. در این صورت تعداد مجهولات دو برابر است و برای تعیین آن‌ها لازم است چندجمله‌ایهای درجه بالاتر هم استفاده شوند. بنابراین فرمول‌هایی با درجه دقت بالاتر بدست می‌آیند.

مثال ۸.۵. نقطه‌ی $x_۰$ و وزن $\omega_۰$ را طوری بدست می‌آوریم که فرمول یک نقطه‌ایِ

$$\int_{-۱}^{۱} f(x)dx = \omega_۰ f(x_۰) + E(f),$$

از ماکزیمم درجه دقت چندجمله‌ای باشد. از آنجا که باید دو مجهول تعیین شوند، فرض می‌کنیم فرمول بالا برای اعضای $\{۱, x\}$ دقیق باشد. داریم

$$\begin{aligned} f(x) = ۱ \ &: \ \omega_۰ \ = \ ۲ \\ f(x) = x \ &: \ \omega_۰ x_۰ \ = \ ۰ \end{aligned}$$

که می‌دهد $x_۰ = ۰$ و $\omega_۰ = ۲$. این همان فرمول نقطه میانی است. همانگونه که قبلاً گفته شد و اینجا هم دیدیم، درجه دقت آن دو است.

حال یک فرمول دو نقطه‌ای با نقاط و ضرایب مجهول بدست می‌آوریم. این فرمول به شکل

$$\int_{-۱}^{۱} f(x)dx = \omega_۰ f(x_۰) + \omega_۱ f(x_۱) + E(f),$$

است، که برای تعیین مجهولات فرض می‌کنیم روی $\{۱, x, x^۲, x^۳\}$ دقیق است. معادلات زیر بدست می‌آیند

$$\begin{aligned}
\omega_۰ &+ \omega_۱ &&= ۲\\
\omega_۰ x_۰ &+ \omega_۱ x_۱ &&= ۰\\
\omega_۰ x_۰^۲ &+ \omega_۱ x_۱^۲ &&= \frac{۱}{۳}\\
\omega_۰ x_۰^۳ &+ \omega_۱ x_۱^۳ &&= ۰
\end{aligned}$$

که یک دستگاه معادلات غیر خطی است. برای حل آن به صورت زیر عمل می‌کنیم. ابتدا به سادگی می‌توان دید هیچ یک از مجهولات $x_۰$، $x_۱$، $\omega_۰$ و $\omega_۱$ صفر نیستند. اگر معادله‌ی دوم را در $x_۰^۲$ ضرب کرده و از معادله‌ی چهارم کم کنیم، داریم $\omega_۱ x_۱(x_۱^۲ - x_۰^۲) = ۰$، که این هم می‌دهد $x_۱ = \pm x_۰$. اگر $x_۱ = x_۰$، معادله‌ی اول و دوم به تناقض می‌رسند. پس $x_۱ = -x_۰$. با جایگذاری در معادله‌ی دوم و استفاده از معادله‌ی اول داریم

$$\omega_۰ = \omega_۱ = ۱.$$

از طرف دیگر با جایگذاری این وزن‌ها در معادله سوم و با توجه به اینکه $x_۱ = -x_۰$ داریم

$$x_۰ = -\frac{\sqrt{۳}}{۳}, \quad x_۱ = \frac{\sqrt{۳}}{۳}.$$

بنابراین فرمول دو نقطه‌ای زیر بدست می‌آید

$$\int_{-۱}^{۱} f(x)dx = f\Big(-\frac{\sqrt{۳}}{۳}\Big) + f\Big(\frac{\sqrt{۳}}{۳}\Big) + E(f), \tag{۱۵.۵}$$

که از درجه دقت ۳ است، یعنی $E(\mathbb{P}_۳) = ۰$. پس اگر $f \in C^۴[-۱, ۱]$، جمله‌ی خطا به صورت $E(f) = cf^{(۴)}(\xi)$ برای یک $\xi \in [-۱, ۱]$ خواهد بود. ضریب c با قرار دادن $f(x) = x^۴$ در (۱۵.۵) با مقدار $c = ۱/۱۳۵$ بدست می‌آید، یعنی داریم

$$E(f) = \frac{۱}{۱۳۵} f^{(۴)}(\xi) \doteq ۰/۰۰۷۴ f^{(۴)}(\xi), \quad \xi \in [-۱, ۱].$$

واضح است که اگر یک فرمول n نقطه‌ای به روش بالا بسازیم دارای درجه دقت $۲n - ۱$ است. به این فرمول‌ها، *فرمول‌های انتگرال‌گیری گاوسی* می‌گویند. بدست آوردن آن‌ها با روش ضرایب نامعین نیاز به حل یک دستگاه معادلات غیرخطی دارد

که برای مقادیر بزرگ n کار دشواری است. در بخش ۵.۵ با دیدگاه متفاوتی فرمول‌های گاوسی را بدست خواهیم آورد. در اینجا تنها یک مثال ارائه می‌دهیم، که در آن

$$I = \int_{-۱}^{۱} \frac{۱}{۱ + x^۲} dx$$

را با روش ذوزنقه‌ای ساده و گاوس دو نقطه‌ای بدست می‌آوریم و جواب‌ها را با جواب دقیق که $I = \frac{\pi}{۴} \doteq ۱/۵۷۰۸$ است، مقایسه می‌کنیم. روش ذوزنقه‌ای می‌دهد

$$I \approx \frac{۲}{۲}\big[f(-۱) + f(۱)\big] = ۱, \quad \text{خطا} \doteq ۰/۵۷۰۸$$

و روش گاوس دو نقطه‌ای نتیجه‌ی زیر را در بر دارد

$$I \approx f(-۱/\sqrt{۳}) + f(۱/\sqrt{۳}) = \frac{۱}{۱ + ۱/۳} + \frac{۱}{۱ + ۱/۳} = ۱/۵, \quad \text{خطا} \doteq ۰/۰۷۰۸.$$

واضح است که روش گاوس تقریب بهتری ارائه می‌دهد. اگر از فرمول‌های گاوسی با تعداد نقاط بیشتر استفاده کنیم تقریب بسیار خوبی بدست خواهیم آورد. ♢

۳.۵ روش انتگرال‌گیری رامبرگ

همانطور که پیش‌تر اشاره کردیم، فرمول‌های نیوتن-کاتس مرتبه بالا به دلیل داشتن وزن‌های منفی مورد استفاده قرار نمی‌گیرند و معمولاً از فرمول‌های مرتبه پایین مرکب استفاده می‌شود. اگر لازم باشد تقریب دقیق‌تری بدست آوریم، باید طول گام h را کوچک و کوچک‌تر کنیم، که منجر به بالا رفتن تعداد ارزیابی‌های تابع و احیاناً بروز خطاهای محاسباتی خواهد شد.

در بحث مشتق‌گیری عددی دیدیم که اگر یک بسط مجانبی برای خطای فرمول مشتق‌گیری داشته باشیم، می‌توان به کمک ایده‌ی برونیابی ریچاردسون، فرمول‌های دقیق‌تری بدست آورد. این ایده را می‌توان برای انتگرال‌گیری عددی هم بکار برد. اما لازم است یک بسط مجانبی بر حسب توان‌های h برای خطا داشته باشیم. در این بخش ایده‌ی ریچاردسون را برای فرمول ذوزنقه‌ای مرکب به کار می‌بریم. با فرض $f \in C^{۲m}[a,b]$، $m \geqslant ۱$، بسط مجانبی

$$\int_a^b f(x)dx - T_n(f) = \alpha_۱ h^۲ + \alpha_۲ h^۴ + \alpha_۳ h^۶ + \cdots + \alpha_{m-۱} h^{۲m-۲} + \mathcal{O}(h^{۲m}),$$

را می‌توان برای فرمول ذوزنقه‌ای مرکب اثبات کرد، که در آن فقط توان‌های زوج h ظاهر شده‌اند. ضرایب α_k وابسته به مشتقات f و مستقل از h می‌باشند. از این پس (در این بخش) بجای نماد $T_n(f)$ از $T_h^۰$ استفاده می‌کنیم و مقدار دقیق انتگرال را همانند قبل با I نشان می‌دهیم. بنابراین بسط بالا به صورت زیر نوشته می‌شود

$$I - T_h^۰ = \alpha_۱ h^۲ + \alpha_۲ h^۴ + \alpha_۳ h^۶ + \cdots + \alpha_{m-۱} h^{۲m-۲} + \mathcal{O}(h^{۲m}), \qquad (۱۶.۵)$$

که نشان می‌دهد خطا از $\mathcal{O}(h^۲)$ است. اگر h را به $h/۲$ تبدیل کنیم به فرمول زیر می‌رسیم

$$I - T^{۰}_{h/۲} = \frac{\alpha_۱}{۴}h^۲ + \frac{\alpha_۲}{۱۶}h^۴ + \frac{\alpha_۳}{۶۴}h^۶ + \cdots + \frac{\alpha_{m-۱}}{۲^{۲m-۲}}h^{۲m-۲} + \mathcal{O}(h^{۲m}). \qquad (۱۷.۵)$$

اکنون بین فرمول‌های (۱۶.۵) و (۱۷.۵) ضریب $h^۲$ را حذف و فرمولی با مرتبه‌ی $h^۴$ بدست می‌آوریم. برای این کار کافی است (۱۶.۵) را از چهار برابر (۱۷.۵) کم کنیم، که می‌دهد

$$I - \underbrace{\frac{۴T^{۰}_{h/۲} - T^{۰}_{h}}{۳}}_{T^{۱}_{h/۲}} = \widehat{\alpha}_۲h^۴ + \widehat{\alpha}_۳h^۶ + \cdots + \widehat{\alpha}_{m-۱}h^{۲m-۲} + \mathcal{O}(h^{۲m}), \qquad (۱۸.۵)$$

که در آن $\widehat{\alpha}_۲ = -\frac{۱}{۴}\alpha_۲$، $\widehat{\alpha}_۳ = -\frac{۵}{۱۶}\alpha_۳$، و بطور کلی $\widehat{\alpha}_k = -\frac{۱}{۳}(۱ - (\frac{۱}{۲})^{۲k-۲})\alpha_k$. فرمول جدید، $T^{۱}_{h/۲}$، از مرتبه‌ی $h^۴$ است. می‌توان اطلاعات را در یک جدول به صورت زیر مرتب کرد

$$\begin{array}{|ll|}\hline T^{۰}_{h} & \\ T^{۰}_{h/۲} & T^{۱}_{h/۲} \\ \hline \end{array}$$

به همین ترتیب می‌توان $T^{۰}_{h/۴}$ را با تبدیل h به $h/۲$ در (۱۷.۵) بدست آورد و با حذف ضریب $h^۲$ بین معادلات شامل $T^{۰}_{h/۲}$ و $T^{۰}_{h/۴}$ به فرمول جدید

$$T^{۱}_{h/۴} = \frac{۴T^{۰}_{h/۴} - T^{۰}_{h/۲}}{۳}$$

رسید که از مرتبه‌ی $h^۴$ است. در واقع پس از انجام محاسبات داریم

$$I - T^{۱}_{h/۴} = \frac{\widehat{\alpha}_۲}{۱۶}h^۴ + \frac{\widehat{\alpha}_۳}{۶۴}h^۶ + \cdots + \frac{\widehat{\alpha}_{m-۱}}{۲^{۲m-۲}}h^{۲m-۲} + \mathcal{O}(h^{۲m}). \qquad (۱۹.۵)$$

جدول قبل اکنون درایه‌های بیشتری دارد و به صورت زیر درآمده است

$$\begin{array}{|ll|}\hline T^{۰}_{h} & \\ T^{۰}_{h/۲} & T^{۱}_{h/۲} \\ T^{۰}_{h/۴} & T^{۱}_{h/۴} \\ \hline \end{array}$$

ستون اولِ جدول که شامل درایه‌های $T^{۰}_{h/۲^k}$، $k \geqslant ۰$، است، فرمول‌هایی از مرتبه‌ی $h^۲$ و ستون دوم که که درایه‌های $T^{۱}_{h/۲^k}$، $k \geqslant ۱$، را در بر دارد، فرمول‌هایی از مرتبه‌ی $h^۴$ ارائه می‌دهند. در اینجا می‌توانیم ستون سوم جدول را هم بسازیم. برای این منظور بین فرمول‌های (۱۸.۵) و (۱۹.۵) ضریب $h^۴$ را حذف کرده و فرمولی با مرتبه‌ی $h^۶$ می‌سازیم. کافی است (۱۹.۵) در $۴^۲ = ۱۶$ ضرب کرده و (۱۸.۵) را از آن کم کنیم. این کار منجر به فرمول جدید

$$T^{۲}_{h/۴} := \frac{۱۶T^{۱}_{h/۴} - T^{۱}_{h/۲}}{۱۵} = \frac{۴^۲T^{۱}_{h/۴} - T^{۱}_{h/۲}}{۴^۲ - ۱}$$

می‌شود که از مرتبه‌ی $h^۶$ است و داریم

$$I - T^{(۲)}_{h/۴} = \widetilde{\alpha}_۳ h^۶ + \widetilde{\alpha}_۴ h^۸ + \cdots + \widetilde{\alpha}_{m-۱} h^{۲m-۲} + \mathcal{O}(h^{۲m}),$$

که مقادیر $\widetilde{\alpha}_k$ صریحاً بر حسب α_k بدست می‌آیند. جای درایه‌ی $T^{(۲)}_{h/۴}$ در ستون سوم و سطر سوم جدول است. جدول جدید به صورت زیر است

$T^{(۰)}_h$		
$T^{(۰)}_{h/۲}$	$T^{(۱)}_{h/۲}$	
$T^{(۰)}_{h/۴}$	$T^{(۱)}_{h/۴}$	$T^{(۲)}_{h/۴}$

محاسبات را می‌توان با اضافه کردن سطرهای بعدی و به تبع آن ستون‌های بعدی ادامه داد. ستون اول، مقادیر روش ذوزنقه‌ای مرکب بازای طول گام‌های مختلف (با نصف شدن در هر سطر) می‌باشد. درایه‌های دیگر از روی درایه قبل و بالا در ستون قبل بدست می‌آیند. بنابراین به غیر از ستون اول، نیازی به محاسبه‌ی مقادیر تابع در ستون‌های دیگر نیست. با افزایش تعداد ستون‌ها، مرتبه‌ی خطا با توان‌های زوج h افزایش می‌یابد. در عمل با تولید تعداد متناهی از سطر و ستون‌های جدول، به دقت مورد نظر دست خواهیم یافت. به این روش انتگرال‌گیری، *روش رامبرگ* می‌گوییم.

برای سادگی نماد جدید

$$R(k,j) := T^{(j)}_{h/۲^k}, \quad k \geqslant ۰, \quad ۰ \leqslant j \leqslant k,$$

را تعریف می‌کنیم و جدول رامبرگ را به شکل زیر در نظر می‌گیریم

$R(۰,۰)$			
$R(۱,۰)$	$R(۱,۱)$		
$R(۲,۰)$	$R(۲,۱)$	$R(۲,۲)$	
$\vdots$	$\vdots$	$\vdots$	$\ddots$

ستون اول با روش ذوزنقه‌ای مرکب به دست می‌آید و درایه‌های دیگر، با فرمول

$$R(k,j) = \frac{۴^j R(k,j-۱) - R(k-۱,j-۱)}{۴^j - ۱}, \quad k \geqslant ۱, \quad ۱ \leqslant j \leqslant k, \qquad (۲۰.۵)$$

تعیین می‌شوند. ستون j-ام تقریب‌هایی از مرتبه $h^{۲j}$ در بر دارد.

مثال ۹.۵. انتگرال زیر، تابع اولیه بر حسب توابع مقدماتی ندارد و برای تخمین آن باید متوسل به روش‌های عددی شویم:

$$\int_۰^۱ \frac{\sin x}{۱+x} dx.$$

در این مثال، تقریب‌هایی از آن به کمک روش رامبرگ بدست می‌آوریم. با $h = ۱$ شروع می‌کنیم و آن را هر بار نصف می‌کنیم. با استفاده از فرمول ذوزنقه‌ای مرکب داریم

$$R(۰,۰) = T_h^۰ = \frac{۱}{۲}[f(۰) + f(۱)] \doteq ۰/۲۱۰۴$$
$$R(۱,۰) = T_{h/۲}^۰ = \frac{۱}{۴}[f(۰) + ۲f(\frac{۱}{۲}) + f(۱)] \doteq ۰/۲۶۵۰$$

به کمک این دو مقدار، درایه‌ی $R(۱,۱)$ را بدست می‌آوریم،

$$R(۱,۱) = \frac{۴R(۱,۰) - R(۰,۰)}{۳} \doteq ۰/۲۸۳۲$$

قطعاً شما هم با ادامه‌ی محاسبات به این صورت موافق نیستید و ترجیح می‌دهید برنامه‌ای برای این کار بنویسید. پیش از آن لازم است شرطی برای توقف محاسبات در نظر بگیریم. اگر تعداد دفعاتی که قرار است طول گام h نصف شود از قبل داده شده باشد، ابتدا ستون اول را به کمک فرمول ذوزنقه‌ای تولید می‌کنیم و سپس ستون‌های بعدی را به ترتیب با فرمول (۲۰.۵) محاسبه می‌کنیم. در این حالت، تعداد سطرها و ستون‌های جدول از قبل مشخص است. اما اگر بخواهیم مقدار انتگرال را با دقت از پیش تعیین‌شده‌ای بدست آوریم، باید جدول رامبرگ را سطر به سطر تولید کنیم و در هر سطر که به دقت مفروض رسیدیم متوقف شویم. فرض کنیم ε دقتِ از پیش تعیین شده باشد. چون معمولاً مقدار دقیق انتگرال را نداریم، در سطر k-ام که شرط

$$|R(k,k) - R(k,k-۱)| \leqslant \varepsilon,$$

برقرار شود، متوقف می‌شویم. در حقیقت الگوریتم وقتی پایان می‌یابد که اندازه‌ی اختلاف دو آخرین مقدار محاسبه شده در سطر k-ام از ε کوچکتر باشد. با این توضیحات برنامه‌ی روش رامبرگ (با تولید سطر به سطر جدول) به صورت زیر است:

```
function int = romberg(a,b,ep,f)
R(1,1)=trapez(a,b,1,f);
k=2;
while(1)
   R(k,1)= trapez(a,b,2^(k-1),f);
   for j=2:k
       R(k,j)=(4^(j-1)*R(k,j-1)-R(k-1,j-1))/(4^(j-1)-1);
   end
   if abs(R(k,k)-R(k,k-1))<ep break; end
   k=k+1;
```

```
end
int=R(k,k); disp(R);
```

در این برنامه a و b کران‌های انتگرال، ep دقت از پیش تعیین شده و f تابع تحت انتگرال است. فراخوانی این تابع برای مثال اخیر به صورت زیر انجام می‌شود:

```
f = @(x) sin(x)./(1+x);
int = romberg(0,1,10^-6,f)
```

در اینجا محاسبات برای دقت $۱۰^{-۶}$ انجام می‌شود. با اجرای این برنامه خروجی زیر را خواهیم داشت، که هم جدول رامبرگ و هم آخرین درایه به عنوان تقریب انتگرال را در بر دارد:

```
 0.210367746201974                  0                  0                  0
 0.264992385969055   0.283200599224748                  0                  0
 0.279353950553051   0.284141138747717   0.284203841382581                  0
 0.283004275424243   0.284221050381307   0.284226377823546   0.284226735544831
int =
 0.284226735544831
```

مقدار تقریبی این انتگرال تا ۱۵ رقم اعشار ۰/۲۸۴۲۲۶۹۸۵۵۱۲۴۱۱ است. اگر در برنامه‌ی بالا مقدار ep را برابر واحد گردکردن یعنی eps قرار دهیم، با تولید هشت سطر و ستون از جدول رامبرگ به این جواب می‌رسیم. ♢

درباره‌ی فرمول رامبرگ نکات زیر قابل توجه هستند. همانطور که قبلاً هم گفته شد، تنها در ستون اول لازم است مقادیر تابع محاسبه شوند. برای محاسبه‌ی درایه‌ی $R(k,j)$ از مقادیر $f(a+ih)$، $h=(b-a)/۲^{k-۱}$، که در ستون اول قرار دارند، استفاده می‌شود. ستون دوم جدول رامبرگ همان مقادیر فرمول سیمسن مرکب بازای طول گام‌های $h/۲$، $h/۴$، ... را می‌دهد. پرسش ۶ را ببینید. می‌توان نشان داد ستون سوم مقادیر فرمول نیوتن-کاتس بسته‌ی مرکب بازای $n=۴$ (روش میلن مرکب) را در بر دارد. اما بقیه‌ی ستون‌ها هیچ ارتباطی با فرمول‌های نیوتن-کاتس ندارند. اما مهمترین نکته در مورد فرمول رامبرگ این است که اگر آن را بر حسب مقادیر تابع مرتب کنیم، وزن‌ها همگی مثبت خواهند بود. در حقیقت داریم

$$R(k,j)=\sum_{i=۰}^{n}\omega_{k,i}^{(j)}f(a+ih),\quad n=۲^{k-۱},\quad h=\frac{b-a}{n}, \tag{۲۱.۵}$$

که در آن

$$\sum_{i=0}^{n} \omega_{k,i}^{(j)} = b - a, \quad \omega_{k,i}^{(j)} > 0. \tag{۲۲.۵}$$

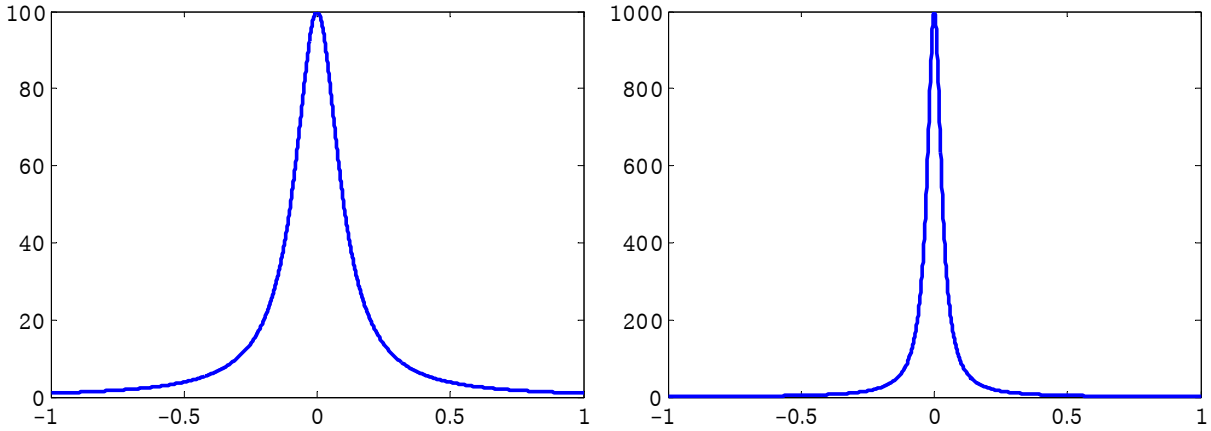
برقرار بودن معادله‌ی (۲۱.۵) با توجه به نحوه‌ی ساختن فرمول رامبرگ، روشن است. قسمت اول (۲۲.۵) هم با جایگذاری تابع ثابت $f(x) = 1$ بدست می‌آید. اما اثباتِ مثبت بودن وزن‌ها قدری پیچیده است و در اینجا ارائه نمی‌شود. علاقه‌مندان می‌توانند مرجع [۱] را ببینند. بنابراین با توجه به مثبت بودن وزن‌ها، فرمول رامبرگ نسبت به اختلال در داده‌های ورودی پایدار است.

۴.۵ انتگرال‌گیری عددی تطبیقی

در فرمول‌های انتگرال‌گیری مرکب، بازه‌ی انتگرال‌گیری، مستقل از تابع تحت انتگرال، به زیربازه‌هایی به طول مساوی تقسیم می‌شود. فرض کنیم اندازه‌ی انتگرالده (یا مشتقات مرتبه پایین آن) روی بازه‌ی $[a, b]$ از نقطه‌ای به نقطه‌ی دیگر تغییرات شدید داشته باشد. در این صورت بهتر است بازه‌ی $[a, b]$ به زیربازه‌هایی با طول‌های متفاوت (بسته به وضعیت انتگرالده) تقسیم شود. به عنوان مثال تابع زیر را در نظر بگیرید

$$f(x) = \frac{1}{10^{-k} + x^2}, \quad x \in [-1, 1], \quad k \geqslant 0.$$

نمودار این تابع بازای $k = 2$ در سمت چپ و بازای $k = 3$ در سمت راست شکل ۵.۵ رسم شده است. این تابع در



شکل ۵.۵: نمودار یک تابع قله‌ای شکل

اطراف صفر شیب تندی دارد و در نزدیکی دو انتها شبیه توابع خطی (یا حتی ثابت) رفتار می‌کند. برای انتگرال‌گیری از چنین تابعی، بهتر است در جاهایی که شیب تند است، طول گام‌ها ریزتر و در جاهای دیگر طول گام‌ها بزرگتر انتخاب شوند. این کار باعث کاهش هزینه‌ی محاسباتی می‌شود. اما برای یک تابع داده شده انتخاب زیربازه‌ها (طول گام‌ها) از پیش معلوم

نیست. در یک فرمول انتگرال‌گیری *تطبیقی*، طول گام‌ها به طور خودکار به نحوی تنظیم می‌شوند که تقریب نهایی $Q(f)$ در خطای از پیش تعیین‌شده‌ی

$$\left|\int_a^b f(x)dx - Q(f)\right| \leqslant \varepsilon, \tag{۲۳.۵}$$

صدق کند. در حقیقت، در یک فرمول تطبیقی، طول زیربازه‌ها با رفتار موضعی انتگرالده تطبیق داده می‌شود. اما در عمل باید رفتار موضعی انتگرالده را از روی مقادیر آن در چند نقطه حدس زد. فرض کنیم زیربازه‌های $[x_k, x_{k+1}]$، $0 \leqslant k \leqslant n-1$، را طوری بدست آورده‌ایم که تقریب انتگرال در هر زیربازه‌ی $[x_k, x_{k+1}]$، (که آن را با $Q_k(f)$ نشان دهیم) در

$$\left|\int_{x_k}^{x_{k+1}} f(x)dx - Q_k(f)\right| \leqslant \frac{h_k}{b-a}\varepsilon, \quad k = 0, 1, \ldots, n, \quad h_k = x_{k+1} - x_k, \tag{۲۴.۵}$$

صدق کند. آنگاه با قرار دادن

$$Q(f) = Q_0(f) + Q_1(f) + \cdots + Q_{n-1}(f),$$

شرط (۲۳.۵) برقرار خواهد بود، زیرا $\sum_{k=0}^{n-1} h_k = b-a$ و $\int_a^b = \int_a^{x_1} + \int_{x_1}^{x_2} + \cdots + \int_{x_{n-1}}^b$.

در اینجا یک روش تطبیقی مبتنی بر روش سیمسن ارائه می‌دهیم. روش کار به این صورت است که ابتدا بازه‌ی $[a, b]$ را به تعداد اندکی زیربازه به طول یکسان h تقسیم می‌کنیم. (می‌توانیم با یک زیربازه که همان $[a, b]$ است نیز شروع کنیم). در هر زیربازه دو تقریب برای انتگرال با فرمول سیمسن یکی با طول گام $h/2$ و دیگری با طول گام $h/4$ بدست می‌آوریم و به کمک این دو تقریب خطای انتگرال‌گیری را تخمین می‌زنیم. اگر خطا در (۲۴.۵) صدق کند، این تقریب به عنوان تقریب زیربازه‌ی مذکور حفظ خواهد شد و به سراغ زیربازه‌ی بعد می‌رویم. اما اگر شرط (۲۴.۵) برآورده نشود، زیربازه را نصف می‌کنیم و محاسبات را تکرار می‌کنیم. این کار را تا آنجا که شرط (۲۴.۵) برای همه‌ی زیربازه‌ها (که اکنون تعداد آن‌ها بیشتر شده است) برقرار شود، تکرار می‌کنیم.

در زیربازه‌ی نوعیِ $[x_k, x_{k+1}]$ فرض کنیم

$$I_k = \int_{x_k}^{x_{k+1}} f(x)dx, \quad h = x_{k+1} - x_k.$$

دو تقریب سیمسن برای این انتگرال یکی با طول گام $h/2$ و دیگری با طول گام $h/4$ بدست می‌آوریم:

$$S = \frac{h}{6}\left[f(x_k) + 4f\left(x_k + \frac{h}{2}\right) + f(x_{k+1})\right], \tag{۲۵.۵}$$

$$S' = \frac{h}{12}\left[f(x_k) + 4f\left(x_k + \frac{h}{4}\right) + 2f\left(x_k + \frac{h}{2}\right) + 4f\left(x_k + \frac{3h}{4}\right) + f(x_{k+1})\right]. \tag{۲۶.۵}$$

فرمول S، فرمول سیمسن ساده روی بازه‌ی $[x_k, x_{k+1}]$ و فرمول S' فرمول سیمسن مرکب روی دو زیربازه از آن است. چون روش ما مبتنی بر روش سیمسن است، از علامت S بجای Q استفاده کرده‌ایم. اگر f به اندازه‌ی کافی هموار باشد،

خطای انتگرال‌گیریِ هر یک از دو فرمول بالا به صورت زیر است

$$I_k - S = -\left(\frac{h}{۲}\right)^۵ \frac{f^{(۴)}(\xi_۱)}{۹۰}, \tag{۲۷.۵}$$

$$I_k - S' = -۲\left(\frac{h}{۴}\right)^۵ \frac{f^{(۴)}(\xi_۲)}{۹۰}. \tag{۲۸.۵}$$

در فرمول (۲۸.۵)، ضریب ۲ پشت جمله‌ی خطا به خاطر مرکب بودن فرمول روی دو زیربازه از $[x_k, x_{k+۱}]$ ظاهر شده است. مقادیر $\xi_۱$ و $\xi_۲$ مجهولاتی احیاناً متفاوت در $[x_k, x_{k+۱}]$ هستند. اگر فرض کنیم برای وقتی که این بازه به اندازه‌ی کافی کوچک باشد، مقادیر $f^{(۴)}(\xi_۱)$ و $f^{(۴)}(\xi_۲)$ تقریباً یکسان و برابر $f^{(۴)}(\xi)$ باشند، با کم کردن (۲۸.۵) از (۲۷.۵) و بعد از ساده‌سازی داریم

$$-۲\left(\frac{h}{۴}\right)^۵ \frac{f^{(۴)}(\xi)}{۹۰} \approx \frac{S' - S}{۱۵}.$$

با جایگذاری در (۲۸.۵)، فرمول زیر را برای خطا بدست می‌آوریم

$$I_k - S' \approx \frac{S' - S}{۱۵}. \tag{۲۹.۵}$$

این معادله بیانگر این است که خطای فرمول S'، تقریباً $\frac{۱}{۱۵}$ اختلاف دو تقریب متوالی است. این همان چیزی است که به دنبالش بودیم: تخمین خطا با استفاده از برخی مقادیر f. حال اگر این تخمین خطا در

$$E := \frac{۱}{۱۵}\left|S' - S\right| \leqslant \frac{h}{b-a}\varepsilon, \tag{۳۰.۵}$$

صدق کند، S' را به عنوان تقریب I_k در نظر می‌گیریم و به سراغ بازه‌ی بعد می‌رویم. اما اگر (۳۰.۵) برقرار نباشد، بازه‌ی $[x_k, x_{k+۱}]$ را نصف کرده و محاسبات را تکرار می‌کنیم. این فرآیند را با یک مثال تشریح می‌کنیم.

مثال ۱۰.۵. تقریبی از

$$I = \int_۰^۱ \exp(-۱۰x)dx$$

به کمک روش تطبیقی بالا با خطای $\varepsilon = ۱۰^{-۴}$ بدست می‌آوریم. تابع تحت انتگرال دارای شیب تندی در اطراف صفر و شیب بسیار ملایم در نزدیکی ۱ است. اما فرض کنیم اطلاعی از شکلِ این تابع نداریم و می‌خواهیم به روش تطبیقی تشریح شده در بالا مقدار این انتگرال را با خطای ε تخمین بزنیم. محاسبات در دقت دو برابر انجام اما تا هشت رقم اعشار نمایش داده می‌شوند. با بازه‌ی $[۰, ۱]$ و $h = ۱$ شروع می‌کنیم. داریم

$$S[۰, ۱] = \frac{۱}{۶}\left[f(۰) + ۴f\left(\frac{۱}{۲}\right) + f(۱)\right] \doteq ۰/۱۷۱۱۶۶۲۰$$

$$S'[۰, ۱] = \frac{۱}{۱۲}\left[f(۰) + ۴f\left(\frac{۱}{۴}\right) + ۲f\left(\frac{۱}{۲}\right) + ۴f\left(\frac{۳}{۴}\right) + f(۱)\right] \doteq ۰/۱۱۲۰۰۶۱۳$$

و تخمین خطا به صورت زیر است

$$E[۰, ۱] = \frac{۱}{۱۵}|S' - S| \doteq ۰/۰۰۳۹ \not\leq \frac{h}{b-a}\varepsilon = ۱۰^{-۴}.$$

پس با تقسیم بازه‌ی $[۰, ۱]$ به دو زیر بازه‌ی $[۰, \frac{۱}{۲}]$ و $[\frac{۱}{۲}, ۱]$ و $h = ۱/۲$ محاسبات را ادامه می‌دهیم. در زیربازه‌ی $[\frac{۱}{۲}, ۱]$ داریم

$$S[\frac{۱}{۲}, ۱] = \frac{۱}{۱۲}\left[f\left(\frac{۱}{۲}\right) + ۴f\left(\frac{۳}{۴}\right) + f(۱)\right] \doteq ۰/۰۰۰۷۴۹۶۴$$
$$S'[\frac{۱}{۲}, ۱] = \frac{۱}{۲۴}\left[f\left(\frac{۱}{۲}\right) + ۴f\left(\frac{۵}{۸}\right) + ۲f\left(\frac{۶}{۸}\right) + ۴f\left(\frac{۷}{۸}\right) + f(۱)\right] \doteq ۰/۰۰۰۶۷۶۸۸$$
$$E[\frac{۱}{۲}, ۱] = \frac{۱}{۱۵}|S' - S| \doteq ۴/۸۵۰۵ \times ۱۰^{-۶} < \frac{۱/۲}{۱}۱۰^{-۴}$$

بنابراین $S'[\frac{۱}{۲}, ۱]$ را به عنوان تقریب انتگرال در زیربازه‌ی $[\frac{۱}{۲}, ۱]$ ذخیره می‌کنیم و به سراغ زیربازه‌ی $[۰, \frac{۱}{۲}]$ می‌رویم. داریم

$$S[۰, \frac{۱}{۲}] \doteq ۰/۱۱۱۲۵۶۴۹, \quad S'[۰, \frac{۱}{۲}] \doteq ۰/۱۰۰۴۵۸۲۵$$
$$E[۰, \frac{۱}{۲}] \doteq ۷/۱۹۸۸ \times ۱۰^{-۴} \not\leq \frac{۱/۲}{۱}۱۰^{-۴},$$

با توجه به اینکه کران خطا برآورده نشده است، باید این زیربازه را به دو زیربازه تقسیم کنیم. با محاسباتی مشابه و با فرض $h = ۱/۴$ داریم

$$S[\frac{۱}{۴}, \frac{۱}{۲}] \doteq ۰/۰۰۷۶۲۰۵۸, \quad S'[\frac{۱}{۴}, \frac{۱}{۲}] \doteq ۰/۰۰۷۵۴۰۸۱$$
$$E[\frac{۱}{۴}, \frac{۱}{۲}] \doteq ۵/۳۱۸۲ \times ۱۰^{-۶} < \frac{۱/۴}{۱}۱۰^{-۴},$$

که نشان می‌دهد معیار خطا برآورده شده است. پس $S'[\frac{۱}{۴}, \frac{۱}{۲}]$ را هم به عنوان تقریب انتگرال در بازه‌ی $[\frac{۱}{۴}, \frac{۱}{۲}]$ ذخیره می‌کنیم. در زیربازه‌ی $[۰, \frac{۱}{۴}]$ داریم

$$S[۰, \frac{۱}{۴}] \doteq ۰/۰۹۲۸۳۷۶۷, \quad S'[۰, \frac{۱}{۴}] \doteq ۰/۰۹۱۸۶۵۸۴$$
$$E[۰, \frac{۱}{۴}] \doteq ۶/۴۷۸۹ \times ۱۰^{-۵} \not\leq \frac{۱/۴}{۱}۱۰^{-۴},$$

که نشان می‌دهد باید این بازه را به دو زیربازه تقسیم کنیم. با فرض $h = ۱/۸$ داریم

$$S[\frac{۱}{۸}, \frac{۱}{۴}] \doteq ۰/۰۲۰۴۵۸۵۳, \quad S'[\frac{۱}{۸}, \frac{۱}{۴}] \doteq ۰/۰۲۰۴۴۳۰۵$$
$$E[\frac{۱}{۸}, \frac{۱}{۴}] \doteq ۱/۰۳۲۲ \times ۱۰^{-۶} < \frac{۱/۸}{۱}۱۰^{-۴},$$

که ملاک خطا را برآورده می‌کند. پس $S'[\frac{۱}{۸},\frac{۱}{۴}]$ نیز به عنوان تقریب انتگرال روی $[\frac{۱}{۸},\frac{۱}{۴}]$ منظور می‌شود. از طرف دیگر داریم

$$S[۰,\frac{۱}{۸}] \doteq ۰/۰۷۱۴۰۷۳۰۲, \quad S'[۰,\frac{۱}{۸}] \doteq ۰/۰۷۱۳۵۳۲۶$$
$$E[۰,\frac{۱}{۸}] \doteq ۳/۶۰۳۰ \times ۱۰^{-۶} < \frac{۱/۸}{۱}۱۰^{-۴},$$

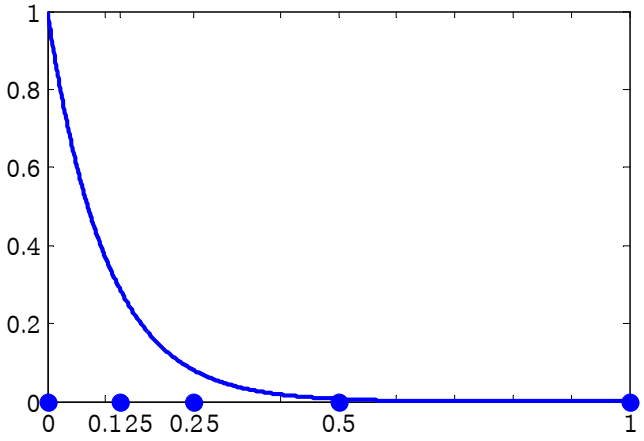
که این هم ملاک خطا را برآورده می‌کند و $S'[۰,\frac{۱}{۸}]$ نیز تخمین انتگرال روی $[۰,\frac{۱}{۸}]$ خواهد بود. فرآیند تطبیقی در اینجا به اتمام می‌رسد و در آخر قرار می‌دهیم

$$Q(f) = S'[\frac{۱}{۲},۱] + S'[\frac{۱}{۴},\frac{۱}{۲}] + S'[\frac{۱}{۸},\frac{۱}{۴}] + S'[۰,\frac{۱}{۸}] \doteq ۰/\,۱۰۰۰۱۴۰۰.$$

خطای کلی با جمع خطاهای متناظر بدست می‌آید. داریم

$$E(f) = E[\frac{۱}{۲},۱] + E[\frac{۱}{۴},\frac{۱}{۲}] + E[\frac{۱}{۸},\frac{۱}{۴}] + E[۰,\frac{۱}{۸}] \doteq ۱/۴۸۰۴ \times ۱۰^{-۵} \leqslant \varepsilon,$$

که نشان می‌دهد معیار خطای کلی برآورده شده است. نمودار این تابع به همراه زیربازه‌هایی که روش انتگرال‌گیری تنظیم کرده است، در شکل ۶.۵ رسم شده است.



شکل ۶.۵: نمودار تابع $\exp(-۱۰x)$ و زیربازه‌های نهایی روش سیمسن تطبیقی برای خطای $\varepsilon = ۱۰^{-۴}$

مشاهده می‌شود که در نزدیکی صفر که تابع دارای قله با شیب تند است، تعداد زیربازه‌ها بیشتر است. ♢

برنامه‌ی روش انتگرال‌گیری تطبیقی مبتنی بر روش سیمسن که در این بخش تشریح شد، به صورت زیر است:

```
function int = SimpAdapt(a,b,L,m,ep,f)
h = (b-a)/m; x = a:h:b;
int = 0;
for k=1:m
    s1 = simpson(x(k),x(k+1),2,f);
    s2 = simpson(x(k),x(k+1),4,f);
    e = abs(s2-s1)/15;
    if e<h/L*ep
        int = int+s2;
    else
        int = int + SimpAdapt(x(k),x(k+1),L,2,ep,f);
    end
end
```

در این تابع m در ابتدا تعداد تقسیمات اولیه‌ی بازه است، اما با توجه به اینکه این تابع خودش را فراخوانی می‌کند، m مقدار ۲ را هم در طول برنامه می‌گیرد. عدد L طول بازه یعنی $b-a$ است. از تابع simpson که قبلاً معرفی شده است هم استفاده می‌شود. این برنامه از نظر تعداد ارزیابی‌های تابع f بهینه نیست، زیرا همانگونه که در محاسبات دستی هم دیدید، مثلاً مقدار $f(0)$ چندین بار محاسبه می‌شود. بهتر است این برنامه به گونه‌ای اصلاح شود که مقدار f در هر نقطه تنها یک بار حساب شود و برای محاسبات بعدی حفظ شود. نوشتن برنامه‌ی اصلاح‌شده را بر عهده‌ی خواننده می‌گذاریم.

برای فراخوانی برنامه‌ی بالا برای مثالِ اخیر کافی است بنویسیم

```
f = @(x) exp(-10*x);
int = SimpAdapt(0,1,1,1,10^-4,f)
```

که پس از اجرا نتیجه‌ی زیر را ارائه خواهد داد

```
int =
   0.100013996957359
```

که در اینجا تا ۱۵ رقم اعشار نمایش داده شده است.

در آخر به این نکته اشاره می‌کنیم که فرمول تطبیقی بالا یکی از چندین روش انتگرال‌گیری تطبیقی است که تا به امروز طراحی شده‌اند. علاقه‌مندان می‌توانند به مراجع [۳، ۵] مراجعه کنند.

۵.۵ فرمول‌های گاوسی

در بخش ۱.۵ دیدیم که همواره می‌توان یک درونیاب- فرمول n نقطه‌ای با درجه دقت $n-1$ و گاهی هم درجه دقت n بدست آورد. از طرفی در بخش ۲.۵ مشاهده کردیم که با دستکاری کردن جایگاه نقاط انتگرال‌گیری می‌توان فرمول‌هایی با درجه دقت بالاتر هم بدست آورد و تعیین یک فرمول n نقطه‌ای با درجه دقت $2n-1$ امکان‌پذیر است که به آن فرمول گاوسی می‌گوییم. این اطلاعات بر پایه‌ی شهودی است که از روش ضرایب نامعین بدست آورده‌ایم، کما اینکه هنوز نمی‌دانیم آیا دستگاه معادلات غیرخطی که منجر به تعیین نقاط و گره‌های انتگرال‌گیری گاوس می‌شود بازای هر n دارای جواب (یکتا) هست. در این بخش با نگاهی دقیق‌تر و متفاوت از روش ضرایب نامعین به این موضوع می‌پردازیم. برای این منظور لازم است ابتدا چندجمله‌ایهای متعامد را معرفی کنیم.

چندجمله‌ایهای متعامد

تعامد نقشی اساسی در آنالیز عددی بازی می‌کند که در این بخش کوتاه نمی‌توان به شایستگی در مورد آن بحث کرد. اما یک کاربرد اساسیِ این مفهوم در بدست آوردن فرمول‌های انتگرال‌گیری را به مختصر توضیح خواهیم داد.

گوییم یک خانواده از چندجمله‌ایها مانند

$$\{p_0, p_1, \ldots, p_n, \ldots\}, \quad p_k \in \mathbb{P}_k$$

نسبت به تابع وزن w روی بازه‌ی $[a,b]$ متعامدند، اگر

$$\int_a^b w(x)p_j(x)p_k(x)dx = 0, \quad j \neq k.$$

تابع وزن w تابعی مثبت و پیوسته روی بازه‌ی باز (a,b) است.

به عنوان مثال چندجمله‌ایهای چبیشف که در بخش ۵.۳ به صورت

$$T_n(x) = \cos(n\cos^{-1}x), \quad n = 0, 1, 2, \ldots, \quad x \in [0, 1],$$

تعریف شدند، نسبت به وزن $w(x) = (1-x^2)^{-1/2}$ روی بازه‌ی $[-1, 1]$ متعامدند. زیرا با تغییر متغیر $x = \cos t$

می‌توان نوشت

$$\begin{aligned}\int_{-۱}^{۱} \frac{۱}{\sqrt{۱-x^۲}} T_j(x)T_k(x)dx &= \int_{-۱}^{۱} \frac{۱}{\sqrt{۱-x^۲}} \cos(k\cos^{-۱} x)\cos(k\cos^{-۱} x)dx \\ &= \int_۰^{\pi} \cos jt \cos kt\, dt = \frac{۱}{۲}\int_۰^{\pi} \big[\cos(j+k)t + \cos(j-k)t\big]dt \\ &= \begin{cases} ۰, & j \neq k, \\ \pi, & j = k = ۰, \\ \frac{\pi}{۲}, & j = k \neq ۰. \end{cases}\end{aligned}$$

همواره با داشتن تابع وزن w و بازه‌ی $[a,b]$ می‌توان به کمک روش زیر که به *الگوریتم گرم-اشمیت* مشهور است، یک خانواده از چندجمله‌ایهای متعامد تولید کرد: قرار می‌دهیم $p_۰(x) \equiv ۱$. فرض می‌کنیم $p_۱(x) = x - a_۰p_۰(x)$ و ضریب $a_۰$ را طوری تعیین می‌کنیم که $p_۰$ بر $p_۱$ عمود باشد. باید داشته باشیم

$$\int_a^b w(x)p_۰(x)p_۱(x) = \int_a^b w(x)(x-a_۰)dx = ۰,$$

که این هم نتیجه می‌دهد

$$a_۰ = \frac{\int_a^b xw(x)dx}{\int_a^b w(x)dx}.$$

با تعیین $a_۰$، چندجمله‌ای $p_۱$ تعیین می‌شود. این فرآیند را برای ساختن چندجمله‌ای‌های درجه بالاتر ادامه می‌دهیم. با استقرا فرض کنیم چندجمله‌ایهای متعامد

$$\{p_۰, p_۱, \ldots, p_\ell\}$$

ساخته شده‌اند و می‌خواهیم چندجمله‌ای بعدی یعنی $p_{\ell+۱}$ را بگونه‌ای بسازیم که بر تمام قبلی‌ها عمود باشد. برای این منظور قرار می‌دهیم

$$p_{\ell+۱}(x) = x^{\ell+۱} - a_۰p_۰(x) - \cdots - a_\ell p_\ell(x).$$

برای اینکه $p_{\ell+۱}$ بر p_k، $۰ \leqslant k \leqslant \ell$، عمود باشد باید داشته باشیم

$$\int_a^b w(x)p_{\ell+۱}(x)p_k(x)dx = \int_a^b w(x)\big(x^{\ell+۱} - a_۰p_۰(x) - \cdots - a_\ell p_\ell(x)\big)p_k(x) = ۰.$$

با توجه به اینکه طبق فرض استقرا p_k بر p_j، $j \neq k$، عمود است، همه انتگرال‌های سمت راست به غیر از ضریب a_k صفرند. پس داریم

$$a_k = \frac{\int_a^b w(x)x^{\ell+۱}p_k(x)dx}{\int_a^b w(x)p_k^۲(x)dx}, \quad k = ۰, ۱, \ldots, \ell.$$

قابل ذکر است که با ضرب کردن هر عضو یک خانواده‌ی متعامد در هر ضریب غیر صفر، تعامد بر هم نمی‌خورد. به همین علت معمولاً هر خانواده را به صورت مناسبی نرمال‌سازی می‌کنند.

مثال ۱۱.۵. می‌خواهیم یک خانواده از چندجمله‌ایهای متعامد $\{p_0, p_1, p_2, p_3\}$ نسبت به وزن $w(x) \equiv 1$ روی $[-1, 1]$ بسازیم. قرار می‌دهیم $p_0(x) \equiv 1$ و $p_1(x) = x - a_0 p_0(x)$. در این صورت

$$a_0 = \frac{\int_{-1}^{1} x p_0(x)dx}{\int_{-1}^{1} p_0^2(x)dx} = \frac{\int_{-1}^{1} x dx}{\int_{-1}^{1} dx} = 0$$

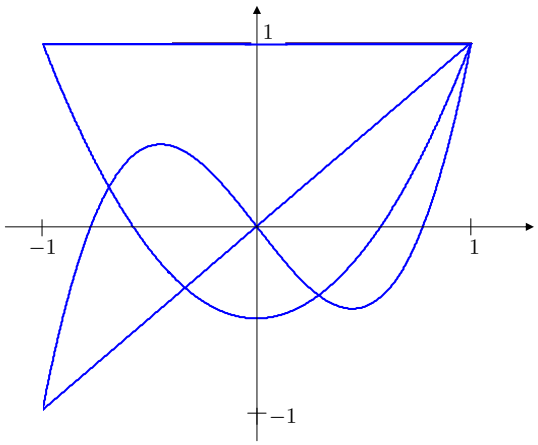
از این‌رو $p_1(x) = x$. همچنین تعریف می‌کنیم $p_2(x) = x^2 - a_0 p_0(x) - a_1 p_1(x)$ که در آن

$$a_0 = \frac{\int_{-1}^{1} x^2 p_0(x)dx}{\int_{-1}^{1} p_0^2(x)dx} = \frac{1}{3}, \quad a_1 = \frac{\int_{-1}^{1} x^2 p_1(x)dx}{\int_{-1}^{1} p_1^2(x)dx} = 0.$$

بنابراین $p_2(x) = x^2 - \frac{1}{3}$. اگر $n = 3$، با ادامه‌ی روند بالا داریم $p_3(x) = x^3 - \frac{3}{5}x$. چندجمله‌ایهای درجه بالاتر به همین ترتیب ساخته می‌شوند. چندجمله‌ایهای متعامد نسبت به وزن $w = 1$، چندجمله‌ایهای لژاندر نامیده می‌شوند. اگر چندجمله‌ایهای p_j بدست آمده در بالا را طوری نرمال‌سازی کنیم که مقدار آنها در $x = 1$ برابر ۱ باشد، چهار چندجمله‌ای ابتدایی لژاندر عبارتند از

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_2(x) &= \frac{3}{2}x^2 - \frac{1}{2}, \\ P_3(x) &= \frac{5}{2}x^3 - \frac{3}{2}x, \end{aligned} \tag{۳۱.۵}$$

که نمودار آن‌ها در شکل ۷.۵ ترسیم شده است. برای نمایش چندجمله‌ای لژاندر درجه n معمولاً نماد $P_n(x)$ استفاده می‌شود.



شکل ۷.۵: چندجمله‌ایهای لژاندر درجه صفر تا سه روی بازه‌ی $[-1, 1]$

می‌توان ثابت کرد (برای مثال فصل .. مرجع ... را ببینید) چندجمله‌ایهای لژاندر در رابطه‌ی بازگشتی سه جمله‌ای زیر صدق می‌کنند

$$P_{n+1}(x) = \frac{2n+1}{n+1} x P_n(x) - \frac{n}{n+1} P_{n-1}(x), \quad n \geqslant 1, \quad P_0(x) = 1, \quad P_1(x) = x. \qquad (۳۲.۵)$$

می‌توان نشان داد همه‌ی خانواده‌های چندجمله‌ایهای متعامد در یک رابطه‌ی بازگشتی سه جمله‌ای صدق می‌کنند. در بخش ۵.۳ رابطه‌ی بازگشتی چندجمله‌ایهای چبیشف را ارائه کردیم.

اکنون چند قضیه مهم در مورد چندجمله‌ایهای متعامد بیان می‌کنیم که می‌توانید اثبات قضیه‌ی اول را در فصل سوم [۶] ببینید. قضیه دوم به سادگی اثبات می‌شود، پرسش ۱۶ را ببینید.

قضیه ۱.۵. *فرض کنیم $\{p_0, p_1, \ldots\}$ یک خانواده از چندجمله‌ایهای متعامد نسبت به تابع وزن w روی بازه‌ی $[a,b]$ باشد. می‌دانیم که هر p_j دقیقاً از درجه‌ی j است. برای هر $j \geqslant 1$ چندجمله‌ای p_j دقیقاً j ریشه دارد که همگی حقیقی، متمایز و در بازه‌ی (a,b) هستند.*

قضیه ۲.۵. *مجموعه‌ی $n+1$ عضوی $\{p_0, p_1, \ldots p_n\}$ از یک خانواده‌ی چندجمله‌ایهای متعامد روی $[a,b]$، یک پایه برای $\mathbb{P}_n$ تشکیل می‌دهد.*

طبق قضیه‌ی بالا هر $q \in \mathbb{P}_n$ را می‌توان به صورت زیر روی بازه‌ی $[a,b]$ بسط داد

$$q(x) = c_0 p_0(x) + c_1 p_1(x) + \cdots + c_n p_n(x), \quad x \in [a,b], \quad c_k \in \mathbb{R},$$

که به آن یک *بسط متعامد* می‌گویند.

ارتباط فرمول‌های گاوسی با چندجمله‌ایهای متعامد

فرمول انتگرال‌گیری وزن‌دار زیر را در نظر می‌گیریم

$$\int_a^b w(x) f(x) dx = \sum_{k=1}^{n} \omega_k f(x_k) + E_n(f), \qquad (۳۳.۵)$$

که در آن اندیس سیگما بر خلاف قبل از یک شروع شده است تا یک فرمول n نقطه‌ای حاصل شود. می‌خواهیم ببینیم نقاط x_k و ضرایب ω_k را چگونه انتخاب کنیم تا یک فرمول از حداکثر درجه دقت چندجمله‌ای داشته باشیم. همه چیز در قضیه‌ی زیر خلاصه می‌شود.

قضیه ۳.۵. *فرض کنیم (۳۳.۵) یک درونیاب- فرمول باشد، یعنی $E_n(\mathbb{P}_{n-1}) = 0$. چندجمله‌ای*

$$\pi_n(x) := (x - x_1)(x - x_2) \cdots (x - x_n)$$

بر تمام اعضای $\mathbb{P}_{n-1}$ نسبت به وزن w روی $[a,b]$ عمود است، اگر و تنها اگر فرمول (۳۳.۵) دارای درجه دقت چندجمله‌ای $2n-1$ باشد، یعنی $E_n(\mathbb{P}_{2n-1}) = 0$.

برهان. ابتدا فرض کنیم فرمول (۳۳.۵) دارای درجه دقت $۲n-۱$ باشد. چون برای هر $p_{n-۱} \in \mathbb{P}_{n-۱}$ داریم $\pi_n p_{n-۱} \in \mathbb{P}_{۲n-۱}$، می‌توان نوشت

$$\int_a^b w(x)\pi_n(x)p_{n-۱}(x)dx = \sum_{k=۱}^{n} \omega_k \pi_n(x_k)p_{n-۱}(x_k) = ۰,$$

زیرا $\pi_n(x_k) = ۰$ برای $۱ \leqslant k \leqslant n$. این نشان می‌دهد π_n بر $\mathbb{P}_{n-۱}$ عمود است. بر عکس فرض کنیم $p_{۲n-۱} \in \mathbb{P}_{۲n-۱}$ چندجمله‌ای دلخواه باشد. تقسیم $p_{۲n-۱}$ بر π_n می‌دهد

$$p_{۲n-۱} = \pi_n q_{n-۱} + r_{n-۱}, \quad q_{n-۱}, r_{n-۱} \in \mathbb{P}_{n-۱}.$$

با انتگرال‌گیری از طرفین داریم

$$\begin{aligned}
\int_a^b w(x)p_{۲n-۱}(x)dx &= \int_a^b w(x)\pi_n(x)q_{n-۱}(x)dx + \int_a^b w(x)r_{n-۱}(x)dx \\
&= ۰ + \sum_{k=۰}^{n} \omega_k r_{n-۱}(x_k) \\
&= \sum_{k=۰}^{n} \omega_k \big[p_{۲n-۱}(x_k) - \pi_n(x_k)q_{n-۱}(x_k)\big] \\
&= \sum_{k=۰}^{n} \omega_k p_{۲n-۱}(x_k).
\end{aligned}$$

در سطر دوم انتگرال اول به دلیل عمود بودن π_n بر $\mathbb{P}_{n-۱}$ صفر است و تساوی به دلیل درونیاب-فرمول بودنِ (۳۳.۵) درست است. تساوی آخر طبق $\pi_n(x_k) = ۰$، $۱ \leqslant k \leqslant n$، برقرار است. معادله‌ی بالا اثبات را تمام می‌کند. □

قضیه‌ی بالا می‌گوید، فرمولِ n نقطه‌ایِ (۳۳.۵) از درجه دقت $۲n-۱$ است، اگر نقاط انتگرال‌گیری x_k ریشه‌های یک چندجمله‌ای درجه n مانند π_n باشد که بر تمام اعضای $\mathbb{P}_{n-۱}$ عمود است. فرض کنیم

$$\{p_۰, p_۱, \ldots, p_{n-۱}\}$$

یک خانواده از چندجمله‌ایهای متعامد نسبت به وزن w روی $[a,b]$ باشد. طبق قضیه‌ی ۲.۵، مجموعه‌ی بالا پایه‌ای برای $\mathbb{P}_{n-۱}$ است. بنابراین π_n باید بر تمام اعضای این پایه عمود باشد. پس π_n حتماً ضریبی از عضو بعدیِ این خانواده یعنی p_n است. پس نتیجه می‌گیریم:

نتیجه ۴.۵. *درونیاب-فرمولِ (۳۳.۵) از درجه دقت $۲n-۱$ است، اگر و تنها اگر نقاط درونیابی ریشه‌های چندجمله‌ای درجه n از خانواده‌ی متعامد نسبت به وزن w روی $[a,b]$ باشند.*

با مشخص شدن نقاط انتگرال‌گیری، وزن‌های ω_k را می‌توان به چند طریق بدست آورد. یک راه استفاده از روش ضرایب نامعین برای توابع $\{۱, x, \dots, x^{n-۱}\}$ است که منجر به حل یک دستگاه معادلات خطی با ماتریس واندرموند می‌شود، که قطعاً برای n های بزرگ این راه پیشنهاد نمی‌شود. روش دوم استفاده از پرسش ۵ فصل ۳ و فرمول (۲.۵) همین فصل است، که نتیجه می‌دهند

$$\omega_k = \int_a^b \frac{\pi_n(x)}{(x-x_k)\pi_n'(x_k)} w(x)dx, \quad k = ۱, ۲, \dots, n. \tag{۳۴.۵}$$

اما این روش هم بهترین روش نیست. به کمک برخی خواص چندجمله‌ایهای متعامد می‌توان روش‌هایی مبتنی بر مقادیر ویژه‌ی ماتریس‌های سه قطری، برای بدست آوردن نقاط و وزن‌های انتگرال‌گیری گاوس بدست آورد. علاقه‌مندان می‌توانند به فصل ششم [۶] مراجعه کنند.

فرمول گاوس-لژاندر

در فرمول گاوس-لژاندر تابع وزن، $w(x) \equiv ۱$ منظور می‌شود. پیش از هر چیز یادآور می‌شویم که انتگرال روی بازه‌ی دلخواه متناهی $[a, b]$ را می‌توان با تغییر متغیر

$$t = \frac{b-a}{۲}x + \frac{b+a}{۲}, \quad x \in [-۱, ۱], \tag{۳۵.۵}$$

به انتگرال روی بازه‌ی $[-۱, ۱]$ تبدیل کرد. داریم

$$\int_a^b g(t)dt = \frac{b-a}{۲}\int_{-۱}^{۱} f(x)dx, \quad f(x) = g\Big(\frac{b-a}{۲}x + \frac{b+a}{۲}\Big).$$

بنابراین بدون اینکه از کلیت کاسته شود، فرض کنیم بازه‌ی انتگرال‌گیری $[-۱, ۱]$ است و به دنبال فرمولی گاوسی به صورت

$$\int_{-۱}^{۱} f(x)dx = \sum_{k=۱}^{n} \omega_k f(x_k) + E_n(f),$$

هستیم که از درجه دقت $۲n-۱$ باشد. چندجمله‌ایهای متعامد نسبت به وزن $w(x) \equiv ۱$ روی $[-۱, ۱]$، چندجمله‌ایهای لژاندر می‌باشند که چندتای آن‌ها در (۳۱.۵) معرفی شدند و بقیه هم طبق فرمول بازگشتی (۳۲.۵) قابل محاسبه هستند. طبق نتیجه‌ی ۴.۵ همین بخش، نقاط x_k باید ریشه‌های چندجمله‌ای لژاندر $P_n(x)$ روی $[-۱, ۱]$ باشند. اگر ω_k هم با یک روش مناسب تعیین شوند، فرمول بالا یک فرمول گاوسی خواهد بود که به آن فرمول *گاوس-لژاندر* می‌گوییم.

مثال ۱۲.۵. فرمول‌های گاوس-لژاندرِ یک، دو، و سه نقطه‌ای را بدست آورید. در فرمول یک نقطه‌ای، $x_۱$ ریشه‌ی $P_۱(x) = x$ است، یعنی $x_۱ = ۰$. همچنین طبق فرمول (۳۴.۵) داریم

$$\omega_۱ = \int_{-۱}^{۱} \frac{P_۱(x)}{(x-x_۱)P_۱'(x_۱)} = \int_{-۱}^{۱} dx = ۲.$$

پس فرمول گاوس-لژاندر یک نقطه‌ای به صورت زیر نوشته می‌شود

$$\int_{-۱}^{۱} f(x)dx = ۲f(۰) + E_۱(f),$$

که همان فرمول نقطه میانی است که از درجه دقت یک (خطی) است. در فرمول دو نقطه‌ای نقاط انتگرال‌گیری ریشه‌های $P_۲(x) = \frac{۳}{۲}x^۲ - \frac{۱}{۲}$ می‌باشند که عبارتند از

$$x_۱ = -\frac{۱}{\sqrt{۳}}, \quad x_۲ = \frac{۱}{\sqrt{۳}}.$$

وزن‌های انتگرال‌گیری $\omega_۱$ و $\omega_۲$ طبق فرمول (۳۴.۵) به صورت زیر محاسبه می‌شوند

$$\omega_۱ = \int_{-۱}^{۱} \frac{\frac{۳}{۲}x^۲ - \frac{۱}{۲}}{(x + \frac{۱}{\sqrt{۳}})\frac{-۳}{\sqrt{۳}}}dx = ۱, \quad \omega_۲ = \int_{-۱}^{۱} \frac{\frac{۳}{۲}x^۲ - \frac{۱}{۲}}{(x - \frac{۱}{\sqrt{۳}})\frac{۳}{\sqrt{۳}}}dx = ۱.$$

انتگرال‌های بالا به سادگی محاسبه می‌شوند، زیرا $x - x_k$ در مخرج، یک عاملِ صورت است. پس فرمول گاوس-لژاندر دو نقطه‌ای به صورت زیر است و درجه دقت آن سه است

$$\int_{-۱}^{۱} f(x)dx = f\Big(\frac{-۱}{\sqrt{۳}}\Big) + f\Big(\frac{۱}{\sqrt{۳}}\Big) + E_۲(f).$$

نقاط انتگرال‌گیری فرمول گاوس-لژاندر سه نقطه‌ای ریشه‌های $P_۳(x) = \frac{۵}{۲}x^۳ - \frac{۳}{۲}x$ هستند که عبارتند از

$$x_۱ = -\sqrt{\frac{۳}{۵}} \doteq -۰/۷۷۴۵۹۶۶۶۹۲, \quad x_۲ = ۰, \quad x_۳ = \sqrt{\frac{۳}{۵}} \doteq ۰/۷۷۴۵۹۶۶۶۹۲.$$

با محاسباتی مشابه قبل ضرایب انتگرال‌گیری به صورت زیر حاصل می‌شوند

$$\omega_۱ = \omega_۳ = \frac{۵}{۹} \doteq ۰/۵۵۵۵۵۵۵۵۵۶, \quad \omega_۲ = \frac{۸}{۹} \doteq ۰/۸۸۸۸۸۸۸۸۸۹,$$

پس فرمول گاوس-لژاندر سه نقطه‌ای که از درجه دقت پنج است، به صورت زیر است

$$\int_{-۱}^{۱} f(x)dx = \frac{۵}{۹}f\Big(-\sqrt{\frac{۳}{۵}}\Big) + \frac{۸}{۹}f(۰) + \frac{۵}{۹}f\Big(\sqrt{\frac{۳}{۵}}\Big) + E_۳(f).$$

$\diamondsuit$

به همین ترتیب می‌توان فرمول‌های بعدی را هم ساخت. اگر n افزایش یابد، تعیین ریشه‌های $P_n(x)$ و ضرایب ω_k مشکل و پرهزینه خواهد شد. همانطور که قبلاً هم اشاره کردیم، در عمل روشِ دیگری برای بدست آوردن ریشه‌های $P_n(x)$ و ضرایب ω_k استفاده می‌شود که برای n بزرگ بسیار کم هزینه و کارا است و در فصل ششم [۶] آمده است. در اینجا تنها جدول ضرایب و نقاط انتگرال‌گیری گاوس-لژاندر برای چند مقدار n را ارائه می‌دهیم. نتایج تا ۱۶ رقم دهدهی در جدول ۳.۵ ارائه شده‌اند.

جدول ۳.۵: نقاط و ضرایب انتگرال‌گیری گاوس-لژاندر

n	x_k	ω_k
۱	۰	۲
۲	±۰/۵۷۷۳۵۰۲۶۹۱۸۹۶۲۶	۱
۳	۰	۰/۸۸۸۸۸۸۸۸۸۸۸۸۸۹
	±۰/۷۷۴۵۹۶۶۶۹۲۴۱۴۸۳	۰/۵۵۵۵۵۵۵۵۵۵۵۵۵۶
۴	±۰/۸۶۱۱۳۶۳۱۱۵۹۴۰۵۳	۰/۳۴۷۸۵۴۸۴۵۱۳۷۴۵۴
	±۰/۳۳۹۹۸۱۰۴۳۵۸۴۸۵۶	۰/۶۵۲۱۴۵۱۵۴۸۶۲۵۴۶
۵	۰	۰/۵۶۸۸۸۸۸۸۸۸۸۸۸۹
	±۰/۹۰۶۱۷۹۸۴۵۹۳۸۶۶۴	۰/۲۳۶۹۲۶۸۸۵۰۵۶۱۸۹
	±۰/۵۳۸۴۶۹۳۱۰۱۰۵۶۸۳	۰/۴۷۸۶۲۸۶۷۰۴۹۹۳۶۶
۶	±۰/۹۳۲۴۶۹۵۱۴۲۰۳۱۵۲	۰/۱۷۱۳۲۴۴۹۲۳۷۹۱۷۰
	±۰/۶۶۱۲۰۹۳۸۶۴۶۶۲۶۴	۰/۳۶۰۷۶۱۵۷۳۰۴۸۱۳۸
	±۰/۲۳۸۶۱۹۱۸۶۰۸۳۱۹۷	۰/۴۶۷۹۱۳۹۳۴۵۷۲۶۹۱
۷	۰	۰/۴۱۷۹۵۹۱۸۳۶۷۳۴۶۹
	±۰/۹۴۹۱۰۷۹۱۲۳۴۲۷۵۸	۰/۱۲۹۴۸۴۹۶۶۱۶۸۸۷۰
	±۰/۷۴۱۵۳۱۱۸۵۵۹۹۳۹۴	۰/۲۷۹۷۰۵۳۹۱۴۸۹۲۷۷
	±۰/۴۰۵۸۴۵۱۵۱۳۷۷۳۹۷	۰/۳۸۱۸۳۰۰۵۰۵۰۵۱۱۹
۸	±۰/۹۶۰۲۸۹۸۵۶۴۹۷۵۳۶	۰/۱۰۱۲۲۸۵۳۶۲۹۰۳۷۶
	±۰/۷۹۶۶۶۴۷۷۴۱۳۶۲۷	۰/۲۲۲۳۸۱۰۳۴۴۵۳۳۷۴
	±۰/۵۲۵۵۳۲۴۰۹۹۱۶۳۲۹	۰/۳۱۳۷۰۶۶۴۵۸۷۷۸۸۷
	±۰/۱۸۳۴۳۴۶۴۲۴۹۵۶۵۰	۰/۳۶۲۶۸۳۷۸۳۳۷۸۳۶۲

در مورد این فرمول‌ها ذکر چند نکته ضروری است. اول اینکه نقاط انتگرال‌گیری به صورت متقارن در بازه‌ی $[-۱, ۱]$ پخش شده‌اند و چگالی آن‌ها در نزدیکی دو انتهای بازه بیشتر است. نقاط ۱ و -۱ جزء نقاط انتگرال‌گیری نیستند و از این رو این فرمول‌ها از نوع فرمول‌های باز محسوب می‌شوند. نقطه‌ی صفر در فرمول‌های فرد وجود دارد و در فرمول‌های زوج وجود ندارد. اینها همگی از خصوصیات ریشه‌های چندجمله‌ایهای متعامد لژاندر هستند. وزن‌های انتگرال‌گیری هم متقارنند و نکته‌ی بسیار مهم در مورد آن‌ها این است که همگی مثبت هستند. اثبات این مهم، در حالت کلی چندان مشکل نیست. فرض کنیم $\ell_j(x)$، $۱ \leqslant j \leqslant n$، چندجمله‌ایهای لاگرانژ درجه $n-۱$ مبتنی بر ریشه‌های $P_n(x)$ باشند. از آنجا که درجه دقت فرمول گاوس n نقطه‌ای برابر $۲n-۱$ است، این فرمول برای انتگرال‌گیری از $\ell_j^۲(x)$ دقیق است. پس داریم

$$۰ < \int_{-۱}^{۱} \ell_j^۲(x)dx = \sum_{k=۱}^{n} \omega_k \ell_j^۲(x_k) = \omega_j, \quad j = ۱, ۲, \ldots, n.$$

تساوی آخر به دلیل خاصیت دلتای کرونکر چندجمله‌ایهای لاگرانژ برقرار است.

مثال ۱۳.۵. مقدار انتگرال

$$I = \int_{-۱}^{۱} \frac{۱}{۱+x^۲}dx$$

را با فرمول‌های گاوس-لژاندر محاسبه می‌کنیم و با توجه به مقدار دقیق $I = \frac{\pi}{۲}$ خطا را در هر حالت گزارش می‌دهیم. در مثال ۸.۵ فرمول‌های یک نقطه‌ای و دو نقطه‌ایِ گاوس-لژاندر را امتحان کردیم. با فرمول گاوس سه نقطه‌ای داریم

$$I \approx G_۳ = \frac{۵}{۹} \times \frac{۱}{۱+۳/۵} + \frac{۸}{۹} \times \frac{۱}{۱+۰} + \frac{۵}{۹} \times \frac{۱}{۱+۳/۵} \doteq ۱/۵۸۳۳, \quad \text{خطا} = ۰/۰۱۲۵.$$

برای فرمول‌های درجه بالاتر می‌توان از یک برنامه‌ی کامپیوتری استفاده کرد. مثلاً برای فرمول شش نقطه‌ای برنامه‌ی زیر را می‌نویسیم:

```
x = [-0.932469514203152 -0.661209386466264 -0.238619186083197 ...
   0.238619186083197 0.661209386466265 0.932469514203152];
w = [0.171324492379170 0.360761573048138 0.467913934572691 ...
   0.467913934572692 0.360761573048138 0.171324492379171];
f = 1./(1+x.^2);
G6 = f*w'
err = abs(pi/2-G6)
```

با اجرای این برنامه مقدار انتگرال 1.5707 و خطای 6.4619e-05 در خروجی چاپ می‌شوند. اگر برای محاسبه‌ی این انتگرال فرمول‌های نیوتن-کاتس مرکب با تعداد نقاط مشابه را به کار برید، خواهید دید که فرمول‌های گاوسی بسیار دقیق‌ترند. برای مثال فرمول ذورنقه‌ای مرکب با شش نقطه به صورت زیر فراخوانی می‌شود

```
T = trapez(-1,1,5,@(x) 1./(1+x.^2));
err = abs(pi/2-T)
```

که در خروجی خطای 0.0133 را چاپ می‌کند، که بسیار نادقیق‌تر از فرمول گاوس شش نقطه‌ای است. ♢

مثال ۱۴.۵. تقریبی از $I = \int_{۰}^{۳} \frac{\sin t}{t} dt$ به کمک فرمول گاوس-لژاندر سه نقطه‌ای بدست می‌آوریم. اگر تغییر متغیر (۳۵.۵) را اعمال کنیم، داریم

$$\int_{۰}^{۳} \frac{\sin t}{t} dt = \frac{۳}{۲} \int_{-۱}^{۱} \underbrace{\frac{\sin(\frac{۳}{۲}x + \frac{۳}{۲})}{\frac{۳}{۲}x + \frac{۳}{۲}}}_{f(x)} dx.$$

با بکارگیری فرمول سه نقطه‌ای روی انتگرال طرف راست داریم

$$I \approx \frac{۳}{۲} \left[\frac{۵}{۹} f\Big(-\sqrt{\frac{۳}{۵}} \Big) + \frac{۸}{۹} f(۰) + \frac{۵}{۹} f\Big(\sqrt{\frac{۳}{۵}} \Big) \right] \doteq ۱/۸۴۸۶۸۹۶۷۳.$$

این انتگرال تابع اولیه ندارد، پس اطلاعی از اندازه‌ی خطا نداریم. برای اطمینان بهتر است از فرمول‌های با درجه بالاتر استفاده شود. این کار به عهده‌ی شما واگذار می‌شود، تنها اشاره می‌کنیم که اگر فرمول گاوس-لژاندر هفت نقطه‌ای را به کار گیرید مقدار تقریبی ۱/۸۴۸۶۵۲۵۲۸ را بدست می‌آورید که با مقدار فرمول سه نقطه تا چهار رقم اعشار یکسان است. ♢

فرمول گاوس-چبیشف

فرمول گاوس-چبیشف برای محاسبه‌ی تقریبی انتگرال‌هایی به فرم

$$\int_{-۱}^{۱} \frac{f(x)}{\sqrt{۱ - x^۲}} dx$$

به کار می‌رود که در آن‌ها تابع وزن $w(x) = (۱ - x^۲)^{-۱/۲}$ در انتگرالده ظاهر شده است. قبلاً دیدیم که چندجمله‌ایهای متعامد نسبت به این وزن، چندجمله‌ایهای چبیشف می‌باشند. بنابراین در فرمول گاوس-چبیشف، نقاط x_k ریشه‌های چندجمله‌ای $T_n(x)$ هستند، که در بخش ۵.۳ دیدیم با

$$x_k = \cos\left(\frac{۲k - ۱}{۲n} \pi \right), \quad k = ۱, ۲, \ldots, n,$$

داده می‌شوند. از طرفی با توجه به تعریفِ چندجمله‌ایهای چبیشف و با استفاده از فرمول (۳۴.۵) می‌توان ثابت کرد، وزن‌های ω_k همگی یکسان و با

$$\omega_k = \frac{\pi}{n}, \quad k = ۱, ۲, \ldots, n,$$

داده می‌شوند. بنابراین فرمول گاوس-چبیشف به صورت زیر است

$$\int_{-۱}^{۱} \frac{f(x)}{\sqrt{۱-x^۲}}dx = \frac{\pi}{n}\sum_{k=۱}^{n} f(x_k) + E_n(f).$$

می‌دانیم که این فرمول برای چندجمله‌ایهای حداکثر درجه $۲n-۱$ دقیق است، یعنی اگر $f \in \mathbb{P}_{۲n-۱}$، آنگاه $E_n(f) = ۰$.

تابع وزن $w(x) = (۱-x^۲)^{-۱/۲} = (۱-x)^{-۱/۲}(۱+x)^{-۱/۲}$، در نقاط انتهایی بازه‌ی $[-۱, ۱]$ دارای تکینگی جبری است. در حالت کلی‌تر یک انتگرال به صورت

$$\int_a^b (b-t)^{-۱/۲}(t-a)^{-۱/۲} f(t)dt \qquad (۳۶.۵)$$

با a و b متناهی که دارای تکینگی جبری از مرتبه‌ی $-۱/۲$ در ابتدا و انتهای بازه است را می‌توان به کمک فرمول گاوس-چبیشف محاسبه کرد. برای این منظور با اعمال تغییر متغیر (۳۵.۵) می‌توان نشان داد

$$\int_a^b (b-t)^{-۱/۲}(t-a)^{-۱/۲} f(t)dt = \int_{-۱}^{۱} \frac{g(x)}{\sqrt{۱-x^۲}}dx, \quad g(x) = f\left(\frac{b-a}{۲}x + \frac{b+a}{۲}\right).$$

اثبات به عنوان تمرین واگذار می‌شود. بنابراین برنامه‌ی روش گاوس-چبیشف برای محاسبه‌ی یک انتگرال به فرم (۳۶.۵) به صورت زیر است.

```
function int = GaussCheb (a,b,n,f)
x = cos((2*(1:n)-1)*pi/(2*n));
fx = f((b-a)/2*x+(b+a)/2);
if length(fx)==1 fx = fx*ones(n,1); end
int = pi/n*sum(fx);
```

سطر چهارم برنامه برای سازگار کردن آن برای توابع ثابت نوشته شده است. در اینجا یک مثال ارائه می‌دهیم که در آن همگرایی سریع فرمول گاوس-چبیشف به تصویر کشیده می‌شود.

مثال ۱۵.۵. انتگرال زیر را در نظر بگیرید،

$$I = \int_{-۱}^{۱} \frac{\cos x}{\sqrt{۱-x^۲}}dx.$$

مقدار دقیق این انتگرال $\pi J_{۰}(۱) \doteq$ ۲٫۴۰۳۹۳۹۴۳۰۶۳۴۴۱۳ است که در آن $J_{۰}(x)$ تابع بسل نوع اول مرتبه صفر است. فرمول‌های گاوس چبیشف برای تقریب این انتگرال به صورت زیر نوشته می‌شوند

$$G_{۱} = \pi \cos(\cos\frac{\pi}{۲}) \doteq ۳٫۱۴۱۵۹۲۶۵۳۵۸۹۷۹۳,$$
$$G_{۲} = \frac{\pi}{۲}[\cos(x_{۱}) + \cos(x_{۲})] \doteq ۲٫۳۸۸۳۷۸۸۴۱۱۰۴۱۳۲, \quad x_{۱} = \cos\frac{\pi}{۴}, \quad x_{۲} = \cos\frac{۳\pi}{۴}.$$

انجام محاسبات به صورت دستی، خسته کننده است و بهتر است از برنامه‌ی روش استفاده کنیم. برای این منظور برنامه‌ی زیر را می‌نویسیم که در آن تابع GaussCheb فراخوانی شده است.

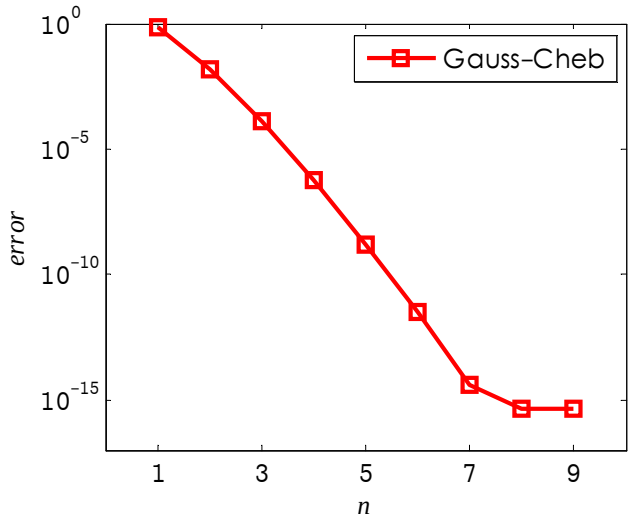
```
f = @(x) cos(x); I = pi*besselj(0,1);
for n=1:9
   int = GaussCheb(-1,1,n,f)
   err(n) = abs(int-I);
end
semilogy(1:n,err,'-sr')
```

این برنامه مقادیر را تا $n = ۹$ محاسبه می‌کند و هر بار خطا را بدست می‌آورد. در آخر نمودار خطا با دستور semilogy رسم می‌شود که محور y (مقادیر خطا) را به صورت لگاریتمی مقیاس می‌کند. در این دستور، محور x (مقادیر n) به صورت معمولی (خطی) است. این نمودار در شکل ۸.۵ رسم شده است.

مشاهده می‌کنید که خطا به سرعت کاهش می‌یابد و برای $n = ۸$ به سطح دقت ماشین (در اینجا دقت دوبرابر) می‌رسد.

اکنون می‌خواهیم مقایسه‌ای با فرمول‌های نیوتن-کاتس مرکب داشته باشیم. برای محاسبه‌ی این انتگرال فرمول‌های بسته قابل استفاده نیستند، زیرا انتگرالده در دو نقاط ± ۱ تعریف نشده است. از این رو فرمول نقطه میانی مرکب را بکار می‌بریم.

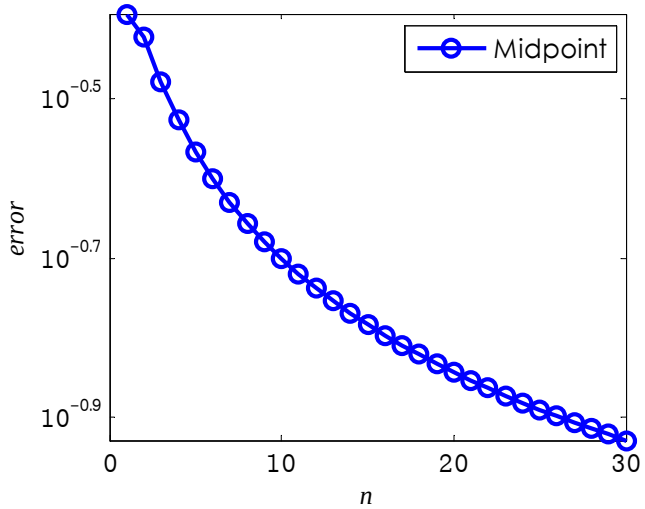
```
f = @(x) cos(x)./sqrt(1-x.^2); I = pi*besselj(0,1);
for n=1:30
    int=midpoint(-1,1,n,f);
    err(n) = abs(int-I);
```



شکل ۸.۵: نمودار خطای روش گاوس-چبیشف برای محاسبه انتگرال مثال ۱۵.۵

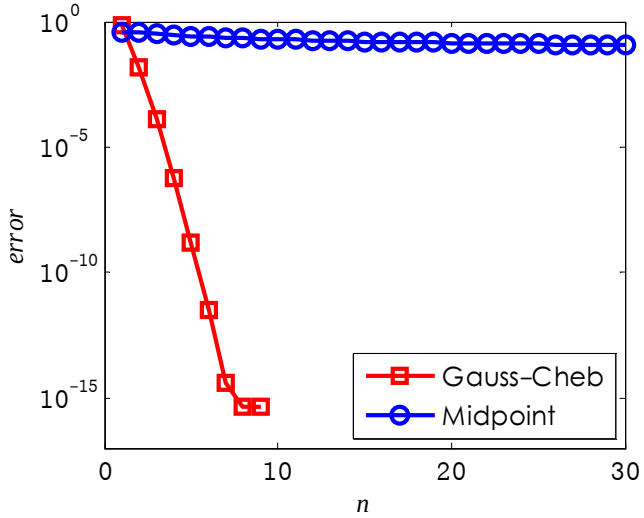
```
end
semilogy(1:n,err,'-ob')
```

نمودار خطا در شکل ۹.۵ رسم شده است. مشاهده می‌کنیم با اینکه تا مقدار $n = ۳۰$ پیش رفته‌ایم، اما خطا به کندی کاهش می‌یابد و به سختی به $۱۰^{-۱}$ می‌رسد. برای مقایسه، هر دو نمودار را در یک شکل (شکل ۱۰.۵) رسم می‌کنیم تا تفاوت دو فرمول بهتر مشاهده شود.



شکل ۹.۵: نمودار خطای روش نقطه میانی برای محاسبه انتگرال مثال ۱۵.۵

ملاحظه ۴.۵. برای ترسیم شکل در متلب، دستور `plot` هر دو محور را خطی مقیاس می‌کند، دستور `loglog` هر دو



شکل ۱۰.۵: نمودارهای خطای روش‌های نقطه میانی و گاوس-چبیشف برای محاسبه انتگرال مثال ۱۵.۵

محور را لگاریتمی و دستور `semilogx` محور x را لگاریتمی و محور y را خطی مقیاس می‌کند. دستور `semilogy` محور y را لگاریتمی و محور x را خطی مقیاس می‌کند.

$\diamondsuit$

دیگر فرمول‌های گاوسی

به ذکر این نکته بسنده می‌کنیم که برای هر تابع وزن دلخواه می‌توان فرمول‌های گاوسی را بدست آورد، که در دو بخش قبل دو نوع آن‌ها را بررسی کردیم. برخی توابع وزن مشهور عبارتند

$$
\begin{aligned}
&w(x) = (1-x)^{\alpha}(1+x)^{\beta}, \quad \alpha,\beta > -1, \quad x \in [-1,1], && \text{وزن ژاکوبی}\\
&w(x) = e^{-x}, \quad x \in [0,\infty), && \text{وزن لاگر}\\
&w(x) = e^{-x^2}, \quad x \in (-\infty,\infty), && \text{وزن ارمیت}
\end{aligned}
$$

در هر حالت می‌توان چندجمله‌ایهای متعامد نظیر را تعیین و فرمول‌های گاوسی را بدست آورد. فرمول‌های گاوس-لژاندر و گاوس-چبیشف حالت خاصی از فرمول‌های گاوس-ژاکوبی هستند، زیرا تابع وزن لژاندر بازای $\alpha = \beta = 0$ و تابع وزن چبیشف بازای $\alpha = \beta = -1/2$ بدست می‌آیند. فرمول گاوس-لاگر برای انتگرال‌گیری روی بازه‌ی $[0,\infty)$ و فرمول گاوس-ارمیت برای انتگرال‌گیری روی $(-\infty,\infty)$ بکار می‌روند. برای مشاهده‌ی توضیحات بیشتر در این زمینه می‌توانید به فصل ششم [۶] مراجعه کنید.

۶.۵ پرسش‌ها

۱. نشان دهید اگر تابع f روی $[a,b]$ پیوسته باشد و تابع g روی $[a,b]$ تغییر علامت ندهد (یا نامنفی باشد یا نامثبت) آنگاه وجود دارد $\xi \in [a,b]$ بطوریکه

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

۲. نشان دهید اگر $f \in C[a,b]$ و ξ_1، ξ_2، ...، ξ_n در $[a,b]$ واقع باشند، آنگاه $\xi \in [a,b]$ وجود دارد بطوریکه

$$f(\xi) = \frac{1}{n}\sum_{k=1}^{n} f(\xi_k).$$

۳. نشان دهید اگر یک فرمول انتگرال‌گیری برای توابع ثابت دقیق باشد، آنگاه

$$\sum_{k=0}^{n} \omega_k = b - a.$$

۴. با توجه به اینکه فرمول نقطه میانی برای اعضای $\mathbb{P}_1$ دقیق است، جمله‌ی خطا را با فرضِ اینکه ضریبی از $f''(\xi)$ است بدست آورید.

۵. فرض کنید مقادیر تابع f و مشتق آن در نقاط هم‌فاصله‌ی $x_0 = a$، x_1، ...، $x_n = b$ داده شده‌اند. فرمول ذوزنقه‌ای اصلاح‌شده‌ی مرکب را بدست آورید و جمله‌ی خطای آن را تعیین کنید.

۶. فرض کنید $T(h)$، $M(h)$ و $S(h)$ به ترتیب فرمول‌های ذوزنقه‌ای، نقطه میانی و سیمسن با طول گام h باشند. نشان دهید

$$T\left(\frac{h}{2}\right) = \frac{1}{2}\big(T(h) + M(h)\big),$$
$$S\left(\frac{h}{2}\right) = \frac{1}{3}\big(T(h) + 2M(h)\big),$$
$$S\left(\frac{h}{2}\right) = \frac{1}{3}\big(4T(h/2) - T(h)\big).$$

قسمت سوم چه نکته‌ای را در مورد ستون دوم جدول رامبرگ یادآوری می‌کند؟

۷. خطای فرمول نقطه میانی را به کمک بسط تیلر دوباره بدست آورید.

۸. فرمول انتگرالگیری ساده سیمسُن تابع $f(x)$ روی بازهٔ $[x_0, x_2]$ با نقطه مرکزی $x_1 = (x_2 + x_1)/2$ را در نظر بگیرید. با توجه به توضیحات زیر، خطای این فرمول را به روش دیگری بدست آورید. با توجه به فرمول خطای

درونیابی فرم نیوتن داریم

$$E(f) = \int_{x_0}^{x_2} (x-x_0)(x-x_1)(x-x_2) f[x_0, x_1, x_2, x]dx.$$

فرض کنیم $f \in C^4[x_0, x_2]$، و تعریف کنیم

$$v(x) = \int_{x_0}^{x} (t-x_0)(t-x_1)(t-x_2)dt,$$

نشان دهید $v(x) \geqslant 0$ و $v(x_0) = v(x_2) = 0$. سپس قرار دهید

$$E(f) = \int_{x_0}^{x_2} v'(x) f[x_0, x_1, x_2, x]dx,$$

و با انتگرال‌گیری جزء به جزء نشان دهید

$$E(f) = -\frac{h^5}{90} f^{(4)}(\xi), \quad \xi \in [x_0, x_2], \quad h = \frac{x_2 - x_0}{2}.$$

۹. برای محاسبه‌ی $\int_{-1}^{1} \frac{1}{1+x^2}$، به ترتیب در روشهای ذوزنقه‌ای مرکب و سیمسُن مرکب بایستی بازه انتگرالگیری به چند زیربازه تقسیم شود تا جواب بدست آمده دارای خطای حداکثر $\varepsilon = 10^{-5}$ باشد؟

۱۰. فرمول سیمسن اصلاح شده روی بازه‌ی $[-h, h]$ با اضافه کردن مشتقات تابع در دو انتها به صورت زیر است

$$\int_{-h}^{h} f(x)dx = h\big[\alpha f(-h) + \beta f(0) + \alpha f(h)\big] + h^2\gamma\big[f'(-h) - f'(h)\big] + E(f),$$

که برای چندجمله‌ایهای تا درجه ۵ دقیق است.

الف: وزن‌های α، β و γ را با استفاده از توابع $f(x) = 1, x^2, x^4$ بدست آورید. تابع $f(x) = x^6$ را برای تعیین جمله‌ی خطا استفاده کنید.

ب: نشان دهید فرمول مرکبِ متناظر روی بازه‌ی $[a, b]$ با $h = (b-a)/n$ (n زوج) ، فقط شامل مشتقات انتهایی یعنی $f'(a)$ و $f'(b)$ است.

۱۱. برنامه‌ی فرمول ذوزنقه‌ای مرکب را روی

$$\int_0^{2\pi} \sin x\, dx, \quad \frac{1}{2\pi}\int_0^{2\pi} e^{\frac{1}{\sqrt{2}}\sin x} dx,$$

اجرا کنید. آنچه در مورد خطاها مشاهده می‌کنید را گزارش کنید. مقدار دقیق انتگرال اول صفر و مقدار دقیق انتگرال دوم $I_0(\frac{1}{\sqrt{2}})$ است که $I_\alpha(x)$ تابع بسل اصلاح‌شده‌ی نوع اول مرتبه صفر است. (در متلب از دستور `besseli(0,1/sqrt(2))` استفاده کنید). در اینجا فقط اشاره می‌کنیم که فرمول ذوزنقه‌ای برای توابع متناوب روی بازه‌ی تناوبشان به طور شگفت‌انگیزی جواب‌های دقیق ارائه می‌دهد. علت این امر را در دروس پیشرفته‌تر خواهید یافت.

۱۲. برای تقریبِ $\int_{x_0}^{x_1} f(x)\,dx$، فرض کنید $p_2(x)$ چندجمله‌ای درجه دومی است که f را در x_0، x_1 و نقطه دیگر x_2 که $x_0 < x_1 < x_2$ درونیابی می‌کند. در حالتی که نقاط هم‌فاصله هستند نشان دهید

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{h}{12}[5f_0 + 8f_1 - f_2] + E(f),$$

و به کمک خطای درونیابی نشان دهید

$$E(f) = \frac{h^4}{4!}f'''(\xi), \quad x_0 \leqslant \xi \leqslant x_2.$$

۱۳. روش ضرایب نامعین را برای ساختن فرمولی به شکل

$$\int_0^1 f(x)\,dx = af(0) + bf(1) + cf''(\alpha) + E(f)$$

که دارای ماکزیمم درجه دقت باشد به کار برید.

۱۴. فرمول انتگرالگیری به شکل

$$\int_{-1}^{1} f(x)dx = \alpha_{-1}\int_{-1}^{-1/2} f(x)dx + \alpha_0 f(0) + \alpha_1 \int_{1/2}^{1} f(x)dx + E(f)$$

بسازید که دارای ماکزیمم درجه دقت باشد. ماکزیمم درجه دقت چقدر است؟

۱۵. با روش انتگرال‌گیری تطبیقی سیمسن، تقریبی از

$$\int_0^1 \frac{1}{0/01 + x^2}dx$$

با خطای $\varepsilon = 0/01$ بدست آورید.

۱۶. قضیه‌ی ۲.۵ را اثبات کنید.

۱۷. به کمک الگوریتم گرم-اشمیت چندجمله‌ایهای متعامد تا درجه‌ی سه نسبت به وزن $w(x) = \ln\frac{1}{x}$ روی $[0, 1]$ را بسازید. به کمک آن فرمول‌های گاوس یک نقطه‌ای، دو نقطه‌ای و سه نقطه‌ای را برای انتگرال

$$\int_0^1 \ln\frac{1}{x} f(x)dx$$

بسازید و به کمک آن‌ها مقدار انتگرال را بازای $f(x) = x^4$ محاسبه کنید. راهنمایی: در محاسبات از $\int_0^1 x^k \ln x dx = -(1+k)^{-2}$ استفاده کنید.

۱۸. فرمول گاوس دو نقطه‌ای با وزن $w(x) = x^{-1/2}$ در بازه $[0, 1]$ بسازید و کران خطای آن را برای $f \in C^4[0, 1]$ بدست آورید.

۱۹. به کمک فرمول گاوس-چبیشف نشان دهید مساحت دایره ای به شعاع واحد برابر π است.

فصل ۶

حل معادلات غیرخطی

در این فصل مسایلی را مورد بررسی قرار می‌دهیم که به‌صورت کلی

$$f(x) = ۰ \qquad (۱.۶)$$

نوشته می‌شود اما با توجه به معنای x و f ممکن است تعبیرهای مختلفی داشته باشد. ساده‌ترین حالت آن است که f یک تابع حقیقی تک‌متغیره از x باشد و ما تلاش کنیم مقادیری از متغیر x را بیابیم که در آن‌ها f صفر شود. چنین مقادیری، ریشه‌های معادله‌ی (۱.۶) یا صفرهای تابع f نامیده می‌شوند. اگر x در (۱.۶) یک بردار به صورت $x = [x_۱, x_۲, \ldots, x_n]^T \in \mathbb{R}^n$ باشد و f نیز یک بردار که هر مولفه‌ی آن یک تابع از x است، باشد آن‌گاه (۱.۶) یک دستگاه معادلات را نمایش می‌دهد. این دستگاه معادلات، غیرخطی گفته می‌شود هرگاه دست‌کم یکی از مولفه‌های f به دست‌کم یکی از متغیرهای $x_۱$، $x_۲$، ... ، x_n به طور غیرخطی وابسته باشد. اگر تمام مولفه‌های f توابعی خطی از $x_۱$، $x_۲$، ... ، x_n باشند، (۱.۶) یک دستگاه از معادلات جبری خطی نامیده می‌شود که بحث جالبی است و درس "جبرخطی عددی" مورد بررسی قرار می‌گیرد. باز خیلی کلی‌تر، (۱.۶) می‌تواند یک معادله‌ی تابعی را نمایش دهد، اگر x یک عنصر در یک فضای تابعی و f یک عملگر (خطی یا غیرخطی) باشد که روی این فضا اثر می‌کند. در هر کدام از حالت‌های توصیف شده، صفر سمت راست (۱.۶) معنی متفاوتی دارد: عدد صفر در حالت اول، بردار صفر در حالت دوم، و تابع متحد با صفر در حالت آخر.

بیشتر مباحث این فصل به بررسی معادلات غیرخطی تک‌متغیره اختصاص دارد. این معادلات اغلب در تجزیه و تحلیل دستگاه‌های ارتعاشی ظاهر می‌شوند و ریشه‌ها با فرکانس‌های بحرانی (تشدید) متناظر می‌باشند. حالت خاصی از معادلات جبری که در آن f در (۱.۶) یک چندجمله‌یی است نیز به طور ویژه بررسی می‌شود.

۱.۰.۶ مسایل نمونه

در این بخش چند مسئله‌ی ساده با کاربردهای جالب را فرمول‌بندی می‌کنیم که به حل معادلات غیرخطی تک‌متغیره منجر می‌شوند.

مثال ۱.۶. (معادله‌ی حالت گاز). گازهای ایده‌آل در حالت تعادل داخلی از معادله‌ی گاز ایده‌آل به‌صورت

$$pV = nRT$$

پیروی می‌کنند که در آن p فشار داخلی سیستم، V حجم سیستم، n تعداد مول‌های ذرات سیستم، R ثابت جهانی گازها و T دمای سیستم با یکای کلوین است. همان‌طور که از نام آن پیداست، قانون گاز ایده آل تنها یک تقریب خام از واقعیت است زیرا اثرات فیزیکی مهم مانند اندازه‌ی غیرصفر مولکول‌های گاز و نیروهای بین آن‌ها را نادیده می‌گیرد، به طوری که تنها در دماهای نسبتاً بالا و فشار کم منطقی است. یک تقریب بهتر که برخی از این اثرات را در نظر می‌گیرد معادله‌ی حالت واندروالس به‌صورت زیر است

$$\big(p + \frac{a}{v^۲}\big)(v - b) = RT \tag{۲.۶}$$

در این‌جا $v = V/n$ حجم ویژه، و a و b دو ثابت وابسته به این گاز خاص می‌باشند. اگر بخواهیم برای مقادیر معلوم از پارامترها حجم ویژه v را بیابیم، بایستی معادله‌ی غیرخطی تک‌متغیره (۲.۶) را برای یافتن ریشه‌ی v حل کنیم.

مثال ۲.۶. (گرافیک کامپیوتری). مسئله‌ی خطوط و سطوح پنهان در ارائه[۱] گرافیک کامپیوتری مهم است. این مسایل شامل یافتن اشتراک اشیاء در ابعاد دو یا سه است و معمولاً به حل دستگاه‌هایی از معادلات غیرخطی نیاز دارد. به‌عنوان مثال، فرض کنید بخواهیم یک شی در صفحه را رسم کنیم که با نابرابری زیر تعریف می‌شود

$$x^۴ + y^۴ \leqslant ۱,$$

همچنین بخواهیم خط راست $y = x + ۰/۵$ را نمایش دهیم که مانند شکل ۱.۶ از پشت این شی می‌گذرد. برای تعیین قطعه‌ای از خط که نمایش داده می‌شود، باید تعیین کرد که اشتراک آن با شی در کجا قرار دارد. خط راست با مرز این شی در نقاطی برخورد می‌کند که مختص اول آنها از معادله‌ی زیر به‌دست می‌آید

$$x^۴ + (x + ۰/۵)^۴ = ۱. \tag{۳.۶}$$

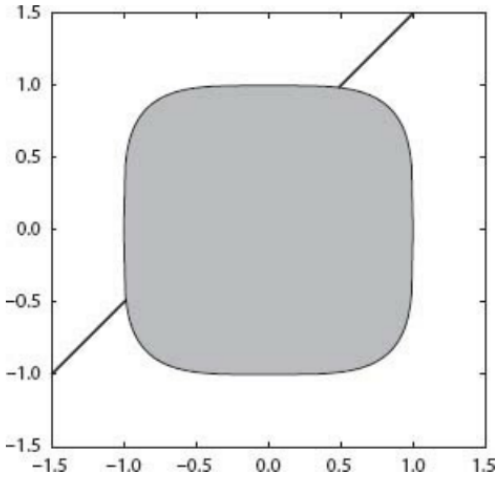
این معادله درجه چهارم دارای ۴ جواب است، اما تنها دو تای آنها حقیقی هستند و با نقاط برخورد واقعی متناظر می‌باشند. در صورتی که مختص x از نقاط برخورد معلوم شود، مختص y آنها به آسانی با $x + ۰/۵$ محاسبه می‌شود. با این حال محاسبه‌ی مقادیر x نیازمند حل معادله‌ی چندجمله‌ای (۳.۶) است.

مثال ۳.۶. (معادله‌ی کپلر[۲]). مسئله‌ی چرخیدن یک ماهواره در مدار زمین و یا یک سیاره در حال چرخش به دور خورشید را در نظر بگیرید. کپلر کشف کرد که مدار حرکت، یک بیضی و جرم سماوی مرکزی در یک کانون از بیضی قرار دارد. مطابق شکل ۲.۶ فرض می‌کنیم طول نیمقطر بزرگ بیضی a و طول نیمقطر کوچک بیضی b، و زمان لازم برای یک دور کامل مدار برابر T باشد.

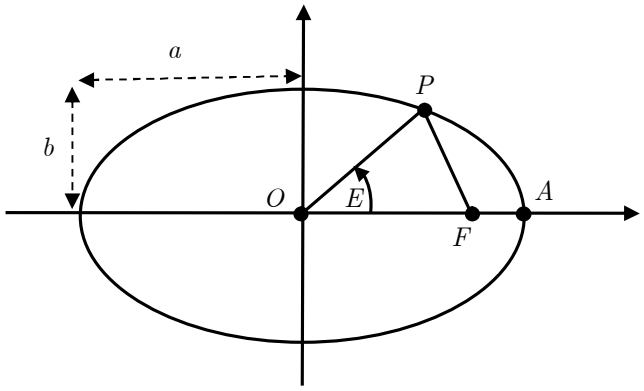
---

[۱] rendering

[۲] Johannes Kepler (1571–1630)



شکل ۱.۶: ارائه در گرافیک کامپیوتری می‌تواند شامل آشکار کردن یک خط باشد زمانی که در پشت یک شی پنهان شده است.



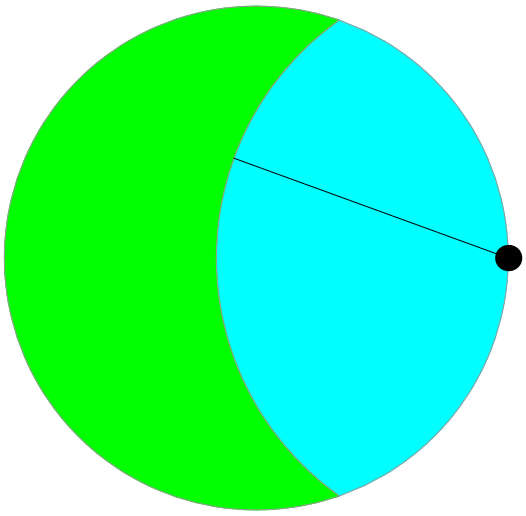
شکل ۲.۶: ماهواره P به دور زمین F می‌چرخد

اگر ماهواره در زمان $t = ۰$ در نقطه‌ی A، نزدیک‌ترین نقطه به زمین، واقع باشد، سوال این است که: مکان ماهواره در زمان t (برای $T > t$) کجاست؟

قانون دوم کپلر بیان می‌کند که زمان حرکت متناسب با مساحت ناحیه‌ای است که بردار شعاعی FP جاروب می‌کند. در نتیجه برای یافتن مدل مسئله، ابتدا لازم است مساحت ناحیه‌ی FAP، ناحیه‌ای که توسط بردار شعاعی جاروب می‌شود، به‌عنوان یک تابع از زاویه E (با نام گریز از مرکز نامتعارف) محاسبه شود. ثابت می‌شود این مساحت با فرمول زیر به‌دست می‌آید

$$S(E) = \frac{۱}{۲}ab(E - e\sin E)$$

در این‌جا $e = \frac{\sqrt{a^۲-b^۲}}{a}$ خروج از مرکز بیضی است. مطابق قانون دوم کپلر، $S(E)$ متناسب با t است. با در نظر گرفتن



شکل ۳.۶: مسئله‌ی کشاورز، بز، و چمن‌زار

ضریب تناسب مناسب، معادله‌ی کپلر به‌صورت زیر به‌دست می‌آید

$$E - e\sin E = \frac{۲\pi}{T}t \qquad (۴.۶)$$

این معادله تابع $E(t)$ را به‌صورت ضمنی تعریف می‌کند، یعنی، رابطه‌ای بین مکان ماهواره (زاویه E) و زمان t ارائه می‌کند. اگر بخواهیم مکان ماهواره را در زمان معین t به‌دست آوریم لازم است معادله‌ی غیرخطی (۴.۶) نسبت به E حل شود.

مثال ۴.۶. (مسئله‌ی کشاورز، بز، و چمن‌زار). یک کشاورز در کناره چمن‌زار دایره‌ای شکل به شعاع واحد بزی را با یک طناب به طول r بسته است. اگر بز دقیقاً به نیمی از مساحت چمن‌زار دسترسی داشته باشد، مقدار r چقدر است؟
ابتدا مسئله را مدل‌بندی می‌کنیم. به مرکز نقطه‌ای در کناره چمن‌زار، دایره‌ای به شعاع r ($r > ۱$) رسم می‌کنیم. مطابق شکل ۳.۶، بخشی از چمن‌زار که بز در آن می‌چرد همان ناحیه‌ی مشترک بین دایره‌ها است.

اگر مساحت این ناحیه را با A نشان دهیم، A را می‌توان با دستوری بر حسب شعاع دایره‌ها و فاصله‌ی بین مرکز دایره‌ها بیان کرد. در این مسئله فاصله‌ی بین مرکز دایره‌ها همان شعاع چمن‌زار است، بنابراین مساحت ناحیه‌ی A برابر است با

$$A(r) = r^۲\arccos(\frac{۱}{۲}r) + \arccos(۱ - \frac{۱}{۲}r^۲) - \frac{۱}{۲}r\sqrt{۴ - r^۲} \qquad (۵.۶)$$

اگر بز به نیمی از چمن‌زار دسترسی داشته باشد بایستی r را طوری بیابیم $A(r) = ۰/۵\pi$ شود که یک معادله‌ی غیرخطی تک‌متغیره است.

۱.۶ تکرار و همگرایی

معادله‌ی غیرخطی تک‌متغیره (۱.۶) را در نظر می‌گیریم طوری که f یک تابع حقیقی‌مقدار از یک متغیر باشد. عدد α را ریشه‌ی (دقیق) معادله‌ی (۱.۶) می‌نامیم هرگاه $f(\alpha) = ۰$. به‌طور کلی، ریشه‌های (۱.۶) را نمی‌توان به فرم بسته بیان کرد، به‌عنوان مثال، هرگاه f یک چندجمله‌یی با درجه‌ی بیش‌تر از چهار باشد، فرمول صریحی برای محاسبه‌ی صفرهای آن وجود ندارد. حتی وقتی یک جواب صریح نیز در دست است (برای مثال، معادله‌ی درجه سه ناقص)، اغلب این جواب چنان پیچیده است که استفاده از روش‌های عددی بسیار عملی‌تر می‌باشد.

روش‌های عددی ماهیت تکراری دارند. با شروع از یک یا چند تقریب اولیه، یک دنباله از مقادیر x_k توسط روش ساخته می‌شود با این امید که به α همگرا شود، یعنی

$$\lim_{k\to\infty} x_k = \alpha.$$

هر چند همگراییِ یک فرایند تکراری بسیار مطلوب است، اما در روش‌های عددی به چیزی بیش از همگرایی نیاز داریم و آن "همگرایی سریع" است. یک مفهوم مناسب برای اندازه‌گیری سرعت همگرایی، مرتبه‌ی همگرایی است که در زیر تعریف می‌شود.

تعریف ۱.۶ (همگرایی از مرتبه p)**.** یک دنباله از تکرارهای $\{x_k\}$ را به نقطه‌ی α همگرا از مرتبه‌ی $p \geqslant ۱$ گوییم، اگر ثابتی مانند $C > ۰$، و عدد صحیح مناسب $N \geqslant ۰$ وجود داشته باشد که به ازای $k \geqslant N$ داشته باشیم

$$\frac{|x_{k+۱} - \alpha|}{|x_k - \alpha|^p} \leqslant C. \tag{۶.۶}$$

در این حالت می‌گوییم مرتبه‌ی همگرایی برابر p است. اگر $p = ۱$، برای همگرایی x_k به α لازم است در (۶.۶) داشته باشیم $C < ۱$. در این حالت، دنباله **همگرای خطی** به α و C **ضریب مجانبی همگرایی** می‌نامیم. اگر $p > ۱$، دنباله $\{x_k\}$ را **همگرای اَبَرخطی** به α می‌گوییم. برای $p = ۲$ همگرایی از مرتبه دو (مربعی) و برای $p = ۳$ همگرایی از مرتبه سه (مکعبی) نامیده می‌شود. در حالت کلی ممکن است p عددی غیر صحیح باشد.

مثال ۵.۶. دنباله‌های زیر به لحاظ نظری هر دو به ریشه‌ی مثبت تابع $f(x) = x^۲ - ۲$ یعنی $\sqrt{۲}$ همگرا می‌شوند.

$$x_۰ = ۱, \quad x_{k+۱} = \frac{۳}{۴}x_k + \frac{۱}{۲x_k}, \quad k = ۰, ۱, \dots$$
$$y_۰ = ۱, \quad y_{k+۱} = \frac{۱}{۲}y_k + \frac{۱}{y_k}, \quad k = ۰, ۱, \dots$$

برای بررسی سرعت همگرایی دنباله‌ها، در ابتدا برخی از جمله‌های آن‌ها را در جدول ۱.۶ می‌آوریم. با توجه به این که تا ۸ رقم اعشار مقدار همگرایی $\sqrt{۲} \doteq ۱/۴۱۴۲۱۳۵۶$ است، به طور شهودی می‌توان گفت سرعت همگرایی دنباله‌ی y_k

جدول ۱.۶: مقایسه سرعت همگرایی دو دنباله به $\alpha = \sqrt{۲}$

k	۱	۲	۳	۴	۵
x_k	۱٫۲۵	۱٫۳۳۷۵	۱٫۳۷۶۹۵۶۷۸	۱٫۳۹۵۸۳۷۱۹	۱٫۴۰۵۰۸۵۸۶
y_k	۱٫۵	۱٫۴۱۶۶۶۶۶۷	۱٫۴۱۴۲۱۵۶۹	۱٫۴۱۴۲۱۳۵۶	۱٫۴۱۴۲۱۳۵۶

بیش‌تر از دنباله‌ی x_k است. اما چه اندازه؟ برای بررسی دقیق‌تر این مسئله از حسابان کمک می‌گیریم. با توجه به تعریف دنباله x_k داریم

$$\sqrt{۲} - x_{k+۱} = \sqrt{۲} - \frac{۳}{۴}x_k - \frac{۱}{۲x_k} = \frac{۱}{۲}(\sqrt{۲} - x_k)(\frac{۳}{۲} - \frac{۱}{\sqrt{۲}x_k}).$$

اکنون نابرابری $۱ \leqslant x_k < \sqrt{۲}$ به‌ازای هر $k \geqslant ۰$ نتیجه می‌دهد که

$$\sqrt{۲} - x_{k+۱} \leqslant \frac{۱}{۲}(\sqrt{۲} - x_k)(\frac{۳}{۲} - \frac{۱}{۲}) = \frac{۱}{۲}(\sqrt{۲} - x_k).$$

بنابراین

$$\frac{|x_{k+۱} - \sqrt{۲}|}{|x_k - \sqrt{۲}|} \leqslant \frac{۱}{۲} =: C,$$

یعنی دنباله‌ی x_k همگرای خطی است.

از طرف دیگر، با توجه به تعریف دنباله y_k داریم

$$y_{k+۱} - \sqrt{۲} = \frac{۱}{۲}y_k + \frac{۱}{y_k} - \sqrt{۲} = \frac{۱}{۲y_n}(y_k - \sqrt{۲})^۲.$$

اکنون نابرابری $\sqrt{۲} < y_k \leqslant ۱٫۵$ به‌ازای هر $k \geqslant ۰$ نتیجه می‌دهد که

$$y_{k+۱} - \sqrt{۲} = \frac{۱}{۲\sqrt{۲}}(y_k - \sqrt{۲})^۲.$$

بنابراین

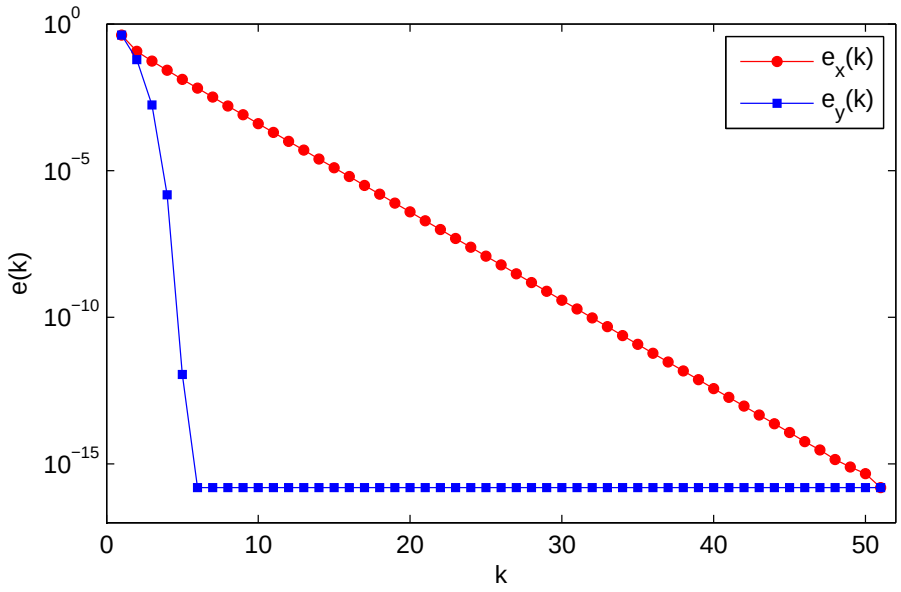
$$\frac{|y_{k+۱} - \sqrt{۲}|}{|y_k - \sqrt{۲}|^۲} \leqslant \frac{۱}{۲\sqrt{۲}} =: C,$$

یعنی دنباله‌ی y_k دارای همگرایی مرتبه دو است و به تعبیری، تعداد ارقام با معنای درست با هر تکرار تقریباً دو برابر می‌شوند.

یک روش شهودی بهتر، نمایش نمودار خطای (مطلق یا نسبی) جمله‌های دنباله است. فرض کنید

$$e_x(k) := \frac{|x_k - \sqrt{۲}|}{\sqrt{۲}}, \qquad e_y(k) := \frac{|y_k - \sqrt{۲}|}{\sqrt{۲}},$$

که در آن $e_x(k)$ خطای دنباله‌ی $\{x_k\}$ و $e_y(k)$ خطای دنباله‌ی $\{y_k\}$ در تکرار k-ام هستند. این نمودارها برای مقادیر مختلف k در شکل ۴.۶ رسم شده‌اند. نمودارها نشان می‌دهند همگرایی دنباله‌ی $\{y_k\}$ به $\sqrt{۲}$ بسیار سریعتر از همگرایی دنباله‌ی $\{x_k\}$ به $\sqrt{۲}$ است. مشاهده می‌کنید که پس از تکرار ششم خطای دنباله‌ی $\{y_k\}$ به سطح دقت ماشین $u \approx ۱۰^{-۱۶}$



شکل ۴.۶: نمودار خطاهای e_x و e_y

رسیده است و پس از آن ثابت مانده است. بنابراین نیازی به ادامه تکرار دنباله‌ی $\{y_k\}$ برای $k \geqslant 6$ نیست. برای رسم این نمودارها و همچنین بدست آوردن اعداد جدول ۱.۶، برنامه زیر در متلب نوشته شده است.

```
x(1) = 1; y(1) = 1;
ex(1) = abs(x(1)-sqrt(2)); ey(1) = abs(y(1)-sqrt(2));
for k = 1:50
    x(k+1) = 3/4*x(k)+1/(2*x(k));
    y(k+1) = 1/2*y(k)+1/y(k);
    ex(k+1) = abs(x(k+1)-sqrt(2))/sqrt(2);
    ey(k+1) = abs(y(k+1)-sqrt(2))/sqrt(2);
end
semilogy(1:k+1,ex,'-ro', 1:k+1,ey,'-bs');
xlabel('k'); ylabel('e(k)');
legend('e_x(k)','e_y(k)');
```

به یافتن ریشه‌ی α از تابع مفروض f با یک روش تکراری مناسب که دنباله‌ی $\{x_k\}$ را می‌سازد، برمی‌گردیم. از لحاظ

نظری دنباله $\{x_k\}$ پس از بی‌نهایت بار تکرار، α را برمی‌گرداند. اما در عمل در بهترین حالت ریشه با خطایی در سطح دقت ماشین بدست می‌آید. گاهی انتظار ما حتی از این هم کمتر است، یعنی می‌خواهیم تقریبی از α تا آستانه‌ی تحمل (با دقت) ε که $\varepsilon \geqslant u$ را بدست آوریم. بنابراین بهتر است در کوچکترین مقدار k که به ازای آن نابرابری زیر برقرار شود تکرارها متوقف شود

$$|e(k)| = |\alpha - x_k| \leqslant \varepsilon. \tag{۷.۶}$$

با این حال شرط توقف روی خطای پیشرو از لحاظ عملی کارایی چندانی ندارد زیرا در عمل مقدار α معلوم نیست. اما گاهی می‌توان با توجه به خصوصیات تابع f و خصوصیات روش و بدون اطلاع از α کران پایین تعداد تکرارها برای برآورده شدن (۷.۶) را از پیش تعیین کرد. اگر چنین کاری امکان نداشته باشد برآوردکننده‌ی مناسب‌تری از خطا مورد نیاز است. یک معیار جایگزین می‌تواند به‌صورت زیر انتخاب شود

$$|x_k - x_{k-1}| \leqslant \varepsilon. \tag{۸.۶}$$

به عبارتی روش تکراری زمانی متوقف می‌شود که تفاوت میان دو تکرار متوالی کمتر از ε شود. سرانجام، یک شرط توقف معمول در حل معادله‌ی $f(x) = 0$ بررسی مانده در x_k است. شرط توقف روی خطای پسین برای دقت داده شده‌ی ε به صورت زیر است

$$|f(x_k)| \leqslant \varepsilon. \tag{۹.۶}$$

لازم به ذکر است حتی اگر باقیمانده بسیار کوچک باشد تضمینی بر نزدیک بودن x_k به α نیست و به وضعیت نگاشت f وابسته است.

در این فصل روش ساختن دنباله‌هایی را فرا می‌گیریم که ریشه‌ی یک تابع مفروض را به صورت تکراری تقریب می‌زنند. روش‌های عددی گوناگون به شکل‌های متفاوت چنین دنباله‌هایی را تولید می‌کنند. از جمله‌ی این روش‌ها می‌توان به روش‌های پایه مانند روش دوبخشی و روش تکرار نقطه‌ی ثابت، روش‌های مبتنی بر درونیابی مانند روش نابجایی، روش وتری و روش مولر، و روش‌های مبتنی بر مشتق مانند روش نیوتن اشاره کرد. لازم به ذکر است که در مثال ۵.۶ در بالا دنباله‌ی $\{x_k\}$ به کمک روش تکرار نقطه‌ی ثابت و دنباله‌ی $\{y_k\}$ به کمک روش نیوتن ساخته شده‌اند که بعداً آن‌ها را مطالعه خواهیم کرد.

در این فصل عمدتاً در مورد ریشه‌های حقیقی صحبت می‌کنیم. اما در برخی حالات و بخصوص در مورد چندجمله‌ایها یافتن ریشه‌های مختلط نیز اهمیت دارد. از این رو اندکی هم در مورد ریشه‌های مختلط چندجمله‌ایها صحبت خواهیم کرد. از این پس منظورمان از ریشه، ریشه‌ی حقیقی است مگر اینکه لفظ "مختلط" ذکر شود.

چندگانگی

پیش از آنکه وارد بحث روش‌های ریشه‌یابی شویم، در مورد چندگانگی ریشه‌ها مختصراً توضیحاتی ارائه می‌دهیم.

ریشه‌ی α از معادله‌ی (۱.۶) را یک ریشه با چندگانگی m می‌نامیم هرگاه f را بتوان به صورت

$$f(x) = (x-\alpha)^m q(x) \qquad (۱۰.۶)$$

نوشت طوری که $\lim_{x\to\alpha} q(x) \neq ۰$. یک ریشه با چندگانگی یک، ریشه‌ی ساده نام دارد.
در معادلاتی که $f(x)$ یک چندجمله‌یی است، تجزیه به عوامل اول می‌تواند چندگانگی ریشه را مشخص کند. به طور مثال، با توجه به

$$x^۶ + x^۵ - ۱۲x^۴ + ۲x^۳ + ۴۱x^۲ - ۵۱x + ۱۸ = (x-۱)^۳(x+۳)^۲(x-۲)$$

نتیجه می‌شود معادله‌ی

$$x^۶ + x^۵ - ۱۲x^۴ + ۲x^۳ + ۴۱x^۲ - ۵۱x + ۱۸ = ۰$$

دارای یک ریشه‌ی با چندگانگی ۳ در $x = ۱$، یک ریشه‌ی با چندگانگی ۲ در $x = -۳$ و یک ریشه‌ی ساده در $x = ۲$ است.
در مورد معادله‌ی $f(x) = ۰$ با

$$f(x) = ۲x + \ln\left(\frac{۱-x}{۱+x}\right)$$

چه می‌توان گفت؟ واضح است $f(۰) = ۰$، یعنی معادله یک ریشه در $x = ۰$ دارد. اما چندگانگی این ریشه چقدر است؟ در قضیه‌ی زیر یک روش ساده برای یافتن چندگانگی ریشه بیان می‌کنیم.

قضیه ۱.۶. *فرض کنید تابع f دارای m مشتقِ پیوسته باشد. معادله‌ی $f(x) = ۰$ دارای ریشه‌ی α با چندگانگی m است اگر و فقط اگر $f(\alpha) = f'(\alpha) = \cdots = f^{(m-۱)}(\alpha) = ۰$، اما $f^{(m)}(\alpha) \neq ۰$ باشد.*

برهان. اگر f به صورت (۱۰.۶) باشد، بوضوح داریم $f(\alpha) = f'(\alpha) = \cdots = f^{(m-۱)}(\alpha) = ۰$. برعکس فرض کنیم مشتقات f تا مرتبه‌ی $m-۱$ در α صفر باشند. بسط تیلر مرتبه m تابع f حول α می‌دهد

$$\begin{aligned} f(x) &= f(\alpha) + f'(\alpha)(x-\alpha) + \cdots + (x-\alpha)^{m-۱}\frac{f^{(m-۱)}(\alpha)}{(m-۱)!} + (x-\alpha)^m \frac{f^{(m)}(\xi(x))}{m!} \\ &= (x-\alpha)^m \frac{f^{(m)}(\xi(x))}{m!}, \end{aligned}$$

که $\xi(x)$ مجهول وابسته به x است. بنابراین (۱۰.۶) با $q(x) = f^{(m)}(\xi(x))/m!$ برقرار است. □

به مسئله‌ی محاسبه‌ی چندگانگی ریشه‌ی $\alpha = ۰$ در مثال قبل از قضیه برمی‌گردیم. با توجه به این‌که $f(۰) = f'(۰) = ۰$ $f''(۰) = ۰$ اما $f'''(۰) = -۴ \neq ۰$، چندگانگی ریشه ۳ است.

۲.۶ روش دوبخشی

یکی از روش‌های بسیار ساده و سرراست برای ساختن دنباله‌ای تکراری برای تقریب ریشه، روش دوبخشی[1] است. در روش دوبخشی لازم است یک بازه مانند $[a,b]$ از پیش تعیین شود طوری که ریشه‌ی $f(x)=۰$ به صورت یکتا در آن قرار داشته باشد. در صورت وجود چندین ریشه، باید برای هر ریشه یک بازه در نظر گرفت که ریشه در آن یکتا باشد. به طور دقیق باید بگوییم بازه‌ی $[a,b]$ را آنقدر کوچک انتخاب می‌کنیم که شامل تنها یک ریشه از $f(x)=۰$ باشد. گاهی با استفاده از خواص f مانند یکنوایی، تغییرات علامت و غیره می‌توان چنین بازه‌ای را یافت. همچنین می‌توان با رسم نمودار تابع به صورت دستی (در حالت‌های خیلی خاص) یا به کمک بسته‌های نرم افزاری (مثلاً با دستور `fplot` در متلب) چنین بازه‌هایی را حداقل به صورت تقریبی مشخص کرد.

بنابراین فرض می‌کنیم f بر روی بازه معلوم $[a,b]$ تابع پیوسته‌ای باشد و $f(a)f(b)<۰$ که تضمین می‌کند حداقل یک ریشه در (a,b) واقع است. همچنین فرض کنیم این ریشه که آن را α می‌نامیم، یکتا باشد. روش دوبخشی بسیار ساده و خط‌مشیِ آن نصف کردن بازه در هر مرحله و انتخاب یکی از دو زیربازه‌ای است که f در آن تغییر علامت می‌دهد (یعنی ریشه در آن واقع است). به‌طور دقیق‌تر، ابتدا قرار می‌دهیم

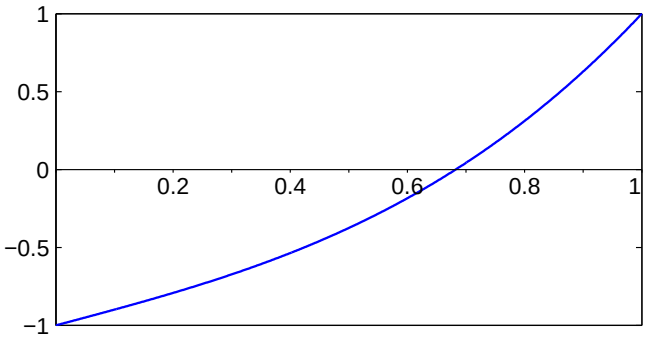
$$a_۰=a,\quad b_۰=b,\quad I_۰=(a_۰,b_۰),\quad x_۰=(a_۰+b_۰)/۲.$$

در این صورت $x_۰$ نقطه‌ی وسط بازه‌ی $(a_۰,b_۰)$ است. واضح است که سه حالت وجود دارد: یا $x_۰$ ریشه است، یعنی $f(x_۰)=۰$، که در این صورت $\alpha=x_۰$ و کار تمام است. و یا ریشه در $(a_۰,x_۰)$ واقع است، یا در $(x_۰,b_۰)$. برای تشخیصِ اینکه ریشه در کدام یک از دو زیربازه قرار دارد، علامت $f(a_۰)f(x_۰)$ را تعیین می‌کنیم. در صورت منفی بودن علامت، ریشه در $(a_۰,x_۰)$ است و در غیر این صورت در $(x_۰,b_۰)$ قرار دارد. بازه مورد نظر را تعیین و آن را $(a_۱,b_۱)$ می‌نامیم و قرار می‌دهیم $x_۱=(a_۱+b_۱)/۲$ و دوباره روند بالا را تکرار می‌کنیم. به طور کلی در گام $k\geqslant ۱$، زیربازه‌ی $I_k=(a_k,b_k)$ از بازه‌ی $I_{k-۱}=(a_{k-۱},b_{k-۱})$ به‌شکل زیر انتخاب می‌شود:

قرار دهید $x_{k-۱}=(a_{k-۱}+b_{k-۱})/۲$، اگر $f(x_{k-۱})=۰$ در این صورت $\alpha=x_{k-۱}$ و روش متوقف می‌شود، در غیر این‌صورت، اگر $f(a_{k-۱})f(x_{k-۱})<۰$ قرار دهید $a_k=a_{k-۱}$ و $b_k=x_{k-۱}$، و اگر $f(x_{k-۱})f(b_{k-۱})<۰$، قرار دهید $a_k=x_{k-۱}$ و $b_k=b_{k-۱}$. سپس تعریف کنید $x_k=(a_k+b_k)/۲$ و k را یک واحد افزایش دهید.

مثال ۶.۶. می‌خواهیم صفر تابع $f(x)=x^۳+x-۱$ واقع در بازه‌ی $[۰,۱]$ را بیابیم. نمودار f در شکل ۵.۶ آمده است.

[1] Bisection



شکل ۵.۶: نمودار تابع $f(x) = x^۳ + x - ۱$ در بازه‌ی $[۰, ۱]$

با فرض $a_۰ = ۰$ و $b_۰ = ۱$، نتیجه‌ی اجرای روش دوبخشی به‌صورت زیر است

$$
\begin{aligned}
&I_۰ = (۰, ۱), && x_۰ = ۰/۵\\
&I_۱ = (۰/۵, ۱), && x_۱ = ۰/۷۵\\
&I_۲ = (۰/۵, ۰/۷۵), && x_۲ = ۰/۶۲۵\\
&I_۳ = (۰/۶۲۵, ۰/۷۵), && x_۳ = ۰/۶۸۷۵\\
&\vdots && \vdots\\
&I_۹ = (۰/۶۸۱۶, ۰/۶۸۳۶), && x_۹ = ۰/۶۸۲۶\\
&I_{۱۰} = (۰/۶۸۱۶, ۰/۶۸۲۶), && x_{۱۰} = ۰/۶۸۲۱
\end{aligned}
$$

با توجه به طول بازه‌ی $I_{۱۰}$، مقدار $x_{۱۰}$ تقریبی برای α به صورت $\alpha = ۰/۶۸۲۱ \pm ۰/۰۰۰۵$ می‌باشد.

مرتبه همگرایی. با توجه به روند ساختن زیربازه‌ها، هر زیربازه I_k شامل α است. همچنین، دنباله‌ی $\{x_k\}$ لازم است به α همگرا شود. زیرا، در هر گام طول I_k یعنی $|I_k| = b_k - a_k$ نصف می‌شود بنابراین $|I_k| = (۱/۲)^k |I_۰|$. اگر e_k خطای مطلق در مرحله‌ی kام باشد، یعنی $e_k = |x_k - \alpha|$، آنگاه

$$e_k \leqslant \frac{۱}{۲}|I_k| = \left(\frac{۱}{۲}\right)^{k+۱}(b-a). \qquad (۱۱.۶)$$

نامساوی بالا نشان می‌دهد

$$\lim_{k\to\infty} x_k = \alpha \quad \text{یا} \quad \lim_{k\to\infty} e_k = ۰$$

یعنی روش دوبخشی همواره همگراست. از طرفی (۱۱.۶) نشان میدهد همگرایی دنباله‌ی $\{x_k\}$ به α، حداقل همانند همگرایی دنباله‌ی $\{۲^{-k}\}$ به صفر است. می‌دانیم که دنباله‌ی $\{۲^{-k}\}$ با مرتبه‌ی خطی ($p = ۱$) همگراست. پس همگرایی روش دوبخشی حداقل خطی است.

همچنین در این روش می‌توان حداکثر تکرارهای لازم برای رسیدن به یک دقت از پیش تعیین شده‌ی ε را تعیین کرد. برای این کار باید داشته باشیم $e_k \leqslant \varepsilon$. با توجه به (۱۱.۶) کافی است

$$\left(\frac{۱}{۲}\right)^{k+۱}(b-a) \leqslant \varepsilon,$$

که این هم نتیجه می‌دهد

$$k+۱ \geqslant \log_۲\left(\frac{b-a}{\varepsilon}\right).$$

بنابراین تعداد تکرار لازم برای رسیدن به دقت از پیش تعیین شده‌ی ε (حداکثر) برابر

$$\left\lfloor \log_۲\left(\frac{b-a}{\varepsilon}\right)\right\rfloor$$

است، که در آن $\lfloor \cdot \rfloor$ کف یک عدد را نشان می‌دهد. توجه داریم که فرمول بالا حداکثر تکرارهای لازم را ارائه می‌دهد که به تابع f بستگی ندارد. ممکن است برای یک تابع خاص یکی از اعضای دنباله‌ی $\{x_k\}$ در همان تکرارهای اولیه برابر α شود و روش خیلی زود متوقف شود. برای مثال روش دوبخشی روی $[-۱, ۱]$ در یک تکرار ریشه‌ی تابع $f(x) = x$ را ارائه می‌دهد.

مثال ۷.۶. می‌خواهیم صفر تابع $f(x) = \cosh x + \cos x - ۳$ واقع در بازه‌ی $[۱, ۳]$ با $\varepsilon = ۱۰^{-۱۰}$ را بیابیم. با توجه به تخمین (۲.۶) داریم

$$\lfloor \log_۲(۲ \times ۱۰^{۱۰}) \rfloor = \lfloor ۳۴٫۲۲ \rfloor = ۳۴.$$

بنابراین با ۳۴ تکرار به دقت مورد نظر دست می‌یابیم. مقدار محاسبه شده با این تعداد تکرار برابر است با $\alpha = ۱٫۸۵۷۹۲۰۸۲۹۱۴۸۵$ و به ازای آن $|f(\alpha)| \doteq ۳٫۶۸۷۷ \times ۱۰^{-۱۲}$.

برنامه‌ی زیر یک پیاده‌سازی از روش دوبخشی را در متلب فراهم می‌کند. ورودی‌ها تابع `fun`، ابتدای بازه `a`، انتهای بازه `b` و دقت از پیش تعیین شده‌ی `tol` می‌باشد. خروجی‌ها تقریب به‌دست آمده‌ی `root`، مقدار تابع به ازای این تقریب (باقیمانده) `res`، و تعداد تکرارهای انجام شده `Niter` است.

```
function [root,res,Niter] = Bisection(fun,a,b,tol)
Niter = 0; I = (b - a)*0.5;
x = [a, (a+b)*0.5, b]; fx = fun(x);
while I > tol
   Niter = Niter + 1;
   if sign(fx(1)) * sign(fx(2))<0
```

```
      x(3) = x(2); x(2) = x(1)+(x(3)-x(1))*0.5;
      fx = fun(x); I = (x(3)-x(1))*0.5;
   elseif sign(fx(2))* sign(fx(3))<0
      x(1) = x(2); x(2) = x(1)+(x(3)-x(1))*0.5;
      fx = fun(x); I = (x(3)-x(1))*0.5;
   else
       x(2) = x(find(fx==0)); I = 0;
   end
end
root = x(2); res = fun(root);
```

با توجه به شرط جلوی `while`، تکرارها تا زمانی ادامه می‌یابند که نصف طول بازه‌ای که ریشه در آن واقع است از `tol` کوچکتر شود؛ به عبارت دیگر $|x_k - \alpha| \leqslant$ `tol`. خروجی مثال ۶.۶ با اجرای دستور زیر به‌دست می‌آید

```
>> [root,res,Niter] = Bisection(@(x) x.^3+x-1,0,1,0.0005)
```

پیش از این، ثابت کردیم اگر بازه‌ی اولیه‌ی $[a, b]$ به درستی انتخاب شود، روش دوبخشی همواره همگراست. این یکی از مزایای این روش است. بجز انتخاب بازه‌ی اولیه، مهمترین چالش در روش دوبخشی تعیین علامت $f(a_k)f(x_k)$ در هر گام است. یک راه محاسبه‌ی $f(a_k)$ و $f(x_k)$ و سپس ضرب کردن آن‌ها در هم و بررسی شرط مثبت یا منفی بودن حاصلضرب آن‌هاست. برای جلوگیری از خطاهای محاسباتی مانند خطای سرریز بهتر است ابتدا علامت $f(a_k)$ و $f(b_k)$ را تعیین و علامت آن‌ها را در هم ضرب کنیم. این همان کاری است که در برنامه بالا هم انجام داده‌ایم.

فرض کنیم می‌خواهیم ریشه را با دقت ماشین u بدست آوریم. در این حالت وقتی به نزدیکیِ ریشه می‌رسیم برای مقادیر $f(x_k)$ که $|f(x_k)| < u$ ممکن است ماشین علامت $f(x_k)$ را درست تشخیص ندهد. در این صورت بازه‌ها به اشتباه انتخاب می‌شوند و ممکن است الگوریتم به بیراهه برود. اما خوشبختانه این امر هیچ مشکلی ایجاد نخواهد کرد. زیرا در این سطح محاسبات داریم $|f(x_k)| < u$، یعنی ما ریشه را با خطای پسین کمتر از u بدست آورده‌ایم و به بیراهه رفتن یا نرفتن روش در ادامه اهمیتی ندارد. به همین خاطر یکی دیگر از خصوصیات مثبت روش دوبخشی این است که در "سطح دقت ماشین" روشی مستحکم [۱] (کارامد) است.

[۱] Robust

یک نقص روش دوبخشی آن است که تنها برای صفرهایی از تابع f به‌کار می‌رود که علامت f در اطراف‌شان عوض می‌شود. به‌ویژه، ممکن است ریشه‌های مضاعف نادیده گرفته شوند. از دیگر نواقص آن کند بودن آن است. شاید چون این روش اولین روشی است که در این فصل فراگرفته‌ایم، فعلاً متوجه این موضوع نشویم، اما در ادامه خواهیم دید که روش‌هایی بسیار سریع‌تر از روش دوبخشی نیز وجود دارند.

۳.۶ روش نیوتن

علامت تابع f در نقاط انتهایی زیربازه‌ها تنها اطلاعاتی است که روش دوبخشی آن را به‌کار می‌گیرد. در صورتی که f مشتق‌پذیر باشد، یک روش کارا با به‌کارگیری مقادیر f و f' ساخته می‌شود که به روش نیوتن معروف است. ایده‌ی اصلی روش نیوتن بسیار سر راست می‌باشد. در این روش، بجای یافتن ریشه‌ی f، در هر تکرار ریشه‌ی یک تقریب خطی از f را بدست می‌آوریم. این تقریب‌های خطی، همان خطوط مماس f در نقاط x_k هستند. گیریم x_k تقریب محاسبه شده‌ی اخیر از یک ریشه‌ی تابع f باشد. معادله‌ی خط مماس بر منحنی $y = f(x)$ در نقطه‌ی x_k به صورت زیر است

$$L_k(x) = f(x_k) + f'(x_k)(x - x_k)$$

اگر محل تقاطع این خط مماس با محور xها (یعنی همان ریشه‌ی L_k) را تقریب جدید x_{k+1} بگیریم (شکل ۶.۶)، در آن صورت

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, \ldots, \tag{۱۲.۶}$$

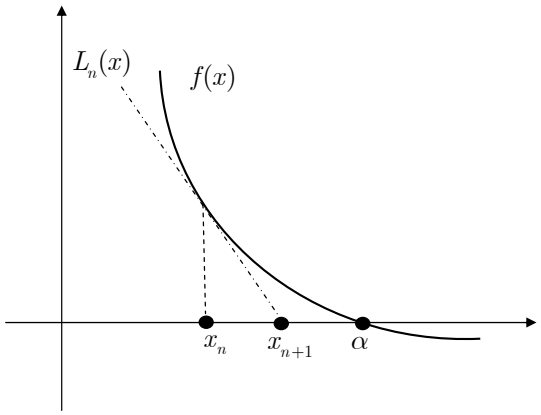
به شرطی که $f'(x_k) \neq 0$، این دستور دنباله‌ی $\{x_k\}$ را با شروع از حدس اولیه‌ی x_0 تولید می‌کند. در هر تکرار مقدار $-f(x_k)/f'(x_k)$ به مقدار قبلی اضافه می‌شود. به قدرمطلق این کسر، طول گام نیوتن می‌گوییم؛ اندازه‌ی اختلاف بین دو گام متوالی برابر طول گام است.

در بالا مشاهده کردیم که روش نیوتن یک "روش خطی‌سازی" است، به این معنی که بجای حل یک مسئله غیرخطی ($f(x) = 0$)، چندین مسئله‌ی خطی ($L_k(x) = 0$ برای $k = 0, 1, \ldots$) حل می‌کند.

به‌عنوان یادداشتی تاریخی، نیوتن در ۱۶۶۹ با مثال‌هایی عددی این روش را شرح داد و از توضیح هندسی شامل تقریب یک منحنی با خط مماسش استفاده نکرد. او کارش را به شکل رابطه بازگشتی بالا ارائه نداد بلکه این کار توسط ریاضیدان انگلیسی جوزف رافسون در ۱۶۹۰ انجام شد. به همین دلیل این روش را اغلب روش نیوتن-رافسون نیز می‌نامند.

مثال ۸.۶. می‌خواهیم صفر تابع $f(x) = x^3 + x - 1$ واقع در بازه‌ی $[0, 1]$ را با روش نیوتن بیابیم. با توجه به این که $f'(x) = 3x^2 + 1$، دستور (۱۲.۶) دنباله‌ی زیر را تولید می‌کند

$$x_{k+1} = x_k - \frac{x_k^3 + x_k - 1}{3x_k^2 + 1} = \frac{2x_k^3 + 1}{3x_k^2 + 1}$$



شکل ۶.۶: تعبیر هندسی روش نیوتن

اگر حدس اولیه را وسط بازه $[۰, ۱]$ یعنی $x_۰ = ۰/۵$ در نظر بگیریم در آن صورت نتایج در ستون دوم جدول ۲.۶ آمده است توجه داشته باشید که روش دوبخشی برای یافتن این تقریب می‌بایست تا گام ۵۲ اجرا شود!

جدول ۲.۶: محاسبه‌ی صفر f با روش نیوتن

k	x_k	$\lvert f(x_k)\rvert$	$\lvert x_k-\alpha\rvert$	نسبت
۱	۰/۷۱۴۲۸۵۷۱۴۲۸۵۷۱۴	۰/۰۷۸۷۲	۰/۰۳۱۹۶	۰/۹۶۱۳۳
۲	۰/۶۸۳۱۷۹۷۲۳۵۰۲۳۰۴	$۲/۰۴۳ \times ۱۰^{-۳}$	$۸/۵۲ \times ۱۰^{-۴}$	۰/۸۳۴۱۵
۳	۰/۶۸۲۳۲۸۴۲۳۳۰۴۵۷۸	$۱/۴۸۵ \times ۱۰^{-۶}$	$۶/۱۹ \times ۱۰^{-۷}$	۰/۸۵۳۵۵
۴	۰/۶۸۲۳۲۷۸۰۳۸۲۸۳۴۷	$۷/۸۵۴ \times ۱۰^{-۱۳}$	$۳/۲۸ \times ۱۰^{-۱۳}$	۰/۸۵۴۰۴
۵	۰/۶۸۲۳۲۷۸۰۳۸۲۸۰۱۹	$۱/۱۱ \times ۱۰^{-۱۶}$	۰	

ستون چهارم جدول ۲.۶ شامل نسبتهای $\frac{|x_{k+۱}-\alpha|}{|x_k-\alpha|^۲}$ است. همچنان‌که در قضیه‌ی ۲.۶ خواهیم دید حد این نسبتها برابر $C = |f''(\alpha)/۲f'(\alpha)| \approx ۰/۸۵۴۰۸$ است و مقادیر ستون آخر جدول ۲.۶ به این مقدار میل می‌کنند. برای تشکیل این نسبتها مقدار α با استفاده از دستور `fzero` در متلب فراهم شده است. ستون آخر نشان می‌دهد که مرتبه‌ی همگرایی روش نیوتن بایستی $p = ۲$ باشد، که البته در ادامه آن را اثبات می‌کنیم.

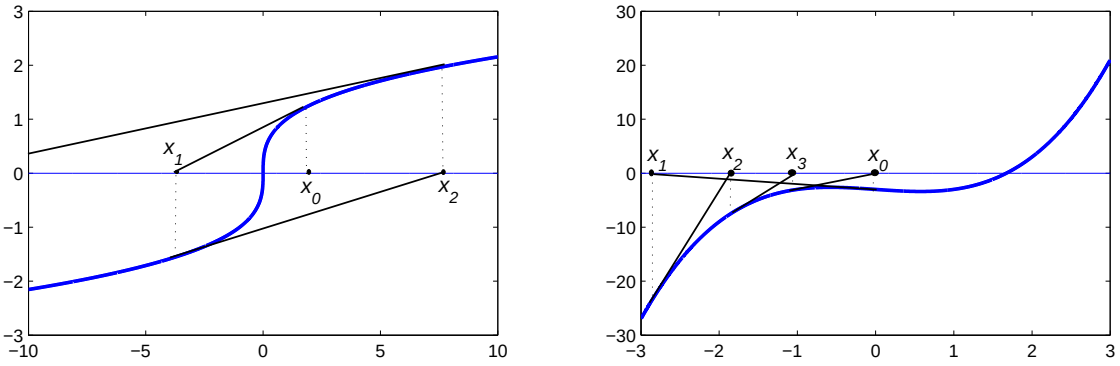
برنامه‌ی زیر یک پیاده‌سازی از روش نیوتن را در متلب فراهم می‌کند. ورودی‌ها تابع `fun`، مشتق تابع `dfun`، حدس اولیه `x0` و دقت از پیش تعیین شده‌ی `tol` می‌باشد. خروجی‌ها به ترتیب تقریب به‌دست آمده `root`، مقدار تابع به ازای این تقریب `res` و تعداد تکرارهای انجام شده `Niter` است.

```
function [root,res,Niter] = Newton(fun,dfun,x0,tol)
```

```
x = x0; Niter = 0; diff = tol+1;
while abs(diff) >= tol
    Niter = Niter + 1;
    diff = -fun(x)/dfun(x); x = x + diff;
end
root = x; res = fun(root);
```

همانطور که در شرط `while` برنامه می‌بینید، تکرارها زمانی خاتمه می‌یابند که طول گام نیوتن از `tol` کوچکتر شود؛ به بیان دیگر $|x_{k+1} - x_k| \leqslant$ `tol`. می‌توان شرط‌های توقف دیگر مانند شرط روی باقیمانده، یا شرط روی حداکثر تعداد تکرار را هم اضافه کرد.

در حالت کلی ممکن است برای انتخاب‌های دلخواه x_0 روش نیوتن همگرا نباشد. آیا برخی موارد را می‌توانید تصور کنید که ممکن است مشکلاتی را برای روش نیوتن بوجود آورند؟ شکل ۷.۶ حالت‌هایی را نشان می‌دهد که روش نیوتن به شکست می‌انجامد: در سمت چپ دنباله‌ی حاصل واگرا می‌شود در حالی که در سمت راست، دنباله در بین چند نقطه دور می‌زند و هرگز به صفر تابع میل نمی‌کند.



شکل ۷.۶: روش نیوتن برای $f(x) = x^3 - x - 3$ در سمت راست و $f(x) = \sqrt[3]{x}$ در سمت چپ

ترکیب روش‌های نیوتن و دوبخشی

روش نیوتن اگر همگرا باشد، همگرایی آن بسیار سریع است، اما همانطور که در نمودارهای بالا مشاهده می‌کنید تکرارهای روش نیوتن ممکن است از ریشه دور شوند یا در یک چرخه‌ی نامتناهی قرار گیرند. اما قبلاً دیدیم که در روش دوبخشی اگرچه همگرایی کند است اما ریشه همواره تحت کنترل است و همگرایی همیشه حاصل می‌شود. برای به‌دست آوردن یک

الگوریتم قابل اعتماد، روش نیوتن را می‌توان با روش دوبخشی ترکیب کرد. برای اجرای این روش، ابتدا یک بازه را تعیین می‌کنیم که تابع f در آن تغییر علامت دهد زیرا پیوستگی تابع وجود ریشه در این بازه را تضمین می‌کند. در هر تکرار از روش ترکیبی، بدست آوردن چنین بازه‌ای ضروری است. فرض کنید در یکی از تکرارهای روش ترکیبی بازه $[a,b]$ شامل ریشه و x_c تقریب کنونی از ریشه باشد. با روش نیوتن تکرار بعدی چنین است

$$x_+ = x_c - \frac{f(x_c)}{f'(x_c)}$$

در صورتی که $x_+ \in [a,b]$، این مقدار به عنوان تکرار جدید روش ترکیبی پذیرفته می‌شود و در غیر این صورت، تکرار جدید با روش دو بخشی بدست می‌آید یعنی

$$x_+ = \frac{a+b}{۲}.$$

تابع `StepIsOK` به شکل زیر نحوه انجام تکرار بعدی را مشخص می‌کند. ورودی‌های این تابع تکرار کنونی `xc`، مقدار تابع f در این نقطه `fx`، مقدار مشتق تابع در این نقطه `dfx` و ابتدا و انتهای بازه $[a,b]$ است. خروجی این تابع `ok` ، صفر یا یک است. اگر تکرار بعدی همچنان درون بازه کنونی $[a,b]$ واقع باشد عدد ۱ را برمی‌گرداند، بنابراین تکرار بعدی با روش نیوتن انجام و x_+ پذیرفته می‌شود. اگر تکرار بعدی خارج بازه کنونی $[a,b]$ قرار گیرد عدد ۰ را برمی‌گرداند و با این خروجی معلوم می‌شود تکرار بعدی را روش دوبخشی انجام می‌دهد.

```
function ok = StepIsOK(xc,fx,dfx,a,b)
if dfx > 0,  ok = ((a-xc)*dfx <= -fx) & (-fx <= (b-xc)*dfx);
elseif dfx < 0,  ok = ((a-xc)*dfx >= -fx) & (-fx >= (b-xc)*dfx);
else  ok = 0;
end
```

پس از تعیین تقریب جدید با در نظر گرفتن اصل تغییر علامت تابع، همانند آنچه در روش دو بخشی داشتیم، بازه‌ی شامل ریشه به‌روز می‌شود. با این اصلاح همگرایی روش ترکیبی تضمین می‌شود. کد روش ترکیبی می‌تواند به صورت زیر نوشته شود.

```
function [root,res,Niter] = NewtonBisection(fun,dfun,a,b,tol)
xc = a; Niter = 0; diff = tol+1;
while (abs(diff) > tol) && (0.5*(b-a) > tol)
```

```
    fxc = fun(xc); dfxc = dfun(xc);
    ok = StepIsOK(xc,fxc,dfxc,a,b);
    if (ok)
        diff = -fxc/dfxc; xc = xc + diff;
    else
        xc = (a+b)/2;
    end
    s = sign(fun(a))*sign(fun(xc));
    if s < 0, b = xc; elseif s > 0, a = xc; else break; end
    Niter = Niter + 1;
end
root = xc; res = fun(xc);
```

شرط توقف هم ترکیبی از شرط توقف برنامه روش دوبخشی و برنامه روش نیوتن است، یعنی تا زمانی که یا گام نیوتن یا نصف طول بازه‌ی دوبخشی از tol بزرگترند تکرارها ادامه می‌یابند. یک اجرا از این کد روی مثال شکل سمت راست ۷.۶ به صورت زیر است

```
>> [root,res,Niter] = NewtonBisection(@(x)x.^3-x-3,@(x)3*x.^2-1,-5,5,10^-14)
```

اجرای این دستور نتایج زیر را به همراه خواهد داشت

```
root =
    1.671699881657161
res =
    -8.881784197001252e-16
Niter =
    11
```

در قسمت بعد ثابت خواهیم کرد که تحت شرایطی روی تابع f و با انتخاب مناسب نقطه‌ی شروع x_0، روش نیوتن با مرتبه دو به ریشه همگراست.

مرتبه همگرایی روش نیوتن

قضیه ۲.۶. (همگرایی روش نیوتن) *فرض کنید f، f' و f'' به ازای جمیع مقادیر x در بازه بسته $I := [\alpha - r, \alpha + r]$ با $r > ۰$، پیوسته باشند طوری که $f(\alpha) = ۰$، $f'(\alpha) \neq ۰$ و $f''(\alpha) \neq ۰$. به علاوه فرض می‌کنیم ثابت مثبت M موجود باشد طوری که برای هر t و s در I داشته باشیم*

$$\frac{|f''(s)|}{|f'(t)|} \leqslant M.$$

اگر $\delta \leqslant \min\{r, ۱/M\}$ و حدس اولیه $x_۰$ در بازه‌ی $I_\delta := \{x \in \mathbb{R} : |x - \alpha| \leqslant \delta\}$ انتخاب شود، آنگاه دنباله $\{x_k\}$ حاصل از روش نیوتن (۱۲.۶) به‌طور مربعی به α همگرا می‌شود.

برهان. گیریم x_k جمله‌ای از دنباله باشد که $|x_k - \alpha| \leqslant \delta = \min\{r, ۱/M\}$، در این صورت $x_k \in I_\delta \subseteq I$. با قضیه تیلر، تابع f را حول نقطه x_k بسط می‌دهیم

$$۰ = f(\alpha) = f(x_k) + (\alpha - x_k)f'(x_k) + \frac{۱}{۲}(\alpha - x_k)^۲ f''(\xi_k), \tag{۱۳.۶}$$

که ξ_k بین α و x_k، و بنابراین در I قرار دارد. با توجه به (۱۲.۶) و (۱۳.۶) داریم

$$x_{k+۱} - \alpha = (x_k - \alpha)^۲ \frac{f''(\xi_k)}{۲f'(x_k)}. \tag{۱۴.۶}$$

چون $|x_k - \alpha| \leqslant \frac{۱}{M}$ بنابراین

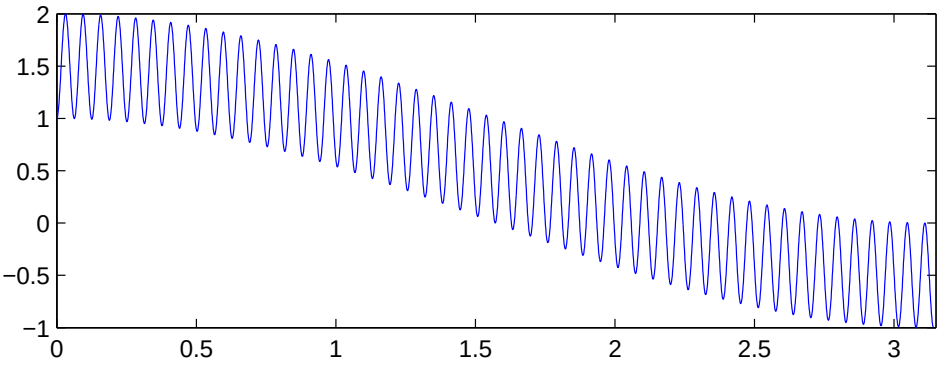
$$|x_{k+۱} - \alpha| = \frac{۱}{۲}|x_k - \alpha|\frac{|f''(\xi_k)|}{|f'(x_k)|}|x_k - \alpha| \leqslant \frac{۱}{۲}\frac{۱}{M}\, M|x_k - \alpha| = \frac{۱}{۲}|x_k - \alpha|.$$

بنابراین اگر $|x_k - \alpha| \leqslant \delta$ آنگاه $|x_{k+۱} - \alpha| \leqslant \delta/۲$. چون $|x_۰ - \alpha| \leqslant \delta$، با استقرای ریاضی می‌توان نشان دارد برای هر $k \geqslant ۰$ داریم $|x_k - \alpha| \leqslant ۲^{-k}\delta$. بنابراین وقتی $k \to \infty$ دنباله $\{x_k\}$ به α همگرا خواهد شد. نقطه ξ_k بین α و x_k قرار دارد، بنابراین وقتی $k \to \infty$ دنباله $\{\xi_k\}$ نیز به α همگرا می‌شود. چون f' و f'' روی I پیوسته هستند بنا به (۱۴.۶) داریم

$$\lim_{k\to\infty} \frac{|x_{k+۱} - \alpha|}{|x_k - \alpha|^۲} = \Big|\frac{f''(\alpha)}{۲f'(\alpha)}\Big|, \tag{۱۵.۶}$$

که با توجه به تعریف ۱.۶ همگرایی مربعی دنباله $\{x_k\}$ به α با $C = |f''(\alpha)/۲f'(\alpha)|$ و $۰ < C \leqslant M/۲$ نتیجه می‌شود. □

برای همگرایی روش نیوتن گاهی لازم است حدس اولیه $x_۰$ خیلی به α نزدیک باشد. به‌عنوان مثال برای یافتن $\alpha = \frac{\pi}{۲}$، صفر تابع $f(x) = \cos(x) + \sin^۲(۵۰x)$، چنین کاری ضروری است (شکل ۸.۶ را ببینید). توجه کنید که حتی اگر انتخاب $x_۰$ به یک همسایگی از $\pi/۲$ به شعاع ۰٫۰۱ محدود شود، باز هم این تابع نوسانی در این همسایگی دو ریشه دارد.



شکل ۸.۶: نمودار $f(x) = \cos(x) + \sin^{۲}(۵۰x)$ و لزوم نزدیکی $x_{۰}$ به α

از جمله فرض‌های قضیه ۲.۶ برای همگرایی x_k به α، انتخاب حدس اولیه $x_{۰}$ به‌صورت زیر است

$$|x_{۰} - \alpha| \leqslant \frac{۱}{M} \qquad (۱۶.۶)$$

بنابراین M مقداری است برای آنکه بدانیم چه اندازه $x_{۰}$ باید به α نزدیک باشد تا همگرایی به α تضمین شود. تمرین ۶.۶ را ببینید.

اگر در مورد علامت مشتق‌های تابع اطلاعاتی در دسترس باشد، می‌توان نشان داد که روش نیوتن روی یک بازه وسیع‌تر همگرا است.

قضیه ۳.۶. *فرض کنید تابع f در شرایط قضیه ۲.۶ صدق کند و عدد حقیقی $A > \alpha$، موجود باشد طوری که f' و f'' هر دو در بازه $J = [\alpha, A]$ مثبت باشند. در این صورت، دنباله $\{x_k\}$ که از روش نیوتن (۱۲.۶) به‌دست می‌آید برای هر حدس اولیه $x_{۰} \in J$ به‌طور مربعی به α همگرا می‌شود.*

اثبات قضیه بالا به خواننده واگذار می‌شود.

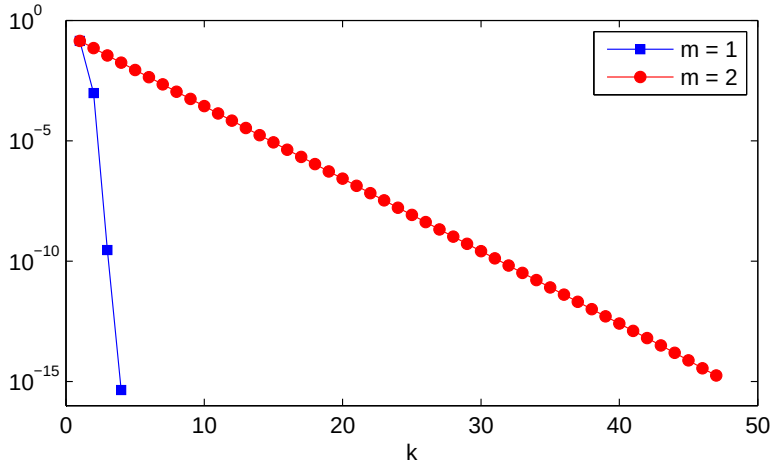
ملاحظه ۱.۶. فرض کنید $f(\alpha) = f'(\alpha) = ۰$، یعنی f یک ریشه مضاعف در α داشته باشد، به علاوه f'' در یک همسایگی از α پیوسته است. اگر $\{x_k\}$ دنباله حاصل از روش نیوتن باشد در این صورت بنا به (۱۴.۶) و به‌کارگیری قضیه مقدار میانگین برای f' داریم

$$x_{k+۱} - \alpha = \frac{۱}{۲}(x_k - \alpha) f''(\xi_k) \frac{x_k - \alpha}{f'(x_k) - f'(\alpha)} = \frac{۱}{۲}(x_k - \alpha)\frac{f''(\xi_k)}{f''(\eta_k)},$$

که ξ_k و η_k هر دو بین α و x_k قرار دارند. اگر برای هر x در بازه $I = [\alpha - \delta, \alpha + \delta]$ با $\delta > ۰$ فرض کنیم $۰ < M' \leqslant |f''(x)| \leqslant M$ که در آن $M \leqslant ۲M'$، آنگاه برای حدس اولیه $x_{۰} \in I$ دنباله‌ی تکرارهای $\{x_k\}$ همگرای خطی به α است. به علاوه داریم

$$\lim_{k\to\infty} \frac{|x_{k+۱} - \alpha|}{|x_k - \alpha|} = \frac{۱}{۲}.$$

!h



شکل ۹.۶: همگرایی روش نیوتن برای ریشه‌های ساده و چندگانه

در حالت کلی، اگر α صفر f با چندگانگی $m > ۱$، و دنباله حاصل از روش نیوتن همگرا باشد در آن صورت

$$\lim_{k\to\infty} \frac{|x_{k+۱} - \alpha|}{|x_k - \alpha|} = ۱ - \frac{۱}{m}.$$

اثبات در تمرینات خواسته شده است.

مثال ۹.۶. توابع $f(x) = \sin x$ و $g(x) = \sin^۲ x$ هر دو در $\alpha = \pi$ برابر صفرند. با توجه به این که $f'(\pi) \neq ۰$ لذا π صفر ساده‌ی f است در حالی که $g'(\pi) = ۰$ و π صفر مضاعف g است. در شکل ۹.۶ همگرایی مرتبه دو برای تکرارهای f و همگرایی خطی برای تکرارهای g با روش نیوتن ملاحظه می‌شود.

در حالتی که چندگانگی ریشه $m > ۱$ باشد، روش نیوتن همچنان همگرا است اگرچه با مرتبه‌ی خطی، با این حال باید $x_۰$ به طور مناسبی انتخاب شود و به ازای هر $x \in I \setminus \{\alpha\}$ داشته باشیم $f'(x) \neq ۰$. در چنین حالت‌هایی برای سرعت بخشیدن به همگرایی، روش متداول (۱۲.۶) را به‌شکل زیر اصلاح می‌کنیم

$$x_{k+۱} = x_k - m\frac{f(x_k)}{f'(x_k)}, \qquad k = ۰, ۱, \ldots \tag{۱۷.۶}$$

به شرطی که $f'(x_k) \neq ۰$، مرتبه همگرای روش نیوتن اصلاح شده (۱۷.۶) برابر دو است. اثبات این ادعا هم در تمرینات خواسته شده است. در مثال ۹.۶، اگر برای $g(x)$ روش نیوتن اصلاح‌شده (۱۷.۶) را با $m = ۲$ به‌کار گیریم، سرعت همگرایی دنباله حاصل شبیه دنباله تولید شده با تابع f خواهد شد.

برای ریشه‌ای که چندگانگی آن معلوم نیست به شکل زیر عمل می‌کنیم: اگر معادله $f(x) = ۰$ در $x = \alpha$ ریشه چندگانه داشته باشد، در آن صورت α یک ریشه ساده معادله زیر است

$$u(x) = ۰ \qquad u(x) = f(x)/f'(x)$$

بنابراین اگر روش نیوتن را برای معادله $u(x) = ۰$ به‌کار گیریم، همگرای مربعی خواهد بود و این مستقل از چندگانگی α به‌عنوان یک ریشه از معادله $f(x) = ۰$ است. روش نیوتن برای معادله $u(x) = ۰$ چنین است

$$x_{k+۱} = x_k - \frac{f(x_k)}{f'(x_k) - f(x_k)f''(x_k)/f'(x_k)} \tag{۱۸.۶}$$

این روش در سال ۱۸۷۰ توسط شرودر[۱] پیشنهاد شده است. برای اجرای این روش محاسبه $f''(x_k)$ مورد نیاز است و این کارایی روش را کم می‌کند.

۴.۶ روش‌های شبه‌نیوتنی

هرگاه حدس اولیه به اندازه کافی نزدیک ریشه باشد، روش نیوتن دارای ویژگی خیلی خوب همگرایی مربعی است. این بدان معناست که این روش خیلی سریع‌تر از روش‌هایی مانند دوبخشی به یک ریشه در نزدیکی نقطه‌ای معلوم همگرا می‌شود. هزینه این کار محاسبه‌ی تابع f و مشتق آن در هر تکرار است و برای مثال ساده‌ای مانند ۸.۶ چندان مشکل نیست. ولی در بسیاری از مسایل تابع f کاملاً پیچیده است، ممکن است به‌طور تحلیلی نوشته نشده باشد، یا به جای آن، ممکن است ناگزیر به اجرای یک برنامه برای محاسبه $f(x)$ باشیم. در چنین حالت‌هایی، مشتق‌گیری f به طور تحلیلی مشکل یا ناممکن است، و حتی اگر بتوان یک دستور به‌دست آورد، محاسبه‌ی آن پرهزینه می‌باشد. برای چنین مسایلی، می‌خواهیم از محاسبه‌ی مشتق‌ها اجتناب یا، دست‌ِکم، در حد امکان تعداد کم‌تری از آن‌ها را محاسبه کنیم، حال آنکه همگرایی تا اندازه‌ای بالا نگه‌داشته شود. تکرارهایی به صورت

$$x_{k+۱} = x_k - \frac{f(x_k)}{d_k}, \qquad d_k \approx f'(x_k) \tag{۱۹.۶}$$

را اغلب روش‌های **شبه‌نیوتنی** می‌نامند.

۱.۴.۶ روش شیب ثابت

فرض کنید f' تنها یک بار در $x_۰$ محاسبه شود و در (۱۹.۶) برای هر k به جای d_k مقدار $f'(x_۰)$ قرار گیرد. اگر شیب f چندان تغییر نکند، از این روش انتظار می‌رود رفتار روش نیوتن را دنبال کند. این روش، روش شیب ثابت نامیده می‌شود

$$x_{k+۱} = x_k - \frac{f(x_k)}{d} \qquad d = f'(x_۰) \tag{۲۰.۶}$$

برای تحلیل این روش، با قضیه تیلر $f(x_k)$ را حول ریشه α بسط می‌دهیم. اگر خطا در x_k را با $e_k := x_k - \alpha$ نمایش دهیم، در آن صورت

$$f(x_k) = f(\alpha) + (x_k - \alpha)f'(\alpha) + \mathcal{O}\big((x_k - \alpha)^۲\big) = e_k f'(\alpha) + \mathcal{O}(e_k^۲)$$

[۱] Ernst Schröder (1841-1902)

به جای نوشتن صریح باقی‌مانده از نماد $\mathcal{O}(\cdot)$ استفاده شده است. از دو طرف (۲۰.۶) ریشه α را کم می‌کنیم و داریم

$$e_{k+1} = e_k - \frac{f(x_k)}{d} = e_k\Big(1 - \frac{f'(\alpha)}{d}\Big) + \mathcal{O}(e_k^2)$$

اگر $|1 - f'(\alpha)/f'(x_0)| < 1$، آنگاه برای x_0 به اندازه کافی نزدیک به α، چنان نزدیک که جمله $\mathcal{O}(e_k^2)$ در بالا قابل صرفنظر باشد، این روش همگرای خطی است. در هر حال از روش مماس ثابت نمی‌توان انتظاری بهتر از همگرایی خطی را داشت.

یک شکل متفاوت از این روش آن است که گاهی اوقات مشتق را روزآمد کنیم. به جای این که در (۱۹.۶) مقدار d_k را برای هر k برابر $f'(x_0)$ بگیریم، هرگاه همگرایی روش تکراری به کندی پیش رفت، یک مشتق جدید $f'(x_k)$ را محاسبه کنیم. اگر چه این روش نسبت به روش شیب ثابت به محاسبات بیشتری از مشتق نیاز دارد با این وجود، ممکن است در تعداد تکرار کمتری همگرا شود. در انتخاب یک روش برای حل معادلات غیرخطی، بین هزینه یک تکرار و تعداد تکرارهای مورد نیاز برای رسیدن به یک دقت مطلوب معمولاً تضاد وجود دارد.

۲.۴.۶ روش وتری

فرض کنید $f'(x_k)$ را با شیب خط قاطع مار بر نقاط $(x_{k-1}, f(x_{k-1}))$ و $(x_k, f(x_k))$ تقریب بزنیم، یعنی

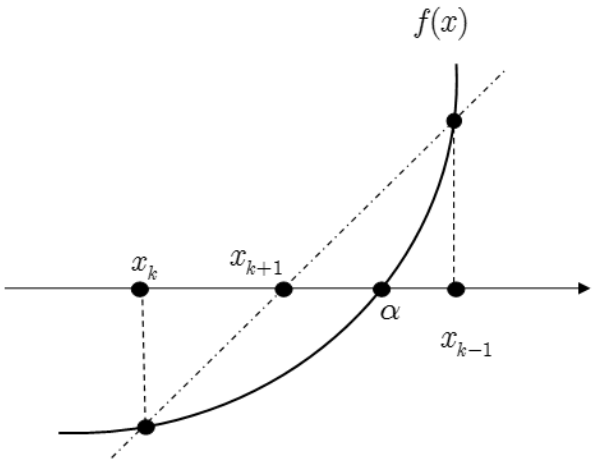
$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

در واقع این نسبت، همان اولین تفاضل تقسیم شده‌ی $f[x_{k-1}, x_k]$ می‌باشد و به‌کارگیری آن باعث اجتناب از مشتق‌گیری است. به این ترتیب d_k در (۱۹.۶) را می‌توان برابر این نسبت تفاضلی در نظر گرفت. توجه کنید که در حد وقتی x_{k-1} به x_k میل کند، d_k به $f'(x_k)$ میل خواهد کرد. بنابراین وقتی x_{k-1} به x_k نزدیک باشد می‌توان انتظار داشت d_k تقریب قابل قبولی برای $f'(x_k)$ باشد. روش وتری یا خط قاطع از این خط‌مشی استفاده می‌کند و برای مقادیر آغازین x_0 و x_1 به‌شکل زیر ساخته می‌شود

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k), \qquad k = 1, 2, \ldots. \tag{۲۱.۶}$$

مثال ۱۰.۶. می‌خواهیم صفر تابع $f(x) = x^3 + x - 1$ واقع در بازه‌ی $[0, 1]$ را با روش وتری بیابیم. اگر مقادیر اولیه را به‌صورت $x_0 = 0$ و $x_1 = 1$ در نظر بگیریم در آن صورت نتایج در ستون دوم جدول ۳.۶ آمده است توجه داشته باشید که روش نیوتن برای یافتن این تقریب از ریشه اندکی سریع‌تر بود.

تعیین مرتبه همگرایی روش وتری از نتایج مثال ۱۰.۶ مشکل است. به نظر می‌رسد سریع‌تر از خطی و کندتر از مربعی باشد. آیا مرتبه همگرایی روش وتری را می‌توانید حدس بزنید؟ در ادامه برآوردی از مرتبه همگرایی ارائه می‌کنیم.



شکل ۱۰.۶: روش وتری یا خط قاطع

جدول ۳.۶: محاسبه‌ی صفر f با روش وتری

k	x_k	$\lvert f(x_k)\rvert$	$\lvert x_k-\alpha\rvert$	p_k
۲	۰/۵۰	۰/۳۷۵۰	۰/۱۸۲۳	
۳	۰/۶۳۶۳۶۳۶۳۶۳۶۳۶۳	۰/۱۰۵۹	۰/۰۴۶۰	۲/۴۸۲
۴	۰/۶۹۰۰۵۲۳۵۶۰۲۰۹۴۲	۰/۰۱۸۶	$۷/۷۲ \times ۱۰^{-۳}$	۱/۲۹۴
۵	۰/۶۸۲۰۲۰۴۱۹۶۴۸۱۸۶	$۷/۳۶ \times ۱۰^{-۴}$	$۳/۰۷ \times ۱۰^{-۴}$	۱/۸۰۸
۶	۰/۶۸۲۳۲۵۷۸۱۴۰۹۸۹۳	$۴/۸۵ \times ۱۰^{-۶}$	$۲/۰۲ \times ۱۰^{-۶}$	۱/۵۵۸
۷	۰/۶۸۲۳۲۷۸۰۴۳۵۹۰۲۶	$۱/۲۷ \times ۱۰^{-۹}$	$۵/۳۱ \times ۱۰^{-۱۰}$	۱/۶۴۱
۸	۰/۶۸۲۳۲۷۸۰۳۸۲۸۰۱۸	$۲/۲۲ \times ۱۰^{-۱۵}$	$۸/۸۸ \times ۱۰^{-۱۶}$	۱/۶۱۳
۹	۰/۶۸۲۳۲۷۸۰۳۸۲۸۰۱۹	$۱/۱۱ \times ۱۰^{-۱۶}$	۰	

فرض کنیم $e_k := \lvert x_k-\alpha\rvert$ خطای مطلق x_k و p مرتبه همگرای روش وتری باشد. با توجه به (۶.۶) برای مقادیر به اندازه کافی بزرگ k داریم

$$e_{k+۱} \approx Ce_k^p \quad \text{و} \quad e_k \approx Ce_{k-۱}^p$$

این دو رابطه را بر هم تقسیم می‌کنیم و رابطه‌ی $e_{k+۱}/e_k \approx (e_k/e_{k-۱})^p$ نتیجه می‌شود. با گرفتن لگاریتم از دو طرف این رابطه برآوردی به صورت زیر برای p موسوم به مرتبه همگرایی تجربی[۱] به‌دست می‌آید

$$p_k = \log(e_{k+۱}/e_k)/\log(e_k/e_{k-۱}), \qquad k = ۱, ۲, \ldots.$$

[۱] Experimental Order of Convergence

در مثال ۱۰.۶ مقادیر p_k در ستون پایانی جدول ۳.۶ آمده است، این مقادیر برآوردی برای مرتبه همگرایی دقیق باشند. می‌خواهیم مقدار p را به‌طور دقیق تعیین کنیم، برای این کار ابتدا لم زیر را می‌آوریم. اثبات این لم به تمرین ۲۲.۶ واگذار می‌شود.

لم ۴.۶. *فرض کنید f، f' و f'' به ازای جمیع مقادیر x در بازه‌ای شامل α پیوسته باشند طوری که $f(\alpha) = ۰$ و $f'(\alpha) \neq ۰$. در این صورت اگر حدس‌های اولیه $x_۰$ و $x_۱$ به اندازه کافی نزدیک به α انتخاب شوند، آنگاه دنباله $\{x_k\}$ حاصل از روش وتری (۲۱.۶) به α همگرا می‌شود. به‌علاوه*

$$\lim_{k\to\infty} \frac{e_{k+۱}}{e_k e_{k-۱}} = \frac{f''(\alpha)}{۲f'(\alpha)} =: \mu. \tag{۲۲.۶}$$

در مثال ۱۰.۶ نسبتهای $e_{k+۱}/(e_k e_{k-۱})$ برای $k = ۴, ۵, ۶$ به ترتیب برابر ۰/۸۶۵۷، ۰/۸۵۱۸ و ۰/۸۵۴۲ است و به خوبی با مقدار حدی ۰/۸۵۴۱، مقدار μ در (۲۲.۶)، قابل مقایسه می‌باشند.
دستور (۲۲.۶) بیان می‌کند که برای مقادیر به اندازه کافی بزرگ k رابطه‌ی زیر برقرار است

$$e_{k+۱} \approx \mu e_k e_{k-۱}. \tag{۲۳.۶}$$

در ادامه نشان می‌دهیم که چگونه از این نتیجه مرتبه همگرایی روش وتری به‌دست می‌آید. فرض کنید ثابت $C > ۰$ موسوم به ضریب همگرایی موجود باشد طوری که

$$e_{k+۱} \approx C e_k^p. \tag{۲۴.۶}$$

در این جا نیز برای مقادیر به اندازه کافی بزرگ k این رابطه را می‌توان به اندازه دلخواه به تساوی نزدیک کرد. با توجه به فرض $e_k \approx C e_{k-۱}^p$ و بنابراین

$$e_{k-۱} \approx \big(e_k/C\big)^{۱/p}.$$

هم‌اکنون با فرض برقراری (۲۳.۶)، دستورهای $e_{k-۱}$ و $e_{k+۱}$ را در (۲۳.۶) جایگزین می‌کنیم و داریم

$$C e_k^p \approx \mu e_k \big(e_k/C\big)^{۱/p} = \mu C^{-۱/p}\, e_k^{۱+۱/p}.$$

با انتقال ثابت‌ها به یک طرف خواهیم داشت

$$C^{۱+۱/p}\mu^{-۱} \approx e_k^{۱+۱/p-p}. \tag{۲۵.۶}$$

سمت چپ این عبارت ثابت و مستقل از k است بنابراین سمت راست نیز باید چنین باشد، یعنی توان e_k باید صفر شود

$$۱ + \frac{۱}{p} - p = ۰.$$

جواب‌های این معادله درجه دوم برابر است با $p = (۱ \pm \sqrt{۵})/۲$. برای همگرایی روش p باید مثبت باشد و بنابراین مرتبه همگرایی روش وتری برابر برابر نسبت طلایی است، یعنی

$$p = \frac{۱ + \sqrt{۵}}{۲} \doteq ۱/۶۱۸۰.$$

مقادیر ستون پایانی جدول ۳.۶ تا حدی به این مقدار نزدیک شده‌اند. اکنون ضریب همگرایی C در (۲۴.۶) را تعیین می‌کنیم. برای $p = (۱ + \sqrt{۵})/۲$ سمت راست (۲۵.۶) مستقل از k و برابر ۱ است بنابراین $C^{۱+۱/p}\mu^{-۱} = ۱$، اما $p = ۱ + ۱/p$ پس $C = \mu^{۱/p}$. به طور خلاصه برای روش وتری نشان دادیم

$$\lim_{k\to\infty} \frac{|x_{k+۱} - \alpha|}{|x_k - \alpha|^p} = \left|\frac{f''(\alpha)}{۲f'(\alpha)}\right|^{۱/p}. \qquad (۲۶.۶)$$

همانگونه که در روش نیوتن اشاره کردیم، اگر یک جفت از نقاط $x_{k-۱}$ و x_k یافت شود که در آن علامت‌های f مختلف باشند، روش وتری نیز می‌تواند با روش دوبخشی ترکیب شود. این بدان معناست که در روش وتری به جای کار کردن خودکار با نقاط جدید $x_{k+۱}$ و x_k، نقطه‌ی $x_{k+۱}$ و یکی از نقاط x_k یا $x_{k-۱}$ بر حسب اینکه علامت $f(x_k)$ یا $f(x_{k-۱})$ با علامت $f(x_{k+۱})$ مخالف باشد، انتخاب می‌شود. به این ترتیب، همواره یک بازه شامل صفر تابع f در دسترس است. این الگوریتم روش نابجایی یا تصحیح خطا نامیده می‌شود و در صورت تضمین همگرایی، دارای همگرایی خطی است.

۳.۴.۶ روش مولر

در روش وتری از درونیابی خطی بین $(x_۰, f(x_۰))$ و $(x_۱, f(x_۱))$ برای یافتن تقریب بعدی $x_۲$ استفاده می‌شود. یک تعمیم ساده برای روش وتری آن است که هر بار، به‌جای دو نقطه از سه نقطه استفاده کنیم. فرض کنیم سه نقطه $x_۰$، $x_۱$ و $x_۲$ داده شده‌اند، چندجمله‌ای درجه دومی می‌سازیم که از سه نقطه $(x_i, f(x_i))$، $i = ۰, ۱, ۲$ بگذرد. این چندجمله‌ای درجه دوم با دستور زیر به‌دست می‌آید

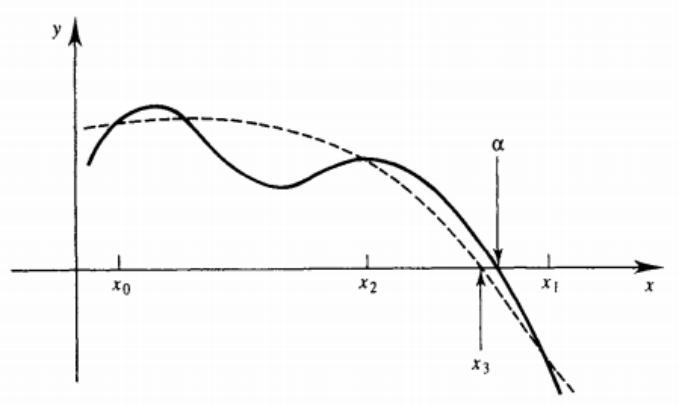
$$p_۲(x; f) = f(x_۲) + (x - x_۲)f[x_۲, x_۱] + (x - x_۲)(x - x_۱)f[x_۲, x_۱, x_۰].$$

فرض کنیم $p_۲(x; f) = ۰$ دارای دو ریشه حقیقی و $x_۳$ ریشه نزدیکتر به $x_۲$ باشد (شکل ۱۱.۶). این روند را با نقاط $x_۱$، $x_۲$ و $x_۳$ تکرار می‌کنیم و الی آخر. این روش به روش مولر[۱] معروف است.

در حالت کلی ریشه‌های چندجمله‌ای درونیاب درجه دوم که از سه نقطه $(x_i, f(x_i))$، $i = k-۲, k-۱, k$ می‌گذرد با یک بازنویسی مناسب، از معادله درجه دوم زیر به‌دست می‌آیند

$$f[x_k, x_{k-۱}, x_{k-۲}]\, h_k^۲ + w\, h_k + f(x_k) = ۰, \qquad (۲۷.۶)$$

[۱] David Eugene Muller (1924 – 2008)



شکل ۱۱.۶: منحنی خطچین نمایش $p_۲(x;f)$ در روش مولر است.

که در آن

$$h_k = x - x_k \qquad w = f[x_k, x_{k-۱}] + (x_k - x_{k-۱})f[x_k, x_{k-۱}, x_{k-۲}].$$

برای اجتناب از خطاهای کاهش ارقام بامعنی از دستور زیر برای محاسبه ریشه‌های معادله (۲۷.۶) استفاده می‌کنیم

$$h_k = -\frac{۲f(x_k)}{w \pm \sqrt{w^۲ - ۴f(x_k)f[x_k, x_{k-۱}, x_{k-۲}]}}, \quad k = ۲, ۳, \ldots$$

نزدیک‌ترین ریشه معادله (۲۷.۶) به x_k متناظر کوچکترین مقدار $|h_k|$ است. بنابراین علامت مخرج را طوری انتخاب می‌کنیم که مخرج را حداکثر و در نتیجه $|h_k|$ را حداقل نماید. با این مقدمه روش مولر چنین است

$$x_{k+۱} = x_k + h_k, \qquad k = ۲, ۳, \ldots, \tag{۲۸.۶}$$

که به سه حدس اولیه $x_۰$، $x_۱$ و $x_۲$ از ریشه نیاز دارد.

مثال ۱۱.۶. می‌خواهیم صفر تابع $f(x) = x^۳ + x - ۱$ واقع در بازه‌ی $[۰, ۱]$ را با روش مولر بیابیم. اگر مقادیر اولیه را به‌صورت $x_۰ = ۰$، $x_۱ = ۱$ و $x_۲ = ۰/۵$ در نظر بگیریم در آن صورت نتایج در ستون دوم جدول ۴.۶ آمده است

برخلاف روش‌های نیوتن یا وتری، حتی اگر حدس‌های اولیه همگی حقیقی باشند، روش مولر می‌تواند صفرهای مختلط تابع f را محاسبه کند. این یک جنبه مهم و یک دلیل برای استفاده از آن است. البته برای محاسبه یک ریشه حقیقی نیز امکان مواجه‌شدن با تقریب‌های مختلط وجود دارد، زیرا ممکن است جواب‌های به‌دست آمده از معادله درجه دوم (۲۷.۶) مختلط باشند. در چنین حالت‌هایی قسمت موهومی معمولاً آنقدر نزدیک صفر است که می‌تواند نادیده گرفته شود. از دیگر سو، می‌توان نشان داد که

$$e_{k+۱} = -\frac{f'''(\xi_k)}{۶p'_۲(c_k)} e_k e_{k-۱} e_{k-۲}, \tag{۲۹.۶}$$

جدول ۴.۶: محاسبه‌ی صفر f با روش مولر

k	x_k	h_{k-1}	$\lvert f(x_k)\rvert$
۳	۰/۶۶۶۶۶۶۶۶۶۶۶۶۶۶۷	۰/۱۶۶۶۶۶۶۶۶۶۶۶۶۶۷	۰/۰۳۷۰
۴	۰/۶۸۱۹۵۸۴۵۶۲۵۶۰۰۹	۰/۰۱۵۲۹۱۷۸۹۵۸۹۳۴۲	۸/۸۵ × ۱۰$^{-۴}$
۵	۰/۶۸۲۳۲۸۲۴۴۴۰۹۳۹۴	۰/۰۰۰۳۶۹۷۸۸۱۵۳۳۸۵	۱/۰۶ × ۱۰$^{-۶}$
۶	۰/۶۸۲۳۲۷۸۰۳۸۲۶۹۵۶	−۰/۰۰۰۰۰۰۴۴۰۵۸۲۴۳۸	۲/۵۵ × ۱۰$^{-۱۲}$
۷	۰/۶۸۲۳۲۷۸۰۳۸۲۸۰۱۹	۰/۰۰۰۰۰۰۰۰۰۰۱۰۶۳	۱/۱۱ × ۱۰$^{-۱۶}$

که در اینجا $e_k = x_k - \alpha$، ξ_k بین α و x_{k+1}، و c_k بین α و x_{k-1} یا بین α و x_k واقع است. با استدلالی که در روش وتری به‌کار برده شد، به شرطی که α صفر ساده f باشد و f در یک همسایگی α سه بار به طور پیوسته مشتق‌پذیر باشد می‌توان نشان داد که

$$\lim_{k\to\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^p} = \left|\frac{f'''(\alpha)}{6f'(\alpha)}\right|^{(p-1)/2}, \qquad (۳۰.۶)$$

که در آن p یعنی مرتبه روش مولر، ریشه مثبت معادله $p^3 - p^2 - p - 1 = 0$ و تقریبا برابر $p \doteq 1.839$ است.

۵.۶ روش‌های تکراری تک نقطه‌ای

مسئله یافتن ریشه‌ای مانند α از معادله‌ی $f(x) = 0$ را پی می‌گیریم. در روش نیوتن برای پیش‌بینی تقریب بهتری از ریشه، تنها از یک مقدار x استفاده می‌کنیم. در واقع روش نیوتن را می‌توان به‌شکل $x_{k+1} = g(x_k)$ با $k \geqslant 0$ در نظر گرفت که در آن تابع g برابر سمت راست (۱۲.۶) است. بنابراین روش نیوتن از جمله روش‌های تک‌نقطه‌ای است. در روش وتری به‌شکل $x_{k+1} = g(x_k, x_{k-1})$ با $k \geqslant 1$ که در آن تابع g برابر سمت راست (۲۱.۶) است، از دو مقدار x استفاده می‌کنیم. پس روش وتری یک روش دونقطه‌ای است. سرانجام بر اساس (۲۸.۶) روش مولر یک روش سه‌نقطه‌ای است. اگر $\ell \geqslant 0$ یک عدد صحیح باشد، روش‌های تکراری را می‌توان با الگوی کلی زیر نمایش داد

$$x_{k+1} = g(x_k, x_{k-1}, \ldots, x_{k-\ell}) \qquad k = \ell, \ell+1, \ldots. \qquad (۳۱.۶)$$

در این روش تقریب جدید x_{k+1} بر حسب $\ell + 1$ نقطه قبلی x_k، x_{k-1}، ... و $x_{k-\ell}$ به‌دست می‌آید، بنابراین (۳۱.۶) روشی $\ell + 1$ نقطه‌ای است.

چارچوب کار در این بخش به روش‌های تکراری تک‌نقطه‌ای محدود است. به عبارت دقیق‌تر، معادله‌ی $f(x) = 0$ را پس از دستکاری‌هایی جبری به صورت هم‌ارز

$$x = g(x)$$

می‌نویسیم که در آن g نیز یک تابع حقیقی‌مقدار، تک‌متغیره و پیوسته روی $[a,b]$ است. به عنوان نمونه تابع $g(x) = x + \lambda f(x)$، برای عدد حقیقی و دلخواه λ، چنین است. تبدیل معادله‌ی $f(x) = ۰$ به صورت هم‌ارز $x = g(x)$ را می‌توانیم به شکل‌های متفاوتی انجام دهیم و $g(x)$های متفاوتی به‌دست آوریم، با این حال همه آنها مناسب نمی‌باشند. در ادامه شرایطی را بررسی می‌کنیم که به کمک آنها تابع مناسب g و حدس اولیه مناسب $x_۰$ را در طرح تکراری تک‌نقطه‌ای $x_{k+۱} = g(x_k)$ با $k \geqslant ۰$ به‌دست آوریم.

تعریف ۲.۶. عدد حقیقی $\alpha \in [a,b]$ را **نقطه‌ی ثابت** تابع معلوم $g : [a,b] \longmapsto \mathbb{R}$ می‌نامیم اگر $\alpha = g(\alpha)$.

در گزاره زیر موسوم به قضیه نقطه ثابت براور[۱] شرایطی کافی را بررسی می‌کنیم که با فرض آنها تابع نقطه ثابت داشته باشد.

قضیه ۵.۶. (*قضیه نقطه ثابت براور*) *فرض کنیم تابع حقیقی‌مقدار g در بازه بسته و کراندار $[a,b]$ پیوسته باشد و برای هر $x \in [a,b]$ داشته باشیم $g(x) \in [a,b]$. در این صورت تابع g دست‌کم یک نقطه‌ی ثابت $\alpha \in [a,b]$ دارد.*

برهان. برد تابع g در بازه $[a,b]$ واقع است پس، $g(a) \geqslant a$ و $g(b) \leqslant b$. تابع $h(x) = g(x) - x$ روی $[a,b]$ پیوسته است و

$$h(a) = g(a) - a \geqslant ۰ \quad \text{و} \quad h(b) = g(b) - b \leqslant ۰$$

بنابراین طبق قضیه مقدار میانی دست‌کم یک $\alpha \in [a,b]$ وجود دارد که $h(\alpha) = ۰$ یا $\alpha = g(\alpha)$ باشد و در نتیجه g دست‌کم یک نقطه‌ی ثابت در $[a,b]$ دارد. □

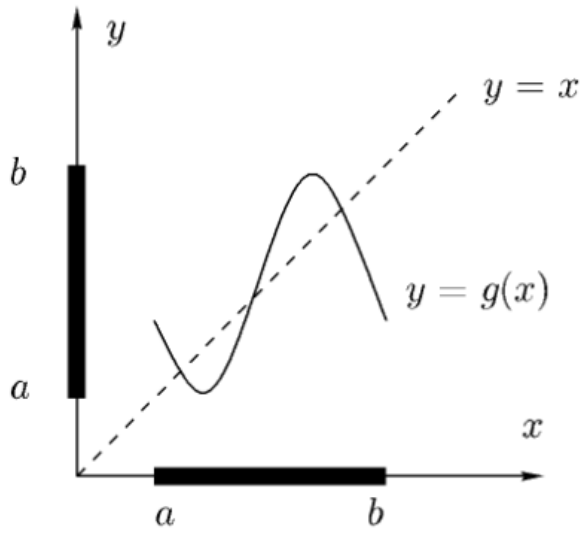
از نظر هندسی، نقاط ثابت در محل برخورد خط $y = x$ و منحنی $y = g(x)$ قرار دارند. برخی تابع‌ها ممکن است نقطه‌ی ثابت نداشته باشند. به عنوان مثال، خط $y = x$ منحنی $y = e^x$ را در هیچ نقطه‌ای قطع نمی‌کند پس تابع نمایی $g(x) = e^x$ نقطه‌ی ثابت ندارد. در شکل ۱۲.۶ نمودار تابع $x \mapsto g(x)$ نمایش داده شده است. تابع g در بازه $[a,b]$ سه نقطه‌ی ثابت دارد و مکان این سه نقطه مولفه‌های x نقاط برخورد نمودار g و خط $y = x$ می‌باشند.

مثال ۱۲.۶. تابع f را در نظر بگیرید که برای هر $x \in [۱,۲]$ به صورت $f(x) = e^x - ۲x - ۱$ تعریف می‌شود. با توجه به اینکه $f(۱) < ۰$ و $f(۲) > ۰$، بنابراین دست‌کم یک α در $[۱,۲]$ وجود دارد که $f(\alpha) = ۰$ باشد. این معادله را به دو شکل متفاوت بازنویسی می‌کنیم.

۱. معادله $e^x - ۲x - ۱ = ۰$ را به صورت $e^x = ۲x + ۱$ و پس از آن $x = \ln(۲x + ۱)$ می‌نویسیم. بنابراین $g_۱(x) = \ln(۲x + ۱)$ اولین تابع به‌دست می‌آید. می‌دانیم

$$g_۱(۱) = \ln(۳) \doteq ۱/۰۹۸۶ \in [۱,۲] \qquad \text{و} \qquad g_۱(۲) = \ln(۵) \doteq ۱/۶۰۹۴ \in [۱,۲]$$

[۱] Luitzen Brouwer (1881-1966)



شکل ۱۲.۶: توصیف هندسی نقاط ثابت g روی بازه $[a,b]$

و چون $g_۱$ روی $[۱,۲]$ پیوسته و اکیداً صعودی است، برای هر $x \in [۱,۲]$ داریم $g_۱(x) \in [۱,۲]$. هم‌اکنون بنا به قضیه ۵.۶ تابع $g_۱$ یک نقطه‌ی ثابت $\alpha \in [۱,۲]$ دارد، یا به طور هم‌ارز، معادله $f(x) = ۰$ یک ریشه در $[۱,۲]$ دارد.

۲. معادله $e^x - ۲x - ۱ = ۰$ را به صورت $۲x = e^x - ۱$ و پس از آن $x = (e^x - ۱)/۲$ می‌نویسیم. بنابراین $g_۲(x) = (e^x - ۱)/۲$ دومین تابع به‌دست می‌آید. می‌دانیم

$$g_۲(۲) = \frac{e^۲ - ۱}{۲} \doteq ۳/۱۹۴۵ \notin [۱,۲] \qquad \text{و} \qquad g_۲(۱) = \frac{e - ۱}{۲} \doteq ۰/۸۵۹۱ \notin [۱,۲]$$

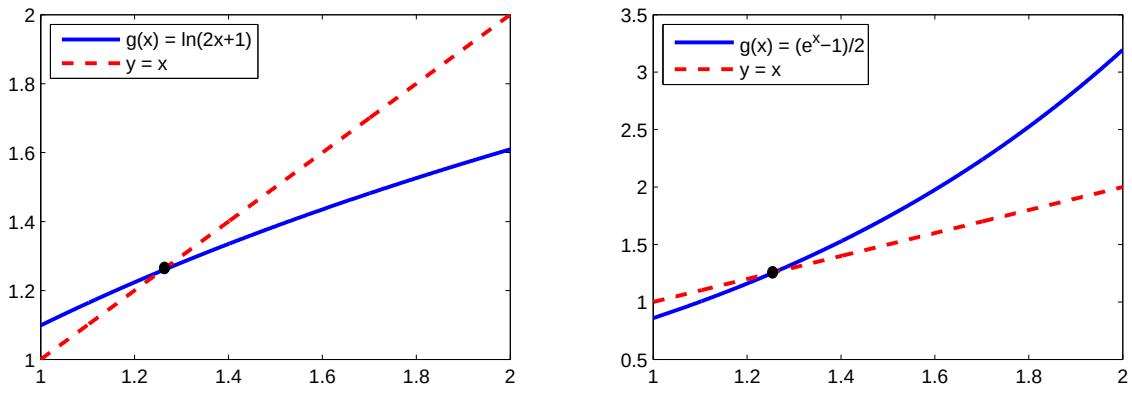
از طرف دیگر $g_۲$ روی $[۱,۲]$ پیوسته و اکیداً صعودی است بنابراین تابع $g_۲$ بازه $[۱,۲]$ را بر روی خودش نمی‌نگارد. در نتیجه قضیه ۵.۶ را نمی‌توان برای تابع $g_۲$ به‌کار برد، اگر چه با توجه به سمت راست شکل ۱۳.۶ این تابع نیز نقطه ثابت دارد! البته این با قضیه براور در تناقض نمی‌باشد زیرا آن قضیه یک شرط کافی برای وجود نقطه ثابت را بیان می‌کند و نه یک شرط لازم.

تعریف زیر اولین گام عملی در این بخش برای ساخت یک الگوریتم برای محاسبه‌ی تقریبی از نقطه‌ی ثابت تابع g، که همان ریشه‌ی معادله $f(x) = ۰$ است، می‌باشد.

تعریف ۳.۶. فرض کنیم تابع حقیقی‌مقدار g در بازه بسته و کراندار $[a,b]$ پیوسته باشد. برای $x_۰ \in [a,b]$ رابطه بازگشتی

$$x_{k+۱} = g(x_k), \quad k = ۰, ۱, ۲, \ldots \tag{۳۲.۶}$$

را **تکرار نقطه‌ی ثابت** یا **تکرار ساده** و تابع g **تابع تکرار** گفته می‌شود.



شکل ۱۳.۶: توصیف هندسی نقاط ثابت g_1 و g_2 روی بازه $[1, 2]$ در مثال ۱۲.۶

در (۳۲.۶) تقریب x_{k+1} تنها به تقریب ماقبل خود یعنی x_k وابسته است بنابراین روش تکرار نقطه‌ی ثابت یک روش تکراری تک‌نقطه‌ای می‌باشد. اگر دنباله $\{x_k\}$ در (۳۲.۶) همگرا باشد، حد آن بایستی نقطه‌ی ثابت تابع g شود، زیرا تابع g روی بازه بسته $[a, b]$ پیوسته است. در واقع، با فرض $\alpha = \lim_{k\to\infty} x_k$ داریم

$$\alpha = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} g(x_k) = g\big(\lim_{k\to\infty} x_k\big) = g(\alpha).$$

مثال ۱۳.۶. ماشین حساب جیبی خود را در حالت رادیان قرار دهید. با وارد کردن عدد ۱ و فشار دادن پی‌درپی کلید کسینوس، دنباله‌ای از اعداد حقیقی به‌شکل زیر تولید می‌شود

$$\begin{aligned} x_1 &= \cos(1) \doteq ۰/۵۴۰۳, & x_2 &= \cos(x_1) \doteq ۰/۸۵۷۶, & &\dots \\ x_9 &= \cos(x_8) \doteq ۰/۷۳۱۴, & x_{10} &= \cos(x_9) \doteq ۰/۷۴۴۲, & &\dots \\ x_{19} &= \cos(x_{18}) \doteq ۰/۷۳۸۹, & x_{20} &= \cos(x_{19}) \doteq ۰/۷۳۹۲, & &\dots \end{aligned}$$

این دنباله، اگرچه به کندی اما، سرانجام به $\alpha \doteq ۰/۷۳۹۰۸۵۱۳۳۲۱۵۱۶۱$ میل می‌کند. با توجه به روش ساخت دنباله‌ی $x_{k+1} = \cos(x_k)$ با $x_0 = ۱$، حد α در معادله‌ی $\alpha = \cos(\alpha)$ صدق می‌کند، یعنی α یک نقطه‌ی ثابت از تابع کسینوس است.

در ادامه یک شرط کافی برای همگرایی دنباله $\{x_k\}$ ارائه می‌کنیم، قبل از آن تعریف زیر را می‌آوریم.

تعریف ۴.۶. (انقباض) فرض کنیم تابع حقیقی مقدار g روی بازه بسته و کراندار $[a, b]$ پیوسته باشد. تابع g یک انقباض روی $[a, b]$ گفته می‌شود اگر ثابت $0 < L < 1$ موجود باشد طوری که برای هر x و y در $[a, b]$ داشته باشیم

$$|g(x) - g(y)| \leqslant L|x - y|. \qquad (۳۳.۶)$$

اصطلاح «انقباض» ریشه در این واقعیت دارد که هرگاه (۳۳.۶) برای $0 < L < 1$ برقرار باشد، فاصله بین تصویر نقاط x و y تحت g یعنی $|g(x) - g(y)|$، کوچکتر از فاصله بین نقاط x و y یعنی $|x - y|$ است. به طور کلی، هرگاه L یک عدد حقیقی مثبت دلخواه باشد، (۳۳.۶) به شرط لیپشیتس موسوم است. هم‌اکنون آماده‌ایم که اساسی‌ترین نتیجه این بخش را ارائه کنیم.

قضیه ۶.۶. (*قضیه نگاشت انقباضی*) *فرض کنیم تابع حقیقی‌مقدار g در بازه بسته و کراندار $[a, b]$ پیوسته باشد و برای هر $x \in [a, b]$ داشته باشیم $g(x) \in [a, b]$، به علاوه g یک انقباض روی $[a, b]$ باشد. در این صورت g یک نقطه ثابت یکتا $\alpha \in [a, b]$ دارد. همچنین دنباله‌ی $\{x_k\}$ در (۳۲.۶) برای هر حدس اولیه‌ی $x_0 \in [a, b]$ به α همگرا می‌شود.*

برهان. وجود نقطه‌ی ثابت α برای g نتیجه‌ای از قضیه ۵.۶ است. یکتایی نقطه ثابت از (۳۳.۶) با برهان خلف نتیجه می‌شود. در واقع، اگر دو نقطه‌ی ثابت α_1 و α_2 هر دو در $[a, b]$ برای g موجود باشد، آنگاه

$$|\alpha_1 - \alpha_2| = |g(\alpha_1) - g(\alpha_2)| \leqslant L|\alpha_1 - \alpha_2| < |\alpha_1 - \alpha_2|,$$

که امکان‌پذیر نمی‌باشد، درنتیجه $\alpha_1 = \alpha_2$. در ادامه نشان می‌دهیم که دنباله‌ی x_k تعریف شده با (۳۲.۶) وقتی $k \to \infty$، برای هر حدس اولیه‌ی $x_0 \in [a, b]$ به نقطه‌ی ثابت و یکتای α همگرا می‌شود. با توجه به (۳۳.۶) داریم

$$|x_{k+1} - \alpha| = |g(x_k) - g(\alpha)| \leqslant L|x_k - \alpha|, \qquad k \geqslant 0,$$

که از آن به استقرای ریاضی نتیجه می‌شود

$$|x_k - \alpha| \leqslant L^k|x_0 - \alpha|, \qquad k \geqslant 1. \tag{۳۴.۶}$$

چون $L \in (0, 1)$، داریم $\lim_{k\to\infty} L^k = 0$، و بنابراین $\lim_{k\to\infty} |x_k - \alpha| = 0$ یعنی دنباله $\{x_k\}$ به نقطه ثابت یکتای α همگرا است. □

در عمل یافتن ثابت L چگونه است؟ فرض کنیم g روی بازه بسته $[a, b]$ پیوسته و درون این بازه یعنی (a, b) مشتق‌پذیر باشد، در این صورت بنا به قضیه مقدار میانگین برای هر x و y در بازه $[a, b]$، نقطه‌ی ξ بین x و y وجود دارد که

$$|g(x) - g(y)| = |g'(\xi)||x - y|. \tag{۳۵.۶}$$

بنابراین یافتن کرانی بالا برای $|g'(\xi)|$ معادل یافتن کران L در (۳۳.۶) است. قضیه‌ی ۷.۶ را هم در ادامه ببینید. اما قبل از آن به مثال زیر توجه کنید.

مثال ۱۴.۶. تابع f را در نظر بگیرید که برای هر $x \in [1, 2]$ به صورت $f(x) = e^x - 2x - 1$ تعریف می‌شود. در مثال ۱۲.۶ نشان دادیم که معادله $f(x) = 0$ دارای جواب $\alpha \in [1, 2]$ است، همچنین α نقطه ثابت $g(x) = \ln(2x + 1)$ می‌باشد. تابع g روی $[1, 2]$ پیوسته و روی $(1, 2)$ مشتق‌پذیر است. به علاوه

$$g'(x) = \frac{2}{2x + 1}, \qquad g''(x) = \frac{-4}{(2x + 1)^2}.$$

چون برای هر $x \in [۱, ۲]$، $g''(x) < ۰$، بنابراین g' روی $[۱, ۲]$ اکیداً نزولی است و برای $۱ < \xi < ۲$ داریم

$$\frac{۲}{۵} = g'(۲) \leqslant g'(\xi) \leqslant g'(۱) = \frac{۲}{۳},$$

و بنا به (۳۵.۶) برای هر x و y در بازه $[۱, ۲]$ نتیجه می‌گیریم

$$|g(x) - g(y)| \leqslant \frac{۲}{۳}|x - y|.$$

هم‌اکنون با توجه به قضیه نگاشت انقباضی، دنباله‌ی $\{x_k\}$ به‌صورت

$$x_{k+۱} = \ln(۲x_k + ۱), \qquad k = ۰, ۱, ۲, \ldots,$$

برای هر $x_۰ \in [۱, ۲]$ به α همگرا است. فرض کنید $x_۰ = ۱$، در این صورت برخی از جملات این دنباله به صورت زیر هستند:

$$x_۱ \doteq ۱/۰۹۸۶۱۲, \quad x_۲ \doteq ۱/۱۶۲۲۸۳, \quad x_۳ \doteq ۱/۲۰۱۳۳۹, \ldots$$
$$x_{۱۳} \doteq ۱/۲۵۶۲۲۷, \quad x_{۱۴} \doteq ۱/۲۵۶۳۱۵, \quad x_{۱۵} \doteq ۱/۲۵۶۳۶۵, \ldots$$

به نظر می‌رسد روش به کندی همگرا می‌شود زیرا پس از ۱۵ تکرار تنها سه رقم اعشار درست از α را یافته است.

می‌خواهیم تعداد تکرارهای لازم $K(\varepsilon)$ برای تضمین یک تقریب از ریشه با دقت ε را تعیین کنیم. به عبارت دقیق‌تر می‌خواهیم $K(\varepsilon)$ را طوری تعیین کنیم که برای هر $k > K(\varepsilon)$، داشته باشیم $|x_k - \alpha| \leqslant \varepsilon$. برای انجام این کار از برآورد (۳۴.۶) در قضیه نگاشت انقباضی استفاده می‌کنیم. با توجه به نابرابری مثلثی داریم

$$|x_۰ - \alpha| \leqslant |x_۰ - x_۱| + |x_۱ - \alpha| \leqslant |x_۰ - x_۱| + L|x_۰ - \alpha|,$$

بنابراین

$$|x_۰ - \alpha| \leqslant \frac{۱}{۱ - L}|x_۰ - x_۱|.$$

با به‌کارگیری این نابرابری در (۳۴.۶) یک **کران پیشرو** به صورت زیر به‌دست می‌آید

$$|x_k - \alpha| \leqslant \frac{L^k}{۱ - L}|x_۰ - x_۱|. \tag{۳۶.۶}$$

با استفاده از کران پیشرو، تنها پس از یک تکرار و یافتن $x_۱$ می‌توانیم تعداد تکرارهای لازم برای رسیدن به دقت مطلوب ε را تعیین کنیم. به عبارت دقیق‌تر $|x_k - \alpha| \leqslant \varepsilon$ نتیجه می‌دهد

$$\frac{L^k}{۱ - L}|x_۰ - x_۱| \leqslant \varepsilon.$$

با لگاریتم طبیعی گرفتن از این نابرابری نتیجه می‌گیریم

$$k \geqslant \frac{\ln|x_۱ - x_۰| - \ln(\varepsilon(۱ - L))}{\ln(۱/L)}.$$

بنابراین عدد صحیح $K(\varepsilon)$ وابسته به ε که برای هر $k \geqslant K(\varepsilon)$ داشته باشیم $|x_k - \alpha| \leqslant \varepsilon$ با دستور زیر به‌دست می‌آید

$$K(\varepsilon) = \left\lceil \frac{\ln|x_۱ - x_۰| - \ln(\varepsilon(۱ - L))}{\ln(۱/L)} \right\rceil. \tag{۳۷.۶}$$

لازم به ذکر است که با توجه روند تعیین $K(\varepsilon)$، این تعداد تکرار، "حداکثر" تکرار مورد نیاز برای رسیدن به دقت پیشرو ε است و در عمل ممکن است با تعداد تکرار کمتری هم این دقت حاصل شود.

مثال ۱۵.۶. به مثال ۱۴.۶ باز می‌گردیم و می‌خواهیم ماکسیمم تعداد تکرارهای مورد نیاز روش تکرار نقطه ثابت را طوری تعیین کنیم که تقریب به‌دست آمده دست‌کم تا پنج رقم اعشار درست باشد. پس داریم $\varepsilon = \frac{۱}{۲} \times ۱۰^{-۵}$. از مثال ۱۴.۶ داریم $L = \frac{۲}{۳}$، پس دستور (۳۷.۶) نتیجه می‌دهد که $K(\varepsilon) = \lceil ۲۷/۱۰۰۱ \rceil = ۲۸$. در واقع ۲۸ حداکثر تکرار لازم برای رسیدن به پنج رقم اعشار درست بر طبق تئوریِ ارائه شده است. نتایج عددی نشان می‌دهند که $x_{۲۰}$ تا پنج رقم اعشار درست است.

در عمل انتخاب بازه‌ی $[a, b]$ که شرایط قضیه ۶.۶ را به طور کامل برآورده کند کار مشکلی است. در چنین حالت‌هایی، مفهوم همگرایی موضعی به‌شکل زیر مفید می‌باشد.

تعریف ۵.۶. یک روش تکراری **موضعاً همگرا** به α نامیده می‌شود اگر روش برای حدس‌های اولیه‌ی به اندازه کافی نزدیک به α همگرا باشد.

به عبارت دیگر، یک روش تکراری به ریشه‌ی α موضعاً همگرا است اگر یک همسایگی از α موجود باشد طوری که برای هر حدس اولیه $x_۰$ در این همسایگی دنباله $\{x_k\}$ حاصل از این روش به α همگرا شود. به عنوان مثال، طبق قضیه‌ی ۲.۶، روش نیوتن در حالت کلی یک روش موضعاً همگراست زیرا لازم است $x_۰$ در بازه‌ی $[\alpha - \delta, \alpha + \delta]$ انتخاب شود که $\delta \leqslant \min\{r, ۱/M\}$.

قضیه‌ی زیر همگرایی موضعی روش‌های نقطه ثابت را بیان می‌کند.

قضیه ۷.۶. *گیریم α یک نقطه‌ی ثابت تابع پیوسته-مشتق‌پذیر g در $I_\delta := \{x \in \mathbb{R} : |x - \alpha| \leqslant \delta\}$ باشد که در آن δ بگونه‌ای انتخاب شده است که*

$$\max_{t \in I_\delta} |g'(t)| \leqslant L < ۱. \tag{۳۸.۶}$$

آنگاه α یکتاست و تکرار ساده‌ی $x_{k+۱} = g(x_k)$ برای هر $x_۰ \in I_\delta$ با مرتبه‌ی حداقل خطی به α همگراست.

برهان. با توجه به قضیه مقدار میانگین برای هر $x, y \in I_\delta$ داریم

$$|g(x) - g(y)| = |g'(\xi)||x - y| \leqslant L|x - y|, \quad \xi \in I_\delta,$$

که نشان می‌دهد g یک انقباض روی I_δ است. اکنون قضیه‌ی ۶.۶ با جایگزینی بازه‌ی $[a, b]$ با I_δ یکتایی نقطه ثابت و همگرایی دنباله نقطه ثابت را نتیجه می‌دهد و داریم

$$|x_{k+۱} - \alpha| = |g(x_k) - g(\alpha)| \leqslant |g(\xi_k)||x_k - \alpha| < |x_k - \alpha|,$$

که ξ_k بین x_k و $x_{k+۱}$ واقع است. این نشان می‌دهد اگر $x_k \in I_\delta$ آنگاه $x_{k+۱} \in I_\delta$، یعنی همه‌ی تکرارها در I_δ باقی می‌مانند. همچنین اگر قسمت اول این رابطه را بدون قدرمطلق بنویسیم داریم

$$\frac{x_{k+۱} - \alpha}{x_k - \alpha} = g'(\xi_k), \quad k = ۰, ۱, \ldots, \tag{۳۹.۶}$$

با حد گرفتن از طرفین و با توجه به فرض پیوستگی g' و قرار گرفتن ξ_k بین x_k و $x_{k+۱}$ نتیجه می‌گیریم

$$\lim_{k\to\infty} \frac{x_{k+۱} - \alpha}{x_k - \alpha} = g'(\alpha). \tag{۴۰.۶}$$

از آنجا که $|g'(\alpha)| < ۱$ نتیجه می‌گیریم همگرایی روش حداقل خطی است. □

با توجه به (۳۴.۶) و (۴۰.۶) می‌توان دریافت که تکرار نقطه‌ی ثابت دست‌کم همگرای خطی است، یعنی برای kی به اندازه کافی بزرگ، رفتار خطا در گام $k + ۱$ همانند خطای گام kام می‌باشد که در یک ثابت (L در (۳۴.۶) و $g'(\alpha)$ در (۴۰.۶)) مستقل از k با قدرمطلق کمتر از ۱ ضرب می‌شود. به همین دلیل، این ثابت ضریب مجانبی همگرایی نامیده می‌شود. توجه کنید ضریب مجانبی همگرایی کوچکتر، همگرایی سریعتری را در پی خواهد داشت.

ملاحظه ۲.۶. اگر در هر همسایگی I_δ از α داشته باشیم $|g'(t)| > ۱$ برای $t \in I_\delta$، و بخصوص $|g'(\alpha)| > ۱$، آنگاه طبق ۳۹.۶ داریم $|x_{k+۱} - \alpha| > |x_k - \alpha|$ و دنباله نمی‌تواند به نقطه‌ی ثابت همگرا شود. اگر $|g'(\alpha)| = ۱$ ممکن است همگرایی یا واگرایی رخ دهد، و این به ویژگی‌های تابع تکرار g وابسته می‌باشد.

مثال ۱۶.۶. می‌خواهیم صفر تابع $f(x) = x^۳ + x - ۱$ واقع در بازه‌ی $[۰, ۱]$ را با تکرار نقطه‌ی ثابت بیابیم. برای انجام این کار، سه تابع تکرار به‌شکل زیر می‌سازیم

$$g_۱(x) = ۱ - x^۳, \quad g_۲(x) = \sqrt[۳]{۱ - x}, \quad g_۳(x) = \frac{۱ + ۲x^۳}{۱ + ۳x^۲},$$

با فرض $x_۰ = ۰/۵$ تکرارهای نقطه ثابت $x_{k+۱} = g_i(x_k)$، $i = ۱, ۲, ۳$ را در جدول ۵.۶ می‌آوریم. ملاحظه می‌شود دنباله تولید شده با تابع تکرار $g_۱$ واگرا است در حالی که دنباله‌های تولید شده با توابع تکرار $g_۲$ و $g_۳$ (البته با سرعت‌های متفاوت) همگرا می‌باشند. برای توجیه این رفتارها از قضیه‌ی ۷.۶ کمک می‌گیریم. مشتق توابع تکرار چنین است

$$g_۱'(x) = -۳x^۳, \quad g_۲'(x) = -\frac{۱}{۳\sqrt[۳]{(۱ - x)^۲}}, \quad g_۳'(x) = \frac{۶x^۴ + ۶x^۲ - ۶x}{(۱ + ۳x^۲)^۲}.$$

جدول ۵.۶: روش تکرار نقطه‌ی ثابت برای یافتن صفر تابع $f(x)$

k	$x_{k+۱} = g_۱(x_k)$	$x_{k+۱} = g_۲(x_k)$	$x_{k+۱} = g_۳(x_k)$
۱	۰/۸۷۵۰	۰/۷۹۳۷۰۰۵۲۵۹۸۴۱۰۰	۰/۷۱۴۲۸۵۷۱۴۲۸۵۷۱۴
۲	۰/۳۳۰۰۷۸۱۲۵۰	۰/۵۹۰۸۸۰۱۱۳۲۷۵۱۷۷	۰/۶۸۳۱۷۹۷۲۳۵۰۲۳۰۴
۳	۰/۹۶۴۰۳۷۴۷۰۵	۰/۷۴۲۳۶۳۹۳۲۱۶۸۰۰۶	۰/۶۸۲۳۲۸۴۲۳۳۰۴۵۷۸
۴	۰/۱۰۴۰۵۴۱۸۸۳	۰/۶۳۶۳۱۰۲۰۳۴۸۱۶۶۱	۰/۶۸۲۳۲۷۸۰۳۸۲۸۳۴۷
۵	۰/۹۹۸۸۷۳۳۷۶	۰/۷۱۳۸۰۰۸۱۴۱۴۴۲۰۷	۰/۶۸۲۳۲۷۸۰۳۸۲۸۰۱۹
۶	۰/۰۰۳۳۷۶۰۶۳۲	۰/۶۵۹۰۰۶۱۴۵۶۲۲۴۰۰	
۷	۰/۹۹۹۹۹۹۹۶۱	۰/۶۹۸۶۳۲۶۰۵۷۳۰۲۱۹	
۸	۰/۰۰۰۰۰۰۱۱۵۴۳	۰/۶۷۰۴۴۸۴۹۶۲۲۸۰۷۲	
۹	۱/۰۰۰۰۰۰۰۰۰۰۰۰۰۰	۰/۶۹۰۷۲۹۱۲۰۵۸۹۱۴۱	
۱۰	۰/۰۰۰۰۰۰۰۰۰۰۰۰۰۰	۰/۶۷۶۲۵۸۹۲۴۹۲۶۸۲۷	
⋮	⋮	⋮	
۹۹	۱/۰۰۰۰۰۰۰۰۰۰۰۰۰۰	۰/۶۸۲۳۲۷۸۰۳۸۲۸۰۲۰	
۱۰۰	۰/۰۰۰۰۰۰۰۰۰۰۰۰۰۰	۰/۶۸۲۳۲۷۸۰۳۸۲۸۰۱۹	

اگر فرض کنیم $\alpha \doteq ۰/۶۸۲۳۲۷۸۰۳۸۲۸۰۱۹$، در آن صورت

$$|g_۱'(\alpha)| = ۱/۳۹۶۷ > ۱, \quad |g_۲'(\alpha)| = ۰/۷۱۶ < ۱, \quad |g_۳'(\alpha)| = ۰ < ۱.$$

در مورد $g_۱$ واضح است که نمی‌توان همسایگی‌ای از α یافت که اندازه‌ی مشتق در آن کمتر از یک باشد. قطعاً چنین همسایگی‌هایی برای $g_۲$ و $g_۳$ وجود دارند، چرا که $g_۲'$ و $g_۳'$ در همسایگی α پیوسته‌اند. ضریب مجانبی همگرایی مربوط به $g_۳$ بسیار کوچکتر از ضریب مجانبی همگرایی مربوط به $g_۲$ است و بنابراین انتظار می‌رود که دنباله تولید شده با تابع تکرار $g_۳$ بسیار سریع‌تر به α همگرا شود و اینگونه نیز هست!

در مثال ۱۶.۶ مشتق دوم تابع تکرار $g_۳$ نیز در α صفر است یعنی $g_۳''(\alpha) = ۰$. آیا صفر شدن مشتق دوم تابع تکرار در α می‌تواند دلیل دیگری برای همگرایی سریع دنباله تولید شده با آن باشد؟

قضیه ۸.۶. *فرض کنیم α یک نقطه‌ی ثابت برای تابع g باشد که در همسایگی I_δ تعریف شده در قضیه‌ی ۷.۶ باشد. همچنین فرض کنیم $g \in C^p(I_\delta)$ که برای $p \geqslant ۲$. به علاوه گیریم*

$$g'(\alpha) = \cdots = g^{(p-۱)}(\alpha) = ۰, \quad g^{(p)}(\alpha) \neq ۰. \tag{۴۱.۶}$$

در این صورت اگر حدس اولیه‌ی $x_۰$ در I_δ انتخاب شود آنگاه دنباله‌ی $x_{k+۱} = g(x_k)$ به α همگرا از مرتبه‌ی p است و

$$\lim_{k\to\infty} \frac{x_{k+۱} - \alpha}{(x_k - \alpha)^p} = \frac{۱}{p!}\, g^{(p)}(\alpha). \tag{۴۲.۶}$$

برهان. یکتایی نقطه ثابت و همگرایی دنباله بدون تغییر همانند قضیه‌ی ۷.۶ است. برای اثبات رابطه‌ی (۴۲.۶) بسط g حول α را به صورت زیر می‌نویسیم

$$g(x) = g(\alpha) + (x-\alpha)g'(\alpha) + \cdots + \frac{(x-\alpha)^{p-۱}}{(p-۱)!}g^{(p-۱)}(\alpha) + \frac{(x-\alpha)^p}{p!}g^{(p)}(\xi),$$

که در آن ξ بین x و α قرار دارد. با توجه به (۴۱.۶) و $\alpha = g(\alpha)$، با جایگزین کردن $x := x_k$ داریم

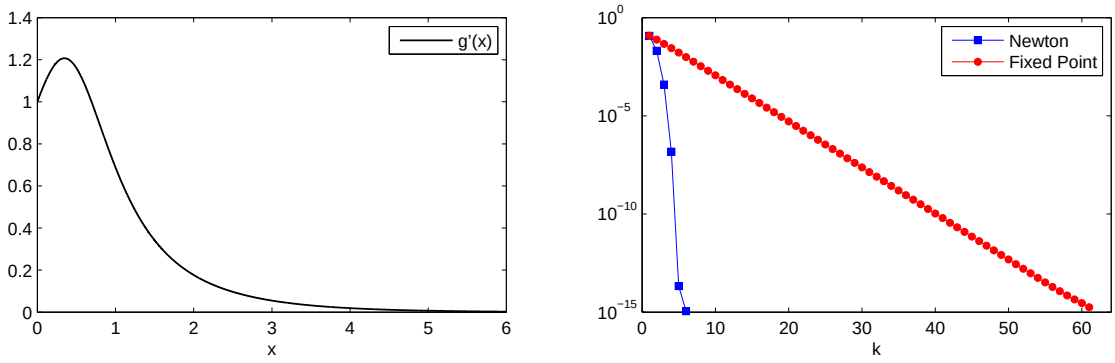
$$x_{k+۱} = \alpha + ۰ + \cdots + ۰ + \frac{(x_k-\alpha)^p}{p!}g^{(p)}(\xi_k),$$

که در آن ξ_k بین x_k و α قرار دارد. بنابراین

$$\frac{x_{k+۱} - \alpha}{(x_k - \alpha)^p} = \frac{۱}{p!}g^{(p)}(\xi_k).$$

چون $x_k \to \alpha$ پس $\lim_{k\to\infty} \xi_k = \alpha$. با توجه به پیوستگی $g^{(p)}$ در همسایگی α، رابطه‌ی (۴۲.۶) نتیجه می‌شود. □

با توجه به این که در مثال ۱۶.۶ داریم $g_۳'(\alpha) = ۰$ و $g_۳''(\alpha) \doteq ۱/۷۰۸۲$، بنابراین همگرایی دنباله‌ی $x_{k+۱} = g_۳(x_k)$ به α از مرتبه‌ی ۲ است.



شکل ۱۴.۶: سمت چپ: نمودار $y = g'(x)$، سمت راست: مقایسه‌ی سرعت همگرایی دو روش تکراری

مثال ۱۷.۶. معادله‌ی $f(x) = x - \arctan(e^x - ۱) = ۰$ را در نظر می‌گیریم. می‌خواهیم ریشه‌ی بزرگتر معادله، $\alpha > ۱$، را با تکرار نقطه‌ی ثابت بیابیم. برای انجام این کار، طرح تکراری زیر را در نظر می‌گیریم

$$x_{k+۱} = g(x_k), \quad g(x) = \arctan(e^x - ۱).$$

تابع g' در بازه‌ی $[۱, \infty)$ نزولی است (شکل ۱۴.۶ سمت چپ) و ماکسیمم آن در $x = ۱$ تقریباً برابر ۰٫۶۸۷۷ می‌باشد. بنابراین

$$۰ \leqslant g'(x) \leqslant g'(۱) \doteq ۰٫۶۸۷۷, \quad x \geqslant ۱,$$

و دنباله تولید شده با تابع تکرار g با هر شروع اولیه در بازه‌ی $[۱, \infty)$ به $\alpha \doteq ۱٫۱۱۹۴۹۹۲۰۹۴۲۶۲۳۳$ همگرا است. اگر چه قضیه‌ی ۷.۶ همگرایی را برای انتخاب‌های $x_۰$ در بازه‌ای حول ریشه تضمین می‌کند، با این حال این مثال برای تمام $x_۰ \in [۱, \infty)$ نیز همگراست. توجه داشته باشید که شرایط آن قضیه کافی هستند نه لازم. حتی اگر $x_۰$ را عددی بسیار بزرگ در نظر بگیرید با یک تکرار به نزدیکی ریشه خواهید رسید، اما پس از آن آنچنان که در شکل ۱۴.۶ سمت راست دیده می‌شود سرعت همگرایی کند می‌باشد و در حدود ۶۰ تکرار به دقت ماشین می‌رسد، در حالی که روش نیوتن در حدود ۱۰ برابر سریع‌تر یعنی با ۶ تکرار جواب این مسئله را با دقت ماشین ارائه می‌دهد! قطعاً این قابل پیش‌بینی بود چرا که مرتبه همگرایی روش نیوتن برای این تابع مربعی است، و روش نقطه ثابت ارائه شده در این مثال دارای مرتبه همگرایی خطی است زیرا $g'(\alpha) \doteq ۰٫۵۸۲۷$.

۶.۶ معادلات جبری

در این بخش حالتی را در نظر می‌گیریم که در آن تابع f یک چندجمله‌ای جبری از درجه‌ی $n \geqslant ۰$ به صورت زیر است

$$p_n(x) = a_۰ + a_۱x + a_۲x^۲ + \cdots + a_nx^n. \qquad (۴۳.۶)$$

در اینجا ضرایب a_j همگی حقیقی می‌باشند. یادآوری می‌کنیم که اگر $\alpha \in \mathbb{C}$ یک ریشه مختلط p_n باشد، آنگاه مزدوج آن یعنی $\overline{\alpha}$ نیز یک ریشه از p_n است. بنابراین تعداد ریشه‌های مختلط همواره زوج است.
قضیه آبل تضمین می‌کند اگر n بزرگتر از چهار باشد دستوری صریح برای محاسبه تمام صفرهای چندجمله‌یی p_n وجود ندارد. این واقعیت انگیزه بیشتری را برای استفاده از روش‌های عددی در حل معادله $p_n(x) = ۰$ ایجاد می‌کند. گرچه معادلات چندجمله‌ای را می‌توان با هر یک از روش‌های تکراری که قبلاً در این فصل بحث شده حل کرد، اما بدلیل کاربردهای فراوان آنها نیاز به بحثی خاص دارند. در واقع قضیه‌های مفید زیادی درباره ریشه‌های یک معادله چندجمله‌ای می‌توان بیان کرد که درباره تابع‌های دیگر درست نخواهند بود.

آن گونه که در بخش‌های قبل آمد انتخاب یک حدس اولیه مناسب $x_۰$ یا یک بازه جستجوی مناسب $[a, b]$ برای ریشه با اهمیت است. در مورد چندجمله‌ایها بر اساس گزاره‌هایی که خواهد آمد این کار تا حدی امکان‌پذیر است. ابتدا قضیه زیر را بدون اثبات بیان می‌کنیم. برای اثبات به ... مراجعه کنید.

قضیه ۹.۶. ***(قاعده علامت‌های دکارت)*** *فرض کنیم ν تعداد تغییر علامت‌ها در ضرایب a_j و k تعداد ریشه‌های حقیقی مثبت چندجمله‌ای p_n (با شمردن چندگانگی آنها) باشد. در این صورت $\nu \geqslant k$ و $\nu - k$ زوج است.*

از آنجایی که ریشه‌های مثبت $p_n(-x)$، همان ریشه‌های منفی $p_n(x)$ هستند، قاعده تغییر علامت‌های دکارت را می‌توان برای محدود کردن تعداد ریشه‌های حقیقی منفی نیز به‌کار برد.

مثال ۱۸.۶. درباره ریشه‌های حقیقی چندجمله‌ای زیر اطلاعاتی به‌دست دهید.

$$p_۴(x) = x^۴ - x - ۱$$

چون ضرایب $p_۴$ دارای $\nu = ۱$ تغییر علامت است برای زوج شدن $k - ۱$ لازم است تعداد ریشه‌های حقیقی مثبت $k = ۱$ باشد. از طرف دیگر ضرایب $p_۴(-x) = x^۴ + x - ۱$ دارای $\nu' = ۱$ تغییر علامت است و برای زوج شدن $k' - ۱$ لازم است تعداد ریشه‌های حقیقی منفی $k' = ۱$ باشد. بنا بر قضیه اساسی جبر $p_۴$ چهار ریشه (حقیقی و مختلط) دارد، پس چندجمله‌ای $p_۴$ یک ریشه حقیقی مثبت، یک ریشه حقیقی منفی و دو ریشه مختلط دارد.

قضیه ۱۰.۶. ***(قضیه کوشی)*** *تمام ریشه‌های چندجمله‌ای p_n درون دایره‌ای به مرکز مبدا و شعاع $r + ۱$ قرار دارند که در آن*

$$r := \max_{۰ \leqslant k \leqslant n-۱} \left| \frac{a_k}{a_n} \right|.$$

برهان. اثبات این قضیه در تمرین ۳۴.۶ به کمک قضیه گرشگورین و ماتریس همراه خواسته شده است. □

در مثال ۱۸.۶ با توجه به این که $r = ۱$، بنابراین تمام ریشه‌ها درون دایره‌ای به مرکز مبدا و شعاع ۲ قرار دارند. شایان ذکر است در صورتی که r بزرگ باشد کران ارائه شده بی‌فایده است.

۱.۶.۶ الگوریتم هورنر

در این بخش روشی کارآمد به نام الگوریتم هورنر[۱] را برای محاسبه یک چندجمله‌ای (و مشتق آن) در نقطه‌ی z شرح می‌دهیم. این الگوریتم یک فرایند خودکار به نام روش تقلیل را برای تقریب تدریجی تمام ریشه‌های یک چندجمله‌ای به وجود می‌آورد. از دیدگاه جبری، (۴۳.۶) با نمایش زیر هم‌ارز است

$$p_n(x) = a_۰ + x(a_۱ + x(a_۲ + \cdots + x(a_{n-۱} + a_n x)\cdots)) \tag{۴۴.۶}$$

عبارت (۴۴.۶) به الگوریتم ضرب تودرتو نیز معروف است و مبنای الگوریتم هورنر می‌باشد. به کمک آن الگوریتمی کارآمد برای محاسبه‌ی چندجمله‌ای p_n در نقطه‌ی z به صورت زیر نوشته می‌شود

$$\begin{aligned} b_n &= a_n \\ b_k &= a_k + b_{k+۱} z, \qquad k = n-۱, n-۲, \ldots, ۰ \end{aligned} \tag{۴۵.۶}$$

در (۴۵.۶) تمام ضرایب b_k برای $k \leqslant n-۱$ به z وابسته هستند و به سادگی می‌توان بررسی کرد که $b_۰ = p_n(z)$. دستور (۴۴.۶) به n جمع و n ضرب برای محاسبه‌ی $p_n(z)$ نیاز دارد که یک صرفه‌جویی قابل ملاحظه نسبت به محاسبه مستقیم با فرمول (۴۳.۶) است. الگوریتم (۴۵.۶) در برنامه زیر اجرا شده است. ورودی‌ها مقدار z و ضرایب a_j است که در بردار a از $a_۰$ تا a_n ذخیره شده‌اند. خروجی‌ها بردار b شامل ضرایب b_j و مقدار $p_n(z)$ است که در pz ذخیره می‌شود.

```
function [pz,b] = Horner(a,z)
n = length(a); b=zeros(n,1);
b(n) = a(n);
for k = n-1:-1:1
    b(k) = a(k) + b(k+1)*z;
end
pz = b(1); b = b(2:end);
```

در ادامه یک الگوریتم کارآمد را معرفی می‌کنیم که با در دست داشتن یک ریشه (یا تقریبی از آن)، تمام ریشه‌های چندجمله‌ای را محاسبه می‌کند. ابتدا چندجمله‌ای

$$q_{n-۱}(x; z) = b_۱ + b_۲ x + \cdots + b_n x^{n-۱} \tag{۴۶.۶}$$

[۱] William George Horner (1786 – 1837)

از درجه‌ی $n-1$ بر حسب x را تعریف می‌کنیم که از راه ضرایب b_k به پارامتر z وابسته است، و به آن چندجمله‌ای وابسته‌ی p_n می‌گوییم. با توجه به

$$\begin{aligned} b_0 + (x-z)q_{n-1}(x;z) &= b_0 + (x-z)(b_1 + b_2 x + \cdots + b_n x^{n-1}) \\ &= (b_0 - b_1 z) + (b_1 - b_2 z)x + (b_2 - b_3 z)x^2 + \cdots + b_n x^n \\ &= a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n, \end{aligned}$$

نتیجه می‌گیریم

$$p_n(x) = b_0 + (x-z)q_{n-1}(x;z).$$

در واقع q_{n-1} خارج قسمت و b_0 باقی‌مانده تقسیم تقسیم p_n بر $x-z$ هستند. حال اگر z یک ریشه p_n باشد در آن صورت $b_0 = p_n(z) = 0$ و بنابراین $p_n(x) = (x-z)q_{n-1}(x;z)$. در این حالت معادله جبری $q_{n-1}(x;z) = 0$ سایر $n-1$ ریشه را فراهم می‌کند. بنابراین برای پیدا کردن ریشه‌های دیگرِ p_n، می‌توانیم جستجوی خود را به ریشه‌های $q_{n-1}(\cdot;z)$ محدود کنیم. به این ترتیب برای محاسبه تمام ریشه‌های p_n معیار تقلیل زیر اتخاذ می‌شود. برای $m = n, n-1, \ldots, 1$، مراحل زیر را انجام دهید:

۱. با یک روش تقریبی مناسب ریشه‌ی α_m از p_m را بیابید؛

۲. با استفاده از (۴۵.۶) و (۴۷.۶) مقدار $q_{m-1}(x;\alpha_m)$ را محاسبه کنید؛

۳. قرار دهید $p_{m-1} =: q_{m-1}$.

در ادامه یکی از روش‌های مشهور در این نوع را ارائه می‌کنیم که از روش نیوتن برای تقریب ریشه‌ها استفاده می‌کند.

۲.۶.۶ روش نیوتن-هورنر

آن گونه که از نامش پیداست، روش نیوتن-هورنر با استفاده از روش نیوتن برای محاسبه ریشه‌های α_m فرآیند تقلیل را اجرایی می‌کند. مزیت این روش در آن است که الگوریتم هورنر (۴۵.۶) را به راحتی در اجرای روش نیوتن به کار می‌گیرد. در واقع، اگر q_{n-1} چندجمله‌ای متناظر با p_n در (۴۷.۶) باشد، چون

$$p_n'(x) = q_{n-1}(x;z) + (x-z)q_{n-1}'(x;z)$$

بنابراین $p_n'(z) = q_{n-1}(z;z)$. با توجه به این اتحاد، روش نیوتن-هورنر برای تقریب یک ریشه (حقیقی یا مختلط) α_m، $m = 1, \ldots, n$ از p_n به صورت زیر است:

برای حدس اولیه داده شده $\alpha_{m,0}$ از ریشه، برای $k \geqslant 0$ تا همگرایی محاسبات زیر را انجام دهید

$$\alpha_{m,k+1} = \alpha_{m,k} - \frac{p_n(\alpha_{m,k})}{p_n'(\alpha_{m,k})} = \alpha_{m,k} - \frac{p_n(\alpha_{m,k})}{q_{n-1}(\alpha_{m,k};\alpha_{m,k})}. \qquad (۴۷.۶)$$

برای ریشه مختلط ($\alpha_m \in \mathbb{C}$) لازم است محاسبات در حساب مختلط انجام شود و حدس اولیه داده شده $\alpha_{m,0}$ نیز قسمت موهومی ناصفر داشته باشد. در غیر این صورت روش نیوتن-هورنر دنباله $\alpha_{m,k}$ از اعداد حقیقی را تولید خواهذ کرد که به هیچ ریشه مختلطی همگرا نخواهد شد.

روش نیوتن-هورنر در برنامه زیر اجرا شده است. ورودی‌ها بردار ضرایب چندجمله‌ای یعنی a، از a_0 تا a_n، حدس اولیه x0 و پارامترهای tol و Nmax برای در اختیار داشتن یک معیار توقف است. خروجی‌ها به ترتیب ریشه‌ها roots و تعداد تکرارهای مورد نیاز برای محاسبه آنها یعنی Niter است.

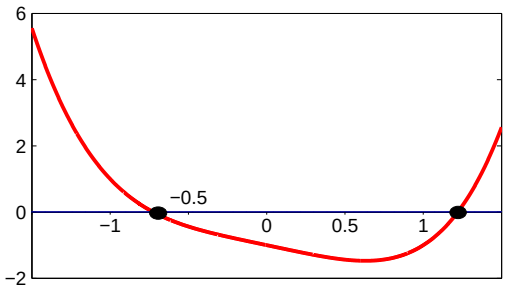
```
function [roots,Niter] = NewtonHorner(a,x0,tol,Nmax)
n = length(a); roots = zeros(n-1,1);
for k = 1:n-1
    iter = 0; x = x0; diff = tol+1;
    while (abs(diff) > tol) && (iter < Nmax)
        [pz,b] = Horner(a,x); [dpz,b] = Horner(b,x);
        diff = -pz/dpz; x = x + diff;
        iter = iter + 1;
    end
    [pz,a] = Horner(a,x); roots(k) = x; Niter(k) = iter;
end
```

مثال ۱۹.۶. چندجمله‌ای $p_4(x) = x^4 - x - 1$ را در نظر می‌گیریم. در مثال ۱۸.۶ ادعا شد که این چند جمله‌ای یک ریشه حقیقی مثبت، یک ریشه حقیقی منفی و دو ریشه مختلط دارد. در شکل ۱۵.۶ این موضوع به خوبی نمایش داده شده است.

مقدار دقیق ریشه‌ها برابر است با

$$r_1 = ۱/۲۲۰۷۴۴۰۸۴۶۶۰۵۷۵۹,$$
$$r_2 = -۰/۷۲۴۴۹۱۹۵۹۰۰۰۵۱۶,$$
$$r_{3,4} = -۰/۲۴۸۱۲۶۰۶۲۸۰۲۶۲۲ \pm ۱/۰۳۳۹۸۲۰۶۰۹۷۵۹۶۸i.$$

برای محاسبه ریشه‌ها با روش نیوتن-هورنر به‌شکل زیر عمل می‌کنیم. در صورتی که حدس اولیه $x_0 = ۱$ باشد، این روش ریشه حقیقی مثبت و ریشه حقیقی منفی را با دقت $۱۰^{-۱۵}$ پس از ۷ تکرار محاسبه می‌کند، در حالی که قادر به محاسبه



شکل ۱۵.۶: نمودار چندجمله‌ای $p_۴(x) = x^۴ - x - ۱$

ریشه‌های مختلط نیست. اگر حدس اولیه $x_۰ = ۱ + i$ باشد روش نیوتن-هورنر ریشه حقیقی مثبت را با ۱۳ تکرار، ریشه حقیقی منفی را با ۱۰ تکرار، همچنین ریشه‌های مختلط را با ۷ و ۲ تکرار و با دقت $۱۰^{-۱۵}$ محاسبه می‌کند. برای اجرا در حالت دوم دستور زیر را نوشته‌ایم

```
>> a = [-1 -1 0 0 1];
>> [roots,Niter] = NewtonHorner(a,1+1i,10^-15,100);
```

که نتایج زیر را در خروجی چاپ کرده است

```
roots =
 -0.724491959000516 - 0.000000000000000i
  1.220744084605759 + 0.000000000000000i
 -0.248126062802622 - 1.033982060975968i
 -0.248126062802622 + 1.033982060975968i
Niter =
    10    13     7     2
```

ملاحظه ۳.۶. فرض کنیم ریشه‌های p_n به صورت زیر مرتب شده باشند

$$۰ \leqslant |\alpha_۱| \leqslant |\alpha_۲| \leqslant \cdots \leqslant |\alpha_n|.$$

برای مینیمم‌سازی انتشار خطای گردکردن، بهتر آن است که هنگام فرآیند تقلیل ابتدا ریشه‌ی با کوچکترین مقدار قدرمطلق، $\alpha_۱$، را به دست آوریم و پس از آن $\alpha_۲$، $\alpha_۳$، ... و سرانجام α_n. در صورتی که تمام ریشه‌های یک چندجمله‌ای، حقیقی و مثبت باشند با حدس اولیه $x_۰ = ۰$ تمام ریشه‌ها به ترتیب صعودی به دست می‌آیند.

شایان ذکر است که مرتبه همگرایی روش نیوتن-هورنر برای ریشه‌های تکراری (دارای چندگانگی) بازهم کاهش می‌یابد و روش کارایی کمتری خواهد داشت.

۳.۶.۶ ماتریس همراه

چندجمله‌ایهای بسیاری هستند که ریشه‌های آنها نسبت به تغییرات کوچک در ضرایب حساسیت زیادی دارند. در واقع، آنچنان که که در فصل دوم اشاره شد، ریشه‌های چندجمله‌ایها ممکن است بدوضع باشند. در این بند می‌خواهیم بین مسئله‌ی یافتن ریشه‌های یک چندجمله‌ای و یافتن ویژه‌مقدارهای یک ماتریس ارتباط برقرار کنیم. چون الگوریتم‌های کارایی (مثلاً، الگوریتم QR یا الگوریتم‌های مبتنی بر زیرفضاهای کرایلف) و در نتیجه نرم‌افزارهای آماده‌ای برای یافتن مقادیر ویژه‌ی یک ماتریس وجود دارند.

به چندجمله‌ای مونیک (با ضریب پیشروی یک)

$$p_n(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

یک ماتریس $n \times n$ به نام ماتریس همراه[1] به شکل زیر وابسته می‌کنیم

$$A = \begin{bmatrix} -a_{n-1} & -a_{n-2} & \dots & -a_1 & -a_0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix}$$

با بسط دترمینان $\det(\lambda I - A)$ می‌توان نشان داد $p_n(\lambda)$ همان چندجمله‌ای مشخصه (سرشت نما) ماتریس A است. بنابراین ریشه‌های چندجمله‌ای p_n همان مقادیر ویژه‌ی ماتریس همراه A می‌باشند و برعکس.

مثال ۲۰.۶. چندجمله‌ای $p_4(x) = x^4 - x - 1$ را در نظر می‌گیریم. ماتریس همراه p_4 چنین است

$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

مقادیر ویژه‌ی ماتریس A، ریشه‌های چندجمله‌ای p_4 می‌باشند. به عنوان مثال می‌توان با دستور

```
eig(A)
```

[1] companion matrix

در متلب آن‌ها را بدست آورد. خروجی، ریشه‌های $p_۴$ است که در مثال ۱۹.۶ به‌دست آمد. لازم به ذکر است که در این دستور از روش‌های جبرخطی عددی برای محاسبه‌ی مقادیر ویژه استفاده می‌شود و از الگوریتم‌هایی که در این فصل مطالعه کردیم (مانند الگوریتم نیوتن-هورنر) استفاده نمی‌شود.

۷.۶ پرسش‌ها

پرسش ۱.۶. معادله‌ی $f(x) = x^۳ - ۲x - ۵ = ۰$ مفروض است. ابتدا نشان دهید این معادله یک ریشه در بازه $[۲, ۲/۲]$ دارد.، سپس ۳ تکرار از روش دوبخشی را به‌دست آورید.

پرسش ۲.۶. فرض کنید معادله‌ی $f(x) = ۰$ یک ریشه به نام α در بازه $[۰, ۱]$ داشته باشد. اگر در گام kام، $k \geqslant ۰$ روش دوبخشی نیمه‌ی اول بازه انتخاب شود قرار دهید $d_k = ۰$ و در غیر این صورت $d_k = ۱$. بین α و ارقام $d_۰$، $d_۱$، ... چه ارتباطی وجود دارد.

پرسش ۳.۶. منحنی‌های $y = e^x$ و $y = ۲x^۲$ را به طور دقیق رسم کنید و به کمک آن نشان دهید معادله $e^x - ۲x^۲ = ۰$ یک ریشه منفی $\alpha_۱$ و دو ریشه‌ی مثبت $\alpha_۲$ و $\alpha_۳$ دارد. برای کدام مقادیر از $x_۰$ روش نیوتن به $\alpha_۱$ همگرا می‌شود.

پرسش ۴.۶. فرض کنید α صفر ساده‌ی $f(x)$ باشد. در روش نقطه‌ی ثابت، تابع تکرار را سمت راست (۱۲.۶) یعنی

$$g(x) = x - \frac{f(x)}{f'(x)}$$

در نظر بگیرید. با استفاده از قضیه‌ی ۸.۶ نشان دهید مرتبه‌ی همگرایی روش نیوتن دست‌کم $p = ۲$ است.

پرسش ۵.۶. فرض کنید α ریشه‌ی مرتبه‌ی m از معادله‌ی $f(x) = ۰$ باشد. نشان دهید مرتبه‌ی همگرایی دنباله‌ی (۱۷.۶) برابر ۲ است.

پرسش ۶.۶. می‌خواهیم با روش نیوتن صفر $\alpha = \frac{\pi}{۲}$ از تابع نوسانی $f(x) = \cos(x) + \sin^۲(۵۰x)$ را بیابیم. یک برآورد تقریبی از بازه‌ی همگرایی I_δ برای تضمین همگرایی با شروع $x_۰ \in I_\delta$ بدست آورید.

پرسش ۷.۶. تابع $f(x) = (x - ۱)\ln x$ دارای صفر $\alpha = ۱$ با چندگانگی $m = ۲$ است. با استفاده از روش نیوتن (۱۲.۶) و روش نیوتن اصلاح‌شده (۱۷.۶) تقریبی از α را محاسبه کنید. خطای حاصل از این دو روش را در یک شکل گزارش نمایید.

پرسش ۸.۶. فرض کنید $p_۳$ یک چندجمله‌ای درجه سوم با سه ریشه حقیقی α، β و γ باشد. اگر روش نیوتن با حدس اولیه $x_۰ = (\alpha + \beta)/۲$ به کار گرفته شود، نشان دهد پس از یک تکرار γ را می‌یابد.

پرسش ۹.۶. تعیین ریشه سوم عدد حقیقی a، با پیدا کردن جواب معادله $x^۳ - a = ۰$ هم‌ارز است. (الف) دنباله $\{x_k\}$ حاصل از روش وتری (۲۱.۶) را برای حل این معادله بنویسید. (ب) فرض کنید $a = ۳$ و مقادیر اولیه $x_۰ = ۱$ و $x_۱ = ۲$ باشد. مقادیر $x_۲$ و $x_۳$ را به‌دست آورید.

پرسش ۱۰.۶. روش استیفنسن[1] یک روش تک‌نقطه‌ای شبه‌نیوتنی از نوع (۱۹.۶) است که در آن d_k به‌صورت زیر تعریف می‌شود

$$d_k = \frac{f(x_k + f(x_k)) - f(x_k)}{f(x_k)}$$

نشان دهید همگرایی این روش به ریشه‌های ساده دست‌کم از مرتبه دو است.

پرسش ۱۱.۶. تابع $f(x) = \sin x + x^۲ \cos x - x^۲ - x$ را در نظر بگیرید. ابتدا چندگانگی $\alpha = ۰$ را به عنوان ریشه‌ی معادله $f(x) = ۰$ تعیین کنید. سپس تعداد تکرارهای مورد نیاز روش نیوتن با حدس اولیه $x_۰ = ۱$ را برای یافتن تقریبی از این ریشه با شش رقم درست به دست آورید.

پرسش ۱۲.۶. نقاط ثابت تابع‌های زیر را به‌دست آورید.

الف) $g(x) = x^۲ - ۴x + ۲$ ب) $g(x) = \dfrac{۸ + ۲x}{۲ + x^۲}$

پرسش ۱۳.۶. کدام‌یک از تابع‌های زیر دارای نقطه ثابت $\alpha = \sqrt{۵}$ می‌باشد

الف) $g(x) = \dfrac{۵ + ۷x}{x + ۷}$ ب) $g(x) = \dfrac{۱۰}{۳x} + \dfrac{x}{۳}$ ج) $g(x) = x^۲ - ۵$

پرسش ۱۴.۶. فرض کنید تابع $g(x)$ به‌طور پیوسته مشتق‌پذیر و دارای دقیقاً سه نقطه‌ثابت $\alpha_۱$، $\alpha_۲$ و $\alpha_۳$ باشد طوری که $\alpha_۱ < \alpha_۲ < \alpha_۳$، $|g'(\alpha_۱)| = ۰/۵$ و $|g'(\alpha_۳)| = ۰/۵$. در مورد مقادیر $|g'(\alpha_۲)|$ چه می‌توان گفت؟

پرسش ۱۵.۶. تابع $g(x) = ax^۲ + bx + c$ را در نظر بگیرید. الف) مجموعه‌ای از ثابت‌های a، b و c را بیابید بطوریکه $x = ۰$ نقطه ثابت g باشد و روش تکرار نقطه ثابت موضعاً همگرا به صفر باشد. ب) مجموعه‌ای از ثابت‌های a، b و c را بیابید بطوریکه $x = ۰$ نقطه ثابت g باشد با این حال روش تکرار نقطه ثابت موضعاً همگرا به صفر نباشد.

پرسش ۱۶.۶. هر یک از معادله‌های زیر را به سه طریق متفاوت به عنوان یک مسئله نقطه ثابت $x = g(x)$ بازنویسی کنید.

الف) $۳x^{-۲} + ۹x^۳ = x^۲$ ب) $x^۳ - x + e^x = ۰$

پرسش ۱۷.۶. تابع $f(x) = x^۴ - ۷x^۳ + ۱۸x^۲ - ۲۰x + ۸$ را در نظر بگیرید. آیا روش نیوتن به ریشه $\alpha = ۲$ همگرای مربعی است؟ اگر e_k خطای تکرار kام باشد، مقدار $\lim_{k\to\infty} e_{k+۱}/e_k$ را بیابید.

[1] Johan Frederik Steffensen (1873 - 1961)

پرسش ۱۸.۶. فرض کنید $g(x) = x - f(x)/f'(x)$ تکرار روش نیوتن برای تابع f و c نقطه ثابت تابع $h(x) = g(g(x))$، اما نه نقطه ثابت g، باشد. نشان دهید اگر c نقطه عطف f باشد در آن صورت تکرار نقطه ثابت h به c موضعاً همگرا است. اگر روش نیوتن با حدس اولیه نزدیک c برای تابع f به کار گرفته شود، نتیجه چه خواهد شد؟

پرسش ۱۹.۶. تحت فرض‌های قضیه نگاشت انقباضی و مشابه کران پیشرو در (۳۶.۶)، یک کران دیگر به صورت زیر به‌دست آورید

$$|x_k - \alpha| \leqslant \frac{L}{۱ - L}|x_k - x_{k-۱}|, \tag{۴۸.۶}$$

این نابرابری خطای مطلق را بر حسب آخرین اطلاعات به‌دست آمده با روش تکراری مقیاس می‌کند.

پرسش ۲۰.۶. یک کاربرد عملی روش نیوتن محاسبه معکوس عدد a در ماشین‌های محاسباتی (اولیه) بدون عمل تقسیم بود. الف) نشان دهید روش نیوتن برای حل معادله

$$f(x) = \frac{۱}{x} - a = ۰$$

بدون عمل تقسیم قابل به‌کارگیری است. ب) دستوری برای جمله خطای $e_k = x_k - a^{-۱}$ بنویسید و نشان دهید همگرایی مربعی است. پ) شرایطی روی حدس اولیه ارائه کنید طوری که دنباله $\{x_k\}$ به $a^{-۱}$ همگرا شود. اگر $۰ < a < ۱$، یک مقدار عددی از $x_۰$ ارائه کنید که با آن همگرایی تضمین شود.

پرسش ۲۱.۶. برای روش وتری (۲۱.۶) دستور خطای زیر را به‌دست آورید

$$x_{k+۱} - \alpha = (x_k - \alpha)(x_{k-۱} - \alpha)\frac{f[x_{k-۱}, x_k, \alpha]}{f[x_{k-۱}, x_k]}$$

پرسش ۲۲.۶. با استفاده از پرسش ۲۱.۶، لم ۴.۶ را ثابت کنید.

پرسش ۲۳.۶. فرض کنیم α ریشه معادله $f(x) = ۰$، $x_۰ < x_۱$ تقریب‌هایی از آن، $y_۰ = f(x_۰)$ و $y_۱ = f(x_۱)$ دارای علامت‌های مختلف و $y = f(x)$ دارای تابع وارون یکتایی به صورت $x = \varphi(y)$ باشد که بر هر بازه $[c, d]$ شامل $y_۰$ و $y_۱$ به طور پیوسته دو بار مشتق‌پذیر است.

۱. اگر $p_۱(y; \varphi)$ چندجمله‌ای درونیاب خطی φ مبتنی بر $y_۰$ و $y_۱$ باشد و تقریب جدید ریشه را به صورت $x_۲ = p_۱(۰; \varphi)$ در نظر بگیریم، خطای $|x_۲ - \alpha|$ را برآورد کنید.

۲. فرض کنید $f(x) = x^۳ + x - ۱$، $x_۰ = ۰$ و $x_۱ = ۱$ باشد. مقدار $x_۲$ و برآوردی از $|x_۲ - \alpha|$ را بیابید.

پرسش ۲۴.۶. فرض کنیم α ریشه معادله $f(x) = ۰$، $x_۰ < x_۲ < x_۱$ تقریب‌هایی از آن، $y_۰ = f(x_۰)$ و $y_۱ = f(x_۱)$ دارای علامت‌های مختلف، $y_۲ = f(x_۲)$ و $y = f(x)$ دارای تابع وارون یکتایی به صورت $x = \varphi(y)$ باشد که بر هر بازه $[c, d]$ شامل $y_۰$، $y_۱$ و $y_۲$ به طور پیوسته سه بار مشتق‌پذیر است.

۱. اگر $p_۲(y;\varphi)$ چندجمله‌ای درونیاب درجه دوم φ مبتنی بر $y_۰$، $y_۱$ و $y_۲$ باشد و تقریب جدید ریشه را به صورت $x_۳ = p_۲(۰;\varphi)$ در نظر بگیریم، خطای $|x_۳ - \alpha|$ را برآورد کنید.

۲. فرض کنید $f(x) = x^۳ + x - ۱$، $x_۰ = ۰$، $x_۱ = ۱$ و $x_۲ = ۰/۵$ باشد. مقدار $x_۳$ و برآوردی از $|x_۳ - \alpha|$ را بیابید.

پرسش ۲۵.۶. روش تکراری $x_{k+۱} = x_k - f(x_k)/g(x_k)$ را در نظر بگیرید. فرض کنید این دنباله به α همگرا شود که صفر ساده‌ای از f است اما صفر g نیست. ارتباط بین f و g را طوری تعیین که مرتبه همگرایی دست‌کم ۳ شود.

پرسش ۲۶.۶. هذلولی $h(x) = b + a/(x - c)$ را در نظر بگیرید.

۱. ثابت‌های a، b و c را به گونه‌ای بیابید که $h(x)$ در نقطه‌ی معلوم x_k بر منحنی $y = f(x)$ مماس و با آن در این نقطه انحنای یکسان داشته باشد. نقطه‌ی تقاطع این هذلولی با محور xها را $x_{k+۱}$ بنامید و نشان دهید

$$x_{k+۱} = x_k - \frac{f(x_k)}{f'(x_k) - f(x_k)f''(x_k)/(۲f'(x_k))} \tag{۴۹.۶}$$

۲. نشان دهید $x_{k+۱}$ در (۴۹.۶) را می‌توان با اجرای روش نیوتن روی تابع $u(x) = f(x)/\sqrt{f'(x)}$ نیز به‌دست آورد.

۳. فرض کنید α ریشه ساده معادله‌ی $f(x) = ۰$ باشد و دنباله $\{x_k\}$ به α همگرا شود. نشان دهید مرتبه‌ی همگرایی دقیقاً سه است، مگر اینکه مشتق شوارتسی

$$(Sf)(x) := \frac{f'''(x)}{f'(x)} - \frac{۳}{۲}\Big(\frac{f''(x)}{f'(x)}\Big)^۲$$

در $x = \alpha$ صفر شود، که در آن صورت مرتبه همگرایی بزرگتر از سه است.

روش معرفی شده در این تمرین **روش هالی**[۱] نامیده می‌شود.

پرسش ۲۷.۶. فرض کنید روش نیوتن در (۱۲.۶) با علامت نادرست به شکل زیر نوشته شود

$$x_{k+۱} = x_k + \frac{f(x_k)}{f'(x_k)}$$

این دنباله به چه مقداری همگرا می‌شود؟

پرسش ۲۸.۶. همگرایی دنباله‌ی تکرار نقطه‌ی ثابت

$$x_{k+۱} = \frac{x_k(x_k^۲ + ۳a)}{۳x_k^۲ + a}, \quad k = ۰, ۱, \ldots$$

[۱] Edmond Halley (1656 – 1742)

را برای محاسبه‌ی ریشه‌ی دوم عدد مثبت a تحلیل کنید. با فرض این‌که $x_۰$ به اندازه‌ی کافی به $\sqrt{a}$ نزدیک باشد، مقدار حد زیر را بیابید

$$\lim_{k\to\infty} \frac{x_{k+۱} - \sqrt{a}}{(x_k - \sqrt{a})^۳}.$$

پرسش ۲۹.۶. معادله‌ی $f(x) = x^۳ + ۴x^۲ - ۱۰ = ۰$ مفروض است. ابتدا با قضیه مقدار میانی نشان دهید این معادله یک ریشه‌ی در بازه $[۱, ۲]$ دارد. برای این مسئله سه تابع تکرار g به‌شکل زیر در نظر بگیرید

$$g_۱(x) = \sqrt{\frac{۱۰}{x} - ۴x}, \quad g_۲(x) = \frac{۱}{۲}\sqrt{۱۰ - x^۳}, \quad g_۳(x) = \sqrt{\frac{۱۰}{۴ + x}}$$

همگرایی طرح‌های تکراری $x_{k+۱} = g_i(x_k)$، $i = ۱, ۲, ۳$ به نقطه ثابت $\alpha \doteq ۱/۳۶۵۲۳۰۰۱۳$ را با توجه به قضیه‌ی ۷.۶ بررسی نمایید.

پرسش ۳۰.۶. برای تقریب $\sqrt{۲}$ طرح تکراری زیر را در نظر بگیرید

$$x_۰ = ۱, \quad x_{k+۱} = \frac{۳}{۸}x_k + \frac{۳}{۲x_k} - \frac{۱}{۲x_k^۳}, \quad k = ۰, ۱, \ldots$$

با فرض $e_x(k) = |x_k - \sqrt{۲}|/\sqrt{۲}$ نشان دهید عدد $\rho \in (۰, ۱)$ وجود دارد طوری که

$$\ln(e_x(k)) \simeq C + \ln(\rho)k^۳$$

و نتیجه بگیرید مرتبه‌ی همگرایی این دنباله ۳ است.

پرسش ۳۱.۶. تکرار نقطه ثابت $x_{k+۱} = g(x_k)$ را با تابع تکرار $g(x) = ax + bx^۲ + cx^۳$ در نظر بگیرید و فرض کنید α عدد مثبت معلومی است. مقادیر a، b و c را طوری تعیین کنید که دنباله به $۱/\alpha$ موضعا همگرا از مرتبه ۳ شود.

پرسش ۳۲.۶. مقادیر p، q و r را طوری بیابید که مرتبه همگرایی روش تکراری

$$x_{k+۱} = px_k + qa/x_k^۲ + ra^۲/x_k^۵$$

برای محاسبه $\sqrt[۳]{a}$ حداکثر باشد. با این مقادیر ارتباط میان خطای $x_{k+۱}$ با خطای x_k را تعیین کنید.

پرسش ۳۳.۶. فرض کنید ماتریس مربعی $A \in \mathbb{R}^{n\times n}$ با درایه‌های a_{kj} داده شده است. نشان دهید اگر

$$r_k = \sum_{j=۱, j\neq k}^{n} |a_{kj}|, \quad k = ۱, ۲, \ldots, n,$$

آنگاه هر مقدار ویژه‌ی A در یکی از گوی‌های $B(a_{kk}, r_k)$ در صفحه‌ی مختلط واقع است. یادآوری می‌کنیم که $B(a, r)$ گویی به شعاع r و به مرکز a است. همچنین نشان دهید اگر قرار دهیم

$$r_j = \sum_{k=۱, k\neq j}^{n} |a_{kj}|, \quad j = ۱, ۲, \ldots, n,$$

باز هم هر مقدار ویژه‌ی A در یکی از گوی‌های $B(a_{jj}, r_j)$ واقع است. (قسمت دوم با توجه به این واقعیت درست است که مقادیر ویژه‌ی A و A^T یکسانند). این تمرین صورتی از قضیه‌ای معروف به نام "قضیه‌ی گرشگورین" می‌باشد.

پرسش ۳۴.۶. با استفاده از تعریف ماتریس همراه و با کمک قضیه گرشگورین (قسمت دوم)، قضیه کوشی ۱۰.۶ را اثبات کنید.

۸.۶ مسئله‌های ماشینی

پرسش ۳۵.۶. ریشه‌های حقیقی معادله‌ی $f(x) = x^۳ - ۲x - ۵ = ۰$ را به روش دوبخشی به‌دست آورید.

پرسش ۳۶.۶. ریشه‌های حقیقی معادله‌ی $f(x) = \tan(x) + \tanh(x) = ۰$ را به روش دوبخشی به‌دست آورید.

پرسش ۳۷.۶. کوچکترین ریشه مثبت معادله زیر را به روش دوبخشی و دقت $\varepsilon = ۱۰^{-۸}$ به‌دست آورید.

$$\cos(۴x\sqrt{۱ - x^۲}) = -۱ + ۸x^۲ - ۸x^۴$$

پرسش ۳۸.۶. ماتریس زیر را در نظر بگیرید. با روش دوبخشی دو مقدار برای x، با شش رقم درست اعشاری، طوری بیابید که دترمینان این ماتریس برابر ۱۰۰ شود.

$$A = \begin{bmatrix} ۱ & ۲ & ۳ & x \\ ۴ & ۵ & x & ۶ \\ ۷ & x & ۸ & ۹ \\ x & ۱۰ & ۱۱ & ۱۲ \end{bmatrix}$$

با مقادیر به‌دست آمده برای x دترمینان ماتریس را با دستور `det` در متلب به دست آورید و با ۱۰۰ مقایسه کنید.

پرسش ۳۹.۶. تابع $f(x;\gamma) = \cosh x + \cos x - \gamma$ را برای $\gamma = ۱, ۲, ۳$ در نظر می‌گیریم. الف) در هر سه حالت بازه‌ای شامل صفر (یا صفرهای) f به دست آورید. ب) به روش دوبخشی و با دقت $۱۰^{-۱۰}$ صفر (یا صفرهای) تابع را بیابید. پ) چرا در حالت $\gamma = ۲$، روش نیوتن دقیق نیست؟

پرسش ۴۰.۶. معادله‌ی کپلر (۴.۶) در مثال ۳.۶ با $e = ۰/۸$ و $T = ۹۰$ دقیقه مفروض است. این معادله را با روش نیوتن (به کمک تابع `Newton` با پارامترهای مناسب) حل کنید.

پرسش ۴۱.۶. مسئله‌ی کشاورز-بُز-چمن‌زار در مثال ۴.۶ را با روش نیوتن حل کنید. حدس اولیه را شعاع چمن‌زار و برابر ۱ در نظر بگیرید.

پرسش ۴۲.۶. معادله زیر برای $x > ۰$ داده شده است

$$\ln(۱ + x^۲) = ۲x \arctan(\frac{۱}{x}) = \ln(۴)$$

به روش نیوتن با حدس اولیه $x_۰ = ۰/۵$ تقریبی برای ریشه این معادله به‌دست آورید.

پرسش ۴۳.۶. چندجمله‌ای لاگر درجه ۴ زیر را در نظر بگیرید

$$p_۴(x) = x^۴ - ۱۶x^۳ + ۷۲x^۲ - ۹۶x + ۲۴$$

این چندجمله‌ای چهار ریشه‌ی حقیقی متمایز و مثبت دارد. با حدس اولیه $x_۰ = ۰$ و روش نیوتن-هورنر ریشه‌های این چندجمله‌ای را بیابید. در طی این فرآیند چندجمله‌ایهای تقلیل یافته $q_۳(x; z)$، $q_۲(x; z)$ و $q_۱(x; z)$ را به طور صریح به‌دست آورید.

پرسش ۴۴.۶. معادله حالت واندروالس در (۲.۶) را در نظر بگیرید. با واحدهای مناسب فرض کنید $R = ۰/۰۸۲۰۵$، و برای دی‌اکسید کربن $a = ۳/۵۹۲$ و $b = ۰/۰۴۲۶۷$ باشد. به کمک الف) روش دوبخشی، ب) روش نیوتن و ج) روش ترکیبی دوبخشی-نیوتن، حجم ویژه v را در دمای ۳۰۰ کلوین و برای فشارهای ۱، ۱۰ و ۱۰۰ اتمسفر برآورد کنید. نتایج به دست آمده را با حالت گاز ایده‌ال $pv = RT$ مقایسه کنید. برای حدس اولیه $v_۰$ در روش‌های تکراری می‌توانید از این معادله استفاده کنید.

پرسش ۴۵.۶. دو دالان با عرض‌های $l_۱$ و $l_۲$ مطابق شکل ۱۶.۶ در نظر بگیرید. می‌خواهیم میله‌ای به طول L را در حالت افقی از دالان به عرض $l_۲$ به دالان به عرض $l_۱$ ببریم. بزرگ‌ترین طول $L_۰$ که اگر $L < L_۰$، این کار میسر باشد به صورت زیر است

$$L_۰ = l_۱ \csc(\alpha) + l_۲ \csc(\pi - \gamma - \alpha).$$

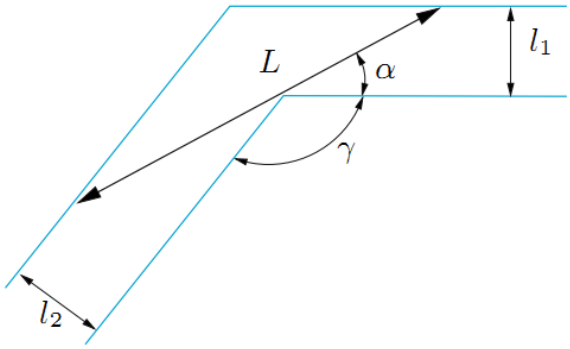
در اینجا α جواب معادله‌ی غیرخطی زیر است

$$l_۱ \cot(\alpha) \csc(\alpha) - l_۲ \cot(\pi - \gamma - \alpha) \csc(\pi - \gamma - \alpha) = ۰.$$

فرض کنید $l_۱ = ۸$، $l_۲ = ۱۰$ و $\gamma = ۳\pi/۵$ باشد. با روش تکراری نیوتن α را به‌دست آورید.

پرسش ۴۶.۶. فرض کنید برای یافتن صفرهای تابع $f(x) = e^x - x - ۲$ روش تک‌نقطه‌ای استیفنسن در تمرین ۱۰.۶ را با حدس اولیه $x_۰ = ۱$ به‌کار گیریم. همگرایی مربعی این روش را راستی‌آزمایی کنید.

پرسش ۴۷.۶. چندجمله‌ای $f(x) = x^۳ + x^۲ - ۱۰x - ۱۰$ را در نظر می‌گیریم. با حدس‌های اولیه $x_۰ = ۱$، $x_۱ = ۲$ و $x_۲ = ۳$ جواب‌های معادله‌ی درجه دوم (۲۷.۶) در روش مولر را به‌دست آورید. فرض کنید $x_۳ = ۳/۱۷۹۷۱۰۸۶$ تقریب جدید باشد. یک بار دیگر معادله‌ی درجه دوم (۲۷.۶) را با $x_۱$، $x_۲$ و $x_۳$ تشکیل دهید. با این که ضرایب این معادله درجه دوم حقیقی هستند آیا ریشه‌های آن نیز حقیقی می‌باشد.



شکل ۱۶.۶: مسئله عبور یک میله به طول L در پیچ یک دالان

پرسش ۴۸.۶. معادله‌ی $f(x) = x^3 + 4x^2 - 10 = 0$ مفروض است. با دستور plotf نشان دهید این معادله یک ریشه‌ی در بازه $[1, 2]$ دارد. برای این مسئله سه تابع تکرار g به‌شکل زیر در نظر بگیرید

$$g_1(x) = \sqrt{\frac{10}{x} - 4x}, \quad g_2(x) = \frac{1}{2}\sqrt{10 - x^3}, \quad g_3(x) = \sqrt{\frac{10}{4 + x}}$$

با فرض $x_0 = 1$ تکرارهای نقطه ثابت $x_{k+1} = g_i(x_k)$، $i = 1, 2, 3$ را در جدولی گزارش کنید. ریشه واقعی در حدود $\alpha \doteq 1/365230013$ است. نتایج را با پیش‌بینی تمرین ۲۹.۶ تطبیق دهید.

پرسش ۴۹.۶. معادله (۳.۶) در مثال ۲.۶ را به کمک روش نیوتن-هورنر حل کنید.

پرسش ۵۰.۶. معادله‌ی $x - e^{-x} = 0$ با ریشه α مفروض است. طرح تکراری $x_{k+1} = g(x_k)$ را به شکل‌های زیر در نظر بگیرید.

۱. فرض کنید $g(x) = e^{-x}$ و حدس اولیه $x_0 = 1$ باشد. جملات دنباله $\{x_k\}$ را تا کوچکترین k که $|x_{k+1} - x_k| \leqslant$ eps به‌دست آورید.

۲. فرض کنید

$$g(x; \omega) = \frac{\omega e^{-x} + x}{1 + \omega}, \quad \omega \neq 0 \text{ و } \omega \neq -1$$

تحت چه شرطی روی ω همگرایی طرح تکراری با این تابع تکرار سریع‌تر از $g(x) = e^{-x}$ است.

۳. انتخاب بهینه از ω چقدر است؟ نتایج را راستی‌آزمایی کنید.

کتاب‌نامه

[1] F. L. Bauer, H. Rutishauser, and E. Stiefel. New aspects in numerical quadrature. In *Proc. of Symposia in Appl. Math.*, volume 15, pages 199–218. Amer. Math. Soc., 1963.

[2] G. Dahlquist and Å. Björck. *Numerical Methods in Scientific Computing, Volume I*. SIAM, Philadelphia, 2008.

[3] W. Gander and W. Gautschi. Adaptive quadrature–revisited. *BIT*, 40:84–101, 2000.

[4] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. second edition. SIAM, Philadelphia, PA, 2002.

[5] J. N. Lyness. Notes on the adaptive simpson quadrature routine. *J. Assoc. Comput. Mach.*, 16:483–495, 1969.

[6] د. میرزائی. *آنالیز عددی پیشرفته*. انتشارات دانشگاه اصفهان, ۱۳۹۲.

[7] C. Runge. Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten. *Z. f. Math. u. Phys.*, 46:224–243, 1901.

[8] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. 3rd edition. Springer-Verlag, New York, 2002.