آنالیز عددی مقدماتی

به شیوهٔ الگوریتمی

سموئل د. کونت، کارل دوبور

ترجمهٔ سراج‌الدین کاتبی

آنالیز عددی مقدماتی

به شیوهٔ الگوریتمی

سموئل د. کونت، کارل دوبور

ترجمهٔ سراج‌الدین کاتبی

مرکز نشر دانشگاهی، تهران

بسم الله الرحمن الرحیم

فهرست

عنوان	صفحه
پیشگفتار	۱
مقدمه	۳
فصل اول دستگاه اعداد و خطاها	۵
۱.۱ نمایش اعداد صحیح	۵
۲.۱ نمایش کسرها	۱۰
۳.۱ حساب با ممیزشناور	۱۳
۴.۱ از دست دادن ارقام با معنی و پخش‌خطا، شرط و ناپایداری	۲۰
۵.۱ روشهای کامپیوتری برای برآورد خطا	۲۷
۶.۱ توضیحاتی درمورد همگرایی دنباله‌ها	۲۹
۷.۱ برخی مقدمات ریاضی	۳۷
فصل دوم درونیابی به‌وسیلهٔ بسجمله‌ایها	۴۴
۱.۲ صورتهای بسجمله‌ای	۴۴
۲.۲ وجود و یکتایی بسجمله‌ای درونیاب	۵۳
۳.۲ جدول تفاضل‌منقسم	۵۸
۴.۲ درونیابی دریك عدهٔ رو به‌افزایش از نقاط درونیابی	۶۴
۵.۲ خطای بسجمله‌ای درونیاب	۶۹
۶.۲ درونیابی یك تابع جدولی براساس نقاط با فواصل‌مساوی	۷۵
۷.۲ تفاضل منقسم به‌عنوان تابعی ازشناسه‌ها و درونیابی بوسانی‌اش	۸۴

عنوان	صفحه
فصل سوم حل معادلات غیرخطی	۹۶
۱.۳ بررسی اجمالی روشهای بازستی	۹۸
۲.۳ برنامه‌های فورترن برای بعضی از روشهای بازستی	۱۰۷
۳.۳ بازست نقطهٔ ثابت	۱۱۴
۴.۳ شتاب همگرایی برای بازست نقطهٔ ثابت	۱۲۳
۵.۳ همگرایی نیوتن و روشهای خط قاطع	۱۲۸
۶.۳ معادلات بسجمله‌ای: ریشه‌های حقیقی	۱۴۱
۷.۳ ریشه‌های همتافت و روش مولر	۱۵۴
فصل چهارم ماتریسها و دستگاههای معادلات خطی	۱۶۴
۱.۴ ویژگیهای ماتریسها	۱۶۴
۲.۴ حل دستگاههای خطی از راه حذف	۱۸۹
۳.۴ تدبیر لولا گزینی	۲۰۲
۴.۴ تجزیه به‌عوامل مثلثی	۲۰۶
۵.۴ خطا و باقیماندهٔ یك جواب تقریبی؛ نرمها	۲۱۷
۶.۴ تحلیل خطای پسرو و اصلاح بازستی	۲۲۷
۷.۴ دترمینانها	۲۳۸
۸.۴ مسئلهٔ ویژه‌مقدارها	۲۴۳
فصل پنجم دستگاه معادلات و بهینه‌سازی نامقید	۲۶۹
۱.۵ بهینه‌سازی و تندترین کاهش	۲۷۰
۲.۵ روش نیوتن	۲۷۹
۳.۵ بازست نقطهٔ ثابت و روشهای واهلش	۲۸۹
فصل ششم تقریب	۳۰۵
۱.۶ تقریب یکنواخت به‌وسیلهٔ بسجمله‌ایها	۳۰۵
۲.۶ برازاندن داده‌ها	۳۱۹
۳.۶ بسجمله‌ایهای متعامد	۳۲۶
۴.۶ تقریب با روش کوچکترین مربعات به‌وسیلهٔ بسجمله‌ایها	۳۳۷
۵.۶ تقریب به‌وسیلهٔ بسجمله‌ایهای مثلثاتی	۳۴۸
۶.۶ تبدیلهای سریع فوریه	۳۶۰

عنوان	صفحه
۷.۶ تقریب به وسیلهٔ بسجمله‌ای-تکه‌ای	۳۶۹
فصل‌هفتم مشتقگیری و انتگرالگیری	۳۸۰
۱.۷ مشتقگیری عددی	۳۸۲
۲.۷ انتگرالگیری عددی: برخی قواعد اساسی	۳۹۲
۳.۷ انتگرالگیری عددی: قواعد گاوسی	۴۰۲
۴.۷ انتگرالگیری عددی: قاعده‌های مرکب	۴۱۳
۵.۷ انتگرال تطبیقی	۴۲۵
۶.۷ برونیابی به‌سمت حد	۴۳۳
۷.۷ انتگرالگیری رامبرگ	۴۴۱
فصل‌هشتم حل معادلات دیفرانسیل	۴۴۹
۱.۸ مقدمات ریاضی	۴۴۹
۲.۸ معادلات تفاضلی‌ساده	۴۵۳
۳.۸ انتگرالگیری عددی به‌وسیلهٔ سری تیلر	۴۶۰
۴.۸ براوردهای خطا و همگرایی روش اویلر	۴۶۷
۵.۸ روش رونگه-کوتا	۴۷۱
۶.۸ کنترل اندازهٔ h در روشهای رونگه-کوتا	۴۷۶
۷.۸ فرمولهای چندمرحله‌ای	۴۸۴
۸.۸ روشهای پیشگو-تصحیحی	۴۹۲
۹.۸ روش ادمز-مولتن	۴۹۶
۱۰.۸ پایداری روشهای عددی	۵۰۵
۱۱.۸ کنترل و پخش خطای گردکردن	۵۱۲
۱۲.۸ دستگاههای معادلات دیفرانسیل	۵۱۶
۱۳.۸ معادلات دیفرانسیل سخت	۵۲۱
فصل‌نهم مسائل با مقدار مرزی درمعادلات دیفرانسیل معمولی	۵۲۶
۱.۹ روشهای تفاضل متناهی	۵۲۷
۲.۹ روشهای نشانه‌گیری	۵۳۳
۳.۹ روشهای پهلوی هم‌گذاری	۵۳۹
ضمیمه: کتابهای زیربرنامه‌ای	۵۴۵
منابع	۵۴۷
فهرست اسامی خاص	۵۴۹
واژه‌نامه	۵۵۰

پیشگفتار

این سومین چاپ کتاب آنالیز عددی مقدماتی است که مخصوصاً برای نیاز دانشجویان سالهای دورهٔ کارشناسی در رشته‌های مهندسی، ریاضی و علوم و به‌ویژه رشتهٔ علوم کامپیوتر طرح‌ریزی شده است. به‌طور کلی دانشجویانی که دروس حساب دیفرانسیل و انتگرال را به‌طور مرتب با پایهٔ محکمی گذرانیده باشند، در دنبال نمودن مطالب این کتاب مشکلی نخواهند داشت. هنگام به‌کاربردن مفاهیم پیشرفتهٔ ریاضی مانند نرمها[1] و تعامد[2] دقت شده است تا عرضهٔ مطالب در سطحی مناسب برای دانشجویان دورهٔ کارشناسی، بافرض اینکه دانشجو هیچ‌گونه آشنایی با آنها ندارد، صورت گیرد. در فصل مربوط به‌معادلات، آشنایی‌مختصری با ماتریسها و در فصلهای ۸ و ۹ آشنایی بامعادلات دیفرانسیل‌مسلم انگاشته شده است. در این چاپ بخشهایی وجود دارندکه نسبت به‌چاپهای قبلی به‌کمال ریاضی بیشتری نیاز دارند. ولی همهٔ این بخشها با «ستاره نماد» مشخص شده‌اند و استاد می‌تواند بی‌آنکه پیوستگی مطالب را از دست دهد، آنها را حذف کند.

این چاپ جدید متضمن مطالبی تازه و تغییراتی قابل‌ملاحظه نسبت به‌چاپ قبلی‌است. ترتیب فصول به‌گونه‌ای‌که فکر می‌کنیم طبیعیتر است درآمده است. درونیابی بسجمله‌ای اکنون حتی مقدم برفصل مربوط به‌حل دستگاههای غیرخطی (فصل ۳) عرضه شده است و بعداً برای برخی مطالب مذکور درتمام فصول به‌کار رفته‌است.

روش حذف گاوس (فصل ۴) ساده شده است. بعلاوه در فصل ۴، اکنون استفادهٔ دامنه‌داری از تحلیل خطای پسرو ویلکینسن[3] به عمل آمده، و شرح مختصری از روشهای بسیار معروف برای مسئلهٔ ویژه-مقدارها و ویژه-بردارها گنجانده شده‌است. فصل ۵، فصل تازه‌ای‌است دربارهٔ دستگاه معادلات و بهینه‌سازی نامقید ومتضمن مقدمه‌ای‌است برروشهای تندترین کاهش، روش نیوتن برای دستگاه معادلات غیرخطی و روشهای واهلش[4] حل دستگاههای بزرگ خطی باروش بارستی. فصل مربوط به‌تقریب (فصل ۶) بسط داده شده

---

1. norms 2. orthogonality 3. Wilkinson
4. relaxation

است. در این فصل بهترین تقریب و تقریب خوب به توسط بسجمله‌ایهـا، و نیز تقریب به‌کمك توابع مثلثاتی از جمله تبدیلهای سریع فوریه[1]، برازاندن داده‌ها با روش کمترین مربعات، بسجمله‌ایهای متعامد و برازاندن منحنی با روش قلمی[2] مورد بررسی قرار گرفته است. مشتقگیری وانتگرالگیری درفصل ۷، که شامل بخش جدیدی درباب انتگرالگیری تطبیقی‌است، آمده‌است. فصل۸ در زمینهٔ معادلات دیفرانسیل معمولی است‌که مقدار زیادی مطالب تازه و چندین بخش جدید به‌آن افزوده شده است. بخش جدیدی در زمینهٔ کنترل اندازه ـ مرحله در روشهای‌رونگه ـ کوتا، وبخش‌جدیدی‌دربارهٔ معادلات دیفرانسیل‌سخت، و نیز بخشی درزمینهٔ ناپایداری عددی که به‌مقدار زیادی‌تجدید نظرشده، دراین فصل آمده است. فصل ۹ حاوی مقدمه‌ای‌است کوتاه درباب پهلوی هم‌گذاری[3] به‌عنوان روشی برای حل مسائل مقدار مرزی.

در این چاپ، همانند چاپ قبلی، فرض براین است که دانشجویان به‌یك کامپیوتر دسترسی، و با برنامه‌سازی در برخی از زبانهای شیوه ـ پرداخته[4] آشنایی دارند. تعداد زیادی الگوریتم در متن گنجانیده شده و برای بسیاری از آنها برنامه‌هـای **فورترن** ارائه شده است. دراین چاپ برنامه‌های‌کامل‌تاحدی کمتر آورده شده است. همهٔ برنامه‌ها به‌زبان فورترن ۷۷ که در آن از مفهوم برنامه‌نویسی ساختیافته جدید استفاده می‌کنند، بازنویسی شده است. تمامی برنامه‌ها روی یك یا چندکامپیوتر امتحان، و در بسیاری از این حالات برونداد‌های‌حاصل ازکامپیوتر عرضه شده‌است. زمانی‌که‌برونداد‌های عددی داده‌می‌شوند، خود متن مشخص خواهد ساخت که برای به‌دست‌آوردن این نتایج از کدام نوع کامپیوتر (یونی‌وك، سی. دی. سی، آی. بی. ام[5]) استفاده شده است.

مطالب این کتاب بیش‌از آنی است کـه معمولا در یك ترم تحصیلی عـادی دورهٔ کارشناسی برای رشته‌های‌علوم تدریس می‌شود. این‌امر درطرح و چگونگی ارائهٔ درس، آزادی بیشتری به‌معلم خواهد داد. به‌همین دلیل توجه به این موارد مهم است‌که مطالبی که در زمینهٔ درونیابی بسجمله‌ای درفصل ۲، دستگاههای خطی درفصل ۴ و مشتقگیری و انتگرالگیری درفصل ۷ بیان می‌شوند مسلماً برای فصول بعدی ضروری هستند. مطالبی‌که در هفت فصل اول کتاب (بـه‌غیراز قسمتهای ستاره‌دار) آمده‌انـد ساختار مطلوبی برای اولین‌دورهٔ این درس را تشکیل می‌دهند.

در اینجا فرصت را مغتنم شمرده از کلیهٔ کسانی که از راه مکاتبه خطاهـای چاپ و اشتباهات موجود درچاپ دوم را متذکر شده‌اند، و همچنین پیشنهاداتی برای بهبود کتاب به‌ما داده‌اند تشکر می‌کنیم.

س. د. کونت

کارل. دوبور

---

1. Fast Fourier Transforms 2. splines 3. collocation
4. procedure-oriented 5. UNIVAC, CDC, IBM

مقدمه

دراین کتاب حل عملی مسائل با کامپیوتر مطرح می‌شود. در روند حل یك مسئله امکان دارد چندین مرحلهٔ کم و بیش متمایز وجـود داشته باشد. نخستین مرحله صورتبندی[1] است. در صورتبندی یك الگوریتم ریاضی ازیك وضعیت فیزیکی، دانشمندان می‌باید ازقبل واقعیت امید به‌حل یك مسئله با کامپیوتر را درنظر داشته باشند. لذا برای هدفهای خاص ونوع و مقدار برونداد، دروندادهای مناسب، آزمونهای کافی‌فراهم سازند.

بعداز آنکه مسئله‌ای صورتبندی شد، برای حل آن روشهای عددی همراه با تحلیل خطای مقدماتی‌می‌باید طرح‌شود. هرروش عددی‌که بتواند برای حل مسئله‌ای به‌کار رود، الگوریتم نامیده می‌شود. یك الگوریتم مجموعهٔ کامل و روشنی از روندهاست کـه به‌حل یك مسئلهٔ ریاضی منجر می‌شود. انتخاب ساختن الگوریتمهـای مناسب به‌ویژه در محدودهٔ آنالیز عددی صورت می‌گیرد. متخصصان آنالیز عددی، پس‌از اینکـه دربارهٔ الگوریتم خاص یا مجموعه‌ای از الگوریتمها، برای حل مسئله‌ای تصمیم گرفتند، می‌باید همهٔ منابع خطا راکه ممکن است بر نتایج اثر گذارند درنظر بگیرند. باید درجهٔ دقت مورد نیاز را رعایت کنند، دامنهٔ خطاهای گردکردن و گسسته‌سازی را برآورد نمایند،نمو مناسبی باتعداد بارست‌های لازم معین وملاکهای کافی برای دقت تهیه‌کنند و عمل تصحیحی را درصورت عدم‌همگرایی منظور نمایند.

مرحلهٔ سوم حل‌مسئلهٔ برنامه‌نویسی است. برنامه‌نویس باید الگوریتمی را که به‌آن عقیده‌مند شده است، به‌مجموعهٔ دستورالعملهای غیرمبهم گـام به‌گامی برای کامپیوتر تبدیل نماید. نخستین گام این شیوه عمل «روند نماسازی[2]» نامیده مـی‌شود. یك روندنما اصولا مجموعه‌ای است از شیوه‌های عمل، معمولا به‌شکل قالبی منطقی، که کامپیوتر دنبال می‌کند و می‌تواند به‌صورت‌دستورالعمل نموداری یا عبارتی باشد. پیچیدگی‌روندنما به‌پیچیدگی مسئله و میزان جزئیات مندرج در آن بستگی‌دارد.ولی‌باید غیراز خود برنامه‌نویس شخص

---

1. formulation 2. flow charting

دیگری هم بتواند جریان اطلاعات را از روی روندنما دنبال کند. روندنما کمك مؤثری است برای برنامه‌نویس که باید توابع اصلی‌را به‌برنامه تبدیل کند، و درعین‌حال وسیله‌ای مؤثر برای ارتباط دیگرانی است که می‌خواهند کاری را که برنامه انجام می‌دهد بفهمند. دراین کتاب،روندنما را گاهی به شکل نموداری ولی اغلب به‌صورت دستورالعمل عبارتی عرضه می‌کنیم. هنگامی کـه روندنما به‌صورت نموداری به‌کار برده شود، قراردادهـای استانده دنبال می‌شوند. ولی وقتی به‌صورت دستورالعمل عبارتی باشد، از زبان بی‌نیاز از توضیح **الگول**-مانندی استفاده خواهد شد. پس‌از تهیهٔ یك روندنما، برنامه‌نویس شیوه‌های مذکور را به‌مجموعه‌ای از دستورالعملهای ماشینی تبدیل می‌کند.این‌کار را می‌توان مستقیماً به‌زبان ماشین یا به‌زبان هم‌گذاری[1] یا یك‌زبان شیوه- پرداخته انجام داد. در این کتاب منحصراً از یك گویش[2] خاص فورترن، به‌نام فورترن ۷۷ استفاده شده است. فورترن ۷۷ گویش جدیدی است از زبان فورترن کـه عبارتهای ممیزی تـازه‌ای را وارد می‌کند و برمفاهیم برنامه‌نویسی ساختیافته[3] جدیدی تکیه‌دارد. درحالی‌که همگردان‌های[4] فورترن۴ تقریباً روی همهٔ کامپیوترها وجود دارند. دستیابی به‌همگردان فورترن ۷۷ به‌آسانی میسر نیست ولی تبدیل برنامه از فورترن ۷۷ به‌فورترن ۴ نسبتاً به‌سادگی انجام می‌گیرد.

زبانهای شیوه- پرداخته مـانند فورترن یـا الگول، گاهی زبان الگوریتمی نامیده می‌شوند. این زبانها امکان بیان یك‌الگوریتم ریاضی را به‌شکل مناسبتری برای ارتباط با کامپیوترها فراهم می‌سازند. یك‌شیوهٔ عمل فورترنی، که یك الگوریتم را پیاده می‌کند،در حالت‌کلی از خود الگوریتم دقیقتر است. مثلا اگراین الگوریتم ریاضی مشخص‌کنندهٔ یك شیوهٔ عمل بارستی برای پیدا کردن یك معادلـه باشد، برنامهٔ فورترن مـی‌باید سه نکتهٔ: (۱) دقت لازم، (۲) تعداد بارستهایی‌که باید انجام گیرد و (۳) درصورت ناهمگرایی چه باید کرد، را مشخص کند. بیشتر الگوریتمهای این‌کتاب به‌شکل عادی ریاضی و به‌شکلی دقیقتر ازیك شیوهٔ عمل فورترنی داده شده‌اند.

در بسیاری از مراکزکامپیوتری هریك از این‌مراحل حل‌مسئله را شخص جداگانه‌ای انجام می‌دهد. در برخی دیگر ممکن است شخص واحدی مسئول هرسه کار باشد. واضح است کـه بین این‌سه مرحله فعل و انفعالهایی وجود دارند. هرچه اجرای برنامه پیشرفت می‌کند، اطلاعات بیشتری بـه‌دست می‌آید و این اطلاعات ممکن است تغییراتی را در صورتبندی، در الگوریتمهای به‌کار رفته و درخود برنامه موجب شوند.

---

1. assembly language
2. dialect
3. structured-programming
4. compilers

۱

دستگاه اعداد و خطاها

در این فصل روشهای نمایش اعداد درکامپیوتر و خطاهای ناشی از این نمایشها را بررسی می‌کنیم. بعلاوه، منشأ انواع خطاهای گوناگون محاسباتی و انتشار بعدی آنها را مورد بررسی قرار می‌دهیم، و اندکی هم به‌بحث درمقدمات ریاضی می‌پردازیم.

۱.۱ نمایش اعداد صحیح

درزندگی‌روزمره اعداد را درپایهٔ دستگاه‌دهدهی به‌کارمی‌بریم. از این‌رو مثلا عدد ۲۵۷ به‌صورت زیر قابل‌بیان است

$$۲۵۷ = ۲\times ۱۰۰ + ۵\times ۱۰ + ۷\times ۱$$

$$= ۲\times ۱۰^{۲} + ۵\times ۱۰^{۱} + ۷\times ۱۰^{۰}$$

عدد ۱۰ را پایه[1] این‌دستگاه‌می‌نامیم. هرعدد صحیح در دستگاه‌دهدهی را می‌توان به‌صورت یك بسجمله‌ای در پایهٔ ۱۰ بیان‌کرد که ضرایب آن صحیح و بین ۰ و ۹ باشند. برای مشخص‌کردن هرعدد صحیح مثبت درپایهٔ ۱۰، نشانگذاری زیر را به‌کار می‌بریم

---

1. base

$$N=(a_n a_{n-1}\cdots a_0)_{10}$$
$$=a_n 10^n+a_{n-1}10^{n-1}+\cdots+a_0 10^0 \qquad (1.1)$$

در به‌کارگیری پایهٔ ۱۰ هیچ دلیل غریزی وجود ندارد. تمدنهای دیگر، پایه‌های دیگری مانند ۱۲، ۲۰ یا ۶۰ را به‌کار برده‌اند. کامپیوترهای جدید تپهای[1] ارسالی ابزارهای الکتریکی را دریافت می‌کنند. حالت یك ضربهٔ الکتریکی به‌یکی از دوصورت وصل یا قطع است. بنابراین نمایش اعداد درکامپیوتر در دستگاه **دودویی**[2] مناسبتر است. پایهٔ این دستگاه ۲ است وضرایب صحیح می‌توانند مقادیر صفر یا یك باشند.

یك عدد صحیح نامنفی N در دستگاه دودویی به‌صورت

$$N=(a_n a_{n-1}\cdots a_1 a_0)_2$$
$$=a_n 2^n+a_{n-1}2^{n-1}+\cdots+a_1 2^1+a_0 2^0 \qquad (2.1)$$

نمایش داده‌می‌شود، که در آن ضرایب a_k صفر یا یك‌هستند. ملاحظه‌می‌کنید که N باز به‌وسیلهٔ یك بسجمله‌ای، اما اکنون در پایهٔ ۲، نمایش داده شده است. بسیاری از کامپیوترهایی که درکارهای علمی مورد استفاده قرار می‌گیرند ذاتاً در دستگاه دودویی عمل می‌کنند. لیکن استفاده‌کنندگان ازاین کامپیوترها ترجیح می‌دهند در دستگاه‌دهدهی، که‌برای آنها مأنوستر است، کار کنند. بنابراین، هنگام دادن اطلاعات بـه کامپیوتر روشهایی برای تبدیل اعداد دهدهی به‌دودویی، و هنگام دریافت‌اطلاعات ازآن، روشهایی برای تبدیل اعداد دودویی به‌دهدهی، مورد نیاز است.

تبدیل اعداد دودویی به‌دهدهی ممکن است مستقیماً طبق‌تعریف (۲.۱) انجام پذیرد. برای مثال داریم:

$$(11)_2=1\times 2^1+1\times 2^0=3$$
$$(1101)_2=1\times 2^3+1\times 2^2+0\times 2^1+1\times 2^0=13$$

تبدیل اعدادصحیح از یك‌پایهٔ β به‌پایهٔ ۱۰ به‌وسیلهٔ الگوریتم زیر، که‌روش به‌دست آوردن آن درفصل دوم آمده، نیز انجام‌پذیر است.

الگوریتم ۱.۱ ضرایب بسجمله‌ای

$$p(x)=a_n x^n+a_{n-1}x^{n-1}+\cdots+a_1 x+a_0 \qquad (3.1)$$

یعنی a_n، ...، a_1، a_0 و عدد β داده شده‌اند. اعداد b_n، b_{n-1}، ...، b_0 را باید به‌طور بازگشتی[3] به‌شکل زیر محاسبه‌کرد

---

1. pulses 2. binary 3. recursively

$$b_n := a_n$$
$$b_{n-1} := a_{n-1} + b_n\beta$$
$$b_{n-2} := a_{n-2} + b_{n-1}\beta$$
$$b_{n-3} := a_{n-3} + b_{n-2}\beta$$
$$\cdots\cdots\cdots\cdots$$
$$b_0 := a_0 + b_1\beta$$

بنابراین

$$b_0 := p(\beta)$$

از آنجا که طبق تعریف (۲.۱)، عدد صحیح دودویی $(a_n a_{n-1}\cdots a_0)_2$ نمایانگر مقدار بسجملـه‌ای (۳.۱) به ازای $x = 2$ است، برای پیدا کردن معادل دهدهی یك عدد صحیح دودویی می‌توانیم الگوریتم (۱.۱) را با $\beta = 2$ به‌کار بریم.

بنابراین معادل‌دهدهی عدد $(1101)_2$ که با استفاده از الگوریتم (۱.۱) محاسبه‌شده به‌صورت زیر است:

$$b_3 = 1$$
$$b_2 = 1 + 1 \times 2 = 3$$
$$b_1 = 0 + 3 \times 2 = 6$$
$$b_0 = 1 + 6 \times 2 = 13$$

و عدد دهدهی معادل $(10000)_2$ چنین است:

$$b_4 = 1$$
$$b_3 = 0 + 1 \times 2 = 2$$
$$b_2 = 0 + 2 \times 2 = 4$$
$$b_1 = 0 + 4 \times 2 = 8$$
$$b_0 = 0 + 8 \times 2 = 16$$

اگر کسی بخواهد حساب دودویی را به‌کار برد،می‌تواند تبدیل عدد صحیح دهدهی N به معادل دودویی آن را نیز به‌وسیلهٔ الگوریتم (۱.۱) انجام دهد. زیرا درصورتی کـه

$N=(a_n a_{n-1}\cdots a_0)_{10}$، آنگاه طبق تعریف (۱.۱) داریم $N=p(10)$، و در اینجا $p(x)$ همان بسجمله‌ای (۳.۱) است. لذا می‌توانیم نمایش دودویی N را چنین محاسبه کنیم که ابتدا ضرایب a_n، ...، a_0 را به‌اعداد صحیح دودویی تبدیل و سپس با استفاده از الگوریتم (۱.۱)، مقدار $p(x)$ را به‌ازای $x=10=(1010)_2$ درحساب دودویی محاسبه کنیم. برای مثال اگر $N=187$، آنگاه:

$$187=(187)_{10}=1\times10^2+8\times10^1+7\times10^0$$
$$=(1)_2(1010)_2^2+(1000)_2(1010)_2^1+(111)_2(1010)_2^0$$

اکنون با استفاده از الگوریتم (۱.۱) و حساب دودویی داریم:

$$b_2=(1)_2$$
$$b_1=(1000)_2+(1)_2(1010)_2=(1000)_2+(1010)_2=(10010)_2$$
$$b_0=(111)_2+(10010)_2(1010)_2=(111)_2+(10110100)_2$$
$$=(10111011)_2$$

بنابراین

$$187=(10111011)_2$$

گرچه اعداد دودویی و حساب دودویی برای کامپیوترهای امروزی به‌طور آرمانی مناسب‌اند، استفاده از آنها به‌سبب تعداد ارقامی که حتی برای نمایش اعداد متوسط لازم است، برای شخص تاحدی خسته‌کننده است. چنانچه ملاحظه شد برای نمایش عدد دهدهی سه‌رقمی ۱۸۷، هشت‌رقم دودویی لازم است. دستگاه **هشت‌هشتی (اکتال)**[1] که از پایهٔ ۸ بهره می‌گیرد، تقریباً واسطه‌ای بین دستگاه دودویی مطلوب کامپیوتر، و دستگاه دهدهی مطلوب افراد است. از آنجا که سه‌رقم دودویی معرف یک رقم هشت‌هشتی است، به‌سادگی می‌توان دستگاه هشت‌هشتی را به‌دستگاه دودویی، و به‌عکس اعداد دودویی را به‌هشت‌هشتی تبدیل کرد. برای تبدیل یک عدد هشت‌هشتی به‌دودویی فقط باید به‌جای تمامی ارقام عدد هشت‌هشتی، اعداد دودویی هم ارزش آنها را گذاشت، بنابراین

$$(347)_8=(011\ \ 100\ \ 111)_2=(11100111)_2$$

بعکس، برای تبدیل اعداد دودویی به‌هشت‌هشتی، رقمهای عدد دودویی را به‌گروههای سه‌تایی (از طرف راست) افراز می‌کنیم و سپس به‌جای هر گروه سه‌تایی معادل هشت‌هشتی آن را قرار می‌دهیم، بنابراین:

---

1. octal

$$(۱۰۱۱۱۰۱۱)_۲=(۰۱۰\ \ ۱۱۱\ \ ۰۱۱)_۲=(۲۷۳)_۸$$

درصورتی که بخواهیم یك‌عدد صحیح دهدهی را با محاسبهٔ دستی به‌دودویی تبدیل کنیم، معمولاً سریعترین راه این است کـه ابتدا با استفاده از الگوریتم ۱.۱، این عدد را به‌هشت‌هشتی و سپس‌از هشت‌هشتی به‌دودویی تبدیل‌کنیم. مثال‌قبلی را درنظر می‌گیریم

$$۱۸۷=(۱۸۷)_{۱۰}=(۱)_۸(۱۲)_۸^۲+(۱۰)_۸(۱۲)_۸^۱+(۷)_۸(۱۲)_۸^۰$$

با استفاده از الگوریتم ۱.۱ و با قراردادن $(۱۲)_۸=۱۰$ به‌جای ۲ و به‌کار گرفتن حساب هشت‌هشتی داریم

$$b_۲=(۱)_۸$$
$$b_۱=(۱۰)_۸+(۱)_۸(۱۲)_۸=(۲۲)_۸$$
$$b_۰=(۷)_۸+(۲۲)_۸(۱۲)_۸=(۷)_۸+(۲۶۴)_۸=(۲۷۳)_۸$$

بنابراین سرانجام داریم:

$$۱۸۷=(۲۷۳)_۸=(۰۱۰۱۱۱۰۱۱)_۲$$

تمرین

۱.۱–۱ اعداد دودویی زیر را به‌دهدهی تبدیل کنید:

$(۱۰۰۰۰۰۰۰۱)_۲$ $(۱۰۰۱۰۱)_۲$ $(۱۰۱۰)_۲$

۱.۱–۲ اعداد دهدهی زیر را به‌دودویی تبدیل‌کنید:

۸۲ ۱۰۹ ۳۴۳۳

۱.۱–۳ اعداد تمرینهـای ۱.۱–۱ و ۱.۱–۲ را ابتدا به‌صورت هشت‌هشتی تبدیل کنید و سپس اعداد به‌دست آمده را در دستگاه خواسته شده بنویسید.

۱.۱–۴ زیربرنامه‌ای[1] به‌زبان فورترن بنویسیدکه عددی با NIN رقم را درپایهٔ مفروض BETIN واقع دریك‌آرایهٔ[2] یك‌بعدی NUMIN بپذیرد و عدد NOUT رقمی هم‌ارز باآن درپایهٔ مفروض BETOUT را در یك آرایهٔ یك‌بعدی NUMOUT ذخیره‌کند. برای سادگی، هم BETIN و هم BETOUT را به‌اعداد ۲، ۴، ۸، ۱۰ محدودکنید.

---

1. subroutine 2. array

۲.۱ نمایش کسرها

اگر x یک‌عدد مثبت حقیقی باشد، قسمت **صحیح** x را که بزرگترین عدد صحیح نابزرگتر از x است، با x_I و قسمت **کسری** x را با x_F نمایش می‌دهیم، بنابراین

$$x_F = x - x_I$$

قسمت کسری را همیشه می‌توان به‌صورت کسر اعشاری زیر نوشت:

$$x_F = \sum_{k=1}^{\infty} b_k 10^{-k} \qquad (۴.۱)$$

که در آن هر b_k یک عدد صحیح نامنفی کمتر از ۱۰ است. اگر به ازای جمیع kهای بزرگتر از یک عدد صحیح معین، داشته باشیم $b_k = 0$، آنگاه کسر بالا را **مختوم**[1] گویند. از این‌رو کسر

$$\frac{1}{4} = 0.25 = 2 \times 10^{-1} + 5 \times 10^{-2}$$

مختوم است، درحالی که کسر

$$\frac{1}{3} = 0.333\ldots = 3 \times 10^{-1} + 3 \times 10^{-2} + 3 \times 10^{-3} + \cdots$$

مختوم نیست.

اگر قسمت صحیح x به‌عنوان یک عدد صحیح در دستگاه دهدهی به‌صورت

$$x_I = (a_n a_{n-1} \cdots a_0)_{10}$$

و قسمت کسری آن طبق فرمول (۴.۱) داده شده باشد، معمولا دو قسمت را پشت سرهم می‌نویسند و با یک «ممیز» آنها را از هم جدا می‌کنند.

$$x = (a_n a_{n-1} \cdots a_0 . b_1 b_2 b_3 \cdots)_{10}$$

عیناً به‌همین‌طریق می‌توان قسمت کسری x را به‌صورت یک کسر دودویی نوشت:

$$x_F = \sum_{k=1}^{\infty} b_k 2^{-k}$$

که در آن هر b_k عدد صحیحی است نامنفی کوچکتر از ۲، یعنی یا صفر یا یک. اگر قسمت صحیح x به‌وسیلهٔ عدد صحیح دودویی

---

1. terminate

$$x_I = (a_n a_{n-1} \cdots a_0)_2$$

داده شده باشد، می‌توان عدد x را با استفاده از «ممیز»، به‌صورت زیر نوشت:

$$x = (a_n a_{n-1} \cdots a_0 . b_1 b_2 b_3 \cdots)_2$$

کسر دودویی $(.b_1 b_2 b_3 \cdots)_2$ را برای عدد مفروض x_F، که بین صفر و یک است، می‌توان به‌صورت زیر محاسبه کرد: اگر

$$x_F = \sum_{k=1}^{\infty} b_k 2^{-k}$$

آنگاه

$$2x_F = \sum_{k=1}^{\infty} b_k 2^{-k+1} = b_1 + \sum_{k=1}^{\infty} b_{k+1} 2^{-k}$$

بنابراین b_1 قسمت صحیح $2x_F$ است، در حالی که

$$(2x_F)_F = 2x_F - b_1 = \sum_{k=1}^{\infty} b_{k+1} 2^{-k}$$

بنابراین اگر این روش را تکرار کنیم خواهیم دید که b_2 قسمت صحیح $2(2x_F)_F$، b_3 قسمت صحیح $2(2(2x_F)_F)_F$ و هکذا برای بقیه می‌باشند.

مثلا اگر $x_F = 0.625$، آنگاه

$2(0.625) = 1.25$	بنابراین	$b_1 = 1$
$2(0.25) = 0.5$	بنابراین	$b_2 = 0$
$2(0.5) = 1.0$	بنابراین	$b_3 = 1$

و کلیهٔ b_kهای بعدی صفرند، درنتیجه

$$0.625 = (0.101)_2$$

این مثال به‌گونه‌ای انتخاب شده بود که کسر دودویی آن مختوم باشد. متأسفانه تمامی کسرهای دهدهی مختوم به کسرهای دودویی مختوم منجر نمی‌شوند. این امر ناشی از این حقیقت است که کسر دودویی معادل $0.1 = 10^{-1} = x_F$ مختوم نیست زیرا داریم

$2(0.1) = 0.2$	بنابراین	$b_1 = 0$
$2(0.2) = 0.4$	بنابراین	$b_2 = 0$

$$۰٫۸ = ۰٫۴)۲ \qquad \text{بنابراین} \qquad b_۲ = ۰$$
$$۱٫۶ = ۰٫۸)۲ \qquad \text{بنابراین} \qquad b_۴ = ۱$$
$$۱٫۲ = ۰٫۶)۲ \qquad \text{بنابراین} \qquad b_۵ = ۱$$

و اکنون به کسر ۰٫۲ باز می‌گردیم، و ارقام دوباره به‌طور دوره‌ای تکرار می‌شوند. در نتیجه

$$۰٫۱ = (۰٫۰\ ۰۰۱۱\ ۰۰۱۱\ \ldots)_۲$$

شیوه‌ای که به اجمال شرح داده شد در الگوریتم زیر به‌صورت فرمول درمی‌آید.

الگوریتم ۲۰۱ عدد x بین ۰ و ۱ و یک‌عدد صحیح β بزرگتر از ۱، داده شده‌اند. اعداد $b_۱$، $b_۲$، $b_۳$، ... را به‌طور بازگشتی به‌طریق زیر تولید می‌کنیم:

$$c_۰ := x$$
$$b_۱ := (\beta c_۰)_I \ , \ c_۱ := (\beta c_۰)_F$$
$$b_۲ := (\beta c_۱)_I \ , \ c_۲ := (\beta c_۱)_F$$
$$\ldots\ldots\ldots\ldots\ldots$$

سپس

$$x = (۰٫b_۱b_۲b_۳\ldots)_\beta = \sum_{k=۱}^{\infty} b_k \beta^{-k}$$

بنابر دو دلیل این الگوریتم را برای یک پایهٔ کلی β بیان کرده‌ایم، و نه برای یک پایهٔ بخصوص دودویی $\beta = ۲$. اگر این تبدیل به دودویی با مداد و کاغذ انجام پذیرد، معمولا اگر عدد فوق اول به هشت‌هشتی، $\beta = ۸$، و سپس از هشت‌هشتی به دودویی تبدیل شود، سرعت عمل بیشتر خواهد بود. همچنین می‌توان با انتخاب $\beta = ۱۰$ و استفاده از محاسبات دودویی (یا هشت‌هشتی)، این الگوریتم را برای تبدیل یک کسر دودویی (یا هشت‌هشتی) به‌دهدهی به‌کار برد.

برای مثال، اگر $x = (۰٫۱۰۱)_۲$، آنگاه به‌ازای $\beta = ۱۰ = (۱۰۱۰)_۲$، و به‌کمک حساب دودویی و با استفاده از الگوریتم ۲۰۱، داریم

$$۱۰(۰٫۱۰۱)_۲ = (۱۰۱۰٫۱۱۰)_۲^* \qquad \text{بنابراین} \qquad b_۱ = (۱۱۰)_۲ = ۶ \text{، } c_۱ = (۰٫۰۱۰)_۲$$
$$۱۰(۰٫۰۱۰)_۲ = (۱۰۰٫۱۰۰)_۲ \qquad \text{بنابراین} \qquad b_۲ = (۱۰)_۲ = ۲ \text{، } c_۲ = (۰٫۰۰۱)_۲$$
$$۱۰(۰٫۰۰۱)_۲ = (۱۰٫۰۱۰۱)_۲ \qquad \text{بنابراین} \qquad b_۳ = (۱۰۱)_۲ = ۵ \text{، } c_۳ = ۰$$

---

* درعبارت $۱۰(۰٫۱۰۱)_۲$ عدد ۱۰ در پایهٔ دهدهی است. ازآنجا که $(۱۰۱۰)_۲ = (۱۰)_{۱۰}$، بنابراین $(۱۰۱۰٫۱۱۰)_۲ = (۱۰۱)(۱۰۱۰)_۲ = (۰٫۱۰۱)_۲ ۱۰$.ـم.

بنابراین b_kهای بعدی صفر خواهند بود. این نشان می‌دهد که

$$(۰٫۱۰۱)_۲ = ۰٫۶۲۵$$

که مؤید محاسبات قبلی است. باید توجه داشت که هرگاه x_β یک کسر دودویی مختوم n رقمی باشد، در دستگاه دهدهی نیز کسری مختوم است و تعداد ارقام آن برابر با n. زیرا

$$(۰٫۱)_۲ = ۰٫۵$$

تمرین

۱-۲۰۱ کسرهای دودویی زیر را به کسرهای دهدهی تبدیل کنید.

$$(۰٫۱۱۱۱۱۱۱۱۱)_۲ \qquad (۰٫۱۱۰۰۰۰۱۱)_۲$$

۲-۲۰۱ پنج رقم اول ۰٫۱ را درصورتی که به‌صورت یک کسر هشت‌هشتی نوشته شود پیدا، وسپس به‌وسیلهٔ این کسر، ۱۵ رقم اول ۰٫۱ را در دستگاه دودویی محاسبه کنید.

۳-۲۰۱ کسرهای هشت‌هشتی زیر را به کسرهای دهدهی تبدیل کنید:

$$(۰٫۷۷۶)_۸ \qquad (۰٫۶۱۴)_۸$$

جوابهای این مسئله را باجوابهای مسئلهٔ ۲۰۱-۱ مقایسه کنید.

۴-۲۰۱ یک عدد دودویی پیدا کنید که π را باتقریب $۱۰^{-۳}$ نشان دهد.

۵-۲۰۱ اگر بخواهیم بااستفاده از الگوریتم ۱۰۱، یک‌عددصحیح دهدهی N را به‌دودویی تبدیل کنیم، بایستی حساب‌دودویی را به‌کار گیریم. نشان دهیدکه چگونه این تبدیل با استفاده از الگوریتم ۲۰۱ و حساب دهدهی[1] انجامپذیر است.

(راهنمایی: N را بر توان مناسبی از ۲ تقسیم، و سپس نتیجه را به‌دودویی تبدیل کنید و «ممیز دودویی» را به‌طور مناسبی تغییرمکان دهید).

۶-۲۰۱ اگر بخواهیم یک کسر مختوم دودویی x را با استفاده از الگوریتم ۲۰۱ به‌یک کسر دهدهی تبدیل کنیم، بایستی حساب دودویی را به‌کار گیریم. نشان دهیدکه چگونه این تبدیل با به‌کار بستن الگوریتم ۱۰۱ و حساب دهدهی انجامپذیر است.

۳۰۱ حساب با ممیز شناور

محاسبات علمی معمولا در حساب باممیز شناور انجام می گیرد.

یک‌عدد با n رقم ممیز شناور در پایهٔ β به‌شکل زیر است:

$$x = \pm(۰٫d_۱d_۲\cdots d_n)_\beta \beta^e \qquad (۵۰۱)$$

1. decimal

که در آن $(0.d_1d_2\ldots d_n)_\beta$ کسری است در پایهٔ β که **جزءکسری یا مانتیس**[1] نامیده‌می‌شود و e عددصحیحی است که آن‌را **نما**[2] می‌نامند. هر گاه $d_1\neq 0$ ویا $d_1=d_2=\ldots=d_n=0$، چنین عدد باممیز شناور را **نرمال‌شده**[3] نامند.

اگر چه در برخی از کامپیوترها $\beta=16$ ودر محاسبات دستی و حسابگرهای جیبی و رومیزی $\beta=10$، لیکن در اکثر کامپیوترها $\beta=2$ گرفته می‌شود.

معمولا در هر کامپیوتری، دقت یاطول n در اعداد با ممیز شناور به‌طول کلمه در آن کامپیوتر بستگی دارد، از این‌رو ممکن‌است دقت این اعداد در بسیاری جاها متفاوت باشند (به‌شکل ۱.۱ مراجعه شود). دردستگاههای محاسباتی که برنامه‌های فورترن را می‌پذیرند قرار براین است که اعداد با ممیزشناور را با دو طول مختلف به‌کار گیرند که یکی تقریباً دوبرابر دیگری است. جز در شرایط خاص کـه به‌اصطلاح به **دقت مضاعف**[4] نیاز است، درحالت عادی از اعداد با ممیزشناور باطول کوتاهتر به‌اصطلاح با **دقت معمولی**[5] استفاده می‌شود. معمولا انجام محاسبات بـادقت مضاعف، در مقایسه بادقت ساده، نیاز به‌حافظه‌ای دوبرابر دارد و زمان محاسبه را به‌بیش‌از دوبرابر افزایش می‌دهد.

نمای e معمولا در محدودهٔ زیر

$$m<e<M \qquad (۶.۱)$$

تغییر می‌کند که m و M اعداد صحیح معینی هستند.

معمولا $m=-M$، لیکن ممکن است این محدوده در بسیاری از جاها تغییر یابد، به‌شکل ۱.۱ نگاه‌کنید.

برای تبدیل یك‌عدد حقیقی x به‌یك عدد $fl(x)$، n رقمی باممیزشناور درپایه‌β، دو روش متداول است یکی **گردکردن**[6] و دیگری **قطع‌کردن**[7]. در روش گردکـردن، عـدد

Computer	β	n	$M=-m$
IBM 7094	2	27	2^7
Burroughs 5000 Series	8	13	2^6
IBM 360/370	16	6	2^6
CDC 6000 and Cyber Series	2	48	2^{10}
DEC 11/780 VAX	2	24	2^7
Hewlett Packard 67	10	10	99

شکل ۱.۱ مشخصه‌های ممیز شناور.

1. mantissa 2. exponent 3. normalized
4. double precision 5. single precision 6. rounding
7. chopping

$fl(x)$ بـه عنوان نزدیکترین عدد نرمال‌شده بـا ممیز شناور به x، انتخاب می‌شود. در حالتی کـه محدودیتی وجود داشته بـاشد، از بعضی قواعد خاص مانند گرد کردن متقارن (گرد کردن به یك عدد زوج) استفاده می‌شود. در روش **قطع کردن** عدد $fl(x)$ به عنوان نزدیکترین عدد نرمال شدهٔ بـا ممیز شناور بین x و ۰ انتخاب می‌شود. برای مثال اگر اعداد دهدهی دورقمی با ممیز شناور به‌کار برده شوند، داریم

$$fl\left(\frac{۲}{۳}\right)=\begin{cases}(۰٫۶۷)\times ۱۰^{۰} & \text{گردشده}\\ (۰٫۶۶)\times ۱۰^{۰} & \text{قطع‌شده}\end{cases}$$

و

$$fl(-۸۳۸)=\begin{cases}-(۰٫۸۴)\times ۱۰^{۳} & \text{گردشده}\\ -(۰٫۸۳)\times ۱۰^{۳} & \text{قطع‌شده}\end{cases}$$

در برخی از کامپیوترها این تعریف عدد $fl(x)$ درحالت $|x|\geqslant\beta^{M}$ (**حالت سرریز**)[1] یا $0<|x|<\beta^{m-n}$ (**حالت پی‌ریز**)[2]، که m و M کرانه‌های نما می‌باشند، اصلاح‌شده است؛ دراین‌حالت یا $fl(x)$ تعریف نشده است که این‌امر موجب قطع برنامه می‌شود و یا $fl(x)$ به‌وسیلهٔ عدد خاصی نمایش داده شده‌است که خود این عدد هنگام ترکیب با اعداد با ممیز شناور معمولی از قواعد متداول حساب پیروی نمی‌کند.

تفاضل بین x و $fl(x)$ **خطای گردکردن** نامیده می‌شود. این خطا به‌مقدار x بستگی دارد و از این‌رو بهترین روش برای اندازه‌گیری آن سنجش آن نسبت به x است. زیرا اگر بنویسیم

$$fl(x)=x(۱+\delta) \qquad (۷.۱)$$

که $\delta=\delta(x)$ عددی است که بـه x بستگی دارد، آنگاه می‌توانیم، حداقل مادامی که x موجب بروز سرریز وپی‌ریز نشود، δ را مستقل از x کراندار سازیم. برای چنین xی (که موجب سرریز وپی‌ریز نشود) دشوار نیست که نشان دهیم

$$|\delta|<\frac{۱}{۲}\beta^{۱-n} \qquad \text{در گردکردن} \qquad (۸.۱)$$

درحالی که

$$-\beta^{۱-n}<\delta\leqslant ۰ \qquad \text{درقطع کردن} \qquad (۹.۱)$$

به‌تمرین ۳۰۱ـ۳ نگاه کنید. حداکثر مقدار ممکن برای $|\delta|$ اغلب «**واحدگردکردن**» نامیده شده و با u نشان داده می‌شود.

هرگاه عملی حسابی روی دوعدد یا ممیز شناور انجام گیرد، نتیجه معمولا یك‌عدد یا

---

1. overflow 2. underflow

ممیز شناور باهمان‌طول نیست. اگر برای مثال با اعداد دهدهی دورقمی کارکنیم وبگیریم

$$x=(۰٫۲۰)۱۰^{۱}=۲ \qquad y=(۰٫۷۷)۱۰^{-۶} \qquad z=(۰٫۳۰)۱۰^{۱}=۳$$

آنگاه

$$x+y=(۰٫۲۰۰۰۰۰۰۷۷)۱۰^{۱} \qquad x\cdot y=(۰٫۱۵۴)۱۰^{-۵}$$

$$\frac{x}{z}=(۰٫۶۶۶\ldots)۱۰^{۰}$$

بنابراین، اگر ω معرف یکی‌از اعمال حسابی (جمع، تفریق، ضرب و یا تقسیم) و ω^* همان عمل باممیز شناور باشدکه به‌توسط کامپیوتر انجام شده است، آنگاه گرچه به‌ازای دوعدد باممیز شناور x، y نتیجهٔ‌محاسبهٔ‌کامپیوتر $x\omega^*y$ می‌باشد، اما می‌توان مطمئن بود که‌معمولا

$$x\omega^*y\neq x\omega y$$

گرچه ممکن است عمل بـا ممیز شناور ω^* متناظر با ω، در ماشینهای مختلف در بعضی از جزئیات کمی متفاوت باشد، اما معمولا ω^* طوری اختیار می‌شود که

$$x\omega^*y=fl(x\omega y) \qquad (۱۰.۱)$$

به‌عبارت دیگر، مجموع (تفاضل، حاصلضرب یا خارج‌قسمت) باممیز شناور دوعدد باممیز شناور معمولا مساوی عـددی است باممیز شناور کـه مجموع (تفاضل، حاصلضرب یـا خارج‌قسمت) دقیق این دوعدد را نمایش دهد. بنابراین (درصورتی‌که سرریز و پی‌ریزی پیش نیاید) داریم:

$$x\omega^*y=(x\omega y)(۱+\delta) \quad \text{برای مقداری از } \delta \text{ با } |\delta|\leqslant u, \qquad (۱۱.۱ \text{ الف})$$

که u واحد گردشده دراین‌فرمول است. دربرخی موارد مناسبتر آن‌است‌که فرمول معادل زیر به‌کار گرفته شود

$$x\omega^*y=(x\omega y)/(۱+\delta) \quad \text{برای مقداری از } \delta \text{ با } |\delta|\leqslant u, \qquad (۱۱.۱ \text{ ب})$$

معادلـهٔ (۱۱.۱) بیانـگر مـوضوع اساسی تحلیل خطاهـا بـه‌طریق «پسرو»[1] مـی‌باشد (به *J. H. Wilkinson [۲۴] نگاه کنید). توضیحاً اضافه می‌کنیم که معادلهٔ (۱۱.۱) امکان می‌دهد کـه یك‌نتیجه باممیز شناور به‌عنوان نتیجه‌ای متناظر باحساب معمولی، اما با داده‌هایی‌که اندکی تغییر یافته است، تلقی شود. ازاین‌رو، تحلیل نتیجه‌باحساب‌ممیزشناور را می‌توان برحسب حساب معمولی انجام داد.

برای مثال، مقدار تابع $f(x)=x^{۲}$* درنقطهٔ x_0 را می‌توان بـا n بار مربع‌کردن

---

1. backward

* عدد درون کروشه مربوط به‌مراجعی است‌که درآخر کتاب داده شده است.

یعنی باانجام مراحل متوالی زیر انجام داد.

$$x_1 := x_0^2,\ x_2 := x_1^2, \dots, x_n := x_{n-1}^2$$

که‌در رابطهٔ فوق $f(x_0) = x_n$. درحساب باممیز شناور، به‌جای آن، طبق‌معادلهٔ (۱۱.۱الف) دنبالهٔ اعداد زیر را محاسبه می‌کنیم.

$$\hat{x}_1 = x_0^2(1+\delta_1),\quad \hat{x}_2 = (\hat{x}_1)^2(1+\delta_2), \dots,\ \hat{x}_n = (\hat{x}_{n-1})^2(1+\delta_n)$$

که در آن به‌ازای جمیع مقادیر i، $|\delta_i| \leqslant u$. بنابراین جواب محاسبه‌شده برابراست با

$$\hat{x}_n = x_0^{2^n}(1+\delta_1)^{2^{n-1}} \cdots (1+\delta_{n-1})^2(1+\delta_n)$$

برای ساده‌کردن عبارت‌فوق‌توجه کنیدکه اگر $|\delta_2| \leqslant u$,$\dots$,$|\delta_1|$، آنگاه به‌ازای‌بعضی از مقادیر δ با $|\delta| \leqslant u$ داریم:

$$(1+\delta_1)\cdots(1+\delta_r) = (1+\delta)^r$$

(به‌تمرین ۳۰۱ـ۶ نگاه‌کنید). همچنین به‌ازای بعضی از مقادیر η باشرط $|\eta| \leqslant u$، داریم

$$(1+\delta)^r = (1+\eta)^{r+1}$$

درنتیجه به‌ازای بعضی از مقادیر δ با $|\delta| \leqslant u$

$$\hat{x}_n = x_0^{2^n}(1+\delta)^{2^n} = f(x_0(1+\delta))$$

یا اگر با عبارت بیان کنیم، مقدار محاسبه شدهٔ $\hat{x}_n$ برای $f(x_0)$ برابر است بامقدار دقیق $f(x)$ در شناسهٔ آشفته[1] $x = x_0(1+\delta)$.

اکنون با بررسی این‌امرکه چگونه مقدار تابع (مقدار دقیق محاسبه شده) $f(x)$ با آشفته‌شدن (تغییرجزئی) شناسهٔ x تغییر می‌کند، می‌توانیم اثری را که محاسبهٔ ممیزشناور بردقت عمل‌مقدارمحاسبه‌شده برای $f(x_0)$ برجا گذاشته‌است اندازه‌گیری‌کنیم، همان‌گونه که دربند بعدی انجام داده‌ایم. بعلاوه توجه داریم‌که درمثال موردبحث این‌خطا اساساً با خطای ناشی ازتبدیل دادهٔ اولیهٔ x_0 به‌یك‌عدد باممیز شناور، قابل‌مقایسه است.

به‌عنوان مثال دوم، که مخصوصاً درفصل ۴ مـوردتوجه قرار خواهدگرفت، محاسبهٔ عدد s درمعادلهٔ

$$a_1b_1 + \cdots + a_rb_r + a_{r+1}s = c \qquad (۱۲.۱)$$

را به‌وسیلهٔ فرمول

---

1. perturbed argument

$$s=\left(c-\sum_{k=1}^{r} a_k b_k\right)/a_{r+1}$$

مورد بررسی قرار می‌دهیم. اگر s از مراحل

$$\begin{aligned} s_0 &:= c \\ s_i &:= s_{i-1} - a_i b_i \qquad i = 1, \dots, r \\ s &:= s_r / a_{r+1} \end{aligned}$$

به‌دست آید، آنگاه اعداد متناظر محاسبه شده درحساب با ممیزشناور، در روابط زیر صدق می‌کنند:

$$\begin{aligned} \hat{s}_0 &= c \\ \hat{s}_i &= [\hat{s}_{i-1} - a_i b_i (1+\delta)](1+\delta), \quad i = 1, \dots, r \\ \hat{s} &= \hat{s}_r / [a_{r+1}(1+\delta)] \end{aligned}$$

دراینجامعادلهٔ (۱۱۰۱ الف) و (۱۱۰۱ب) را به‌کار برده‌ایم وبه‌مشخص کردن δهای مختلف به‌وسیلهٔ زیرنمایه[1] اعتنا نکرده‌ایم. درنتیجه داریم:

$$\begin{aligned} a_{r+1}(1+\delta)\hat{s} = \hat{s}_r &= \hat{s}_{r-1}(1+\delta) - a_r b_r (1+\delta)^2 \\ &= \hat{s}_{r-2}(1+\delta)^2 - a_{r-1} b_{r-1}(1+\delta)^3 - a_r b_r (1+\delta)^2 \\ &\vdots \\ &= \hat{s}_0 (1+\delta)^r - a_1 b_1 (1+\delta)^{r+1} - \dots - a_r b_r (1+\delta)^2 \end{aligned}$$

این رابطه نشان می‌دهد که مقدار محاسبه شدهٔ s برای $\hat{s}$، درمعادلهٔ آشفتهٔ زیر صدق می‌کند.

$$a_1 b_1 (1+\delta)^{r+1} + \dots + a_r b_r (1+\delta)^2 + a_{r+1}(1+\delta)\hat{s} = c(1+\delta)^r \qquad (۱۳۰۱)$$

توجه داریم که درحالت $a_{r+1} = 1$، یعنی درصورتی که نیازی به‌انجام آخرین تقسیم نباشد، می‌توانیم کلیهٔ نماها را به‌اندازهٔ ۱ کاهش دهیم.

تمرین

۱-۳۰۱ دریك کامپیوتر دهدهی* با مانتیس نرمال شدهٔ چهار رقمی، اعداد زیر داده شده‌اند

(الف) $۰٫۴۵۲۳ \times 10^4$ (ب) $۰٫۲۱۱۵ \times 10^{-3}$ (پ) $۰٫۲۵۸۳ \times 10^1$

---

1. subscripts

* منظور کامپیوتری است که با اعداد در دستگاه دهدهی کار می‌کند.ـم.

عملیات زیر را با اتخاذ روش گرد کردن متقارن انجام دهید و خطای موجود در نتیجه را تعیین کنید.

(الف) $(a)+(b)+(c)$ (ب) $(a)/(c)$

(پ) $(a)-(b)$ (ت) $(a)-(b)-(c)$

(ث) $(a)(b)/(c)$ (ج) $(a)\cdot(b)/(c)$

۲-۳۰۱ فرض می‌کنیم عدد $fl(x)$ ازراه قطع کردن داده شده باشد. نشان دهید که (جز درحالت سرریز وپی‌دیز) روابط زیر برقرارند

$$fl(-x)=-fl(x)$$

$$fl(\beta^r x)=\beta^r fl(x)$$

۳-۳۰۱ فرض می‌کنیم $fl(x)$ به‌وسیلهٔ قطع کردن x داده شده باشد. $\delta=\delta(x)$ را طوری انتخاب می‌کنیم که $fl(x)=x(1+\delta)$. (درحالت $x=0$، $\delta=0$ انتخاب شود) نشان دهیدکه دراین‌صورت کرانه‌های δ ازفرمول (۹۰۱) به‌دست می‌آید.

۴-۳۰۱ با ذکرچند مثال نشان دهیدکه درحساب با ممیز شناور، اکثر قواعد حساب صدق نمی‌کنند. (راهنمایی: به‌قواعدی‌که مستلزم سه عملوند[1] هستند توجه‌کنید).

۵-۳۰۱ یك **تابع فورترن** FL(X) بنویسیدکه‌نتیجهٔ بازگشت آن‌یك‌عدد n رقمی‌اعشاری‌با ممیزشناور، حاصل‌ازگردکردن X باشد. n را ۴ فرض‌کنیدومحاسبات‌خود درتمرین ۳۰۱-۱ راکنترل کنید. (برای‌تعیین e با شرط $10^{e-1}\leqslant|x|<10^e$، از ALOG 10(ABS(X)) استفاده‌کنید).

۶-۳۰۱ گیریم $U=\{1+\delta : |\delta|\leqslant u\}$. نشان دهید که به ازای جمیع مقادیر $\alpha_1,\ldots,\alpha_r\in U$ مقداری مانند $\alpha\in U$ وجود دارد به طوری‌که $\alpha_1\alpha_2\cdots\alpha_r=\alpha^r$. همچنین‌نشان دهیدکه اگر $a_1,a_2,\ldots,a_r$ همگی دارای یك‌علامت باشند، به‌ازای‌مقداری چون $\alpha\in U$ رابطهٔ

$$a_1\alpha_1+a_2\alpha_2+\cdots+a_r\alpha_r=(a_1+a_2+\cdots+a_r)\alpha$$

برقراراست.

۷-۳۰۱ برای محاسبهٔ حاصلضرب داخلی $s=a_1b_1+a_2b_2+\cdots+a_nb_n$، یك تحلیل خطای پسرو انجام دهید. تحلیل را با این فرض‌که انباشتگی با دقت مضاعف به‌کارگرفته شده‌است تکرارکنید. این‌فرض بدین‌معنی‌است‌که نتایج هرضرب بادقت مضاعف‌نگهداری

---

1. operand

و با دقت مضاعف به‌حاصلجمع افزوده می‌شود وفقط درپایان، حاصلجمع نهایی با دقت معمولی گرد می‌گردد.

۴.۱ ازدست دادن ارقام بامعنی وپخش‌خطا، شرط و پایداری

اگرعدد x^* تقریبی‌برای جواب دقیقx باشد، آنگاه تفاضل $x-x^*$ را **خطای**x^*می‌نامیم. لذا

$$\text{مقدار دقیق} = \text{مقدار تقریبی} + \text{خطا} \qquad (۱۴.۱)$$

بنا برتعریف، عدد $(x-x^*)/x$، **خطای نسبی**[1] در x^*، تقریبی برای x است. توجه‌کنیدکه این‌عدد (اگر اصلا بسیارکوچك باشد) به عدد $(x-x^*)/x^*$ نزدیك است [تحقیقاً، اگر $\alpha=(x-x^*)/x$، آنگاه $(x-x^*)/x^*=\alpha/(۱-\alpha)$.]

دریك روند محاسباتی ازهرعمل باممیز شناور ممکن است خطایی حاصل شودکه میزان آن در عملیات بعدی احتمالا وسعت یاکاهش می‌یابد.

یکی از عادیترین راههای افزایش اهمیت یك خطا (که غالباً اجتناب‌پذیر است) به‌اصطلاح **از دست دادن ارقام بامعنی** است. اگر x^* تقریبی ازx باشد، x^* را تقریب x تا r رقم بامعنی درپایهٔ β گوییم، مشروط برآنکه خطای مطلق $|x-x^*|$ حداکثر برابر با $۱/۲$ رقم rام x در پایهٔ β باشد. این‌مطلب را می‌توان با رابطهٔ

$$|x-x^*| \leqslant \frac{۱}{۲}\beta^{s-r+۱} \qquad (۱۵.۱)$$

بیان کرد، که‌در آن s بزرگترین‌عدد صحیحی‌است‌که $\beta^s \leqslant |x|$. برای‌مثال، $x^*=۳$ تا یك رقم بامعنی(اعشاری) تقریبی برای $x=\pi$ است،درحالی‌که $x^*=۲۲/۷=۳$ر۱۴۲۸۰۰۰ به‌عنوان تقریبی برای π تاسه رقم با معنی صحیح است. اکنون فرض‌کنید می‌خواهیم عدد

$$z=x-y$$

را محاسبه‌کنیم وتقریبهای x^* وy^* را که هرکدام تا r رقم اعشاری‌صحیح است به‌ترتیب برای x وy دردست داریم. در این‌صورت

$$z^*=x^*-y^*$$

تقریبی است ازz که تا r رقم معتبر است، مگرآنکه یك یا چند رقم از x^* وy^* با هم یکی باشند. درحالت اخیر، هنگام عمل تفریق برخی از ارقام، صفر و در نتیجه تعداد ارقام صحیح z^* کمتر ازr خواهد شد.

---

1. relative error

برای مثال، فرض می‌کنیم

$$x^* = (0.76542421)10^1 \qquad y^* = (0.76544200)10^1$$

به‌ترتیب تقریبی از x و y تاهفت رقم با معنی صحیح باشند. دراین‌صورت تفاضل **دقیق** بین x^* و y^* درحساب باممیزشناور هشت رقمی برابراست با

$$z^* = x^* - y^* = (0.12210000)10^{-3}$$

اما به‌عنوان تقریبی برای $z = x - y$، z^* فقط تاسه رقم صحیح است زیرا رقم چهارم در z^* از هشتمین ارقام x^* و y^* به‌دست‌آمده است که هر دواینها احتمالا متضمن خطا هستند. بنابراین درعین‌اینکه حداکثرمقدار خطا در z^* (به‌عنوان تقریبی برای $z = x - y$) برابر با مجموع خطاها در x^* و y^* است، خطای نسبی در z^* احتمالا ۱۰۰۰۰ برابر خطای نسبی در x^* یا y^* است. بنابراین ازدست دادن ارقام با معنی زمانی خطرناک است که بخواهیم خطای نسبی راکوچك نگهداریم.

اغلب می‌توان با پیشبینی، جلوی ازدست رفتن ارقام بامعنی راگرفت. برای مثال، محاسبهٔ مقدارتابع

$$f(x) = 1 - \cos x$$

را درنظرمی‌گیریم. عملیات را باحساب با شش‌رقم اعشاری انجام می‌دهیم. ازآنجاکه برای xهای‌نزدیك‌به‌صفرداریم $\cos x \approx 1$، لذا اگر $f(x)$ را بدین‌طریق محاسبه‌کنیم که اول مقدار $\cos x$ را پیدا و سپس این مقدار را ازیك‌کم‌کنیم، احتمال از دست دادن‌ارقام با معنی بسیار خواهد بود. ازآنجا که نمی‌توانیم $\cos x$ را تا بیش از شش رقم محاسبه کنیم، پس خطای محاسبه برای xهای نزدیك به‌صفرممکن است به اندازهٔ 5×10^{-7} و بنابراین به‌اندازه $f(x)$ ویا بزرگتر ازآن باشد. اگر بخواهیم مقدار تابع $f(x)$ را در نزدیکی $x = 0$ تاشش‌رقم بامعنی درحساب شش‌رقمی محاسبه‌کنیم، باید ازفرمول دیگری مانند

$$f(x) = 1 - \cos x = \frac{1 - \cos^2 x}{1 + \cos x} = \frac{\sin^2 x}{1 + \cos x}$$

استفاده‌کنیم، که با این فرمول می‌توان $f(x)$ را به‌طور کاملا دقیق به‌ازای مقادیرکوچك x محاسبه‌کرد؛ ویا ازبسط $f(x)$ به‌سری تیلر (به‌بخش ۷.۱ نگاه‌کنید)

$$f(x) = \frac{x^2}{2} - \frac{x^4}{24} + \cdots$$

استفاده‌کنیم. رابطهٔ بالانشان می‌دهدکه برای مثال، به ازای $|x| \leqslant 10^{-3}$، $x^2/2$ حداقل تاشش رقم با معنی $f(x)$ تطبیق می‌کند.

تعیین ریشه‌های معادلهٔ درجهٔ دوم

$$ax^2+bx+c=0 \qquad (16.1)$$

مثال دیگری برای این مطلب است. دردرس جبر دیده‌ایم که جواب معادلهٔ فوق از دستور درجهٔ دوم

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a} \qquad (17.1)$$

به‌دست می‌آید. فرض کنید $b^2-4ac>0$ و $b>0$ باشد و بخواهیم با به‌کاربستن فرمول (۱۷.۱)، ریشه‌ای را که قدر مطلقش کمتر است یعنی

$$x_1=\frac{-b+\sqrt{b^2-4ac}}{2a} \qquad (18.1)$$

را پیدا کنیم. اگر $4ac$ درمقایسه با b^2 کوچك باشد، آنگاه $\sqrt{b^2-4ac}$ در تعدادزیادی از ارقام با b یکی خواهد بود. بنابراین با فرض آنکه $\sqrt{b^2-4ac}$ تنها با همان تعداد ارقام صحیحی که درمحاسبات به‌کار می‌رود محاسبه شود، تعداد ارقام صحیح صورت کسر (۱۸.۱) و در نتیجه تعداد ارقام صحیح ریشهٔ محاسبه شده نیز کمتر از تعداد ارقام به‌کاربرده شده درمحاسبه خواهد بود. برای اینکه مورد خاصی بررسی شود، معادلهٔ

$$x^2+111.11x+1.2121=0 \qquad (19.1)$$

را درنظر می‌گیریم. با به‌کار گرفتن معادلهٔ (۱۸.۱) وحساب قطع شده با ممیزشناور با پنج رقم‌دهدهی داریم:

$$b^2=12345$$

$$b^2-4ac=12340$$

$$\sqrt{b^2-4ac}=111.09$$

$$x_1=\frac{-b+\sqrt{b^2-4ac}}{2a}=-0.01000$$

درحالی که ریشهٔ دقیق معادله تا ۵ رقم دهدهی برابر با $x_1=-0.010910$ است. در اینجا نیز ازدست دادن ارقام بامعنی قابل اجتناب است و این‌کار با به‌کار گرفتن فرمول دیگری برای محاسبهٔ ریشه‌ای که قدرمطلقش کوچکتر است، یعنی

$$x_1=\frac{-2c}{b+\sqrt{b^2-4ac}} \qquad (20.1)$$

انجام می‌پذیرد. بااستفاده ازاین فرمول ومحاسبه تاپنج رقم دهدهی، ریشهٔ x را تا پنج‌رقم صحیح به‌دست می‌آوریم

$$x_1 = -0.510910$$

هنگامی‌که خطایی صورت گرفته باشد، این خطا نهایتاً به‌نتایج بعدی سرایت‌می‌کند. **انتشارخطا**[1] درمحاسبات بعدی را می‌توان به‌آسانی برحسب دومفهوم مرتبط به‌هم، "شرط" و "ناپایداری" مورد مطالعه قرارداد.

واژهٔ **شرط** برای بیان حساسیت مقدارتابع $f(x)$ در رابطه باتغییرات شناسهٔ x به‌کار برده می‌شود. شرط معمولا به وسیلهٔ حداکثر تغییرنسبی درمقدار تابع $f(x)$ که به‌توسط یك واحد تغییر نسبی درشناسهٔ x صورت گرفته اندازه‌گیری می‌شود. با یك فرمول نسبتاً غیررسمی شرط یك تابع را چنین تعریف می‌کنند.

$$\text{شرط } f \text{ در } x = \max\left\{ \left| \frac{f(x)-f(x^*)}{f(x)} \right| \Big/ \left| \frac{x-x^*}{x} \right| : \quad |x-x^*| \text{ "کوچك"} \right\}$$

$$\approx \left| \frac{f'(x)x}{f(x)} \right| \qquad (۲۱.۱)$$

هرقدرکه این شرط بزرگتر باشد، به‌همان اندازه تابع f، به‌اصطلاح بد شرط‌تر[2] است.در اینجا ازاین حقیقت استفاده شده است که

$$f(x)-f(x^*) \approx f'(x)(x-x^*)$$

یعنی با تبدیل شناسهٔ x به x^*، مقدار تابع تقریباً به‌اندازهٔ $f'(x)(x-x^*)$ تغییر می‌یابد. (به‌بخش ۷.۱ نگاه‌کنید.) اگر، برای مثال، داشته باشیم

$$f(x)=\sqrt{x}$$

در این صورت $f'(x)=\frac{1}{2}\Big/\sqrt{x}$، و بنابراین شرط f تقریباً چنین‌است

$$\left| \frac{f'(x)x}{f(x)} \right| = \frac{\left[\frac{1}{2}\Big/\sqrt{x}\right]x}{\sqrt{x}} = \frac{1}{2}$$

رابطهٔ فوق نشان می‌دهد که جذرگیری یك روند "بهشرط[3]" است زیرا خطای نسبی را واقعاً کاهش می‌دهد. بعکس اگرداشته باشیم:

---

1. error propagation
2. ill-conditioned
3. well-conditioned

$$f(x)=\frac{۱۰}{۱-x^۲}$$

دراین صورت $f'(x)=۲۰x/(۱-x^۲)^۲$ و بنابراین

$$\left|\frac{f'(x)x}{f(x)}\right|=\left|\frac{[۲۰x/(۱-x^۲)^۲]x}{۱۰/(۱-x^۲)}\right|=\frac{۲x^۲}{|۱-x^۲|}$$

هرگاه $|x|$ به ۱ نزدیك شود عدد فوق می‌تواند کاملا بزرگ گردد. بنابراین به ازای x نزدیك به ۱ یا ۱ —، تابع فوق کاملا بدشرط است. این فرمول، خطاهای نسبی حاصل به‌توسط شناسه در $f(x)$ را خیلی بزرگ می‌کند.*

ناپایداری، مفهوم وابسته به‌شرط، مبین حساسیت یك فرایند عددی درمحاسبهٔ $f(x)$ به‌ازای x است، درخطاهای اجتناب‌ناپذیری که هنگام اجرای عمل گردکردن درمحاسبات با دقت محدود مرتکب می‌شویم. تأثیردقیق این خطاها را برمقدارمحاسبه شدهٔ $f(x)$ به دشواری می‌توانیم تعیین کنیم، مگر اینکه محاسبات را عملا برای حساب ویژه با دقت محدود انجام دهیم وسپس نتیجه را باجواب دقیق‌مقایسه‌کنیم. اما این تأثیرها را می‌توانیم تقریباً از بررسی گردکردن خطاها یك به‌یك تخمین بزنیم. بدین‌معنی که هریك از مراحل محاسباتی راکه‌کل فرایند را تشکیل می‌دهند تك‌تك مورد بررسی قراردهیم. فرض کنیدکه n چنین مرحله‌ای وجود داشته باشد. برونداد مرحلهٔ iام را با x_i نشان می‌دهیم ومی‌گیریم $x=x_۰$. سپس این x_iها به عنوان درونداد برای یك یا چند مرحلهٔ بعدی به‌کارگرفته می‌شوند، وبدین‌طریق روی جواب‌نهایی $x_n=f(x)$ اثر می‌گذارند. حال تابعی را که وابستگی جواب نهایی را به‌نتیجهٔ میانی x_i بیان می‌کند با f_i نشان می‌دهیم. به ویژه $f_۰$ همان تابع f خواهد بود. در این حال‌کل فرایند را ناپایدارگوییم هرگاه یك یا چند تا از این توابع f_i بد شرط باشند. دقیقتر بگوییم، فرایند ناپایدار است هرگاه یك یا چند تا ازتوابع f_i شرطی بسیار بزرگتر ازشرط تابع $f=f_۰$ داشته باشند. زیرا این‌شرط f_i است‌که اثر نسبی خطای حاصل از گردکردن اجتناب‌ناپذیر درمرحله iام را برجواب‌نهایی می‌رساند. به‌عنوان یك مثال ساده، تابع

$$f(x)=\sqrt{x+۱}-\sqrt{x}$$

را به ازای xهای «بزرگ» مثلا، $x\approx ۱۰^۴$ بررسی می‌کنیم. شرط این تابع به‌صورت زیراست

$$\left|\frac{f'(x)x}{f(x)}\right|=\frac{۱}{۲}\frac{|۱/\sqrt{x+۱}-۱/\sqrt{x}|x}{\sqrt{x+۱}-\sqrt{x}}=\frac{۱}{۲}\frac{x}{\sqrt{x+۱}\sqrt{x}}\approx\frac{۱}{۲}$$

* یعنی هر خطای نسبی در شناسهٔ x موجب ایجاد خطایی به مراتب بزرگتر در مقدار $f(x)$ می‌شود. —م.

که شرط کاملا خوبی است. اما اگر $f(۱۲۳۴۵)$ را با حساب شش رقم اعشاری محاسبه کنیم، مشاهده خواهیم کرد که

$$f(۱۲۳۴۵)=\sqrt{۱۲۳۴۶}-\sqrt{۱۲۳۴۵}$$
$$=۱۱۱٫۱۱۳-۱۱۱٫۱۰۸=۰٫۰۰۵$$

درحالی‌که مقداردقیق $f(۱۲۳۴۵)$ عبارت است از

$$f(۱۲۳۴۵)=۰٫۰۰۴۵۰۰۰۳۲۶۲۶۲۷۷۵۱\ldots$$

بنابراین جواب محاسبه شده دارای ۱۰ درصد خطاست. اکنون روند محاسباتی را مورد تحلیل قرارمی‌دهیم. این روند شامل چهارمرحلهٔ محاسباتی به‌شرح زیراست

$$\begin{aligned} x_۰ &:= ۱۲۳۴۵ \\ x_۱ &:= x_۰+۱ \\ x_۲ &:= \sqrt{x_۱} \\ x_۳ &:= \sqrt{x_۰} \\ x_۴ &:= x_۲-x_۳ \end{aligned} \qquad (۲۲.۱)$$

اکنون برای مثال تابع $f_۳$، یعنی تابعی راکه چگونگی وابستگی جواب نهایی $x_۳$ به $x_۴$ رابیان می‌کند درنظرمی‌گیریم. داریم

$$f_۳(t)=x_۲-t$$

بنابراین شرط تابع $f_۳$ تقریباً چنین است

$$\left|\frac{f_۳'(t)t}{f_۳(t)}\right|=\left|\frac{t}{x_۲-t}\right|$$

این عدد معمولا نزدیک به ۱ است، یعنی $f_۳$ معمولا خوش شرط است، مگراینکه t نزدیک به $x_۲$ باشد. درحالت اخیر $f_۳$ می‌تواند کاملا بد شرط باشد. مثلا در مورد خاص مسئلهٔ ما $t\approx ۱۱۱٫۱۱$، وحال آنکه $x_۲-t\approx ۰٫۰۰۵$، در نتیجه شرط تابع $f_۳$ تقریباً برابر با ۲۲۲۲۲ یا ۴۰۰۰۰ مرتبه بزرگتر ازشرط خود f است.

ازاینجا نتیجه می‌گیریم که روند مذکور در (۲۲.۱) راه ناپایداری برای محاسبهٔ f است. البته اگر ابتدای این قسمت را با دقت مطالعه کرده باشید می‌دانیدکه یك راه پایدار برای محاسبهٔ این تابع همان استفاده ازفرمول هم ارزآن، یعنی فرمول زیراست

$$f(x)=\frac{۱}{\sqrt{x+۱}+\sqrt{x}}$$

درحساب باشش رقم اعشاری داریم

$$f(12345)=\frac{1}{\sqrt{12346}+\sqrt{12345}}=\frac{1}{222{,}221}=0{,}00450002$$

که فقط دارای خطایی برابر ۰٫۰۰۰۰۳ درصد است. روند محاسباتی به‌شرح زیراست

$$\begin{aligned} x_0 &:= 12345 \\ x_1 &:= x_0+1 \\ x_2 &:= \sqrt{x_1} \\ x_3 &:= \sqrt{x_0} \\ x_4 &:= x_2+x_3 \\ x_5 &:= 1/x_4 \end{aligned} \qquad (23.01)$$

برای‌مثال، دراینجا $f_3(t)=1/(x_2+t)$، وشرط این‌تابع تقریباً به‌ازای $t=x_2$ عبارت است از

$$\left|\frac{f_3'(t)t}{f_3(t)}\right|=\left|\frac{t}{x_2+t}\right|\approx\frac{1}{2}$$

بنابراین شرط f_3 کاملا خوب، و به‌خوبی شرط خود f است.

مثالهایی از ناپایداریهای بزرگ را، بخصوص در مبحث حل معادلات دیفرانسیل، مشاهده خواهیدکرد.

تمرین

۱-۴۰۱ با استفاده ازفرمول (۱۸.۱) و(۲۰.۱) کوچکترین ریشهٔ معادلهٔ

$$x^2+0{,}4002\times10^{0}x+0{,}8\times10^{-4}=0$$

را به‌دست آورید. محاسبات را با ممیز شناور ومانتیس چهاررقمی اعشاری انجام دهید.

۲-۴۰۱ درصورتی‌که خطای مطلق x برابر با 10^{-6} باشد، خطای ناشی از محاسبهٔ $f(x)=(\cos x)\exp(10x^2)$ را درحول نقطهٔ $x=2$ برآوردکنید.

۳-۴۰۱ برای محاسبهٔ عبارتهای

$$f(x)=\frac{x-\sin x}{\tan x} \qquad \text{(الف)}$$

$$f(x)=(\alpha+x)^n-\alpha^n \qquad \text{(ب)}$$

$$f(x)=\sin(\alpha+x)-\sin\alpha \qquad \text{(پ)}$$

$$f(x)=x-\sqrt{x^2-\alpha} \qquad \text{(ت)}$$

راهی پیدا کنید که $f(x)$ را درست باهمان تعداد ارقام x به‌دست دهد، وقتی که مقدار x در (الف) تا (پ) نزدیك به‌صفر و در(ت) به‌مراتب بزرگتر از α باشد.

۴-۴۰۱ فرض کنید کامپیوتری با مانتیس چهاررقمی اعشاری در دست باشد. اعداد زیر را اول به‌ترتیب صعودی (ازکوچکتر تا بزرگتر) و بعد به‌ترتیب نزولی با هم جمع کنید. در حین عمل حاصلجمعهای جزئی را گرد کنید. نتایج خود را با جواب درست $x=۰٫۱۰۷۱۰۱۰۲۳\times10^5$، مقایسه نمایید.

$۰٫۸۹۹۹\times10^4$	$۰٫۶۲۶۶\times10^2$	$۰٫۱۵۸۰\times10^0$
	$۰٫۷۵۵۵\times10^2$	$۰٫۲۶۵۳\times10^0$
	$۰٫۷۸۸۹\times10^3$	$۰٫۲۵۸۱\times10^1$
	$۰٫۷۷۶۷\times10^3$	$۰٫۴۲۸۸\times10^1$

۵-۴۰۱ یك روش آشکارا ناپایدار برای محاسبهٔ $f(x)=e^x$، به ازای x منفی، بسط آن به‌سری تیلر (۳۶۰۱) است. مقدار e^{-12} را از راه بسط به‌سری تیلر (۳۶۰۱) و به‌ازای $x=-12$ محاسبه کنید ومقدارمحاسبه شده را با مقداردقیق

$$e^{-12}=۰٫۰۰۰۰۰\ ۶۱۴۴۲\ ۱۲۳۵۴\ \ldots$$

مقایسه کنید. [راهنمایی: به‌موجب(۳۶۰۱)اختلاف بین e^x وحاصلجمع جزئی $s_n=\sum_0^n x^j/j!$ وقتی x منفی باشد، ازقدرمطلق جملهٔ بعدی یعنی $|x^{n+1}/(n+1)!|$ کوچکتراست. بنابراین بجا خواهد بودکه سری را تا $s_n=s_{n+1}$ جمع کنیم.]

۶-۴۰۱ نتیجهٔ تمرین ۴۰۱-۵ را ازراه مقایسهٔ شرط $f(x)=e^x$ درمجاورت $x=-12$ باشرط برخی ازتوابع f_i موجود در روند محاسباتی فوق، بیان کنید. سپس یك روش پایدار برای محاسبهٔ e^{-12} از راه سری‌تیلر(۳۶۰۱)پیدا کنید. (راهنمایی: $e^{-x}=1/e^x$.)

۵۰۱ روشهای کامپیوتری برای براورد خطا

هدف ما در این فصل آگاه ساختن دانشجویان به منشأ ممکن خطا و ارائهٔ تکنیکهایی است

که می‌توان برای اجتناب از آنها به‌کار گرفت. در ارزیابی نتایج حاصله از کامپیوتر این خطاها را باید منظور کرد. برآورد واقعی خطای کلی در یك مسئلهٔ عملی کاری است دشوار و هنوز یك نظریهٔ ریاضی مناسب برای آن وجود ندارد. یك فکر جالب استفاده از خود کامپیوتر برای تهیهٔ این گونه برآوردهاست. روشهای گوناگونی ازاین نوع پیشنهادشده‌اند که پنج‌تای آنها را مختصراً شرح می‌دهیم. ساده‌ترین روش استفاده از **دقت مضاعف** است. در این روش یك مسئله اساساً دوبارحل می‌شود ــ یکبار با دقت معمولی و یکبار با دقت مضاعف. از تفاضل این دونتیجه می‌توان کل خطای ناشی از گردکردن‌را (بافرض اینکه همهٔ خطاهای‌دیگر ناچیز باشند) تخمین‌زد. دراین‌صورت می‌توان فرض کردکه درمسائل‌دیگری که با همین زیربرنامه حل می‌شوند همان انباشتگی خطای ناشی از گردکردن پدید خواهد آمد. این روش در رابطه باوقت مصرفی‌کامپیوتر فوق‌العاده پرهزینه است زیراکه دربرخی ازکامپیوترهاحساب با دقت مضاعف وقت مورد نیاز را هشت برابرافزایش می‌دهد وبعلاوه جداکردن خطاهای دیگرهمواره ممکن نیست.

یك روش‌دیگر **حساب بازه‌ای**[1] است. دراینجا هرعدد به‌زبان ماشین به‌وسیلهٔ دوعدد بیان می‌شود که ماکسیمم و مینیمم مقادیری هستندکه آن عدد ممکن است اختیارکند. هر وقت عملی انجام گیرد، این مقادیر ماکسیمم و مینیمم محاسبه می‌شوند. بنابراین اصلا در هرمرحله دوجواب به‌دست می‌آیدکه جواب واقعی لزوماً درمحدودهٔ این دومقدارماکسیمم ومینیمم جواب قراردارد. این‌روش مستلزم بیش از دوبرابر وقت‌کامپیوتر وتقریباً دوبرابر ذخیره‌سازی استانده است. بعلاوه این فرض متداول، که جواب واقعی دروسط محدودهٔ ما قراردارد، درحالت کلی صادق نیست. ازاین‌رو ممکن است این محدوده به‌اندازه‌ای وسیع باشدکه هربرآوردی ازخطای گردکردن برپایهٔ این‌روش بسیاراغراق‌آمیز باشد.

نگرش‌سوم حساب **رقم ــ بامعنی** است. همان‌طوری‌که قبلا اشاره‌شد هرگاه دوعدد تقریباً نزدیك به‌هم درزبان ماشین، ازهم تفریق شوند این خطروجود دارد کــه برخی از ارقام بامعنی ازدست بروند. درحساب رقم‌ـبامعنی سعی براین است‌که رد اعداد بامعنی‌که بدین‌طریق ازدست می‌روند حفظ شود. دریکی ازاین شیوه‌ها فقط ارقام بــا معنی هرعدد نگهداشته و کلیهٔ ارقــام دیگر حذف می‌شود. بدین‌ترتیب در انتهای محاسبه مطمئنیم که تمام ارقام حفظ شده بامعنی هستند. اشکال اصلی این‌روش این است‌که برخی اطلاعات با حذف ارقام از بین می‌روند و بانتیجهٔ به‌دست‌آمده احتمالا باید بسیارمحتاطانه عمل‌کرد. گرچه تجربیات به‌دست آمده بااین تکنیك نویدبخش نبوده‌است، اماآزمایش بااین‌تکنیك هنوزادامه دارد.

چهارمین روش که نویدبخش فراهم‌کردن یك نظریهٔ ریــاضی مناسب برای انتشار خطای گردکردن است، بر **نگرش آماری** استوارشده‌است. این نگرش بافرض مستقل بودن خطاهای گردکردن آغازمی‌شود. البته این‌فرض معتبری نیست، زیرا اگرمسئله‌ای چندین بار

---

1. interval arithmetic

روی یك ماشین حل شود جوابها همواره یکسان خواهند بود. اما می‌توانیم یك الگوی تصادفی[1] ازانتشار خطای گردکردن، که در آن به خطاهای موضعی به‌صورت متغیرهای تصادفی نگریسته می‌شود، درنظر بگیریم. بنابراین می‌توانیم فرض کنیم که مقادیر خطاهای موضعی به‌طور یکنواخت و یا به‌طور نرمال بین مقادیر فرینه[2] توزیع شده‌اند. با استفاده از روشهای آماری می‌توان انحراف معیار[3] واریانس[4] توزیع را به‌دست آورد ودرنتیجه خطای گردکردن انباشته را برآورد کرد. این نگرش آماری به‌طور نسبتاً مشروح به توسط همینگ[5] [۱] و هنریتسی[6] [۲] بررسی شده‌است. این روش مستلزم تحلیل اساسی و وقت اضافی برای کامپیوتر است، اما درآزمایشهایی که تااین تاریخ انجام شده نتایجی از برآورد خطاها به‌دست آمده است که تا حد زیادی با مدارك تجربی موجود توافق دارد.

روش پنجم، روش **تحلیل خطای پسرو** است که در بخش ۳.۱ معرفی شد. همان‌طوری که مشاهده کردیم، این‌روش تحلیل اثرات خطای گردکردن را به‌بررسی آشفتگیها درحساب دقیق و درنهایت به‌یك مسئلهٔ مربوط به‌شرط، بدل می‌کند. درفصل ۴ این روش به‌نحو مؤثری مورد استفاده قرارخواهد گرفت.

۶.۱ توضیحاتی درمورد همگرایی دنباله‌ها

حساب دیفرانسیل وانتگرال و به‌طورکلی آنالیز، برپایهٔ مفهوم همگرایی استوار است. مفاهیم پایه‌ای‌مانند مشتق وانتگرال وپیوستگی، برحسب دنباله‌های همگرا وتوابع مقدماتی مانند $\ln x$ یا $\sin x$ تعریف می‌شوند. ازطرفی برای مسائل‌مهندسی وعلمی هرگزجوابهای عددی کاملا دقیق مورد نیاز نیستند، بلکه جوابهایی تقریبی، لازم‌اندکه «تاعدهٔ معینی ارقام اعشاری» یا درحد قابل تحملی‌مانند ε دقیق باشند.

ازاین‌رو شگفت‌آور نیست اگر درپیداکردن α، جواب یك مسئلهٔ مفروض، بسیاری از روشهای عددی، تنها (چند جملهٔ اول) α_1 و α_2 و α_3 و... از یك دنباله راکه همگرا بودنش به‌سمت جواب مطلوب نشان داده شده‌است، به‌دست دهند. یادآوری می‌کنیم که طبق تعریف:

> یك دنبالهٔ اعداد (حقیقی یا مختلط) $\alpha_1, \alpha_2, \alpha_3, \ldots$ به سمت α همگراست اگروفقط اگر، به‌ازای جمیع مقادیر $\varepsilon > 0$ عدد صحیحی مانند $n_0(\varepsilon)$ وجود داشته باشد به‌طوری‌که به‌ازای هر $n \geqslant n_0$، نامساوی $|\alpha - \alpha_n| < \varepsilon$ برقرار باشد.

بنابراین اگر یك‌روش عددی‌داشته باشیم که دنبالهٔ $\alpha_1, \alpha_2, \alpha_3, \ldots$ را تولیدکند و این‌دنباله به‌سمت جواب مطلوب α همگرا باشد، آنگاه می‌توان α را تا هردقت‌دلخواهی فقط با محاسبهٔ α_n به‌ازای مقدار "به‌اندازهٔ کافی بزرگ" n محاسبه‌کرد.

---

1. stochastic
2. extreme
3. standard deviation
4. variance
5. Hamming
6. Henrici

ازدیدگاه محاسباتی این تعریف به‌دلایل زیر رضایتبخش نیست:

(۱) غالباً (بدون دانستن جواب α) امکان ندارد که دریابیم چه موقع n "به اندازهٔ کافی بزرگ" است. به عبارت دیگر پیداکردن تابع $n_0(\varepsilon)$ که در تعریف همگرایی ذکر شد، بسیار مشکل است.

(۲) حتی وقتی‌هم اطلاعاتی دربارهٔ $n_0(\varepsilon)$ در دست باشد، ممکن است n موردنیاز به‌اندازه‌ای بزرگ شود که محاسبهٔ α_n را غیرممکن سازد.

□ **مثال:** عدد $\pi/4$ برابر بامقدار رشتهٔ نامتناهی[1] زیراست

$$\sum_{i=0}^{\infty}\frac{(-1)^i}{2i+1}=1-\sum_{j=1}^{\infty}\frac{2}{16j^2-1}$$

ازاین‌رو با

$$\alpha_n=1-\sum_{j=1}^{n}\frac{2}{16j^2-1}\qquad n=1,2,\ldots$$

دنبالهٔ $\alpha_1, \alpha_2, \ldots$ درحالی‌که یکنواکم می‌شود به‌حد خود یعنی $\pi/4$ میل می‌کند. بعلاوه داریم

$$0\leqslant\alpha_n-\frac{\pi}{4}\leqslant\frac{1}{4n+3}\qquad n=1,2,\ldots$$

برای اینکه ازدنبالهٔ بالا برای محاسبهٔ $\pi/4$ تا 10^{-6} رقم صحیح استفاده‌کنیم، لازم است داشته باشیم $10^6\leqslant 4n+3$ یا تقریباً $n=250000$. اما درکامپیوتری‌که با ممیزشناور هشت رقم اعشاری کار می‌کند خطای گردکردن در محاسبهٔ α_{250000} احتمالا خیلی بزرگتر از 10^{-6} خواهدشد. بنابراین $\pi/4$ را نمی‌توان با استفاده از دنبالهٔ فوق تا 10^{-6} رقم صحیح محاسبه نمود (مگر، احتمالا با اضافه‌کردن جملات به یکدیگر ازکوچکترین تا بزرگترین). □

برای پرداختن به‌این گونه مسائل، برخی قراردادهای نمادگذاری مفیدند. بخصوص آنکه می‌خواهیم سرعت همگرایی دنباله‌هارا اندازه‌گیری کنیم. مانند تمام اندازه‌گیریهای دیگر، این عمل را نیز به‌وسیلهٔ مقایسه بادنباله‌های استانده[2]، مانند دنباله‌های

$$\left.\begin{array}{l}1/n\\1/n^2\\r^n\\1/(\ln n)\end{array}\right\}\qquad n=1,2,3,\ldots$$

---

1. infinite series 2. standard

انجام می‌دهیم. مقایسه به‌گونهٔ زیر انجام می‌گیرد: گوییم α_n از مرتبهٔ β_n (یا α_n مهصفر[1] β_n) است و می‌نویسیم

$$\alpha_n = \mathcal{O}(\beta_n) \qquad (۲۴۰۱)$$

وقتی به‌ازای مقدار ثابتی مانند K وجمیع مقادیر به‌اندازهٔ‌کافی بزرگ n داشته باشیم

$$|\alpha_n| \leqslant K|\beta_n| \qquad (۲۵۰۱)$$

بنابراین

$$\left.\begin{array}{l} ۱/n \\ ۱۰,۰۰۰/n \\ ۱۰/n - ۴۰/n^۲ + e^{-n} \\ ۱/n^۲ \end{array}\right\} = \mathcal{O}(۱/n)$$

وانگهی هرگاه به‌محض اینکه n به‌قدرکافی بزرگ شد، بتوان ثابت K را در (۲۵۰۱) به دلخواه کوچك انتخاب کرد، یعنی اگرحالت

$$\lim_{n\to\infty} \alpha_n/\beta_n = ۰$$

پیش آید، آنگاه گفته می‌شود که مرتبهٔ α_n ازمرتبهٔ β_n بیشتر (یـا α_n، کهصفر[2] β_n) است و می‌نویسیم

$$\alpha_n = o(\beta_n) \qquad (۲۶۰۱)$$

بنابراین وقتی $\sin(۱/n) \neq o(۱/n)$، داریم

$$\left.\begin{array}{l} ۱/n^۲ \\ ۱/(n \ln n) \end{array}\right\} = o(۱/n)$$

معمولا علامت مرتبه فقط در سمت راست یك معادله ظاهر می‌شود وهدف آن‌تشریح جنبهٔ اصلی یك جملهٔ خطا، بدون توجه به‌ضرب ثابتها یا سایرجزئیات است. بـرای مثال، می‌توانیم حالت نامطلوب موضوع را در مثال قبلی، احتمالا به‌وسیلهٔ روابط زیربیان‌کنیم

$$۱ - \sum_{j=۲}^{n} ۱/(j^۲ - ۱) = \pi/۴ + \mathcal{O}(۱/n)$$

اما همچنین

---

1. α_n is big-oh of β_n 2. α_n is little-oh of β_n

$$۱-\sum_{j=۲}^{n} ۱/(j^۲-۱)\neq \pi/۴+o(۱/n)$$

یعنی، این‌دنباله با همان‌سرعت $۱/n$ (سرعتی که $۱/n$ به‌سمت صفرمیل می‌کند) و نه‌سریعتر، به سمت $\pi/۴$ همگرا می‌شود. مرتبه یا میزان همگرایی $۱/n$ به‌اندازه‌ای کم است که در محاسبات نمی‌تواند مورد استفاده قرار گیرد.

□ **مثال:** اگر $\alpha_n=\alpha+o(۱)$، آنگاه بنابرتعریف داریم

$$\lim_{n\to\infty}\frac{\alpha_n-\alpha}{۱}=۰$$

بنابراین $\alpha_n=\alpha+o(۱)$ فقط نوعی سلیقه برای بیان همگرایی دنبالهٔ $\alpha_۱, \alpha_۲, \ldots$ به سمت α است. □

□ **مثال:** اگر $|r|<۱$، آنگاه مجموع سری هندسی $\sum_{i=۰}^{\infty} r^i$ برابر است با $۱/(۱-r)$. با $s_n=\sum_{i=۰}^{n} r^i$ داریم

$$s_n=(۱-r^{n+۱})/(۱-r)=\frac{۱}{(۱-r)}-\frac{r^{n+۱}}{(۱-r)}$$

بنابراین

$$s_n=\frac{۱}{۱-r}+\mathcal{O}(r^n)$$

بعلاوه اگر $|r|<\bar{r}$؛ آنگاه

$$s_n=\frac{۱}{۱-r}+o(\bar{r}^n)$$

بنابراین هرگاه تساوی $\alpha_n=\alpha+\mathcal{O}(r^n)$ به ازای یك $|r|<۱$ برقرار باشد، گوییم همگرایی (حداقل) هندسی است، زیرا دراین‌صورت مرتبهٔ همگرایی (حداقل) مساوی مرتبهٔ همگرایی سری هندسی خواهد بود. □

با وجودی که دانستن رابطهٔ $\alpha_n=\alpha+\mathcal{O}(\beta_n)$ از ندانستن آن بهتر است، آگاهی ازمرتبهٔ همگرایی‌زمانی‌کاملا مفید است که به‌طور دقیقتری رابطهٔ $\alpha_n=\alpha+\beta_n+o(\beta_n)$ را بشناسیم. رابطهٔ بالا بدین‌معنی است که به‌ازای مقدار «به‌اندازهٔ‌کافی بزرگ» n، $\alpha_n\approx\alpha+\beta_n$. به‌عبارت دیگر

$$\alpha_n=\alpha+\beta_n+o(\beta_n)$$
$$=\alpha+\beta_n+\beta_n o(1)$$
$$=\alpha+\beta_n(1+\varepsilon_n)$$

در رابطهٔ بالا ε_1، ε_2، ... دنباله‌ای است که به‌صفر همگراست. گرچه «به‌اندازهٔ کافی بزرگ» بودن مقداری از n را نمی‌توانیم ثابت کنیم، اما می‌توانیم از مقایسهٔ $\alpha_{k+1}-\alpha_k$ با $\beta_{k+1}-\beta_k$، این فرض را که n «به‌اندازهٔ کافی بزرگ است» آزمایش کنیم. اگر برای k نزدیک به n مثلا $k=n-2, n-1, n$ داشته باشیم

$$\frac{|\alpha_{k+1}-\alpha_k|}{|\beta_{k+1}-\beta_k|}\approx 1$$

آنگاه می‌توانیم درستی این فرض را، که n «به‌قدر کافی بزرگ» است، برای

$$\alpha_n\approx\alpha+\beta_n$$

بپذیریم و بنابراین $|\beta_n|$ را به‌عنوان یك برآورد مطلوب ازخطای $|\alpha-\alpha_n|$ قبول کنیم.

□ **مثال:** گیریم $p>1$. پس سری $\sum_j 1/(p^j+1)$ مانند سری هندسی $\sum_j 1/p^j$ به‌سمت حد خود α همگراست، یعنی

$$\alpha_n=\sum_{j=1}^{n} 1/(p^j+1)=\alpha+\theta(1/p^{n+1})$$

برای اینکه حکم دقیقتری به‌دست آوریم، معادلهٔ زیر را درنظر می‌گیریم

$$\beta_n=\sum_{j=n+1}^{\infty} 1/p^j=(1/p^{n+1})/(1-1/p)=1/[p^n(p-1)]$$

پس

$$\alpha_n=\alpha-\sum_{j=n+1}^{\infty} 1/(p^j+1)=\alpha-\beta_n+\sum_{j=n+1}^{\infty}[1/p^j-1/(p^j+1)]$$
$$=\alpha-\beta_n+o(\beta_n)$$

زیرا که

$$0\leqslant\sum_{j=n+1}^{\infty}[1/p^j-1/(p^j+1)]=\sum_{j=n+1}^{\infty}1/(p^j(p^j+1))\leqslant\sum_{j=n+1}^{\infty}(1/p^2)^j$$
$$=\theta((1/p^2)^{n+1})$$

برای نسبت فوق داریم:

$$\left|\frac{\alpha_{n+1}-\alpha_n}{\beta_{n+1}-\beta_n}\right| = p^{n+1}/(p^{n+1}+1)$$

که مثلا برای $n=3$ و $p=2$ درحدود ۱/۱۰ از ۱ است. بنابراین ۰٫۱۲۵ $=\beta_3=1/8$ شاخص‌خوبی‌است از خطا در ۰٫۶۴۴۴۴ ۴۴۴۴۴ $=\alpha_3$. زیرا ۰٫۷۶۴۴۹ ۹۷۸۰ $=\alpha$، که دراین صورت خطا در α_3 برابراست با ۰٫۱۲۰۰۵. □

این نمادگذاری می‌تواند برای توابعی از یك متغیرحقیقی بسط داده شود. اگر

$$\lim_{h\to 0} T(h)=A$$

همگرایی را $\mathcal{O}(f(h))$ گوییم به‌شرطی‌که برای‌مقدارثابت متناهی K وکلیهٔ hهای به‌اندازهٔ کافی‌کوچك، داشته باشیم

$$\frac{|T(h)-A|}{|f(h)|}\leqslant K$$

اگر رابطهٔ فوق به‌ازای جمیع مقادیر $K>0$ صادق باشد، یعنی اگر

$$\lim_{h\to 0}\frac{T(h)-A}{f(h)}=0$$

آنگاه همگرایی را $o(f(h))$ می‌نامیم.

□ **مثال:** به‌ازای h «نزدیك» به‌صفر داریم

$$\frac{\sin h}{h}=1-\left(\frac{1}{3!}\right)h^2+\left(\frac{1}{5!}\right)h^4-\cdots=1+\mathcal{O}(h^2)$$

$$=1-\frac{1}{6}h^2+o(h^2)$$

بنابراین، به‌ازای جمیع مقادیر $\gamma<2$ داریم

$$\frac{\sin h}{h}=1+o(h^\gamma)$$

□

□ **مثال:** اگر تابع $f(x)$ ریشه‌ای از مرتبهٔ γ در $x=\xi$ داشته باشد، آنگاه:

$$f(\xi+h)=\mathcal{O}(h^\gamma) \quad \text{اما} \quad f(\xi+h)\neq o(h^\gamma)$$

□

قواعد محاسبه با نماد[1]های مرتبه‌ای در لم[2] زیرگردآوری شده‌است.

---

1. symbols 2. lemma

لم ۱۰۱ اگر $\alpha_n=\alpha+\Theta(f(n))$ و $\lim_{n\to\infty} f(n)=0$ و c عدد ثابتی باشد، آنگاه:

$$c\alpha_n=c\alpha+\Theta(f(n))$$

همچنین اگر $\beta_n=\beta+\Theta(g(n))$، و $g(n)=\Theta(f(n))$، در این‌صورت داریم:

$$\alpha_n+\beta_n=\alpha+\beta+\Theta(f(n)) \quad \text{و} \quad \alpha_n\beta_n=\alpha\beta+\Theta(f(n)) \qquad (۲۷۰۱)$$

بعلاوه اگر $\beta\neq 0$، همچنین داریم

$$\frac{\alpha_n}{\beta_n}=\frac{\alpha}{\beta}+\Theta(f(n))$$

درصورتی که اگر $\alpha=\beta=0$، آنگاه

$$\alpha_n\beta_n=\Theta(f(n)g(n))$$

بالاخره کلیهٔ احکام فوق با قراردادن o به‌جای Θ درهمه‌جا، صحیح خواهد ماند.

محاسبهٔ تقریبی یک عدد α به‌وسیلهٔ دنبالهٔ $\alpha_1, \alpha_2, \ldots$ که به سمت α همگراست، خواه‌مرتبهٔ‌همگرایی‌معلوم باشدوخواه‌نباشد، همیشه امری‌است‌توأم با اطمینان. اگرهمگرایی این‌دنباله به‌سمت α معلوم باشد، یك تحلیلگرمجرب روشهای عددی، پس‌از اطمینان به‌اینکه به‌ازای مقادیر کوچك i، تفاوت α_{n-i} با α_n «به‌اندازهٔ‌کافی‌کوچك است»، «به‌اندازهٔ کافی بزرگ» بودن عدد n را تحقیق‌خواهدکرد. همچنین‌درصورتی‌که بداند همگرایی به‌صورت $\beta_n+o(\beta_n)$ است، بررسی می‌کندکه‌آیا وضعیت این دنباله در نزدیکی n به‌همان‌ترتیب است یاخیر. همچنین اگربداندکه α در معادلات ویا نامعادلاتی‌صدق می‌کند ــ α می‌تواند جواب یك معادله باشدــ بررسی می‌کندکه‌آیا α_n در آن معادلات ویانامعادلات «به‌اندازهٔ کافی‌خوب» صدق می‌کندیانه. کوتاه سخن آنکه یك تحلیلگرمجرب روشهای عددی، مطمئن می‌شودکه n درتمام شرایط متصورلازم برای «به‌اندازهٔ‌کافی بزرگ» بودن صدق می‌کند، در صورتی‌که تمامی این شرایط برقرار باشند و شرط کافی برای اینکه n «به‌اندازهٔ کافی بزرگ» باشد دردست نباشد، با اطمینان α_n راتقریب «به‌اندازهٔ‌کافی‌خوبی» برای α قبول می‌کند. به‌عبارت دیگر، یك تحلیلگر تمامی روشهایی راکه دراختیار دارد برای تمیز یك تقریب «به‌اندازهٔ‌کافی خوب» از یك تقریب بد به‌کارمی‌گیرد. بیش از این نمی‌تواندکاری انجام دهد (وکمترهم نباید انجام دهد).

از اینجا معلوم می‌شودکه نتایج عددی به‌دست‌آمده از این طریق را نباید باجواب نهایی اشتباه‌گرفت، بلکه چنانچه بررسیهای بعدی صحت آنها را مورد تردید قرار دهند، بایستی با خیال راحت در درستی آنها تأمل‌کرد.

دانشجویان باید این مطلب را به‌عنوان مثال دیگری از تفاوت اساسی بین آنالیز

عددی و آنالیز تلقی کنند. آنالیز زمانی به‌صورت یك نظام دقیق درآمد که محدودیتهای محاسبات عملی را کلا به‌عهدهٔ مسائلی گذاشت که برحسب الگوی مجرد[1]ی ازدستگاه‌عددی، به‌نام اعداد حقیقی، طرح شده بودند. این الگوی مجرد چنان طرح شده‌است که دادن یك تعریف دقیق و مفید ازحد را ممکن می‌سازد و درصورتی که این گونه مسائل به‌مصطلحات چنین الگویی ترجمه شوند، این تعریف راه را برای حل تجزیدی یا نمادی تعدادقابل توجهی از مسائل عملی می‌گشاید. و این امر به‌نوبهٔ خود کار ترجمهٔ حل تجزیدی یا نمادی را به‌حل عملی بازمی‌گرداند. آنالیزعددی این وظیفه را به‌عهده می‌گیرد وهمراه با آن، محدودیتهای موجود در محاسبات عملی راکه آنالیز با ظرافت ازآنها اجتناب می‌ورزد، تقبل می‌کند. از این‌رو آنالیزعددی الزاماً غیردقیق است و جوابهای عددی آن معمولا آزمایشی هستند و در نهایت فقط تا حدود معینی صحیح‌اند.

بنابراین، کار آنالیزعددی فقط ساختن روشهای عددی نیست، بلکه قسمت اعظم آن استخراج کرانهای خطای مفید یا برآوردهای خطا برای جوابهای حاصل‌از الگوریتمهای عددی‌است. درسراسراین کتاب دانشجویان قبل‌ازهرچیز با این کرانهای خطاکه درآنالیز عددی فراوان هستند مواجه خواهند شد.

تمرین

۱-۶۰۱ $\ln ۲$ را می‌توان از سری

$$\ln ۲ = ۱ - \frac{۱}{۲} + \frac{۱}{۳} - \frac{۱}{۴} + \cdots$$

محاسبه کرد. در درس آنالیز دیده‌ایم کـه این‌سری همگرا و اندازهٔ خطا در هرحاصلجمع جزئی آن کمتر از اندازهٔ اولین‌جملهٔ صرف‌نظرشده‌است. تعداد جملات لازم برای محاسبهٔ $\ln ۲$ را تا ده رقم اعشاری برآوردکنید.

۲-۶۰۱ به‌ازای h نزدیك به‌صفر می‌توان نوشت

$$\frac{\tan h}{h} = ۱ + \Theta(h^{۲})$$

و

$$\frac{\tan h}{h} = ۱ + o(h^{۶})$$

مقادیر γ و δ را طوری پیداکنیدکه تساویهای فوق برقرارباشند.

---

1. abstract model

۳-۶.۱ سعی کنید حد دنبالهٔ زیر را به‌وسیلهٔ یک کامپیوتر محاسبه نمایید

$$\alpha_n = (\tan \lambda^{-n} - \sin \lambda^{-n}) \cdot \lambda^{3n} \qquad n = 0, 1, 2, \cdots$$

از لحاظ نظری دنبالهٔ $\alpha = \lim_{n \to \infty} \alpha_n$ و مرتبهٔ همگرایی آن را پیدا کنید.

۷.۱ برخی مقدمات ریاضی

فرض بر این است که دانشجویان با مباحث آشنای دروس حساب دیفرانسیل و انتگرال و هندسهٔ تحلیلی در دورهٔ لیسانس آشنایی دارند. این درس شامل مفاهیم مقدماتی دستگاههای اعداد حقیقی و مختلط، پیوستگی، مفهوم حد، دنباله‌ها، رشته‌ها و مشتقگیری و انتگرالگیری است. برای فصل چهارم آشنایی کمی با دترمینانها مسلم فرض شده‌است. برای فصل هشتم و نهم نیز داشتن آشنایی اندکی با حل معادلات دیفرانسیل معمولی مسلم انگاشته شده‌است، گرچه این فصول ممکن است تدریس نشوند.

به‌ویژه، کراراً از قضیه‌های زیر استفاده خواهیم کرد.

قضیهٔ ۱.۱: قضیهٔ مقدار میانه برای توابع پیوسته. فرض می‌کنیم تابع $f(x)$ در بازهٔ $[a, b]$ پیوسته باشد. اگر به‌ازای اعداد α و $\bar{x} \in [a, b]$ و $\underline{x}$ داشته باشیم $f(\underline{x}) \leqslant \alpha \leqslant f(\bar{x})$، آنگاه حداقل یک عدد $\xi \in [a, b]$ وجود دارد به‌طوری که $\alpha = f(\xi)$. این قضیه غالباً به‌شکل زیر به‌کار برده می‌شود:

قضیهٔ ۲.۱: گیریم $f(x)$ تابعی پیوسته در بازهٔ $[a, b]$، و x_1، ...، x_n نقاطی بر $[a, b]$ باشند. همچنین فرض می‌کنیم g_1، ...، g_n اعداد حقیقی همعلامت باشند، در این صورت عددی مانند $\xi \in [a, b]$ وجود دارد به‌طوری که

$$\sum_{i=1}^{n} f(x_i) g_i = f(\xi) \sum_{i=1}^{n} g_i$$

به‌منظور اشاره به‌اثبات قضیه، بی‌آنکه خللی در کلیت حاصل آید، فرض می‌کنیم $g_i \geqslant 0$، $i = 1, \ldots, n$. اگر $f(\underline{x}) = \min_i f(x_i)$ و $f(\bar{x}) = \max_i f(x_i)$، آنگاه

$$f(\underline{x}) \sum_{i=1}^{n} g_i = \sum_{i=1}^{n} f(\underline{x}) g_i \leqslant \sum_{i=1}^{n} f(x_i) g_i \leqslant \sum_{i=1}^{n} f(\bar{x}) g_i = f(\bar{x}) \sum_{i=1}^{n} g_i$$

بنابراین $\alpha = \sum_i f(x_i) g_i$ عددی است بین دو مقدار $f(\underline{x}) \sum_i g_i$ و $f(\bar{x}) \sum_i g_i$ از تابع پیوستهٔ $f(x) \sum_i g_i$، با توجه به‌قضیهٔ ۱.۱ حکم قضیهٔ ۲.۱ نتیجه می‌شود.

به‌نحو مشابه می‌توان حکم متناظری را برای حاصلجمعهای نامتناهی[1] یا انتگرالها ثابت کرد.

---

1. Infinite

قضیهٔ ۳.۱: قضیهٔ مقدار میانگین برای انتگرالها. گیریم $g(x)$ تابعی نامنفی یا نامثبت و انتگرالپذیر در $[a, b]$ باشد. هرگاه $f(x)$ روی $[a, b]$ پیوسته باشد، به‌ازای مقداری مانند $\xi \in [a, b]$ داریم

$$\int_a^b f(x)g(x)\,dx = f(\xi)\int_a^b g(x)\,dx \qquad (۲۸.۱)$$

توجه: چنانچه از مثال سادهٔ

$$f(x)=g(x)=x,\ [a, b]=[-۱, ۱]$$

برمی‌آید، فرض یك علامت داشتن $g(x)$ درقضیهٔ ۳.۱، فرضی‌است اساسی.

قضیهٔ ۴.۱: گیریم $f(x)$ تابعی پیوسته دربازهٔ بسته وکراندار $[a, b]$ باشد. پس $f(x)$ دارای «مقادیرماکسیمم ومینیممی‌در روی $[a, b]$ است»، یعنی نقاط $\underline{x}$ و $\bar{x}$ دربازهٔ $[a, b]$ طوری وجود دارندکه به‌ازای هر $x \in [a, b]$ نامساویهای

$$f(\underline{x}) \leqslant f(x) \leqslant f(\bar{x})$$

برقرارند.

قضیهٔ ۵.۱: قضیهٔ رول[1] فرض می‌کنیم $f(x)$ دربازهٔ (بسته و متناهی[2]) $[a, b]$ پیوسته ودربازهٔ (a, b) مشتقپذیر باشد. اگر $f(a)=f(b)=۰$، آنگاه مقداری مانند $\xi \in [a, b]$ وجود داردکه $f'(\xi)=۰$.

در اثبات این قضیه عمدتاً ازقضیهٔ ۴.۱ استفاده می‌شود. زیرا طبق قضیهٔ ۴.۱ نقاطی مانند $\underline{x}$ و $\bar{x}$ در $[a, b]$ وجود دارند به‌طوری‌که به‌ازای جمیع مقادیر $x \in [a, b]$ نامساویهای $f(\underline{x}) \leqslant f(x) \leqslant f(\bar{x})$ برقرارند. اما اگرنه $\underline{x}$ و نه $\bar{x}$ هیچکدام در (a, b) نباشند، آنگاه $f(x) \equiv ۰$، و هر $\xi \in (a, b)$ دراین نامساویها صدق می‌کند. درغیراین‌صورت یا $\underline{x}$ و یا $\bar{x}$ در (a, b) واقع‌است، مثلا $\bar{x} \in (a, b)$. دراین‌صورت $f'(\bar{x})=۰$، زیرا

$$۰ \leqslant \lim_{h \to ۰-} \frac{f(\bar{x})-f(\bar{x}+h)}{-h} = f'(\bar{x}) = \lim_{h \to ۰+} \frac{f(\bar{x}+h)-f(\bar{x})}{h} \leqslant ۰$$

$f(\bar{x})$ بزرگترین مقداری است که $f(x)$ روی $[a, b]$ اختیار کرده‌است.

قضیهٔ ۶.۱: قضیهٔ مقدار میانگین برای مشتقها. اگر $f(x)$ دربازهٔ (بسته ومتناهی) $[a, b]$ پیوسته و در بازهٔ (a, b) مشتقپذیر باشد، آنگاه مقداری مانند $\xi \in (a, b)$ وجود دارد به‌طوری‌که

---

1. Rolle's theorem 2. Finite

$$\frac{f(b)-f(a)}{b-a}=f'(\xi) \qquad (۲۹.۱)$$

اگر درقضیهٔ ۵.۱ به‌جای تابع $f(x)$ تابع

$$F(x)=f(x)-f(a)-\frac{f(b)-f(a)}{b-a}(x-a)$$

را قراردهیم، قضیهٔ ۶.۱ به‌دست خواهد آمد. روشن‌است که $F(x)$، هم در a و هم در b صفرمی‌شود.

از قضیهٔ ۶.۱ مستقیماً نتیجه می‌شود که اگر $f(x)$ روی $[a, b]$ پیوسته و روی (a, b) مشتقپذیر و c نقطه‌ای از $[a, b]$ باشد، آنگاه مقداری مانند $\theta \in (۰, ۱)$ وجود دارد به‌طوری‌که

$$f(x)=f(c)+(x-c)f'(c+\theta(x-c)) \qquad (۳۰.۱)^*$$

قضیهٔ اساسی حساب دیفرانسیل و انتگرال حکم دقیقتری را دراختیار ما قرارمی‌دهد: اگر $f(x)$ پیوسته و مشتقپذیر باشد، آنگاه به‌ازای هر $x \in [a, b]$ داریم

$$f(x)=f(c)+\int_c^x f'(s)\,ds \qquad (۳۱.۱)$$

ازآنجاکه $f'(x)$ پیوسته‌است، می‌توانیم با استفاده از رابطهٔ فوق وقضیهٔ میانگین‌انتگرال (۲۸.۱)، رابطهٔ (۳۰.۱) را به‌دست‌آوریم.**

قضیهٔ ۷.۱: فرمول تیلر با باقیمانده (انتگرالی). اگر $f(x)$ دارای $n+1$ مشتق‌پیوسته روی $[a, b]$ و c یك نقطه در $[a, b]$ باشد، آنگاه به‌ازای هر $x \in [a, b]$ داریم:

---

* چنانچه درقضیهٔ ۶.۱ به‌جای بازهٔ $[a, b]$ بازهٔ $[c, x]$ را درنظربگیریم، داریم

$$\frac{f(x)-f(c)}{x-c}=f'(\xi)$$

اگر $\xi=c+\theta(x-c)$، رابطهٔ ۳۰.۱ نتیجه می‌شود. ــم.

** فرض‌کنیم $g(x)=۱$. طبق قضیهٔ ۳.۱ داریم،

$$\int_c^x f'(s)\,ds=f'(\xi)\int_c^x ds=(x-c)f'(\xi)=(x-c)f'(c(x-\theta)\theta)$$

و در نتیجه رابطهٔ ۳۰.۱ به‌دست می‌آید. ــم.

$$f(x)=f(c)+f'(c)(x-c)+\frac{f''(c)(x-c)^{۲}}{۲!}+\cdots \qquad (۳۲.۱)$$

$$+\frac{f^{(n)}(c)(x-c)^{n}}{n!}+R_{n+۱}(x)$$

که در فرمول بالا $R_{n+۱}(x)$ برابر است با

$$R_{n+۱}(x)=\frac{۱}{n!}\int_{c}^{x}(x-s)^{n}f^{(n+۱)}(s)\,ds \qquad (۳۳.۱)$$

رابطهٔ (۳۲.۱) را می‌توان با در نظرگرفتن تابع زیر به‌جای $f(x)$ از رابطهٔ (۳۱.۱) به‌دست آورد

$$F(x)=f(x)+f'(x)(c-x)+\frac{f''(x)(c-x)^{۲}}{۲!}+\cdots+\frac{f^{(n)}(x)(c-x)^{n}}{n!}$$

زیرا در این صورت $F'(x)=f^{n+۱}(x)(c-x)^{n}/n!$ و بنابراین با استفاده از(۳۱.۱) داریم

$$F(x)=F(c)+\frac{۱}{n!}\int_{c}^{x}(c-s)^{n}f^{(n+۱)}(s)\,ds$$

اما از آنجاکه $F(c)=f(c)$، داریم

$$f(c)=F(x)+\frac{۱}{n!}\int_{x}^{c}(c-s)^{n}f^{(n+۱)}(s)\,ds$$

که بعد ازقراردادن x به‌جای c و c به‌جای x، رابطهٔ فوق همان رابطهٔ(۳۲.۱) خواهدشد.

درعمل لازم نیست‌که $f^{(n+۱)}(x)$ پیوسته باشد تا (۳۲.۱) برقرار گردد. گرچه اگر در (۳۲.۱) تابع $f^{(n+۱)}(x)$ پیوسته باشد با استفاده از قضیهٔ ۳.۱ شکلی‌آشناتر، ولی با سودمندی‌کمتری برای باقیمانده یعنی:

$$R_{n+۱}(x)=\frac{f^{(n+۱)}(\xi)(x-c)^{n+۱}}{(n+۱)!},\qquad \xi=c+\theta(x-c) \qquad (۳۴.۱)$$

به‌دست می‌آید. با قراردادن $h=x-c$، روابط (۳۲.۱) و (۳۴.۱) شکل‌زیر را به‌خود می‌گیرند:

$$f(c+h)=f(c)+hf'(c)+\frac{h^{۲}}{۲!}f''(c)+\cdots+\frac{h^{n}}{n!}f^{(n)}(c)$$

$$+\frac{h^{n+۱}}{(n+۱)!}f^{(n+۱)}(c+\theta h) \qquad (۳۵.۱)$$

به‌ازای مقداری مانند $\theta \in (0, 1)$.

□ **مثال:** تابع $f(x)=e^x$ در نزدیکی $c=0$ بسط تیلری به‌صورت زیر دارد

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \frac{x^{n+1}e^{\xi}}{(n+1)!} \qquad (۳۶.۱)$$

به‌ازای مقداری مانند ξ بین 0 و x.

بسط $f(x)=\ln x = \log_e x$ پیرامون $c=1$ عبارت‌است از

$$\ln x = (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} + \cdots$$

$$- \frac{(-1)^n(x-1)^n}{n} + \frac{(-1)^n(x-1)^{n+1}\xi^{-(n+1)}}{n+1}$$

در فرمول بالا $0 < x \leqslant 2$ و ξ بین 1 و x است. □

فرمول مشابهی برای توابع چند متغیره برقرار است. این فرمول از قضیهٔ ۷.۱ به‌کمك قضیهٔ زیر به‌دست می‌آید.

قضیهٔ ۸.۱: قاعدهٔ مشتق تابع مضاف[1]. اگر تابع $f(x, y, \ldots, z)$ دارای مشتق اول جزئی پیوسته نسبت به هر یك از متغیرهایش باشد و توابع $x=x(t)$، $y=y(t)$، ...، $z=z(t)$ نیز به‌طور پیوسته نسبت به t مشتقپذیر باشند، آنگاه

$$g(t) = f(x(t), y(t), \ldots, z(t))$$

نیز به‌طور پیوسته مشتقپذیر خواهد بود و

$$g'(t) = \frac{\partial f}{\partial x}x'(t) + \frac{\partial f}{\partial y}y'(t) + \cdots + \frac{\partial f}{\partial z}z'(t)$$

از این قضیه و با دخالت تابع

$$g(t) = f(a + t(x-a), b + t(y-b), \ldots, c + t(z-c))$$

رابطه‌ای برای بیان $f(x, y, \ldots, z)$ برحسب مقدار تابع و مشتقهای جزئی آن در $(a, b, \ldots, c)$ به‌دست می‌آید وسپس سری بسط تیلر این تابع پیرامون $t=0$ در $t=1$ ارزیابی می‌شود. برای مثال توجه به‌این نکته قضیهٔ زیر را می‌دهد

---

1. Chain rule

قضیهٔ ۹۰۱: اگر $f(x, y)$ درهمسایگی D از نقطهٔ (a, b) واقع درصفحهٔ (x, y) دارای مشتقهای جزئی اول و دوم پیوسته باشد، آنگاه به‌ازای کلیهٔ مقادیر (x, y) نزدیك به D و نقطه‌ای مانند $(\xi, \eta) \in D$ وابسته به (x, y) داریم:

$$f(x, y)=f(a, b)+f_x(a, b)(x-a)+f_y(a, b)(y-b)+R_2(x, y) \qquad (۳۷۰۱)$$

و

$$R_2(x, y)=\frac{f_{xx}(\xi, \eta)(x-a)^2}{2}+f_{xy}(\xi, \eta)(x-a)(y-b)+\frac{f_{yy}(\xi, \eta)(y-b)^2}{2}$$

در فرمول بالا زیرنمایه‌های f معرف مشتقات جزئی هستند.

برای مثال بسط $e^{x \sin y}$ پیرامون $(a, b)=(0, 0)$ چنین است:

$$e^{x \sin y}=1+0 \times x+0 \times y+R_2(x, y) \qquad (۳۸۰۱)$$

بالاخره دربحث ویژه‌مقدارهای[1] ماتریسها و درجاهای دیگرقضیهٔ زیرلازم می‌شود.

قضیهٔ ۱۰۰۱: قضیهٔ بنیادی جبر. اگر $p(x)$ یك بسجمله‌ای از درجهٔ $n \geqslant 1$ باشد، یعنی اگر

$$p(x)=a_0+a_1x+a_2x^2+\cdots+a_nx^n$$

و a_0، a_1، ...، a_n اعداد حقیقی یا مختلط باشند و $a_n \neq 0$، آنگاه $p(x)$ حداقل دارای یك ریشه‌است، یعنی حداقل یك عدد مختلط مانند ξ وجود دارد که $p(\xi)=0$.

این قضیهٔ نسبتاً پر محتوا را با قضیهٔ سرراست «یك بسجمله‌ای از درجهٔ n حداکثر n ریشه (با احتساب ریشه‌های مکرر) دارد»، که درفصل دوم اثبات ومثلا درمبحث درونیابی بسجمله‌ای از آن استفاده می‌کنیم، نباید اشتباه گرفت.

تمرین

۱-۷۰۱ در قضیهٔ مقدار میانگین برای انتگرالها، قضیهٔ ۳۰۱، می‌گیریم $f(x)=e^x$ و $g(x)=x$ و $[a, b]=[0, 1]$. نقطهٔ ξ را که به‌وسیلهٔ این قضیه مشخص‌شده‌است پیدا و تحقیق کنید که این نقطه دربازهٔ $(0, 1)$ قراردارد.

۲-۷۰۱ درقضیهٔ مقدار میانگین برای مشتق، قضیهٔ ۶۰۱، می‌گیریم $f(x)=x^2$. نقطهٔ ξ را

---

1. eigenvalues

که به‌وسیلهٔ این قضیه مشخص شده‌است پیدا و تحقیق کنید که این نقطه در بازهٔ (a, b) قرار دارد.

۳-۷۰۱ در بسط (۳۶۰۱) برای e^x، n را طوری پیدا کنید که به‌ازای جمیع مقادیر x روی $[0, 1]$ حاصلجمع توانی نتیجه، تقریبی به‌دست دهد که تا پنج رقم با معنی دقیق باشد.

۴-۷۰۱ با استفاده از فرمول تیلر (۳۲۰۱) بسط سری توانی $\sin(\pi x/2)$ را پیرامون $c = 0$ پیدا کنید. عبارتی برای باقیمانده تعیین نمایید و به‌وسیلهٔ آن تعداد جملات ضروری برای تضمین دقت عملی تا شش رقم با معنی در محاسبهٔ $\sin(\pi x/2)$ را، به‌ازای جمیع مقادیر x متعلق به بازهٔ $[-1, 1]$، برآورد کنید.

۵-۷۰۱ باقیماندهٔ $R_2(x, y)$ در مثال (۳۸۰۱) را پیدا و حداکثر مقدار آن را در ناحیهٔ D که به‌وسیلهٔ

$$\left[0 \leqslant x \leqslant \frac{\pi}{2}, \quad 0 \leqslant y \leqslant \frac{\pi}{2}\right]$$

تعریف می‌شود، تعیین کنید.

۶-۷۰۱ ثابت کنید که جملهٔ باقیمانده در (۳۵۰۱) را می‌توان به‌صورت زیر نیز نوشت

$$\frac{h^{n+1}}{(n+1)!} f^{(n+1)}(c) + o(h^{n+1})$$

[اگر $f^{(n+1)}(x)$ در $x = c$ پیوسته باشد].

۷-۷۰۱ صحت حکم تمرین (۶-۷۰۱) را از راه محاسبهٔ

$$R_2(h) = e^h - \left(1 + h + \frac{h^2}{2}\right), \qquad \frac{h^3}{3!} f^{(3)}(0) = \frac{h^3}{3!}$$

به‌ازای $f(x) = e^x$ و $c = 0$، و به‌ازای مقادیر مختلف h، مثلا $h = 2^{-k}$ $k = 1, 2, \ldots,$ و مقایسهٔ $R_2(h)$ با $f^{(3)}(0)\,(h^3/3!)$ نشان دهید.

۸-۷۰۱ قضیهٔ ۹۰۱ را با استفاده از قضیه‌های ۷۰۱ و ۸۰۱ ثابت کنید.

۹-۷۰۱ فرمول اویلر[1] یعنی

$$e^{i\theta} = \cos\theta + i \sin\theta$$

را ($i = \sqrt{-1}$) به‌وسیلهٔ مقایسهٔ سری نمایی e^x، به‌ازای $x = i\theta$، با حاصلجمع سری توانی $\cos\theta$ و i برابر سری توانی $\sin\theta$ اثبات کنید.

---

1. Euler

۲

درونیابی به‌وسیلهٔ بسجمله‌ایها

بسجمله‌ایها تقریباً در همهٔ زمینه‌های آنالیز عددی به‌عنوان وسیلهٔ اصلی تقریب زدن به‌کار برده می‌شوند. بسجمله‌ایها درحل معادلات، و تقریب‌زدن توابع، تقریب‌زدن انتگرالها و مشتقها، تقریب‌زدن جوابهای انتگرال و معادلات دیفرانسیل و غیره به‌کار برده می‌شوند. مقبولیت بسجمله‌ایها به‌علت ساخت سادهٔ آنهاست که امر ساختن تقریبهای مؤثر و استفاده‌از آنها را آسان می‌سازد.

بدین دلیل، نمایش و ارزیابی بسجمله‌ایها یك مبحث اساسی در آنالیزعددی است. ما این‌موضوع را در این فصل، ضمن گفتگو از درونیابی بسجمله‌ای، که ساده‌ترین ومطمئناً گسترده‌ترین تکنیکی‌است که برای به‌دست آوردن تقریبهای بسجمله‌ای به‌کاربرده می‌شود، مورد بحث قرارخواهیم داد. برای به‌دست آوردن تقریبهای خوب به‌کمك بسجمله‌ایها و توابع تقریبی‌دیگر، روشهای پیشرفته‌تری درفصل۶ داده شده‌اند. اما درهمان‌جا نشان داده خواهد شدکه حتی بهترین تقریب، نتایجی چندان بهتر از یك طرح مناسب درونیابی بسجمله‌ای عاید ما نمی‌سازد.

پایهٔ عمل ما برای بسجمله‌ای درونیاب، تفاضلات منقسم‌است. این خود به‌ما امکان می‌دهدکه با درونیابی بوسانی (یا هرمیتی[1]) به صورت یك حالت حدی خاص درونیابی بسجمله‌ای در نقاط متمایز، عمل کنیم.

۲.۱ صورتهای بسجمله‌ای

دراین بخش، اشاره می‌کنیم که راه معمولی برای توصیف یك بسجمله‌ای ممکن‌است همیشه

1. Hermite

بهترین راه برای محاسبات نباشد، و لذا ما راه دیگر، به‌ویژه صورت نیوتنی را پیشنهاد می‌کنیم. همچنین نشان می‌دهیم که یك بسجمله‌ای را که به‌صورت نیوتنی داده شده چگونه باید ارزیابی کنیم. سرانجام، در مقدمهٔ درونیابی بسجمله‌ای، چگونگی شمارش ریشه‌های یك بسجمله‌ای را نیز مورد بحث قرار می‌دهیم.

بنابر تعریف، **یك بسجمله‌ای $p(x)$ از درجهٔ** نابزرگتر از n، تابعی است به‌صورت

$$p(x)=a_0+a_1x+a_2x^2+\cdots+a_nx^n \qquad (۱۰۲)$$

با ضرایب معین a_0، a_1، ...، a_n، و در مواردی که ضریب جمله پیشرو[1]، a_n، غیرصفر باشد، بسجمله‌ای دقیقاً از **درجهٔ** n، خواهد بود.

در مباحث ریاضی، راه استانداردِ مشخص کردن یك بسجمله‌ای، داشتن **صورت توانی**[2] (۱۰۲) آن است. برای مشتقگیری و انتگرالگیری یك بسجمله‌ای این صورت بسیار مناسب است. اما، در زمینه‌های خاص مختلف صورتهای دیگر مناسبترند.

□ **مثال ۱۰۲: صورت توانی ممکن است به ازدست دادن ارقام با معنی منجر شود.** اگر صورت توانی یك خط مستقیم $p(x)$ را که برای آن داریم، $p(۶۰۰۱)=-۲/۳$ و $p(۶۰۰۰)=۱/۳$، را درنظر بگیریم، آنگاه در حساب با ممیزشناور با پنج رقم اعشاری خواهیم داشت $p(x)=۶۰۰۰ر۳-x$. اگر مقدار این خط راست را درهمان حساب ارزیابی کنیم، خواهیم دید که $p(۶۰۰۰)=۰ر۳$ و $p(۶۰۰۱)=-۰ر۷$، که فقط اولین رقم مقادیر تابع داده شده را به‌دست می‌دهد و درنتیجه موجب از دست رفتن چهار رقم اعشاری می‌شود. □

یك راه چاره برای جلوگیری از این‌گونه از دست دادن ارقام با معنی، استفاده از **صورت توانی منتقله**[3] به‌شرح زیر است:

$$p(x)=a_0+a_1(x-c)+a_2(x-c)^2+\cdots+a_n(x-c)^n \qquad (۲۰۲)$$

اگر **مرکز** c را مساوی ۶۰۰۰ انتخاب کنیم، آنگاه در مثال فوق خواهیم داشت $p(x)=۰ر۳۳۳۳۳-(x-۶۰۰۰ر۰)$، و با محاسبهٔ مقدار تابع درحساب با ممیزشناور با پنج رقم اعشاری، خواهیم داشت

$$p(۶۰۰۱)=-۰ر۶۶۶۶۷ \quad و \quad p(۶۰۰۰)=۰ر۳۳۳۳۳$$

یعنی دراین‌صورت مقادیر تابع تا پنج رقم صحیح‌اند.

هرگاه یك بسجمله‌ای دریك بازهٔ $[a, b]$ موردنظر باشد، بهتر است از صورت‌توانی منتقله با مرکز $c\in[a, b]$ استفاده کنیم. یك راه چارهٔ عاقلانه‌تر برای جلوگیری از به‌هدر رفتن ارقام با معنی (یا بدشرطی[4]) بسط به‌بسجمله‌ایهای چبیشف[5] یا بسجمله‌ایهای متعامد

1. leading 2. Power form 3. Shifted power form
4. illconditioning 5. Chebyshev

دیگر است. (به‌بخش ۳.۶ نگاه کنید).

اگر $p(x)$ به‌وسیلهٔ (۲.۲) داده شده باشد، ضرایب صورت توانی منتقله (۲.۲) مقادیر مشتق زیر را به‌دست می‌دهند، یعنی

$$a_i=\frac{p^{(i)}(c)}{i!} \qquad i=0,\dots,n$$

درحقیقت صورت توانی منتقله همان بسط تیلر[1] $p(x)$ پیرامون مرکز c است.

یك تعمیم دیگر صورت توانی منتقله، صورت نیوتنی آن‌است:

$$\begin{aligned}p(x)=a_0+a_1(x-c_1)+a_2(x-c_1)(x-c_2)\\+a_3(x-c_1)(x-c_2)(x-c_3)+\cdots\\+a_n(x-c_1)(x-c_2)\cdots(x-c_n)\end{aligned} \qquad (۳.۲)$$

این صورت نقش عمده‌ای در ساختن یك بسجمله‌ای درونیاب ایفا می‌کند. اگر مراکز c_1، ...، c_n همگی مساوی c باشند، این‌صورت نیز به‌همان صورت توانی بدل می‌شود، و اگر مراکز c_1، ...، c_n همگی‌مساوی صفر باشند، صورت توانی به‌دست می‌آید. لذا بحث زیر دربارهٔ ارزیابی صورت نیوتنی، مستقیماً برای این صورتهای ساده نیز به‌کار خواهد رفت.

محاسبهٔ جداگانهٔ هر یك از $n+1$ جمله[2] در (۳.۲)، و بعد جمع کردن آنها با هم، بیهوده‌است. این‌کارمتضمن $n+n(n+1)/2$ عمل جمع و $n(n+1)/2$ عمل‌ضرب‌است. درعوض ملاحظه می‌شودکه عامل $(x-c_1)$ درهمهٔ جملات هست، بجز درجمله اول، یعنی

$$\begin{aligned}p(x)=a_0+(x-c_1)\{a_1+a_2(x-c_2)+a_3(x-c_2)(x-c_3)\\+\cdots+a_n(x-c_2)(x-c_3)\cdots(x-c_n)\}\end{aligned}$$

و به‌همین ترتیب، تمامی جملات دیگر داخل ابرو، به‌استثنای اولین جملهٔ آن، شامل عامل $(x-c_2)$ می‌باشند، یعنی

$$\begin{aligned}p(x)=a_0+(x-c_1)\{a_1+(x-c_2)[a_2+a_3(x-c_3)\\+\cdots+a_n(x-c_3)\cdots(x-c_n)]\}\end{aligned}$$

با ادامهٔ عملیات فوق، $p(x)$ را به‌صورت تودرتو[3] به‌دست خواهیم آورد:

$$\begin{aligned}p(x)=a_0+(x-c_1)\{a_1+(x-c_2)[a_2+(x-c_3)<a_3+\cdots\\+(x-c_{n-1})(a_{n-1}+(x-c_n)a_n)\cdots>]\}\end{aligned}$$

---

1. Taylor expansion 2. Term 3. Nested form

که محاسبهٔ آن به‌ازای هرمقدار خاص α تنها شامل $2n$ عمل‌جمع و n عمل ضرب است. برای مثال اگر $p(x)$ به‌صورت

$$p(x)=1+2(x-1)+3(x-1)(x-2)+4(x-1)(x-2)(x-3)$$

باشد و بخواهیم $p(4)$ را محاسبه کنیم، محاسبات به‌صورت زیر انجام می‌گیرند:

$$\begin{aligned}p(4)&=1+(4-1)\{2+(4-2)[3+(4-3)<4>]\}\\&=1+(4-1)\{2+(4-2)[7]\}\\&=1+(4-1)\{16\}\\&=49\end{aligned}$$

این شیوهٔ عمل در الگوریتم زیر، به‌شکل رسمی بیان شده‌است.

الگوریتم ۱.۲: ضرب تودرتو برای صورت نیوتنی. $n+1$ ضریب a_0، ...، a_n برای صورت نیوتنی (۳.۲) از بسجمله‌ای $p(x)$، همراه با مراکز c_1، ...، c_n داده شده‌اند. عدد z نیز مفروض است

$a'_n := a_n$

به‌ازای $i=n-1, n-2, \ldots, 0$ می‌شود:

$\quad a'_i := a_i+(z-c_{i+1})a'_{i+1}$

پس، $a'_0=p(z)$. بعلاوه، کمیتهای کمکی[1] a'_1، ...، a'_n مستقلا مورد علاقهٔ ما هستند. زیرا، داریم

$$\begin{aligned}p(x)=a'_0&+a'_1(x-z)+a'_2(x-z)(x-c_1)\\&+a'_3(x-z)(x-c_1)(x-c_2)\\&+\cdots+a'_n(x-z)(x-c_1)(x-c_2)\cdots(x-c_{n-1})\end{aligned}\qquad (۴.۲)$$

یعنی a'_1، ...، a'_n، ضرایب نیوتنی برای $p(x)$ با مراکز z، c_1، c_2، ...، c_{n-1} هستند.

حال به‌اثبات (۴.۲) می‌پردازیم. بنابرالگوریتم فوق داریم

$$a_n=a'_n$$

$$a_i=a'_i+a'_{i+1}(c_{i+1}-z) \qquad i=n-1, n-2, \ldots, 0$$

---

1. auxilliary

از قراردادن این روابط در (۳.۲) به‌دست می‌آوریم

$$\begin{aligned}
p(x) &= a_0 + a_1(x-c_1) + a_2(x-c_1)(x-c_2) \\
&\quad + \cdots + a_n(x-c_1)\cdots(x-c_n) \\
&= a_0' + a_1'(c_1 - z) \\
&\quad + [a_1' + a_2'(c_2 - z)](x-c_1) \\
&\quad + [a_2' + a_3'(c_3 - z)](x-c_1)(x-c_2) \\
&\quad \vdots \\
&\quad + [a_{n-1}' + a_n'(c_n - z)](x-c_1)\cdots(x-c_{n-1}) \\
&\quad + a_n'(x-c_1)\cdots(x-c_{n-1})(x-c_n) \\
&= a_0' + a_1'(x-z) + a_2'(x-z)(x-c_1) \\
&\quad + \cdots + a_n'(x-z)(x-c_1)\cdots(x-c_{n-1})
\end{aligned}$$

که مؤید صحت رابطهٔ (۴.۲) است.

علاوه بر تعیین مقدار بسجمله‌ای (۳.۲) درهر نقطهٔ خاص z، که ازنظر اقتصادی بسیار مؤثر است، الگوریتم ضرب تودرتو درعبور از یك صورت نیوتنی به‌صورت دیگر، بسیار مفید است. برای مثال، فرض کنید می‌خواهیم بسجمله‌ای

$$p_2(x) = ۱٫۵۷۰۹ + ۰٫۰۰۰۶(x-۱) + ۰٫۰۰۰۰۱۲(x-۱)(x-۴)$$

را برحسب توانهای x، یعنی به‌صورت نیوتنی که کلیهٔ مراکز آن صفرند درآوریم. دراین حال، اگر از الگوریتم ۱.۲ با $z=۰$ (و $n=۲$) استفاده کنیم، خواهیم داشت

$$a_2' = a_2 = ۰٫۰۰۰۰۱۲$$

$$a_1' = a_1 + (z-c_2)a_2' = ۰٫۰۰۰۶ + (۰-۴)(۰٫۰۰۰۰۱۲) = ۰٫۰۰۰۰۱۲$$

$$a_0' = a_0 + (z-c_1)a_1' = ۱٫۵۷۰۹ + (۰-۱)(۰٫۰۰۰۰۱۲) = ۱٫۵۷۰۷۸$$

بنابراین

$$p_2(x) = ۱٫۵۷۰۷۸ + ۰٫۰۰۰۰۱۲(x-۰) + ۰٫۰۰۰۰۱۲(x-۰)(x-۱)$$

با به‌کارگیری الگوریتم ۱.۲ برای این بسجمله‌ای، دوباره به‌ازای $z=۰$، داریم

$$a_2' = a_2 = ۰٫۰۰۰۰۱۲$$

$$a_1' = a_1 + (z-c_2)a_2' = ۰٫۰۰۰۰۱۲ + (۰-۱)(۰٫۰۰۰۰۱۲) = ۰٫۰۰$$

$$a_0' = a_0 + (z-c_1)a_1' = ۱٫۵۷۰۷۸ + (۰-۰)(۰٫۰) = ۱٫۵۷۰۷۸$$

بنابراین

$$p_۲(x)=۱٫۵۷۰۷۸+۰٫۰(x-۰)+۰٫۰۰۰۰۱۲(x-۰)(x-۰)$$
$$=۱٫۵۷۰۷۸+۰٫۰۰۰۰۱۲\,x^۲$$

در این مثال ساده،می‌توانیم صحت نتیجه را سریعاً ازضرب جمله‌های رابطهٔ اولیه، تحقیق کنیم.

$$p_۲(x)=۱٫۵۷۰۹+۰٫۰۰۰۰۶(x-۱)+۰٫۰۰۰۰۱۲(x^۲-۵x+۴)$$
$$=[۱٫۵۷۰۹-۰٫۰۰۰۰۶+(۰٫۰۰۰۰۱۲)(۴)]$$
$$+[۰٫۰۰۰۰۶+(۰٫۰۰۰۰۱۲)(-۵)]\,x+۰٫۰۰۰۰۱۲\,x^۲$$
$$=۱٫۵۷۰۷۸+۰٫۰۰۰۰۱۲\,x^۲$$

استفادهٔ مکرر از الگوریتم‌ضرب تودرتو در محاسبهٔ مشتقهای یک بسجمله‌ای که به‌صورت نیوتنی داده شده باشد (به تمرین ۱.۲ــ۲ تا ۱.۲ــ۵ نگاه‌کنید) نیز مفید است. الگوریتم‌فوق در اثبات مطلب مسلم پایه‌ای زیر نیز مؤثر است.

لم ۱.۲ اگر $z_۱,\ldots,z_k$ ریشه‌های متمایز بسجمله‌ای $p(x)$ باشند، آنگاه

$$p(x)=(x-z_۱)(x-z_۲)\ldots(x-z_k)r(x)$$

که در آن $r(x)$ یک بسجمله‌ای است.

برای اثبات این لم، $p(x)$ را به‌صورت توانی (۱.۲)،یعنی به‌صورت نیوتنی که همهٔ مراکز آن صفرند می‌نویسیم و سپس یکبار الگوریتم (۱.۲) را به‌کار می‌بریم وداریم

$$p(x)=p(z)+(x-z)q(x)$$

[زیرا $a'_۰=p(z)$] که $q(x)=a'_۱+a'_۲x+\ldots+a'_nx^{n-۱}$ یک بسجمله‌ای با درجهٔ کمتر از n است. درواقع، $p(x)$ را بر بسجمله‌ای خطی $(x-z)$ تقسیم کرده‌ایم و $q(x)$ بسجمله‌ای خارج‌قسمت وعدد $p(z)$ باقیماندهٔ آن شده‌است. اکنون به‌طور اخص $z=z_۱$ را انتخاب می‌کنیم و آنگاه طبق فرض مسئله $p(z_۱)=۰$، یعنی

$$p(x)=(x-z_۱)q(x)$$

بدین‌ترتیب به‌ازای $k=۱$ اثبات تمام شده‌است. بعد به‌ازای $k>۱$، نتیجه می‌شود که $z_۲,\ldots,z_k$ الزاماً ریشه‌های $q(x)$ هستند، زیرا بنابرفرض $p(x)$ دراین‌نقاط برابر باصفر می‌شود، درحالی که $x-z_۱$ دراین نقاط برابر با صفر نمی‌شود. بنابراین برای اتمام برهان می‌توان از استقراء نسبت به k، تعداد ریشه‌ها، استفاده‌کرد.

فرع: اگر $p(x)$ و $q(x)$ دو بسجمله‌ای از درجهٔ نابزرگتر از k باشندکه در $k+۱$ نقطهٔ

متمایز $z_0, \ldots, z_k$ برهم منطبق باشند، آنگاه $p(x)=q(x)$.

زیرا، تفاضل آنها $d(x)=p(x)-q(x)$ یك بسجمله‌ای است از درجهٔ نابزرگتر از k، و بنابر لم ۱.۲ می‌توان آن را به‌صورت

$$d(x)=(x-z_0)\ldots(x-z_k)r(x)$$

که در آن $r(x)$ یك بسجمله‌ای است، نوشت. فرض کنید که به‌ازای ضرایبی مانند $c_0, \ldots, c_m$ با شرط $c_m \neq 0$، داشته باشیم $r(x)=c_0+\ldots+c_m x^m$. دراین صورت رابطهٔ

$$k \geqslant d \text{ درجه} = k+1+m$$

برقرار می‌گردد که بیمعنی است. بنابراین $r(x)=0$، و لذا $p(x)=q(x)$.

این فرع به‌این سؤال که «چند بسجمله‌ای از درجهٔ نابزرگتر از k وجود دارد که در $k+1$ نقطهٔ بخصوص مقادیر خاصی داشته باشند؟» جواب «حداکثر یك» را می‌دهد.

این بررسیهای مربوط به‌ریشه‌های بسجمله‌ای را می‌توان درطی بحثی دربارهٔ مفهوم تکرر اصلاح نمود. و این مطلب در بحثهای بعدی‌ما دربارهٔ درونیابی بوسانی بسیار مهم خواهد بود. گوییم z را یك ریشهٔ (دقیق[1]) با تکرر jام[2] یا از مرتبهٔ jام تابع $f(x)$ گوییم به‌شرط آنکه داشته باشیم:

$$f(z)=f'(z)=\ldots=f^{(j-1)}(z)=0 \neq f^{(j)}(z)$$

□ **مثال:** برای مثال، بسجمله‌ای

$$(x-z)^j$$

دارای یك ریشهٔ z از مرتبهٔ j است. قابل‌قبول‌است که یك چنین ریشه‌ای، j مرتبه به‌حساب آید، زیرا که می‌تواند به‌عنوان حد بسجمله‌ای

$$(x-z_1)\ldots(x-z_j)$$

بیان شود که دارای j ریشهٔ متمایز و یا ساده است که همگی در نقطهٔ z برهم منطبق و یا باهم متحد می‌شوند.

مثال‌دیگر آنکه به‌ازای $0<A<1$، تابع $\sin x - Ax$ دربازهٔ $-\pi<x<\pi$ دارای سه ریشهٔ (ساده) است که وقتی $A \to 1$، ریشه‌ها به‌سمت‌عدد صفر همگرا می‌شوند. همین‌طور، تابع (حدی) $\sin x - x$ در صفر دارای یك ریشهٔ سه‌گانه‌است. □

با این برداشت از تکرر یك ریشه، می‌توان به‌لم ۱.۲ صورت قویتری بخشید.

---

1. exact 2. multiplicity

لم ۲.۲ اگر $z_1, z_2, \ldots, z_k$ یك دنباله از ریشه‌های بسجمله‌ای $p(x)$ با احتساب تکرر آنها باشد، آنگاه

$$p(x)=(x-z_1)(x-z_2)\cdots(x-z_k)r(x)$$

که در آن $r(x)$ یك بسجمله‌ای است.

برای اثبات لم ۲.۲ به‌تمرین ۱.۲ـ۶ نگاه کنید. باید توجه داشت در حالتی که z یك ریشهٔ $p(x)$ ازمرتبهٔ j ام باشد، z می‌تواند دردنبالهٔ $z_1, z_2, \ldots, z_k$ به‌تعداد j مرتبه ظاهر شود. از لم ۲.۲ و بحث قبلی نتیجه می‌گیریم

فرع: اگر $p(x)$ و $q(x)$ دو بسجمله‌ای از درجهٔ نابزرگتر از k، با $k+1$ نقطهٔ $z_0, \ldots, z_k$ منطبق برهم باشند، یعنی تفاضل آنها، $r(x)=p(x)-q(x)$، $k+1$ ریشهٔ $z_0, \ldots, z_k$ (با احتساب تکرر آنها) داشته باشد، آنگاه $p(x)\equiv q(x)$.

تمرین

۱.۲ـ۱ مقدار بسجمله‌ای درجهٔ سوم

$$p(x)=(x-۹۹\pi)(x-۱۰۰\pi)(x-۱۰۱\pi)$$

را در $x=۳۱۴٫۱۵$ حساب کنید، سپس با استفاده ازضرب تودرتو برای به‌دست‌آوردن $p(x)$ به‌شکل‌توانی مقدار بسجمله‌ای حاصل را به‌شکل‌توانی در $x=۳۱۴٫۱۵$ به‌دست آورید و نتایج را با هم مقایسه کنید!

۲.۲ـ۲ گیریم $p(x)=a_0+(x-c_1)(a_1+\cdots+(x-c_n)(a_n)\cdots)$ یك‌بسجمله‌ای به‌صورت نیوتنی باشد. ثابت کنید که اگر $c_1=c_2=\cdots=c_{r+1}$، آنگاه به ازای $j=0,\ldots,r$، رابطهٔ $p^{(j)}(c_1)=j!\,a_j$ برقراراست. [راهنمایی: درشرایط فوق، $p(x)$ را می‌توانید به‌صورت

$$p(x)=\sum_{j=0}^{r} a_j(x-c_1)^j+(x-c_1)^{r+1}q(x)$$

که در آن $q(x)$ یك بسجمله‌ای است بنویسید. حال از $p(x)$ مشتق بگیرید.]

۳.۲ـ۳ مشتق اول بسجمله‌ای

$$p(x)=۳-(x-۱)(۴-(x+۱)\{۵-x[۶-(x+۲)]\})$$

را در $x=۲$ پیدا کنید. [راهنمایی: برای به دست آوردن صورت نیوتنی با مراکز ۲، ۲، ۱، ۱ ـ، الگوریتم ۱.۲ را دوبار به‌کار گیرید و سپس طبق تمرین ۱.۲ـ۲ عمل کنید.]

۴-۱۰۲ مشتق دوم بسجمله‌ای $p(x)$ در تمرین ۱۰۲-۳ را نیز در $x=۲$ پیدا کنید.

۵-۱۰۲ بسط تیلر بسجمله‌ای $p(x)$ در تمرین ۱۰۲-۳ را پیرامون $c=۳$ پیدا کنید. [راهنمایی: بسط تیلر برای یك بسجمله‌ای پیرامون نقطهٔ c، همان صورت نیوتنی آن بسجمله‌ای است با مراکز c، c، c، ... ، c، c.]

۶-۱۰۲ لم ۲.۲ را ثابت کنید. [راهنمایی: به موجب الگوریتم ۱.۲، داریم $p(x)=(x-z_۱)q(x)$. اما برای اتمام برهان به‌استقراء نسبت به عدد k، تعداد ریشه‌ها در دنبالهٔ داده شده، باید ثابت نمایید که $z_۲, z_۳, \ldots, z_k$ الزاماً دنباله‌ای از ریشه‌های $q(x)$ (با احتساب تکرر) است. بدین منظور فرض کنید که عدد z دقیقاً j مرتبه در دنبالهٔ $z_۲, z_۳, \ldots, z_k$ ظاهر شود و موارد $z=z_۱$ و $z\neq z_۱$ متمایز باشند. و نیز از رابطهٔ

$$p^{(j)}(x)=(x-z_۱)q^{(j)}(x)+jq^{(j-۱)}(x)$$

استفاده کنید.]

۷-۱۰۲ ثابت کنید که، به‌زبان فرع لم ۲.۲، بسجمله‌ای تیلر $\sum_{i=۰}^{j-۱} f^{(i)}(a)(x-a)^i/i!$ در نقطهٔ $x=a$، j مرتبه[1] بر تابع $f(x)$ منطبق است (یعنی a یك ریشهٔ مرتبهٔ j ام تفاضل آنهاست).

۸-۱۰۲ فرض کنید یك **تابع** $F(X)$ که قراراست مقدار یك بسجمله‌ای خاص از درجهٔ کمتر از r را در نقطهٔ X به‌ما بدهد، داده شده‌است. بعد، فرض کنید که از راه تجسس دریافتید که این تابع واقعاً مقدار یك بسجمله‌ای از درجهٔ کمتر از r را به‌شما می‌دهد (مثلا می‌بینید که فقط حاصلجمع (یا تفاضل) جملاتی پیدا شده‌اند که از ضرب اعداد در توانهای مختلف X کمتر از r بار به‌دست آمده‌اند). مقادیر چند تابع را قبلا باید امتحان کنید تا بتوانید مطمئن شوید که این تابع آنچه را که قرار بوده انجام دهد واقعاً انجام می‌دهد یا نه؟ (البته با فرض نبودن خطای گرد کردن.)

۹-۱۰۲ برای هر یك از سریهای توانی زیر از ضرب تودرتو استفاده کنید تا راه مؤثری برای ارزیابی آنها پیدا کنید. (البته مجبورید فرض کنید که این سریها فقط باید به‌ازای مقداری از N که از قبل تعیین‌شده، روی دامنهٔ $n\leqslant N$ جمع شوند.)

(الف) $$e^x=\sum_n x^n/n!$$

(ب) $$\ln x=۲\sum_{n\ \text{فرد}} \frac{۱}{n}\left(\frac{x-۱}{x+۱}\right)^n$$

(پ) $$\sin^{-۱}x=\sum_{n\ \text{فرد}} \frac{۱\cdot۳\cdots(n-۲)}{۲\cdot۴\cdots(n-۱)}x^n/n$$

---

1. j-fold

۲.۲ وجود و یکتایی[1] بسجمله‌ای درونیاب

گیریم x_0، x_1، ...، x_n، $n+1$ نقطهٔ متمایز روی محور حقیقی xها باشند و $f(x)$ تابعی با مقدار حقیقی باشد که در بازهٔ $I=[a, b]$، که این نقاط را دربردارد، تعریف شده‌است. می‌خواهیم بسجمله‌ای $p(x)$ از درجهٔ نابیشتر از n را طوری به‌دست‌آوریم که تابع $f(x)$ را در نقاط x_0، ...، x_n، $p(x_i)=f(x_i)$، $i=0, 1, \ldots, n$، درونیابی کند، یعنی شرایط برقرار باشند.

به‌طوری که مشاهده خواهیم کرد راههای زیادی برای نوشتن چنین بسجمله‌ای وجود دارد، و بنابراین لازم‌است در آغاز بدخواننده یادآوری کنیم که به‌موجب فرع ۱.۲ حداکثر یك بسجمله‌ای از درجهٔ نابیشتر از n وجود دارد که $f(x)$ را در $n+1$ نقطهٔ متمایز x_0، ...، x_n درونیابی می‌کند.

سپس نشان می‌دهیم که حداقل یك بسجمله‌ای از درجهٔ نابیشتر از n وجود دارد که $f(x)$ را در $n+1$ نقطهٔ متمایز x_0، x_1، ...، x_n درونیابی می‌کند. برای این امر، باز از شکل دیگری از بسجمله‌ای به‌نام **صورت لاگرانژی**[2] استفاده کنیم

$$p(x)=a_0 l_0(x)+a_1 l_1(x)+\cdots+a_n l_n(x) \qquad (۵.۲)$$

که در آن بسجمله‌ایهای

$$l_k(x)=\prod_{\substack{i=0\\ i\neq k}}^{n}\frac{x-x_i}{x_k-x_i} \qquad k=0, \ldots, n \qquad (۶.۲)$$

بسجمله‌ای لاگرانژ برای نقاط x_0، ...، x_n نامیده می‌شوند. تابع $l_k(x)$ حاصلضرب n عامل خطی است، بنابراین یك بسجمله‌ای است دقیقاً از درجهٔ n. لذا رابطهٔ (۵.۲) واقعاً معرف یك بسجمله‌ای است از درجهٔ نابیشتر از n. وانگهی $l_k(x)$، به‌ازای جمیع مقادیر $i\neq k$ در x_i برابر صفر و در x_k برابر با یك‌است. یعنی

$$l_k(x_i)=\begin{cases}1 & ,i=k\\ 0 & ,i\neq k\end{cases} \qquad i=0, \ldots, n$$

این، نشان می‌دهد که

$$p(x_i)=\sum_{k=0}^{n} a_k l_k(x_i)=a_i \qquad i=0, \ldots, n$$

یعنی ضرایب a_0، ...، a_n در صورت لاگرانژی، مقادیر بسجمله‌ای $p(x)$ در نقاط x_0، ...، x_n هستند. درنتیجه برای یك تابع دلخواه $f(x)$، بسجمله‌ای

---

1. Existence and uniqueness 2. Lagrange form

$$p(x)=\sum_{k=0}^{n} f(x_k)l_k(x) \qquad (۷.۲)$$

از درجهٔ نابیشتر از n است که $f(x)$ را در نقاط x_0، ...، x_n درونیابی می‌کند و این مطلب قضیهٔ زیر را ثابت می‌کند.

قضیهٔ ۱.۲ یك تابع حقیقی مقدار $f(x)$ و $n+1$ نقطهٔ متمایز x_0، ...، x_n داده شده‌اند. دقیقاً یك بسجمله‌ای با درجهٔ نابیشتر از n وجود دارد که $f(x)$ را در نقاط x_0، ...، x_n درونیابی می‌کند.

معادلهٔ (۷.۲) **فرمول لاگرانژ** برای بسجمله‌ای درونیاب نامیده می‌شود.

به‌عنوان یك کاربرد ساده، مورد $n=1$ را بررسی می‌کنیم. یعنی تابع $f(x)$ و دو نقطهٔ متمایز x_0، x_1 داده شده‌اند. آنگاه داریم

$$l_0(x)=\frac{x-x_1}{x_0-x_1} \qquad l_1(x)=\frac{x-x_0}{x_1-x_0}$$

و

$$p(x)=f(x_0)l_0(x)+f(x_1)l_1(x)=f(x_0)\frac{x-x_1}{x_0-x_1}+f(x_1)\frac{x-x_0}{x_1-x_0}$$

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

$$=f(x_0)+\frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_0)$$

این همان مورد آشنای درونیابی خطی[1] است که به‌یکی از صورتهای هم ارزش نوشته شده‌است.

□ **مثال ۲.۲:** انتگرالی مربوط به‌انتگرال بیضوی کامل به‌شرح زیر تعریف می‌شود

$$K(k)=\int_0^{\pi/2}\frac{dx}{[1-(\sin k)^2\sin^2 x]^{1/2}} \qquad (۸.۲)$$

از جدول مقادیر این انتگرالها به‌ازای مقادیر مختلف k که برحسب درجه اندازه‌گیری شده‌اند، نتیجه می‌گیریم

$$K(1)=1.۵۷۰۹$$

---

1. linear interpolation

$$K(۴) = ۱٫۵۷۲۷$$

$$K(۶) = ۱٫۵۷۵۱$$

با استفاده از بسجمله‌ای درونیاب درجهٔ دوم، مقدار $K(۳٫۵)$ را محاسبه کنیم. داریم:

$$l_۰(۳٫۵) = \frac{(۳٫۵-۴)(۳٫۵-۶)}{(۱-۴)(۱-۶)} = \frac{۱٫۲۵}{۱۵} = ۰٫۰۸۳۳۳$$

$$l_۱(۳٫۵) = \frac{(۳٫۵-۱)(۳٫۵-۶)}{(۴-۱)(۴-۶)} = \frac{-۶٫۲۵}{-۶} = ۱٫۰۴۱۶۷$$

$$l_۲(۳٫۵) = \frac{(۳٫۵-۱)(۳٫۵-۴)}{(۶-۱)(۶-۴)} = \frac{-۱٫۲۵}{۱۰} = -۰٫۱۲۵۰۰$$

لذا

$$K(۳٫۵) \approx (۱٫۵۷۰۹)(۰٫۰۸۳۳۳) + (۱٫۵۷۲۷)(۱٫۰۴۱۶۷) + (۱٫۵۷۵۱)(-۰٫۱۲۵۰۰) = ۱٫۵۷۲۱۲۵$$

رقم آخر این تقریب توأم با خطاست. □

صورت لاگرانژی (۷.۲) برای بسجمله‌ای درونیاب، اثبات وجود یک بسجمله‌ای درونیاب را آسان می‌سازد. اما محاسبهٔ مقدار آن در یک نقطه مستلزم حداقل $۲(n+۱)$ عمل ضرب یا تقسیم و $(۲n+۱)$ عمل جمع و تفریق می‌باشد، این بعد از آن است که مخرج بسجمله‌ای لاگرانژ یکبار برای همیشه محاسبه و بر مقادیر تابع متناظر تقسیم شده‌است. این تعداد عملیات باید با n عمل ضرب و n عمل جمع که برای محاسبهٔ مقدار یک بسجمله‌ای درجهٔ nام به‌صورت توانی، به‌وسیلهٔ ضرب تودرتو لازم است، مقایسه شود (به‌الگوریتم ۱.۲ نگاه کنید).

یک ایراد جدیتر به‌روش لاگرانژ ایراد زیر است: درعمل اغلب معلوم نیست که چه تعداد نقاط درونیابی باید به‌کار گرفته شود. بنابراین به‌کمک $p_j(x)$ که معرف یک بسجمله‌ای از درجهٔ نابیشتر از j است و $f(x)$ را در نقاط $x_۰$، ...، x_j درونیابی می‌کنند، مقادیر $p_۰(x)$، $p_۱(x)$، $p_۲(x)$، ...، را محاسبه می‌کنیم و تعداد نقاط درونیابی و درنتیجه درجهٔ بسجمله‌ای درونیاب را افزایش می‌دهیم، به‌این امید که $p_k(x)$ تقریب موردقبولی از $f(x)$ را به‌دست دهد. درچنین موردی به‌کار گرفتن صورت لاگرانژی بیفایده به‌نظر می‌رسد. زیرا که در محاسبهٔ $p_k(x)$ از این واقعیت که $p_{k-۱}(x)$ در دست است هیچگونه استفاده‌ای

تمی‌شود. بدین دلیل و به‌دلایل دیگر صورت نیوتنی برای بسجمله‌ای درونیاب بسیار مناسبتر است.

درحقیقت، با استفاده از نقاط درونیابی $x_0, \ldots, x_{n-1}$ به‌عنوان مراکز، بسجمله‌ای درونیاب $p_n(x)$ را به‌صورت نیوتنی، یعنی به‌شرح زیر می‌نویسیم

$$p_n(x) = A_0 + A_1(x-x_0) + A_2(x-x_0)(x-x_1) + \cdots + A_n(x-x_0)\cdots(x-x_{n-1}) \qquad (۹.۲)$$

به‌ازای هر عدد صحیح k بین ۰ و n، گیریم تابع $q_k(x)$ مجموع $k+1$ جملهٔ اول این صورت باشد،

$$q_k(x) = A_0 + A_1(x-x_0) + A_2(x-x_0)(x-x_1) + \cdots + A_k(x-x_0)\cdots(x-x_{k-1})$$

دراین‌صورت هر کدام از جمله‌های[1] باقیمانده در (۹.۲) دارای عامل $(x-x_k)\cdots(x-x_0)$ است و می‌توانیم رابطهٔ (۹.۲) را به‌صورت

$$p_n(x) = q_k(x) + (x-x_0)\cdots(x-x_k)r(x)$$

که در آن، صورت بسجمله‌ای $r(x)$ چندان مهم نیست بنویسیم. نکتهٔ مهم این است که این جملهٔ آخر یعنی $r(x)(x-x_k)\cdots(x-x_0)$ در نقاط $x_0, \ldots, x_k$ صفر می‌شود. از این‌رو خود $q_k(x)$ می‌باید $f(x)$ را در $x_0, \ldots, x_k$ درونیابی کند [زیرا که $p_n(x)$ درونیابی می‌کند]. از آنجا که $q_k(x)$ نیز یك بسجمله‌ای از درجهٔ نابیشتر از k است، نتیجه می‌شود که $q_k(x) = p_k(x)$، یعنی $q_k(x)$ می‌باید بسجملهٔ یکتایی از درجهٔ نابیشتر از k باشد که $f(x)$ را در نقاط $x_0, \ldots, x_k$ درونیابی می‌کند.

این نشان می‌دهد که صورت نیوتنی (۹.۲) برای بسجمله‌ای درونیاب $p_n(x)$ را می‌توان مرحله به‌مرحله همان‌طوری که دنبالهٔ $p_0(x), p_1(x), p_2(x), \ldots$، ساخته می‌شود بنا کرد و $p_k(x)$ را با افزودن جملهٔ بعدی درصورت نیوتنی (۹.۲)، یعنی

$$p_k(x) = p_{k-1}(x) + A_k(x-x_0)\cdots(x-x_{k-1})$$

از $p_{k-1}(x)$ به‌دست آورد. همچنین رابطهٔ بالا نشان می‌دهد که ضریب A_k در صورت نیوتنی (۹.۲)، برای بسجمله‌ای درونیاب ضریب جملهٔ پیشرو، یعنی ضریب x^k در بسجمله‌ای $p_k(x)$ است که $p_k(x)$ از درجهٔ نابیشتر از k و در $x_0, \ldots, x_k$ بر $f(x)$ منطبق است. این ضریب فقط به‌مقادیر $f(x)$ در نقاط $x_0, \ldots, x_k$ بستگی دارد، این ضریب k امین تفاضل منقسم $f(x)$ در نقاط $x_0, \ldots, x_k$ نامیده شده (که دلایل آن در بخش بعدی آمده‌است) و

1. terms

با نماد

$$f[x_0, \cdots, x_k]$$

نشان داده شده است. با این تعریف، به‌صورت **نیوتنی** برای بسجمله‌ای درونیاب می‌رسیم

$$p_n(x)=f[x_0]+f[x_0, x_1](x-x_0)+f[x_0, x_1, x_2](x-x_0)(x-x_1)$$
$$+\cdots+f[x_0, x_1, \cdots, x_n](x-x_0)(x-x_1)\cdots(x-x_{n-1})$$

فرمول بالا را می‌توان به‌صورت فشرده‌تر:

$$\boxed{p_n(x)=\sum_{i=0}^{n} f[x_0, \cdots, x_i]\prod_{j=0}^{i-1}(x-x_j)} \qquad (۱۰.۲)$$

نوشت. اگر از قرارداد

$$\prod_{m=r}^{s} a_m=\begin{cases} a_r a_{r+1}\cdots a_s & \text{برای } r\leqslant s \\ 1 & \text{برای } r>s \end{cases}$$

استفاده کنیم، به‌ازای $n=1$، رابطهٔ (۱۰.۲) به‌صورت

$$p_1(x)=f[x_0]+f[x_0, x_1](x-x_1)$$

درمی‌آید و درمقایسه با فرمول $p_1(x)=f(x_0)+[f(x_1)-f(x_0)]/[x_1-x_0]$ که قبلا به‌دست آمده‌است، دیده می‌شود که

$$f[x_0]=f(x_0)$$

$$f[x_0, x_1]=[f(x_1)-f(x_0)]/(x_1-x_0)=\frac{f(x_0)-f(x_1)}{x_0-x_1} \qquad (۱۱.۲)$$

درهرحال نخستین تفاضل منقسم، نسبت تفاضلهاست.

تمرین

۱-۲۰۲ با درنظرگرفتن رابطهٔ $w(x)=(x-x_0)(x-x_1)\cdots(x-x_n)$، ثابت کنید که $f[x_0, x_1, \cdots, x_n]=\sum_{i=0}^{n} f(x_i)/w'(x_i)$. [راهنمایی: ضریب پیشرو بسجمله‌ای (۷.۲) را پیدا کنید.]

۲-۲۰۲ حد فرمولی راکه برای $f[x_0, x_1, \cdots, x_n]$ درتمرین ۲۰۲-۱ داده شده، در حالتی که $x_1 \to x_2$، وکلیهٔ نقاط دیگر ثابت بمانند، محاسبه کنید.

۳-۲-۲ در صورتی که $f(x)$ یك بسجمله‌ای نابیشتر از درجهٔ nام باشد، ثابت کنید آن بسجمله‌ای از درجهٔ نابیشتر از n که $f(x)$ را در $n+1$ نقطهٔ متمایز درونیابی کند، همان خود $f(x)$ است.

۴-۲-۲ ثابت کنید که kامین تفاضل منقسم $p[x_0, \cdots, x_k]$ از بسجمله‌ای درجهٔ نابیشتر از kی $p(x)$، به‌نقاط درونیابی x_0، x_1، ...، x_k بستگی‌ندارد.

۵-۲-۲ ثابت کنید که kامین تفاضل منقسم یك بسجمله‌ای با درجهٔ کمتر از k، برابر صفر است.

۳.۲ جدول تفاضل منقسم

تفاضلهای منقسم از مرتبهٔ بالاتر را می‌توان از فرمول

$$f[x_0, \cdots, x_k] = \frac{f[x_1, \cdots, x_k] - f[x_0, \cdots, x_{k-1}]}{x_k - x_0} \qquad (۱۲.۲)$$

به‌دست آورد که صحت آن را می‌توان به‌شرح زیر بررسی کرد.

گیریم $p_i(x)$ یك بسجمله‌ای از درجهٔ نابیشتر از i باشد که درنقاط x_0، ...، x_i بر $f(x)$ منطبق است. فرض می‌کنیم $q_{k-1}(x)$ یك بسجمله‌ای از درجهٔ نابیشتر از $k-1$ باشد که در نقاط x_1، ...، x_k بر $f(x)$ منطبق‌است. دراین‌صورت

$$p(x) = \frac{x - x_0}{x_k - x_0} q_{k-1}(x) + \frac{x_k - x}{x_k - x_0} p_{k-1}(x) \qquad (۱۳.۲)$$

یك بسجمله‌ای از درجهٔ نابیشتر از k است و به آسانی دیده می‌شود که به‌ازای $i = 0, \cdots, k$، $p(x_i) = f(x_i)$. درنتیجه به‌علت یکتا بودن بسجمله‌ای درونیاب، باید داشته باشیم $p(x) = p_k(x)$. بنابراین

به‌موجب تعریف $f[x_0, \cdots, x_k] =$ ضریب پیشرو $p_k(x)$

بنابر (۱۳.۲) $= \dfrac{\text{ضریب پیشرو } q_{k-1}(x)}{x_k - x_0} - \dfrac{\text{ضریب پیشرو } p_{k-1}(x)}{x_k - x_0}$

بنابر تعریف $= \dfrac{f[x_1, \cdots, x_k] - f[x_0, \cdots, x_{k-1}]}{x_k - x_0}$

که اثبات فرمول مهم (۱۲.۲) است.

□ **مثال ۳.۲:** با استفاده از فرمول نیوتن مثال ۲.۲ را حل کنید.

در این مثال باید بسجمله‌ای $p_۲(x)$ از درجهٔ نابیشتر از ۲ را طوری تعیین کنیم که روابط زیر برقرار باشند

$$p_۲(۱)=۱٫۵۷۰۹ \qquad p_۲(۴)=۱٫۵۷۲۷ \qquad p_۲(۶)=۱٫۵۷۵۱$$

به‌موجب (۱۱.۲) می‌توانیم محاسبات زیر را انجام دهیم

$$K[۱,\ ۲]=\frac{۱٫۵۷۰۹-۱٫۵۷۲۷}{۱-۴}=۰٫۰۰۰۶$$

$$K[۴,\ ۶]=\frac{۱٫۵۷۲۷-۱٫۵۷۵۱}{۴-۶}=۰٫۰۰۱۲$$

بنابراین، به‌موجب رابطهٔ (۱۲.۲) داریم

$$K[۱,\ ۴,\ ۶]=\frac{۰٫۰۰۰۶-۰٫۰۰۱۲}{۱-۶}=۰٫۰۰۰۱۲$$

و بنابر (۱۰.۲) خواهیم داشت

$$p_۲(x)=۱٫۵۷۰۹+۰٫۰۰۰۶(x-۱)+۰٫۰۰۰۱۲(x-۱)(x-۴)$$

با قراردادن $x=۳٫۵$ دراین فرمول خواهیم داشت

$$p_۲(۳٫۵)=۱٫۵۷۰۹+(۰٫۰۰۰۶)(۲٫۵)+(۰٫۰۰۰۱۲)(۲٫۵)(-۰٫۵)$$
$$=۱٫۵۷۲۲۵$$

که با نتایج به‌دست‌آمده در مثال ۲.۲ یکی است. □

معادلهٔ (۱۲.۲) نشان می‌دهد که kامین تفاضل منقسم همان تفاضل خارج‌قسمت $(k-۱)$امین تفاضل منقسم است، که مؤید مناسب بودن نام «تفاضل منقسم» است. همچنین معادلهٔ (۱۲.۲) به‌ما امکان می‌دهد که کلیهٔ تفاضلهای منقسم را که برای فرمول نیوتن (۱۰.۲) لازم‌است با یك روش ساده به‌کمك جدولی موسوم به‌جدول تفاضل منقسم تولید کنیم.

چنین جدولی درشکل ۱.۲، به‌ازای $n=۴$ داده شده‌است.

x_i	$f[\ ]=f(\ )$	$f[,]$	$f[,,]$	$f[,,,]$	$f[,,,,]$
x_0	$f[x_0]$				
		$f[x_0, x_1]$			
x_1	$f[x_1]$		$f[x_0, x_1, x_2]$		
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$	
x_2	$f[x_2]$		$f[x_1, x_2, x_3]$		$f[x_0, x_1, x_2, x_3, x_4]$
		$f[x_2, x_3]$		$f[x_1, x_2, x_3, x_4]$	
x_3	$f[x_3]$		$f[x_2, x_3, x_4]$		
		$f[x_3, x_4]$			
x_4	$f[x_4]$				

شکل ۱۰۲ جدول تفاضل منقسم.

درایه‌های جدول، مثلا ستون به‌ستون، طبق الگوریتم زیر محاسبه شده‌اند.

الگوریتم ۲۰۲: جدول تفاضل منقسم. دوستون اول جدول شامل x_0، x_1، ...، x_n و مقادیر متناظر آنها، $f[x_0]$، $f[x_1]$، ...، $f[x_n]$، داده شده‌اند.

For $k = 1, \ldots, n$, do:
  For $i = 0, \ldots, n-k$, do:

$$f[x_i, \ldots, x_{i+k}] := \frac{f[x_{i+1}, \ldots, x_{i+k}] - f[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}$$

اگر عملیات مربوط به‌این الگوریتم با دست انجام پذیرد، دستورالعملهای زیر ممکن است مفید واقع شوند. از درایه‌ای که باید محاسبه شود دو قطر از دو درایهٔ مجاورش به‌طرف چپ رسم کنید. اگر این خطوط به‌ترتیب در $f[x_i]$ و $f[x_j]$ خاتمه پذیرند، تفاضل دو درایهٔ مجاور را بر تفاضل متناظر آنها $x_j - x_i$ تقسیم کنید تا درایهٔ مطلوب را به‌دست آورید. این امر در شکل ۱۰۲ برای درایهٔ $f[x_1, \ldots, x_4]$ نشان داده شده‌است.

پس از آنکه جدول تفاضل منقسم تکمیل گردید به‌ازای $i = 0, \ldots, n$، ضرایب $f[x_0, \ldots, x_i]$ را می‌توان برای فرمول نیوتن (۱۰۰۲) در صدر ستونهای مربوطه پیدا کرد.

به‌دلیل نیاز به‌حافظه و همچنین به‌علت آنکه در بسیاری از زبانهای **فورترن تنها** متغیرهای موجود در عبارت DO می‌توانند زیاد شوند، الگوریتم ۲۰۲ را به‌صورت اصلاح شده در برنامه‌های فورترن به‌کار می‌برند. اولا برای ارزیابی صورت نیوتنی به موجب الگوریتم ۱۰۲ آسانتر است که از صورت

$$p_n(x) = \sum_{i=0}^{n} f[x_i, \ldots, x_n] \prod_{j=i+1}^{n} (x - x_j)$$

یعنی، صورت نیوتنی با مراکز x_1، x_{n-1}، ...، x_n استفاده کنیم، زیرا مقدار $v = p_n(z)$

را می‌توان بعداً طبق الگوریتم ۱.۲ و

$v := f[x_0, \ldots, x_n]$
For $i = 1, \ldots, n$, do:
  $v := f[x_i, \ldots, x_n] + (z - x_i)v$

محاسبه کرد. ثانیاً از آنجا که به‌ازای $i = 0, \ldots, n$ تنها اعداد $f[x_i, \ldots, x_n]$ مورد نظر هستند نیازی به‌ذخیره کردن تمامی جدول تفاضل منقسم (که مستلزم یك آرایهٔ دوبعدی[1] است که در آن تقریباً از نیمی از درایه‌ها درهرحال، به‌علت ویژگی مثلثی جدول تفاضل‌منقسم، استفاده نخواهد شد) نیازی نیست. زیرا اگر از نماد اختصاری

$$d_{ij} = f[x_i, \ldots, x_{i+j}]$$

استفاده کنیم، محاسبات الگوریم ۲.۲ به‌صورت زیر درمی‌آید

For $k = 1, \ldots, n$, do:
  For $i = 0, \ldots, n - k$, do:
    $d_{ik} := (d_{i+1,k-1} - d_{i,k-1})/(x_{i+k} - x_i)$

ومخصوصاً به‌محض اینکه d_{ik} محاسبه شد، عدد $d_{i,k-1}$ دیگر مورد استفاده قرار نخواهد گرفت، لذا با اطمینان می‌توان d_{ik} را روی $d_{i,k-1}$ ذخیره کرد.

الگوریتم ۳.۲: محاسبهٔ ضرایب فرمول نیوتن. $n+1$ نقطهٔ متمایز x_0، ...، x_n واعداد متناظر آنها $f(x_0)$، ...، $f(x_n)$، که $f(x_i)$، $(i = 0, \ldots, n)$، در بردار d_i ذخیره شده‌است

For $k = 1, \ldots, n$, do:
  For $i = 0, \ldots, n - k$, do:
    $d_i := (d_{i+1} - d_i)/(x_{i+k} - x_i)$

مفروض‌اند. پس به‌ازای $i = 0, \ldots, n$، داریم $d_i = f[x_i, \ldots, x_n]$.

□ **مثال ۴.۲:** گیریم $f(x) = (1 + x^2)^{-1}$. به‌ازای $n = 2, 4, \ldots, 16$ مطلوب‌است محاسبهٔ بسجمله‌ای $p_n(x)$ از درجهٔ نابیشتر از n که $f(x)$ را در $n+1$ نقطهٔ

$$x_i = i\frac{10}{n} - 5, \qquad i = 0, \ldots, n$$

با فواصل مساوی درونیابی کند. سپس حداکثر خطای درونیابی

$$E_n = \max_{-5 \le x \le 5} |f(x) - p_n(x)| \qquad n = 2, 4, \ldots, 16$$

---

1. Two-dimensional Array

را دربازهٔ $[-۵, ۵]$ از راه محاسبهٔ

$$E_n \approx \max_i |f(y_i) - p_n(y_i)|$$

که درآن

$$y_i = \frac{i}{۱۰} - ۵ \qquad i = ۰, \cdots, ۱۰۰$$

به‌دست آورید. برای حل این مسئله در برنامهٔ **فورترن** زیر، ازالگوریتم ۱.۲ و ۳.۲ استفاده شده‌است.

برنامهٔ فورترن برای مثال۴.۲

```
C  PROGRAM FOR EXAMPLE 2.4
      INTEGER I,J,K,N,NP1
      REAL D(17),ERRMAX,H,PNOFY,X(17),Y
C  POLYNOMIAL INTERPOLATION AT EQUALLY SPACED POINTS TO THE FUNCTION
           F(Y)  =  1./(1. + Y*Y)
C
      PRINT 600
  600 FORMAT('1   N',5X,'MAXIMUM ERROR')
      DO 40 N=2,16,2
         NP1 = N+1
         H = 10./FLOAT(N)
         DO 10 I=1,NP1
            X(I) = FLOAT(I-1)*H - 5.
            D(I) = F(X(I))
   10    CONTINUE
C              CALCULATE DIVIDED DIFFERENCES BY ALGORITHM 2.3
         DO 20 K=1,N
            DO 20 I=1,NP1-K
               D(I) = (D(I+1) - D(I))/(X(I+K) - X(I))
   20    CONTINUE
C              ESTIMATE MAXIMUM INTERPOLATION ERROR ON (-5,5)
         ERRMAX = 0.
         DO 30 J=1,101
            Y = FLOAT(J-1)/10. - 5.
C              CALCULATE PN(Y) BY ALGORITHM 2.1
            PNOFY = D(1)
            DO 29 K=2,NP1
               PNOFY = D(K) + (Y - X(K))*PNOFY
   29       CONTINUE
            ERRMAX = MAX(ABS(F(Y) - PNOFY), ERRMAX)
   30    CONTINUE
         PRINT 630, N,ERRMAX
  630    FORMAT(I5,E18.7)
   40 CONTINUE
                                          STOP
      END
```

برونداد (نتایج)کامپیوتری برای مثال ۴.۲

N	MAXIMUM ERROR
2	6.4615385E − 01
4	4.3813387E − 01
6	6.1666759E − 01
8	1.0451739E + 00
10	1.9156431E + 00
12	3.6052745E + 00
14	7.1920080E + 00
16	1.4051542E + 01

ملاحظه می‌کنید که اگرچه در روند درونیابی خود از اطلاعات موجود در باب $f(x)$ بیش از پیش استفاده کردیم، ولی خطای درونیابی به‌سرعت با درجهٔ رو به‌افزایشی ازدیاد پیدا کرد. علت آن، استفادهٔ ما از نقاط درونیابی با فاصلهٔ یکنواخت است. به تمرین ۱۰.۶–۱۲ و معادلهٔ (۲۰.۶) نگاه کنید. □

تمرین

۱–۳.۲ از یك جدول لگاریتم، مقادیر $\log x$ در نقاط x طبق جدول زیر به‌دست آمده است

x	$\log x$
۱٫۰	۰٫۰
۱٫۵	۰٫۱۷۶۰۹
۲٫۰	۰٫۳۰۱۰۳
۳٫۰	۰٫۴۷۷۱۲
۳٫۵	۰٫۵۴۴۰۷
۴٫۰	۰٫۶۰۲۰۶

جدول تفاضل منقسم این نقاط را تشکیل دهید.

۲–۳.۲ با استفاده از جدول تفاضل منقسم در تمرین ۲.۳–۱ و به‌کارگرفتن صورت نیوتنی یك بسجمله‌ای درونیاب از درجهٔ سوم، مقادیر $\log$ ۲٫۵، $\log$ ۱٫۲۵، $\log$ ۳٫۲۵ را درونیابی کنید.

۳–۳.۲ خطای حاصل در نتیجهٔ محاسبهٔ $\log$ ۲٫۵ در تمرین ۲.۳–۲ را از راه محاسبهٔ جملهٔ بعدی در بسجمله‌ای درونیاب، برآورد کنید. همچنین از مقایسهٔ مقدار تقریبی $\log$ ۲٫۵ با حاصلجمع $\log$ ۲ و مقدار تقریبی $\log$ ۱٫۲۵، این خطا را تخمین بزنید.

۴–۳.۲ فرمول

$$f[x_0, \cdots, x_j] = f[x_1, \cdots, x_{j+1}] + (x_0 - x_{j+1}) f[x_0, \cdots, x_{j+1}]$$

را به‌دست آورید. سپس از آن به‌عنوان راهی برای محاسبهٔ $p[z, x_0, \cdots, x_{n-1}]$، $p[z, x_0, \cdots, x_{n-2}]$، ...، $p[z, x_0]$ و $p[z]$، یعنی راهی برای به‌دست آوردن قطر دیگری در جدول تفاضل منقسم برای $p(x)$، در تعبیر الگوریتم ضرب تودرتوی ۱.۲ که در بسجمله‌ای (۱۰.۲) به‌کار برده شده استفاده کنید.

۵-۳.۲ طبق تمرین ۲.۲-۳، یك بسجمله‌ای از درجهٔ نابیشتر از k که یك تابع $f(x)$ را در نقاط $x_0, \ldots, x_k$ درونیابی کند، خود $f(x)$ است هرگاه بسجمله‌ای از درجهٔ نابیشتر از k باشد. این‌حقیقت را می‌توان برای بررسی‌دقت بسجمله‌ای درونیابی که حساب شده[1] به‌کار گرفت. برنامهٔ **فورترن** درمثال۴.۲ را برای انجام این بررسی به‌طریق زیر تغییر دهید. به‌ازای $n = 4, 8, 12, \ldots, 32$، بسجمله‌ای $p_n(x)$ از درجهٔ نابیشتر از n را چنان‌پیدا کنیدکه تابع $f_n(x) = \Pi_{j=1}^{n}(x - j - 1/3)$ را در $0, 1, 2, \ldots, n$ درونیابی کند. سپس $\max|f_n(y_i) - p_n(y_i)|$ را به‌وسیلهٔ $E_n = \max_{x_0 \leqslant x \leqslant x_n}|f_n(x) - p_n(x)|$ برآورد کنید. y_i دراینجا، تعداد زیادی نقطهٔ مناسب در $[0, n]$ است.

۶-۳.۲ ثابت کنید که $p_2'(x)$، مشتق سهمی درونیاب تابع $f(x)$ در $x_2 < x_1 < x_0$، خط مستقیمی است که مقدار آن در نقطهٔ $(x_{i-1} + x_i)/2$ به‌ازای $i = 1, 2$، برابر $f[x_{i-1}, x_i]$ است. در موردی که $p_n(x)$ تابع $f(x)$ را در $x_0 < x_1 < \ldots < x_n$ درونیابی کند، از تعمیم این مطلب برای بیان $p_n'(x)$ به‌عنوان درونیاب داده‌های $\{\xi_i, f(x_i, x_{i+1})\}$ به‌ازای ξ_iی مناسب در $x_i \leqslant x \leqslant x_{i+1}$، استفاده نمایید.

۴.۲* درونیابی دریك عدهٔ روبه‌افزایش از نقاط درونیابی

اکنون با استفاده از درونیابی بسجمله‌ای درنقاط متمایز $x_0, x_1, x_2, \ldots$ به‌مسئلهٔ برآورد $f(x)$ درنقطهٔ $x = \bar{x}$ می‌پردازیم. با استفاده از بسجمله‌ای $p_k(x)$ از درجهٔ نابیشتر از k که $f(x)$ رادرنقاط $x_0, \ldots, x_k$ درونیابی می‌کند، متوالیاً مقادیر $p_0(\bar{x})$، $p_1(\bar{x})$، $p_2(\bar{x})$، ... را تا آنجا محاسبه می‌کنیم که احتمالاً، تفاضل بین $p_{k+1}(\bar{x})$ و $p_k(\bar{x})$ به‌اندازهٔ کافی کوچك شود. صورت نیوتنی برای بسجملهٔ درونیاب

$$p_k(x) = \sum_{i=0}^{k} f[x_0, \ldots, x_i]\psi_i(x)$$

با

$$\psi_i(x) = (x - x_0)\cdots(x - x_{i-1})$$

بخصوص برای چنین‌محاسباتی‌طرح شده‌است. اگر $p_k(\bar{x})$، $\psi_k(\bar{x})$ و $f[x_0, \ldots, x_{k+1}]$ در دست باشند، آنگاه می‌توانیم $p_{k+1}(\bar{x})$ را به‌وسیلهٔ زیرمحاسبه کنیم،

$$p_{k+1}(\bar{x}) = p_k(\bar{x}) + f[x_0, \ldots, x_{k+1}]\psi_k(\bar{x})(\bar{x} - x_k)$$

الگوریتم ۴.۲: درونیابی با استفاده از یك عدهٔ روبه‌افزایش از نقاط درونیاب. نقاط متمایز $x_0, x_1, x_2, \ldots$ و مقادیر $f(x_0), f(x_1), f(x_2), \ldots$ از تابع $f(x)$ در این نقاط و همچنین یك نقطهٔ $\bar{x}$ داده شده‌اند.

---

1. Computed

$$
\begin{aligned}
&f[x_0] := f(x_0),\ p_0(\bar{x}) := f(x_0),\ \psi_0(\bar{x}) := 1\\
&\text{For } k = 0, 1, 2, \ldots, \text{ until satisfied, do:}\\
&\quad f[x_{k+1}] := f(x_{k+1})\\
&\quad \text{For } i = k, \ldots, 0, \text{ do:}\\
&\quad\quad f[x_i, \ldots, x_{k+1}] := \frac{f[x_{i+1}, \ldots, x_{k+1}] - f[x_i, \ldots, x_k]}{x_{k+1} - x_i}\\
&\quad \psi_{k+1}(\bar{x}) := \psi_k(\bar{x})(\bar{x} - x_k)\\
&\quad p_{k+1}(\bar{x}) := p_k(\bar{x}) + f[x_0, \ldots, x_{k+1}]\psi_{k+1}(\bar{x})
\end{aligned}
$$

این الگوریتم، هر بار درایه‌های یك قطر از جدول تفاضل منقسم برای $f(x)$ رادر x_0، x_1، x_2، ... تولید می‌کند. هنگام محاسبهٔ $p_{k+1}(\bar{x})$ قطر بالاروندهٔ ماربر $f[x_{k+1}]$ با استفاده از عدد $f[x_{k+1}] = f(x_{k+1})$ و درایه‌های قبلا حساب شدهٔ $f[x_k]$، $f[x_{k-1}, x_k]$، ...، $f[x_0, \ldots, x_{k+1}]$ درقطر پیشین، تا با عدد $f[x_0, x_1, \ldots, x_k]$ محاسبه شد. حتی اگر فقط جدیدترین قطر محاسبه شده (مثلا، دریك برنامهٔ **فورترن**)، حفظ شود، این الگوریتم اتفاقاً ضرایب لازم برای صورت نیوتنی تابع $p_{k+1}(x)$ با مراکز x_1، ...، x_{k+1} را به‌دست می‌دهد:

$$p_{k+1}(x) = \sum_{i=0}^{k+1} f[x_i, \ldots, x_{k+1}] \prod_{j=i+1}^{k+1} (x - x_j) \qquad (۱۴.۲)$$

□ **مثال ۵.۲:** الگوریتم ۴.۲ را برای مسئلهٔ مثال ۲.۲ و ۳.۲ به‌کار می‌بریم و از $x_0 = ۱$، $x_1 = ۴$، $x_2 = ۶$، و نیز از $x_3 = ۰$ استفاده می‌کنیم. در این مثال داریم $\bar{x} = ۳٫۵$. از آنجا خواهیم داشت $p_0(\bar{x}) = K[x_0] = ۱٫۵۷۰۹$ و $\psi_0(\bar{x}) = ۱$. سپس با $K[x_1] = ۱٫۵۷۲۷$، $K[x_0, x_1] = ۰٫۰۰۰۶$ را به‌دست می‌آوریم و با داشتن $\psi_1(\bar{x}) = (\bar{x} - x_0)\psi_0(\bar{x}) = ۲٫۵$ به‌دست می‌آوریم

$$p_1(\bar{x}) = ۱٫۵۷۰۹ + ۰٫۰۰۱۵ = ۱٫۵۷۲۴$$

با افزودن نقطهٔ $x_2 = ۶$، داریم $K[x_2] = ۱٫۵۷۵۱$، در نتیجه $K[x_1, x_2] = ۰٫۰۰۱۲$، $K[x_0, x_1, x_2] = ۰٫۰۰۰۱۲$، بنابراین چون $\psi_2(\bar{x}) = (۲٫۵)(-۰٫۵) = -۱٫۲۵$، لذا

$$p_2(\bar{x}) = ۱٫۵۷۲۴ - ۰٫۰۰۰۱۵ = ۱٫۵۷۲۲۵$$

که همان عددی‌است که قبلا درمثال ۳.۲ محاسبه‌شده بود. به‌منظور بررسی خطا دراین‌تقریب برای $K(۳٫۵)$، نقطهٔ $x_3 = ۰$ را به‌نقاط موجود اضافه می‌کنیم. با $K[x_3] = ۱٫۵۷۰۸$، اعداد $K[x_2, x_3] = ۰٫۰۰۰۷۱۷$ و $K[x_1, x_2, x_3] = -۰٫۰۰۰۱۲۱$ و $K[x_0, x_1, x_2, x_3] = -۰٫۰۰۰۰۰۱$ را محاسبه می‌کنیم و به‌ازای $\psi_3(\bar{x}) = (-۱٫۲۵)(-۲٫۵) = ۳٫۱۲۵$ مقدار $p_3(\bar{x})$ را به‌دست می‌آوریم

$$p_۳(\bar{x}) = ۱٫۵۷۲۲۵ - ۰٫۰۰۰۰۰۳$$

مقدار فوق نشان‌دهندهٔ این است که ۱٫۵۷۲۲ و یا ۱٫۵۷۲۳ احتمالا مقدار $K(۳٫۵)$ در چارچوب دقت مقادیرداده شده برای مقدار $K(x)$ است.

اگر این محاسبات با دست انجام گیرد به‌آسانی می‌تواند به‌صورت جدولی که در شکل ۲.۲ نشان داده شده تنظیم شود، ودرضمن چگونگی تولید تدریجی تفاضل‌منقسم بر اثر الگوریتم (۴.۲) را نیز نشان می‌دهد. □

در زیر فهرستی از یك **تابع فورترن** به‌نام تابع TABLE (تابع جدولی) داده شده است، که برای درونیابی در یك جدول مفروض از طول وعرضهای X(I), F(I) به‌ازای I = ۱, ..., NTABLE، با F(I) = f(X(I)) و X(۱) < X(۲) < ... در پیداکردن یك تقریب خوب از $f(x)$ در نقطهٔ x = XBAR از الگوریتم(۴.۲) مورد استفاده قرار می‌گیرد. این برنامه $p_۰$(XBAR) و $p_۱$(XBAR) و ... را تا برقراری شرط

$$|(p_k(\text{XBAR}) - p_{k-۱}(\text{XBAR})| \leqslant \text{TOL}$$

(TOL حد خطای مطلوب) یا تا برقراری $k+۱ = \min(۲۰, \text{NTABLE})$ تولید می‌کند و آنگاه p_k(XBAR) را باز می‌گرداند. دنبالهٔ $x_۰$، $x_۱$، $x_۲$، ... از نقاط درونیابی جدولی X(۱)، X(۲)، ...، X(NTABLE) به‌شرح زیر انتخاب می‌شوند. اگر X(I) < XBAR ⩽ X(I+۱) آنگاه $x_۰$ = X(I+۱) و $x_۱$ = X(I) و $x_۲$ = X(I+۲) و $x_۳$ = X(I−۱) و... انتخاب می‌شوند، جزنزدیك به‌ابتدا و انتهای جدول داده شده، که سرانجام فقط نقاط سمت راست یا سمت چپ XBAR به‌کار گرفته می‌شوند. برای حفاظت برنامه (و استفاده‌کننده!) در قبال انتخاب مقدار غیرمعقول برای XTOL، برنامه باید طوری اصلاح شود که تعیین کند که هنگام افزایش k، چه موقع تفاضلهای متوالی $|p_{k+۱}(\text{XBAR}) - p_k(\text{XBAR})|$ شروع به‌افزایش می‌کنند و آیا این افزایش صورت می‌گیرد یا نه. به‌تمرین ۴.۲–۱ نیز نگاه کنید).

k	$p_k(\bar{x})$	$\psi_k(\bar{x})$	x_k	K[ ]	K[,]	K[,,]	K[,,,]
0	1.5709	1.	1	1.5709			
	+ 15				0.0006		
1	1.5724	2.5	4	1.5727		0.00012	
	− 15				0.0012		−0.000001
2	1.57225	−1.25	6	1.5751		0.000121	
	− 3				0.000717		
3	1.572247	3.125	0	1.5708			

شکل ۲.۲

زیربرنامهٔ فورترن برای درونیابی در یك تابع TABLE (تابع جدولی)

```
      REAL FUNCTION TABLE (XBAR, X, F, NTABLE, TOL, IFLAG )
C  RETURNS AN INTERPOLATED VALUE  TABLE  AT  XBAR  FOR THE FUNCTION
C  TABULATED AS (X(I),F(I)), I=1,...,NTABLE.
      INTEGER IFLAG,NTABLE,   J,NEXT,NEXTL,NEXTR
      REAL F(NTABLE),TOL,X(NTABLE),XBAR,   A(20),ERROR,PSIK,XK(20)
C******  I N P U T  ******
C  XBAR  POINT AT WHICH TO INTERPOLATE .
C  X(I), F(I), I=1,...,NTABLE  CONTAINS THE FUNCTION TABLE .
C     A S S U M P T I O N ...  X IS ASSUMED TO BE INCREASING.
C  NTABLE  NUMBER OF ENTRIES IN FUNCTION TABLE.
C  TOL  DESIRED ERROR BOUND .
C******  O U T P U T  ******
C  TABLE  THE INTERPOLATED FUNCTION VALUE .
C  IFLAG  AN INTEGER,
C     = 1 , SUCCESSFUL EXECUTION ,
C     = 2 , UNABLE TO ACHIEVE DESIRED ERROR IN 20 STEPS,
C     = 3 , XBAR LIES OUTSIDE OF TABLE RANGE. CONSTANT EXTRAPOLATION IS
C         USED.
C******  M E T H O D  ******
C  A SEQUENCE OF POLYNOMIAL INTERPOLANTS OF INCREASING DEGREE IS FORMED
C  USING TABLE ENTRIES ALWAYS AS CLOSE TO  XBAR  AS POSSIBLE. EACH IN-
C  TERPOLATED VALUE IS OBTAINED FROM THE PRECEDING ONE BY ADDITION OF A
C  CORRECTION TERM (AS IN THE NEWTON FORMULA). THE PROCESS TERMINATES
C  WHEN THIS CORRECTION IS LESS THAN  TOL  OR, ELSE, AFTER  20  STEPS.
C
C            LOCATE  XBAR  IN THE X-ARRAY.
      IF (XBAR .GE. X(1) .AND. XBAR .LE. X(NTABLE)) THEN
         DO 10 NEXT=2,NTABLE
            IF (XBAR .LE. X(NEXT))       GO TO 12
   10    CONTINUE
      END IF
      IF (XBAR .LT. X(1)) THEN
         TABLE = F(1)
      ELSE
         TABLE = F(NTABLE)
      END IF
      PRINT 610,XBAR
  610 FORMAT(E16.7,' NOT IN TABLE RANGE.')
      IFLAG = 3
                                          RETURN
   12 XK(1) = X(NEXT)
      NEXTL = NEXT-1
      NEXTR = NEXT+1
      A(1) = F(NEXT)
      TABLE = A(1)
      PSIK = 1.
C        USE ALGORITHM 2.4, WITH THE NEXT XK ALWAYS THE TABLE
C        ENTRY NEAREST   XBAR  OF THOSE NOT YET USED.
      KP1MAX = MIN(20,NTABLE)
      DO 20 KP1=2,KP1MAX
         IF (NEXTL .EQ. 0) THEN
            NEXT = NEXTR
            NEXTR = NEXTR+1
         ELSE IF (NEXTR .GT. NTABLE) THEN
            NEXT = NEXTL
            NEXTL = NEXTL-1
         ELSE IF (XBAR - X(NEXTL) .GT. X(NEXTR) - XBAR) THEN
            NEXT = NEXTR
            NEXTR = NEXTR+1
         ELSE
            NEXT = NEXTL
            NEXTL = NEXTL-1
         END IF
         XK(KP1) = X(NEXT)
         A(KP1) = F(NEXT)
         DO 13 J=KP1-1,1,-1
            A(J) = (A(J+1) - A(J))/(XK(KP1) - XK(J))
   13    CONTINUE
C         FOR I=1,...,KP1,  A(I)  NOW CONTAINS THE DIV.DIFF. OF
C         F(X) OF ORDER  K-I  AT  XK(I),...,XK(KP1).
         PSIK = PSIK*(XBAR - XK(KP1-1))
         ERROR = A(1)*PSIK
```

```
C                TEMPORARY PRINTOUT
          PRINT 613,KP1,XK(KP1),TABLE,ERROR
 613      FORMAT(I10,3E17.7)
       TABLE = TABLE + ERROR
       IF (ABS(ERROR) .LE. TOL)  THEN
          IFLAG = 1
                                          RETURN
       END IF
  20 CONTINUE
     PRINT 620,KP1MAX
 620 FORMAT(' NO CONVERGENCE IN ',I2,' STEPS.')
     IFLAG = 2
                                          RETURN
     END
```

تمرین

۱-۴۰۲ تابع TABLE (تابع جدولی) فورترن که در متن داده شده است به‌محض اینکه شرط $|p_{k+1}(\text{XBAR})-p_k(\text{XBAR})|\leqslant \text{TOL}$ برقرار گردد، تمام می‌شود. با مثالهای زیر نشان دهید که این امر قرار گرفتن مقدار $p_{k+1}(\text{XBAR})$ را، که به‌وسیلهٔ TABLE برگردانده می‌شود، در محدودهٔ TOL از عدد مطلوب $f(\text{XBAR})$ تضمین نمی‌کند.

(الف) $\text{X(I)}=-10,\ \text{X(I+1)}=10,\ \text{XBAR}=0,\ \text{TOL}=10^{-5}$ به‌ازای $f(x)=x^2$؛ I

(ب) $\text{X(I)}=-100, \text{X(I+1)}=0, \text{X(I+2)}=100$ به‌ازای $f(x)=x^3$؛ I
$\text{XBAR}=-50,\ \text{TOL}=10^{-5}$.

۲-۴۰۲ درونیابی خطی بارسته[1] برپایهٔ ملاحظات زیر که به‌نویل[2] منسوب است استوار می‌باشد: بسجمله‌ای از درجهٔ نابیشتر از $j-i$ را که $f(x)$ را به‌ازای $i\leqslant j$ در نقاط $x_i, x_{i+1}, \ldots, x_j$ درونیابی می‌کند با $p_{i,j}(x)$ نشان می‌دهیم. پس داریم

$$p_{ij}(x)=\frac{x-x_i}{x_j-x_i}\,p_{i+1,j-1}(x)+\frac{x_j-x}{x_j-x_i}\,p_{i,j-1}(x)$$

صحت این همانی[3] را تحقیق کنید. [راهنمایی: ما از این همانی در بخش ۳.۲ استفاده کرده‌ایم. به‌معادلهٔ (۱۳.۲) نگاه کنید.]

۳-۴۰۲ درونیابی خطی بارسته (ادامه) . همانی نویل، که در تمرین ۲-۴۰۲ اثبات شد این امکان را به‌دست می‌دهد که درایه‌های جدول مثلثی زیر

$$\begin{array}{lllll}
f(x_0)=p_{00}(\bar{x}) & & & & \\
 & p_{01}(\bar{x}) & & & \\
f(x_1)=p_{11}(\bar{x}) & & p_{02}(\bar{x}) & & \\
 & p_{12}(\bar{x}) & & & \\
f(x_2)=p_{22}(\bar{x}) & & \vdots & ; & p_{0n}(\bar{x}) \\
 & \vdots & p_{n-2,n}(\bar{x}) & & \\
\vdots \quad \vdots & p_{n-1,n}(\bar{x}) & & & \\
f(x_n)=p_{nn}(\bar{x}) & & & &
\end{array}$$

1. Iterated Linear Interpolation 2. Neville 3. identity

را ستون به‌ستون تولید کنیم، این عمل به‌وسیلهٔ عملیاتی نظیر درونیابی خطی انجام می‌گیرد و در نهایت عدد موردنظر $p_{0n}(\bar{x})$ را، که با مقدار بسجمله‌ای درونیاب در $\bar{x}$ برابر و در $n+1$ نقطهٔ $x_0, \ldots, x_n$ بر $f(x)$ منطبق است به‌دست می‌دهد. این الگوریتم، **الگوریتم نویل** نامیده می‌شود. تفاوت **الگوریتم ایتکن**[1] با الگوریتم نویل در این است که در الگوریتم ایتکن یک جدول مثلثی تولید می‌شود که ستون j ام آن شامل اعداد

$$p_{0,1,\ldots,j,j+1}(\bar{x})$$

$$p_{0,1,\ldots,j,j+2}(\bar{x})$$

$$p_{0,1,\ldots,j,n}(\bar{x})$$

می‌باشد که در آن $p_{0,1,\ldots,j,r}(x)$ ($r > j$ به‌ازای) یک بسجمله‌ای است از درجهٔ نابیشتر از $j+1$، منطبق بر $f(x)$ در نقاط $x_0, x_1, \ldots, x_j$ و x_r.

با شمارش تعداد عملیات نشان دهید که الگوریتم نویل پرهزینه‌تر از الگوریتم ۴.۲ است. (همچنین باید توجه کرد که الگوریتم ۴.۲، بدون هزینهٔ اضافی، صورت نیوتنی بسجمله‌ای درونیاب را تولید می‌کند که برای محاسبهٔ مقادیر بعدی تابع در نقاط دیگر مورد استفاده قرار می‌گیرد. درحالی که اطلاعات حاصل از الگوریتم نویل یا ایتکن برای محاسبهٔ مقادیر در نقاط دیگر کمکی نمی‌کنند.)

۴-۴۰۲ در **درونیابی معکوس**[2] در یک جدول، عدد $\bar{y}$ داده شده‌است و می‌خواهیم نقطهٔ $\bar{x}$ را طوری پیدا کنیم که داشته باشیم $f(\bar{x}) = \bar{y}$. در اینجا $f(x)$ یک تابع جدول‌بندی‌شده است. اگر $f(x)$ (پیوسته و) اکیداً یکنوای صعودی یا نزولی[3] باشد، همواره می‌توان این مسئله را بدین‌صورت حل کرد که جدول داده شدهٔ x_i، $f(x_i)$، $(i = 0, 1, 2, \ldots)$، را با اختیار کردن $y_i = f(x_i)$، $g(y_i) = x_i$، $(i = 0, 1, 2, \ldots)$، یک جدول y_i، $g(y_i)$، $(i = 0, 1, 2, \ldots)$، برای تابع معکوس $g(y) = f^{-1}(y) = x$ انگاشت و برای مقدار مجهول $g(\bar{y})$ در این جدول درونیابی کرد. از تابع **جدولی فورترن** برای پیدا کردن $\bar{x}$ به‌طوری که $\sin \bar{x} = 0.6$، استفاده کنید.

۵۰۲ خطای بسجمله‌ای درونیاب

گیریم $f(x)$ تابعی حقیقی مقدار در بازهٔ $I = [a, b]$، و $x_0, \ldots, x_n, x_{n+1}$ نقطهٔ متمایز از بازهٔ I باشند. اگر $p_n(x)$ یک بسجمله‌ای باشد از درجهٔ نابیشتر از n که $f(x)$ را در نقاط $x_0, \ldots, x_n$ درونیابی می‌کند، آنگاه خطای درونیابی در $p_n(x)$ عبارت است از:

---

1. Aitken's Algorithm
2. Inverse Interpolation
3. strictly monotone-increasing or-decreasing

$$e_n(x)=f(x)-p_n(x) \qquad (۱۵.۲)$$

اکنون فرض کنید که $\bar{x}$ نقطه‌ای غیر از x_0، ...، x_n باشد. اگر $p_{n+1}(x)$ بسجمله‌ایی باشد از درجهٔ نابیشتر از $(n+1)$ که $f(x)$ را در x_0، ...، x_n و در $\bar{x}$ درونیابی می‌کند، آنگاه $p_{n+1}(\bar{x})=f(\bar{x})$، درحالی که طبق رابطهٔ (۱۵.۲) داریم

$$p_{n+1}(x)=p_n(x)+f[x_0,\cdots,x_n,\bar{x}]\prod_{j=0}^{n}(x-x_j)$$

در نتیجه

$$f(\bar{x})=p_{n+1}(\bar{x})=p_n(\bar{x})+f[x_0,\cdots,x_n,\bar{x}]\prod_{j=0}^{n}(\bar{x}-x_j)$$

بنابراین به‌ازای جمیع نقاط $\bar{x}\neq x_0,\cdots,x_n$

$$e_n(\bar{x})=f[x_0,\cdots,x_n,\bar{x}]\prod_{j=0}^{n}(\bar{x}-x_j) \qquad (۱۶.۲)$$

رابطهٔ بالا نشان می‌دهد که خطا «مشابه جملهٔ بعدی» درصورت نیوتنی است.

بدون داشتن عدد $f(\bar{x})$ نمی‌توانیم سمت راست رابطهٔ (۱۶.۲) را محاسبه کنیم. اما به‌طوری که اکنون ثابت می‌کنیم، عدد $f[x_0,\cdots,x_n,\bar{x}]$ با $(n+1)$امین مشتق $f(x)$ ارتباط نزدیك دارد و با به‌کارگرفتن این اطلاعات اکنون می‌توانیم، $e_n(\bar{x})$ را برآورد کنیم.

قضیهٔ ۲.۲ گیریم $f(x)$ تابعی است حقیقی مقدار که در بازهٔ $[a,b]$ تعریف شده و kمرتبه مشتقپذیر است. اگر x_0، ...، x_k، $k+1$ نقطهٔ متمایز در $[a,b]$ باشند، آنگاه نقطه‌ای مانند $\xi\in(a,b)$ وجود دارد به‌طوری که رابطهٔ زیر برقرار است

$$f[x_0,\cdots,x_k]=\frac{f^{(k)}(\xi)}{k!} \qquad (۱۷.۲)$$

به‌ازای $k=1$، رابطهٔ فوق، همان قضیهٔ مقدارمیانگین برای مشتق است (به بخش ۷.۱ نگاه کنید). برای اثبات قضیهٔ بالا در حالت کلی، باید توجه داشت که تابع خطای $e_k(x)=f(x)-p_k(x)$ در $I=[a,b]$ دارای (حداقل) $k+1$ ریشهٔ متمایز x_0، ...، x_k است. بنابراین اگر $f(x)$، ودر نتیجه $e_k(x)$، kمرتبه روی (a,b) مشتقپذیر باشد از قضیهٔ رول[1] نتیجه می‌شود (به بخش ۷.۱ نگاه کنید) که $e'(x)$ حداقل k ریشه در (a,b) دارد و بنابراین حداقل $e''(x)$ دارای $k-1$ ریشه در (a,b) است، و اگر عمل را به‌همین ترتیب ادامه دهیم سرانجام به‌این نتیجه می‌رسیم که $e_k^{(k)}(x)$ حداقل یك ریشه در (a,b) دارد. گیریم ξ چنین ریشه‌ای باشد. بنابراین داریم:

1. Rolle

$$0 = e_k^{(k)}(\xi) = f^{(k)}(\xi) - p_k^{(k)}(\xi)$$

از طرف دیگر می‌دانیم که به‌ازای هر x

$$p_k^{(k)}(x) = f[x_0, \cdots, x_k]k!$$

زیرا بنابرتعریف، $f[x_0, \cdots, x_k]$ ضریب پیشرو $p_k(x)$ است و بدین‌ترتیب رابطهٔ (۱۷.۲) به‌دست خواهد آمد.

به‌فرض آنکه $a = \min_i x_i$، $b = \max_i x_i$، نتیجه می‌شود که نقطهٔ مجهول ξ در رابطهٔ (۱۷.۲) در بین x_iها واقع است.

اگر قضیهٔ ۲.۲ را برای (۱۶.۲) به‌کار بریم، قضیهٔ ۳.۲ به‌دست می‌آید.

قضیهٔ ۳.۲ گیریم $f(x)$ تابعی است حقیقی‌مقدار[1] که روی $[a, b]$ تعریف شده و در (a, b)، $n+1$ مرتبه مشتقپذیر است. اگر $p_n(x)$ یك بسجمله‌ای باشد از درجهٔ نابیشتر از n که $f(x)$ را در $n+1$ نقطهٔ متمایز x_0، ...، x_n و متعلق به‌بازهٔ $[a, b]$ درونیابی می‌کند، آنگاه به‌ازای جمیع مقادیر $\bar{x} \in [a, b]$ نقطه‌ای مانند $\xi(\bar{x}) \in (a, b)$ وجود دارد به‌طوری که

$$e_n(\bar{x}) = f(\bar{x}) - p_n(\bar{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{j=0}^{n} (\bar{x} - x_j) \qquad (۱۸.۲)$$

لازم است توجه کنیم که $\xi(\bar{x})$ به‌نقطهٔ $\bar{x}$، یعنی به‌جایی که برآورد خطا لازم است، بستگی دارد. این بستگی حتی نیاز ندارد که پیوسته باشد. از آنجا که در فصل ۷ به‌انتگرالگیری و مشتقگیری از $e_n(x)$ نسبت به x احتیاج داریم، معمولا برای چنین منظوری فرمول (۱۶.۲) ترجیح داده می‌شود. زیرا همان‌طوری که در بخش ۷.۲ نشان داده خواهد شد $f[x_0, \cdots, x_n, x]$ تابعی است خوشرفتار[2] از x.

فرمول خطای (۱۸.۲) از فایدهٔ عملی محدودی برخوردار است، زیرا درحالت کلی به‌ندرت از مقدار $f^{(n+1)}(x)$ آگاهیم. تقریباً هیچوقت جای نقطهٔ ξ را نمی‌دانیم. اما هرگاه یك کران[3] برای $|f^{(n+1)}(x)|$ درتمامی بازهٔ (a, b) در دست باشد، آنگاه می‌توانیم از (۱۸.۲) برای پیداکردن کران خطای بسجمله‌ای درونیاب در آن بازه استفاده کنیم.

□ **مثال ۶.۲:** کران خطایی در درونیابی خطی پیدا کنید.

بسجملهٔ خطیی که $f(x)$ را در x_0 و x_1 درونیابی می‌کند به‌شرح زیر تعریف شده است

$$p_1(x) = f(x_0) + f[x_0, x_1](x - x_0) = \frac{(x_1 - x)f(x_0) + (x - x_0)f(x_1)}{x_1 - x_0}$$

---

1. real valued function
2. well-behaved function
3. bound

پس معادلهٔ (۱۸۰۲) فرمول خطای

$$f(\bar{x})-p_1(\bar{x})=(\bar{x}-x_0)(\bar{x}-x_1)\frac{f''(\xi)}{2!}$$

را که در آن ξ بستگی به $\bar{x}$ دارد، به‌دست می‌دهد. اگر $\bar{x}$ نقطه‌ای بین x_0 و x_1 باشد، ξ نیز بین x_0 و x_1 خواهد بود. بنابراین اگر بدانیم که در $[x_0, x_1]$، نامساوی $|f''(x)|\leqslant M$ برقرار است، آنگاه خواهیم داشت

$$|f(\bar{x})-p_1(\bar{x})|\leqslant|(\bar{x}-x_0)(\bar{x}-x_1)|\frac{M}{2}$$

ماکسیمم مقدار $|(\bar{x}-x_0)(\bar{x}-x_1)|$ به ازای $\bar{x}\in[x_0, x_1]$ در $\bar{x}=(x_0+x_1)/2$ حاصل می‌شود و مقدار آن برابر $(x_1-x_0)^2/4$ است. در نتیجه به‌ازای هر $\bar{x}\in[x_0, x_1]$، داریم:

$$|f(\bar{x})-p_1(\bar{x})|\leqslant(x_1-x_0)^2\frac{M}{8}$$ □

□ **مثال ۷۰۲:** بازهٔ h را در جدولی که مقادیر تابع $f(x)=\sqrt{x}$ را بین ۱ و ۲ در فواصل مساوی به‌دست می‌دهد، طوری تعیین کنید که درونیابی با بسجمله‌ای درجهٔ دوم در این جدول دقت عمل مطلوب را نتیجه دهد.

بنابر فرض، این جدول با شرط $x_i=1+ih$ به‌ازای $i=0, \ldots, N$ و $N=(2-1)/h$ شامل $f(x_i)$ است. اگر $\bar{x}\in[x_i, x_{i+1}]$، آنگاه $f(\bar{x})$ را با $p_2(\bar{x})$ تقریب می‌زنیم که $p_2(x)$ یک بسجمله‌ای است از درجهٔ دوم که $f(x)$ را در x_{i-1}، x_i، x_{i+1} درونیابی می‌کند. پس به‌موجب ۱۸۰۲، به‌ازای نقطه‌ای مانند ξ در (x_{i-1}, x_{i+1})، خطا برابر است با

$$f(\bar{x})-p_2(\bar{x})=(\bar{x}-x_{i-1})(\bar{x}-x_i)(\bar{x}-x_{i+1})\frac{f'''(\xi)}{3!}$$

از آنجا که مقدار ξ را نمی‌دانیم فقط می‌توانیم $f'''(\xi)$ را برآورد کنیم

$$|f'''(\xi)|\leqslant\max_{1\leqslant x\leqslant 2}|f'''(x)|$$

طبق محاسبه داریم: $f'''(x)=(3/8)x^{-5/2}$، بنابراین $|f'''(\xi)|\leqslant 3/8$. بعلاوه با استفاده از تغییر متغیر خطی $y=x-x_i$ داریم:

$$\max_{x\in[x_{i-1},x_{i+1}]}|(x-x_{i-1})(x-x_i)(x-x_{i+1})|$$
$$=\max_{y\in[-h,h]}|(y+h)y(y-h)|$$
$$=\max_{y\in[-h,h]}|y(y^2-h^2)|$$

از آنجا که مقدار تابع $\psi(y)=y(y^2-h^2)$ در $y=-h$ و $y=h$ برابر صفر می‌شود، ماکسیمم $|\psi(y)|$ در بازهٔ $[-h,h]$ باید در یکی از نقاط فرین[1] $\psi(y)$ پیش آید. این نقاط فرین را با حل معادلهٔ $\psi'(y)=3y^2-h^2=0$ می‌توان پیدا کرد. که از این معادله $y=\pm h/\sqrt{3}$ به‌دست می‌آید. بنابراین

$$\max_{x\in[x_{i-1},x_{i+1}]}|(x-x_{i-1})(x-x_i)(x-x_{i+1})|=\frac{2h^3}{3\sqrt{3}}$$

اکنون اگر $p_2(x)$ به عنوان یك بسجملهٔ درجهٔ دومی که $f(x)=\sqrt{x}$ را در سه نقطهٔ جدولی نزدیکتر به $\bar{x}$ درونیابی می‌کند انتخاب شود، مطمئن هستیم که به‌ازای هر $\bar{x}\in[1,2]$ خواهیم داشت

$$|f(\bar{x})-p_2(\bar{x})|\leqslant\frac{(2h^3/[3\sqrt{3}])(3/8)}{6}=\frac{h^3}{24\sqrt{3}}$$

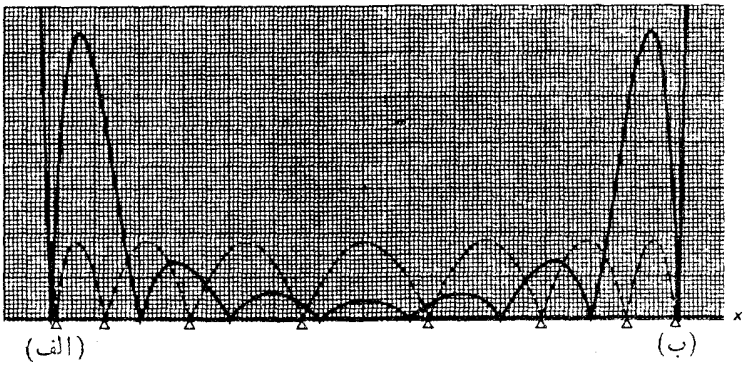
اگر بخواهیم با روش فوق، دقتی تا هفت رقم به‌دست آوریم می‌بایستی h را طوری انتخاب کنیم که

$$\frac{h^3}{24\sqrt{3}}<5\times10^{-8}$$

و در نتیجه خواهیم داشت $h\approx0.0128$ یا $N\approx79$. □

البته تابع $\psi_{n+1}(x)=\Pi_{j=0}^{n}(x-x_j)$ که در (۱۸.۲) پیدا می‌شود، قویاً به‌تعیین جای نقاط درونیابی بستگی دارد. این امکان وجود دارد که نقاط مذکور به‌ازای مقدار مفروض n در بازهٔ مفروض $a\leqslant x\leqslant b$ طوری انتخاب شوند که مقدار $\max|\psi_{n+1}(x)|$ درحد امکان کوچك شود. این انتخاب نقاط، که نقاط چبیشف[2] نامیده می‌شوند، مشروحاً در بخش ۱.۶ موردبحث قرار خواهد گرفت. درصورتی که نقاط درونیابی با فواصل مساوی یا انتخاب معمولی صورت گیرند، هنگامی که از وسط بازه به‌طرف راست آن حرکت می‌کنیم ماکسیمم‌های موضعی[3] $|\psi_{n+1}(x)|$ افزایش می‌یابند، و این افزایش با افزایش n محسوستر می‌شود (به‌شکل ۳.۲ نگاه کنید). به‌موجب (۱۸.۲) مناسبتر است (حداقل وقتی که

1. Extrema 2. Chebyshev 3. Local maxima



شکل ۳.۲ تابع $|\psi_{n+1}(x)|$ به‌ازای $n=8$ و (الف) نقاط درونیابی با فواصل مساوی (خط پر)؛ (ب) نقاط چبیشف برای همان بازه (خط‌چین)

داده‌هایی با فواصل یکنوا درونیابی می‌شوند) که از بسجمله‌ای درونیاب فقط نزدیک به‌وسط نقاط داده شده استفاده شود. با نزدیک شدن به انتهای سمت چپ یا انتهای سمت راست نقاط داده شده، درونیاب کمتر مورد اطمینان است. البته بیرون از فاصلهٔ موردنظر درونیابی این عدم اطمینان بازهم بیشتر می‌شود. این عمل را برونیابی[1] نامند و فقط باید این عمل با احتیاط زیاد انجام گیرد.

تمرین

۱-۵.۲ جدولی از مقادیر $\cos x$ لازم است، تا درونیابی خطی به‌ازای هرمقدار از x در $[0, \pi]$، دقتی تا شش‌رقم به‌دست دهد. به‌فرض اینکه نقاط جدولی با فواصل مساوی باشند، مینیمم تعداد درایه‌های لازم در جدول چقدر است؟

۲-۵.۲ تابعی که با ضابطهٔ

$$f(x)=\int_0^x \sin s^2\, ds$$

تعریف شده‌است، برای مقادیر متساوی‌الفاصلهٔ x به‌ازای $h=0.1$ جدول‌بندی شده است. درصورتی که به‌ازای هر نقطهٔ $\bar{x}$ از بازهٔ $[0, \pi/2]$ مقدار $f(\bar{x})$ به‌وسیلهٔ درونیابی درجهٔ سوم محاسبه شود، ماکسیمم خطایی که با آن مواجه می‌شویم چقدر است؟

۳-۵.۲ ثابت کنید که اگر مقادیر $f(x_0)$، ...، $f(x_n)$ تنها اطلاعات ما دربارهٔ تابع $f(x)$ باشند، آنگاه دربارهٔ خطای $e_n(\bar{x})=f(\bar{x})-p_n(\bar{x})$ در نقطهٔ $\bar{x}\neq x_0, \ldots, x_n$

1. extrapolation

هیچ‌گونه اظهارنظری نمی‌توانیم بکنیم، یعنی خطا ممکن است «بسیار بزرگ» یا «بسیار کوچك» باشد. [راهنمایی: درونیابی تابع $f(x)=K(x-x_0)\cdots(x-x_n)$ را که در آن K ثابت مجهولی است در نقاط x_0، x_1، ...، x_n درنظر بگیرید.] این موضوع چه تأثیری روی برنامه‌هایی نظیر **تابع جدولی**[1] در بخش ۴.۲ یا الگوریتم ۴.۲ دارد؟

۴-۵.۲ رابطهٔ (۱۸.۲) را برای به‌دست آوردن کران پایین در خطای درونیابی $|f(\bar{x})-p_n(\bar{x})|$ به‌کار برید وقتی‌که $f(x)=\ln x$، $n=3$، $x_0=1$، $x_1=4/3$، $x_2=5/3$، $x_3=2$ و $\bar{x}=3/2$.

۶.۲ درونیابی یك تابع جدولی بر اساس نقاط با فواصل مساوی

در بیشتر محاسبات مهندسی و علمی، توابعی مانند $\sin x$، e^x، $J_n(x)$، $\operatorname{erf}(x)$ و غیره مورد استفاده قرار می‌گیرند که تعریف آنها به‌وسیلهٔ یك‌سری نامتناهی، یا به‌صورت جواب یك معادلهٔ دیفرانسیل، یا به‌وسیلهٔ فرایندهای مشابهی که مستلزم حد می‌باشند، و بنابراین در حالت کلی نمی‌توانند در تعداد مراحل متناهی به‌دست آیند، صورت می‌گیرد. مؤسسات کامپیوتری برنامه‌هایی برای به‌دست آوردن این‌گونه توابع در اختیار دارند که به‌کمك بسجمله‌ایها یا نسبت آنها تقریبهایی برای توابع فوق به‌دست می‌آورند. اما پیش‌از پیدایش کامپیوترهای با سرعت زیاد، تنها وسیلهٔ استفاده ازچنین توابعی درمحاسبات، تابع جدولی بوده است. جدولی از این‌گونه، شامل مقادیر تابع $f(x_i)$ برای چند نقطهٔ x_0، ...، x_n است و استفاده‌کننده بایستی با درونیابی (که معنی تحت‌اللفظی آن «پرکردن رخنه‌ها و صاف کردن» ولذا «اثبات نادرستی» نیزهست) مقادیرمفروض، مقدار $f(x)$ را در نقطه‌ای که درجدول موجود نیست به‌دست‌آورد. درونیابی بسجمله‌ای اصلا برای تسهیل این‌روند گسترش پیداکرده‌است. ازآنجاکه معمولا در چنین جداولی $f(x)$ به‌صورت یك دنبالهٔ رو‌به‌افزایشی ازنقاط با فواصل مساوی داده شده‌اند، در محاسبهٔ بسجمله‌ای درونیاب ساده‌سازیهایی می‌تواند انجام گیردکه دراین قسمت موردبحث قرارخواهیم داد.

در سراسر این بخش، فرض براین است‌که $f(x)$ **به‌ازای** $x=a(h)b$ **جدولبندی** شده‌است، یعنی اعداد $f(x_i)$ به‌ازای $i=0,\ldots,N$ دراختیار ما هستند. دراینجا داریم

$$x_i=a+ih \qquad i=0,\ldots,N \quad \text{با} \quad N=\frac{b-a}{h} \qquad (۱۹.۲)$$

بهتراست‌که یك تبدیل متغیرخطی به‌صورت زیر را دخالت دهیم

$$x=x(s)=x_0+sh \qquad \text{به‌طوری‌که} \qquad s=s(x)=\frac{x-x_0}{h} \qquad (۲۰.۲)$$

---

1. Function Table

و یا به‌اختصار

$$f(x)=f(x_0+sh)=f_s \qquad (۲۱.۲)$$

این امرموجب مقایسه‌کردن این‌حالت با حالتی می‌شودکه درآن مقدار $f(x)$ در اولین $N+1$ عدد صحیح غیرمنفی‌معلوم‌است،ودرنتیجه موجب ساده‌ترشدن نمادگذاری[1]می‌شود. باید توجه داشت‌که تبدیل متغیر خطی (۲۰.۲) بسجمله‌ای درجهٔ nام برحسب x را به‌بسجمله‌ای درجهٔ nام برحسب s بدل می‌کند.

در این حالت محاسبهٔ یك بسجمله‌ای از درجهٔ نابیشتر از nکه $f(x)$ را در x_k، ... ، x_{k+n} درونیابی کند، به‌محاسبهٔ جدول تفاضل‌منقسم نیازی نیست. بلکه بهتراست یك جدول تفاضل‌را محاسبه‌کنیم. برای توضیح این‌مطلب،تفاضل پیشرو[2] را وارد می‌کنیم

$$\Delta^i f_s=\begin{cases} f_s & i=0 \\ \Delta(\Delta^{i-1}f_s)=\Delta^{i-1}f_{s+1}-\Delta^{i-1}f_s & i>0 \end{cases} \qquad (۲۲.۲)$$

تفاضل پیشرو و تفاضل منقسم به‌شرح زیربا هم در ارتباط‌اند.

لم ۳.۲ به‌ازای جمیع مقادیر $i\geqslant 0$

$$f[x_k,\ldots,x_{k+i}]=\frac{1}{i!h^i}\Delta^i f_k \qquad (۲۳.۲)$$

از آنجاکه هردو طرف رابطهٔ (۲۳.۲) به‌استقرا نسبت به i تعریف شده‌اند، اثبات لم ۳.۲ باید بـا استقرا انجام پذیرد. بـه‌ازای $i=0$، معادلهٔ (۲۳.۲) فقط مدعی صحت قرارداد

$$f[x_k]=f(x_k)=f_k=\Delta^0 f_k$$

و بنابراین درست‌است. با فرض اینکه (۲۳.۲) به‌ازای $i=n\geqslant 0$ برقرار باشد، داریم

$$f[x_k,\ldots,x_{k+n+1}]=\frac{f[x_{k+1},\ldots,x_{k+n+1}]-f[x_k,\ldots,x_{k+n}]}{x_{k+n+1}-x_k}$$

$$=\frac{\Delta^n f_{k+1}/(n!h^n)-\Delta^n f_k/(n!h^n)}{(n+1)h}$$

$$=\frac{\Delta^{n+1}f_k}{(n+1)!h^{n+1}}$$

---

1. notation 2. forward difference

که نشان می‌دهد (۲۳.۲) به‌ازای $i=n+1$ نیز برقراراست.

بدین‌ترتیب بسجمله‌ای از درجهٔ نابیشتر از n که $f(x)$ را درنقاط $x_k, \ldots, x_{k+n}$ درونیابی می‌کند به‌صورت زیردرمی‌آید

$$p_n(x)=\sum_{i=0}^{n}\frac{1}{i!h^i}\Delta^i f_k\prod_{j=0}^{i-1}(x-x_{k+j}) \qquad (۲۴.۲)$$

و برحسب s، داریم

$$x-x_{k+j}=x_0+sh-[x_0+(k+j)h]=(s-k-j)h$$

بنابراین

$$p_n(x)=p_n(x_0+sh)=\sum_{i=0}^{n}\Delta^i f_k\prod_{j=0}^{i-1}\frac{s-k-j}{j+1}$$

با یك تعریف نهایی بازهم این عبارت ساده‌ترمی‌شود. به‌ازای مقدار حقیقی y وعددصحیح نامنفی i، یك **تابع دوجمله‌ای**[1] را به‌طریق زیر تعریف می‌کنیم

$$\binom{y}{i}=\begin{cases}1 & i=0\\ \prod_{j=0}^{i-1}\frac{y-j}{j+1}=\frac{(y)(y-1)\cdots(y-i+1)}{1\cdot 2\cdots i} & i>0\end{cases} \qquad (۲۵.۲)$$

واژهٔ «دوجمله‌ای» قابل توجیه است، زیرا وقتی y یك عدد صحیح باشد، (۲۵.۲) تنها ضریب دوجمله‌ای $\binom{y}{i}$ است. بدین‌ترتیب معادلهٔ (۲۴.۲) به‌شکل سادهٔ

$$\begin{aligned}p_n(x_0+sh)&=\sum_{i=0}^{n}\Delta^i f_k\binom{s-k}{i}\\ &=f_k+(s-k)\Delta f_k+\frac{(s-k)(s-k-1)}{2}\Delta^2 f_k\\ &\quad+\cdots+\frac{(s-k)\cdots(s-k-n+1)}{n!}\Delta^n f_k \qquad (۲۶.۲)\end{aligned}$$

نوشته می‌شودکه **فرمول نیوتنی تفاضل پیشرو** برای بسجمله‌ای از درجهٔ نابیشتر از n که $f(x)$ را در x_k+ih، $i=0,\ldots,n$ درونیابی می‌کند نامیده می‌شود.

اگر در (۲۶.۲)، k را مساوی صفرقراردهیم، که این یك امر معمولی است، فرمول نیوتنی تفاضل پیشرو به‌صورت:

---

1. binomial

$$\boxed{p_n(x_0+sh)=\sum_{i=0}^{n}\Delta^i f_0\binom{s}{i}} \qquad (۲۷.۲)$$

درمی‌آید. اگر s یك عدد صحیح بین صفر و n باشد، آنگاه فرمول فوق به‌فرمول زیر بدل می‌شود

$$f_s=p_n(x_0+sh)=\sum_{j=0}^{n}\binom{s}{j}\Delta^j f_0=\sum_{j=0}^{s}\binom{s}{j}\Delta^j f_0 \qquad (۲۸.۲)$$

شباهت عجیب این فرمول با قضیهٔ دوجمله‌ای

$$(a+b)^s=\sum_{j=0}^{s}\binom{s}{j}a^j b^{s-j}$$

تصادفی نیست. زیرا با معرفی عملگر **تعویض پیشرو**[1]

$$Ef_i=f_{i+1} \qquad \text{به‌ازای جمیع مقادیر } i,$$

می‌توانیم بنویسیم $\Delta+1=E$، یعنی

$$(\Delta+1)f_i=(f_{i+1}-f_i)+f_i=f_{i+1}=Ef_i$$

بنابراین

$$f_s=E^s f_0=(\Delta+1)^s f_0=\sum_{j=0}^{s}\binom{s}{j}\Delta^j 1^{s-j} f_0=\sum_{j=0}^{s}\binom{s}{j}\Delta^j f_0$$

که همان رابطهٔ (۲۸.۲) است.

ما، دراینجا، ازدست زدن به‌حساب عملی وسیع برای تفاضلات، بر اساس فرمولهایی نظیر $E=\Delta+1$، خودداری می‌کنیم، لیکن فرمولی را که بلاواسطه مورد استفاده قرار می‌گیرد، از آن نتیجه می‌گیریم. از آنجا که $\Delta=E-1$، از قضیهٔ دوجمله‌ای نتیجه می‌گیریم

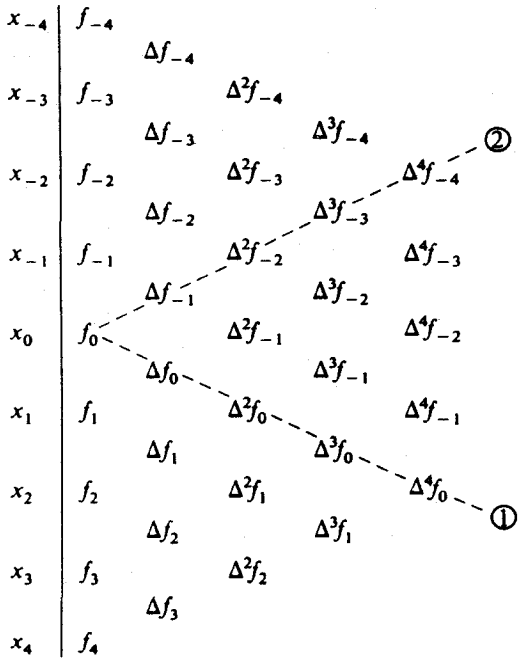
$$\Delta^s=(E-1)^s=\sum_{j=0}^{s}\binom{s}{j}E^j(-1)^{s-j}$$

یا

$$\Delta^s f_i=\sum_{j=0}^{s}(-1)^{s-j}\binom{s}{j}f_{i+j} \qquad (۲۹.۲)$$

ضرایب $\Delta^i f_k$ در (۲۶.۲) به‌آسانی از جدول تفاضلهای (پیشرو) برای $f(x)$ به‌دست

1. forward-shift



شکل ۴.۲ جدول تفاضل-پیشرو

می‌آید. این جدول درشکل ۴.۲ نشان داده شده است. به‌موجب معادلهٔ (۲۲.۲) هر درایه تنها تفاضل بین درایهٔ سمت چپ پایینی و درایهٔ سمت چپ بالایی‌است. تفاضلهایی‌که در (۲۷.۲) ظاهر می‌شوند، درطول قطری قرار دارندکه درشکل ۴.۲ با علامت ① مشخص شده‌است.

جدولهای تفاضل برای بررسی همواری یك تابع مندرج در جدول، پیداکردن خطاهای منفرد وتعیین درجهٔ بسجمله‌ای درونیاب مناسب برای جدول موردنظربه‌کارمی‌رود. این نکات را با مثال زیر روشن می‌کنیم.

□ **مثال ۸.۲:** از کتابی در مختصات درون-سیاره‌ای، مختص x مریخ را در دستگاه مختصات خورشید-مرکزی درتاریخهای داده شده، به‌طور غلط (به‌منظور نشان‌دادن‌مطلبی) برداشته‌ایم. این‌مختصات درفواصل‌ده روزه داده شده‌اند وبه‌توسط ستاره‌شناسان با وسایل مختلف به‌دست آمده‌اند. درشکل ۵.۲، یك جدول تفاضل (پیشرو) برای این داده‌ها درست کرده‌ایم.

در سه ستون اول تفاضلها، علامتهای اعداد هرستون یکی هستند، بنابراین دوستون اول تفاضلها یکنوا هستند. از ستون سوم به‌بعد، تفاضلها رفتارنوسانی محسوسی نشان

t	$x = f(t)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$	$\Delta^5 f$	$\Delta^6 f$
1,250.5	1.39140						
		−1444					
1,260.5	1.37696		−1469				
		−2913		55			
1,270.5	1.34783		−1414		−3		
		−4327		52		97	
1,280.5	1.30456		−1362		94		−302
		−5689		146		−205	
1,290.5	1.24767		−1216		−111		408
		−6905		35		203	
1,300.5	1.17862		−1181		92		−311
		−8086		127		−108	
1,310.5	1.09776		−1054		−16		128
		−9140		111		20	
1,320.5	1.00636		−943		4		
		−10083		115			
1,330.5	0.90553		−828				
		−10911					
1,340.5	0.79642						

شکل ۵۰۲ x-مختص مریخ، در دستگاه‌استوایی خورشیدمرکزی (تا حدی همراه با خطا)

می‌دهند. اگر بپذیریم که تابع جدولبندی‌شده هموار[1]، یعنی دارای تغییر آهسته وکند است، آنگاه این رفتار تفاضلهای درجات بالاتر باید نتیجهٔ خطا باشد.

فرض کنید که خطا در iامین مقدار تابع برابر با ε_i (به‌ازای کلیهٔ مقادیر iها) باشد، دراین‌صورت جدول شکل ۵۰۲ شامل اعداد $\Delta^s(f_i+\varepsilon_i)$ است که تفاوت این‌اعداد با اعداد دقیق $\Delta^s f_i$، که بنابرفرض تغییر آهسته داشتند، مقدار $\Delta^s\varepsilon_i$ است. از رابطهٔ (۲۹۰۲) نتیجه می‌گیریم

$$\Delta^s\varepsilon_i=\sum_{j=0}^{s}\binom{s}{j}(-1)^{s-j}\varepsilon_{i+j} \qquad (۳۰۰۲)$$

ولذا باتوجه به $\varepsilon:=\max_j|\varepsilon_j|$ داریم

$$|\Delta^s\varepsilon_i|\leqslant\sum_{j=0}^{s}\binom{s}{j}|\varepsilon_{i+j}|\leqslant\varepsilon\sum_{j=0}^{s}\binom{s}{j}=\varepsilon(1+1)^s=\varepsilon 2^s$$

اگر مقادیرمندرج درجدول،مقادیردقیقاً گردشده باشند، آنگاه $\varepsilon\leqslant$ ۰٫۰۰۰۰۰۵، وخطاها در تفاضلهای چهارم نباید بزرگتر از ۸ واحد درآخرین رقم اعشاری باشند. اما اگررفتار نوسانی را مربوط به‌خطا بدانیم، خطاها خیلی بیشتر ازاین‌است.

1. smooth

بررسی دقیقتر این تفاضلهای چهارم نشان می‌دهدکـه رفتار نوسانها **منظم واصولی** است. اگر عدد ۱۰، یعنی میانگین ستون تفاضلهای چهـارم را از هریك از درایه‌های این ستون کم کنیم، خواهیم داشت

$$-۱۳ \quad ۸۴ \quad -۱۲۱ \quad ۸۲ \quad -۲۶ \quad -۶$$

که یك فرد با تجربه از اعداد فوق درمی‌یابدکه اشتباهی تقریباً برابر ۲۰ واحد درآخرین رقم اعشاری، در درایهٔ جدول متناظر با عدد ۱۲۱ —، یعنی در درایهٔ ۱٫۲۴۷۶۷ به‌ازای $t = ۱۲۹۰۵$، رخ داده‌است. درحقیقت یك تغییر تنها به‌اندازهٔ ۲۰ — واحد در آخرین رقم اعشاری این درایه، طبق رابطهٔ (۳۰.۲) ستون تفاضلهای چهارم را به‌صورت‌زیر تغییر می‌دهد

$$-۲۰ \quad ۸۰ \quad -۱۲۰ \quad ۸۰ \quad -۲۰ \quad ۰$$

که این اساساً همان مقدارکل نوسانها درآن ستون‌است. □

خلاصه: خطاهای منفرد[1] در یك جدول تابع به‌وسیلهٔ نوسانهای منظم و اصولی در تفاضلهای درجات بـالاتر آشکار می‌شوند. از مقایسهٔ این نوسانها پیرامون مقدار میانگین (موضعی) بـا نوسانهای حاصل از یك خطای تنها، بنابـر (۳۰.۲)، برآوردی از خطا را می‌توان به‌دست آورد و جدول را تصحیح نمود.

درمثال فوق، با اصلاح $f(۱۲۹۰۵)$ به۱٫۲۴۷۸۷،جدول تفاضل‌شکل۶.۲ به‌دست می‌آیدکـه درآن حتی تفاضلهای چهارم دارای یـك علامت‌اند. تفاضلهای پنجم نوسان دارند، اما مقدار آنهاکوچکتر از حداکثر خطای مربوط به‌گردکردن مقادیر تابع یعنی $۲^۵/۲ = ۱۶$ واحداست. ازاینجا نتیجه‌گیری می‌کنیم که‌تفاضلهای پنجم اصولا اختلالات[2] ناشی از گردکردن مقادیر تابع هستند و درونیابی به‌وسیلهٔ یك بسجمله‌ای درجهٔ چهارم نتایج رضایتبخش (و قابل دفاع) می‌دهد.

به‌علت‌اهمیت قبلی‌جدولهای تابعی، مقدارنسبتاً زیادی از مطالب مربوط به‌درونیابی در جدولهای تابعی در طی سده‌های گذشته گسترش یافته است. عملگرهای تفاضلی غیر از عملگر تفاضل پیشرو Δ (مانند تعویض پیشرو E) برای نشانگذاری فشرده برای صورتهای مختلف بسجمله‌های درونیاب معرفی شده‌اند و تفاوت این شکلهای مختلف فقط در ترتیب نقاط درونیابی درآنهاست. این صورتها، اغلب نه به‌علت حقایق تاریخی بلکـه بیشتر از لحاظ سنتی، با نامهایی مانند: نیوتن[3]، گاوس[4]، بسل[5]، ستیرلینگ[6]، گرگوری[7]، اورت[8] و غیره همراه‌اند. نحوهٔ استفادهٔ کامل از انواع صورتها را می‌توان درکتاب هیلدبراند[9][۵] پیداکرد.

---

1. Isolated error 2. noise 3. Newton 4. Gauss
5. Bessel 6. Stirling 7. Gregory 8. Everett
9. Hildebrand

t	$x = f(t)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$	$\Delta^5 f$	$\Delta^6 f$
1,250.5	1.39140						
		−1444					
1,260.5	1.37696		−1469				
		−2913		55			
1,270.5	1.34783		−1414		17		
		−4327		72		−3	
1,280.5	1.30456		−1342		14		−2
		−5669		86		−5	
1,290.5	1.24787		−1256		9		8
		−6925		95		3	
1,300.5	1.17862		−1161		12		−11
		−8086		107		−8	
1,310.5	1.09776		−1054		4		8
		−9140		111		0	
1,320.5	1.00636		−943		4		
		−10083		115			
1,330.5	0.90553		−828				
		−10911					
1,340.5	0.79642						

شکل ۶.۲ مختص استوایی x سیارهٔ مریخ در دستگاه خورشید-مرکزی.

در اینجا صورتهای مختلف را مورد بحث قرار نمی‌دهیم. احساس می‌کنیم که الگوریتم ۴.۲ و برنامهٔ فرعی **فورترن** TABLE که در بخش ۴.۲ داده شده، برای چند موردی که احتمالا دانشجویان از جداول استفاده می‌کنند کافی است.

تمرین

۱-۶.۲ ثابت کنید که یك خطای منفرد در یك جدول تابع، میانگین چند ستون اول تفاضل را تغییر نمی‌دهد.

۲-۶.۲ مقادیر $f(x)$ که ذیلا داده شده‌اند مقادیر یك بسجمله‌ای معین درجهٔ چهارم هستند. یك جدول تفاضل برای آن تشکیل دهید و از این جدول $f(5)$ را پیدا کنید (به تمرین ۲.۶-۶ نگاه کنید).

x	۰	۱	۲	۳	۴
$f(x)$	۱	۵	۳۱	۱۲۱	۳۴۱

۳-۶.۲ یك جدول تفاضل برای داده‌های زیر تشکیل دهید و درجهٔ بسجمله‌ای درونیاب لازم برای ایجاد مقادیر درونیافته را دقیقاً تا تعداد ارقام اعشاری با معنی داده شده برآورد کنید:

x	$f(x)$
۱٫۰	۱٫۵۷۰۹
۲٫۰	۱٫۵۷۱۳
۳٫۰	۱٫۵۷۱۹
۴٫۰	۱٫۵۷۲۷
۵٫۰	۱٫۵۷۳۸
۶٫۰	۱٫۵۷۵۱
۷٫۰	۱٫۵۷۶۷
۸٫۰	۱٫۵۷۸۵
۹٫۰	۱٫۵۸۰۵

۴-۶۰۲ با استفاده از جدول تفاضل درشکل۶.۲ مقادیر زیر را به‌دست آورید و درهردو مورد خطا را برآورد کنید

(الف) $f(۱۲۵۲٫۵)$ (ب) $f(۱۳۳۲٫۵)$

۵-۶۰۲ ثابت کنید که اگر $p_n(x)$ یك بسجمله‌ای از درجهٔ nام با ضریب پیشرو a_n باشد و x_0 یك نقطهٔ دلخواه، آنگاه

$$\Delta^n p_n(x_0) = a_n n! h^n$$

و

$$\Delta^{n+1} p_n(x_0) = 0$$

[راهنمایی: از تعریف (۲۲.۲) در مورد عملگر تفاضل پیشرو، یا لم ۱.۲ رابطهٔ (۱۷.۲) استفاده کنید.]

۶-۶۰۲ فرض کنیم تساوی $x_i = x_0 + ih$، $i = 0, 1, 2, \ldots$، دردست باشد وفرض کنیم که اعداد $\Delta^0 p_n(x_n)$، $\Delta^1 p_n(x_{n-1})$، ...، $\Delta^n p_n(x_0)$ برای یك بسجمله‌ای ازدرجهٔ نابیشتر از n معلوم باشد. نشان دهیدکه چگونه می‌توان با این اطلاعات مقادیر $p_n(x_{n+1})$، $p_n(x_{n+2})$، ... را فقط با n عمل‌جمع برای هر مقدار به‌دست آورد. [راهنمایی: بنابر تمرین۲.۶۰-۵، $\Delta^n p_n(x_i)$ به i بستگی ندارد درحالی‌که به‌موجب تعریف تفاضل پیشرو به‌ازای‌جمیع‌مقادیر i و j داریم $\Delta^j p_n(x_i) = \Delta^j p_n(x_{i-1}) + \Delta^{j+1} p_n(x_{i-1})$.] این‌روش برای رسم بسجمله‌ایها مفیداست. چه رابطه‌ای بین این روش و الگوریتم ۱.۲ وجوددارد؟

۶.۲-۷ چگونه صورت لاگرانژی بسجمله‌ای درونیاب را، وقتی نقاط متساوی‌الفاصله باشند، می‌توانید ساده کنید؟

۶.۲-۸ فرمول تفاضل پسرو[1] نیوتن

$$p_n(x_0+sh)=\sum_{i=0}^{n}(-1)^i\Delta^i f_{-i}\binom{-s}{i}$$

را برای استفاده در نزدیکی انتهای راست جدول به‌دست آورید. در این فرمول از تفاضلهای واقع در امتداد قطری از شکل ۴.۲ که با (۲) مشخص شده، استفاده شده‌است.

۷.۲* تفاضل منقسم به‌عنوان تابعی از شناسه‌ها و درونیابی بوسانی‌اش

تاکنون با تفاضلهای منقسم تنها در نقشی که به‌عنوان ضرایب صورت نیوتنی برای بسجمله‌ای درونیاب داشته‌اند، یعنی به‌عنوان ثابتهایی که به‌وسیلهٔ اعداد مفروض $f(x_i)$، $i=0,\dots,n$، محاسبه می‌شوند، برخورد کرده‌ایم. اما پیدایش تابع $g_n(x)=f[x_0, x_1, \dots, x_n, x]$ در عبارت خطای (۱۸.۲)، مربوط به درونیابی بسجمله‌ای (۱۸.۲)، مستلزم درک چگونگی رفتار تفاضل منقسم $f[x_0, \dots, x_k]$ هنگام تغییر یك یا چندتا از نقاط x_0، $\dots$، x_k است.

در ابتدا تعریف kامین تفاضل منقسم $f[x_0, \dots, x_k]$ را برای کلیهٔ انتخابهای x_0، $\dots$، x_k تعمیم می‌دهیم، یعنی از شرط دو به‌دو متمایز بودن نقاط x_0، $\dots$، x_k صرفنظر می‌کنیم. از آنجا که kامین تفاضل منقسم $f[x_0, \dots, x_k]$ از f در نقاط x_0، $\dots$، x_k، به‌عنوان ضریب پیشرو (یعنی ضریب x^k) در بسجمله‌ای $p_k(x)$، نابیشتر از درجهٔ k، که با $f(x)$ در $k+1$ نقطهٔ x_0، $\dots$، x_k منطبق است تعریف می‌شود. لذا باید منظور خود را از عبارت «$p_k(x)$ بر $f(x)$ در نقاط x_0، $\dots$، x_k منطبق است»، وقتی برخی از این نقاط برهم منطبق‌اند، توضیح دهیم.

تعریف ما از این جمله چنین است. **دو تابع $f(x)$ و $g(x)$ را در نقاط x_0، $\dots$، x_k برهم منطبق گوییم** هرگاه به‌ازای هر نقطهٔ z که m بار در دنبالهٔ x_0، $\dots$، x_k آمده‌است، داشته باشیم

$$f^{(j)}(z)=g^{(j)}(z) \qquad \text{به‌ازای } j=0,1,\dots,m-1$$

درحقیقت $f(x)$ و $g(x)$ در نقاط x_0، $\dots$، x_k برهم منطبق‌اند هرگاه، با احتساب تکرر ریشه‌ها، تفاضل آنها ریشه‌های x_0، $\dots$، x_k داشته باشند (به بخش ۱.۲ نگاه کنید).

□ **مثال:** $f(x)$ و $g(x)$ در نقاط ۲، ۱، ۲، ۴، ۲، ۵، ۴ برهم منطبق‌اند، درحالتی که

$$f(1)=g(1),\quad f(2)=g(2),\quad f'(2)=g'(2),\quad f''(2)=g''(2)$$

---

1. backward

$$f(۴)=g(۴),\quad f'(۴)=g'(۴),\quad f(۵)=g(۵)$$

بنابراین تعریف، بسجمله‌ای تیلر

$$\hat{p}_n(x)=\sum_{j=۰}^{n} f^{(j)}(c)(x-c)^j/j! \qquad (۳۱.۲)$$

درنقطهٔ c، $n+۱$مرتبه بر $f(x)$ منطبق‌است

چون $$(d/dx)^i(x-c)^j|_{x=c}=\begin{cases} j! & i=j \\ ۰ & i\neq j \end{cases}$$

بنابراین

$$\hat{p}_n^{(i)}(c)=f^{(i)}(c) \qquad i=۰,\ldots,n$$ □

زمانی از **درونیابی بوسانی**[1] صحبت می‌کنیم که درجهٔ تماس بسجمله‌ای درونیاب با $f(x)$ در یکی از نقاط درونیابی از مرتبهٔ اول[2] بیشتر باشد.

صحبت از این بسجمله‌ای نابیشتر ازدرجهٔ kکه در $k+۱$ نقطه‌بر تابع $f(x)$ منطبق باشد بجاست، زیراکه به‌موجب فرع ۲.۲ (دربخش ۱.۲) دو بسجمله‌ای نابیشتر از درجهٔ k که در $k+۱$ نقطهٔ (متمایز یا نامتمایز، ولی با احتساب تکرر) برهم منطبق‌اند، بایستی یکی باشند. اگریك چنین بسجمله‌ای درونیاب $p_k(x)$ از درجهٔ نابیشتر از k برای $f(x)$ درنقاط $x_۰$، ...، x_k وجود داشته باشد، آنگاه بنابر تعریف، ضریب پیشرو این‌بسجمله‌ای، kامین تفاضل منقسم $f[x_۰,\ldots,x_k]$ خواهد شد. بنابراین بسجمله‌ای

$$p(x)=p_k(x)-f[x_۰,\ldots,x_k](x-x_۰)\cdots(x-x_{k-۱})$$

نابیشتر ازدرجهٔ $k-۱$ است. ازآنجاکه $(x-x_۰)\cdots(x-x_{k-۱})$ در $x_۰$، ...، $x_{k-۱}$ برتابع صفر منطبق‌است، لذا نتیجه می‌شودکه $p(x)$ درنقاط $x_۰$، ...، x_k بر$p_k(x)$ ودر نتیجه بر$f(x)$ منطبق‌است، یعنی $p(x)$ می‌باید یك بسجمله‌ای از درجهٔ نابیشتر از $k-۱$ باشدکه در نقاط $x_۰$، ...، $x_{k-۱}$ بر $f(x)$ منطبق است. بنابراین، استقرا روی n منجر به‌اثبات

$$p_n(x)=\sum_{j=۰}^{n}(f[x_۰,\ldots,x_j]\prod_{i=۰}^{j-۱}(x-x_i) \qquad (۳۲.۲)$$

برای بسجمله‌ای نابیشتر از درجهٔ n می‌شودکه در $x_۰$، ...، x_n بر$f(x)$ منطبق‌است. البته این فرمول همان فرمول (۱۰.۲) است،که تمامی موضوع این بخش را تشکیل‌می‌دهد.

---

1. osculatory interpolation 2. First order

در پایان، می‌بایستی مطمئن می‌شدیم که به‌ازای هر انتخابی از نقاط درونیابی $x_0, \dots, x_k$ و تابع $f(x)$، یك بسجمله‌ای نابیشتر از درجهٔ k وجود دارد که در این نقاط بر $f(x)$ منطبق است ولی این امر را نمی‌توانیم تضمین کنیم، زیرا ممکن است $f(x)$ به‌تعداد x_iهایی که برهم منطبق می‌شوند، مشتق نداشته باشد. اما اگر $f(x)$ به‌اندازهٔ کافی دارای مشتق باشد، آنگاه به‌استقراء نسبت به k می‌توانیم وجود بسجمله‌ای درونیاب $p_k(x)$ را ثابت کنیم و دوباره یك فرمول مفید [که اساساً همان فرمول ۱۲.۲ خواهد بود] برای تفاضل منقسم به‌دست آوریم.

قضیهٔ ۴.۲ اگر $f(x)$ دارای n مشتق پیوسته باشد و هیچ نقطه‌ای از دنبالهٔ $x_0, \dots, x_n$ بیش از $m+1$ مرتبه تکرار نشود، آنگاه دقیقاً یك بسجمله‌ای $p_n(x)$ از درجهٔ نابیشتر از n وجود دارد که در نقاط $x_0, \dots, x_n$ بر $f(x)$ منطبق است. برای اثبات وجود چنین بسجمله‌ای، همچنین می‌توانیم فرض کنیم که دنبالهٔ نقاط درونیابی غیرنزولی است، یعنی

$$x_0 \leqslant \dots \leqslant x_n$$

به‌ازای $n=0$ مسئله محرز است. فرض کنید که مطلب به‌ازای $n=k-1$ صحیح باشد و صحت آن را به‌ازای $n=k$ بررسی می‌کنیم، دو حالت وجود دارد.

حالت $x_k = x_0$، که در این حالت $x_0 = \dots = x_k$ و، بنابر فرض بالا باید داشته باشیم $m \geqslant k$، یعنی $f(x)$ حداقل k مشتق پیوسته دارد. در این صورت، همان‌طور که قبلا اشاره شد [به‌رابطهٔ (۳۱.۲) نگاه کنید]، بسجمله‌ای تیلر $\hat{p}_k(x)$ برای $f(x)$ پیرامون مرکز $c=x_0$ همان بسجمله‌ای مورد نظر است. باید توجه داشت که ضریب پیشرو آن عدد $f^{(k)}(x_0)/k!$ است، بنابراین

$$f[x_0, \dots, x_k] = \frac{f^{(k)}(x_0)}{k!} \qquad \text{اگر } x_0 = x_1 = \dots = x_k, \quad (۳۳.۲)$$

حالت $x_k < x_0$، که در این حالت بنابرفرض استقرا، می‌توانیم یك بسجمله‌ای $p_{k-1}(x)$ از درجهٔ نابیشتر از $k-1$ پیدا کنیم که در نقاط $x_0, \dots, x_k$ بر $f(x)$ منطبق باشد و یك بسجمله‌ای $q_{k-1}(x)$ از درجهٔ نابیشتر از $k-1$ که در $x_0, \dots, x_{k-1}$ بر $f(x)$ منطبق باشد. در این صورت بسجمله‌ای

$$p_k(x) = \frac{x-x_0}{x_k-x_0} q_{k-1}(x) + \frac{x_k-x}{x_k-x_0} p_{k-1}(x) \qquad (۳۴.۲)$$

از درجهٔ نابیشتر از k خواهد بود، و اکنون ادعا می‌کنیم که این بسجمله‌ای همان بسجمله‌ای مطلوب است، یعنی $p_k(x)$ در نقاط $x_0, \dots, x_k$ بر $f(x)$ منطبق است. داریم

$$p_k^{(j)}(x) = \frac{x-x_0}{x_k-x_0} q_{k-1}^{(j)}(x) + \frac{x_k-x}{x_k-x_0} p_{k-1}^{(j)}(x) + j \frac{q_{k-1}^{(j-1)}(x) - p_{k-1}^{(j-1)}(x)}{x_k-x_0} \qquad (۳۵.۲)$$

فرض کنید $z=x_i=\cdots=x_{i+r}$. اگر $z=x_0$، آنگاه به‌ازای $j=0,\ldots,r-1$ داریم

$$q_{k-1}^{(j)}(z)=f^{(j)}(z)=p_{k-1}^{(j)}(z) \quad \text{و همچنین} \quad p_{k-1}^{(r)}(z)=f^{(r)}(z) \quad (\text{زیرا که } z<x_k)$$

سپس با توجه به‌رابطهٔ (۳۵.۲)، به‌ازای $j=0,\ldots,r$ داریم

$$p_k^{(j)}(z)=0\times q_{k-1}^{(j)}(z)+f^{(j)}(z)+j\,[f^{(j-1)}(z)-f^{(j-1)}(z)]/(x_k-x_0)=f_{(z)}^{(j)}$$

استدلال درحالت $z=x_k$ نیز مشابه این است. بالاخره اگر x_k و $z\neq x_0$، آنگاه به‌ازای $j=0,\ldots,r$ داریم $q_{k-1}^{(j)}(z)=f^{(j)}(z)=p_{k-1}^{(j)}(z)$ و نیز بنابر (۳۵.۲) داریم

$$p_k^{(j)}(z)=\frac{z-x_0}{x_k-x_0}\,f^{(j)}(z)+\frac{x_k-z}{x_k-x_0}\,f^{(j)}(z)+j\,\frac{f^{(j-1)}(z)-f^{(j-1)}(z)}{x_k-x_0}$$

$$=f^{(j)}(z) \qquad j=0,\ldots,r \text{ به‌ازای}$$

که موضوع را برای $n=k$ ثابت می‌کند.

از مقایسهٔ ضرایب پیشرو در دوطرف رابطهٔ (۳۴.۲) دوباره فرمول (۱۲.۲) را به‌دست خواهیم آورد، یعنی با شرط $x_k\neq x_0$ داریم

$$f[x_0,\ldots,x_k]=(f[x_1,\ldots,x_k]-f[x_0,\ldots,x_{k-1}])/(x_k-x_0) \qquad (۳۶.۲)$$

اگر $x_k\neq x_0$

با تعمیم تعریف $f[x_0,\ldots,x_k]$ برای هر انتخاب دلخواه از $x_0,\ldots,x_k$ اکنون چگونگی بستگی $f[x_0,\ldots,x_k]$ را به نقاط $x_0,\ldots,x_k$ مورد بررسی قرار می‌دهیم. این بررسی آشکار می‌سازدکه انگیزهٔ تعمیم تعریف، ملاحظات پیوستگی بوده‌است.

مسئله را با این ملاحظه آغاز می‌کنیم که $f[x_0,\ldots,x_k]$ نسبت به‌شناسه‌هایش یك تابع متقارن[1] است، یعنی $f[x_0,\ldots,x_k]$ فقط به‌اعداد $x_0,\ldots,x_k$ بستگی دارد، نه به‌ترتیبی که این اعداد در فهرست شناسه‌ها ظاهر می‌شوند. این موضوع کاملا واضح‌است، زیرا که کلیهٔ بسجمله‌ایهای درونیاب $p_k(x)$ مستقل از ترتیب نوشتن نقاط درونیابی‌هستند. این بدان‌معنی‌است که هرگاه مناسب باشد می‌توانیم بی‌آنکه خللی در استدلال وارد آید، فرض کنیم که ترتیب شناسه‌های $x_0,\ldots,x_k$ از $f[x_0,\ldots,x_k]$ صعودی‌است.

در مرحلهٔ بعد نشان خواهیم داد که $f[x_0,\ldots,x_k]$ تابعی است پیوسته از شناسه‌هایش.

قضیهٔ ۵.۲ فرض کنیدکه $f(x)$، nبار به‌طور پیوسته روی $[a,b]$ مشتقپذیر باشد وگیریم

1. Symmetric function

$y_0, \ldots, y_n$ نقاط متمایز یا غیرمتمایزی روی $[a, b]$ باشند. دراین‌صورت

(i) نقطه‌ای مانند $\xi \in [\min_i y_i \text{ و } \max_i y_i]$ وجود دارد به‌طوری که تساوی $f[y_0, \ldots, y_n] = f^{(n)}(\xi)/n!$ برقراراست.

(ii) اگر به‌ازای جمیع مقادیر r، $x_0^{(r)}$، $\ldots$، $x_n^{(r)}$، $n+1$ نقطه روی $[a, b]$ باشند و به‌ازای $i = 0, \ldots, n$، رابطهٔ $\lim_{r \to \infty} x_i^{(r)} = y_i$ برقرار باشد. آنگاه داریم

$$\lim_{r \to \infty} f[x_0^{(r)}, \ldots, x_n^{(r)}] = f[y_0, \ldots, y_n]$$

اثبات به‌استقرا نسبت به n انجام می‌گیرد. به‌ازای $n = 0$ هردو حکم به‌طور بدیهی صحیح‌اند. فرض کنید این احکام به‌ازای $n = k - 1$ نیز صحیح باشد و حالت $n = k$ را بررسی کنید.

ابتدا به‌اثبات حالت (ii) درموردی که همهٔ $n+1$ نقطهٔ $y_0, \ldots, y_n$ یکی نیستند می‌پردازیم. سپس با فرض $y_0 \leqslant \ldots \leqslant y_n$، داریم $y_0 < y_n$ و بنابراین به‌ازای همهٔ مقادیر بزرگ r، رابطهٔ $x_0^{(r)} < x_n^{(r)}$ برقراراست و بنابراین طبق (۳۶.۲) داریم

$$\lim_{r \to \infty} f[x_0^{(r)}, \ldots, x_n^{(r)}] = \lim_{r \to \infty} \frac{f[x_1^{(r)}, \ldots, x_n^{(r)}] - f[x_0^{(r)}, \ldots, x_{n-1}^{(r)}]}{x_n^{(r)} - x_0^{(r)}}$$

$$= \frac{\lim f[x_1^{(r)}, \ldots, x_n^{(r)}] - \lim f[x_0^{(r)}, \ldots, x_{n-1}^{(r)}]}{y_n - y_0}$$

$$= \frac{f[y_1, \ldots, y_n] - f[y_0, \ldots, y_{n-1}]}{y_n - y_0}$$

تساوی آخر از فرض استقرایی نتیجه می‌شود. اما رابطهٔ آخر بنابر (۳۶.۲) برابر $f[y_0, \ldots, y_n]$ است که درنتیجه برای این، حالت (ii) ثابت می‌شود.

سپس حالت (i) را ثابت می‌کنیم. اگر $y_0 = y_1 = \ldots = y_n$، آنگاه (i) فقط بیان دیگری از (۳۳.۲) است. در غیراین صورت فرض کنید که

$$y_0 \leqslant y_1 \leqslant \ldots \leqslant y_n$$

و سپس $y_0 < y_n$. اما در این صورت خواهیم دید که برای کلیهٔ rها، می‌توان نقاط $x_0^{(r)} < \ldots < x_n^{(r)}$ را در $[a, b]$ طوری پیدا کرد که به‌ازای $i = 0, \ldots, n$، تساوی $\lim_{r \to \infty} x_i^{(r)} = y_i$ برقرار باشد. سپس، بنابر قضیهٔ ۲.۲ می‌توانیم $\xi^{(r)} \in [x_0^{(r)}, x_n^{(r)}]$ را طوری پیدا کنیم که

$$f[x_0^{(r)}, \ldots, x_n^{(r)}] = \frac{f^{(n)}(\xi^{(r)})}{n!} \qquad \text{به‌ازای } r = 1, 2, 3, \ldots,$$

اما بنابرحالت (ii) که در بالا برای آن ثابت شد، به‌ازای نقطه‌ای مانند

$$\xi \in [\lim x_0^{(r)}, \lim x_n^{(r)}] = [y_0, y_n]$$

داریم

$$f[y_0, \cdots, y_n] = \lim_{r \to \infty} f[x_0^{(r)}, \cdots, x_n^{(r)}] = \frac{\lim_{r \to \infty} f^{(n)}(\xi^{(r)})}{n!}$$

$$= f^{(n)}\frac{(\xi)}{n!}$$

که به‌موجب پیوستگی $f^{(n)}(x)$، (i) ثابت می‌شود.

و بالاخره اثبات (ii) در موردی که $y_n = \cdots = y_1 = y_0$. اکنون از (i) برای اثبات وجود $\xi^{(r)} \in [\min_i x_i^{(r)}, \max_i x_i^{(r)}]$ که به ازای کلیهٔ r ها در رابطهٔ $f[x_0^{(r)}, \ldots, x_n^{(r)}] = f^{(n)}(\xi^{(r)})/n!$ صادق باشد، استفاده می‌کنیم. در این‌صورت چون $y_0 = \cdots = y_n$ و به‌ازای کلیهٔ مقادیر i، رابطهٔ $\lim x_i^{(r)} = y_i$ برقرار است، داریم $\lim \xi^{(r)} = y_0$ وبنابراین طبق (۳۶.۲) وپیوستگی $f^{(n)}(x)$ داریم

$$f[y_0, \cdots, y_n] = \frac{f^{(n)}(y_0)}{n!} = \frac{\lim f^{(n)}(\xi^{(r)})}{n!} = \lim f[x_0^{(r)}, \cdots, x_n^{(r)}]$$

که این رابطه هردو حالت (i) و (ii) را به‌ازای $n = k$ و برای کلیهٔ انتخابهای $y_0, \ldots, y_n$ در بازهٔ $[a, b]$، ثابت می‌کند.

این قسمت را با نتایج جالبی از قضیهٔ ۵.۲ خاتمه می‌دهیم. بلافاصله نتیجه‌گیری می‌شود که تابع

$$g_n(x) = f[x_0, \cdots, x_n, x]$$

که در عبارت خطای مربوط به‌درونیابی بسجمله‌ای ظاهر می‌شود، برای کلیهٔ xها تعریف شده و تابعی پیوسته از x است، اگر $f(x)$ به‌اندازهٔ‌کافی هموار باشد. بنابراین نتیجه می‌شود که

$$f(x) = \sum_{i=0}^{n} f[x_0, \cdots, x_i] \prod_{j=0}^{i-1} (x - x_j) + f[x_0, \cdots, x_n, x] \prod_{j=0}^{n} (x - x_j) \qquad (۳۷.۲)$$

به‌ازای همهٔ xها و نه‌فقط به‌ازای $x \neq x_0, \ldots, x_n$ [به ۱۶.۲ نگاه‌کنید] و نیز به‌ازای کلیهٔ نقاط متمایز یا نامتمایز $x_0, \ldots, x_n$، درحالتی که $f(x)$، به‌اندازهٔ کافی مشتق داشته باشد، برقراراست.

بعلاوه اگر $f(x)$ به‌اندازهٔ‌کافی مشتقپذیر باشد، آنگاه $g_n(x)$ مشتقپذیرخواهد بود. زیرا که طبق تعریف مشتق داریم

$$g_n'(x)=\lim_{h\to 0} g_n[x, x+h]$$

اگر این حد وجود داشته باشد. از طرف دیگر بنابرقضیهٔ ۵.۲ داریم

$$g_n[x, x+h]=f[x_0, \cdots, x_n, x, x+h] \xrightarrow[h\to 0]{} f[x_0, \cdots, x_n, x, x]$$

بنابراین

$$\frac{d}{dx} f[x_0, \cdots, x_n, x]=f[x_0, \cdots, x_n, x, x] \qquad (۳۸.۲)$$

در پایان، قضیهٔ بالا مبین تعریف درونیابی مماسی به‌صورت درونیابی مکرر[1] است. زیرا نشان می‌دهدکه بسجمله‌ای درونیاب در نقاط x_0، ...، x_n همراه با $x_i \longrightarrow y_i$ برای کلیهٔ iها به‌سمت بسجمله‌ای درونیاب درنقاط y_0، ...، y_n همگرا می‌شود، بنابراین kمرتبه درونیابی‌دریك نقطه، حد حالتی‌است که k نقطهٔ درونیابی متمایزیکی‌می‌شوند. دانشجویان با این پدیده در مورد $n=1$، کـه همان درونیابی خطی‌است، آشنا هستند. دراین حالت خط مستقیم $p_1(x)=f(x_0)+f[x_0, x_1](x-x_0)$ قاطعی‌است برای (نمودار) $f(x)$ کـه هرگـاه دو نقطهٔ x_1 و x_0 به‌سمت نقطـهٔ y نزدیك شوند، این قاطـع به‌خـط مماس $\hat{p}_1(x)=f(y)+(x-y)f'(y)$ تبدیل می‌شود ومقدار $\hat{p}_1(x)$ و ضریب زاویهٔ‌آن در نقطهٔ $x=y$ با $f(x)$ یکی می گردد.

□ **مثال ۹.۲:** بافرض $f(x)=\ln x$، مطلوب‌است محاسبهٔ $f(۱٫۵)$ به‌توسط‌درونیابی تابـع درجـهٔ سوم، بـا استفـاده از $f(۱)=۰$، $f(۲)=۰٫۶۹۳۱۴۷$، $f'(۱)=۱$، $f'(۲)=۰٫۵$.

دراین موردچهار نقطهٔ درونیابی عبارت‌انـد از: $y_0=y_1=۱$، $y_2=y_3=۲$. محاسبات زیر را انجام می‌دهیم

$$f[y_0, y_1]=f'(y_0)=۱ \qquad f[y_1, y_2]=۰٫۶۹۳۱۴۷$$

$$f[y_2, y_3]=f'(y_2)=۰٫۵$$

$$f[y_0, y_1, y_2]=\frac{۰٫۶۹۳۱۴۷-۱}{۲-۱}=-۰٫۳۰۶۸۵۳$$

$$f[y_1, y_2, y_3]=\frac{۰٫۵-۰٫۶۹۳۱۴۷}{۲-۱}=-۰٫۱۹۳۱۴۷$$

---

1. repeated interpolation

$$f[y_0, y_1, y_2, y_3] = \frac{-0.193147 + 0.306853}{2-1} = 0.113706$$

جدول کامل تفاضل منقسم به‌صورت زیر نوشته می‌شود

y_i	$f[\ ]$	$f[,\ ]$	$f[,,]$	$f[,,,]$
۱	۰٫۰			
		۱٫۰		
۱	۰٫۰		−۰٫۳۰۶۸۵۳	
		۰٫۶۹۳۱۴۷		۰٫۱۱۳۷۰۶
۲	۰٫۶۹۳۱۴۷		−۰٫۱۹۳۱۴۷	
		۰٫۵		
۲	۰٫۶۹۳۱۴۷			

با استفاده از جدول فوق، تابع

$$p_3(x) = 0.0 + (1.0)(x-1) + (-0.306853)(x-1)^2 + (0.113706)(x-1)^2(x-2)$$

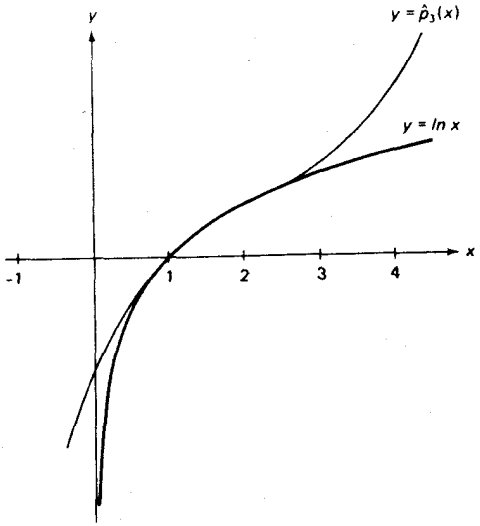
یك بسجمله‌ای درجهٔ سه‌است که از لحاظ مقدار و شیب با $\ln x$ در دونقطهٔ $x=1$ و $x=2$ یکی است. مشخصهٔ بوسانی تقریب $\ln x$ به‌توسط $p_3(x)$ ازشکل ۷.۲ آشکاراست. ازبه‌کار بستن الگوریتم ۱.۲ برای محاسبهٔ $P_3(x)$ در ۱٫۵ خواهیم داشت

$$\ln 1.5 \approx p_3(1.5) = 0.409074$$

برآورد زیر برای خطای $e_3(x) = f(x) - p_3(x)$، با استفاده از (۳۷.۲) و حالت (i) قضیهٔ ۵.۲ به‌دست می‌آید

$$|e_3(1.5)| \leqslant \frac{1}{4!} \max_{1 \leqslant \xi \leqslant 2} |f^{(4)}(\xi)| (1.5-1)^2 (1.5-2)^2 = 0.015625$$

از آنجا که $\ln 1.5 = 0.405465$، عملا مقدار خطا فقط برابر با ۰٫۰۰۳۶۱ است. این مقدار باردیگر نشان می‌دهدکه نامعلوم بودن جای ξ خطا را بر اساس رابطهٔ (۱۸.۲) تقریباً محتاطانه ــ درست بگوییم ــ برآورد می‌کند. □



شکل ۷۰۲ درونیابی بوسانی.

این بخش را با برنامـهٔ فورترنی که ضرایب صورت نیوتنی $p_n(x)$ را محاسبه و سپس $p_n(x)$ را درمجموعه‌ای از نقاط داده شده با فواصل مساوی، ارزیابی می‌کند، خاتمه می‌دهیم.

```
C  CONSTRUCTION OF THE NEWTON FORM FOR THE POLYNOMIAL OF DEGREE
C  .LE. N , WHICH AGREES WITH  F(X)  AT  Y(I), I=1,...,NP1.
C  SOME OR ALL OF THE INTERPOLATION POINTS MAY COINCIDE, SUBJECT
C  ONLY TO THE FOLLOWING RESTRICTIONS.
C  (1) IF  Y(I) = Y(I+K),  THEN  Y(I) = Y(I+1) = ... = Y(I+K) .
C  (2) IF ALSO  Y(I-1) .NE. Y(I) , OR IF   I = 1 , THEN
C      F(I+J) = VALUE OF J-TH DERIVATIVE OF  F(X)  AT  X = Y(J),
C                                                    J=0,...,K.
C
      INTEGER I,J,K,N,NPOINT,NP1
      REAL DX,DY,F(30),FLAST,PNOFX,REALK,X,Y(30)
      READ 500,NP1,(Y(I),F(I),I=1,NP1)
  500 FORMAT(I2/(2F10.3))
C             CONSTRUCT DIVIDED DIFFERENCES
      N = NP1 - 1
      DO 10 K=1,N
         REALK = K
         FLAST = F(1)
         DO 9 I=1,NP1-K
            DY = Y(I+K) - Y(I)
            IF (DY .EQ. 0.) THEN
               F(I) = F(I+1)/REALK
            ELSE
               F(I) = (F(I+1) - FLAST)/DY
               FLAST = F(I+1)
            END IF
    9    CONTINUE
         F(NP1-K+1) = FLAST
   10 CONTINUE
C        CALCULATE PN(X) FOR VARIOUS VALUES OF X.
      READ 501,NPOINT,X,DX
```

```
501 FORMAT(I3/2F10.3)
    DO 30 J=1,NPOINT
       PNOFX = F(1)
       DO 29 I=2,NP1
          PNOFX = F(I) + (X - Y(I))*PNOFX
 29    CONTINUE
       PRINT 629,J,X,PNOFX
629    FORMAT(I10,2E20.7)
       X = X + DX
 30 CONTINUE
                                              STOP
    END
```

اگر کلیهٔ نقاط درونیابی متمایز باشند، محاسبهٔ تفاضلهای منقسم، به‌الگوریتم ۳.۲ مربوط می‌شود. اگر بعضی از نقاط درونیابی برهم منطبق باشند، ورودی برنامه بایستی شامل مشتق تابع درونیاب باشد. به‌خصوص، فرض بر این است که ورودی از آرایه‌ای[1] از نقاط درونیابی $Y(I)$، به‌ازای $I=1,\ldots,NP1=n+1$ تشکیل شده است. برای ساده‌تر شدن برنامه‌نویسی فرض شده‌است که دنبالهٔ نقاط درونیابی در شرط زیر صدق می‌کند

اگر $Y(I)=Y(I+K)$، آنگاه $Y(I)=Y(I+1)=\ldots=Y(I+K)$

یعنی تمامی نقاط درونیابی تکراری با هم ظاهر می‌شوند. علاوه بر این شرط، فرض بر این است که به‌ازای هر I

اگر $Y(I)=Y(I-j)\neq Y(I-j-1)$، آنگاه $F(I)=f^{(j)}(Y(I))$

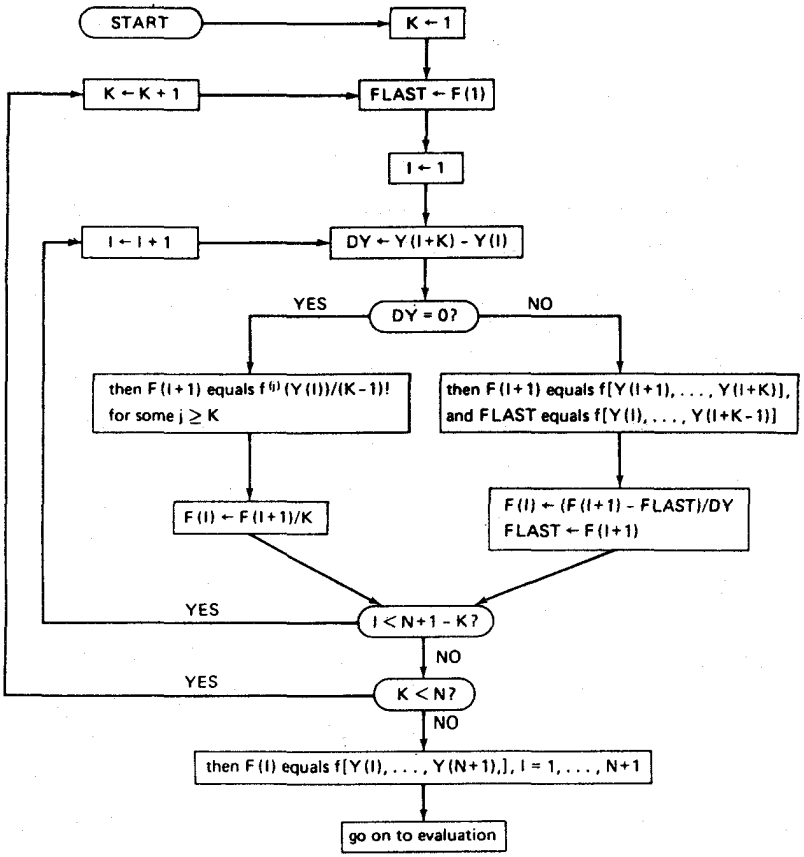
بنابراین با $f(x)=1/x$ و $n=6$ ورودیهای زیر صحیح خواهند بود، بدین‌معنی که یك بسجمله‌ای از درجهٔ نابیشتر از ۶ تولید خواهد شد که $f(x)=1/x$ را در $Y(I)$ داده شده به‌ازای $I=1,\ldots,7$، درونیابی می‌کند

I	۱	۲	۳	۴	۵	۶	۷
Y(I)	۲٫۰	۲٫۰	۲٫۰	۱٫۰	۴٫۰	۴٫۰	۵٫۰
F(I)	۰٫۵	−۰٫۲۵	۰٫۲۵	۱٫۰	۰٫۲۵	−۰٫۰۶۲۵	۰٫۲

دانشجویان را تشویق می‌کنیم که مثال مشابهی انتخاب و محاسبات را از راه برنامهٔ **فورترن** دنبال کنند. روند نمای[2] زیر که معرف محاسبات تفاضلهای منقسم است ممکن است در مساعی فوق مفید واقع شود.

---

1. array 2. Flowchart



تمرین

۱-۷۰۲ با استفاده از درونیابی درجهٔ دوم ومقادیر داده شدهٔ $f(۰)=۱$ و $f'(۰)=۱$ و $f(۱)=۲٫۷۱۸۳$، مقدار $f(۰٫۵)$ را برای تابع $f(n)=e^x$ محاسبه کنید. این مقدار را با $f(۰٫۵)=۱٫۶۴۸۷$ که به‌طور صحیح گردشده‌است، مقایسه کنید.

۲-۷۰۲ مقادیر

$$f(۰)=۰ \qquad f'(۰)=۱ \qquad f(۱)=۱٫۱۷۵۲ \qquad f'(۱)=۱٫۵۴۳۱$$

برای تابع $f(x)=\sinh x$ داده شده‌اند. جدول تفاضل منقسم را تشکیل دهید و با استفاده از درونیابی تابع درجهٔ سوم، مقدار $f(۰٫۵)$ را محاسبه نمایید و آن را با $\sinh ۰٫۵=۰٫۵۲۱۱$ مقایسه کنید.

۳-۷۰۲ تابع $f(x)$ یك ریشهٔ دوگانه در z_1 و یك ریشهٔ سه‌گانه در z_2 دارد. شکل بسجمله‌ای

نابیشتر از درجهٔ پنجم راکه $f(x)$ را دوبار در z_1 و سه‌بار در z_2 و یك بار درنقطهٔ غیر مشخص z_3 درونیابی کند، تعیین نمایید.

۴-۷.۲ ضرایب a_0، a_1، a_2، a_3 از بسجمله‌ای درجهٔ سه

$$p_3(x)=a_0+a_1(x-y)+a_2(x-y)^2+a_3(x-y)^3$$

را چنان پیداکنیدکه داشته باشیم

$$p_3(y)=f_y \qquad p_3'(y)=f_y' \qquad p_3(z)=f_z \qquad p_3'(z)=f_z'$$

دراینجا، y، z، f_y، f_y'، f_z، f_z' اعداد مفروضی هستند (و $y\neq z$).

۵-۷.۲ عبارت ساده‌ای برای $p_3[(y+z)/2]$ برحسب اعداد مفروض y، z، f_y، f_y'، f_z، f_z' به‌دست آورید، در اینجا $p_3(x)$، بسجمله‌ای حاصل درتمرین ۲.۷-۴ است.

۶-۷.۲ گیریم $f(x)$ و $g(x)$، توابع هموار باشند. ثابت‌کنیدکه تابع $f(x)$ در نقطهٔ $x=c$، kمرتبه بر $g(x)$ منطبق‌است، اگر و فقط اگر، به‌ازای x نزدیك به c داشته باشیم $f(x)=g(x)+\mathcal{O}(|x-c|^k)$.

۷-۷.۲ گیریم $g(x)=f[x_0,\ldots,x_k,x]$. ثابت‌کنید (با استفاده از استقرا) که

$$g[y_0,\ldots,y_n]=f[x_0,\ldots,x_k,y_0,\ldots,y_n]$$

۸-۷.۲ با استفاده از تمرین ۲.۷-۷، ثابت‌کنیدکه اگر $g(x)=f[x_0,\ldots,x_k,x]$، آنگاه

$$g^{(n)}(x)=n!f[x_0\ldots,x_k,\underbrace{x,\ldots,x}_{(n+1)\text{مرتبه}}]$$

۹-۷.۲ گیریم $f(x)=g(x)h(x)$. ثابت‌کنیدکه

$$f[x_0,\ldots,x_k]=\sum_{i=0}^{k} g[x_0,\ldots,x_i]h[x_i,\ldots,x_k]$$

(از روش استقرا استفاده‌کنید، یا سمت راست را به‌صورت ضریب پیشرو بسجمله‌ای از درجهٔ نابیشتر از k که در نقاط x_0، ...، x_k، $g(x)h(x)$ را درونیابی می‌کند، مشخص سازید). این‌حالت $x_0=\ldots=x_k$، کدام فرمول معروف دیفرانسیل و انتگرال را به‌دست می‌آورید؟

۳

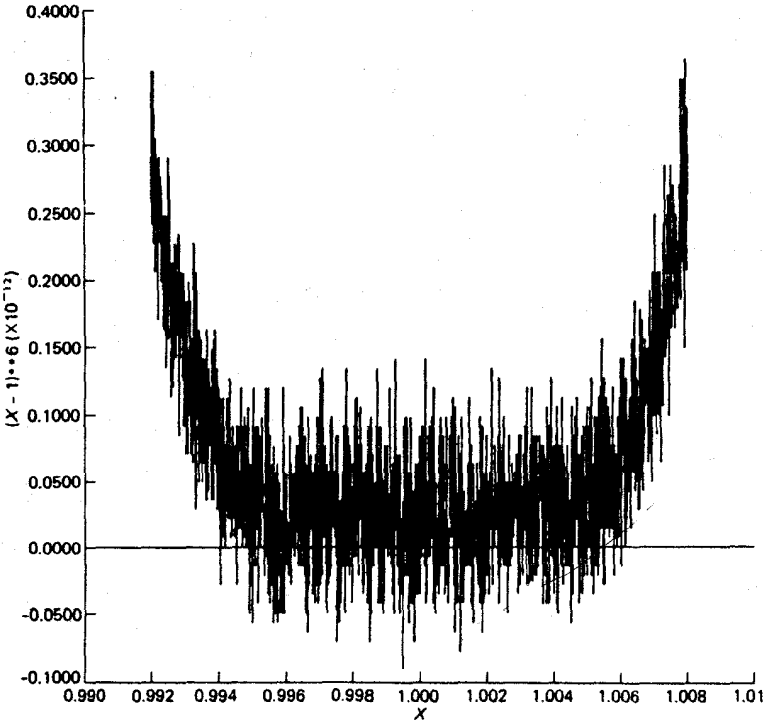
حل معادلات غیرخطی

یکی از مسائلی که غالباً درکارهای عملی پیش می‌آید پیدا کردن ریشه‌های معادلاتی به‌صورت

$$f(x)=0 \qquad (۱.۳)$$

یعنی، صفرهای تابع $f(x)$ است. تابع $f(x)$ ممکن است صریحاً، مثلا به‌صورت یك بسجمله‌ای برحسب x یا به‌صورت یك تابع متعالی[1] داده شده باشد. اما اغلب ممکن است $f(x)$ فقط به‌صورت ضمنی داده شده باشد، یدین‌معنی که ممکن‌است قاعده‌ای برای ارزیابی $f(x)$ به‌ازای هرشناسه در دست باشد، ولی صورت سادهٔ آن معلوم نباشد. ازاین‌رو $f(x)$ ممکن‌است معرف مقادیرجواب یك معادلهٔ دیفرانسیل دریك نقطهٔ مشخص باشد، درحالی‌که x ممکن‌است بیانگر یك‌شرط اولیهٔ آن معادله باشد. درموارد نادر ممکن‌است بتوان ریشه‌های دقیق معادلهٔ (۱.۳) را پیدا کرد، که یك مثال آن بسجمله‌ای‌هایی‌هستند که به‌صورت حاصلضرب چند عامل درمی‌آیند. لیکن درحالت کلی می‌توانیم امیدوار باشیم که، با اتکا به‌تکنیکهای محاسباتی، فقط جوابهای تقریبی را به‌دست آوریم. بسته به‌اینکه زمینه‌کار چه باشد، «جواب تقریبی» ممکن‌است بدین‌معنی باشدکه یا x^* در معادلهٔ (۱.۳) به‌طور تقریبی صدق می‌کند یعنی، $|f(x^*)|$ «بسیار کوچك» است، ویا نقطهٔ x^* به‌جواب (۱.۳) «نزدیك» است. متأسفانه مفهوم یك «جواب تقریبی» تقریباً نامعلوم‌است. یك جواب تقریبی که ازیك کامپیوتر به‌دست

1. transcendental



شکل ۱.۳

آمده‌است، به‌علت گرد کردن یا ناپایداری و یا به‌سبب حساب خاصی که به‌کار گرفته می‌شود، تقریباً همیشه دارای خطـاست. درحقیقت ممکن است «جوابهای تقریبی» بسیاری وجـود داشته باشند کـه آنها هم صحیح باشند ولو اینکه جـواب مطلوب یگانه باشد. به منظور نشان دادن عدم اطمینانی که درپیدا کردن ریشه‌ها وجود دارد، درشکل ۱.۳ نمودار تابع

$$P_6(x)=(1-x)^6=1-6x+15x^2-20x^3+15x^4-6x^5+x^6$$

را می‌آوریم. البته این تابـع تنها دارای ریشهٔ $x=1$ است. یك برنامـهٔ **فورترن** برای ارزیابی منبسط این تابع نوشته شده‌است. این برنامه برای محاسبهٔ مقدار $P_6(x)$ درتعداد زیادی نقاط $x_1<x_2<\cdots<x_N$ نزدیك به $x=1$ روی یك کامپیوتر CDC ۶۵۰۰ مورد استفاده قرارگرفته است. سپس یك رسام[1] کامپیوتری برای رسم نمودار پاره‌خطهای تکه‌ای، که درشکل ۱.۳ مشاهده می‌شود، به‌کار گرفته شده‌است. دراین نمودار دیده می‌شود که $P_6(x)$ آشکارا دارای صفرهای زیادی‌است زیرا که این نمودار چندین‌بار تغییر علامت

1. plotter

نشان می‌دهد. این صفرهای آشکار درمحدودهٔ ۰٫۹۹۴ تا ۱٫۰۰۶ هستند. بنابراین به‌کار گرفتن شکل منبسط برای برآورد صفر در $x=1$، آشکارا به‌برآوردهای قابل‌قبولی منجر می‌شود که فقط تا دو رقم اعشاری صحیح‌اند، درحالی‌که کامپیوتر CDC۶۵۰۰ درحساب با ۱۴ رقم ممیز شناور کار کرده‌است. دلیل این‌رفتار را باید درخطای گردکردن، از دست رفتن و حذف ارقام با معنی در برنامهٔ **فورترن** هنگام محاسبهٔ $P_6(x)$ جستجو کرد. این مثال برخی از مخاطرات موجود در ریشه‌یابی را نشان می‌دهد.

در بقیهٔ این‌فصل روشهای بارستی‌گونا گونی رابرای پیداکردن تقریبهای ریشه‌های سادهٔ معادلهٔ (۱.۳) مورد بررسی‌قرارمی‌دهیم. به‌علت‌اهمیت‌کاربردهای معادلات بسجمله‌ای در مهندسی، دراین فصل توجه خاصی به‌آنها خواهیم داشت.

۱.۳ بررسی اجمالی روشهای بارستی

در این بخش برخی از روشهای مقدماتی برای تعیین یك جواب معادلهٔ

$$f(x)=0 \qquad (۱.۳)$$

را مطرح می‌کنیم ومورد استفادهٔ آنها را با به‌کاربردن این روشها درحل معادلهٔ بسجمله‌ای سادهٔ

$$x^3-x-1=0 \qquad (۲.۳)$$

نشان می‌دهیم. در اینجا داریم $f(x)=x^3-x-1$. برای این مثال پیدا می‌کنیم که

$$f(1)=-1<0<5=f(2) \qquad (۳.۳)$$

بنابراین، از آنجاکه $f(x)$ پیوسته‌است، با در نظرگرفتن قضیهٔ مقدار میانگین در توابع پیوسته (به‌بخش ۷.۱ نگاه‌کنید)، $f(x)$ باید در بازهٔ $[1,2]$ درنقطه‌ای صفرشود. هرگاه $f(x)$ در دو نقطه یا بیشتر دربازهٔ $[1,2]$ صفر شود، با توجه به‌قضیهٔ **ژول** (به‌بخش ۷.۱ نگاه‌کنید)، می‌بایستی $f'(x)$ نیزدربازهٔ $[1,2]$ صفرشود. اما ازآنجاکه $f'(x)=3x^2-1$ روی $[1,2]$ مثبت است، بنابراین $f(x)$ دربازهٔ $[1,2]$ دقیقاً یك‌ریشه دارد. اگر این ریشه‌را ξ بنامیم، داریم:

$$\xi \approx 1.5 \qquad \text{با خطای مطلق} \leqslant 0.5$$

برای شناخت بیشتر این ریشه، $f(x)$ را در نقطهٔ میانی بازهٔ $[1,2]$ یعنی در نقطهٔ ۱٫۵ ارزیابی می‌کنیم. داریم:

$$f(1.5)=0.875>0>-1=f(1)$$

در نتیجه اکنون می‌دانیم که ریشهٔ ξ دربازهٔ کوچکتر $[1,1.5]$ یعنی

$$\xi = ۱٫۲۵ \quad \text{با خطای مطلق} \quad \leqslant ۰٫۲۵$$

قرار دارد. دوباره $f(x)$ را در نقطهٔ میانی ۱٫۲۵ آزمایش می‌کنیم. داریم

$$f(۱٫۲۵) = -۰٫۲۹۶\ldots < ۰ < ۰٫۸۷۵ = f(۱٫۵)$$

بنابراین می‌دانیم که ξ حتی در بازهٔ کوچکتر $[۱٫۲۵, ۱٫۵]$ قرار دارد یعنی

$$\xi \approx ۱٫۳۷۵ \quad \text{با خطای مطلق} \quad \leqslant ۰٫۱۲۵$$

این شیوهٔ جایابی یك ریشه از معادلهٔ $f(x) = ۰$ در یك دنباله از بازه‌ها، که اندازهٔ آنها سیرنزولی دارد، به‌روش **تنصیف**[1] معروف است.

الگوریتم ۱.۳: روش تنصیف. فرض می‌کنیم تابع $f(x)$ در بازهٔ $[a_۰, b_۰]$ پیوسته باشد به‌طوری که $f(a_۰)f(b_۰) \leqslant ۰$.

For $n = 0, 1, 2, \ldots,$ until satisfied, do:
  Set $m = (a_n + b_n)/2$
  If $f(a_n)f(m) \leq 0$, set $a_{n+1} = a_n$, $b_{n+1} = m$
  Otherwise, set $a_{n+1} = m$, $b_{n+1} = b_n$
  Then $f(x)$ has a zero in the interval $[a_{n+1}, b_{n+1}]$

از این به‌بعد غالب الگوریتمها را به‌صورت فشردهٔ فوق بیان خواهیم کرد. برای دانشجویانی که با زبان **الگول**[2] آشنایی دارند، نشانگذاریهای فوق باید کاملا طبیعی باشد. بعلاوه در الگوریتم بالا، اصطلاح «until satisfied» به‌کار برده شده است تا تأکید شود که این شرح الگوریتم کامل نیست، و استفاده‌کننده از الگوریتم باید ملاك دقیق ختم آن را خود تعیین کند. بخشی از این ملاکها به‌مسئلهٔ خاصی که با الگوریتم فوق حل می‌شود بستگی دارد. بعضی از ملاکهای اختتامی ممکن در بخش بعد مورد بحث قرار خواهند گرفت.

در هر مرحله از الگوریتم تنصیف ۱.۳، طول بازه‌ای که می‌دانیم شامل یك صفر $f(x)$ است، نصف می‌شود. بنابراین هر مرحله یك رقم دودویی صحیح جدید دیگری برای ریشهٔ ξ از $f(x) = ۰$ تولید می‌کند. در مثال بالا با همان مقادیر اولیهٔ $a_۰ = ۱$ و $b_۰ = ۲$ و پس از اجرای ۲۰ مرحله از این الگوریتم خواهیم داشت:

$$a_{۲۰} = ۱٫۳۲۴۷۱۷۵\ldots \leqslant \xi \leqslant b_{۲۰} = ۱٫۳۲۴۷۱۸۴\ldots$$

$$f(a_{۲۰}) = (-۱٫۸۵۷\ldots)۱۰^{-۶} < ۰ < (۲٫۲۰۹\ldots)۱۰^{-۶} = f(b_{۲۰})$$

آشکار است که با این الگوریتم و کوشش کافی همیشه می‌توان یك ریشه را تا دقت عمل مورد

---

1. Bisection 2. Algol

نظری جایابی کرد. اما در مقایسه با روشهای دیگری که بحث خواهد شد، روش تنصیف دارای سرعت همگرایی نسبتاً اندکی‌است.

می‌توان امیدوار بودکه با به‌کار گرفتن کاملتر اطلاعاتی که درهرمرحله از $f(x)$ در دست‌است، سریعتر به‌ریشه رسید. درمثال (۲.۳) با اطلاعات زیرشروع کردیم:

$$f(۱)=-۱<۰<۵=f(۲)$$

از آنجاکه $|f(۱)|$ نسبت به $|f(۲)|$ نزدیکتر به‌صفراست، ریشهٔ ξ احتمالا به۱ نزدیکتر است تا ۲ (دست‌کم اگر $f(x)$ «تقریباً» خطی باشد). بنابراین به‌جای اینکـه نقطهٔ میانی یا متوسط ۱ و ۲ راکه نقطهٔ ۱٫۵ است آزمایش کنیم، اکنون $f(x)$ را درمیانگین وزنی یعنی

$$w=\frac{|f(۲)|\times ۱+|f(۱)|\times ۲}{|f(۲)|+|f(۱)|} \qquad (۴.۳)$$

آزمایش‌می‌کنیم. ازآنجاکه علامتهای $f(۱)$ و$f(۲)$ مختلف هستند، می‌توانیم‌رابطهٔ(۴.۳) را به‌صورت ساده‌تر زیربنویسیم:

$$w=\frac{f(۲)\times ۱-f(۱)\times ۲}{f(۲)-f(۱)} \qquad (۵.۳)$$

و برای مثال فوق جواب

$$w=\frac{۵\times ۱+۱\times ۲}{۶}=۱٫۱۶۶۶۶۶\ldots$$

و

$$f(w)=-۰٫۵۷۸۷۰۳\ldots<۰<۵=f(۲)$$

به‌دست می‌آید. ازاین‌رو ξ دربازهٔ $[۲, ۱٫۱۶۶۶۶۶\ldots]$ قراردارد. با تکرار روند فوق برای این بازه خواهیم داشت:

$$w=\frac{۵\times(۱٫۱۶۶۶۶۶\ldots)+(۰٫۵۷۸۷۰۳\ldots)\times ۲}{۵٫۵۷۸۷۰۳\ldots}=۱٫۲۵۳۱۱۲\ldots$$

$$f(w)=-۰٫۲۸۵۳۶۳\ldots<۰<۵=f(۲)$$

در نتیجه $f(x)$ دربازهٔ $[۲, ۱٫۲۵۳۱۱۲\ldots]$ یك ریشه دارد. ایـن الگوریتم به‌روش **تصحیح خطا**[1] معروف است.

---

1. false_position or regula_falsi

الگوریتم ۲.۳: تصحیح خطا. تابع $f(x)$ که در بازهٔ $[a_0, b_0]$ پیوسته است و $f(a_0)f(b_0) < 0$، داده شده‌است.

For $n = 0, 1, 2, \ldots,$ until satisfied, do:
  Calculate $w = [f(b_n)a_n - f(a_n)b_n]/[f(b_n) - f(a_n)]$
  If $f(a_n)f(w) \leq 0$, set $a_{n+1} = a_n$, $b_{n+1} = w$
  Otherwise, set $a_{n+1} = w$, $b_{n+1} = b_n$

بعد از ۱۶ بار استفاده از این الگوریتم در مثال قبلی و با همان نقاط شروع $a_0 = 1$ و $b_0 = 2$ خواهیم داشت:

$$1.32471740\ldots = a_{16} \leqslant \xi \leqslant b_{16} = 2$$

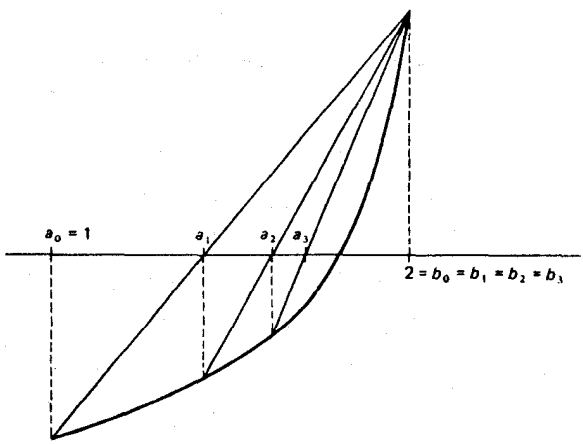
$$f(a_{16}) = (-1.95\ldots)10^{-6} < 0 < 5 = f(b_{16})$$

بنابراین، گرچه روش تصحیح خطا تا حدی سریعتر از روش تنصیف نقطه‌ای راکه درآن نقطه مقدار $|f(x)|$ «کوچک» است به‌دست می‌دهد، اما به‌طور قطع نمی‌تواند بازهٔ «کوچکی» راکه بدانیم ریشه درآن وجود دارد پدید آورد.

نگاهی اجمالی به‌شکل۲.۳ دلیل این امر را نشان می‌دهد. همان‌طورکه به‌آسانی دیده می‌شود، میانگین وزنی

$$w = \frac{f(b_n)a_n - f(a_n)b_n}{f(b_n) - f(a_n)}$$

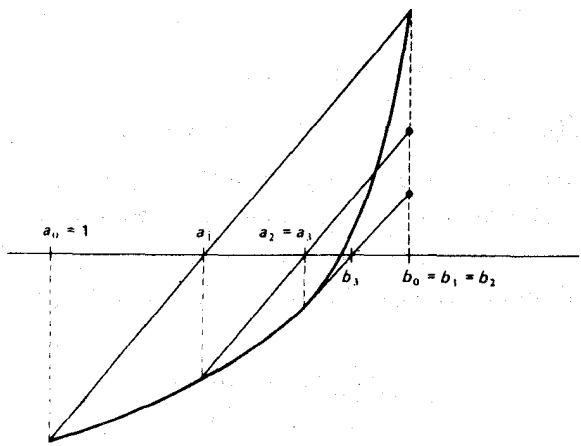
نقطهٔ تقاطع‌خط مستقیم ماربر نقاط $\{a_n, f(a_n)\}$ و $\{b_n, f(b_n)\}$ است با محورxها.این خط مستقیم قاطع $f(x)$ است، و در مثال فوق $f(x)$ دربازهٔ موردنظر$[1, 2]$ صعودی و



شکل۲.۳ موقعیت اشتباهی.

دارای تقعری به‌سمت بالاست. بنابراین قاطع همیشه در بالای نمودار $f(x)$ قراردارد. در نتیجه، دراین‌مثال w همیشه درطرف چپ صفرقراردارد. اگر $f(x)$ صعودی و دارای تقعر به‌سمت پایین می‌بود، x همیشه درسمت راست ریشه قرارمی‌گرفت.

الگوریتم تصحیح خطا را می‌توان به‌چندین طریق بهبود بخشیدکه دوطریقه ازآنها را ذیلا مورد بحث قرار می‌دهیم. طریقهٔ اول «**تصحیح خطای اصلاح شده**» نامیده می‌شود که در آن خطهای مستقیمی‌که شیب آنها پیوسته‌کمتر می‌شود جایگزین قاطع می‌شوند تا جایی‌که w در طرف دیگر ریشه قرار می‌گیرد. این نکته درنمودار ۳.۳ نشان داده شده است.



شکل ۳.۳ تصحیح خطای اصلاح‌شده.

الگوریتم ۳.۳: تصحیح خطای اصلاح‌شده. فرض‌می‌کنیم $f(x)$ روی $[a_0, b_0]$ پیوسته باشد و $f(a_0)f(b_0) < 0$، پس

Set $F = f(a_0)$, $G = f(b_0)$, $w_0 = a_0$

For $n = 0, 1, 2, \ldots,$ until satisfied, do:

  Calculate $w_{n+1} = (Ga_n - Fb_n)/(G - F)$

  If $f(a_n)f(w_{n+1}) \leq 0$, set $a_{n+1} = a_n$, $b_{n+1} = w_{n+1}$, $G = f(w_{n+1})$

    If also $f(w_n)f(w_{n+1}) > 0$, set $F = F/2$

  Otherwise, set $a_{n+1} = w_{n+1}$, $F = f(w_{n+1})$, $b_{n+1} = b_n$

    If also $f(w_n)f(w_{n+1}) > 0$, set $G = G/2$

Then $f(x)$ has a zero in the interval $[a_{n+1}, b_{n+1}]$

اگر این الگوریتم برای مثال قبلی با $a_0 = 1$ و $b_0 = 2$ به‌کار گرفته شود، بعد از ۶ مرحله جواب زیر به‌دست می‌آید:

$$1.324717950\ldots = a_6 \leqslant \xi \leqslant b_6 = 1.324717960\ldots$$

$$f(a_6) = (-1.7360\ldots)10^{-8} < 0 < (1.7350\ldots)10^{-8} = f(b_6)$$

که بهبود قابل‌توجهی نسبت به‌روش تنصیف نشان می‌دهد.

یك طریقهٔ دیگر (روش دوم) تصحیح خطا، که بسیار متداول است، روش «**خط‌قاطع**[1]» نامیده می‌شود، که درعمل همه‌جا ازخطوط قاطع استفاده، ولی از **تخصیص حاشیه** به‌ریشه، صرفنظر می‌کنند.

الگوریتم ۴.۳: روش خط قاطع. تابع $f(x)$ و دونقطهٔ x_0 ، x_{-1} داده شده‌اند

For $n = 0, 1, 2, \ldots,$ until satisfied, do:
  Calculate $x_{n+1} = [f(x_n)x_{n-1} - f(x_{n-1})x_n]/[f(x_n) - f(x_{n-1})]$

اگر دومین روش اصلاح‌شده را برای مثال قبلی با $x_{-1} = 1$ و $x_0 = 2$ به‌کار بریم، پس از ۶ مرحله خواهیم داشت:

$$x_6 = 1.324717900\ldots, \quad f(x_6) = (3.4580\ldots)10^{-8}$$

ظاهراً روش خط قاطع سریعاً نقطه‌ای را که در آن $|f(x)|$ «کوچك» است جایابی می‌کند اما درحالت کلی هیچگونه احساسی نسبت به‌فاصلهٔ این نقطه از ریشهٔ $f(x)$ به‌شخص نمی‌دهد. همچنین از آنجا که مختلف‌العلامه بودن توابع $f(x_n)$ و $f(x_{n-1})$ الزامی نیست، لذا عبارت

$$x_{n+1} = \frac{f(x_n)x_{n-1} - f(x_{n-1})x_n}{f(x_n) - f(x_{n-1})} \qquad (6.3)$$

مهیای نتایج خطای گردکردن است. درموارد استثنایی ممکن است تساوی $f(x_n) = f(x_{n-1})$ برقرار باشد، که دراین صورت محاسبهٔ x_{n+1} غیرممکن می‌شود. از این‌رو ــ اگرچه این کار نیز چاره‌ساز نیست ــ برای محاسبهٔ x_{n+1} بهتر است از عبارت معادل زیر استفاده شود

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \qquad (7.3)$$

که طبق این فرمول x_{n+1} با افزودن «جملهٔ تصحیحی» زیر از x_n به‌دست می‌آید:

---

1. secant

$$\frac{-f(x_n)}{[f(x_n)-f(x_{n-1})]/(x_n-x_{n-1})} \qquad (۸.۳)$$

دانشجویان به‌یاد دارند که نسبت $[f(x_n)-f(x_{n-1})]/(x_n-x_{n-1})$، همان اولین تفاضل منقسم $f(x)$ است، و طبق (۱۰.۲) این نسبت برابر است با شیب خط قاطع ماربر نقاط $\{x_{n-1}, f(x_{n-1})\}$ و $\{x_n, f(x_n)\}$ برای خم $f(x)$. بعلاوه بنابر (۱۷.۲) ملاحظه می‌کنیم که اگر $f(x)$ مشتقپذیر باشد، نسبت فوق برابر با شیب $f(x)$ در نقطه‌ای بین x_{n-1} و x_n خواهد بود. بنابراین بجاست که در نقطه‌ای «نزدیک» به x_n و x_{n-1} به‌جای نسبت فوق، $f'(x)$ گذاشته شود البته درصورتی که $f'(x)$ قابل محاسبه باشد.

اگر $f(x)$ مشتقپذیر باشد، دراین‌صورت اگر در رابطهٔ (۷.۳) به‌جای شیب قاطع، شیب مماس برمنحنی در نقطهٔ x_n راقراردهیم. فرمول بارستی روش **نیوتن** به‌دست می‌آید:

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)} \qquad (۹.۳)$$

الگوریتم ۵.۳: روش نیوتن. تابع پیوسته-مشتقپذیر $f(x)$ و نقطهٔ x_0 داده شده‌اند.

For $n = 0, 1, 2, \ldots,$ until satisfied, do:
  Calculate $x_{n+1} = x_n - f(x_n)/f'(x_n)$

اگر این الگوریتم برای مثال قبلی با $x_0=1$ به‌کار برده شود، بعد از چهار مرحله خواهیم داشت:

$$x_4=1.32471810\ldots, \qquad f(x_4)=(9.24\ldots)10^{-7}$$

سرانجام به‌روش «**بارستی نقطهٔ ثابت**» که مورد خاصی از روش نیوتن است اشاره می‌کنیم. اگرقراردهیم

$$g(x)=x-\frac{f(x)}{f'(x)} \qquad (۱۰.۳)$$

آنگاه فرمول بارستی روش نیوتن (۹.۳) شکل سادهٔ زیر را پیدا خواهد کرد:

$$x_{n+1}=g(x_n) \qquad (۱۱.۳)$$

اگر دنبالهٔ $x_1, x_2, x_3, \ldots$ که این‌گونه تولید می‌شود به‌سمت نقطهٔ ξ همگرا شود و $g(x)$ پیوسته باشد، آنگاه خواهیم داشت:

$$\xi=\lim_{n\to\infty} x_{n+1}=\lim_{n\to\infty} g(x_n)=g\left(\lim_{n\to\infty} x_n\right)=g(\xi) \qquad (۱۲.۳)$$

ویا $g(\xi)=\xi$، یعنی ξ نقطهٔ ثابتی‌است از $g(x)$. واضح‌است که اگر ξ یك نقطهٔ ثابت تابع

بارستی $g(x)$ برای روش نیوتن باشد، آنگاه ع یـك ریشهٔ معادلهٔ $f(x)=0$ نیزخواهد بود. اما برای معادلهٔ داده شدهٔ $f(x)=0$ این امکان وجود داردکه توابع بارستی مختلف، $g(x)$ با این‌ویژگی‌انتخاب شوندکه هر نقطهٔ ثابت$g(x)$ یك صفر$f(x)$باشد. برای هرچنین انتخابی می‌بایستی دنبالهٔ x_1، x_2، x_3، ... به‌وسیلهٔ فرمول

$$x_{n+1}=g(x_n) \qquad n=0, 1, 2, \ldots$$

محاسبه شود و امیدوار بودکه این دنباله همگرا باشد. اگرچنین باشد، حد این دنباله‌ریشهٔ معادلهٔ $f(x)=0$ خواهـد بود. در بخشهای ۳.۳ و ۴.۳ بارست نقطـهٔ ثابت را به‌طور مشروح موردبحث قرار می‌دهیم.

□ **مثال ۱.۳:** تابع$f(x)=x-0.2\sin x-0.5$ دقیقاً یك ریشه بین $x_0=0.5$ و $x_1=1.0$ دارد، زیراکه رابطهٔ $f(0.5)f(1.0)<0$ برقراراست درحالی‌که$f'(x)$ بربازهٔ $[0.5, 1]$ صفرنمی‌شود. با استفاده از الگوریتمهای ۱.۳، ۳.۳، ۴.۳ و ۵.۳ ریشهٔ معادلهٔ فوق را تا ۶ رقم با معنی تعیین کنید.

محاسبات زیر باکامپیوتر IBM ۷۰۹۴، ۲۷ بیت، با ممیزشناور و دقت ساده انجام گرفته‌است.

n	Algorithm 3.1 x_n	e_n	Algorithm 3.3 x_n	e_n	Algorithm 3.4 x_n	Algorithm 3.5 x_n
−1					1.	
0	0.75	$3\cdot10^{-1}$	0.75	$.3\cdot10^{-1}$	0.5	0.5
1	0.625	$2\cdot10^{-1}$	0.80606124	$2\cdot10^{-1}$	0.61212248	0.61629718
2	0.5625	$6\cdot10^{-2}$	0.61534080	$3\cdot10^{-3}$	0.61549349	0.61546820
3	0.59375	$3\cdot10^{-2}$	0.61701328	$2\cdot10^{-3}$	0.61546816	0.61546816
4	0.609375	$2\cdot10^{-2}$	0.61701363	$2\cdot10^{-3}$		
5	0.6171875	$8\cdot10^{-3}$	0.61546816	0		
6	0.61328125	$4\cdot10^{-3}$				
...	.....	.....				
10	0.61547852	$4\cdot10^{-4}$				
...	.....	.....				
19	0.61546850	$5\cdot10^{-7}$				

در الگوریتم۱.۳ و ۳.۳، x_n نقطهٔ وسط بین کرانهای پایینی و بالایی a_n و b_n، بعد ازnبارست‌است، درحالی‌که ε_n معرف کران مربوط به‌خطا درx_n است‌که به‌وسیلهٔ الگوریتم محاسبه شده‌است. همگرایی‌سریع ومنظم[1] الگوریتمهای ۴.۳ و ۵.۳ قابل توجه‌است. روش تنصیف خیلی به‌کندی اما به‌طور یکنوا همگرا می‌شود. درحالی‌کـه ظاهراً روش تصحیح خطای اصلاح شده با «پرش» همگرا می‌شود،اگرچه جواب صحیح را نسبتاً سریع به‌دست می‌دهد. □

1. systematic

تمرین

۱-۱۰۳ مطلوب بازه‌ای است که ریشهٔ حقیقی و مثبت معادلهٔ $f(x)=x^۲-۲x-۲$ در آن واقع باشد. سپس با استفاده از الگوریتمهای ۱.۳ و ۲.۳ این ریشه را تا دو رقم با معنی صحیح محاسبه کنید. آیا می‌توانید برآورد کنید که در هر روش چند مرحله لازم است تا ۶ رقم با معنی صحیح نتیجه شود؟

۲-۱۰۳ برای مثالی که در متن داده شده است دو مرحله از روش تصحیح خطای اصلاح‌شده، (الگوریتم ۳.۳) را انجام دهید.

۳-۱۰۳ بسجمله‌ای $۱-۲x-x^۳$، یك ریشه بین ۱ و ۲ دارد. با استفاده از روش خط قاطع (الگوریتم ۴.۳) این ریشه را تا سه رقم با معنی صحیح محاسبه نمایید.

۴-۱۰۳ در الگوریتم ۱.۳، گیریم M معرف طول بازهٔ اولیهٔ $[a_۰, b_۰]$ باشد. همچنین نقاط وسط متوالی تولید شده به توسط روش تنصیف را با $\{x_۰, x_۱, x_۲, \ldots\}$ نشان می‌دهیم. نشان دهید که

$$|x_{i+۱}-x_i|=\frac{M}{۲^{i+۲}}$$

همچنین نشان دهید که تعداد بارستهای لازم، I، برای تضمین یك تقریب از یك ریشه با دقت ε با فرمول زیر تعیین می‌شود:

$$I>-۲-\frac{\log(\varepsilon/M)}{\log ۲}$$

۵-۱۰۳ روش تنصیف راهنگامی می‌توان به‌کار برد که نامساوی $f(a)f(b)<۰$ برقرار باشد. درصورتی که $f(x)$ در بازهٔ (a,b) بیش از یك ریشه داشته باشد، الگوریتم ۱.۳ معمولا جای کدامیك را مشخص می‌کند؟

۶-۱۰۳ با $a=۰$ و $b=۱$ کلیهٔ توابع زیر در بازهٔ (a,b) تغییر علامت می‌دهند یعنی $f(a)f(b)<۰$. الگوریتم ۱.۳ (روش تنصیف) چه نقطه‌ای را مشخص می‌کند؟ آیا این نقطه یك ریشهٔ $f(x)$ هست؟

$$f(x)=(۳x-۱)^{-۱} \qquad f(x)=\cos ۱۰x$$

$$f(x)=\begin{cases} ۱ & x\geqslant ۰.۳ \\ -۱ & x<۰.۳ \end{cases}$$

۷-۱۰۳ تابع $f(x)=e^{۲x}-e^x-۲$ دارای یك ریشه در بازهٔ $[۰,۱]$ است. با استفاده از روش نیوتن (الگوریتم ۵.۳) این ریشه را تا چهار رقم با معنی صحیح پیدا کنید.

۸-۱۰۳ تابع $f(x)=4\sin x-e^x$ روی بازهٔ [۰٫۵ , ۰] یك ریشه دارد. با استفاده از الگوریتم خط قاطع (۴.۳) این ریشه را تا چهار رقم با معنی صحیح پیدا کنید.

۹-۱۰۳ با استفاده از الگوریتم قاطع کوچکترین ریشهٔ مثبت بسجمله‌ای

$$p(x)=2x^3-3x-4$$

را تا سه رقم با معنی صحیح پیدا کنید.

۲۰۳ برنامه‌های فورترن برای بعضی از روشهای بارستی

هر گاه الگوریتمهای مذکور در بخشهای پیشین برای محاسبات به‌کار گرفته شوند، به‌جای اصطلاح مبهم «until satisfied» باید ملاك دقیقی برای پایان‌کار گذاشته شود. در این بخش تعدادی از راههای ممکن برای تعیین بارست پایانی معقول مورد بحث قرار می‌گیرند و ترجمه‌های الگوریتمهای ۱.۳ و ۳.۳ به‌زبان فورترن داده می‌شوند.

زیر برنامهٔ فورترن برای الگوریتم تنصیف ۱.۳

```
      SUBROUTINE BISECT ( F, A, B, XTOL, IFLAG )
C******  I N P U T  ******
C  F  NAME OF FUNCTION WHOSE ZERO IS SOUGHT. NAME MUST APPEAR IN AN
C       E X T E R N A L  STATEMENT IN THE CALLING PROGRAM.
C  A,B  ENDPOINTS OF THE INTERVAL WHEREIN A ZERO IS SOUGHT.
C  XTOL  DESIRED LENGTH OF OUTPUT INTERVAL.
C******  O U T P U T  ******
C  A,B  ENDPOINTS OF INTERVAL KNOWN TO CONTAIN A ZERO OF  F  .
C  IFLAG  AN INTEGER,
C         = -1, FAILURE SINCE  F  HAS SAME SIGN AT INPUT POINTS A AND B
C         = 0 , TERMINATION SINCE  ABS(A-B)/2 .LE. XTOL
C         = 1 , TERMINATION SINCE  ABS(A-B)/2 IS SO SMALL THAT ADDITION
C               TO  (A+B)/2  MAKES NO DIFFERENCE .
C******  M E T H O D  ******
C  THE BISECTION ALGORITHM 3.1 IS USED, IN WHICH THE INTERVAL KNOWN TO
C  CONTAIN A ZERO IS REPEATEDLY HALVED .
C
      INTEGER IFLAG
      REAL A,B,F,XTOL,   ERROR,FA,FM,XM
      FA = F(A)
      IF (FA*F(B) .GT. 0.)  THEN
         IFLAG = -1
         PRINT 601,A,B
  601    FORMAT(' F(X) IS OF SAME SIGN AT THE TWO ENDPOINTS',2E15.7)
                                          RETURN
      END IF
C
      ERROR = ABS(B-A)
C     DO WHILE  ERROR .GT. XTOL
    6    ERROR = ERROR/2.
         IF (ERROR .LE. XTOL)             RETURN
         XM = (A+B)/2.
C              CHECK FOR UNREASONABLE ERROR REQUIREMENT
         IF (XM + ERROR .EQ. XM) THEN
            IFLAG = 1
                                          RETURN
         END IF
         FM = F(XM)
C              CHOOSE NEW INTERVAL
         IF (FA*FM .GT. 0.) THEN
            A = XM
            FA = FM
         ELSE
            B = XM
         END IF
                                          GO TO 6
      END
```

در برنامه‌زیر از این‌زیر برنامه برای پیداکردن ریشهٔ معادلهٔ (۲.۳) که در بخش قبل مورد بحث قرار گرفت، استفاده می‌شود.

```
C  MAIN PROGRAM FOR TRYING OUT BISECTION ROUTINE
      INTEGER IFLAG
      REAL A,B,ERROR,XI
      EXTERNAL FF
      A = 1.
      B = 2.
      CALL BISECT ( FF, A, B, 1.E-6, IFLAG )
      IF (IFLAG .LT. 0)                    STOP
      XI = (A+B)/2.
      ERROR = ABS(A-B)/2.
      PRINT 600, XI,ERROR
  600 FORMAT(' THE ZERO IS ',E15.7,' PLUS/MINUS ',E15.7)
                                           STOP
      END
      REAL FUNCTION FF(X)
      REAL X
      FF = -1. - X*(1. - X*X)
      PRINT 600,X,FF
  600 FORMAT(' X, F(X) = ',2E15.7)
                                           RETURN
      END
```

اکنون راجع به‌زیر برنامهٔ نصف‌کردن، مذکور در بالا، توضیح بیشتری می‌دهیم. دراینجا زیر نمایه‌هایی‌راکه در الگوریتم۱.۳ مورد استفاده قرار گرفته‌اند حذف کرده‌ایم. مقادیر اولیه به‌وسیلهٔ برنامهٔ اصلی به‌زیر برنامه داده می‌شوندودرهرمرحله از برنامه، متغیرهای A و B مقادیرجاری کرانهای پایینی و بالایی ریشهٔ موردنظر هستند. به‌خصوص نقطهٔ وسط $XM=(A+B)/2$ همیشه بهترین براورد جاری برای ریشه‌است و اختلاف مطلق آن با ریشه، همیشه به‌وسیلهٔ فرمول زیر محدود می‌شود:

$$\text{خطا} = \frac{|A-B|}{2}$$

و به‌محض اینکه نامساوی زیر برقرار شود محاسبه خاتمه می‌یابد:

$$\text{خطا} \leqslant XTOL$$

در رابطهٔ بالا XTOL کرانهٔ خطای مطلق مفروض است. سپس برنامهٔ اصلی برای براورد ریشه مرتباً از مقادیر جاری A و B استفاده می‌کند. برنامهٔ اصلی بایستی علاوه بر A، B و XTOL، نام تابع $f(x)$ را هم‌که ریشهٔ آن موردنظر است در زبان **فورترن** ادامه دهد. ازآنجاکه فرض مختلف‌العلامه بودن $f(A)$ و $f(B)$ برای اجرای الگوریتم شرط عمده‌است، بنابراین‌درابتدا یك آزمایش برای برقراری این‌شرط انجام‌می‌شود. درصورتی‌که $f(A)$ و $f(B)$ مختلف‌العلامه نباشند، اجرای زیربرنامه فوراً خاتمه می‌یابد. متغیرخروجی IFLAG برای نشان دادن (فهماندن) این پیشامد نامطلوب به‌برنامهٔ اصلی به کار می‌رود.

زیربرنامه، به‌علت مذکور، هرگز مقادیر تابع را برای یك شناسه بیش از یك بار، محاسبه نمی‌کند، اما مقادیری از تابع راکه ممکن‌است درگامهای بعدی موردنیاز باشند، ضبط می‌کند. این‌خط‌مشی‌معقولی‌است زیرا ممکن‌است این زیربرنامه برای محاسبهٔ مقادیر

توابعی که بسیار وقت‌گیرند به‌کار رود. وبالاخره زیر برنامه درمقابل خطای مجاز نامعقول محافظت می‌شود: برای سادگی فرض کنید که کلیهٔ محاسبات با حساب با ممیز شناور وچهار رقم اعشاری انجام شوند وکرانهای A و B به‌حدی به‌ریشه نزدیك شده باشندکه

$$A = ۱۳۱٫۶ \quad و \quad B = ۱۳۱٫۷$$

درنتیجه

$$\text{خطا} = \frac{B-A}{۲} = ۰٫۰۵$$

دراین‌صورت

$$XM = \frac{A+B}{۲} = \frac{۲۶۳٫۳}{۲} = ۱۳۱٫۶ \quad یا \quad ۱۳۱٫۷$$

دو جواب فوق به‌چگونگی گردکردن تا چهار رقم بستگی دارد. درهرصورت داریم:

$$XM = A \quad یا \quad XM = B$$

به‌طوری‌که در آخر این‌مرحله نه A و نه B هیچ‌کدام تغییری نکرده‌اند. اگر اغماض XTOL کمتر از ۰٫۰۵ باشد، این زیر برنامه هیچگاه پایان نمی‌یابد. زیرا $|B-A|/۲$ هرگز کوچکتر از ۰٫۰۵ نخواهد شد. برای جلوگیری از این‌گونه حلقه‌های بی‌پایان که از خطای مجاز نامعقول ناشی می‌شوند (نامعقول، زیرا این امر ایجاب می‌کندکه A و B، بیش از آنچه‌که برای دوعدد با آن دقت ممکن‌است، به‌هم نزدیك شوند بی‌آنکه برهم منطبق شوند) این روال مقدار جاری خطا را به‌طریق زیرمحاسبه می‌کند. درابتدا داریم:

$$\text{خطا} = |B-A|$$

اما درآغازهر مرحله خطا نصف می‌شود، زیرا این مقدارکاهش‌خطا در هرمرحله در روش تنصیف وجود دارد. به‌محض‌اینکه خطا به‌اندازه‌ای کوچك شودکه جمع ممیزشناور آن با مقدار جاری XM، مقدار XM را تغییرندهد، دراین صورت برنامه خاتمه می‌یابد.

در قسمت بعدی، روش تصحیح خطای اصلاح‌شده (الگوریتم ۳.۳) را موردبررسی قرار می‌دهیم. برعکس‌روش‌تنصیف، روش‌تصحیح خطای اصلاح شده ایجاد یك بازهٔ تا حدممکن کوچك رابا استفاده ازحساب دقیق متناهی، تضمین‌نمی‌کند (به‌تمرین ۲.۳–۱ نگاه کنید). بنابراین برای این الگوریتم باید ملاك اختتام دیگری به‌کار گرفته شود.

برنامهٔ فورترن برای الگوریتم تصحیح خطای اصلاح‌شدهٔ ۳.۳

```
      SUBROUTINE MRGFLS ( F, A, B, XTOL, FTOL, NTOL, W, IFLAG )
C******  I N P U T  ******
C  F  NAME OF FUNCTION WHOSE ZERO IS SOUGHT. NAME MUST APPEAR IN AN
C       E X T E R N A L  STATEMENT IN THE CALLING PROGRAM .
C  A,B  ENDPOINTS OF INTERVAL WHEREIN ZERO IS SOUGHT.
C  XTOL  DESIRED LENGTH OF OUTPUT INTERVAL
C  FTOL  DESIRED SIZE OF  F(W)
C  NTOL  NO MORE THAN  NTOL  ITERATION STEPS WILL BE CARRIED OUT.
C******  O U T P U T  ******
C  A,B  ENDPOINTS OF INTERVAL CONTAINING THE ZERO .
C  W  BEST ESTIMATE OF THE ZERO .
C  IFLAG  AN INTEGER,
C     =-1, FAILURE, SINCE  F  HAS SAME SIGN AT INPUT POINTS  A, B .
C     = 0, TERMINATION BECAUSE  ABS(A-B) .LE. XTOL .
C     = 1, TERMINATION BECAUSE  ABS(F(W)) .LE. FTOL .
C     = 2, TERMINATION BECAUSE  NTOL  ITERATION STEPS WERE CARRIED OUT .
C******  M E T H O D  ******
C  THE MODIFIED REGULA FALSI ALGORITHM 3.3 IS USED. THIS MEANS THAT,
C  AT EACH STEP, LINEAR INTERPOLATION BETWEEN THE POINTS  (A,FA)  AND
C  (B,FB)  IS USED, WITH  FA*FB .LT. 0 ,TO GET A NEW POINT  (W,F(W))
C  WHICH REPLACES ONE OF THESE IN SUCH A WAY THAT AGAIN  FA*FB .LT. 0.
C  IN ADDITION, THE ORDINATE OF A POINT STAYING IN THE GAME FOR MORE
C  THAN ONE STEP IS CUT IN HALF AT EACH SUBSEQUENT STEP.
C
      INTEGER IFLAG,NTOL,   N
      REAL A,B,F,FTOL,W,XTOL,   FA,FB,FW,SIGNFA,PRVSFW
      FA = F(A)
      SIGNFA = SIGN(1., FA)
      FB = F(B)
      IF (SIGNFA*FB .GT. 0.) THEN
         PRINT 601,A,B
  601    FORMAT(' F(X) IS OF SAME SIGN AT THE TWO ENDPOINTS',2E15.7)
         IFLAG = -1
                                        RETURN
      END IF
C
      W = A
      FW = FA
      DO 20 N=1,NTOL
C                 CHECK IF INTERVAL IS SMALL ENOUGH.
         IF (ABS(A-B) .LE. XTOL) THEN
            IFLAG = 0
                                        RETURN
         END IF
C                 CHECK IF FUNCTION VALUE AT  W  IS SMALL ENOUGH .
         IF (ABS(FW) .LE. FTOL) THEN
            IFLAG = 1
                                        RETURN
         END IF
C                 GET NEW GUESS  W  BY LINEAR INTERPOLATION .
         W = (FA*B - FB*A)/(FA - FB)
         PRVSFW = SIGN(1.,FW)
         FW = F(W)
C                           CHANGE TO NEW INTERVAL
         IF (SIGNFA*FW .GT. 0.) THEN
            A = W
            FA = FW
            IF (FW*PRVSFW .GT. 0.)  FB = FB/2.
         ELSE
            B = W
            FB = FW
            IF (FW*PRVSFW .GT. 0.)  FA = FA/2.
         END IF
   20 CONTINUE
      PRINT 620,NTOL
  620 FORMAT(' NO CONVERGENCE IN ',I5,' ITERATIONS')
      IFLAG = 2
                                        RETURN
      END
```

اولا، این راهواره هنگامی پایان می‌پذیرد که قدرمطلق آخرین مقدار محاسبه شدهٔ تابع، بزرگتر از اغماض مفروض FTOL نباشد. درنتیجه به‌نظر می‌رسد که یك «ریشهٔ تقریبی» معادلهٔ $f(x)=0$ نقطه‌ای مثل x است که در آن $|f(x)|$ «کوچك» باشد. همچنین، ازآنجایی‌که این راهواره مرتباً عمل تقسیم برمقدار تابع را انجام می‌دهد، چنین پایان‌دادنی برای برنامه ضروری به‌نظر می‌رسد تا درحالتهای نهایی، تقسیم برصفرپیش نیاید.

ثانیاً، هنگامی که تعداد مرحله‌های بارستی بیشتر از عدد داده شدهٔ NTOL باشد، راهواره ختم می‌شود. به‌تعبیری NTOL مقدار محاسبه‌ای را ـ که برنامه‌نویس مایل است برای حل‌مسئلهٔ خود به‌کاربرد ـ مشخص‌می‌کند. به‌کارگرفتن این ملاك اختتام، برنامه‌نویس را در مقابل خطای غیرمعقول و خطاهای برنامه‌نویسی و همچنین احتمال اینکه مسئله‌ای را که می‌خواهد حل‌کند به‌طورکامل نفهمیده باشد، محافظت می‌کند. بنابراین بهتر است که چنین ملاك اختتام در هر روش بارستی مورداستفاده قرارگیرد.

زیربرنامهٔ MRGFLS مانند راهوارهٔ روش تنصیف، عدد صحیح IFLAG را برگشت می‌دهد، که نشان‌دهندهٔ علت اختتام بارست‌است و نیز جدیدترین مقدارکرانهای A و B را برای ریشهٔ موردنظر به‌دست می‌دهد. بالاخره این راهواره، مانند راهوارهٔ تنصیف هرگزمقدار تابع مفروض را برای یك شناسه بیشتر از یك‌بار محاسبه نمی‌کند.

الگوریتمهای ۴.۳ و ۵.۳ به‌ترتیب برای روش خط قاطع و روش نیوتن الزاماً ریشه‌را محصورنمی‌کنند، بلکه‌هردو، دنبالهٔ x_0، x_1، x_2، x_3، ... را تولیدمی‌کنندکه‌انتظار می‌رود به‌سمت ریشهٔ مطلوب ξ از تابع داده شدهٔ $f(x)=0$ همگرا باشد. بنابراین باید مقدمتاً به‌هردو الگوریتم به‌عنوان روشهایی برای پیداکردن نقاطی که درآنها قدرمطلق $f(x)$ «کوچك» باشد نگاه‌کرد. بارست هنگامی خاتمه می‌یابدکه قدرمطلق آخرین مقدار محاسبه شدهٔ تابع، کمتر از یك FTOL داده شده باشد.

همچنین بارست ممکن‌است زمانی خاتمه یابد که قدرمطلق‌تفاضل‌مقادیربارستهای متوالی کمتر از یك عدد مفروض XTOL باشد. بنابراین عادت براین‌است‌که برای‌اختتام روش خط قاطع یا روش نیوتن از یك یا هردو ملاك زیر استفاده می‌کنند:

$$|f(x_n)|\leqslant \text{FTOL} \quad \text{یا} \quad |x_n-x_{n-1}|\leqslant \text{XTOL} \qquad (۱۳.۳)$$

اگر اندازهٔ اعداد موردنظر را از قبل ندانیم، معمولا روش بهتر آن‌است‌که ملاك خطای نسبی مجاز به‌کارگرفته شود یعنی بارستها هنگامی خاتمه یابندکه داشته‌باشیم:

$$\frac{|f(x_n)|}{\text{FSIZE}}\leqslant \text{FTOL} \quad \text{یا} \quad |x_n-x_{n-1}|\leqslant \text{XTOL}*|x_n| \qquad (۱۴.۳)$$

درفرمول بالا FSIZE برآوردی‌است از مقدار $f(x)$ درهمسایگی‌ریشه‌ای‌که می‌خواهیم از راه بارست به‌آن برسیم.

در بخش ۶.۱ دربارهٔ خطراین نتیجه‌گیری که دنبالهٔ مفروضی همگراست، فقط به‌علت اینکه اختلاف دو جملهٔ متوالی آن «بسیارکم» است، بحث کرده‌ایم. با این‌حال دقیقاً از

همین ملاك درهردو راهوارهٔ فوق استفاده کرده‌ایم، زیرا در روش خط قاطع چنین ملاکی برای اجتناب از تقسیم برصفر ضروری‌است. همچنین در هردو روش، هنگامی‌که بارستها «به‌اندازهٔ کافی نزدیك» به‌ریشهٔ «معادله» باشند، تفاضل‌مابین‌آخرین دوبارست محاسبه شده، یك‌کران نسبتاً مناسبی برای خطای‌آخرین‌بارست‌است. به‌بیان ساده: اگر بارستهای متوالی اختلاف زیادی بـا هم نداشته باشند، آنگاه دلیل‌کمی برای ادامـهٔ بارستها وجود دارد. زیربرنامه‌های مربوط به‌روش نیوتن و روش خط قاطـع درمتن داده نشده‌اند و به‌عنوان تمرین به‌عهدهٔ دانشجویان واگذارشده‌اند.

□ **مثال۲.۳ (الف)**: ریشهٔ حقیقی و مثبت معادلهٔ زیر را پیداکنید:

$$x^۳ - x - ۱ = ۰$$

نتایج الگوریتمهای ۱.۳، ۲.۳، ۳.۳، ۴.۳ و ۵.۳ در جدول زیرکه شبیه جدول مثال۱.۳ است، داده شده‌اند.

n	Bisection x_n	e_n	Modified regula falsi x_n	e_n	Secant x_n	Newton x_n
					1.0	
0	1.5	$5 \cdot 10^{-1}$	1.5	$5 \cdot 10^{-1}$	2.0	1.0
1	1.25	$3 \cdot 10^{-1}$	1.5833333	$4 \cdot 10^{-1}$	1.1666667	1.5
2	1.375	$1 \cdot 10^{-1}$	1.6616541	$3 \cdot 10^{-1}$	1.2531120	1.3478261
3	1.3125	$6 \cdot 10^{-2}$	1.3249256	$2 \cdot 10^{-3}$	1.3372064	1.3252004
4	1.34375	$3 \cdot 10^{-2}$	1.3256293	$9 \cdot 10^{-4}$	1.3238501	1.3247182
5	1.328125	$2 \cdot 10^{-2}$	1.3256305	$9 \cdot 10^{-4}$	1.3247079	1.3247180
6	1.3203125	$8 \cdot 10^{-3}$	1.3247180	$4 \cdot 10^{-9}$	1.3247180	
...	.....	.....				
10	1.3247070	$5 \cdot 10^{-4}$				
...	.....	.....				
20	1.3247180	$5 \cdot 10^{-7}$				

□

□ **مثال۲.۲ (ب)**: مسئلهٔ به‌اصطلاح پیشقدری[1] در طرح مدارهای الکتریکی مستلزم حل معادله‌ای به‌شکل $f(v) = ۵۰\, I(e^{qv} - ۱) + v - ۲۰$ است، کـه در آن v ولتاژ، I شدت جریان و q پارامتری‌است‌که بارالکتریکی را با دمای مطلق مرتبط می‌سازد. در یك مسئلهٔ مهندسی نمونه‌ای بایستی این‌معادله به‌ازای مقادیرمختلف I و q حل شود تا چگونگی‌تغییر کوچکترین ریشهٔ مثبت $f(v)$ با تغییر پارامترها معلوم شود.

با استفاده از روش نیوتن‌کوچکترین ریشهٔ مثبت $f(v)$ را برای دومجموعهٔ مختلف ازمقادیرپارامتر $(I, q) = (۱۰^{-۸}, ۴۰)$ و $(I, q) = (۱۰^{-۶}, ۲۰)$ پیداکنید. فرض کنید $\mathrm{XTOL} = ۱۰^{-۸}$ و $\mathrm{FTOL} = ۱۰^{-۷}$.

نتایج با مقادیر اولیهٔ مذکور در زیر داده شده‌اند:

1. Biasing

$I = 10^{-8}, q = 40$		$I = 10^{-6}, q = 20$	
XN	FXN	XN	FXN
0.35000000	−1.0486984E + 00	0.55000000	1.5436571E + 00
0.39186072	1.6002561E + 00	0.52464183	3.2740616E − 01
0.37948785	3.3539839E − 01	0.51580645	2.6559725E − 02
0.37525497	2.6518162E − 02	0.51495562	2.1750479E − 04
0.37485947	2.0554242E − 04	0.51494853	1.4904231E − 08
0.37485636	1.2605184E − 08		

در این مثال انتخاب مقادیر اولیه کمتر موجب واگرایی[1] خواهد شد. □

تمرین

۱-۲-۳ کوشش کنید که ریشهٔ $x = ۱٫۳۳۳۳$ از معادلهٔ $(x - ۱٫۳۳۳۳)^۳ = ۰$ را با استفاده از الگوریتم تصحیح خطای اصلاح‌شدهٔ ۳.۳، تا پنج رقم اعشار صحیح پیدا کنید. این محاسبه را با بازهٔ $[۱ , ۲]$ شروع کنید. توضیح دهید که چرا در این مورد این روش نمی‌تواند یک بازهٔ «کوچکی» را که در برگیرندهٔ ریشه باشد به‌دست دهد؟

۲-۲-۳ در زیربرنامهٔ تنصیف، حتی با وجودی که مقادیر تابع FA و FM اعداد با ممیز شناور به خوبی تعریف شده‌ای هستند، اما به علت استفاده از حاصلضرب FA*FM ممکن است طی اجرای این زیربرنامه سرریز یا پی‌ریز رخ دهد. این نقص زیربرنامه را با به‌کاربستن تابع **فورترن** SIGN تصحیح کنید. آیا لازم است هر دفعه که A تغییر می‌کند، مقدار FA نیز به‌تناسب تغییر داده شود؟

۳-۲-۳ ثابت کنید تابع $f(x) = e^x - ۱ - x - x^۲/۲$ دقیقاً یک ریشهٔ $x = ۰$ دارد. (راهنمایی: از باقیماندهٔ بسط تیلر e^x پیرامون صفر استفاده کنید.) سپس تابع **فورترن**

```
F(X)=EXP(X)−۱−X−X*X/۲
```

را به‌ازای مقادیر مختلف از شناسهٔ X «نزدیک» به صفر ارزیابی کنید تا نشان دهید این تابع تغییر علامت زیادی می‌دهد، بنابراین دارای ریشه‌های زیادی «نزدیک» $X = ۰$ می‌باشد. از این حقیقت چند نتیجه‌گیری می‌توانید بکنید؟ بخصوص در رابطه با روش تنصیف و به‌طور کلی در رابطه با مفهوم (نظری) «ریشهٔ یک تابع»؟

۴-۲-۳ فرض کنید می‌خواهید نزدیکترین ریشهٔ معادلهٔ $\tan x = x$ به‌عدد ۵۰ را با استفاده از روش خط قاطع و ممیز شناور با نه رقم اعشاری محاسبه کنید. آیا منطقی خواهد بود که برای ختم محاسبات، از ملاک $|f(x_n)| \leqslant ۱۰^{-۸}$ استفاده کنیم؟

۵-۲-۳ «**تجسس دودویی**»[2] مسئلهٔ جدول مراجعه‌ای عبارت است از تعیین یک عدد I به‌ازای

1. divergence 2. Binary search

یك X مفروض به‌طوری‌که X مابین TABLE(I) و TABLE(I+۱) قرار گیرد، که در اینجا TABLE آرایه‌ای‌است یك‌بعدی متضمن یك دنبالهٔ صعودی (یا نزولی). یك زیربرنامهٔ **فورترن** بنویسید که در آن برای انجام این تجسس به‌نحو مؤثری از تنصیف استفاده شود. با فرض آنکه TABLE دارای n درایه باشد، در این راهواره چندبار X با یکی از درایه‌های TABLE مقایسه می‌شود؟

۲.۳-۶ یك زیربرنامه برای روش خط قاطع که برپایهٔ فرم (۷.۳) باشد بنویسید. برای ختم عملیات از یکی از ملاکهای خطای نسبی (۱۴.۳) استفاده کنید. همچنین برای مقایسهٔ خطای نسبی $|x_n|$ با XTOL* $|x_n - x_{n-1}| \leqslant$ تفاضل $x_n - x_{n-1}$ را دوباره محاسبه نکنید بلکه مقدار تصحیحی آن از بارست قبلی را به‌کار گیرید.

۲.۳-۷ یك زیربرنامه برای روش نیوتن بنویسید. مطمئن شوید که یك خروجی برای موردی‌که $f'(x_n)$ مساوی صفراست تهیه می‌شود. بعلاوه با به‌کارگرفتن ملاکهای (۱۳.۳) یا (۱۴.۳) باید ترتیبی‌داده شودکه درصورت برقرار نبودن همگرایی، برنامه بعد از تعداد معینی بارست NTOL خاتمه یابد.

۲.۳-۸ کوچکترین ریشهٔ مثبت هریك از معادلات زیررا با استفاده از الگوریتمهای ۱.۳، ۳.۳، ۴.۳ و ۵.۳ تا حداکثر دقت موجود روی کامپیوتر مورد استفاده‌تان پیدا کنید. نتایج خود، و تعداد بارستها و دقت عمل حاصله از الگوریتمها را با هم مقایسه کنید.

(الف) $e^{-x} - \sin x = 0$

(ب) $x - e^{-x^2} = 0$

(پ) $x^3 - x - 2 = 0$

۲.۳-۹ معادلهٔ داده شده‌در (مثال ۲.۳ب) را با استفاده از روش نیوتن و به‌ازای مقادیری از پارامترها برابر با $(I, q) = (30, 10^{-7})$ حل نمایید. سعی کنید این معادله را برای مقادیر اولیهٔ مختلف بین ۰ و ۴ حل کنید و به‌اثرات این مقادیر اولیه روی همگرایی و یا واگرایی توجه نمایید.

۳.۳ بارست نقطهٔ ثابت

در بخش ۱.۳ به‌بارست نقطهٔ ثابت به‌عنوان یك روش ممکن برای به‌دست آوردن ریشهٔ معادلهٔ

$$f(x) = 0 \qquad (۱۵.۳)$$

اشاره کردیم. در این روش از معادلهٔ (۱۵.۳) معادله‌ای به‌صورت

$$x = g(x) \qquad (۱۶.۳)$$

به‌دست می‌آوریم به‌گونه‌ای که هر جواب (۱۶۰۳)، یعنی هر «**نقطهٔ ثابت**» از $g(x)$، یك ریشهٔ (۱۵۰۳) باشد. برای انجام این‌عمل راههای مختلفی وجود دارد. برای مثال، اگر داشته باشیم

$$f(x)=x^{۲}-x-۲ \qquad (۱۷۰۳)$$

آنگاه نمونه‌هایی از انتخابهای ممکن برای تابع $g(x)$ عبارت‌اند از

$$g(x)=x^{۲}-۲ \qquad (الف)$$

$$g(x)=\sqrt{۲+x} \qquad (ب)$$

$$g(x)=۱+\frac{۲}{x} \qquad (پ)$$

(۱۸۰۳)

به‌ازای یك ثابت غیرصفر m

$$g(x)=x-\frac{x^{۲}-x-۲}{m} \qquad (ت)$$

هریك از $g(x)$های فوق‌یك **تابع بارست**[1] برای حل‌معادلهٔ(۱۵۰۳) نامیده می‌شود. [$f(x)$ به‌توسط (۱۷۰۳) داده شده‌است]. پس از انتخاب تابع بارست $g(x)$، برای حل معادلهٔ (۱۵۰۳) الگوریتم زیر را اجرا می‌کنیم.

الگوریتم ۶۰۳: بارست نقطهٔ ثابت. یك تابع بارست $g(x)$ و یك نقطهٔ شروع $x_۰$ داده شده‌اند.

For $n = 0, 1, 2, \ldots$, until satisfied, do:

  Calculate $x_{n+1} = g(x_n)$

برای اینکه این الگوریتم مفید واقع شود باید ثابت‌کنیم که:

(i) به‌ازای یك نقطهٔ شروع مفروض $x_۰$، می‌توانیم $x_۱$، $x_۲$، $x_۳$، ... را متوالیاً محاسبه کنیم.

(ii) دنبالهٔ $x_۱$، $x_۲$، ... به‌سمت نقطه‌ای مانند ξ همگراست.

(iii) حد ξ، یك نقطهٔ ثابت $g(x)$ است، یعنی $g(\xi)=\xi$.

مثال مربوط به‌تابع حقیقی مقدار $g(x)=-\sqrt{x}$ نشان می‌دهد که (i) یك شرط پیش‌پاافتادهٔ ساده‌ای نیست، زیرا دراین‌حالت $g(x)$ فقط به‌ازای $x \geqslant ۰$ تعریف شده‌است. با شروع ازیك نقطهٔ دلخواه $x_۰ > ۰$ خواهیم داشت $x_۱=g(x_۰) < ۰$. بنابراین $x_۲$را

1. Iteration function

نمی‌توانیم محاسبه کنیم. برای پرداختن به (i) از اصل (۱.۳) استفاده می‌کنیم:

اصل۱.۳ یك بازهٔ $I=[a, b]$ وجود دارد چنانکه به‌ازای جمیع نقاط $x \in I$، $g(x)$ معین‌است و $g(x) \in I$، یعنی تابع $g(x)$، I را بر روی خودش می‌نگارد.

به استقراء نسبت به n، از این اصل نتیجه می‌شود که اگر $x_0 \in I$، آنگاه به‌ازای جمیع مقادیر nداریم $x_n \in I$؛ بنابراین $x_{n+1}=g(x_n)$ معین و در I است.

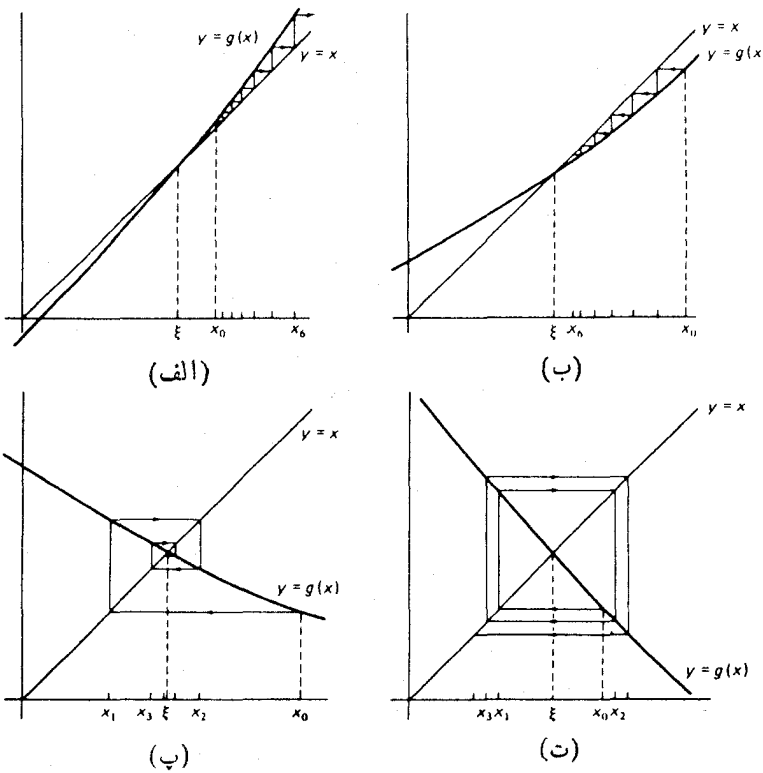
در بخش۱.۳، قبلا (iii) را موردبحث قراردادیم و ثابت کردیم که اگر $g(x)$ پیوسته باشد، شرط (iii) صادق است. برای پرداختن به (iii) از اصل ۲.۳ استفاده می‌کنیم.

اصل۲.۳ تابع بارست $g(x)$ در $I=[a, b]$ پیوسته‌است.

ملاحظه می‌کنیم که اصلهای ۱.۳ و ۲.۳ با هم، ایجاب می‌کنند که $g(x)$ دارای یك نقطهٔ ثابت در $I=[a, b]$ باشد. زیرا اگر $g(a)=a$ و یا $g(b)=b$، آنگاه وجود نقطهٔ ثابت روشن است. در غیر این صورت داریم $g(a) \neq a$ و $g(b) \neq b$. اما به‌موجب اصل۱.۳، $g(a)$ و $g(b)$ هردو در $I=[a, b]$ قرار می‌گیرند، بنابراین $g(a) > a$ و $g(b) < b$. این امر ایجاب می‌کند که تابع $h(x)=g(x)-x$ درشرطهای $h(a) > 0$ و $h(b) < 0$ صدق کند. از آنجاکه طبق اصل۲.۳، $h(x)$ روی I پیوسته‌است، بنابراین طبق قضیهٔ مقدارمیانگین برای توابع پیوسته(به‌بخش۷.۱نگاه‌کنید)، $h(x)$ باید در نقطه‌ای واقع در I صفرشود. اما این امر مبین این‌است که $g(x)$ یك نقطهٔ ثابتی در I دارد و این حکم ما را ثابت می‌کند.

برای بحث دربارهٔ (ii) که مربوط به‌همگرایی‌است، بهتر است که بارست را به‌طور نموداری انجام دهیم. این‌کار به‌نحو زیر صورت می‌گیرد. ازآنجاکه $x_n=g(x_{n-1})$، نقطهٔ $\{x_{n-1}, x_n\}$ برنمودار $g(x)$ قرار دارد. برای‌تعیین $\{x_n, x_{n+1}\}$از روی $\{x_{n-1}, x_n\}$، از نقطهٔ اخیر خطی به‌موازات محور xها رسم می‌کنیم. این خط، خط $y=x$ را درنقطهٔ $\{x_n, x_n\}$ قطع می‌کند. از این نقطه خطی به‌موازات محور yها رسم می‌کنیم تا نمودار $y=g(x)$ را درنقطهٔ $\{x_n, g(x_n)\}$ قطع کند. اما ازآنجاکه $g(x_n)=x_{n+1}$، این همان نقطهٔ مطلوب $\{x_n, x_{n+1}\}$ است. در شکل ۴.۳، چند مرحلهٔ اول از بارست نقطهٔ ثابت را برای چهار مورد نمونه، نشان داده‌ایم. توجه‌کنید که ξ یك نقطهٔ ثابت $g(x)$ است اگر، وتنها اگر، $y=g(x)$ و $y=x$ همدیگر را در (ξ, ξ) قطع کنند.

به‌طوری‌که شکل۴.۳ نشان می‌دهد، بارست نقطهٔ ثابت ممکن‌است اصلاهمگرا نباشد مانند همگرا نبودن درشکلهای ۴.۳ (الف) و ۴.۳ (د). به‌نظر می‌رسدکه همگرا بودن یا نبودن بارستها (درصورتی‌که $g(x)$ نقطهٔ ثابتی داشته باشد) بستگی به‌شیب $g(x)$ دارد. اگرقدرمطلق شیب $g(x)$ در نزدیکی نقطهٔ ثابت ξ از $g(x)$ خیلی بزرگ باشد، نمی‌توان به‌همگرایی امیدوار بود. بنابراین اصل۳.۳ را اختیار می‌کنیم.



شکل ۴۰۳ بازست نقطهٔ ثابت.

اصل ۳۰۳ تابع بازست در $I=[a, b]$ مشتقپذیر است. بعلاوه یك عدد ثابت غیرمنفی مانند $K<1$ وجود دارد به‌طوری‌که

$$|g'(x)| \leqslant K \qquad x \in I$$ به‌ازای جمیع نقاط

توجه شود که اصل ۳۰۳، اصل ۲۰۳ را ایجاب می‌کند. زیرا یك تابع مشتقپذیر حتماً پیوسته است.

قضیهٔ ۱۰۳ گیریم $g(x)$ تابع بازستی باشد که اصلهای ۱۰۳ و ۳۰۳ در آن صدق کنند. در این‌صورت $g(x)$ دقیقاً یك نقطهٔ ثابت ξ در I دارد و با شروع از هرنقطهٔ x_0 در I، دنبالهٔ x_1، x_2، x_3، ... که به‌توسط بازست نقطهٔ ثابت الگوریتم ۶۰۳ تولید می‌شود، در ξ همگراست.

برای اثبات این قضیه یادآور می‌شویم که قبلا وجود یك نقطهٔ ثابت برای $g(x)$ را در بازهٔ I ثابت کرده‌ایم. اکنون فرض کنید x_0 نقطهٔ دلخواهی در I باشد. پس چنانکه قبلا

اشاره شد، بارست نقطهٔ ثابت دنبالهٔ x_1، x_2، ... را تولید می‌کندکه بنابراصل۱.۳، همهٔ این نقاط در I قراردارند. اگرخطا در بارستن nام را با e_n نشان دهیم

$$e_n=\xi-x_n \qquad n=0,1,2,\dots$$

در این‌صورت چون $x_n=g(x_{n-1})$ و $g(\xi)=\xi$، بنابرقضیهٔ مقدار میانگین برای مشتق (به‌بخش۷.۱ نگاه‌کنید) مقداری مثل η_n بین ξ و x_{n-1} وجود دارد به‌طوری‌که

$$e_n=\xi-x_n=g(\xi)-g(x_{n-1})=g'(\eta_n)e_{n-1} \qquad (19.3)$$

بنابراین طبق اصل۳.۳، داریم .

$$|e_n|\leqslant K|e_{n-1}|$$

با استقراء نسبت به n نتیجه می‌شودکه

$$|e_n|\leqslant K|e_{n-1}|\leqslant K^2|e_{n-2}|\leqslant\cdots\leqslant K^n|e_0|$$

از آنجاکه $0\leqslant K<1$، داریم $\lim_{n\to\infty}K^n=0$، بنابراین صرفنظر از خطای اولیهٔ e_0 داریم

$$\lim_{n\to\infty}|e_n|=\lim_{n\to\infty}K^n|e_0|=0$$

اما این تساوی بیانگراین‌است‌که x_1، x_2، ... به‌سمت ξ همگراست. وهمچنین ثابت می‌کند که ξ تنها نقطهٔ ثابت $g(x)$ دربازهٔ I است. زیرا اگر ζ نیز یك نقطهٔ ثابت از $g(x)$ در I باشد،آنگاه با $x_0=\zeta$ خواهیم‌داشت $x_1=g(x_0)=\zeta$، بنابراین $|e_1|\leqslant K|e_0|=|e_0|$. چون $K<1$، نتیجه می‌گیریم‌که $|e_0|=0$ یا $\xi=\zeta$،که اثبات‌کامل می‌شود.

غالباً تعیین صحت وسقم اصل ۱.۳، کار بسیار مشکلی است. درچنین وضعیتی حکم ضعیفتر زیرممکن‌است، حداقل در صورتی‌که بارست «به‌اندازهٔ‌کافی نزدیك» به‌نقطهٔ ثابت شروع شده باشد، موفقیت را تضمین‌کند.

فرع. اگر $g(x)$ در یك بازهٔ بازی‌که متضمن نقطهٔ ثابت ξ است به‌طور پیوسته مشتقپذیر باشد، و هرگاه $|g'(\xi)|<1$، آنگاه مقداری چون $\varepsilon>0$ وجود داردکه هروقت شرط $|x_0-\xi|\leqslant\varepsilon$ برقرار باشد، بارست نقطهٔ ثابت با $g(x)$ همگراست.

در حقیقت چون $g'(x)$ نزدیك به ξ پیوسته‌است و $|g'(\xi)|<1$، به‌ازای هر K با شرط $|g'(\xi)|<K<1$، مقداری‌چون $\varepsilon>0$ وجود دارد. چنانکه‌به‌ازای هرx به‌شرط برقراری $|x-\xi|\leqslant\varepsilon$ نامساوی $|g'(x)|\leqslant K$ برقراراست. با تثبیت این K و ε مربوطه‌اش، به‌ازای $I=[\xi-\varepsilon,\xi+\varepsilon]$ اصل۳.۳ برقراراست. همچنین برای اصل۱.۳، گیریم x نقطهٔ دلخواهی‌از I باشد، لذا $|x-\xi|\leqslant\varepsilon$. پس مانند برهان قضیهٔ۱.۳ به‌ازای نقطه‌ای چون η بین x و ξ داریم

$$g(x)-\xi=g(x)-g(\xi)=g'(\eta)(x-\xi)$$

اما دراین صورت داریم

$$|g(x)-\xi|\leqslant|g'(\eta)|\ |x-\xi|\leqslant K\varepsilon<\varepsilon$$

این رابطه نشان می‌دهد که اگر x در I باشد، $g(x)$ نیز در I قراردارد. این امر صحت اصل ۱.۳ را نشان می‌دهد و نتیجه از قضیهٔ ۱.۳ حاصل می‌شود.

به‌علت این فرع، یك نقطهٔ ثابت ξ برای $g(x)$، که به‌ازای آن نامساوی $g'(\xi)<1$ برقرار باشد، غالباً یك **نقطهٔ جذب**[1] [برای بازست با $g(x)$] نامیده می‌شود.

دوباره تابع درجهٔ دوم (۱۷.۳) یعنی $f(x)=x^2-x-2$ را در نظر می‌گیریم. ریشه‌های این تابع ۲ و ۱ — هستند. فرض کنید می‌خواهیم ریشهٔ $\xi=2$ را با استفاده از بازست نقطهٔ ثابت محاسبه کنیم. اگر از تابع بازستی که به‌وسیلهٔ (۱۸.۳ الف) داده شده بود، یعنی

$$g(x)=x^2-2$$

استفاده کنیم، دراین‌صورت به‌ازای $x>1/2$، خواهیم داشت $g'(x)>1$. از اینجا نتیجه می‌شود که اصل ۳.۳ برای هربازهٔ متضمن $\xi=2$ برقرار نیست. یعنی $\xi=2$ یك نقطهٔ جذب نیست. در حقیقت برای این مثال می‌توان ثابت کرد که با شروع از یك نقطهٔ دلخواه x_0، دنبالهٔ x_1، x_2، ... که به‌وسیلهٔ بازست نقطهٔ ثابت تولید شده، تنها درصورتی به‌سمت $\xi=2$ همگرا می‌شود که به‌ازای مقداری چون n_0 و به‌ازای جمیع مقادیر $n\geqslant n_0$ تساوی $x_n=2$ برقرار باشد، یعنی اگر «تصادفاً» $\xi=2$ گرفته شده باشد (به‌تمرین ۳.۳-۱ نگاه کنید).

از سوی دیگر، اگر (۱۸.۳ ب)، یعنی $g(x)=\sqrt{2+x}$ را انتخاب کنیم، بنابراین

$$g'(x)=\frac{1}{2\sqrt{2+x}}$$

اما $x\geqslant 0$ ایجاب می‌کند که داشته باشیم $g(x)\geqslant 0$ و $g'(x)\leqslant 1/\sqrt{8}<1$ در حالی که مثلاً، $x\leqslant 7$ ایجاب می‌کند که $g(x)=\sqrt{2+x}\leqslant\sqrt{2+7}=3$، بنابراین با $I=[0,7]$ هردو اصل ۱.۳ و ۳.۳ صحیح‌اند و هرنقطهٔ $x_0\in[0,7]$ یك دنبالهٔ همگرا به‌دست می‌دهد. زیرا اگر $x_0=0$ را انتخاب کنیم داریم

$$\begin{aligned}
x_1&=\sqrt{2} &&=1{,}414221\\
x_2&=\sqrt{3{,}414221} &&=1{,}847775\\
x_3&=\sqrt{3{,}847775} &&=1{,}961857\\
x_4&=\sqrt{3{,}961857} &&=1{,}990036\\
x_5&=\sqrt{3{,}990036} &&=1{,}997589
\end{aligned}$$

1. Point af attraction

که آشکارا دیده می‌شود که به‌سمت ریشهٔ $\xi = ۲$ همگراست.

به‌عنوان مثال واقعگرایانه‌تر دیگر، معادلهٔ غیرجبری زیر را درنظر می‌گیریم

$$f(x) = x - ۲\sin x = ۰ \qquad (۲۰.۳)$$

طبیعیترین ترتیب در اینجا

$$x = ۲\sin x$$

است یعنی $g(x) = ۲\sin x$. مطالعهٔ خمهای $y = g(x)$ و $y = x$ نشان می‌دهد که یك ریشه بین $\pi/۳$ و $۲\pi/۳$ وجود دارد. بعلاوه اگر $\pi/۳ \leqslant x \leqslant ۲\pi/۳$، آنگاه $\sqrt{۳} \leqslant g(x) \leqslant ۲$. بنابراین اگر $\pi/۳ \leqslant a \leqslant \sqrt{۳}$ و $\sqrt{۳} \leqslant b \leqslant ۲\pi/۳$، آنگاه اصل۱.۳، صحیح‌است. بالاخره $g'(x) = ۲\cos x$ اکیداً[1] از ۱ تا ۱ — تنزل می‌کند وقتی‌که x از $\pi/۳$ تا $۲\pi/۳$ ترقی کند. در نتیجه هنگامی که $\pi/۳ < a \leqslant \sqrt{۳}$ و $\sqrt{۳} \leqslant b < ۲\pi/۳$، اصل۳.۳ صادق‌است و در نتیجه بارست نقطهٔ ثابت، بـا $g(x) = ۲\sin x$ به‌طرف ریشهٔ یکتای (۲۰.۳) در $[\pi/۳, ۲\pi/۳]$ همگرا می‌شود، هرگاه $x_۰ \in (\pi/۳, ۲\pi/۳)$.

□ **مثال۳.۳:** یك برنامهٔ کامپیوتری بنویسید که در آن برای پیداکردن کوچکترین ریشهٔ مثبت تابع $f(x) = e^{-x} - \sin x$ از بارست نقطهٔ ثابت استفاده شود.

اولین گام، انتخاب یك تابع بارست و یك مقدار آغازینی است که به یك بارست همگرا منجرشود. تابع $f(x) = ۰$ را دوباره به‌شکل زیر می‌نویسیم

$$x = x + e^{-x} - \sin x =: g(x)$$

اما چون داریم $f(۰ر۵) = ۰ر۱۲۷\ldots$ و $f(۰ر۷) = -۰ر۱۴۷\ldots$ لذا، کوچکترین ریشهٔ مثبت در بازهٔ $I = [۰ر۵, ۰ر۷]$ قراردارد. برای اینکه تحقیق کنیم $g(x)$ یك تابع بارست همگراست ملاحظه می‌کنیم که چون داریم

$$g'(x) = ۱ - e^{-x} - \cos x$$

لذا $g'(۰ر۵) = -۰ر۴۸\ldots$، $g'(۰ر۷) = -۰ر۲۶\ldots$ وچون $g'(x)$ یك تابع یکنوا روی I است، به‌ازای $x \in I$ داریم $|g'(x)| < ۱$. به‌همین‌ترتیب می‌توان ثابت کرد که به‌ازای جمیع مقادیر $x \in I$ داریم $۰ر۵ < g(x) < ۰ر۷$. بنابراین اگر $x_۰$ در I انتخاب شود بارست نقطهٔ ثابت همگرا خواهد بود.

برنامهٔ‌زیر با یك کامپیوتر CDC ۶۵۰۰ اجرا شده‌است. توجه کنید که اختتام نتیجه بخش این برنامه مستلزم برقراربودن هردو رابطهٔ زیر برای خطاست

$$|x_n - x_{n-۱}| < \mathrm{XTOL}|x_n|$$

---

1. Strictly

$$|f(x_n)| < \text{FTOL}$$

همچنین اگـر آزمونهای همگرایی پس از ۲۰ بارست صدق نکنند، برنامه مختومه تلقی می‌شود.

```
C  PROGRAM FOR EXAMPLE 3.3
      INTEGER J
      REAL ERROR,FTOL,XNEW,XOLD,XTOL,Y
C  THIS PROGRAM SOLVES THE EQUATION
C     EXP(-X) = SIN(X)
C  BY FIXED POINT ITERATION, USING THE ITERATION FUNCTION
      G(X) = EXP(-X) - SIN(X) + X
C
      DATA XTOL, FTOL /  1.E-8, 1.E-8 /
      PRINT 600
  600 FORMAT(9X,'XNEW',12X,'F(XNEW)',10X,'ERROR')
      XOLD = .6
      Y = G(XOLD) - XOLD
      PRINT 601, XOLD,Y
  601 FORMAT(3X,3E16.8)
      DO 10 J=1,20
         XNEW = G(XOLD)
         Y = G(XNEW) - XNEW
         ERROR = ABS(XNEW - XOLD)/ABS(XNEW)
         PRINT 601, XNEW,Y,ERROR
         IF (ERROR .LT. XTOL .OR. ABS(Y) .LT. FTOL) STOP
         XOLD = XNEW
   10 CONTINUE
      PRINT 610
  610 FORMAT(' FAILED TO CONVERGE IN 20 ITERATIONS')
                                          STOP
      END
```

برونداد برای مثال۳.۳

XN	F(XN)	ERROR
6.00000000E − 01	−1.58308373E − 02	
5.84169163E − 01	6.06240576E − 03	2.70997483E − 02
5.90231568E − 01	−2.35449276E − 03	1.02712326E − 02
5.87877076E − 01	9.09583240E − 04	4.00507667E − 03
5.88786659E − 01	−3.52118178E − 04	1.54484349E − 03
5.88434541E − 01	1.36203144E − 04	5.98398213E − 04
5.88570744E − 01	−5.27011849E − 05	2.31413378E − 04
5.88518043E − 01	2.03892661E − 05	8.95489706E − 05
5.88538432E − 01	−7.88865463E − 06	3.46438992E − 05
5.88530543E − 01	3.05208415E − 06	1.34039851E − 05
5.88533595E − 01	−1.18084550E − 06	5.18591321E − 06
5.88532415E − 01	4.56865632E − 07	2.00642389E − 06
5.88532871E − 01	−1.76760146E − 07	7.76278869E − 07
5.88532695E − 01	6.83880224E − 08	3.00340402E − 07
5.88532763E − 01	−2.64591478E − 08	1.16200876E − 07
5.88532737E − 01	1.02369739E − 08	4.49578182E − 08
5.88532747E − 01	−3.96065403E − 09	1.73940600E − 08
5.88532743E − 01	1.53236357E − 09	6.72970888E − 09

□

تمرین

۱-۳.۳ تحقیق کنید که بارست

$$x_{i+1} = x_i^2 - 2$$

به‌سمت جواب $\xi = 2$ از معادلهٔ

$$x^2 - x - 2 = 0$$

همگراست، فقط اگر به‌ازای مقداری مانند n_0 همهٔ بارستهای x_n با $n \geqslant n_0$، مساوی ۲ باشند، یعنی فقط «به‌طور تصادفی» همگراست.

۲-۳.۳ برای هریك از معادلات زیریك تابع بارست (ویك بازهٔ I) تعیین کنید به‌طوری که شرایط قضیهٔ ۱.۳ برقرار باشد (فرض کنید می‌خواهیم کوچکترین ریشهٔ مثبت را پیدا کنیم).

(الف) $x^3 - x - 1 = 0$ (ب) $x - \tan x = 0$

(پ) $e^{-x} - \cos x = 0$.

۳-۳.۳ برنامه‌ای بر مبنای الگوریتم ۶.۳ بنویسید و ازاین برنامه برای محاسبهٔ کوچکترین ریشه‌های معادله‌های مذکور در تمرین ۳.۳-۲ استفاده کنید.

۴-۳.۳ مطلوب است تعیین بزرگترین بازهٔ I با ویژگی زیر: به‌ازای جمیع مقادیر $x_0 \in I$، بارست نقطهٔ ثابت با تابع بارست

$$g(x) = x(2 - ax)$$

در صورت شروع با x_0، همگرا باشد. آیا اصلهای ۱.۳ و ۳.۳ که برای این I انتخاب شده‌اند صادق‌اند؟ حدود این بارست چه اعدادی می‌توانند باشند؟ آیا می‌توانید دلیل خوبی برای استفاده از این تابع بارست خاص بیاورید؟ (توجه کنید که بازهٔ I به‌ثابت a بستگی دارد).

۵-۳.۳ همان تمرین ۳.۳-۴ را برای تابع بارست $g(x) = (x + a/x)/2$ حل کنید.

۶-۳.۳ تابع $g(x) = \sqrt{1 + x^2}$ از اصل ۱.۳ برای بازهٔ $I = (-\infty, \infty)$ و در اصل ۳.۳ برای هر بازهٔ متناهی صادق است. ولی بارست نقطهٔ ثابت با این تابع بارست همگرا نیست. چرا؟

۷-۳.۳ معادلهٔ $e^x - 4x^2 = 0$ یك ریشه بین $x = 4$ و $x = 5$ دارد. نشان دهید که این ریشه را با استفاده از بارست نقطهٔ ثابت، با تابع بارست «طبیعی» $x = (1/2)e^{x/2}$ نمی‌توان پیدا کرد. آیا می‌توانید یك تابع بارستی پیدا کنید که جای صحیح این ریشه را تعیین کند؟

۸-۳.۳ معادلهٔ $e^x - 4x^2 = 0$ نیز یك ریشه بین $x = 0$ و $x = 1$ دارد. نشان دهید که اگر x_0 در بازهٔ $[0, 1]$ انتخاب شود تابع بارست $x = (1/2)e^{x/2}$ همگرا خواهد بود.

۴.۳ شتاب همگرایی برای بارست نقطهٔ ثابت

در این بخش میزان همگرایی[1] بارست نقطهٔ ثابت بررسی، ونشان داده می‌شودکه چگونه گاهی اوقات ازاطلاعات مربوط به‌میزان همگرایی می‌توان برای شتاب دادن به‌همگرایی استفاده‌کرد.

فرض‌کنیم که تابع بارست $g(x)$ به‌طور پیوسته مشتقپذیر باشد و با شروع ازهرنقطهٔ x_0 دنبالهٔ تولید شدهٔ $x_1, x_2, \ldots$ به‌وسیلهٔ بارست نقطهٔ ثابت به‌سمت یك نقطهٔ ξ همگرا باشد. دراین‌صورت نقطهٔ ξ یك نقطهٔ ثابت $g(x)$ خواهد بـود و طبق (۱۹.۳) نقطه‌ای مانند η_n بین ξ و x_n، $n = 1, 2, \ldots$، وجود دارد به‌طوری‌که داریم

$$e_{n+1} = \xi - x_{n+1} = g'(\eta_n)e_n \qquad (۲۱.۳)$$

ازآنجاکه $\lim_{n\to\infty} x_n = \xi$، درنتیجه $\lim \eta_n = \xi$. بنابراین، چون $g'(x)$ بنابرفرض پیوسته است، داریم

$$\lim_{n\to\infty} g'(\eta_n) = g'(\xi)$$

و در نتیجه

$$e_{n+1} = g'(\xi)e_n + \varepsilon_n e_n \qquad (۲۲.۳)$$

که دراینجا $\lim \varepsilon_n = 0$. بنابراین اگر $g'(\xi) \neq 0$، آنگاه بـه‌ازای n «به‌اندازهٔ کافی بزرگ» داریم

$$e_{n+1} \approx g'(\xi)e_n \qquad (۲۳.۳)$$

یعنی، خطای e_{n+1} در بارست $(n+1)$ام (کمابیش) به‌طورخطی به‌خطای e_n در بارست nام بستگی دارد. از این‌رو می‌گوییم دنبالـهٔ $x_0, x_1, x_2, \ldots$ **به‌طورخطی** بـه‌سمت ξ همگراست.

حال توجه‌کنیدکه می‌توان (۲۱.۳) را نسبت به ξ حل‌کرد. زیرا از

$$\xi - x_{n+1} = g'(\eta_n)(\xi - x_n) \qquad (۲۴.۳)$$

نتیجه می‌شود

$$\begin{aligned}\xi(1 - g'(\eta_n)) &= x_{n+1} - g'(\eta_n)x_n \\ &= [1 - g'(\eta_n)]x_{n+1} + g'(\eta_n)(x_{n+1} - x_n)\end{aligned}$$

بنابراین

---

1. Rate of convergence

$$\xi = x_{n+1} + \frac{g'(\eta_n)(x_{n+1} - x_n)}{1 - g'(\eta_n)} = x_{n+1} + \frac{x_{n+1} - x_n}{g'(\eta_n)^{-1} - 1} \qquad (۲۵.۳)$$

البته، عدد $g'(\eta_n)$ برما معلوم نیست. اما بنا برقضیهٔ مقدار میانگین برای مشتق، نسبت

$$r_n := \frac{x_n - x_{n-1}}{x_{n+1} - x_n} = \frac{x_n - x_{n-1}}{g(x_n) - g(x_{n-1})} = g'(\zeta_n)^{-1} \qquad (۲۶.۳)$$

به‌ازای مقداری مانند ζ_n بین x_n و x_{n-1} برما معلوم‌است. بنابراین بـه‌ازای n «به‌اندازهٔ کافی بزرگ» داریم

$$r_n = \frac{1}{g'(\zeta_n)} \approx \frac{1}{g'(\xi)} \approx \frac{1}{g'(\eta_n)}$$

و دراین‌صورت نقطهٔ

$$\hat{x}_n = x_{n+1} + \frac{x_{n+1} - x_n}{r_n - 1} \qquad \text{با شرط} \qquad r_n = \frac{x_n - x_{n-1}}{x_{n+1} - x_n} \qquad (۲۷.۳)$$

تقریبی به‌مراتب بهتر از x_n یا x_{n+1} برای ξ است.

این مطلب را از لحاظ نموداری نیزمی‌توان دید. زیرا در واقع، ما معادلهٔ(۲۷.۳) را از حل(۲۴.۳) نسبت بـه ξ، پس از قراردادن عدد $g[x_{n-1}, x_n]$ به‌جای $g'(\eta_n)$ و نشان دادن جواب حاصل با $\hat{x}_n$، به‌دست آورده‌ایم. بنابراین داریم

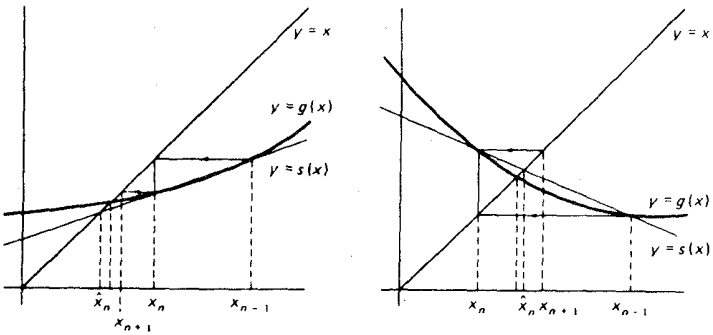
$$\hat{x}_n - x_{n+1} = g[x_{n-1}, x_n](\hat{x}_n - x_n)$$

از آنجاکه $x_{n+1} = g(x_n)$، این رابطه نشان می‌دهدکه $\hat{x}_n$ یك نقطهٔ ثابت ازخط‌مستقیم

$$s(x) = g(x_n) + g[x_{n-1}, x_n](x - x_n)$$

است. ما این رابطه را به‌عنوان درونیاب خطی‌برای $g(x)$ در x_{n-1} و x_n می‌شماریم.اما اگرشیب $g(x)$ بین‌نقاط x_{n-1} و ξ اندکی تغییرکند، یعنی‌اگر$g(x)$ تقریباً یك خط‌مستقیم واصل بین دونقطهٔ x_{n-1} و ξ باشد، دراین صورت وتر $s(x)$ می‌باید تقریب بسیار خوبی برای $g(x)$ درآن بازه باشد، ازاین‌رونقطهٔ ثابت $\hat{x}_n$ از این‌خط قاطع می‌باید یك تقریب بسیار خوبی برای نقطهٔ ثابت ξ از $g(x)$ باشد؛ به‌شکل۵.۳ نگاه‌کنید.

درعمل نمی‌توان ثابت‌کردکه یك x_n خاصی،«به‌اندازهٔ کافی نزدیك» به ξ است تا بتوان $\hat{x}_n$ را به‌عنوان تقریبی بهتر از x_n یا x_{n+1} برای ξگرفت. اما می‌توان بـا بررسی نسبتهای r_{n-1} و r_n این فرض راکـه x_n «به‌اندازهٔ کافی نزدیك» است، آزمایش کـرد. درصورتی که این نسبتها تقریباً ثابت باشند، این فرض راکه شیب $g(x)$ دربازهٔ موردنظر خیلی کم‌تغییرمی‌کند، می‌پذیریم. بنابراین می‌توانیم عقیده داشته باشیم‌که وتر$s(x)$ تقریب «به‌اندازهٔ‌کافی خوبی» برای $g(x)$ است تا بتواند $\hat{x}_n$ را تقریبی خیلی بهتر از x_n برای ξ



شکل ۵.۳ شتاب همگرایی برای بارست نقطهٔ ثابت.

بسازد. به‌ویژه دراین حالت قبول می‌کنیم که $|\hat{x}-x_n|$ برآورد خوبی برای خطای $|e_n|$ است.

□ **مثال ۴.۳:** معادلهٔ

$$1.5\,x-\tan x=0.1 \qquad (28.3)$$

دارای یك ریشهٔ ... ۰.۲۰۵۹۲۱۶۹۵۱۰ = ξ است. تابع بارست زیر را انتخاب

$$g(x)=\frac{0.1+\tan x}{1.5}$$

و با شروع از $x_0=0$ دنبالهٔ x_1، x_2، ... را، با استفاده از بارست نقطهٔ ثابت تولید می‌کنیم. بعضی از x_nها در جدول زیر فهرست شده‌اند. به‌نظر می‌رسد که این دنباله آهسته اما به‌طور مطمئن به‌سمت ξ همگراست. دنبالهٔ نسبتهای r_n را نیز محاسبه می‌کنیم. این دنباله نیز درهمین جدول داده شده‌است.

n	x_n	$\hat{x}_n$	r_n	y_n
	0.0			0.0
1	0.0666667	0.2005954	1.4978	0.2005954
2	0.1111771	0.2024021	1.4879	0.2059125
3	0.1410916	0.2039180	1.4761	0.2059217
4	0.1613569	0.2048536	1.4659	
5	0.1751813	0.2053721	1.4579	
...	.....	.....	.....	
10	0.2009941	0.2059055	1.4408	
15	0.2051206	0.2059213	1.4379	
20	0.2057911	0.2059217	1.4374	
25	0.2059004		1.4373	
30	0.2059182		1.4372	

بخصوص با مشاهدهٔ

$$r_1 = 1.49780\ldots \qquad r_2 = 1.48790\ldots \qquad r_3 = 1.47610\ldots$$

فکر می‌کنیم که این مقادیر «به‌اندازهٔ کافی» ثابت هستند تا نتیجه بگیریم که به‌ازای جمیع مقادیر $n \geqslant 1$، $\hat{x}_n$ تقریبی بهتر از x_n برای ξ است. این امر با درج مقادیر $\hat{x}_n$ در جدول نیز تأیید شده‌است. □

خواه یك $\hat{x}_n$ خاص تقریبی بهتر از x_n برای ξ باشد یا نه، می‌توان ثابت کرد که دنبالهٔ $\hat{x}_1, \hat{x}_2, \ldots$ سریعتر از دنبالهٔ اصلی $x_0, x_1, \ldots$ به‌سمت ξ همگرا می‌شود، یعنی داریم

$$\hat{x}_n = \xi + o(e_n) \qquad (۲۹.۳)$$

[برای تعریف $o(\ )$ به بخش ۶.۱ نگاه کنید.]

این روند استخراج یك دنبالهٔ سریعتر همگرا $\hat{x}_1, \hat{x}_2, \ldots$ با استفاده از رابطهٔ (۲۷.۳) از یك دنبالهٔ به‌طور خطی همگرای $x_0, x_1, x_2, \ldots$ را معمولاً **روند Δ^2ی ایتکن**[1] نامند. با استفاده از علامات اختصاری

$$\Delta x_k = x_{k+1} - x_k \qquad \Delta^2 x_k = \Delta(\Delta x_k) = \Delta x_{k+1} - \Delta x_k$$

از بخش (۶.۲)، رابطهٔ (۲۷.۳) را می‌توان به‌شکل زیر نوشت

$$\hat{x}_n = x_{n+1} - \frac{(\Delta x_n)^2}{\Delta^2 x_{n-1}} \qquad (۳۰.۳)$$

و از این‌روست که نام «روند Δ^2» بر آن نهاده شده‌است. این روند برای هر دنبالهٔ به‌طور خطی همگرا کاربرد دارد، خواه این دنباله بر اثر بارست نقطهٔ ثابت تولید شده یا نشده باشد.

الگوریتم ۷.۳: روند Δ^2ی ایتکن. دنبالهٔ $x_0, x_1, x_2, \ldots$ که به‌سمت ξ همگراست داده شده است. با استفاده از (۳۰.۳) دنبالهٔ $\hat{x}_1, \hat{x}_2, \ldots$ را محاسبه کنید.

اگر دنبالهٔ $x_0, x_1, x_2 \ldots$ به‌طور خطی به‌سمت ξ همگرا شود، یعنی اگر به‌ازای مقداری مانند $K \neq 0$، داشته باشیم

$$\xi - x_{n+1} = K(\xi - x_n) + o(\xi - x_n)$$

در این‌صورت

$$\hat{x}_n = \xi + o(\xi - x_n)$$

---

1. Aitken

بعلاوه اگر از مقدار معینی از k به‌بعد، دنبالهٔ نسبتهای تفاضلی، یعنی $\Delta x_k/\Delta x_{k+1}$، $\Delta x_{k-1}/\Delta x_k$، ... تقریباً ثابت باشد، آنگاه می‌توان فرض کرد که $\hat{x}_k$ تقریبی بهتر از x_k برای ξ است. در این صورت بخصوص $|\hat{x}_k - x_k|$ برآورد مطلوبی برای خطای $|x_k - \xi|$ است.

اگر درمورد بارست نقطهٔ ثابت، تشخیص داده شودکه $\hat{x}_k$ تقریبی بهتر از x_k برای ξ است، دراین‌صورت قطعاً محاسبهٔ دنبالهٔ x_{k+1}، x_{k+2}، ... اتلاف وقت خواهد بود. در این حالت بجا به‌نظر می‌رسدکه بارست نقطهٔ ثابت تازه، با حدس اولیهٔ $\hat{x}_k$ شروع شود. این امر به‌الگوریتم زیرمنجرمی‌شود.

الگوریتم ۸۰۳: بارست استفنسن[1] . تابع بارست $g(x)$ و نقطهٔ y_0 داده شده‌اند.

For $n = 0, 1, 2, \ldots,$ until satisfied, do:
- $x_0 := y_n$
- Calculate $x_1 = g(x_0), x_2 = g(x_1)$
- Calculate $d = \Delta x_1, r = \Delta x_0/d$
- Calculate $y_{n+1} = x_2 + d/(r - 1)$

یك مرحلهٔ این الگوریتم شامل دومرحله از بارست نقطهٔ ثابت و به‌دنبال آن یکبار کاربرد (۲۷.۳) می‌باشد. نتیجهٔ این سه بارست موجود برای محاسبهٔ مقدار شروع برای مرحلهٔ بعدی به‌کارمی‌رود.

درجدول فوق y_n حاصل ازکاربرداین‌الگوریتم برای مثال (۴.۳) درج شده‌است. مشاهده می‌شودکه‌کلیهٔ رقمهای اعشاری y_3 صحیح‌اند.

تمرین

۱-۴.۳ فرض‌کنید خطای بارست نقطهٔ ثابتی، به‌ازای مقداری از k، $|k| < 1$، در رابطهٔ بازگشتی

$$e_{n+1} = ke_n$$

صدق‌کند. برای تعداد بارستها، N، عبارتی‌چنان پیداکنیدکه خطای اولیهٔ e_0 را به‌اندازهٔ مضربی از 10^{-m} $(m > 0)$کاهش دهد.

۲-۴.۳ بارست نقطهٔ ثابت برای معادلهٔ

$$f(x) = 0.5 - x + 0.2 \sin x = 0$$

---

1. Steffensen

به‌کار رفته و تقریبهای متوالی زیر که به‌ترتیب درجدول زیر داده شده‌اند، پدید آمده‌اند.

با استفاده از الگوریتم ۷.۳ ایتکن دنبالهٔ شتابدار $\hat{x}_k$ ونسبتهای r_k را محاسبه کنید. با استفاده از نسبتهای r_k مقدار تقریبی $g'(\xi)$ را نیز محاسبه کنید.

k	x_k
0	0.5000 0000
1	0.5958 8511
2	0.6122 4830
3	0.6149 4176
4	0.6153 8219
5	0.6154 5412
6	0.6154 6587
7	0.6154 6779
8	0.6154 6810
9	0.6154 6815

۳-۴.۳ برنامه‌ای برای روش شتابدار استفنسن (الگوریتم ۸.۳) بنویسید. با استفاده از این برنامه کوچکترین ریشهٔ مثبت تابع مذکور در تمرین ۳.۴-۲ را با استفاده از تابع بارست $g(x)=$ ۰٫۵ + ۰٫۲ $\sin x$ و $x_0=$ ۰٫۵ پیدا کنید.

۴-۴.۳ در بخش ۳.۳ نشان داده شد که بارست نقطهٔ ثابت برای معادلهٔ

$$x_{i+1}=\sqrt{2+x_i}$$

دنبالهٔ تقریبهای زیر را برای ریشهٔ مثبت تابع $f(x)=x^2-x-2$ تولید می‌کند:

$$x_0=0 \qquad x_1=1.414214 \qquad x_2=1.847776$$

$$x_3=1.961570 \qquad x_4=1.990037 \qquad x_5=1.997589$$

با استفاده از الگوریتم ایتکن ۷.۳ دنبالهٔ شتابدار مربوطه را محاسبه و به‌بودمیزان همگرایی به‌طرف ریشهٔ $\xi=2$ توجه کنید.

۵-۴.۳ تابع بارست $g(x)=x-x^3$ را درنظر بگیرید. نقطهٔ ثابت و یکتای $g(x)$ را پیدا کنید. ثابت کنید که اگر $x_0\in(-1,1)$، بارست نقطهٔ ثابت با این تابع بارست به‌طرف نقطهٔ ثابت یکتای ξی $g(x)$ همگرا می‌شود. (راهنمایی: ازاین واقعیت که اگر روابط $x_n\leqslant x_{n+1}\leqslant\cdots\leqslant c$ به‌ازای یك مقدار ثابت c برقرار باشد، آنگاه این دنباله همگرا می‌شود استفاده کنید). آیا نامساوی $|e_n|\leqslant k|e_{n-1}|$ به‌ازای $k<1$ و جمیع مقادیر n صحیح است؟

۵.۳* همگرایی نیوتن و روشهای خط قاطع

در بخش قبل، ثابت شد که برای بارست نقطهٔ ثابت، خطای e_n در nامین بارستن x_n در رابطهٔ

$$e_{n+1} \approx g'(\xi)e_n \qquad (۳۱.۳)$$

صدق می‌کند، به‌شرط آنکه n به‌اندازهٔ کافی بزرگ و $g(x)$ به‌طور پیوسته مشتقپذیر باشد. ظاهراً هرچه $|g'(\xi)|$ کوچکتر باشد وقتی $n \to \infty$، e_n سریعتر به‌سمت صفر میل می‌کند. بنابراین حداکثر سرعت همگرایی بارست نقطهٔ ثابت باید هنگامی باشد که $g'(\xi) = ۰$.

اگر $g(x)$ دوبار مشتقپذیر باشد، به‌ازای مقداری از ζ_n بین ξ و x_n، از فرمول تیلر خواهیم داشت

$$e_{n+1} = \xi - x_{n+1} = g(\xi) - g(x_n)$$

$$= -g'(\xi)(x_n - \xi) - \frac{۱}{۲}g''(\zeta_n)(x_n - \xi)^۲$$

یعنی

$$e_{n+1} = g'(\xi)e_n - \frac{۱}{۲}g''(\zeta_n)e_n^۲ \qquad (۳۲.۳)$$

بنابراین اگر $g'(\xi) = ۰$ و $g''(x)$ در ξ پیوسته باشد، آنگاه داریم

$$e_{n+1} \approx -\frac{۱}{۲}g''(\xi)e_n^۲ \qquad \text{برای } n \text{ به‌اندازهٔ کافی بزرگ} \qquad (۳۳.۳)$$

در این حالت، e_{n+1} (کمابیش) تابع درجهٔ دومی از e_n است. بنابراین گوییم که دنبالهٔ $x_۱, x_۲, \ldots$ **به‌طور مربعی** به‌سمت ξ همگرا می‌شود.

بدیهی است که یك چنین تابع بارستی بسیار مطلوب است. مقبولیت روش نیوتن را می‌توان به‌این حقیقت منتسب دانست که تابع بارست آن یعنی

$$g(x) = x - \frac{f(x)}{f'(x)} \qquad (۳۴.۳)$$

معمولا از این نوع (مربعی) است.

پیش از اینکه ثابت کنیم روش نیوتن (در صورت همگرا بودن) به طور مربعی همگرا می‌شود، مثال سادهٔ زیر را درنظر می‌گیریم.

□ **مثال:** تعیین جذر مثبت یك عدد مثبت A، با پیداکردن جواب معادلهٔ

$$f(x) = x^۲ - A = ۰$$

هم‌ارز است. در اینجا $f'(x) = ۲x$، و با قراردادن آن در (۳۴.۳)، تابع بارست

$$g(x) = x - \frac{x^۲ - A}{۲x} = \frac{۱}{۲}\left(x + \frac{A}{x}\right) \qquad (۳۵.۳)$$

به‌دست می‌آید. پیداکردن جذر A به‌تابع بازست زیر منجرمی‌شود

$$x_{n+1}=\frac{1}{2}\left(x_n+\frac{A}{x_n}\right) \qquad (۳۶.۳)$$

به‌ویژه چنانچه $A=2$ و $x_0=2$ اختیار شود، نتیجهٔ بازست نقطهٔ ثابت با (۳۶.۳) به‌صورت زیردرمی‌آید

$x_0=2$

$x_1=1.5$ $\qquad r_1=6$

$x_2=1.4166666666\ldots$ $\qquad r_2=34$

$x_3=1.4142156862\ldots$ $\qquad r_3=1,154$

$x_4=1.4142135623\ldots$ $\qquad r_4=1,333171$

$x_5=1.4142135623\ldots$

دنبالهٔ بازستها آشکارا به‌سرعت همگرامی‌شود، ودنبالهٔ متناظرآن $r_1, r_2, \ldots$ که ازنسبتهای $r_n=\Delta x_{n-1}/\Delta x_n$ به‌دست می‌آید، به‌سمت ∞ همگرا می‌شود. از آنجا کـه همگرایی بازست نقطهٔ ثابت مستلزم برقراری تساوی $\lim_{n\to\infty} r_n^{-1}=g'(\xi)$ است، مثال فوق این ادعا را روشن می‌کند وهمگرایی سریع و مطلوب روش نیوتن را نشان می‌دهد. □

می‌توان همگرایی مربعی روش نیوتن را این‌گونه نشان دادکـه اگر $f(\xi)=0$ و $f'(\xi)\neq 0$، آنگاه تابع بازست

$$g(x)=x-\frac{f(x)}{f'(x)}$$

در یك بازهٔ باز در همسایگی ξ به‌طور پیوسته مشتقپذیـر است و $g'(\xi)=0$. در نتیجه با درنظرگرفتن فرع قضیهٔ ۱.۳ مقداری مانند $\varepsilon>0$ وجود دارد، به‌طوری که بـه‌ازای هر انتخاب x_0 که در رابطهٔ $|x_0-\xi|\leqslant\varepsilon$ صدق‌کند بازست نقطهٔ ثابت بـا $g(x)$ به‌سمت ξ همگرا می‌شود. اما کاراتر به‌نظر می‌رسدکه اثبات همگرایی مربعی نیوتن، مستقیماً صورت گیرد و همزمان با آن برهانی برای همگرایی روش خط قاطع فراهم‌گردد.

خطا در روش نیوتن و در روش خط قاطع را می‌توان همزمان استخراج کرد. در هردو روش تابع $f(x)$ در دونقطه مثل، α و β، به‌وسیلهٔ یك خط مستقیم

$$p(x)=f(\alpha)+f[\alpha,\beta](x-\alpha)$$

که صفرآن یعنی

$$\hat{\xi}=\alpha-\frac{f(\alpha)}{f[\alpha,\beta]}$$

به‌عنوان تقریب بعدی برای ریشهٔ دقیق $f(x)$ مورد استفاده قرار می‌گیرد، درونیابی می‌شود. در روش خط قاطع با انتخاب $\alpha=x_n$ و $\beta=x_{n-1}$، مقدار $x_{n+1}=\hat{\xi}$ را به‌دست می‌آوریم، درحالی‌که در روش نیوتن می‌گیریم $\alpha=\beta=x_n$.

در هردوحالت به‌موجب (۳۷.۲) داریم

$$f(x)=f(\alpha)+f[\alpha,\beta](x-\alpha)+f[\alpha,\beta,x](x-\alpha)(x-\beta)$$

که این معادله به‌ازای جمیع مقادیر x برقرار است. اکنون اگر $x=\xi$ انتخاب گردد که همان ریشهٔ مطلوب است، دراین‌صورت

$$0=f(\xi)=f(\alpha)+f[\alpha,\beta](\xi-\alpha)+f[\alpha,\beta,\xi](\xi-\alpha)(\xi-\beta)$$

و بنابراین

$$f[\alpha,\beta](\xi-\alpha)=-f(\alpha)-f[\alpha,\beta,\xi](\xi-\alpha)(\xi-\beta)$$

حال از حل معادلهٔ فوق نسبت به ξ، خواهیم داشت:

$$\xi=\alpha-\frac{f(\alpha)}{f[\alpha,\beta]}-\frac{f[\alpha,\beta,\xi]}{f[\alpha,\beta]}(\xi-\alpha)(\xi-\beta)$$

یا

$$\xi=\hat{\xi}-\frac{f[\alpha,\beta,\xi]}{f[\alpha,\beta]}(\xi-\alpha)(\xi-\beta) \qquad (۳۷.۳)$$

اما معادلهٔ (۳۷.۳) را می‌توان برای به‌دست‌آوردن معادله‌های خطا در روشهای نیوتن و خط قاطع به‌کار گرفت. برای روش نیوتن قرار می‌دهیم $\alpha=\beta=x_n$؛ و با توجه به‌اینکه $e_j=\xi-x_j$ و $x_{n+1}=\hat{\xi}$، از معادلهٔ (۳۷.۳) چنین به‌دست می‌آوریم:

$$e_{n+1}=-\frac{f[x_n,x_n,\xi]}{f[x_n,x_n]}e_n^2 \qquad (۳۸.۳)$$

باتوجه به‌اینکه‌روابط $f[x_n,x_n]=f'(x_n)$ و $f[x_n,x_n,\xi]=(1/2)f''(\eta_n)$ به‌ازای مقداری چون η_n بین x_n و ξ برقرار هستند، می‌توانیم (۳۸.۳) را دوباره به‌صورت زیر بنویسیم

$$e_{n+1}=-\frac{1}{2}\frac{f''(\eta_n)}{f'(x_n)}e_n^2 \qquad (۳۸.۳\text{ الف})$$

این معادله نشان می‌دهد که روش نیوتن به‌طور مربعی همگرا می‌شود، زیرا که e_{n+1} تقریباً

با مربع e_n متناسب است.

برای به‌دست آوردن معادله‌ای برای خطا در روش خط قاطع، در معادلهٔ (۳۷.۳) قرار می‌دهیم $\alpha = x_n$ و $\beta = x_{n-1}$ و خواهیم داشت

$$e_{n+1} = -\frac{f[x_{n-1}, x_n, \xi]}{f[x_{n-1}, x_n]} e_n e_{n-1} \qquad (۳۹.۳)$$

این معادله نشان می‌دهد که خطا در $(n+1)$امین بار ستن تقریباً متناسب با حاصلضرب خطا در مراحل nام و $(n-1)$ام است. همچنین از آنجا که $f[x_{n-1}, x_n, \xi] = (1/2) f''(\zeta_n)$ و $f[x_{n-1}, x_n] = f'(\eta_n)$، ($\eta_n$ و ζ_n نقاطی بین x_n و x_{n-1} و ξ هستند.) پس برای n به‌اندازهٔ کافی بزرگ معادلهٔ (۳۹.۳) به‌صورت زیر درخواهد آمد

$$e_{n+1} \approx -\frac{1}{2}\frac{f''(\xi)}{f'(\xi)} e_n e_{n-1} \qquad (۳۹.۳ \text{ الف})$$

برای دقیقتر ساختن مفهوم **مرتبهٔ همگرایی**، تعریف زیر آورده می‌شود:

تعریف ۱۰۳: مرتبهٔ همگرایی. گیریم x_0، x_1، x_2، ... دنباله‌ای باشد که به‌سمت عدد ξ همگرا می‌شود، و قرار می‌دهیم $e_n = \xi - x_n$. اگر عددی مانند p و ثابتی چون $C \neq 0$ طوری وجود داشته باشد که داشته باشیم

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = C$$

آنگاه p **مرتبهٔ همگرایی** این دنباله و C، **ثابت مجانبی خطا** خوانده می‌شود.

در حالت کلی برای بار ست نقطهٔ ثابت که برپایهٔ $x = g(x)$ استوار است، داریم

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|} = |g'(\xi)|$$

لذا مرتبهٔ همگرایی مساوی یك و ثابت مجانبی خطا برابر $|g'(\xi)|$ است. از رابطهٔ (۳۸.۳ الف) برای روش نیوتن، درصورتی که $f'(\xi) \neq 0$، داریم

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^2} = \frac{1}{2}\left|\frac{f''(\xi)}{f'(\xi)}\right|$$

لذا بنابر تعریف، مرتبهٔ همگرایی برابر ۲ و ثابت مجانبی خطا برابر با $|f''(\xi)|/(2|f'(\xi)|)$ است.

برای تعیین مرتبهٔ همگرایی روش خط قاطع، ابتدا ملاحظه می‌کنیم که به‌موجب (۳۹.۳ الف) داریم

$$\lim_{n\to\infty} c_n = c_\infty = \frac{1}{2}\left|\frac{f''(\xi)}{f'(\xi)}\right| \quad \text{به‌ازای} \quad |e_{n+1}| = c_n|e_n e_{n-1}| \qquad (۴۰.۳)$$

عدد p را طوری پیدا می‌کنیم که رابطهٔ

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = C$$

به‌ازای یك مقدار غیرصفر C برقرار باشد.

اما از (۴۰.۳) نتیجه می‌گیریم

$$\frac{|e_{n+1}|}{|e_n|^p} = c_n|e_n|^{1-p}|e_{n-1}| = c_n\left(\frac{|e_n|}{|e_{n-1}|^p}\right)^\alpha \qquad (۴۱.۳)$$

به‌شرط آنکه داشته باشیم $\alpha = 1 - p$ و نیز $\alpha p = -1$؛ یعنی درصورتی‌که

$$p - p^2 = \alpha p = -1$$

معادلهٔ $p^2 - p - 1 = 0$ دارای ریشهٔ ساده ومثبت $p = (1+\sqrt{5})/2 = 1.618\ldots$ است. با این انتخاب p و $\alpha = -1/p$ ملاحظه می‌کنیم که معادلهٔ (۴۱.۳) تعریفی است برای «بارست شبه نقطهٔ ثابت»

$$y_{n+1} = c_n y_n^{-1/p}$$

که در آن

$$\lim_{n\to\infty} c_n = c_\infty \quad \text{و} \quad y_{n+1} = \frac{|e_{n+1}|}{|e_n|^p}$$

از اینجا نتیجه می‌شود که y_n به‌سمت نقطهٔ ثابت معادلهٔ

$$x = c_\infty x^{-1/p}$$

که جواب آن $c_\infty^{1/p}$ می‌باشد، همگرا می‌شود زیرا داریم $p = 1 + 1/p$. این تساوی نشان می‌دهد که برای روش خط قاطع با $p = 1.618\ldots$ داریم

$$\frac{|e_{n+1}|}{|e_n|^p} \approx \left|\frac{f''(\xi)}{2f'(\xi)}\right|^{1/p} \quad \text{به‌ازای مقدار بزرگ } n \qquad (۴۲.۳)$$

یعنی مرتبهٔ همگرایی روش خط قاطع برابر با $p = 1.618\ldots$ است وثابت مجانبی خطا برابراست با $|f''(\xi)/(2f'(\xi))|^{1/p}$. این بدان معنی‌است که روش خط قاطع سریعتر از روش معمول بارست نقطهٔ ثابت همگرا می‌شود اما سرعت همگرایی آن کمتر از روش نیوتن‌است.

□ **مثال ۵.۳:** با استفاده از داده‌های مثال ۲.۳، درستی فرمولهای خطای (۳۹.۳ الف) و (۴۲.۳) را برای روش خط قاطع تحقیق کنید.

در مثال (۲.۳ الف) بارستهای روش‌خط قاطع ریشهٔ مثبت $x^3 - x - 1 = 0$ داده شده و در جدول زیر مقادیر $|e_n|$ و $|e_{n+1}|/|e_n e_{n-1}|$ به‌ازای $n = 2, 3, 4$ محاسبه شده‌اند. فرض کنید که مقدار ξ تا هشت رقم اعشاری $\xi = 1.32471800$ صحیح باشد.

n	$\lvert e_n\rvert$	$\lvert e_{n+1}\rvert/\lvert e_n e_{n-1}\rvert$	$\lvert e_{n+1}\rvert/\lvert e_n\rvert^{1.618\cdots}$
1	0.1580513		1.41684
2	0.0716060	1.1034669	0.88969
3	0.0124884	0.9705400	1.04325
4	0.0008679	0.9318475	0.90778
5	0.0000101		

اگر ثابت $f''(\xi)/(2f'(\xi))$ را مستقیماً حساب کنیم، مقدار آن برابر با ۰٫۹۳۱۸۸۰۰۰ خواهد شد که با نسبت $|e_{n+1}/e_n e_{n-1}|$ به‌ازای $n = 4$ بسیار توافق دارد. □

می‌توان مستقیماً نشان داد که اگر $f(\xi) = 0$، $f'(\xi) \neq 0$، و $f''(x)$ دوبار به‌طور پیوسته مشتقپذیر باشد، آنگاه داریم $g'(\xi) = 0$، که

$$g(x) = x - \frac{f(x)}{f'(x)}$$

تابع بارست نیوتن است. سپس بنابر فرع قضیهٔ ۱.۳ نتیجه می‌شود که اگر x_0 «به‌اندازهٔ کافی نزدیک» به ξ انتخاب گردد، بارست نیوتن همگرا می‌شود. اصطلاح «به‌اندازهٔ کافی نزدیک» خیلی دقیق تعریف نشده‌است، ودرحقیقت روش نیوتن غالباً واگرا می‌شود و اگر هم همگرا شود به‌سمت ریشهٔ دیگری غیر از آن ریشه‌ای که خواسته شده همگرا خواهدشد. آنچه که ما می‌خواهیم برقراری شرایطی‌است که همگرایی را برای هر انتخاب بارستن اولیه در یك بازهٔ داده شده تضمین کند. مجموعه‌ای از چنین‌شرایط در قضیهٔ زیر داده شده است.

قضیهٔ ۲.۳ گیریم $f(x)$ در بازهٔ بسته و معین $[a, b]$ دوبار به‌طور پیوسته مشتقپذیر باشد و شرایط زیر برقرار

(i) $f(a)f(b) < 0$

(ii) $f'(x) \neq 0$, $x \in [a, b]$

(iii) به‌ازای تمام $x \in [a, b]$ $\quad f''(x)$ یا $\geqslant 0$ یا $\leqslant 0$

(iv) در دونقطهٔ a و b

$$\frac{|f(a)|}{|f'(a)|} < b-a \qquad \frac{|f(b)|}{|f'(b)|} < b-a$$

در شرایط فوق روش نیوتن به‌سمت جواب یگانهٔ ξ از معادلهٔ $f(x)=0$ در بازهٔ $[a, b]$ برای هر انتخابی از $x_0 \in [a, b]$ همگرا می‌شود.

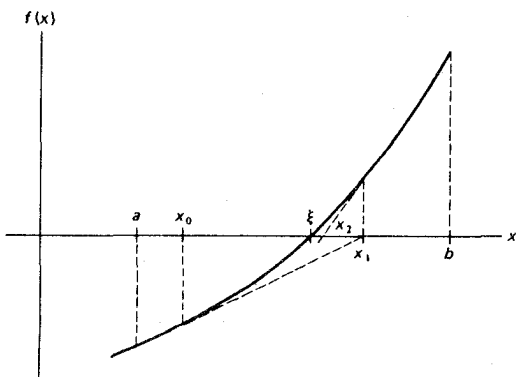
ذکر چند نکته دربارهٔ این شرایط مناسب به‌نظر می‌رسد.

شرایط (i) و(ii) تضمین می‌کنند که فقط یك جواب در $[a, b]$ وجود دارد. شرط (iii) مبین آن‌است که نمودار $f(x)$ تقعر[1] به‌سمت بالا یا تقعر به‌سمت پایین دارد،و بعلاوه همراه با شرط (ii) ایجاب می‌کند که $f'(x)$ روی $[a, b]$ یکنوا باشد.علاوه بر اینها،شرط (iv) مبین این‌است که خط مماس برمنحنی درهریك از دونقطهٔ a و b محور xها را در بازهٔ $[a, b]$ قطع می‌کند. برهان این قضیه دراینجا داده نخواهد شد(به‌تمرین ۳.۵–۷ نگاه کنید). اما اشاره می‌کنیم که چرا این قضیه درست‌است. بی‌آنکه لطمه‌ای به‌کلیت وارد آید، فرض می‌کنیم که $f(a) < 0$، پس دوحالت می‌توانیم تشخیص دهیم:

حالت (الف) $f''(x) \geqslant 0$

حالت (ب) $f''(x) \leqslant 0$

اگر به‌جای f، f — بگذاریم حالت(ب) به‌حالت(الف) تبدیل می‌شود. بنابراین بررسی حالت (الف) کفایت می‌کند. در اینجا نمودار $f(x)$، به‌گونه‌ای‌است که در شکل ۶.۳ داده شده‌است. از این نمودار آشکار است که به‌ازای $\xi > x_0$، بارستن‌های حاصل به‌طور یکنوا به‌سمت ξ کاهش می‌یابند، درحالی که به‌ازای $\xi < x_0$، $a \leqslant x_1$ بین ξ و b قرار می‌گیرد و آنگاه بارستن‌های بعدی به‌طور یکنوا به‌سمت ξ همگرا می‌شوند.



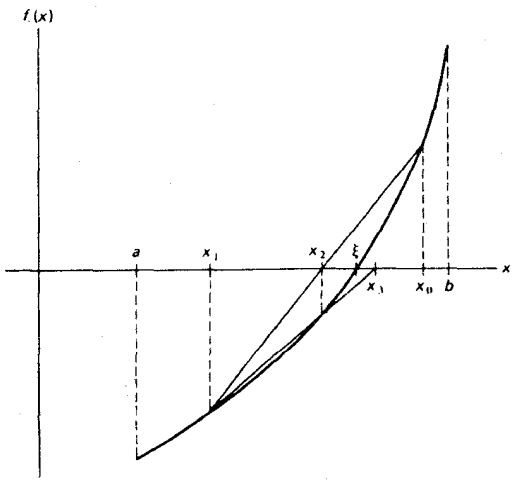
شکل ۶.۳ همگرایی نیوتن.

1. concave

□ **مثال ۶.۳:** بازه‌ای پیدا کنید که کوچکترین ریشهٔ مثبت تابع $f(x)=e^{-x}-\sin x$ را در برگیرد و درشرایط قضیهٔ ۲.۳ برای همگرایی روش نیوتن صدق نماید.

با توجه به‌عبارت $f(x)=e^{-x}-\sin x$، داریم $f'(x)=-e^{-x}-\cos x$ و $f''(x)=e^{-x}+\sin x$. بازهٔ خود را به‌صورت $[a, b]=[0, 1]$ انتخاب می‌کنیم. پس چون $f(0)=1$، $f(1)=-0.347$، داریم $f(a)f(b)<0$ لذا شرط (i) برقرار است. از آنجاکه برای کلیهٔ مقادیر $x\in[0, 1]$، نامساوی $f'(x)<0$ برقرار است، شرط (ii) برقرار بوده و چون به‌ازای جمیع مقادیر $x\in[0, 1]$، داریم $f''(x)>0$، شرط (iii) نیز برقرار می‌باشد. در نهایت چون $f(0)=1$، $f'(0)=-2$، داریم $f(1)=-0.347$ ، $b-a=1<2=|f'(0)|/|f(0)|$ و چون $f'(1)=-0.909000$، داریم $|f(1)|/|f'(1)|=0.520\ldots<1$ که در شرط (iv) صدق می‌کند. بنابراین بارست نیوتن برای هر انتخابی از x_0 در $[0, 1]$ همگرا می‌شود. □

همچنین شرایط قضیهٔ ۲.۳ برای اثبات همگرایی روش خط قاطع کافی است، ولو اینکه حالات ممکن همگرایی با حالات همگرایی نیوتن کاملا متفاوت باشد. اگر دوباره فرض کنیم که در بازهٔ $[a, b]$، همان‌طوری که درشکل (۶.۳ الف) نشان داده شده‌است، $f'(x)>0$ و $f''(x)\geqslant 0$، آنگاه اساساً دوحالت همگرایی وجود خواهد داشت که بستگی به گزینش نقاط اولیهٔ x_0 و x_1 دارند. درحالت اول که ساده‌تر است، اگر x_0 و x_1 دربازهٔ $[\xi, b]$ برگزیده شوند، مانند روش نیوتن همگرایی ازطرف راست یکنوا خواهد بود. دانشجویان می‌توانند این امر را برای روش نیوتن با رسم منحنیهای نمونه‌ای که درشرایط قضیهٔ ۲.۳ صدق می‌کنند بررسی نمایند.



شکل ۶.۳ (الف) همگرایی خط قاطع.

از سوی دیگر اگـر یکی از نقاط، مثلا x_0 را، دربازهٔ $[\xi, b]$ و نقطهٔ دیگر x_1 را در بازهٔ $[a, \xi]$ انتخاب کنیم، آنگاه بارست بعدی x_2 نیز دربـازهٔ $[a, \xi]$ خواهد بود و بارستن x_3 در طرف راست ξ قرارخواهدگرفت. دراین حالت دوباره دو نقطـهٔ بارست پیاپی x_2 وx_3 را خواهیم داشت که ریشهٔ ξ رادربرمی گیرند، و تمامی دنباله تکرارخواهد شد. بنابراین همگرایی به‌اصطلاح، حالت والسی[1] دارد، یعنی دربازه‌ای روی می‌دهدکـه یك بارست در یك طرف آن و دو بارست درطرف دیگر آن‌است، برای نشان دادن این نوع همگرایی به‌شکل(۶.۳ الف) نگاه‌کنید.

□ **مثال۷.۳:** حالت همگرایی‌روش‌خط‌قاطع‌را هنگامی که‌برای‌حل‌تابع $f(x) = e^x - ۳$ به‌کاربرده می‌شود، بررسی کنید.

بدیهی است‌که به‌ازای جمیع مقادیر x، نامساویهای $f'(x) > ۰$ و $f''(x) > ۰$ برقرارند. بعلاوه، برای مثال، در بازهٔ $[۰, ۵]$ شرایط نقاط انتهایی قضیهٔ ۲.۳ برقرارند. بنابراین $f(x)$ یك ریشه درآن بازه داردکه عبارت‌است‌از $\xi = \ln ۳ = ۱٫۰۹۸۶۱۲۲۸۹$ و اگـر $\xi < ۰ = x_0$ و $\xi > ۵ = x_1$ انتخاب شوند، آنگاه انتظار همگرایـی می‌رود. پس بارستن‌های زیر به‌دست خواهند آمد، که حالت همگرایی به‌اصطلاح والسی را نشان می‌دهند:

$$x_0 = ۰$$

$$x_1 = ۵$$

$$x_2 = ۰٫۰۶۷۸۳۶۵۴۹۰۰$$

$$x_3 = ۰٫۱۳۲۴۳۵۰۶۰۹$$

$$x_4 = ۱٫۸۱۳۴۵۹۸۴۳$$

$$x_5 = ۰٫۷۵۸۴۸۶۲۶۵۰$$

$$x_6 = ۰٫۹۸۶۸۰۸۲۷۸۲$$

$$x_7 = ۱٫۱۱۹۱۱۹۹۱۸$$

$$x_8 = ۱٫۰۹۷۴۴۸۶۵۳$$

$$x_9 = ۱٫۰۹۸۶۰۰۳۹۶$$

$$x_{10} = ۱٫۰۹۸۶۱۲۲۹۶$$

---

1. waltz

اگر بگیریم $\xi < ۳ = x_۰$ و $\xi > ۲ = x_۱$، آنگاه بارستهای زیر را خواهیم داشت

$$x_۰ = ۳$$

$$x_۱ = ۲$$

$$x_۲ = ۱٫۶۵۴۳۰۹۲۴۰$$

$$x_۳ = ۱٫۲۹۷۴۳۳۱۵۸$$

$$x_۴ = ۱٫۱۴۷۳۹۳۲۵۹$$

$$x_۵ = ۱٫۱۰۳۲۶۵۴۰۶$$

$$x_۶ = ۱٫۰۹۸۷۲۴۷۷۲$$

$$x_۷ = ۱٫۰۹۸۶۱۲۵۵۰$$

$$x_۸ = ۱٫۰۹۸۶۱۲۲۸۹$$

که معرف حالت یکنوا بودن همگرایی هستند. □

از دیدگاه محاسباتی، دقت حاصل با روش نیوتن بستگی به‌دقت عمل در محاسبهٔ $f(x)/f'(x)$ دارد. مثلا ممکن‌است‌که در همسایگی ریشه، مقدار $f'(x)$ صفر نباشد اما بسیار کوچك باشد. دراین صورت اگر $f(x)$ با خطا همراه باشد انتظار می‌رود که مقدار این خطا در محاسبهٔ $f(x)/f'(x)$ بزرگ شود. درچنین مواردی به‌دست آوردن دقت مناسب دشوارخواهد بود.

در روش نیوتن دو اشکال مهم وجود دارد. اولا برای اینکه بارستها به‌سمت ξ که ریشهٔ $f(x)$ است همگرا باشند، باید ازریشه‌ای که «به‌اندازهٔ‌کافی» به ξ نزدیك‌است شروع کنیم. (به‌تمرینهای ۵.۳–۶ و همچنین ۳.۳–۴ و ۳.۳–۵ نگاه‌کنید.) از آنجاکه معمولا مقدار ξ معلوم نیست انجام این‌کار درعمل مشکل‌است، مگر آنکه با استفاده از روشهای دیگر یك تخمین‌بهتر $\hat{\xi}$ برای ξ براورد شود. مثلا اگر با استفاده از روش تنصیف یا سایر روشهای بارستی، تقریب $\hat{\xi}$ برای ξ که تا دو یا سه رقم اعشاری صحیح باشد محاسبه شود، آنگاه می‌توان روش نیوتن را با تقریب بهتر $x_۰ = \hat{\xi}$ شروع کرد و پس از انجام دو یا سه بارست سریعاً تقریب صحیحتری از ξ به‌دست آورد. بدین‌ترتیب، روش نیوتن غالباً برای **بهبودبخشی** یك تقریب مناسب از ریشه که خود از روش دیگری به‌دست آمده، به‌کار گرفته می‌شود.

اشکال دوم روش نیوتن لزوم محاسبهٔ $f'(x)$ است. دربرخی‌موارد ممکن‌است $f'(x)$ صریحاً در دست نباشد و یا حتی اگر مقدار $f'(x)$ قابل ارزیابی باشد این‌کارممکن‌است

مستلزم محاسبات پرکار وقابل‌ملاحظه‌ای باشد. درحالت اخیرممکن‌است مقدار $f'(x_n)$ در هر kمرحله یکبار از الگوریتم نیوتن‌محاسبه شود و درهرمرحله آخرین مقدار محاسبه‌شده به‌کاربرده شود. اما دربرخورد با هردو حالت بهتر است کـه از روش **خط قاطع** به‌جای روش نیوتن استفاده شود.

در روش خط قاطع از مقادیر مختلف $f(x)$ استفاده می‌شود و در هر بارست فقط محاسبهٔ یك مقدارتابع لازم‌است، درصورتی‌که در روش‌نیوتن درهرمرحله به‌محاسبهٔ مقدار دوتابع* نیازاست. از طرف دیگر اگر روش خط قاطع همگرا باشد سرعت همگرایی‌آن به‌اندازهٔ روش نیوتن نیست‌گرچه‌که معمولا خیلی سریعتر از همگرایی خطی‌است.

سرعت بیشتر همگرایی‌روش نیوتن نسبت به‌روش‌خط قاطع درمثال ۲.۳ نشان داده شده‌است.

دراین فصل، برای پیداکردن ریشه‌های توابع شش الگوریتم را در نظرگرفته‌ایم. درمقایسهٔ بین الگوریتمها، برای استفادهٔ‌کامپیوتری‌آنها، ملاکهای مختلفی را می‌باید درنظر گرفت،که مهمترین‌آنها اطمینان از همگرایی، میزان همگرایی وکارایـی محاسباتی است. همواره نمی‌توان یك روش را برتر از روش دیگردانست. برای مثال، درحالی‌کـه روش تنصیف از نظـر سرعت همگرایی، آهسته‌است ولی اگر این روش به‌طور صحیح به‌کار گرفته شود، روش مطمئنی خواهد بود، درحالی‌که روش نیوتن غالباً واگراست، مگرآنکه تقریب اولیه با دقت انتخاب شود. اصطلاح «کارایی محاسباتی»که دربالا به‌کاربرده شد، سعی داردکه مقدار کار لازم برای به‌دست آوردن دقت عمل مفروضی را در نظر بگیرد. اگرچه به‌طورکلی‌روش نیوتن‌نسبت به‌روش خط قاطع،سریعتر همگرا می‌شود، ولی‌معمولا به‌اندازهٔ این‌روش‌کارایی ندارد. زیرا روش‌نیوتن درهر بارست،نیاز به‌محاسبهٔ هردوتابع $f(x)$ و $f'(x)$ دارد. درمواردی‌کـه $f'(x)$ در دسترس و به‌آسانی قابل محاسبه باشد، روش نیوتن نسبت به‌روش‌خط قاطع‌کاراتر است، اما برای‌کارهای معمولی، روش تنصیف معمولا کاراتر وکاربرد آن ارجح‌است.

الگوریتمهای۱.۳ تا ۳.۳، این مزیت را دارندکه ریشه را محصورکرده وبنابراین تضمین‌کنندهٔ‌کرانهای خطـا روی ریشه هستند. از میان این الگوریتمها، الگوریتـم ۲.۳ (تصحیح خطا) هرگزنباید به‌کار رود. زیرا این الگوریتم فاصله‌ای راکـه ریشه را دربر می‌گیرد کوچکتر نمی‌کنـد. به‌طورکلـی از بین این سه الگوریتم، روش تصحیح خطای اصلاح‌شده (الگوریتم ۳.۳) ارجح‌است.

بارست نقطهٔ ثابت زمانی‌کار است‌که مانند روش‌نیوتن به‌طورمربعی همگرا شود.در حالت‌کلی بارست نقطهٔ ثابت تنها به‌طورخطی همگرا می‌شود، از این‌رو رقابت واقعی‌بین این روش و روش خط قاطع یا روش تصحیح خطای اصلاح‌شده، وجود ندارد. حتی بـا برونیابی مکرر، مانند بارست استفنسن، الگوریتم ۸.۳، همگرایی فقط در بهترین حالت

---

* دوتابع موردنظر $f(x)$ و $f'(x)$ می‌باشند.ـم.

مربعی خواهد بود. از آنجاکه هرمرحله از بارست استفنسن معادل با دو ارزیابی تابع بارست $g(x)$ است، بنابراین بارست استفنسن با روش نیوتن قابل‌قیاس است. اما چون قسمت برونیابی دریك مرحله از بارست استفنسن، با یك مرحله از روش خط قاطع که برای تابع $f(x)=x-g(x)$ به‌کارگرفته شده مشابه‌است، به‌نظر می‌رسدکه‌کاراتر است که روش استفنسن راکناربگذاریم و روش خط قاطع را برای تابع $f(x)=x-g(x)$ به‌کارگیریم.

اساساً، هدف اصلی‌ازبحث دربارهٔ بارست نقطهٔ ثابت این بوده‌است‌که مدل‌ساده‌ای برای یك شیوهٔ بارستی‌که بتوان به‌آسانی تجزیه وتحلیل‌کرد، به‌دست آورد. بینش حاصل درمبحث مربوط به‌چندمعادلهٔ چند مجهولی‌که موضوع فصل۵است، بسیارمفید خواهد بود.

تمرین

۵.۳-۱ بنابرتعریف بارست نقطهٔ ثابت باتابع بارست $g(x)$، باید خطادربارستن nام در معادلهٔ زیرصدق‌کند

$$e_n=\xi-x_n=g(\xi)-g(x_{n-1})$$

در متن نشان داده شد کـه اگـر $g'(\xi)=0$ و $g''(x)$ در $x=\xi$ پیوسته باشـد، بارست $x=g(x)$ به‌طور مربعی همگرا خواهد بود. شرایطی راکـه تحت آنها، می‌توان انتظار داشت‌که یك بارست همگرایی مکعبی‌داشته باشد پیداکنید.

۵.۳-۲ نشان دهیدکه در روش نیوتن، اگر $f(\xi)=0$ و $f'(\xi)\neq 0$، و اگر $f(x)$ دوبار به‌طور پیوسته مشتقپذیرباشد، آنگاه $g'(\xi)=0$. همچنین نشان دهیدکه

$$g''(\xi)=\frac{f''(\xi)}{f'(\xi)}$$

۵.۳-۳ برای هریك ازتوابع زیر، بازه‌ای راکه‌کوچکترین ریشهٔ مثبت را دربرگیردمعین کنید و نشان دهیدکه شرایط قضیهٔ۲.۳ برقراراست

(الف) $e^{-x}-x=0$

(ب) $x^3-x-1=0$

(پ) $e^{-x^2}-\cos x=0$

۵.۳-۴ هریك از معادله‌های تمرین ۵.۳-۳ را با دو روش خط قاطع و روش نیوتن‌حل و جوابها را مقایسه‌کنید.

۵.۳-۵ اگر $x=\xi$ یك ریشهٔ $f(x)$ ازمرتبهٔ۲ باشد،آنگاه، $f'(\xi)=0$ و $f''(\xi)\neq 0$.

نشان دهیدکه دراین‌حالت، روش نیوتن به‌طورمربعی‌همگرا نخواهد بود [یعنی نشان دهید که $g'(\xi)=1/2\neq 0$]. همچنین نشان دهیدکه اگر $f'(\xi)=0$، $f''(\xi)\neq 0$ و $f'''(x)$ درهمسایگی ξ پیوسته باشد، بارست

$$x_{n+1}=x_n-\frac{2f(x_n)}{f'(x_n)}=:g(x_n)$$

به‌طور مربعی همگرا می‌شود. {راهنمایی: برای محاسبهٔ $g'(\xi)$، از رابطهٔ زیر و قانون لوپیتال[1]، استفاده‌کنید.}

$$\lim_{x\to\xi}\frac{f(x)f''(x)}{[f'(x)]^2}=\lim_{x\to\xi}\frac{f(x)}{[f'(x)]^2}\lim_{x\to\xi}f''(x)$$

۶-۵-۳ با استفاده از روش نیوتن، نزدیکترین ریشهٔ معادلهٔ زیر به‌عدد ۱۰۰ را محاسبه کنید:

$$x=\tan x$$

(توجه: چنانچه x_0 با دقت‌کافی انتخاب نشود، روش‌نیوتن یك دنبالهٔ واگرا تولید خواهد کرد.)

۷-۵-۳ جزئیات اثبات قضیهٔ ۲.۳ را بیان‌کنید.

۸-۵-۳ ثابت‌کنیدکه درشرایط قضیهٔ۲.۳، به‌ازای جمیع مقادیر x_0 و x_1 دربازهٔ $[a,b]$، روش خط قاطع همگراست. همچنین نشان دهیدکه بسته به‌محل دو بارست پیاپی حالت همگرایی یا یکنوا یا والسی‌است. [راهنمایی: از معادلهٔ خطا (۳۹.۳) استفاده‌کنید ومانند اثبات همگرایی در روش نیوتن‌کار را ادامه دهید.]

۹-۵-۳ نشان دهیدکه اگر $x=\xi$ یك ریشهٔ مکرر مرتبهٔ m از تابع $f(x)$ باشد، آنگاه بارست

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

در شرایط پیوستگی مناسب، به‌طور مربعی همگرا خواهد بود.

۶.۳ معادلات بسجمله‌ای: ریشه‌های حقیقی

گرچه معادلات بسجمله‌ایها را می‌توان با هریك از روشهای بارستی‌که قبلا بحث شده‌حل

1. L'Hospital

نمود، اما از آنجاکه کراراً درکاربردهای فیزیکی پدید می‌آیند نیاز به‌بحثی خاص دارند. بخصوص ما چند الگوریتم کارآمد را برای پیداکردن ریشه‌های حقیقی وهمتافت بسجمله‌ایها ارائه خواهیم داد. در این بخش به‌دست آوردن اطلاعات (معمولا تقریبی) دربارهٔ جای ریشه‌های بسجمله‌ایها را موردبحث قرار می‌دهیم و سپس روش نیوتن را برای پیداکردن ریشهٔ مثبت بسجمله‌ایها ذکرخواهیم کرد.

یك بسجمله‌ای از درجهٔ (دقیق) n معمولا به‌شکل زیر نوشته می‌شود:

$$p(x)=a_0+a_1x+a_2x^2+\cdots+a_nx^n \qquad a_n\neq 0 \qquad (۴۳.۳)$$

قبل از شروع به‌بحث دربارهٔ روشهای ریشه‌یابی، چند تذکر در بارهٔ ریشه‌های بسجمله‌ای ممکن‌است مناسب باشد. به‌ازای $n=2$، $p(x)$ یك بسجمله‌ای از درجهٔ دوم است و البته ریشه‌های آن را می‌توان صریحاً با استفاده از دستور درجهٔ دوم، همان‌طوری‌که در فصل یك انجام شد، پیداکرد. دستورهای مشابه ولی پیچیده‌تر دقیقی برای بسجمله‌ایها ازدرجهٔ ۳ و ۴ وجود دارند. اما به‌ازای $n\geqslant 5$ به‌طورکلی دستورهای صریحی‌برای ریشه‌ها وجود ندارند. بنابراین مجبوریم کـه روشهای بارستی را بـرای محاسبهٔ ریشهٔ بسجمله‌ایها مورد بررسی قراردهیم. همهٔ روشهایی راکه دراین‌فصل بررسی شدند می‌توان برای پیداکردن ریشه‌های حقیقی بسجمله‌ایها به‌کارگرفت وبرخی ازآنها را می‌توان برای محاسبهٔ ریشه‌های همتافت تعمیم داد. غالباً مـا به‌پیدا کردن همهٔ ریشه‌های بسجمله‌ای علاقه‌مندیم. عده‌ای از قضایای درس جبر برای رده‌بندی و جایابی انواع ریشه‌های بسجمله‌ایها مفیدند.

در ابتدا قضیهٔ اساسی جبر (قضیهٔ ۱۰.۱ را ببینید) را داریـم کـه بیان مـی‌دارد هربسجمله‌ای ازدرجهٔ n با شرط $a_n\neq 0$، دقیقاً n ریشهٔ حقیقی یا همتافت دارد، به‌شرطی که ریشه‌های مکرر از مرتبهٔ r، rمرتبه شمرده شوند. اگرکلیـهٔ ضرایب a_k از بسجمله‌ای $p(x)$، حقیقی باشند و اگر $z=a+ib$ یك ریشه باشد، آنگاه عدد $\bar{z}=a-ib$ نیز یك ریشه خواهد بود. یك روش مناسب برای تعیین تعداد ریشه‌های حقیقی یك بسجمله‌ای بـا ضرایب حقیقی، **قاعدهٔ علامتهای دکارت**[1] است. این قانون بیان می‌کند که تعداد ریشه‌های مثبت یـك بسجمله‌ای، n_p، نابزرگتر از تعـداد **تغییرعلامتهای** ضرایب $p(x)$، v، است. بعلاوه تفاضل $v-n_p$ یك عددزوج‌است. برای تعیین‌تعداد تغییرعلامتها، کافی‌است‌که تعداد تغییر علامتهای ضرایب غیـر صفـر $p(x)$ را بشماریـم. بنابراین اگـر داشته باشیم $p(x)=x^4+2x^2-x-1$، چون تعداد تغییر علامتها یکی است، بنابر قانون دکارت، $p(x)$ حداکثردارای یك ریشهٔ مثبت‌است، اما چون $v-n_p$ می باید یك عدد زوج‌غیرمنفی باشد، بنابراین‌معادلهٔ فوق دقیقاً دارای یك ریشهٔ مثبت‌است. به‌طور مشابه، تعداد ریشه‌های منفی‌حقیقی $p(x)$ برابر با تعداد تغییرعلامتها در ضرایب بسجمله‌ای

$$p(-x)=-x^3-2x^2-x-1$$

1. Descartes' rule of signs

می‌باشد و چون تغییر علامتی در $p(-x)$ وجود ندارد، بنابراین ریشهٔ منفی‌حقیقی نخواهیم داشت

□ **مثال ۸.۳:** تا حدی که می‌توانید، در بارهٔ ریشه‌های حقیقی بسجمله‌ای زیر اطلاعاتی به‌دست دهید

$$p(x)=x^4-x^3-x^2+x-1$$

چون ضرایب $p(x)$ دارای سه تغییر علامت می باشند بنابراین یا سه ریشهٔ مثبت‌حقیقی یا یك ریشهٔ مثبت حقیقی وجود دارد، اکنون $p(-x)=x^4+x^3-x^2-x-1$ را در نظر می گیریم وچون تنها یك تغییر علامت داریم بنابراین می باید یك ریشهٔ منفی‌حقیقی وجود داشته باشد. بنابراین، یا سه ریشهٔ مثبت حقیقی و یك ریشهٔ منفی حقیقی ویا یك ریشهٔ مثبت حقیقی، یك ریشهٔ منفی حقیقی و دو ریشهٔ مزدوج همتافت، داریم. □

اکنون به‌ذکر چندین قضیه، کـه کرانهای ریشه‌های بسجمله‌ایها را معین مـی کنند، می‌پردازیم. یکی ازاین قضایا این است کـه اگر $p(x)$ یك بسجمله‌ای مانند (۴۳.۳) بـا ضرایب a_k باشد، آنگاه $p(x)$ در داخل دایره‌ای که با $\min\{\rho_1, \rho_n\}$ تعریف می‌شود حداقل یك ریشه دارد که:

$$\rho_1=n\frac{|a_0|}{|a_1|} \qquad (۴۴.۳)$$

و

$$\rho_n=\sqrt[n]{\frac{|a_0|}{|a_n|}}$$

□ **مثال:** اگر یك بسجمله‌ای به‌شکل

$$p(x)=x^5-۳٫۷x^4+۷٫۴x^3-۱۰٫۸x^2+۱۰٫۸x-۶٫۸ \qquad (۴۵.۳)$$

باشد، آنگاه $a_5=1$، $a_1=۱۰٫۸$ و $a_0=-۶٫۸$. از رابطهٔ (۴۴.۳) به‌دست می‌آوریم

$$\rho_1=5\times\frac{۶٫۸}{۱۰٫۸}=۳٫۱۴\ldots$$

$$\rho_n=\sqrt[5]{\frac{۶٫۸}{1}}=۱٫۴۶\ldots$$

بنابراین، حداقل یك ریشهٔ حقیقی یا همتافت درداخل دایرهٔ $|x|\leqslant ۱٫۴۶\ldots$ وجود دارد.

در بخش بعدی، این بسجمله‌ای (۴۵.۳) را با جزئیات بیشتر مورد بررسی قرار و نشان می‌دهیم که ریشه‌های آن تحقیقاً برابر i، 1، $\pm i$، $\pm\sqrt{2}\,i$ و 1.7 هستند. □

دومین قضیهٔ مفیدی که به کوشی[1] منسوب است، به‌ما امکان می‌دهد که کرانهای ریشه‌های $p(x)$ را به‌شرح زیر مشخص سازیم. اگر $p(x)$ یك بسجمله‌ای مانند (۴۳.۳) باشد، دو بسجمله‌ای جدید به‌صورت زیر تعریف می‌کنیم:

$$P(x)=|a_n|x^n-|a_{n-1}|x^{n-1}-\cdots-|a_0| \qquad (۴۶.۳)$$

$$Q(x)=|a_n|x^n+|a_{n-1}|x^{n-1}+\cdots+|a_1|x-|a_0| \qquad (۴۶.۳\text{الف})$$

بنا بر قاعدهٔ علامتهای دکارت بسجمله‌ای (۴۶.۳) دقیقاً یك ریشهٔ مثبت حقیقی R و بسجمله‌ای (۴۶.۳الف)، دقیقاً یك ریشهٔ مثبت حقیقی r دارد. دراین حال قضیهٔ کوشی بیان می‌کند که همهٔ ریشه‌های $p(x)$ در ناحیهٔ طوقی[2] $r\leqslant|x|\leqslant R$ قراردارند.

□ **مثال:** مجدداً بسجمله‌ای (۴۵.۳) را درنظر می‌گیریم. پس داریم

$$P(x)=x^5-3.7x^4-7.4x^3-10.8x^2-10.8x-6.8$$

$$Q(x)=x^5+3.7x^4+7.4x^3+10.8x^2+10.8x-6.8$$

که ریشهٔ مثبت آنها به‌ترتیب برابر با $r=0.6300\ldots$ و $R=5.6000\ldots$ است. از این‌رو کلیهٔ ریشه‌های $p(x)$ می‌باید در رابطهٔ زیرصدق کنند

$$0.6300\ldots<|x|\leqslant 5.6000\ldots \qquad \square$$

آخرین قضیه ازاین‌نوع بیان می‌داردکه اگر $p(x)$ یك بسجمله‌ای به‌شکل(۴۳.۳) باشد و اگر داشته باشیم

$$r:=1+\max_{0\leqslant k\leqslant n-1}\left|\frac{a_k}{a_n}\right|$$

آنگاه هر ریشه‌ای از $p(x)$ در ناحیهٔ مستدیری قراردارد که با رابطهٔ $|x|\leqslant r$ تعریف شده است.

□ **مثال:** اگر بسجمله‌ای (۲.۳) را به‌صورت

$$p(x)=x^3-x-1$$

درنظر بگیریم، آنگاه $r=1+1/1=2.0$، لذا همهٔ ریشه‌های $p(x)$ درون قرصی به‌شعاع ۲ و مرکز مبدأ مختصات قرار دارند. در بخش ۱.۳ دیدیم که یك ریشهٔ حقیقی $1.3247\ldots$ بود. دو ریشهٔ دیگر مختلط‌اند ولی باز درون دایرهٔ $|x|\leqslant 2$ قرار دارند. □

1. Cauchy 2. annular

اکنون به بررسی روشهای بارستی برای یافتن ریشه‌های حقیقی بسجمله‌ایها می‌پردازیم. در هر روش بارستی، غالباً مقدار بسجمله‌ای می‌باید ارزیابی شود، و این کار باید تاحد ممکن به‌قدر کافی انجام پذیرد. همان گونه که در فصل ۲ نشان داده شد، مؤثرترین روش برای ارزیابی یك بسجمله‌ای روش **ضرب‌تودرتو** است که در الگوریتم ۱.۲ شرح داده شد.

در الگوریتم ۱.۲، فرض شده بود که بسجمله‌ای به‌شکل نیوتنی (۳.۲) با مراکز $c_1, \ldots, c_n$ داده شده‌است. اگر این مراکز، همگی برابر صفر باشند، شکل نیوتنی (۳.۲) به‌شکل استانده توانی (۴۳.۳) بدل می‌شود. اکنون اگر یك نقطهٔ معین z را به ما بدهند، الگوریتم ۱.۲، برای تعیین $p(z)$ به‌شکل خاص زیردرخواهد آمد

$$\begin{aligned} a'_n &:= a_n \\ a'_{n-1} &:= a_{n-1} + z a'_n \\ &\cdots \\ a'_1 &:= a_1 + z a'_2 \\ p(z) = a'_0 &:= a_0 + z a'_1 \end{aligned} \qquad (۴۷.۳)$$

کمیتهای کمکی $a'_1, a'_2, \ldots, a'_n$ جداگانه مورد علاقهٔ ما هستند زیرا که با دوباره صفر گذاردن تمام c_kها در (۴.۲) داریم

$$\begin{aligned} p(x) &= a'_0 + (x-z)\{a'_1 + a'_2 x + a'_3 x^2 + \cdots + a'_n x^{n-1}\} \\ &= a'_0 + (x-z)q(x) \end{aligned} \qquad (۴۸.۳)$$

از این‌رو، $a'_1, \ldots, a'_n$ ضرایب بسجمله‌ای خارج‌قسمت $q(x)$ هستند که از تقسیم $p(x)$ بر بسجمله‌ای خطی $(x-z)$ به‌دست می‌آیند و a'_0 باقیماندهٔ این تقسیم است. به‌ویژه اگر در (۴۸.۳) قرار دهیم $x=z$، از نو به‌دست می‌آوریم $p(z)=a'_0$.

□ **مثال ۹.۳: تبدیل یك عدد صحیح دودویی به‌یك عدد صحیح دهدهی.** در بخش ۱.۱، برای تبدیل عدد صحیح دودویی به‌عدد صحیح دهدهی الگوریتم ۱.۱ ارائه گردید. بنابرقرارداد، عدد صحیح دودویی

$$\alpha = (a_n a_{n-1} a_{n-2} \cdots a_0)_2$$

که در آن a_i برابر با صفر یا یك‌است، نمایانگر عدد زیر است

$$\alpha = a_n 2^n + a_{n-1} 2^{n-1} + \cdots + a_0 2^0$$

بنابراین معادل دهدهی این عدد را می‌توان با ارزیابی بسجمله‌ای

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

به‌ازای $x=2$ و با استفاده از الگوریتم ضرب‌تودرتو (۱.۲) به‌دست‌آورد. این‌امر نشان می‌دهد که الگوریتم ۱.۱، حالت خاصی از الگوریتم ۱.۲ است. به‌عنوان مثال، عددصحیح دودویی $\alpha=(110011)_2$ به‌شرح زیر به‌هم‌ارز دهدهی خود تبدیل می‌شود

$$a'_5=a_5 \qquad =1$$
$$a'_4=a_4+2a'_5=3$$
$$a'_3=a_3+2a'_4=6$$
$$a'_2=a_2+2a'_3=12$$
$$a'_1=a_1+2a'_2=25$$
$$\alpha=a'_0=a_0+2a'_1=51 \qquad \square$$

هدف مستقیم ما تطابق روش نیوتن برای یافتن‌ریشه‌های حقیقی بسجمله‌ایهاست. بدین منظور باید بتوانیم که نه‌فقط مقدار $p(x)$ بلکه مقدار $p'(x)$ را نیز محاسبه کنیم. برای محاسبهٔ $p'(x)$ درنقطهٔ $x=z$، از (۴۸.۳) نسبت به x مشتق می‌گیریم و داریم

$$p'(x)=q(x)+q'(x)(x-z)$$

بنابراین با قراردادن $x=z$ داریم

$$p'(z)=q(z)$$

از آنجاکه $q(x)$ خود نیزیك بسجمله‌ای‌است که ضرایب آن معلوم‌اند، می‌توانیم الگوریتم ۱.۲ رایکباردیگر برای محاسبهٔ $q(z)$ ونتیجتاً برای $p'(z)$ به‌کار گیریم. این عمل‌الگوریتم زیر را به‌دست می‌دهد.

الگوریتم۹.۳: روش‌نیوتن‌برای پیداکردن ریشه‌های‌حقیقی بسجمله‌ایها. $n+1$ضریب $a_0, \ldots, a_n$ از بسجمله‌ای $p(x)$ در (۴۳.۳) و یك نقطهٔ شروع x_0 داده شده‌اند.

For $m = 0, 1, \ldots,$ until satisfied, do:
  $z := x_m,\ a'_n := a_n,\ a''_n := a'_n$
  For $k = n-1, \ldots, 1,$ do:
    $a'_k := a_k + za'_{k+1}$
    $a''_k := a'_k + za''_{k+1}$
  $a'_0 := a_0 + za'_1$
  $x_{m+1} := x_m - a'_0/a''_1$

□ **مثال۳.۱۰:** کلیهٔ ریشه‌های معادلهٔ بسجمله‌ای $p(x)=x^3+x-3=0$ را پیدا کنید.

این معادله یك ریشهٔ حقیقی و دو ریشهٔ همتافت دارد. از آنجا که $p(1)=-1$ و $p(2)=7$، ریشهٔ حقیقی باید بین $x=1$ و $x=2$ باشد. x_0 را مساوی ۱٫۱ انتخاب می‌کنیم و الگوریتم۹.۳ را به‌کارمی‌بندیم، با انجام محاسبات به‌توسط ماشین‌حساب دستی و احتساب پنج رقم اعشاری نتایج زیرحاصل می‌شود.

$x_0 = 1.1$				$x_1 = 1.1 - (-0.569)/4.63 = 1.22289$	
k	a_k	a'_k	a''_k	a'_k	a''_k
3	1	1	1	1	1
2	0	1.1	2.2	1.22289	2.44578
1	1	2.21	4.63	2.49546	5.48638
0	−3	−0.569		0.05167	
$x_2 = 1.22289 - a'_0/a''_1 = 1.21347$				$x_3 = 1.21347 - a'_0/a''_1 = 1.21341$	
k	a_k	a'_k	a''_k	a'_k	a''_k
3	1	1	1	1	1
2	0	1.21347	2.42694	1.21341	2.42682
1	1	2.47251	5.41753	2.47236	5.41709
0	−3	0.000317		−0.00001	

توجه کنید که a'_0 به‌سمت ریشه نزدیك،و a'_kها همگرا می‌شوند.با توجه به‌دقت عملی که در نظر گرفته شده است، بهسازی جواب و یا بهسازی a'_k ممکن نیست. بنابراین $x_3=1.21341$ را که تا پنج رقم با معنی صحیح است به‌عنوان ریشهٔ مطلوب قبول می‌کنیم. برای محاسبهٔ بقیهٔ ریشه‌های همتافت، از دستور درجهٔ دوم برای حل بسجمله‌ای زیر استفاده می‌کنیم

$$x^2+a'_2x+a'_1=x^2+1.21341\,x+2.47236=0$$

که نتایج زیر حاصل می‌شود

$$x=\frac{-a'_2 \pm (a'^2_2-4a'_1)^{1/2}}{2}$$

$$=\frac{-1.21341 \pm 2.90122\,i}{2}=-0.60671 \pm 1.45061\,i$$

به منظور مقایسه، دربخش ۷.۳ ریشه‌های این بسجمله‌ای دوباره با روشهای خاص مربوط به‌محاسبهٔ ریشه‌های همتافت محاسبه خواهد شد. □

□ **مثال۳.۱۱:** ریشهٔ مثبت حقیقی بسجمله‌ای زیر را محاسبه کنید

$$x^5 - ۳٫۷x^4 + ۷٫۴x^3 - ۱۰٫۸x^2 + ۱۰٫۸x - ۶٫۸ = ۰$$

به‌آسانی می‌توان تحقیق کرد که ریشهٔ موردنظر بین ۱ و ۲ قرار دارد. بنابراین $x_0 = ۱٫۵$ را انتخاب می‌کنیم. برنامهٔ **فورترن** و نتایج حاصله ازکامپیوتر در زیر داده شده‌اند. ریشهٔ دقیق برابر با ۱٫۷ و نتایج حاصله ازکامپیوتر تا هشت رقم اعشاری صحیح است.

برنامهٔ فورترن برای مثال ۱۱.۳

```
C  NEWTON'S METHOD FOR FINDING A REAL ZERO OF A CERTAIN POLYNOMIAL.
C  THE COEFFICIENTS ARE SUPPLIED IN A DATA STATEMENT. A FIRST GUESS
C   X  FOR THE ZERO IS READ IN .
      PARAMETER N=6
      INTEGER J,K
      REAL A(N),B,C,DELTAX,X
      DATA A /-6.8, 10.8, -10.8, 7.4, -3.7, 1./
    1 READ 500, X
  500 FORMAT(E16.8)
      PRINT 601
  601 FORMAT('1NEWTONS METHOD FOR FINDING A REAL ZERO OF A POLYNOMIAL'
     *       //4X,'I',10X,'X',14X,'AP(0)',12X,'APP(1)'/)
      DO 10 J=1,20
         B = A(N)
         C = B
         DO 5 K=N,3,-1
            B = A(K-1) + X*B
            C = B + X*C
    5    CONTINUE
         B = A(1) + X*B
         PRINT 605,J,X,B,C
  605    FORMAT(I5,3(1PE17.7))
         DELTAX = B/C
         IF (ABS(DELTAX) .LT. 1.E-7 .OR. ABS(B) .LT. 1.E-7) STOP
         X = X - DELTAX
   10 CONTINUE
      PRINT 610
  610 FORMAT(' FAILED TO CONVERGE IN 20 ITERATIONS')
                                        GO TO 1
      END
```

نتایج به‌دست‌آمده ازکامپیوتر برای مثال ۱۱.۳

I	X	AP	APP
1	1.5000000E 00	−1.0625001E − 00	3.7124998E 00
2	1.7861953E 00	7.2393334E − 01	9.6004875E 00
3	1.7107894E 00	8.0013633E − 02	7.5470622E 00
4	1.7001875E 00	1.3663173E − 03	7.2905675E 00
5	1.7000000E 00	4.7683716E − 07	7.2861013E 00
6	1.7000000E 00	−1.1920929E − 07	7.2860994E 00
7	1.7000000E 00	−5.9604645E − 08	7.2860998E 00

□

گرچه در مثال فوق برای به‌دست‌آوردن جوابهای دقیق به‌مشکل مهمی برنخوردیم، اما دانشجویان را برحذرمی‌داریم ازاینکه تصورکنندکه همواره محاسبهٔ ریشه‌های بسجمله‌ایها بدون اشکال صورت می‌گیرد. بعضی ازدشواریهایی که ممکن‌است با آنها مواجه شوید در

زیرشرح می‌دهیم.

۱. در روش نیوتن دقت ریشه بستگی به‌دقت محاسبهٔ جملهٔ تصحیحی $p(x_i)/p'(x_i)$ دارد. برای مثال اگرخطای ناشی ازگردکردن یا عوامل دیگر برابر ε باشد، آنگاه ریشهٔ محاسبه شده می‌تواند حداکثر دقتی برابر بـا ریشهٔ واقعـی بعلاوهٔ $\varepsilon/p'(x_i)$ داشته باشد. شکل۱.۳، مقدارخطاهای غیرمنتظره را به‌وضوح نشان می‌دهد. خطاهای جانبی نیز به‌طور قابل ملاحظه‌ای افزایش می‌یابند اگر $p(x)$ ریشهٔ مکرری درنقطهٔ $x=\xi$ داشته باشد،زیرا دراین‌صورت وقتی $x_i \to \xi$، $p'(x)$ به‌صفر نزدیك، و هرگونه خطای ناشی ازگردکردن درمحاسبهٔ $p(x_i)$ بزرگ می‌شود.

برای روشن‌کردن رفتار روش نیوتن پیرامون یك ریشهٔ مکرر، بسجمله‌ای

$$p(x)=x^3-3x^2+4$$

راکـه یك ریشهٔ مکرر در نقطـهٔ $x=2$ دارد، بررسی می‌کنیم. با انتخاب $x_0=1.5$ و به‌کارگیری کامپیوتر IBM 7094، نتایج به‌دست‌آمده درجدول۱.۳ داده شده‌اند.

جدول۱.۳

i	x_i	$p(x_i)$	$p'(x_i)$	$p(x_i)/p'(x_i)$
0	1.5	0.625 E + 0	−0.22499999E + 1	−0.27777778E + 0
1	1.7777777	0.13717422E + 0	−0.11851852E + 1	−0.11574074E + 0
2	1.8935185	0.32807648E − 1	−0.60487403E + 0	−0.54238810E − 1
3	1.9477573	0.80453157E − 2	−0.30526827E − 0	−0.26354902E − 1
4	1.9741122	0.19932091E − 2	−0.15331630E − 0	−0.13000633E − 1
5	1.9871128	0.49611926E − 3	−0.76824840E − 1	−0.64577974E − 2
6	1.9935706	0.12376904E − 3	−0.38452353E − 1	−0.32187638E − 2
7	1.9967893	0.30934811E − 4	−0.19232938E − 1	−0.16084287E − 2
8	1.9983977	0.77188015E − 5	−0.96056932E − 2	−0.80356526E − 3
9	1.9992013	0.19371510E − 5	−0.47900614E − 2	−0.40441045E − 3
10	1.9996057	0.47683716E − 6	−0.23651228E − 2	−0.20161200E − 3
11	1.9998073	0.11920929E − 6	−0.11558611E − 2	−0.10313461E − 3
12	1.9999104	0.59604645E − 7	−0.53713301E − 3	−0.11096813E − 3
13	2.0000214	0.29802322E − 7	+0.12850899E − 3	+0.23190846E − 3
14	1.9997895	0.14901161E − 6	−0.12628894E − 2	−0.11799259E − 3
15	1.9999074	0.59604645E − 7	−0.55501277E − 3	−0.10739328E − 3

اعداد بعد از حرف E معرف توانهای ۱۰ هستند، و ارقامی‌که زیرآنها خط‌کشیده شده، به‌علت از دست رفتن ارقام با معنی طی‌محاسبهٔ $p(x_i)$ و $p'(x_i)$، غیرصحیح می‌باشند. از این‌جدول می‌توان به‌نکات زیرپی برد.(در رابطه با این مسئله،تمرین۵.۳–۵ را ببینید).

الف. بارستها علی‌رغم این حقیقت‌که $p'(2)=0$، همگرا هستند.

ب. میزان همگرایی خطی است و مانند حالت عادی روش نیوتن مربعی نیست. بررسی عامل تصحیحی $p(x_i)/p'(x_i)$، نشان می‌دهد که تا بارست دوازدهم، خطا در هر بارست با ضریب ۲/۱ کاهش می‌یابد.

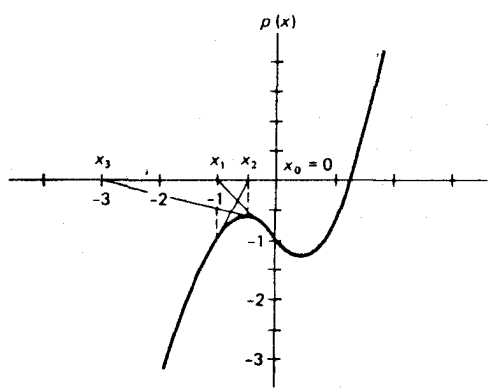
پ. بعد از ۱۳ بارست انتظار بهبود بیشتری در جواب را نمی‌توان داشت، علت آن این است که هیچ رقم صحیحی در $p(x_i)$ باقی نمانده و در عین حال مقدار $p'(x_i)$ از مرتبهٔ 10^{-3} شده است. بنابراین خارج‌قسمت $p(x_i)/p'(x_i)$ نتیجهٔ غیر صحیحی در پنجمین رقم اعشاری تولید می‌کند و بهبود جواب را غیرممکن می‌سازد.

۲. در برخی موارد انتخاب نامناسب تقریب اولیه موجب همگرایی به‌سمت ریشه‌ای غیر از ریشهٔ موردنظر می‌شود.

۳. برای برخی از بسجمله‌ایها انتخاب نامناسب x_0 ممکن است به‌تولید یك دنبالهٔ واگرا منجر شود. به‌عنوان نمونه در مثال ۲.۳، اگر $x_0 = 0$ انتخاب شود تقریبهای متوالی $x_1 = -1$، $x_2 = -1/2$، $x_3 = -3$، $x_4 = -2{,}04$ و $x_5 = -1{,}40$ به دست می‌آید، که قطعاً به‌سمت ریشه‌ای که قبلاً محاسبه شد همگرا نیست. بررسی نمودار بسجمله‌ای $p(x) = x^3 - x - 1$ (شکل ۷.۳ را نگاه کنید) به‌توضیح این عملکرد کمك خواهد کرد. بارستهای پیاپی ممکن است به‌طور نامتناهی پیرامون نقطهٔ $x = (-1/3)\sqrt{3}$، که در آن مقدار $p(x)$ ماکسیمم است، نوسان کنند.

۴. برخی از بسجمله‌ایها به‌ویژه آنهایی که از درجات بالاتر هستند، بسیار ناپایدارند، بدین معنی که تغییرات جزئی و کوچك در ضرایب آنها منجر به تغییرات بزرگی در ریشه‌های آنها می‌شوند (به‌مثال ۱۲.۳، که در زیر آمده، نگاه کنید).

۵. وقتی ریشه‌ای از $p(x)$ را پیدا کردیم، الگوریتم ضرب تودرتوی (۴۷.۳)، ضرایب بسجمله‌ای $q(x)$، یعنی a'_1، ...، a'_n را چنان معین می‌کند که بقیهٔ ریشه‌های $p(x)$ ریشه‌های $q(x)$ باشند. لذا برای یافتن این ریشه‌ها بررسی بسجمله‌ای **تقلیل یافته** یا پایین



شکل ۷.۳

آمدهٔ $q(x)$ آسانتر از بررسی $p(x)$ خواهد بود. اما درمورد ریشه‌های اخیر، انتظار می‌رود که دقت آنها کاهش یابد، زیرا ضرایب بسجمله‌ای تقلیل‌یافته متضمن خطاهایی هستند که از همگرایی غیرکامل و گردکردن این ریشه‌ها ناشی شده‌اند. برای به‌حداقل رساندن میزان کاهش در دقت عمل، ریشه‌ها می‌باید به‌ترتیب صعودی محاسبه شوند (مثال ۱۲.۳ را ببینید). همچنین از بارستن بسجمله‌ای اصلی، می‌توان دقت ریشهٔ به‌دست‌آمده از بسجمله‌ای تقلیل‌یافته را بهتر نمود.

□ **مثال ۱۲.۳:** به‌منظور تشریح برخی مخاطرات موجود در محاسبهٔ ریشهٔ بسجمله‌ایها، در زیر دو بسجمله‌ای را بررسی می‌کنیم

$$x^{۷}-۲۸x^{۶}+۳۲۲x^{۵}-۱۹۶۰x^{۴}+۶۷۶۹x^{۳}-۱۳۱۳۲x^{۲}+۱۳۰۶۸x-۵۰۴۰ \qquad (۴۹.۳)$$

و

$$x^{۵}-۱۵٫۵x^{۴}+۷۷٫۵x^{۳}-۱۵۵x^{۲}+۱۲۴x-۳۲ \qquad (۵۰.۳)$$

برای یافتن کلیهٔ ریشه‌های این بسجمله‌ایها، روش نیوتن‌روی یك کامپیوتر CDC ۶۵۰۰ اجرا و درهرمرحله بسجمله‌ایهای تقلیل‌یافته به‌کار گرفته شده، تقریبهای اولیه همان جواب دقیق با تقریباً ۱۰ درصد خطا درنظر گرفته شده وملاك خاتمه $|x_i-x_{i-۱}|<۱۰^{-۷}|x_i|$ انتخاب شده‌است.

ریشه‌های بسجمله‌ای اولی (۴۹.۳)، به‌ترتیب صعودی ۱، ۲، ۳، ۴، ۵، ۶، و ۷ می‌باشند. ستون A درجدول زیرمتضمن تقریبهای پیدا شده‌است که به‌ترتیب با حدسهای اولیهٔ ۹٫۵، ۹٫۱، ۹٫۲، ۹٫۳، ۹٫۴، ۹٫۵ و ۹٫۶ شروع شده‌اند. تعداد بارستهای ضروری در سمت راست هر ریشه چاپ شده‌است.

ستون B شامل ریشه‌هایی است که پس از تبدیل ضریب $x^{۲}$ در (۴۹.۳) به ۱۳۱۳۳۴ — به‌دست آمده‌اند، یعنی، بعد از آنکه یك تغییر یك واحدی در رقم پنجم اعشاری یکی از ضرایب داده شده است. فقط پنج ریشه محاسبه شده و برخی از آنها در رقم دوم اعشاری با ریشه‌های متناظر خود در ستون A متفاوت‌اند. برای اینکه تأیید شود که این تغییرات فقط ناشی از خطای گردکردن نیست و برای اینکه سرنوشت دو ریشهٔ از دست رفته‌روشن شود، روش مولر[1] (این روش دربخش بعد توضیح داده خواهد شد) به‌کار گرفته شده که این روش هرهفت ریشهٔ چاپ شده در ستون C را تولیدکرده است. تمام ارقام چاپ شدهٔ این اعداد، صحیح‌اند. توجه کنیدکه ریشه‌های ۵ و ۶ به‌یك جفت ریشه‌های مزدوج همتافت تبدیل شده‌اند. بنابراین تغییری معادل ۱/۱۰۰ در یکی از ضرایب به‌یك تغییر ۱۰ درصد دربرخی از ریشه‌ها منجرشده است. هنگامی که ضرایب بسجمله‌ای به‌وسیلهٔ آزمایش به‌دست

1. Müller

آمده باشند، به‌آسانی خطا در چنین حدودی در ضرایب مشاهده می‌شود. بنابراین باید به‌ریشه‌های بسجمله‌ایهای درجهٔ بالا که این گونه پیدا شده‌اند، بسیار محتاطانه نگاه کرد، به‌ویژه هنگامی‌که در دقت ضرایب آنها تردید باشد.

ریشه‌های بسجمله‌ای دوم،(۵۰.۳)، به ترتیب ۵ر۰، ۱، ۲، ۴ و ۸ هستند. با شروع عملیات از حدسهای اولیهٔ ۴۵ر۰، ۹ر۰، ۸ر۱، ۶ر۳ و ۲ر۷ ریشه‌های بسجمله‌ای را، به‌ترتیب صعودی، همان‌گونه‌که درستون D می‌بینیم،محاسبه‌کردیم. سرانجام در ستون E، نتایج محاسبهٔ این ریشه‌ها را به‌ترتیب نزولی ذکرکرده‌ایم، یعنی با حدسهای اولیهٔ ۲ر۷ شروع کرده و ریشهٔ ۸ را محاسبه‌کرده‌ایم، سپس با به‌کار گرفتن بسجمله‌ای تبدیل یافته و تقریب‌اولیهٔ ۶ر۳ ریشهٔ ۴ را به‌دست‌آوردیم،و هکذا. گرچه اولین ریشهٔ پیدا شده تا ۹رقم اعشاری صحیح‌است، اما ریشه‌های بعدی فقط تا شش رقم اعشارصحیح‌اند. بعلاوه دراینجا تعداد بارستهای موردنیاز زیادتراست. این‌امر نشان می‌دهدکه بهتراست درابتدا ریشه‌هایی با قدرمطلق کوچکتر، محاسبه شوند.

نتایج‌کامپیوتری برای مثال ۱۲.۳

A	B	C	D	E
1.00000000 5	1.00139755 5	1.0013976	0.50000000 4	8.00000000 7
2.00000000 5	1.96892082 4	1.9689208	1.00000008 4	3.99999862 6
3.00000000 5	3.31832477 7	3.3183233	2.00000007 4	2.00000552 6
4.00000000 5	3.50505891 7	3.5050604	4.00000005 4	0.99999079 5
5.00000000 4		5.5731849 +0.2641298i	7.99999999 2	0.50000485 2
6.00000014 4		5.5731849 −0.2641298i		
6.99999993 2	7.05992816 40	7.0599281		

□

مهلی[1] راهی برای استفاده از بسجمله‌ای تقلیل‌یافته پیشنهاد نموده است‌که در آن به‌مشکلات مذکور دربالا برخورد نمی‌شود. گیریم ξ_1، ...، ξ_k، k ریشهٔ قبلا به‌دست آمده از یك بسجمله‌ای باشند. برای یافتن‌ریشهٔ بعدی، یك بارست نیوتن روی بسجمله‌ای تقلیل یافتهٔ $(x-\xi_1)\cdots(x-\xi_k)$ انجام می‌دهند، اما $\tilde{p}(x)=p(x)/(x-\xi_1)(x-\xi_2)\cdots(x-\xi_k)$ را با تقسیم ترکیبی مکرر به‌دست نمی‌آورند. بلکه $\tilde{p}(x)$ را به‌شکل‌فوق رها می‌کنند و در این حالت، تابع بارست به‌گونهٔ زیر می‌شود

$$x_{i+1}=x_i-\frac{\tilde{p}(x_i)}{\tilde{p}'(x_i)}$$

$$=x_i-\frac{p(x_i)}{p'(x_i)-\sum_{r=1}^{k}\frac{p(x_i)}{(x_i-\xi_r)}}$$

1. Maehly

به‌نظر می‌رسد که این تکنیک در محاسبهٔ ریشه‌های دقیق و متوالی کاملا کارا باشد. تمرین ۶.۳-۷ را ببینید.

تمرین

۶.۳-۱ با استفاده از الگوریتم ۹.۳ و به‌کمك یك ماشین حساب دستی، ریشهٔ حقیقی معادلهٔ زیر را تا هفت رقم با معنی صحیح محاسبه کنید

$$x^3+2x-1=0$$

بقیهٔ ریشه‌ها را از بسجمله‌ای تقلیل یافته با استفاده از فرمول مربعی تعیین کنید. میزان دقت این ریشه‌ها چقدر است؟

۶.۳-۲ با استفاده از الگوریتم ۹.۳، ریشه‌های مثبت و حقیقی معادلات بسجمله‌ای زیر را به‌دست آورید

(الف) $x^5-3x^3+x^2-1=0$

(ب) $x^3+3x-1=0$

۶.۳-۳ بسجمله‌ای $x^4+2{,}8x^3-6{,}38x^2-3{,}42x+4{,}2$، دارای چهار ریشهٔ حقیقی است. با استفاده از الگوریتم ۹.۳، این ریشه‌ها را محاسبه کنید.

۶.۳-۴ بسجمله‌ای

$$p(x)=x^8-170x^6+7392x^4-39712x^2+51200$$

دارای ریشه‌های $\pm\sqrt{2}$، ± 2، ± 8 و ± 10 است. با انتخاب حدسهای اولیه‌ای که درحدود ۱۰ درصد با ریشه‌های دقیق فاصله داشته باشد، ریشه‌های بسجمله‌ای فوق را به‌ترتیب مقادیر صعودی با کامپیوتر پیدا کنید. سپس ضریب x^2 را به -39710 تبدیل و مسئله را دوباره حل کنید. تغییر حاصل درجوابها را ملاحظه نمایید.

۶.۳-۵ با استفاده از قاعدهٔ علایم دکارت وقضایای مربوط به کرانهای متضمن ریشهٔ بسجمله‌ایها و تا آنجاکه امکان دارد، اطلاعاتی دربارهٔ محل و نوع ریشه‌های بسجمله‌ای زیر به‌دست آورید

$$p(x)=x^4-x^3+x^2-x+1$$

۶.۳-۶ بسجمله‌ای

$$p(x)=x^4-5{,}85x^2+6{,}3504$$

دارای یك ریشهٔ $\xi_1=1{,}2$ است. ریشهٔ مثبت حقیقی دیگری نزدیك $x=2$ وجود دارد.

با استفاده از تکنیك مهلی و با شروع عملیات از $x_0 = ۲$ این ریشه را بهدست آورید.

۷-۶.۳ برمبنای روش مهلی برنامه‌ای برای یافتن ریشه‌های حقیقی و متوالی یك بسجمله‌ای $p(x)$ بنویسید.

۸-۶.۳ با استفاده از روش مهلی، ریشه‌های بسجمله‌ای داده شده درمثال ۱۲.۳ را بهدست آورید و با نتایج مربوط به آن مسئله مقایسه کنید.

۷.۳* ریشه‌های همتافت و روش مولر

روشهایی که تاکنون موردبحث قرارگرفتند به ما امکان می‌دهند که یك ریشهٔ تکین[1] از یك تابع را، وقتی که یك تقریب اولیه از آن ریشه معلوم باشد، محاسبه کنیم. زمانی کـه کلیهٔ ریشه‌های یك تابع لازم باشند و یـا یك تقریب اولیهٔ خوبی در دست نباشد، این روشها چندان رضایتبخش نیستند. برای توابع بسجمله‌ای روشهایی وجود دارند که تمامی ریشه‌ها را به‌طورهمزمان بهدست می‌دهند، و پس از آن می‌توان روشهای بارستی این فصل را برای محاسبهٔ دقیقتر این ریشه‌هـا بهکاربرد. از جملهٔ این روشها می‌توان به الگوریتم تفاضلهای منقسم [۲] و روش گرافه[2] [۵] اشاره کرد.

یك روش ریشه‌دار و تازه‌ای به‌وسیلهٔ مولر بیان و بـا موفقیت قابل‌ملاحظه‌ای روی کامپیوترها بهکاربرده شده‌است. این روش را می‌توان برای محاسبهٔ هر تعداد ریشهٔ حقیقی یا همتافت خواسته شده از هرتابـع دلخواهی بهکارگرفت. این روش، روش بارستی‌است و پیرامون ریشه تقریبـاً به‌طور مربعی همگراست و احتیاج به‌محاسبهٔ مشتق تابـع ندارد و ریشه‌های حقیقی وهمتافت را، حتی درحالتی که این ریشه‌ها ساده نباشند، بهدست می‌دهد.

وانگهی، روشی‌است کلی، بدین‌معنی که احتیاج به‌تقریب اولیه ندارد. در این بخش با انصراف از بحث در بارهٔ همگرایی، مختصراً چگونگی بهدست آوردن آن شرح داده می‌شود وسپس کاربرد آن برای محاسبهٔ ریشه‌های حقیقی وهمتافت موردبحث قرارمی‌گیرد. به‌ویژه، روی محاسبهٔ ریشه‌های همتافت یك بسجمله‌ای باضرایب حقیقی تکیه می‌شود، زیرا این مسئله با تمامی شاخه‌های مهندسی مربوط‌است.

روش مولر توسیع روش خط قاطع‌است. یادآوری می‌کنیم که در روش خط‌قاطع از x_{i-1} و x_i بهمثابـهٔ تقریبهایی برای یك ریشهٔ $f(x) = ۰$، تقریب بعدی x_{i+1} محاسبه می‌شود که یك ریشهٔ بسجمله‌ای خطی $p(x)$ ماربرنقاط $\{x_i, f(x_i)\}$ و $\{x_{i-1}, f(x_{i-1})\}$ است. در روش مولـر تقریب بعدی x_{i+1}، به‌عنوان ریشـهٔ یـك سهمی ماربر سه نقطـهٔ $\{x_i, f(x_i)\}$، $\{x_{i-1}, f(x_{i-1})\}$ و $\{x_{i-2}, f(x_{i-2})\}$ پیدا می‌شود.

چنانچه درفصل ۲ نشان داده شد، تابع

$$p(x) = f(x_i) + f[x_i, x_{i-1}](x - x_i) + f[x_i, x_{i-1}, x_{i-2}](x - x_i)(x - x_{i-1})$$

---

1. isolated 2. Graeffe

سهمی منحصر به‌فردی‌است که با تابع $f(x)$ درسه نقطهٔ x_i، x_{i-1} و x_{i-2} منطبق می‌باشد. و چون داریم

$$(x-x_i)(x-x_{i-1})=(x-x_i)^2+(x-x_i)(x_i-x_{i-1})$$

می‌توانیم $p(x)$ را به‌صورت زیرهم بنویسیم

$$p(x)=f(x_i)+(x-x_i)c_i+f[x_i, x_{i-1}, x_{i-2}](x-x_i)^2 \qquad (۵۱.۳)$$

در رابطهٔ بالا c_i عبارت‌است از

$$c_i=f[x_i, x_{i-1}]+f[x_i, x_{i-1}, x_{i-2}](x_i-x_{i-1})$$

بنابراین هرریشهٔ α از سهمی $p(x)$ براساس یک شکل خاص از فرمول مربعی استانده [(۲۰.۱) را ببینید]، در رابطهٔ زیرصدق می‌کند

$$\alpha-x_i=\frac{-2f(x_i)}{c_i\pm\{c_i^2-4f(x_i)f[x_i, x_{i-1}, x_{i-2}]\}^{1/2}} \qquad (۵۲.۳)$$

اگر در رابطه(۵۲.۳) علامت به‌گونه‌ای‌انتخاب شودکه مخرج ازنظر قدرمطلق بزرگترین مقدار ممکنه را داشته باشد و اگرطرف راست رابطهٔ (۵۲.۳) را با h_{i+1} نمایش دهیم، تقریب بعدی برای یك ریشهٔ $f(x)$ به‌صورت زیر درمی‌آید

$$x_{i+1}=x_i+h_{i+1}$$

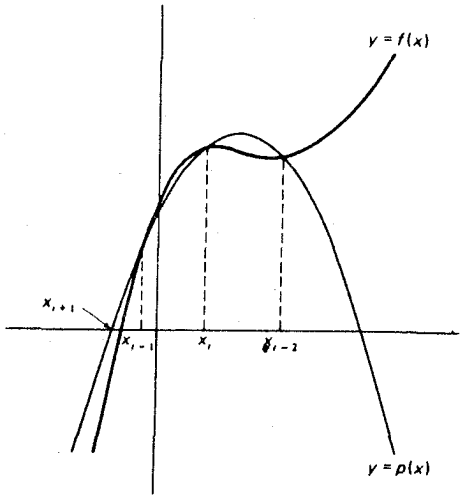
دراین‌صورت بااستفاده از x_{i-1}، x_i و x_{i+1} به‌عنوان سه‌تقریب‌پایه،روند فوق تکرار می‌شود. اگر ریشه‌های به‌دست‌آمده از (۵۲.۳) حقیقی باشند، وضعیت به‌گونه‌ای خواهد شدکه درشکل۸.۳ نمایش‌داده شده‌است. ولی باید توجه داشت‌که حتی‌اگرریشهٔ موردنظر حقیقی باشد امکان مواجه‌شدن با تقریبهای همتافت نیزمی‌رود. زیرا ممکن‌است‌که‌جوابهای به‌دست‌آمده از (۵۲.۳)، همتافت باشند. در چنین حالتهایی مؤلفهٔ همتافت معمولا از نظر مقدار آنقدرکوچك‌است‌که می‌تواند نادیده‌گرفته شود. درحقیقت در زیربرنامه‌ای‌که در زیرداده شده‌است، هنگام جستجو برای یك ریشهٔ حقیقی، هر مؤلفهٔ همتافتی‌که به‌دست‌آید می‌تواند حذف شود.

در الگوریتم۱۰.۳ دنبالهٔ مراحل خواسته شده در روش مولر به‌صورت رسمی بیان شده‌است.

الگوریتم ۱۰.۳: روش مولر

۱. گیریم x_0، x_1، x_2 سه تقریب برای یك ریشهٔ ξ از $f(x)$ باشد.$f(x_0)$، $f(x_1)$ و $f(x_2)$ را محاسبه‌کنید.

۲. مقادیر زیر را محاسبه‌کنید



شکل ۸۰۳

$$h_2 := x_2 - x_1,\ h_1 := x_1 - x_0$$

$$f[x_2, x_1] = \frac{f(x_2) - f(x_1)}{h_2}$$

$$f[x_1, x_0] = \frac{f(x_1) - f(x_0)}{h_1}$$

۳. قرار دهید $i = 2$.

۴. مقادیر زیر را محاسبه کنید

$$f[x_i, x_{i-1}, x_{i-2}] = \frac{f[x_i, x_{i-1}] - f[x_{i-1}, x_{i-2}]}{(h_i + h_{i-1})}$$

$$c_i := f[x_i, x_{i-1}] + h_i f[x_i, x_{i-1}, x_{i-2}]$$

۵. مقدار زیر را محاسبه کنید

$$h_{i+1} := \frac{-2f(x_i)}{(c_i \pm \sqrt{c_i^2 - 4f(x_i)f[x_i, x_{i-1}, x_{i-2}]})}$$

علامت راطوری انتخاب کنید که مخرج رابطهٔ بالا از لحاظ قدرمطلق بیشترین مقدار را داشته باشد.

۶. قراردهید $x_{i+1}=x_i+h_{i+1}$.

۷. مقادیر زیر را محاسبه کنید

$$f[x_{i+1}, x_i]=\frac{f(x_{i+1})-f(x_i)}{h_{i+1}} \quad \text{و} \quad f(x_{i+1})$$

۸. قراردهید $i=i+1$، مراحل ۴ تا ۷ را تکرار کنید تا آنکه یکی ازملاکهای تعیین‌شدهٔ زیر به‌ازای مقادیر مفروض ε_1 و ε_2 برقرار گردد:

(الف) $|x_i-x_{i-1}|<\varepsilon_1|x_i|$

(ب) $|f(x_i)|<\varepsilon_2$

یا آنکه تعداد بارستها از یك حداکثر، بیشترشود.

یك زیربرنامهٔ کاملی براساس این الگوریتم در زیرداده شده‌است. پارامترهای لازم برای زیربرنامه در کارتهای توضیحی[1] ارائه شده‌اند. ZEROS(I)، یك آرایهٔ یك‌بعدی متضمن براوردهای اولیهٔ ریشه‌های موردنظر است. زیربرنامه به‌طور خودکار دوتقریب اضافی برای ZEROS(I) را به‌صورت ZEROS(I)+۰٫۵ و ZEROS(I)−۰٫۵ حساب می‌کند و آنگاه برطبق الگوریتم مولر پیش می‌رود.

```
      SUBROUTINE MULLER ( FN, FNREAL, ZEROS, N, NPREV, MAXIT, EP1, EP2 )
C  DETERMINES UP TO  N  ZEROS OF THE FUNCTION SPECIFIED BY  FN , USING
C  QUADRATIC INTERPOLATION, I.E., MUELLER'S METHOD .
      EXTERNAL FN
      LOGICAL FNREAL
      INTEGER MAXIT,N,NPREV,   KOUNT
      REAL EP1,EP2,   EPS1,EPS2
      COMPLEX ZEROS(N),   C,DEN,DIVDF1,DIVDF2,DVDF1P,FZR,FZRDFL
     *       ,FZRPRV,H,ZERO,SQR
C******  I N P U T  ******
C  FN  NAME OF A SUBROUTINE, OF THE FORM  FN(Z, FZ)  WHICH, FOR GIVEN
C      Z , RETURNS  F(Z) . MUST APPEAR IN AN  E X T E R N A L  STATE-
C      MENT IN THE CALLING PROGRAM .
C  FNREAL  A LOGICAL VARIABLE. IF .TRUE., ALL APPROXIMATIONS ARE TAKEN
C      TO BE REAL, ALLOWING THIS ROUTINE TO BE USED EVEN IF  F(Z)  IS
C      ONLY DEFINED FOR REAL  Z .
C  ZEROS(1),...,ZEROS(NPREV)  CONTAINS PREVIOUSLY FOUND ZEROS (IF
C      NPREV .GT. 0).
C  ZEROS(NPREV+1),...,ZEROS(N)  CONTAINS FIRST GUESS FOR THE ZEROS TO BE
C      FOUND. (IF YOU KNOW NOTHING, 0  IS AS GOOD A GUESS AS ANY.)
C  MAXIT  MAXIMUM NUMBER OF FUNCTION EVALUATIONS ALLOWED PER ZERO.
C  EP1  ITERATION IS STOPPED IF ABS(H) .LT. EP1*ABS(ZR), WITH
C       H = LATEST CHANGE IN ZERO ESTIMATE  ZERO .
C  EP2  ALTHOUGH THE  EP1  CRITERION IS NOT MET, ITERATION IS STOPPED IF
C       ABS(F(ZERO)) .LT. EP2 .
C  N  TOTAL NUMBER OF ZEROS TO BE FOUND .
C  NPREV  NUMBER OF ZEROS FOUND PREVIOUSLY .
C******  O U T P U T  ******
C  ZEROS(NPREV+1), ..., ZEROS(N)  APPROXIMATIONS TO ZEROS .
C
C              INITIALIZATION
      EPS1 = MAX(EP1, 1.E-12)
      EPS2 = MAX(EP2, 1.E-20)
```

1. comment cards

```
C
      DO 100 I=NPREV+1,N
         KOUNT = 0
C              COMPUTE FIRST THREE ESTIMATES FOR ZERO AS
C                 ZEROS(I)+5., ZEROS(I)-.5, ZEROS(I)
    1    ZERO = ZEROS(I)
         H = .5
         CALL DFLATE(FN, ZERO+.5, I, KOUNT, FZR, DVDF1P, ZEROS, *1)
         CALL DFLATE(FN, ZERO-.5, I, KOUNT, FZR, FZRPRV, ZEROS, *1)
         HPREV = -1.
         DVDF1P = (FZRPRV - DVDF1P)/HPREV
         CALL DFLATE(FN, ZERO, I, KOUNT, FZR, FZRDFL, ZEROS, *1)
C        DO WHILE  KOUNT.LE.MAXIT OR  H  IS RELATIVELY BIG
C            OR  FZR = F(ZERO)  IS NOT SMALL
C            OR  FZRDFL = FDEFLATED(ZERO)  IS NOT SMALL OR NOT MUCH
C             BIGGER THAN ITS PREVIOUS VALUE  FZRPRV .
   40       DIVDF1 = (FZRDFL - FZRPRV)/H
            DIVDF2 = (DIVDF1 - DVDF1P)/(H + HPREV)
            HPREV = H
            DVDF1P = DIVDF1
            C = DIVDF1 + H*DIVDF2
            SQR = C*C - 4.*FZRDFL*DIVDF2
            IF (FNREAL .AND. REAL(SQR) .LT. 0.)  SQR = 0.
            SQR = SQRT(SQR)
            IF (REAL(C)*REAL(SQR)+AIMAG(C)*AIMAG(SQR) .LT. 0.) THEN
               DEN = C - SQR
            ELSE
               DEN = C + SQR
            END IF
            IF (ABS(DEN) .LE. 0.)  DEN = 1.
            H = -2.*FZRDFL/DEN
            FZRPRV = FZRDFL
            ZERO = ZERO + H
            IF (KOUNT .GT. MAXIT)        GO TO 99
C
   70       CALL DFLATE(FN, ZERO, I, KOUNT, FZR, FZRDFL, ZEROS, *1)
C                    CHECK FOR CONVERGENCE
            IF (ABS(H) .LT. EPS1*ABS(ZERO)) GO TO 99
            IF (MAX(ABS(FZR),ABS(FZRDFL)) .LT. EPS2)  GO TO 99
C                    CHECK FOR DIVERGENCE
            IF (ABS(FZRDFL) .GE. 10.*ABS(FZRPRV))  THEN
               H = H/2.
               ZERO = ZERO - H
                                         GO TO 70
            ELSE
                                         GO TO 40
            END IF
   99    ZEROS(I) = ZERO
  100 CONTINUE
                                         RETURN
      END
      SUBROUTINE DFLATE ( FN, ZERO, I, KOUNT, FZERO, FZRDFL, ZEROS, * )
C  TO BE CALLED IN   M U L L E R
      INTEGER I,KOUNT,  J
      COMPLEX FZERO,FZRDFL,ZERO,ZEROS(I),  DEN
      KOUNT = KOUNT + 1
      CALL FN(ZERO, FZERO)
      FZRDFL = FZERO
      IF (I .LT. 2)                      RETURN
      DO 10 J=2,I
         DEN = ZERO - ZEROS(J-1)
         IF (ABS(DEN) .EQ. 0.) THEN
            ZEROS(I) = ZERO*1.001
                                         RETURN 1
         ELSE
            FZRDFL = FZRDFL/DEN
         END IF
   10 CONTINUE
                                         RETURN
      END
```

روش مولر مانند الگوریتمهای دیگری که دراین فصل موردبحث قرار گرفته‌اند، در هرمرحله یك ریشه را محاسبه می‌کند. برای محاسبهٔ بیش ازیك ریشه، دراین‌روش ازشیوهٔ

معروف **پایین‌آوری**[1] استفاده می‌شود. برای‌مثال،اگر یك ریشهٔ ξ_1 قبلا محاسبه شده باشد، زیربرنامهٔ ریشهٔ بعدی را با استفاده از تابع زیر محاسبه می‌کند

$$f_1(x)=\frac{f(x)}{x-\xi_1} \qquad (۵۳.۳)$$

ما قبلا هنگام حل معادله‌های بسجمله‌ای با روش‌نیوتن، با این تکنیك که درآن تابع پایین آمده یا تقلیل یافتهٔ $f_1(x)$، نتیجهٔ جانبی الگوریتم بود برخورد داشتیم. در روش مولر، اگر r ریشهٔ ξ_1، ξ_2، ...، ξ_r قبلا محاسبه شده باشند، آنگاه ریشهٔ بعدی با به‌کار گرفتن تابع تقلیل‌یافتهٔ زیرمحاسبه می‌شود

$$f_r(x)=\frac{f(x)}{(x-\xi_1)(x-\xi_2)\cdots(x-\xi_r)} \qquad (۵۴.۳)$$

اگر هیچ تخمینی داده نشده باشد، زیربرنامه همواره ریشه را به‌ترتیب صعودی تجسس می‌کند، زیرا این‌عمل معمولا رشد خطای ناشی از گردکردن را حداقل می‌سازد.همچنین این زیربرنامه کلیهٔ ریشه‌هایی راکه ازبه‌کارگرفتن تابع تقلیل‌یافته به‌دست می‌آیند درتابع اصلی $f(x)$ قرارمی‌دهد و بدین‌وسیله دقت‌آنها راکنترل می‌کند. درعمل، استفاده ازتابع تقلیل‌یافته برای تعیین ریشه تاحدی از دقت عمل می‌کاهد. ریشه‌های تقریبی حاصل‌ازروش پایین‌آوری می‌توانند‌اصلاح شوند، هرگاه این‌ریشه‌ها به‌عنوان حدسهای اولیهٔ روش‌نیوتن که به‌تابع اصلی اعمال می‌گردد، به‌حساب آیند. در کاربرد زیربرنامهٔ مولر، استفاده‌کننده می‌تواند تعداد ریشه‌های موردنظر را مشخص نماید. برای مثال برخی توابع ممکن است دارای بینهایت ریشه باشندکه تنها چند ریشهٔ اول آنها موردنظر باشند.

□ **مثال۱۳.۳:** تابع بسل[2] $J_0(x)$، به‌توسط سری نامتناهی زیرداده شده‌است

$$J_0(x)=1-\frac{x^2}{2^2\times1\times1}+\frac{x^4}{2^4\times2!\times2!}-\frac{x^6}{2^6\times3!\times3!}+\cdots$$

می‌دانیم که $J_0(x)$ دارای بینهایت ریشهٔ حقیقی‌است.با استفاده از الگوریتم ۱۰.۳، سه تا از اولین ریشه‌های مثبت را پیدا کنید. نتایج‌کامپیوتری‌که در زیر ارائه می‌شود با استفاده از زیربرنامهٔ استانداردِ موجود در خودکامپیوتر،که براساس سری فوق برای $J_0(x)$ تهیه شده، روی یك‌کامپیوتر IBM ۷۰۹۴ به‌دست‌آمده‌است. دراین برنامه مقادیر$J_0(x)$ با حداکثردقت عمل محاسبه شده‌اند.

بارستها همگی با تقریبهای اولیهٔ $x_0=-1$، $x_1=1$، $x_2=0$ شروع شده، تا برقراری یکی از ملاکهای زیر ادامه یافته‌اند:

---

1. deflation 2. Bessel

$$\frac{|x_{i+1}-x_i|}{|x_{i+1}|} < 10^{-6} \qquad \text{(الف)}$$

$$|J_0(x_i)| < 10^{-20} \qquad \text{(ب)}$$

مقادیر همگراشده تا شش رقم با معنی صحیح‌اند. نکتهٔ دیگر اینکه مقادیر ریشه‌ها به‌ترتیب صعودی محاسبه شده‌اند.

نتایج کامپیوتری برای مثال ۱۳.۳

Zero 1	Zero 2	Zero 3
−0.09999999E 01	−0.09999999E 01	−0.09999999E 01
0.09999999E 01	0.09999999E 01	0.09999999E 01
0.	0.	0.
0.20637107E 01	0.36557332E 01	0.47983123E 01
0.23167706E 01	0.44416171E 01	0.59396663E 01
0.23970029E 01	0.50863190E 01	0.70758440E 01
0.24047983E 01	0.55024961E 01	0.88981197E 01
0.24048255E 01	0.55202182E 01	0.92976399E 01
0.24048255E 01	0.55200780E 01	0.86854592E 01
	0.55200780E 01	0.86529856E 01
		0.86537299E 01
		0.86537278E 01

کلیهٔ مثالهای زیرروی یك کامپیوتر CDC ۶۵۰۰ با استفاده از الگوریتم ۱۰.۳ اجرا شده‌اند. ملاکهای خطا برای این مثالها $\varepsilon_1 = \varepsilon_2 = 10^{-8}$ می‌باشند وهمگی ازهمان مقادیر اولیه (۰٫۵ ,۰٫۵ —) و سپس پایین‌آوری استفاده کرده‌اند. گرچه نتایج تا هشت رقم با معنی چاپ شده‌اند ولی به‌یاد داشته باشید که در کامپیوتر CDC ۶۵۰۰ طول کلمه درمحاسبات با ممیز شناور برابر با ۱۴ رقم اعشاری است. برونداد عبارت است از اجزای حقیقی و انگاری تقریبات همگرا به‌ریشه‌ها، واجزای حقیقی و انگاری مقادیر تابع به‌ازای آن ریشه‌ها.

□ **مثال ۱۴.۳:** کلیهٔ ریشه‌های بسجمله‌ای زیر را پیدا کنید

$$p(x) = x^3 + x - 3$$

ROOT		F(ROOT)	
جزء حقیقی	جزء انگاری	جزء حقیقی	جزء انگاری
1.2134117E + 00	0.	−4.2632564E − 14	0.
−6.0670583E − 01	1.4506122E + 00	2.8421709E − 14	4.2632564E − 14
−6.0670583E − 01	−1.4506122E + 00	2.8421709E − 14	−2.6290081E − 13

نتایج فوق را با نتایج مثال ۱۰.۳، که با استفاده ازماشین حساب به‌دست‌آمده‌است،

مقایسه کنید. توجه کنید که چون $p(x)$ دارای ضرایب حقیقی است، ریشه‌های همتافت به‌صورت جفتهای مزدوج ظاهر می‌شوند. همچنین توجه کنید که در اینجا هیچ نیازی به برآورد ریشه‌های همتافت نیست. در حالی که می‌توان روش نیوتن را برای محاسبهٔ ریشه‌های همتافت به‌کار گرفت، ولی این روش احتیاج به یك برآورد مناسبی دارد که به‌دست آوردن آن معمولا بسیار مشکل است. توجه شود که خطای موجود در F(ROOT)، همان گونه که ملاك خطا آن را تجویز کرده است، به‌طور قابل ملاحظه‌ایی کوچکتر از 10^{-8} است. درحقیقت در آخرین بار ست می بایستی خطا از مقداری مانند 10^{-7} به 10^{-14} کاهش یابد و این امر نشان می‌دهد که روش مورد بحث تقریباً به‌طور مربعی همگرا است. □

□ **مثال ۱۵.۳:** ریشه‌های بسجمله‌ای زیر را پیدا کنید

$$f(x)=x^5-3.7x^4+7.4x^3-10.8x^2+10.8x-6.8$$

این همان مثال ۱۱.۳ است که قبلا به‌وسیلهٔ روش نیوتن حل شده بود. ریشه‌های دقیق این بسجمله‌ای، i، $1\pm i$، و $1.7\pm\sqrt{2}\,i$، هستند. با وجوداینکه ریشهٔ انگاری محض، $\pm\sqrt{2}\,i$، دارای قسمت حقیقی کوچکی است، نتایج زیر تا هشت رقم با معنی صحیح اند.

ROOT		F(ROOT)	
جزء حقیقی	جزء انگاری	جزء حقیقی	جزء انگاری
1.0000000E + 00	−1.0000000E + 00	−2.8421709E − 14	2.8421709E − 14
1.0000000E + 00	1.0000000E + 00	1.4210855E − 13	−1.5631940E − 13
−9.0964472E − 12	1.4142136E + 00	−8.5330498E − 10	−1.2773698E − 09
8.3306265E − 11	−1.4142136E + 00	−8.5339025E − 10	−1.2785158E − 09
1.7000000E + 00	1.3036419E − 10	6.3431571E − 10	9.4984654E − 10

□

□ **مثال ۱۶.۳:** ریشه‌های بسجمله‌ای زیر را محاسبه کنید

$$f(x)=x^7-28x^6+322x^5-1960x^4+6769x^3-13132x^2+13068x-5040$$

این مثال با روش نیوتن در مثال ۱۲.۳ حل شد و در به‌دست آوردن جوابهای دقیق آن با اشکالاتی مواجه شدیم. ریشه‌های این بسجمله‌ای برابر با $x=1, 2, 3, 4, 5, 6, 7$ هستند. نتایج زیر بسیار دقیق‌اند، گرچه طول نسبتاً بزرگ کلمه در کامپیوتر CDC ۶۵۰۰ در این دقت عمل نقش اساسی داشته است. همچنین باید توجه کرد که گرچه در حالت کلی روش مولر ریشه‌ها را به‌ترتیب صعودی تجسس می‌کند، اما در این مثال به‌چنین امری توفیق نیافته است.

ROOT		F(ROOT)	
جزء حقیقی	جزء انگاری	جزء حقیقی	جزء انگاری
2.0000000E + 00	−2.6080092E − 16	−5.8207661E − 11	3.1296110E − 14
3.0000000E + 00	7.4093893E − 11	−7.8580342E − 10	3.5565069E − 09
1.0000000E + 00	−1.7030067E − 16	0.	−1.2261648E − 13
6.0000000E + 00	1.6284031E − 15	−1.0710210E − 08	−1.9540837E − 13
5.0000000E + 00	−7.2393906E − 13	2.4156179E − 09	−3.4749075E − 11
4.0000000E + 00	−2.3682266E − 10	2.0954758E − 09	8.5256156E − 09
7.0000000E + 00	−8.1000834E − 20	4.0745363E − 10	−5.8320601E − 17

□ **مثال ۱۷۰۳:** ریشه‌های بسجمله‌ای زیر را پیدا کنید

$$f(x)=x^{8}-۱۷۰\,x^{6}+۷۳۹۲\,x^{4}-۳۹۷۱۲\,x^{2}+۵۱۲۰۰$$

ریشه‌های این بسجمله‌ای برابر با $\pm ۱۰, \pm ۲, \pm ۸, \pm \sqrt{۲}$ هستند. این برنامه در حالت همتافت اجرا شده و ریشه‌هایی با دقت هشت رقم با معنی صحیح به‌دست داده بود. این مثال نشان می‌دهد که الگوریتم فوق نمی‌تواند برای حل بسجمله‌ای از درجهٔ نسبتاً بالا به‌کار برده شود و نتایج خوبی به‌دست دهد (تمرین ۳.۶–۴ را نگاه کنید).

ROOT		F(ROOT)	
جزء حقیقی	جزء انگاری	جزء حقیقی	جزء انگاری
−1.4142136E + 00	0.	−2.3282064E − 10	0.
1.4142136E + 00	0.	−2.3283064E − 10	0.
2.0000000E + 00	0.	0.	0.
−2.0000000E + 00	0.	0.	0.
8.0000000E + 00	0.	2.9336661E − 08	0.
−8.0000000E + 00	0.	2.9336661E − 08	0.
1.0000000E + 01	0.	2.3352914E − 07	0.
−1.0000000E + 01	0.	1.8742867E − 07	0.

تمرین

۱–۷۰۳ ریشه‌های حقیقی و همتافت بسجمله‌ایهای زیر را با استفاده از روش مولر پیدا کنید

(الف) $x^{۷}-۱$

(ب) $x^{۴}-۷x^{۳}+۱۸x^{۲}-۲۰x+۸$

(پ) $x^{۶}+۲x^{۵}+x^{۴}+۳x^{۳}+۵x^{۲}+x+۱$

۲–۷۰۳ معادلهٔ $x-\tan x=۰$ دارای بی‌نهایت ریشه حقیقی‌است. با استفاده از روش مولر اولین سه ریشهٔ مثبت آن را پیدا کنید.

۳–۷۰۳ انتگرال فرنل[1] $C(x)$، با سری زیر تعریف می‌شود

1. Fresnel

$$C(x)=\sum_{n=0}^{\infty}\frac{(-1)^n\left(\frac{\pi}{2}\right)^{2n}}{(2n)!(4n+1)}x^{4n+1}$$

با استفاده از روش مولر، اولین سه ریشهٔ مثبت حقیقی این تابع را به‌دست آورید. عملیات را به‌ازای $n=3$ و کنار گذاشتن سایر جملات در سری فوق، شروع کنید و سپس n را افزایش دهید تا آنکه ریشه‌های صحیحی که شما را قانع سازد به‌دست آید.

۴-۷.۳ تابع بسل از مرتبهٔ ۱ با سری زیر تعریف می‌شود

$$J_1(z)=\frac{z}{2}\sum_{k=0}^{\infty}\frac{(-z^2/4)^k}{k!(k+1)!}$$

اولین چهار ریشهٔ این تابع را مانند تمرین ۳-۷.۳ را به‌دست آورید.

۴

ماتریسها و دستگاههای معادلات خطی

بسیاری از مسائل آنالیز عددی را می‌توان به‌مسئلهٔ حل دستگاههای معادلات خطی تبدیل کرد. از آن جمله می‌توان مسئلهٔ حل معادلات دیفرانسیل معمولی یا معادلات با مشتقات جزئی به‌وسیلهٔ روشهای تفاضل متناهی، حل دستگاههای معادلات، مسائل ویژه مقداری فیزیك ریاضی، برازاندن داده‌ها با روش کوچکترین توانهای دوم و تقریب با بسجمله‌ایها را نام برد. کاربرد نمادگذاری ماتریسی نه‌تنها مناسب، بلکه درنشر روابط بنیادی بسیار تواناست. در بخش ۱.۴ برخی از ویژگیهای سادهٔ ماتریسها راکه در قسمتهای بعدی به‌کار برده می‌شوند معرفی خواهیم کرد. برخی از قضایا و ویژگیها بدون برهان بیان خواهند شد.

۱.۴ ویژگیهای ماتریسها

یك دستگاه m معادلهٔ خطی n مجهولی دارای شکل کلی زیر است

$$\begin{aligned}
a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n&=b_1\\
a_{21}x_1+a_{22}x_2+\cdots+a_{2n}x_n&=b_2\\
&\cdots\cdots\cdots\cdots\\
a_{m1}x_1+a_{m2}x_2+\cdots+a_{mn}x_n&=b_m
\end{aligned}\qquad(۱.۴)$$

ضرایب a_{ij} $(i=۱,\dots,m, j=۱,\dots,n)$ و مقادیر **سمت راست** معادله، b_i $(i=۱,\dots,m)$، اعدادی مفروض‌اند. مسئلهٔ ما عبارت است از پیداکردن، اعداد x_j $(j=۱,\dots,n)$ به‌طوری‌که m معادلهٔ (۱.۴) همزمان برقرار باشند. بحث در باب این مسئله و درك آن زمانی بسیار ساده خواهد شدکه مفاهیم جبری ماتریسها و بردارها رامورد استفاده قراردهیم.

تعریف ماتریس و بردار

یك **ماتریس** یك آرایهٔ مستطیلی‌ازاعداد (معمولاحقیقی)است‌که درسطرها و ستونها، مرتب شده باشند. ضرایب (۱.۴) ماتریسی تشکیل می‌دهندکه آن را A می‌نامیم. مرسوم است‌که ماتریس A به‌صورت زیرنمایش داده شود

$$A=\begin{bmatrix} a_{۱۱} & a_{۱۲} & \cdots & a_{۱n} \\ a_{۲۱} & a_{۲۲} & \cdots & a_{۲n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m۱} & a_{m۲} & \cdots & a_{mn} \end{bmatrix} \qquad (۲.۴)$$

گاهی اوقات ماتریس فوق به‌صورت خلاصه‌تر زیرنوشته می‌شود

$$A=(a_{ij}) \qquad (۳.۴)$$

ماتریس A در (۲.۴) دارای m **سطر** و n **ستون** و یا **از مرتبهٔ** $m\times n$ است. a_{ij}، **درایهٔ** (i,j)ام، درمحل تقاطع سطر iام و ستون jام A قرار دارد. اگر A یك ماتریس $n\times n$ باشد می گویندکه A **ماتریس مربع و از مرتبهٔ** nاست. اگرماتریسی فقط یك ستون داشته باشد آن را یك **بردار ستونی** و اگـر فقط یك سطرداشته باشد، آن را یك **بردار سطری** می‌نامند. بردارهای ستونی را با حروف کوچك سیاه نشان می‌دهیم و بدین‌ترتیب آنها را از بردارهای سطری متمایزمی‌سازیم و مختصراً آنها را **بردار** می گوییم. بنابراین ثابتهای سمت راست یعنی b_iها$(i=۱,\dots,m)$ومجهولات x_j،$(j=۱,\dots,m)$ هردوبردارهای زیر را تشکیل می‌دهند

$$\mathbf{b}=\begin{bmatrix} b_۱ \\ b_۲ \\ b_۳ \\ \vdots \\ b_m \end{bmatrix} \qquad \mathbf{x}=\begin{bmatrix} x_۱ \\ x_۲ \\ x_۳ \\ \vdots \\ x_n \end{bmatrix} \qquad (۴.۴)$$

گوییم‌که $\mathbf{b}$ یك m-برداری و $\mathbf{x}$ یك n-برداری‌است.

تساوی

اگر $A=(a_{ij})$ و $B=(b_{ij})$ دو ماتریس باشند، گوییم A **مساوی** B است یا $A=B$، به‌شرط آنکه A و B هم‌مرتبه باشند و به‌ازای کلیهٔ مقادیر i و j، رابطهٔ $a_{ij}=b_{ij}$ برقرار باشد.

ضرب ماتریسها

در اصطلاحاتی که تاکنون وارد شده‌اند، (۱.۴) گویای آن‌است که اگر ماتریس A به‌طریق خاصی با ماتریس یك ستونی، یا بردار $\mathbf{x}$، ترکیب شود حاصل می‌بایستی مساوی ماتریس یك ستونی، یا بردار $\mathbf{b}$، شود. روند ترکیب ماتریسها که در اینجا دخالت کرده‌اند ضرب ماتریس نام دارد و به‌طور کلی به‌صورت زیر تعریف می‌شود: گیریم $A=(a_{ij})$ یك ماتریس $m\times n$ و $B=(b_{ij})$ یك ماتریس $n\times p$ باشد، آنگاه ماتریس $C=(c_{ij})$، حاصلضرب (ماتریسهای) A و B (با همین ترتیب) و یا $C=AB$ است، درصورتی که C ازمرتبهٔ $m\times p$ باشد و داشته باشیم

$$c_{ij}=\sum_{k=1}^{n} a_{ik}b_{kj} \qquad j=1,\dots,p \quad و \quad i=1,\dots,m \quad \text{به‌ازای} \qquad (۵.۴)$$

به‌طور لفظی درایهٔ (i,j)ام حاصلضرب A و B یا $C=AB$، بدین‌صورت محاسبه می‌شود که n درایهٔ سطر iام ماتریس A و n درایهٔ ستون jام ماتریس B نظیر به‌نظیر درهم ضرب و این n حاصلضربها با هم جمع می‌شوند.

□ **مثال:**

$$B=\begin{bmatrix} 2 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \quad و \quad A=\begin{bmatrix} 3 & 0 & 2 \\ 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix} \quad \text{اگر}$$

$$AB=\begin{bmatrix} 8 & 3 \\ 2 & 3 \\ 1 & 1 \end{bmatrix} \qquad \text{آنگاه}$$

برای مثال درایهٔ (۱، ۲) از AB، همان‌طوری که به‌وسیلهٔ پیکانها نشان داده شده‌اند، از ترکیب سطر دوم A و ستون اول B به‌دست آمده‌است:

$$1\times 2+2\times 0+0\times 1=2$$

□

با این تعریف از حاصلضرب ماتریسها و تعاریف (۲.۴) و (۴.۴)، می‌توان دستگاه (۱.۴) را به‌صورت سادهٔ زیر نوشت

$$A\mathbf{x}=\mathbf{b} \qquad (۶.۴)$$

فعلا، چنین به‌نظر می‌رسد که این سادگی بیان به‌بهای چندین تعریف به‌دست‌آمده، که یکی از آنها کاملا پیچیده‌است، ولـی بسیاری از مزایای قرارداد ماتریسی طی این فصل آشکار می‌شوند.

ضرب ماتریسها به‌هیچ‌وجه مشابه باضرب اعداد نیست. برای مثال، تعیین حاصلضرب ماتریس A در ماتریس B تنها وقتی امکان‌پذیر است که تعداد ستونهای A مساوی با تعداد سطرهای B باشد. بنابراین حتی وقتی که حاصلضرب AB تعریف شده باشد، حاصلضرب B با A الزاماً تعریف شده نیست. وانگهی، حتی اگر هردو حاصلضرب AB و BA تعریف شده باشند این‌دو الزاماً مساوی نخواهند بود.

□ **مثال:**

اگر $A=\begin{bmatrix}۲ & ۱\\ ۱ & ۳\end{bmatrix}$ و $B=\begin{bmatrix}۲ & ۱\\ ۰ & ۱\end{bmatrix}$، آنگاه

$$AB=\begin{bmatrix}۴ & ۳\\ ۲ & ۴\end{bmatrix}\neq\begin{bmatrix}۵ & ۵\\ ۱ & ۳\end{bmatrix}=BA$$ □

ازطرف دیگر ضرب ماتریس ویژگی شرکتپذیری[1] دارد: اگر A ، B و C به‌ترتیب ماتریسهایی از مرتبهٔ $m\times n$ و $n\times p$ و $p\times q$ باشند، آنگاه داریم

$$(AB)C=A(BC) \qquad (۷.۴)$$

این مطلب را می‌توان به‌صورت زیر توجیه کرد: چون A از مرتبهٔ $m\times n$ و B از مرتبـهٔ $n\times p$ است، پس AB تعریف شده وازمرتبهٔ $m\times p$ است وبنابراین $(AB)C$ تعریف‌شده و از مرتبهٔ $m\times q$ است. به‌طریق مشابه می‌توان نشان داد که $A(BC)$ نیز تعریف‌شده و از مرتبهٔ $m\times q$ است، بنابراین حداقل یك شرط برای تساوی برقرار است. بعلاوه:

$$(AB)C \text{ ام در } (i,j) \text{ درایهٔ} = \sum_{k=۱}^{p} [AB \text{ ام در } (k,i) \text{ درایهٔ}]\times c_{kj}$$

---

1. associative

$$= \sum_{k=1}^{p} \left[\sum_{r=1}^{n} a_{ir} \times b_{rk} \right] \times c_{kj}$$

$$= \sum_{r=1}^{n} a_{ir} \times \left[\sum_{k=1}^{p} b_{rk} \times c_{kj} \right]$$

$$= \sum_{r=1}^{n} a_{ir} \times [BC \text{ درایهٔ } (r, j) \text{ام در}]$$

$$= A(BC) \text{ درایهٔ } (i, j) \text{ام در}$$

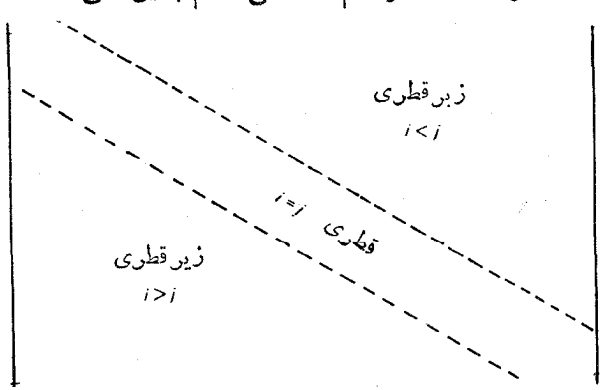
که ثابت می‌کند $(AB)C = A(BC)$. درحالت خاصی که C یك بردار (از مرتبهٔ مناسب باشد) کراراً از تساوی بالا استفاده خواهیم کرد یعنی

$$(AB)\mathbf{x} = A(B\mathbf{x})$$

ماتریسهای قطری و مثلثی

اگر $A = (a_{ij})$ یك ماتریس مربعی از مرتبهٔ n باشد، درایه‌های a_{11}، a_{22}، ... a_{nn} را درایه‌های **قطری** و درایه‌های دیگر را درایه‌های **غیرقطری** گویند. تمامی درایه‌های a_{ij}ی A به‌ازای $i < j$ درایه‌های **زیر قطری** و تمام درایه‌های a_{ij} به‌ازای $i > j$ درایه‌های **زیرقطری** نامیده می‌شوند (به‌شکل ۱.۴ نگاه کنید).

اگر کلیهٔ درایه‌های غیرقطری ماتریس مربعی A صفر باشند، A **ماتریس‌قطری** نامیده می‌شود. اگر کلیهٔ درایه‌های زیرقطری ماتریس مربعی A صفر باشند، آنگاه A یك **ماتریس بالامثلثی** (یا راست مثلثی) نامیده می‌شود، درحالی که اگر تمام درایه‌های زیرقطری ماتریس مربعی A صفر باشند A را **ماتریس پایین مثلثی** (یا چپ مثلثی) گویند. واضح است که یك ماتریس قطری است اگر، و فقط اگر، هم بالامثلثی و هم پایین‌مثلثی باشد.



شکل ۱.۴

□ **مثال:**

درنمونه‌های زیرماتریسهای A و C قطری و A، B و C ماتریسهای بالامثلثی و A، C و D ماتریسهای پایین‌مثلثی هستند و E هیچکدام از این ویژگیها را ندارد

$$A=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} \qquad B=\begin{bmatrix}0&1&2\\0&3&1\\0&0&4\end{bmatrix} \qquad C=\begin{bmatrix}0&0&0\\0&0&0\\0&0&0\end{bmatrix}$$

□

$$D=\begin{bmatrix}10&0&0\\8&7&0\\1&1&1\end{bmatrix} \qquad E=\begin{bmatrix}1&0&0\\3&2&6\\0&0&4\end{bmatrix}$$

ماتریس واحد و وارون ماتریس

یك ماتریس قطری از مرتبهٔ n راکه تمام درایه‌های قطری‌آن یك باشند **ماتریس واحد** از مرتبهٔ nگویند وبا حرف مخصوص I یا I_n (درصورتی‌که مرتبه مهم باشد) نشان می‌دهند. این نام برای ماتریس واحد بدین‌منظور انتخاب شده‌است که

به‌ازای هرماتریس A از مرتبهٔ $n\times p$ داریم: $I_nA=A$

برای هرماتریس B از مرتبهٔ $m\times n$ داریم: $BI_n=B$

مشاهده می‌شودکه ماتریس I دقیقاً همان نقش عدد یك را درضرب معمولی ایفا می‌کند.

تقسیم ماتریسها، در حالت‌کلی تعریف نشده‌است. لیکن برای ماتریسهای مربعی در ارتباط با مفهوم تقسیم، مفهومی موسوم به‌وارون ماتریس تعریف شده است. گویند ماتریس مربعی A ازمرتبهٔ n **وارونپذیر** است به‌شرط آنکه ماتریسی مانند B ازمرتبهٔ n وجود داشته باشد به‌طوری‌که

$$AB=I=BA \qquad (۸.۴)$$

به‌عنوان مثال، ماتریس $A=\begin{bmatrix}1&1\\0&1\end{bmatrix}$ وارونپذیر است، زیرا داریم

$$\begin{bmatrix}1&1\\0&1\end{bmatrix}\begin{bmatrix}1&-1\\0&1\end{bmatrix}=\begin{bmatrix}1&0\\0&1\end{bmatrix}=\begin{bmatrix}1&-1\\0&1\end{bmatrix}\begin{bmatrix}1&1\\0&1\end{bmatrix}$$

از طرف دیگرماتریس $A=\begin{bmatrix}1&2\\2&4\end{bmatrix}$ وارونپذیرنیست، برای اینکه اگر B ماتریسی

باشد به‌طوری‌که $BA=I$، آنگاه نتیجه می‌شود که

$$\begin{bmatrix} b_{11}+2b_{12} & 2b_{11}+4b_{12} \\ b_{21}+2b_{22} & 2b_{21}+4b_{22} \end{bmatrix} = BA = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

بنابراین می‌بایستی هم داشته باشیم $b_{11}+2b_{12}=1$ و همزمان با آن

$$2(b_{11}+2b_{12}) = 2b_{11}+4b_{12} = 0$$

و این، غیرممکن است.

توجه داشته باشید که رابطهٔ (۸.۴) می‌تواند حداکثر برای یك ماتریس B برقرار باشد. زیرا اگر داشته باشیم

$$AB=I \quad و \quad CA=I$$

که در آن B و C ماتریسهای مربعی و از مرتبهٔ مشابه A باشند، آنگاه داریم

$$C=CI=C(AB)=(CA)B=IB=B$$

و این بدان معنی است که B و C باید مساوی باشند. بنابراین اگر A وارونپذیر باشد، دقیقاً یك ماتریس B وجود دارد که رابطهٔ (۸.۴) را برقرار می‌کند، این ماتریس، **وارون** A نامیده شده و با A^{-1} نشان داده می‌شود.

بلافاصله از (۸.۴) نتیجه می‌شود که اگر A وارونپذیر باشد، A^{-1} نیز چنین است و وارون آن A می‌باشد، یعنی

$$(A^{-1})^{-1}=A \qquad (۹.۴)$$

بعلاوه اگر ماتریسهای A و B هردو مربعی و وارونپذیر و از یك مرتبه باشند، حاصلضرب آنها نیز وارونپذیر است و داریم

$$(AB)^{-1}=B^{-1}A^{-1} \qquad (۱۰.۴)$$

به تغییر در ترتیب ماتریسها توجه کنید! اثبات (۱۰.۴) به شرکتپذیری ضرب ماتریسها منوط است

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AA^{-1} = I$$

$$(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}B = I$$

□ **مثال:** وارون ماتریس $A=\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ ماتریس $A^{-1}=\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$ و وارون ماتریس $B=\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ ماتریس $B^{-1}=\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$ است. بعلاوه $AB=\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$،

و بنابراین بـا استفـاده از (۱۰.۴) داریم $(AB)^{-1}=\begin{bmatrix} ۱ & -۱ \\ -۱ & ۲ \end{bmatrix}=B^{-1}A^{-1}$. از طرف دیگر

$$A^{-1}B^{-1}=\begin{bmatrix} ۲ & -۱ \\ -۱ & ۱ \end{bmatrix}$$

و

$$(AB)(A^{-1}B^{-1})=\begin{bmatrix} ۲ & ۱ \\ ۱ & ۱ \end{bmatrix}\begin{bmatrix} ۲ & -۱ \\ -۱ & ۱ \end{bmatrix}=\begin{bmatrix} ۳ & -۱ \\ ۱ & ۰ \end{bmatrix}\neq I$$

یعنی $A^{-1}B^{-1}$ نمی‌تواند وارون AB باشد. □

جمع ماتریسها و ضرب ماتریس دریك اسکالر

ممکن‌است یك ماتریس را در یك اسکالر[1] (= عدد) ضرب و دو ماتریس از مرتبه‌هـای مساوی را به‌طور مناسبی بـاهم جمع کرد. اول اگر $A=(a_{ij})$ و $B=(b_{ij})$ دو ماتریس باشند و d یك عدد، گوییم B **حاصلضرب** d در A یا $B=dA$ درصورتی‌که B و A دارای مرتبه‌های مساوی باشند و به‌ازای جمیع مقادیر i و j داشته باشیم

$$b_{ij}=da_{ij}$$

بعلاوه اگر $A=(a_{ij})$ و $B=(b_{ij})$ مـاتریسهایی از مرتبه‌های مساوی باشند و $C=(c_{ij})$ یك ماتریس باشد، گوییم C **حاصلجمع** A و B است، یا $C=A+B$، درصورتی‌که مرتبهٔ C مساوی مرتبه‌های A و B باشد و به‌ازای جمیع مقادیر i و j داشته باشیم:

$$c_{ij}=a_{ij}+b_{ij}$$

بنابراین ضرب یك ماتریس دریك عدد و جمع ماتریسها، درایه به‌درایه انجـام می‌گیرد. قواعد امر دررابطه باعملیات فوق و همچنین ضرب ماتریسها به‌آسانی قابل‌تحقیق هستند. فرض‌کنید A ، B و C ماتریسهایی باشندکه تمام حاصلجمعها و حاصلضربهای مذکور در زیربرای‌آنها تعریف شده باشند و a و b نیز دوعدد باشند، آنگاه

---

1. scalar

$$
\begin{aligned}
&\text{(i)} \quad A+B=B+A \\
&\text{(ii)} \quad (A+B)+C=A+(B+C) \\
&\text{(iii)} \quad a(A+B)=aA+aB \\
&\text{(iv)} \quad (a+b)A=aA+bA \\
&\text{(v)} \quad (A+B)C=AC+BC \\
&\text{(vi)} \quad A(B+C)=AB+AC \\
&\text{(vii)} \quad a(AB)=(aA)B=A(aB)
\end{aligned}
\tag{۱۱.۴}
$$

(viii) اگر $a \neq 0$ و A وارونپذیر باشد، آنگاه aA وارونپذیر است و داریم $(aA)^{-1}=(1/a)A^{-1}$

برای توضیح بیشتر برهان (vi) در زیر داده می‌شود. اگر ماتریس A از مرتبهٔ $m \times n$ و B و C ازمرتبهٔ $n \times p$ باشند، هردو طرف (vi)، ماتریسهایی تعریف شده و از مرتبهٔ $m \times p$ خواهند بود. بعلاوه

$$
\begin{aligned}
A(B+C) \text{ در } (i,j) \text{ام } &= \text{درایهٔ } \sum_{k=1}^{n} a_{ik} \times [B+C \text{ در } (k,j) \text{ام درایهٔ}] \\
&= \sum_{k=1}^{n} a_{ik} \times [b_{kj}+c_{kj}] \\
&= \sum_{k=1}^{n} a_{ik}\, b_{kj} + \sum_{k=1}^{n} a_{ik} c_{kj} \\
&= [AB \text{ در } (i,j) \text{ام درایهٔ}] + [AC \text{ در } (i,j) \text{ام درایهٔ}] \\
&= (AB+AC) \text{ در } (i,j) \text{ام درایهٔ}
\end{aligned}
$$

و بالاخره اگر کلیهٔ درایه‌های ماتریس A از مرتبهٔ $m \times n$، صفر باشند، آنگاه این ماتریس را **ماتریس صفر** از مرتبهٔ $m \times n$ گویند و با حرف مخصوص O نشان می‌دهند. ماتریس صفر دارای ویژگی آشکار زیراست که به‌ازای کلیهٔ ماتریسهای هم‌مرتبهٔ خودش مانند B، داریم

$$B+O=B$$

ترکیبهای خطی

تعریف جمع ماتریسها وحاصلضرب اعداد در ماتریسها، بخصوص، امکان جمع n-بردارها

با هم و ضرب n-برداریها در اعداد را به‌وجود می‌آورد. اگر $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$ و k n-برداری و $b_1, b_2, \ldots, b_k$، k عدد باشند، آنگاه حاصلجمع وزن[1] دار

$$b_1\mathbf{x}^{(1)}+b_2\mathbf{x}^{(2)}+\ldots+b_k\mathbf{x}^{(k)}$$

نیز یك n-برداری‌است که به ترکیب خطی بردارهای $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$ با وزنها یا ضریبهای $b_1, \ldots, b_k$ موسوم‌است.

اکنون باردیگر دستگاه معادلات(۱.۴) را بررسی می‌کنیم. به‌ازای $j=1,2,\ldots,n$، گیریم $\mathbf{a}_j$ معرف ستون jام ماتریس ضرایب A ازمرتبهٔ $m\times n$ باشد، یعنی $\mathbf{a}_j$ یك m-برداری باشدکه درایهٔ iام آن عدد a_{ij} به‌ازای $i=1,\ldots,m$ باشد، دراین‌صورت می‌توان m-برداری $A\mathbf{x}$ را به‌صورت

$$A\mathbf{x}=x_1\mathbf{a}_1+x_2\mathbf{a}_2+\ldots+x_n\mathbf{a}_n$$

یعنی به‌عنوان یك ترکیب خطی‌از n ستون A با درایه‌های وزن‌دار $\mathbf{x}$ نوشت. بنابراین‌حل مسئلهٔ (۱.۴) باییان زیرهم‌ارز است: وزنهای $x_1, \ldots, x_n$ را طوری پیداکنیدکه‌ترکیب خطی nستون A با این وزنها مساوی m-برداری $\mathbf{b}$ی سمت راست شود.

در رابطه با نشانگزاریهای سازگار فوق، ستون jام ماتریس واحد I با نمادخاص

$$\mathbf{i}_j$$

نشان داده می‌شود. واضح‌است که‌تمامی درایه‌های $\mathbf{i}_j$ صفرند به‌استثنای عنصر jام‌که یك است. چنین مرسوم‌است که $\mathbf{i}_j$ را jامین بردار**واحد** می‌نامند (همان‌گونه‌که باماتریس‌واحد عمل‌کردیم، طول و یا مرتبهٔ $\mathbf{i}_j$ راصریحاً نشان نمی‌دهیم زیراکه مرتبهٔ آن ازمتن مشخص می‌شود). با این نشانگزاریها برای هر n-برداری $\mathbf{b}=(b_i)$ داریم

$$b_1\mathbf{i}_1+b_2\mathbf{i}_2+\ldots+b_n\mathbf{i}_n=\mathbf{b}$$

بعلاوه، ستونjام ماتریسA، یعنی $\mathbf{a}_j$ را می‌توان به‌وسیلهٔ ضرب A در $\mathbf{i}_j$ نیز به‌دست‌آورد، یعنی

$$\mathbf{a}_j=A\mathbf{i}_j$$

بنابراین اگر $C=AB$، آنگاه

$$C\mathbf{i}_j=(AB)\mathbf{i}_j=A(B\mathbf{i}_j)$$

لذا ستون jام حاصلضرب AB از ضرب اولین عاملA در ستون jام عامل دوم B به‌دست می‌آید.

---

1. weighted sum

وجود و یکتایی جوابهای دستگاه معادلات (۱.۴)

در بخشهای بعد، مختصراً با دستگاه معادلات خطی، که ماتریس‌ضرایب مربعی دارند، سروکار خواهیم داشت. این مطلب بدین‌صورت توجیه می‌شود که نشان می‌دهیم دستگاه معادلات (۱.۴) نمی‌تواند به‌ازای هرعامل سمت راست دقیقاً یك جواب داشته باشد مگر اینکه ماتریس ضرایب مربعی باشد.

لم۱.۴ اگر $\mathbf{x}=\mathbf{x}_1$ یك جواب دستگاه معادلهٔ خطی $A\mathbf{x}=\mathbf{b}$، باشد، آنگاه هر جواب $\mathbf{x}=\mathbf{x}_2$ از این دستگاه به‌شکل زیر است

$$\mathbf{x}_2=\mathbf{x}_1+\mathbf{y}$$

در رابطهٔ بالا $\mathbf{x}=\mathbf{y}$ یك جواب از دستگاه همگن[1] $A\mathbf{x}=0$ است.

درحقیقت اگر $\mathbf{x}_1$ و $\mathbf{x}_2$ هردو، ریشهٔ دستگاه $A\mathbf{x}=\mathbf{b}$ باشند آنگاه داریم:

$$A(\mathbf{x}_2-\mathbf{x}_1)=A\mathbf{x}_2-A\mathbf{x}_1=\mathbf{b}-\mathbf{b}=0$$

یعنی تفاضل آنها $\mathbf{y}=\mathbf{x}_2-\mathbf{x}_1$ ریشهٔ دستگاه همگن $A\mathbf{x}=0$ است.

□ **مثال:** دستگاه خطی

$$x_1+2x_2=3$$

$$2x_1+4x_2=6$$

دارای جواب (ریشهٔ) $x_1=x_2=1$ است. دستگاه همگن مربوطه به‌صورت

$$x_1+2x_2=0$$

$$2x_1+4x_2=0$$

است که دارای جوابهای $x_1=-2a$ و $x_2=a$ است که در آن a عددی‌است دلخواه. بنابراین هر جواب دستگاه اولیه دارای شکل $x_1=1-2a$ و $x_2=1+a$ است که در آن a عددی‌است دلخواه. □

قضایای زیر از لم فوق نتیجه می‌شوند.

قضیهٔ ۱.۴ دستگاه خطی $A\mathbf{x}=\mathbf{b}$ حداکثر دارای یك جواب‌است (یعنی اگرجواب داشته باشد این جواب یکتاست) اگر،و فقط اگر، دستگاه همگن مربوطه $A\mathbf{x}=0$ فقط دارای جواب «بدیهی» $\mathbf{x}=0$ باشد.

بعد ثابت خواهیم کردکه نمی‌توان امیدوار بودکه دستگاه خطی فوق یك جواب

---

1. homogeneous

یکتا داشته باشد، مگر اینکه تعداد معادلات دستگاه خطی حداقل برابر با تعداد مجهولات آن باشد.

قضیهٔ ۲.۴ هر دستگاه خطی همگن با تعداد معادلاتی کمتر از تعداد مجهولات، دارای جوابهای نابدیهی (یعنی غیرصفر) است.

باید ثابت کنیم که اگر A یك ماتریس $m\times n$ با $m<n$ باشد آنگاه می‌توان $\mathbf{y}\neq \mathbf{0}$ را طوری پیدا کرد که $A\mathbf{y}=\mathbf{0}$. این عمل با استقرا نسبت، به n انجام می‌گیرد. اول حالت $n=2$ را بررسی می‌کنیم. دراین حالت می‌توانیم فقط یك معادله داشته باشیم

$$a_{11}x_1+a_{12}x_2=0$$

و این معادله در صورت $a_{12}=0$، دارای جواب نابدیهی $x_1=0$، $x_2=1$، و در غیر این‌صورت دارای جوابهای نابدیهی $x_1=a_{12}$، $x_2=-a_{11}$ می‌باشد. بدین‌صورت قضیه را برای $n=2$ اثبات کردیم. حال گیریم $n>2$، وفرض کنید ثابت شده است که هر دستگاه همگنی که تعداد معادلاتش کمتر از تعداد مجهولات و تعداد مجهولاتش کمتر از n باشد، دارای جوابهای نابدیهی است و بعلاوه، گیریم $A\mathbf{x}=\mathbf{0}$ یك دستگاه خطی همگن با n معادله و n مجهول، $m<n$ باشد. باید ثابت کنیم که این دستگاه دارای جوابهای نابدیهی است. اگر ستون n ام A صفر باشد، یعنی اگر $\mathbf{a}_n=\mathbf{0}$، این دستگاه قطعاً جوابهای نابدیهی دارد. زیرا دراین‌حالت n-برداری غیرصفر $\mathbf{x}=\mathbf{i}_n$ یك جواب دستگاه است. درغیراین‌صورت یکی ازعناصر $\mathbf{a}_n$ باید غیرصفر باشد، مثلا فرض کنیم

$$a_{in}\neq 0$$

دراین حالت، ماتریس B از مرتبهٔ $m\times(n-1)$ را، که ستون j ام آن به‌صورت

$$\mathbf{b}_j=\mathbf{a}_j-\frac{a_{ij}}{a_{in}}\mathbf{a}_n \qquad j=1,\cdots,n-1$$

است مورد بررسی قرارمی‌دهیم. اگر بتوانیم نشان دهیم که دستگاه همگن $B\mathbf{x}=\mathbf{0}$ دارای جوابهای نابدیهی است، اثبات انجام گرفته است. زیرا اگر بتوان اعداد x_1، x_2، ...، x_{n-1} راکه همگی صفر نیستند، به گونه‌ای یافت که رابطهٔ

$$x_1\mathbf{b}_1+x_2\mathbf{b}_2+\cdots+x_{n-1}\mathbf{b}_{n-1}=\mathbf{0}$$

برقرار باشد دراین صورت از تعریف بردارهای $\mathbf{b}_j$ نتیجه می‌شود که

$$x_1\mathbf{a}_1+x_2\mathbf{a}_2+\cdots+x_{n-1}\mathbf{a}_{n-1}+\left(-\sum_{j=1}^{n-1}x_j\frac{a_{ij}}{a_{in}}\right)\mathbf{a}_n=\mathbf{0}$$

رابطهٔ بالا یك جواب نابدیهی برای معادلهٔ $A\mathbf{x}=\mathbf{0}$ به‌دست می‌دهد. بنابراین تنها باید

نشان دادکه $B\mathbf{x}=\mathbf{0}$ دارای جوابهای نابدیهی‌است. برای اثبات این موضوع توجه‌کنید که به‌ازای هر j، درایهٔ iام بردار $\mathbf{b}_j$ برابراست با

$$a_{ij}-\frac{a_{ij}}{a_{in}}a_{in}=0$$

به‌طوری‌که معادلهٔ iام دستگاه $B\mathbf{x}=\mathbf{0}$ به‌صورت زیردرمی‌آید

$$0\times x_1+0\times x_2+\cdots+0\times x_{n-1}=0$$

و بنابراین به‌ازای هرانتخابی از $x_1,\ldots,x_{n-1}$ برقرار است. ازاینجا نتیجه می‌شودکه $\mathbf{x}=\mathbf{y}$، ریشهٔ $B\mathbf{x}=\mathbf{0}$ است اگر، و فقط اگر، $\mathbf{x}=\mathbf{y}$ ریشهٔ دستگاه همگن $\hat{B}\mathbf{x}=\mathbf{0}$ باشد، که از $B\mathbf{x}=\mathbf{0}$ فقط با صرفنظرکردن‌ازمعادلهٔ iام به‌دست می‌آید. اما اکنون $\hat{B}\mathbf{x}=\mathbf{0}$ یك دستگاه خطی همگنی‌است بـا $m-1$ معادله و $n-1$ مجهول یعنی بـا معادلاتی‌کمتر از مجهولات و مجهولاتی‌کمتر از n. بنابراین بـا فرض استقرا، $\hat{B}\mathbf{x}=\mathbf{0}$ دارای جوابهای نابدیهی‌است‌که این امر خاتمهٔ اثبات‌است.

□ **مثال:** دستگاه خطی همگن $A\mathbf{x}=\mathbf{0}$ راکه به‌صورت زیرداده شده، بررسی‌کنید

$$x_1+2x_2-x_3=0$$

$$x_1-x_2+x_3=0$$

برای این دستگاه $m=2$ و $n=3$، به‌دنبال برهـان قضیهٔ ۲.۴ به‌طریق زیر یك جواب نابدیهی می‌سازیم: از آنجاکه $a_{23}\neq 0$، $i=2$ را انتخاب می‌کنیم و داریم

$$\mathbf{b}_1=\mathbf{a}_1-\frac{a_{21}}{a_{23}}\mathbf{a}_3=\begin{bmatrix}1\\1\end{bmatrix}-(1/1)\begin{bmatrix}-1\\1\end{bmatrix}=\begin{bmatrix}2\\0\end{bmatrix}$$

$$\mathbf{b}_2=\mathbf{a}_2-\frac{a_{22}}{a_{23}}\mathbf{a}_3=\begin{bmatrix}2\\-1\end{bmatrix}-(-1/1)\begin{bmatrix}-1\\1\end{bmatrix}=\begin{bmatrix}1\\0\end{bmatrix}$$

بنابراین دستگاه همگن کوچکتر $B\mathbf{x}=\mathbf{0}$ عبارت‌است از

$$2x_1+x_2=0$$

$$0x_1+0x_2=0$$

می‌توانیم از معادلهٔ دوم صرفنظرکنیم، دراین‌صورت دستگاه همگن $\hat{B}\mathbf{x}=\mathbf{0}$ فقط شامل‌یك معادلهٔ زیرخواهد شد

$$۲x_۱ + x_۲ = ۰$$

یك جواب نابدیهی برای این معادله، $x_۱ = ۱$ و $x_۲ = -۲$ است. بنابراین با احتساب ۳-برداری

$$x_۳ = -[۱(۱/۱) + (-۲)(-۱/۱)] = -۳$$

$\mathbf{x} = (x_j)$ یك جواب نابدیهی ازدستگاه اولیه است. □

اکنون ثابت می‌کنیم که نمی‌توان انتظارداشت که به‌ازای هرمقدار دلخواه بردار سمت راست $\mathbf{b}$، این‌دستگاه یك جواب داشته باشد، مگر اینکه تعداد معادلات بیشتر ازتعداد مجهولات نباشد.

لم۲.۴ اگر A یك ماتریس $m \times n$ و دستگاه خطی $A\mathbf{x} = \mathbf{b}$ دارای یك جواب به‌ازای هر m-برداری $\mathbf{b}$ باشد، دراین‌صورت یك ماتریس $n \times m$ مانند C وجود دارد به‌طوری‌که

$$AC = I_m$$

چنین ماتریسی را می‌توان به‌صورت زیر تشکیل داد: طبق فرض قضیه، بدون‌توجه به چه بود $\mathbf{b}$ می‌توانیم یك جواب برای $A\mathbf{x} = \mathbf{b}$ پیدا کنیم. بنابراین اگر $\mathbf{b}$ را مساوی ستون jام I انتخاب کنیم، می‌توانیم یك n-برداری $\mathbf{c}_j$ طوری پیداکنیم که

$$A\mathbf{c}_j = \mathbf{i}_j \qquad j = ۱, \ldots, m$$

در این‌صورت با ماتریس C از مرتبهٔ $n \times m$ که ستون jام آن عبارت است از $\mathbf{c}_j$، $j = ۱, \ldots, m$

$$(AC)\mathbf{i}_j = A(C\mathbf{i}_j) = A\mathbf{c}_j = \mathbf{i}_j = I\mathbf{i}_j \qquad j = ۱, \ldots, m$$

رابطهٔ بالا نشان می‌دهدکه به‌ازای $j = ۱, \ldots, m$، ستون jام حاصلضرب AC، با ستون jام ماتریس I برابراست و این به‌معنی $AC = I$ است.

لم۳.۴ اگر B و C ماتریسهایی چنان باشندکه تساوی

$$BC = I$$

برقرار باشد، آنگاه دستگاه همگن $C\mathbf{x} = \mathbf{o}$ فقط دارای جواب بدیهی $\mathbf{x} = \mathbf{o}$ است. زیرا، اگر $C\mathbf{x} = \mathbf{o}$، آنگاه داریم:

$$\mathbf{x} = I\mathbf{x} = (BC)\mathbf{x} = B(C\mathbf{x}) = B\mathbf{o} = \mathbf{o}$$

قضیهٔ ۳.۴ اگر A یك ماتریس $m \times n$ ودستگاه خطی $A\mathbf{x} = \mathbf{b}$ دارای یك جواب به‌ازای هر m-برداری دلخواه $\mathbf{b}$ باشد، آنگاه $m \leqslant n$.

برای اثبات، از لم ۲.۴ نتیجه می‌گیریم که یك ماتریس C از مرتبهٔ $n \times m$ وجود دارد به‌طوری که

$$AC = I$$

اما از لم ۳.۴ نتیجه می‌شود که دستگاه همگن $C\mathbf{x} = \mathbf{0}$ فقط یك جواب بدیهی $\mathbf{x} = \mathbf{0}$ دارد. بنابراین طبق قضیهٔ ۲.۴، C باید حداقل به‌تعداد سطرها، ستون داشته باشد، یعنی $n \geqslant m$، که این امر پایان اثبات است.

اما می‌دانیم که، نمی‌توان انتظار داشت که به‌ازای هر بردار ممکن سمت راست دقیقاً یك جواب برای دستگاه (۱.۴) به‌دست آورد، مگر اینکه دستگاه دقیقاً به‌تعداد معادلات مجهول داشته باشد؛ یعنی، مگر اینکه ماتریس ضرایب، مربعی باشد. بنابراین ازاین به‌بعد فقط دستگاههای خطی با ماتریس ضرایب مربعی بررسی خواهند شد. برای چنین ماتریسهای مربعی یك قضیهٔ نهایی را ثابت خواهیم کرد.

قضیهٔ ۴.۴ گیریم A یك ماتریس $n \times n$ باشد. احکام زیر هم‌ارزند:

(i) دستگاه همگن $A\mathbf{x} = \mathbf{0}$ فقط یك جواب بدیهی $\mathbf{x} = \mathbf{0}$ دارد.

(ii) به‌ازای هر بردار سمت راست $\mathbf{b}$، دستگاه $A\mathbf{x} = \mathbf{b}$ یك جواب دارد.

(iii) A وارونپذیر است.

اول ثابت می‌کنیم که صحت (i) منجر به‌صحت (ii) می‌شود. گیریم $\mathbf{b}$ یك n-برداری داده شده باشد. باید ثابت کنیم که معادلهٔ $A\mathbf{x} = \mathbf{b}$ یك جواب دارد. برای این امر گیریم D یك ماتریس از مرتبهٔ $m \times (n+1)$ باشد که n ستون اول آن با ستونهای A مساوی و ستون $(n+1)$ام آن بردار $\mathbf{b}$ باشد. از آنجاکه تعداد ستونهای D بیش از تعداد سطرهای آن است، با استفاده از قضیهٔ ۲.۴ می‌توان یك $(n+1)$-برداری غیرصفر $\mathbf{y}$ طوری پیدا کرد که تساوی $D\mathbf{y} = \mathbf{0}$ برقرار باشد، یعنی

$$y_1\mathbf{a}_1 + y_2\mathbf{a}_2 + \ldots + y_n\mathbf{a}_n + y_{n+1}\mathbf{b} = \mathbf{0} \qquad (۱۲.۴)$$

واضح است که y_{n+1} نمی‌تواند صفر باشد. زیرا اگر y_{n+1} صفر باشد، آنگاه به‌ازای $\mathbf{y} \neq \mathbf{0}$ حداقل یکی از اعداد $y_1, \ldots, y_n$ باید غیرصفر باشد، و درعین‌حال رابطهٔ

$$y_1\mathbf{a}_1 + \ldots + y_n\mathbf{a}_n = \mathbf{0}$$

برقرار باشد. اما این رابطه بیان می‌کند که معادلهٔ $A\mathbf{x} = \mathbf{0}$ به‌ازای $i = 1, \ldots, n$، جواب نابدیهی $x_i = y_i$ را قبول می‌کند، که مغایر با (i) است. بنابراین چون $y_{n+1} \neq 0$، می‌توانیم (۱۲.۴) را نسبت به $\mathbf{b}$ حل کنیم، که خواهیم داشت:

$$-\frac{y_1}{y_{n+1}}\mathbf{a}_1 - \ldots - \frac{y_n}{y_{n+1}}\mathbf{a}_n = \mathbf{b}$$

که این رابطه مبین آن است که معادلهٔ $A\mathbf{x}=\mathbf{b}$ یك جواب $x_i=-(y_i/y_{n+1})$ به‌ازای $i=1,\ldots,n$، دارد که این امر (ii) را ثابت می‌کند.

در مرحلهٔ بعد ثابت می‌کنیم که صحت (ii) منجر به‌صحت (iii) می‌شود. با فرض درست بودن (ii)، از لم ۲.۴ نتیجه می‌گیریم که یك ماتریس C از مرتبهٔ $n\times n$ وجود دارد به‌طوری که رابطهٔ

$$AC=I$$

برقرار است. بنابراین با توجه به‌لم ۳.۴، معادلهٔ $C\mathbf{x}=\mathbf{0}$ تنها یك جواب بدیهی $\mathbf{x}=\mathbf{0}$ دارد. این نکته مبین این‌است که ماتریس C از مرتبهٔ $n\times n$، در (i) صدق می‌کند و بنابر برهان پیش، C در (ii) نیز صدق می‌کند. بنابراین، با استفاده از لم ۲.۴ یك ماتریس D از مرتبهٔ $n\times n$ وجود دارد به‌طوری که

$$CD=I$$

بدین‌ترتیب اثبات به‌انجام رسیده‌است. قبلا نشان داده شده بود که اگر رابطهٔ زیر برای ماتریسهای مربعی A و C و D برقرار باشد

$$AC=I=CD$$

آنگاه C وارونپذیر است و

$$A=D=C^{-1}$$

بدین‌ترتیب A وارون یك ماتریس وارونپذیر است و بنابراین خود نیز وارونپذیر است.

سرانجام لم ۳.۴، نشان می‌دهد که از (iii) صحت (i) نتیجه می‌شود.

□ **مثال:** درمثال قبلی نشان داده شد که ماتریس 2×2 ی

$$A=\begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix}$$

وارونپذیر نیست، و در مثال دیگری نشان دادیم که دستگاه همگن $A\mathbf{x}=\mathbf{0}$ دارای جواب نابدیهی است. بنابراین به‌موجب قضیهٔ ۴.۴ دستگاه خطی $A\mathbf{x}=\mathbf{b}$ بایستی به‌ازای یك ۲-برداری قابل‌حل نباشد. در واقع با $\mathbf{b}=\mathbf{i}_1$ دستگاه زیر به‌دست می‌آید

$$x_1+2x_2=1$$

$$2x_1+4x_2=0$$

این دستگاه جوابی ندارد، زیرا معادلهٔ دوم مستلزم $2(x_1+2x_2)=0$ است، درحالی‌که معادلهٔ اول مستلزم $x_1+2x_2=2$. □

اکنون به‌عنوان یك نمونهٔ ساده ازکاربرد قضیهٔ ۴.۴، ثابت می‌کنیم که با داشتن ماتریس مربعی A و صحت رابطهٔ $AB=I$، درستی $B=A^{-1}$ و $BA=I$ نتیجه می‌شود. در واقع

اگر A یك ماتریس از مرتبهٔ $n \times n$ باشد، آنگاه رابطهٔ $AB = I$ بیان می‌کند که B ازمرتبهٔ $n \times n$ است و به‌ازای هر n-برداری $\mathbf{b}$، داریم $A(B\mathbf{b}) = \mathbf{b}$. اما این تساوی بیانگر این است که بدون توجه به $\mathbf{b}$، می‌توان $A\mathbf{x} = \mathbf{b}$ را نسبت به $\mathbf{x}$ حل نمود و از این‌روطبق‌قضیهٔ ۴.۴، A وارونپذیراست و $\mathbf{x} = B\mathbf{b}$ یـك جواب آن. از این‌رو به‌ازای جمیـع مقادیر $\mathbf{b}$ خواهیم داشت $B\mathbf{b} = A^{-1}\mathbf{b}$ یا $B = A^{-1}$ و سرانجام $BA = I$.

استقلال خطی و پایه‌ها

گیریم $\mathbf{a}_1, \ldots, \mathbf{a}_n$ n تا m-برداری باشند و A یك ماتریس $m \times n$ کـه ستون jام آن $\mathbf{a}_j$، $j = 1, \ldots, n$، است. ایـن m-بـردارهـا را **مستقل خطی**[1] گوییم اگر از رابطـهٔ $x_1\mathbf{a}_1 + \ldots + x_n\mathbf{a}_n = 0$ روابط $x_1 = \ldots = x_n = 0$ نتیجه شوند. در غیر این صورت بردارها را وابستهٔ خطی گویند. واضح است که این n تا m-برداری، مستقل‌خطی‌هستند اگر، و فقط اگـر، دستگاه همگن $A\mathbf{x} = 0$ تنهـا دارای جواب بدیهی $\mathbf{x} = 0$ باشد. بنابراین می‌توان ازقضیهٔ ۲.۴ نتیجه‌گرفت که هرمجموعه‌ای‌که بیش‌از m تا m-برداری داشته باشد باید وابستهٔ خطی باشد.

گیریم $\mathbf{a}_1, \ldots, \mathbf{a}_n$ مستقل خطی باشند. اگر بتوان هر m-برداری $\mathbf{b}$ را بـه‌صورت یك ترکیب خطی ازاین n تـا m-برداری نوشت، آنگاه $\mathbf{a}_1, \ldots, \mathbf{a}_n$ را یك پایه (برای همهٔ m-برداریهـا) می‌نامند. واضح است که $\mathbf{a}_1, \ldots, \mathbf{a}_n$ یك پایه‌است اگر، و فقط اگر، دستگاه خطی $A\mathbf{x} = \mathbf{b}$ برای هـر m-برداری $\mathbf{b}$ فقط دارای یك جواب باشد، یعنی اگـر، و فقط اگـر، هر m-برداری را فقـط بـه‌یك طریق بتوان به‌صورت یك ترکیب خطی از m-برداریهـای $\mathbf{a}_1, \ldots, \mathbf{a}_n$ نوشت. به‌ویژه آنکه یك پایه (برای همهٔ m-برداریها) دقیقاً شامل m تا m-برداری می‌شود (یعنی $n = m$) وماتریس مربوطه وارونپذیر است.

□ **مثال:** بردارهای

$$\mathbf{a}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad \mathbf{a}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

به‌طورخطی مستقل‌اند؛ اما یك پایه تشکیل نمی‌دهند، زیراکه فقط دو تا ۳-برداری هستند. بعلاوه هر ۲-برداری را می‌توان به‌صورت یك ترکیب خطی از سه تا ۲-برداری نوشت:

$$\mathbf{a}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{a}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{a}_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

اما این سه تا ۲-برداری تشکیل یك پایه نمی‌دهند، زیرا می‌باید به‌طورخطی مستقل باشند.

1. linearly independent

بالاخره سه تا ۳-برداری

$$\mathbf{a}_1=\begin{bmatrix}1\\0\\0\end{bmatrix} \qquad \mathbf{a}_2=\begin{bmatrix}1\\1\\0\end{bmatrix} \qquad \mathbf{a}_3=\begin{bmatrix}1\\1\\1\end{bmatrix}$$

یك پایه تشکیل می‌دهند زیرا، ماتریس مربوطه وارونپذیر است. برای مشاهدهٔ این امر کافی است که بنا بر قضیهٔ ۴.۴ ثابت کنیم که دستگاه

$$\begin{aligned} x_1+x_2+x_3&=0\\ x_2+x_3&=0\\ x_3&=0 \end{aligned}$$

فقط دارای جواب بدیهی $x_1=x_2=x_3=0$ است، ولی این امر ازخود معادلات آشکار است. □

ماتریس ترانهاده

بالاخره، در ماتریسها عملی انجامپذیر است که هم‌ارز آن درحساب معمولی وجودندارد و آن تشکیل ماتریس ترانهاده[1] است. اگر $A=(a_{ij})$ و $B=(b_{ij})$ دو ماتریس باشند، B را **ترانهادهٔ** A گویند و با $B=A^T$ نشان می‌دهند، به‌شرط آنکه تعداد سطرهای B برابر تعداد ستونهای A و تعداد ستونهای آن برابر تعداد سطرهای A باشد و تساوی $b_{ij}=a_{ji}$ به‌ازای جمیع مقادیر i و j درست باشد. به‌طور لفظی برای به‌دست آوردن A^T، ترانهادهٔ A، «قرینهٔ آن را نسبت به‌قطر» آن به‌دست می‌آوریم. اگر رابطهٔ

$$A^T=A$$

برقرار باشد، A را **متقارن**[2] نامند

ماتریسهای

$$A=\begin{bmatrix}1&3&2\\3&0&4\\2&4&5\end{bmatrix} \qquad B=\begin{bmatrix}3&-1\\2&6\\0&8\end{bmatrix} \qquad C=[3\quad 4\quad 7]$$

دارای ترانهاده‌های زیرند.

---

1. transposed matrix 2. symmetric

$$A^T=\begin{bmatrix}۱ & ۳ & ۲\\ ۳ & ۰ & ۴\\ ۲ & ۴ & ۵\end{bmatrix} \qquad B^T=\begin{bmatrix}۳ & ۲ & ۰\\ -۱ & ۶ & ۸\end{bmatrix} \qquad C^T=\begin{bmatrix}۳\\ ۴\\ ۷\end{bmatrix}$$

به‌ویژه ترانهادهٔ $\mathbf{b}^T$ی یك بردار ستونی $\mathbf{b}$، یك بردار سطری‌است.

به‌آسانی می‌توان در رابطه با ترانهش، صحت قواعد زیر را اثبات کرد:

۱. اگر A و B دوماتریس چنان باشندکه AB تعریف شده‌باشد، آنگاه B^TA^T نیز تعریف شده‌است و داریم:

$$(AB)^T=B^TA^T$$

به‌تغییر در ترتیب آنها توجه‌کنید!

۲. به‌ازای هرماتریس A داریم $(A^T)^T=A$

۳. اگر ماتریسA وارونپذیر باشد، A^T نیز وارونپذیر است و داریم

$$(A^T)^{-۱}=(A^{-۱})^T$$

برای اثبات قاعدهٔ (۱)،گوییم اگر A یك ماتریس $m\times n$ و B یك ماتریس$n\times p$ باشد به‌طوری‌که AB یك ماتریس $m\times p$ و $(AB)^T$ یك ماتریس $p\times m$ باشند، آنگاه A^T از مرتبهٔ $n\times m$ و B^T از مرتبهٔ $p\times n$ است، وبنابراین حاصلضرب B^TA^T ماتریسی است به‌خوبی تعریف شده و از مرتبهٔ $p\times m$ است. سرانجام

درایهٔ (i,j)ام در $(AB)^T$ = درایهٔ (j,i)ام در AB

$$=\sum_{k=۱}^{n}\text{[درایهٔ }(k,i)\text{ام در }B\text{]}\times\text{[درایهٔ }(j,k)\text{ام در }A\text{]}$$

$$=\sum_{k=۱}^{n}\text{[درایهٔ }(k,j)\text{ام در }A^T\text{]}\times\text{[درایهٔ }(i,k)\text{ام در }B^T\text{]}$$

$$=\text{درایهٔ }(i,j)\text{ام در }B^TA^T$$

برای اثبات قاعدهٔ (۳) از قاعدهٔ (۱) نتیجه می‌گیریم که

$$A^T(A^{-۱})^T=(A^{-۱}A)^T=I^T=I$$

$$(A^{-۱})^TA^T=(AA^{-۱})^T=I^T=I$$

که قاعدهٔ (۳) را ثابت می‌کند.

اگر $\mathbf{a}$ و $\mathbf{b}$ دو n-برداری باشند، آنگاه $\mathbf{b}^T\mathbf{a}$ یك ماتریس ۱×۱ یا یك‌عدد است، که آن را **حاصلضرب عددی** $\mathbf{a}$ و $\mathbf{b}$ گویند درحالتی‌که $\mathbf{a}$ و $\mathbf{b}$ دوبردار حقیقی باشند.

درباره ماتریسهایی که دارای درایه‌های همتافت (که در بحث مربوط به ویژه مقدارها حائز اهمیت است) هستند، مفهوم ترانهادهٔ مزدوج یا هرمیتی[1] ماتریس A، A^H، پیش می‌آید. برای بیان این مطلب اشاره می‌کنیم که مزدوج یك عدد همتافت z، یعنی $\bar{z}$، از تغییر علامت بخش انگاری آن، به‌دست می‌آید. اگر $z \neq 0$، آنگاه $\bar{z}$ عدد یگانه‌ای است مانند α به‌طوری که $\alpha z = |z|^2$. تعیین هرمیتی A یعنی A^H، درست‌شمه به‌تعیین ترانهادهٔ A^T است، جز آنکه به‌جای همهٔ درایه‌های A^T، مزدوجهای همتافت آنها قرارداده می‌شوند. بنابراین درحالت

$$b_{ij} = \bar{a}_{ji} \qquad j, i \text{ به‌ازای جمیع مقادیر}$$

رابطهٔ $A^H = (b_{ij})$ برقرار است.

از این‌رو در حالتی که A یك ماتریس حقیقی* باشد، خواهیم داشت $A^H = A^T$. توجه کنید که برای n-بردارهای $\mathbf{a}$ و $\mathbf{b}$ که دارای درایه‌های همتافت هستند، حاصلضرب عددی معمولی برابر با عدد $\mathbf{b}^H\mathbf{a}$ است نه $\mathbf{b}^T\mathbf{a}$، زیرا عدد $\mathbf{a}^H\mathbf{a}$ است که مربع طول بردار $\mathbf{a}$ را به‌دست می‌دهد.

جایگشتها و ماتریسهای جایگشتی

یك **جایگشت**[2] از **درجهٔ** n، بازآرایشی است از اولین n عدد صحیح، یعنی یك توالی از n عدد صحیح که در آن هر عدد صحیح بین ۱ تا n حداقل یکبار، حداکثر نیز یکبار و بنابراین دقیقاً یکبار در توالی مزبور ظاهر شود. یك جایگشت از درجهٔ n را می‌توان به‌شکلهای مختلفی نوشت. برای اهداف ما کافی است (به‌تعبیری کاملا دقیق) که یك جایگشت، به‌عنوان یك n-برداری $\mathbf{p} = (p_i)$ با $p_i \in \{1, 2, \ldots, n\}$ به‌ازای جمیع مقادیر i و j به ازای $i \neq j$، $p_i \neq p_j$، نگریسته شود. دراینجا تعداد $n!$ جایگشت از درجهٔ n وجود دارد. بسته به‌اینکه تعداد **برگردانها**[3]ی $\mathbf{p}$ زوج یا فرد باشد، جایگشت $\mathbf{p}$ را زوج یا فرد نامند. در اینجا منظور از برگردانهای یك جایگشت $\mathbf{p} = (p_i)$، تعداد دفعاتی است که یك عدد صحیح بزرگتر پیش از یك عدد کوچکتر آمده باشد. برای مثال در جایگشت $\mathbf{p}$ با $\mathbf{p}^T = [5, 1, 4, 3, 6, 2, 7]$

۷ پیش از ۲، ۶، ۳، ۴، ۱، ۵	آمده که	۶ برگردانی به‌دست می‌دهد
۲ پیش از ۱	آمده که	۱ برگردانی به‌دست می‌دهد
۶ پیش از ۳، ۴، ۱، ۵	آمده که	۴ برگردانی به‌دست می‌دهد
۳ پیش از ۱	آمده که	۱ برگردانی به‌دست می‌دهد
۴ پیش از ۱	آمده که	۱ برگردانی به‌دست می‌دهد
بنابراین $\mathbf{p}$ جمعاً		۱۳ برگردانی دارد

1. Hermitian

* توضیح مترجم: منظور آن است که درایه‌های ماتریس A حقیقی باشند.

2. permutation 3. inversions

باید توجه داشت که هر تعویض دو درایه با هم در یک جایگشت، تعداد برگردانها را به‌تعداد فردی تغییر می‌دهد.

یک ماتریس جایگشتی از مرتبهٔ n، یک ماتریس P از مرتبهٔ $n\times n$ است که ستونها (سطرها)ی آن یک بازآرایش یا جایگشتی از ستونها (سطرها)ی ماتریس واحد از مرتبهٔ n باشد. دقیقاً بگوییم، ماتریس P از مرتبهٔ $n\times n$ یک ماتریس جایگشتی است اگر به‌ازای یک جایگشت $\mathbf{p}=(p_i)$ از درجهٔ n داشته باشیم

$$P\mathbf{i}_j=\mathbf{i}_{p_j} \qquad j=1,\dots,n \qquad (۱۳.۴)$$

قضیهٔ ۵.۴ اگر P یک ماتریس جایگشتی باشد که در (۱۳.۴) صدق کند، آنگاه

(i) P^T یک ماتریس جایگشتی است که به‌ازای $j=1,\dots,n$، در تساوی:

$$P^T\mathbf{i}_{p_j}=\mathbf{i}_j$$

صدق می‌کند. از این‌رو $P^TP=I$، و بنابراین وارونپذیر است و داریم $P^{-1}=P^T$.

(ii) اگر A یک ماتریس $m\times n$ باشد، آنگاه AP یک ماتریس $m\times n$ است که ستون jام آن، به‌ازای $j=1,\dots,n$، مساوی است با ستون p_jام ماتریس A.

(iii) اگر A یک ماتریس $n\times m$ باشد، آنگاه P^TA ماتریسی است $n\times m$ که سطر iام آن، به‌ازای $i=1,\dots,n$، مساوی است با سطر p_iام ماتریس A.

□ **مثال:** ماتریس

$$P=\begin{bmatrix}0&0&1\\1&0&0\\0&1&0\end{bmatrix}$$

یک ماتریس جایگشتی متناظر با جایگشت $\mathbf{p}^T=[2\ \ 3\ \ 1]$ است، زیرا داریم

$$P\mathbf{i}_3=\mathbf{i}_1,\ P\mathbf{i}_2=\mathbf{i}_3,\ P\mathbf{i}_1=\mathbf{i}_2$$

بعلاوه

$$P^T=\begin{bmatrix}0&1&0\\0&0&1\\1&0&0\end{bmatrix}$$

از این‌رو تساویهای $P^T\mathbf{i}_1=\mathbf{i}_3$، $P^T\mathbf{i}_2=\mathbf{i}_1$ و $P^T\mathbf{i}_3=\mathbf{i}_2$ برقرارند که قسمت (i) قضیهٔ (۵.۴)را نشان می‌دهد. بعلاوه می توان محاسبه کرد که مثلا

$$AP=\begin{bmatrix}۱ & ۲ & ۳\\ ۴ & ۵ & ۶\\ ۷ & ۸ & ۹\end{bmatrix}\begin{bmatrix}۰ & ۰ & ۱\\ ۱ & ۰ & ۰\\ ۰ & ۱ & ۰\end{bmatrix}=\begin{bmatrix}۲ & ۳ & ۱\\ ۰ & ۶ & ۴\\ ۸ & ۹ & ۷\end{bmatrix}$$

بنابراین ستون دوم ماتریس AP مساوی ستون سوم ماتریس A، یعنی $P_۲=۳$ است که قسمت (ii) قضیهٔ (۵.۴) را نشان می‌دهد. □

حل عددی دستگاه معادلات خطی

در اینجا فقط دستگاههای خطی

$$A\mathbf{x}=\mathbf{b}$$

را بررسی خواهیم کرد که به‌ازای هرمقدار سمت راست، $\mathbf{b}$، یك و فقط یك جواب دارند. بنابر قضایای ۲.۴ و ۳.۴ باید خود را به‌دستگاههایی که تعداد مجهولات آنها دقیقاً با تعداد معادلات برابرند، یعنی به‌دستگاههایی که ماتریس ضرایبشان ماتریس مربعی است، محدود سازیم. برای این گونه دستگاهها، حکم قضیهٔ ۴.۴ این است که شرط اینکه دستگاه به‌ازای جمیع مقادیر سمت راست، $\mathbf{b}$، فقط یك جواب داشته باشد، A می‌باید وارونپذیر باشد. بنابراین فرض می‌کنیم که همهٔ دستگاههای خطی مورد بحث ماتریس ضرایبی وارونپذیر دارند.

آزمونی که غالباً برای وارونپذیری به‌کار برده می‌شود مبتنی برمفهوم **دترمینان** است. قضیهٔ مربوطه چنین بیان می‌کند که ماتریس A وارونپذیر است اگر، و فقط اگر، $\det(A)\neq ۰$. اگر $\det(A)\neq ۰$، حتی ممکن است جواب دستگاه $A\mathbf{x}=\mathbf{b}$ برحسب دترمینانها از راه به‌اصطلاح قاعدهٔ کرامر[1]، داده شود. ولی دترمینانها درحل دستگاههای خطی ارزش عملی ندارند، زیرا درحالت کلی محاسبهٔ یك دترمینان، خود به‌اندازهٔ حل یك دستگاه خطی دشواری دارد. به‌همین دلیل در اینجا برای حل دستگاههای خطی از دترمینان استفاده نمی‌شود و برای تعریف دترمینان نیز کوششی به‌عمل نمی‌آید. با این حال در بخش ۷.۴ برای ارزیابی دترمینانها (بر اساس روش مستقیم حل دستگاههای خطی) برای استفاده در موارد دیگر، روشی ارائه خواهد شد.

روشهای عددی برای حل دستگاههای خطی را ممکن است به‌دو نوع، **مستقیم** و **بارستی** تقسیم کرد. روشهای مستقیم آنهایی هستند که درصورت نبودن خطای گردکردن و خطاهای دیگر، جواب دقیق را با تعداد متناهی ازعملیات مقدماتی حسابی به‌دست می‌دهند. درعمل، چون یك کامپیوتر با طول کلمهٔ متناهی سروکار دارد، روشهای مستقیم معمولا جوابهای دقیق نمی‌دهند. زیرا خطاهای ناشی از گردکردن، ناپایداری، و از دست رفتن ارقام بامعنی ممکن

1. Cramer

است به‌نتایج خیلی نامطلوب و یا حتی بی‌معنی منجر شود. سروکار قسمت اعظم آنالیزعددی با چرا و چگونه پدید آمدن این خطاها و بررسی روشهایی است که این خطاها را برروی هم حداقل می‌سازند. روش اساسی که برای حل مستقیم به‌کار گرفته می‌شود روش حذف گاوس[1] است، ولی حتی در این دسته (روش مستقیم) نیزمی‌توان روشهای متفاوتی انتخاب کرد که کارایی محاسباتی و دقت آنها متفاوت است. برخی ازاین روشها در بخشهای بعدی بررسی خواهند شد.

روشهای بارستی روشهایی هستند که با یك تقریب اولیه شروع می‌شوند و با به‌کار بستن یك الگوریتم انتخابی مناسب، متوالیاً به‌تقریبهای بهتری دست می‌یابند. حتی اگراین روند همگرا باشد، می‌توان تنها امیدوار بود که از راه روشهای بارستی به یك حل تقریبی رسید. روشهای بارستی از لحاظ الگوریتمهای انتخابی و میزان همگرایی، متفاوت‌اند. برخی از روشهای بارستی ممکن است عملا واگرا باشند و برخی دیگرممکن است آنقدر آهسته همگرا شوند که کارایی محاسباتی نداشته باشند. امتیاز مهم روشهای بارستی در سادگی آنها و یکنواختی عملیاتی است که باید انجام شوند، و درنتیجه موجب می‌شوند برای کاربرد کامپیوتری بسیار مناسب باشند ودرقبال افزایش خطاهای گردکردن عدم‌حساسیت نسبی داشته باشد.

ماتریسهای مربوط به‌دستگاههای خطی نیزبه‌دو ردهٔ **چگال**[2] و **تنك**[3] تقسیم می‌شوند. در ماتریسهای چگال تعداد بسیار کمی از درایه‌ها صفرند و مرتبهٔ این‌گونه ماتریسها نسبتاً کم است ــ شاید از مرتبهٔ ۱۰۰ یا کمتر. معمولا کاراتر این است که مسائل مشتمل براین نوع ماتریسها با روش مستقیم حل شوند. ماتریسهای تنك معمولا تعداد خیلی کمی درایه‌های غیر صفر دارند. این‌گونه ماتریسها معمولا در حل معادلات دیفرانسیل با روش تفاضلات متناهی به‌دست می‌آیند. مرتبهٔ این‌گونه ماتریسها ممکن است بسیار بزرگ باشد و معمولا درحل با روشهای بارستی بسیار مناسب‌اند و این امتیازی است در قبال ماهیت تنك بودن ماتریسهای موردنظر. روشهای بارستی برای دستگاههای خطی درفصل ۵ مورد بحث قرار خواهند گرفت.

تمرین

۱-۱۰۴ گیریم

$$A=\begin{bmatrix} ۲ & ۳ & ۱ \\ ۱ & -۱ & ۱ \\ ۰ & ۲ & ۲ \end{bmatrix} \qquad B=\begin{bmatrix} ۱ & ۲ & ۱ \\ -۱ & ۲ & -۱ \\ ۲ & ۰ & ۲ \end{bmatrix} \qquad C=\begin{bmatrix} ۱ & ۰ & ۱ \\ ۰ & ۱ & ۱ \\ ۱ & ۰ & ۲ \end{bmatrix}$$

(الف) حاصلضرب AB و BA را محاسبه کنید و نشان دهید که $AB \neq BA$.

---

1. Gauss elimination 2. dense 3. sparse

(ب) $(A+B)+C$ و $A+(B+C)$ را پیدا کنید.
(پ) نشان دهید که $A(BC)=(AB)C$.
(ت) تحقیق کنید که $(AB)^T=B^TA^T$.

۲-۱۰۴ نشان دهید که ماتریس A که در زیر داده شده، وارونپذیر نیست (به‌قضیهٔ ۴.۴ نگاه کنید):

$$A=\begin{bmatrix}۱ & ۲ & ۱\\ ۲ & -۱ & -۱\\ ۶ & ۲ & ۰\end{bmatrix}$$

۳-۱۰۴ برای ماتریس A که در زیر داده شده‌است، ماتریس جایگشتی P را طوری پیدا کنید که

(الف) ضرب A از سمت چپ[1] در P، ستونهای اول و چهارم A را با هم عوض کند.

(ب) ضرب A از سمت راست[2] در P، سطرهای اول وسوم A را با هم‌عوض کند.

$$A=\begin{bmatrix}۴ & ۱ & ۲ & ۱\\ ۳ & ۲ & ۱ & ۱\\ ۱ & ۲ & ۰ & ۱\\ ۱ & ۱ & ۰ & ۱\end{bmatrix}$$

۴-۱۰۴ در ماتریس Aی تمرین ۱۰۴-۳ یك دنباله از ماتریسهای جایگشتی پیدا کنید که A را به‌شکل زیردر آورد

$$A'=\begin{bmatrix}۱ & ۰ & ۱ & ۱\\ ۱ & ۰ & ۱ & ۲\\ ۱ & ۱ & ۳ & ۲\\ ۱ & ۲ & ۴ & ۱\end{bmatrix}$$

۵-۱۰۴ دستگاه زیر را به‌شکل ماتریسی بنویسید وماتریس A و بردار $\mathbf{b}$ را مشخص کنید

$$\begin{aligned}۲x_۱+۳x_۲+۴x_۳+x_۴&=۱\\ x_۱+۲x_۲\qquad\quad+x_۴&=۰\end{aligned}$$

---

1. Postmultiplication 2. Premultiplication

$$۲x_۱+۳x_۲+x_۳-x_۴=۲$$

$$x_۱-۲x_۲-x_۳-x_۴=۳$$

۶-۱۰۴ با اثبات عبارت زیر خود را متقاعد سازید که مفهوم وارونپذیری تنها برای ماتریسهای مربعی بامعنی است: گیریم A یك ماتریس $m\times n$ باشد. اگر B و C ماتریسهای $n\times m$ باشند به‌طوری که $AB=I_m$ و $CA=I_n$، آنگاه $B=C=A^{-۱}$. به‌ویژه اینکه در این صورت $m=n$. [راهنمایی: اول ثابت کنید $B=C$، سپس نشان دهید که $m=\text{trace}\,(AB)=\text{trace}\,(BA)=n$. که **trace** (اثر) یك ماتریس به‌صورت حاصلجمع درایه‌های قطری آن تعریف شده‌است].

۷-۱۰۴ با استفاده از قضیهٔ ۴.۴ ثابت کنید که ماتریس جایگشتی وارونپذیر است.

۸-۱۰۴ با استفاده از قضیهٔ ۴.۴ ثابت کنید که اگر A و B ماتریسهای مربعی باشند به‌طوری که حاصلضرب آنها وارونپذیر باشد، آنگاه A و B باید هردو وارونپذیر باشند.

۹-۱۰۴ آیا بردارهای زیر تشکیل یك پایه می‌دهند؟

$$\begin{bmatrix}۰\\-۱\\۰\\۰\end{bmatrix}\quad\begin{bmatrix}۰\\۱\\-۱\\۰\end{bmatrix}\quad\begin{bmatrix}۰\\۰\\۱\\-۱\end{bmatrix}\quad\begin{bmatrix}۱\\۰\\۰\\۱\end{bmatrix}$$

۱۰-۱۰۴ ثابت کنید که سه بردار

$$\begin{bmatrix}۱\\۲\\۰\end{bmatrix}\quad\begin{bmatrix}۲\\۱\\۰\end{bmatrix}\quad\begin{bmatrix}۰\\۰\\-۱\end{bmatrix}$$

تشکیل یك مجموعهٔ مستقل خطی می‌دهند. آیا یك پایه تشکیل می‌دهند؟

۱۱-۱۰۴ برای هریك از سه عمل ماتریسی، یعنی جمع دوماتریس، ضرب دوماتریس و ضرب یك اسکالر دریك ماتریس، یك زیربرنامهٔ **فورترن** بنویسید که با ورودی مناسب محاسبه را انجام دهد و ماتریس حاصل را برگرداند.

۱۲-۱۰۴ اگر $p(x)=c_۰+c_۱x+c_۲x^۲+\cdots+c_kx^k$ یك بسجمله‌ای باشد و A یك ماتریس داده شده $n\times n$، آنگاه ماتریس $p(A)$ چنین تعریف می‌شود

$$p(A)=c_۰A^۰+c_۱A^۱+c_۲A^۲+\cdots+c_kA^k$$

دراینجا $A^{1}=A$، $A^{0}=I$ و به‌ازای $j>1$، $A^{j}=A(A^{j-1})$. یك زیر برنامهٔ **فورترن** کارآمد با شناسه‌های N، KPI، A، C، PA بنویسید، که در آن N مرتبه ماتریس A است و PA درهنگام بازگشت از زیر برنامه، باید ماتریس $p(A)$ را در برداشته باشد و C یك آرایهٔ یك‌بعدی است که شامل $C(i)=c_{i-1}$، $i=1,\ldots,KPI$ است. در این زیر برنامه بجز آرایه‌های A، C، و PA هیچ آرایهٔ دیگری را به‌کار نگیرید. (راهنمایی: الگوریتم ۱.۲ را به‌یاد آورید.)

۱۳-۱.۴ فرض کنید برای یك ماتریس A از مرتبهٔ n یك بسجمله‌ای $p(x)$، به‌گونه‌ای وجود دارد که $p(0)\neq 0$، در عین حال $p(A)$ یك ماتریس صفر است. ثابت کنید که A باید وارونپذیر باشد.

۱۴-۱.۴ قواعد مذکور در (۱۱.۴) را بررسی کنید.

۱۵-۱.۴ طبق تعریف، **ماتریس واندرموند**[1] برای نقاط $x_0,\ldots,x_n$ ماتریسی است از مرتبهٔ $n+1$ که با $V=(x_i^j)_{i,j=0}^{n}$ بیان می‌گردد. ماتریس فوق در برخی از عملیات مربوط به‌درونیابی بسجمله‌ای نقش مهمی دارد، زیرا در دستگاه خطی

$$\sum_{j=0}^{n} a_j x_i^j = f(x_i) \qquad i=0,\ldots,n$$

برای ضرایب توانی بسجمله‌ای درونیاب، ماتریس ضرایب است. بسجمله‌ایهای لاگرانژ (۶.۲) را برای پیدا کردن وارون V درحالت $n=3$ به‌کار گیرید. رابطهٔ بین شکل توانی بسجمله‌ای لاگرانژ برای $x_0,\ldots,x_n$ و درایه‌های وارون V چیست؟

۲.۴ حل دستگاههای خطی از راه حذف

گیریم A یك ماتریس مربعی مفروض از مرتبهٔ n، و $\mathbf{b}$ یك n-برداری مفروض باشد. می‌خواهیم دستگاه خطی زیر را نسبت به n-برداری مجهول $\mathbf{x}$ حل کنیم

$$A\mathbf{x}=\mathbf{b} \qquad (۱۴.۴)$$

در موردی که A یك ماتریس بالامثلثی با همهٔ درایه‌های قطری غیرصفر باشد. بردار جواب $\mathbf{x}$ را به‌آسانی می‌توان به‌دست آورد. زیرا دراین‌صورت دستگاه (۱۴.۴) شکل زیر را دارد

$$\begin{aligned}
a_{11}x_1+a_{12}x_2+\cdots+a_{1,n-1}x_{n-1}\quad &+a_{1n}x_n=b_1\\
a_{22}x_2+\cdots+a_{2,n-1}x_{n-1}\quad &+a_{2n}x_n=b_1\\
&\cdots\cdots\cdots\cdots\\
a_{n-1,n-1}x_{n-1}&+a_{n-1,n}x_n=b_{n-1}\\
a_{nn}x_n&=b_n
\end{aligned} \qquad (۱۵.۴)$$

---

1. Vandermonde matrix

به‌ویژه اینکه معادلهٔ آخر فقط شامل x_n است. بنابراین از آنجاکه $a_{nn} \neq 0$، باید داشته باشیم

$$x_n = \frac{b_n}{a_{nn}}$$

و چون حالا x_n در دست‌است معادلهٔ ماقبل آخر یعنی

$$a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1}$$

تنها یك مجهول داردکه x_{n-1} است. چون $a_{n-1,n-1} \neq 0$، خواهیم داشت

$$x_{n-1} = \frac{b_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}}$$

با در دست داشتن x_n و x_{n-1} اکنون از معادلهٔ سوم از آخر، یعنی

$$a_{n-2,n-2}x_{n-2} + a_{n-2,n-1}x_{n-1} + a_{n-2,n}x_n = b_{n-2}$$

استفاده می‌کنیم که تنها یك مجهول واقعی x_{n-2} دارد. باز چون $a_{n-2,n-2} \neq 0$ معادلهٔ فوق را می‌توان نسبت به x_{n-2} حل کرد

$$x_{n-2} = \frac{b_{n-2} - a_{n-2,n-1}x_{n-1} - a_{n-2,n}x_n}{a_{n-2,n-2}}$$

درحالت کلی که مقادیر x_{k+1}، x_{k+2}، ...، x_n ازقبل‌محاسبه شده‌اند و باتوجه به $a_{kk} \neq 0$، معادلهٔ kام را می‌توان به‌تنهایی نسبت به x_k حل کرد، که در این‌صورت خواهیم داشت

$$x_k = \frac{b_k - \sum_{j=k+1}^{n} a_{kj}x_j}{a_{kk}}$$

این روند تعیین جواب دستگاه (۱۵.۴) را روند پسجایگذاری[1] می‌نامند.

الگوریتم ۱.۴: پسجایگذاری ماتریس بالا مثلثی A ازمرتبهٔ $n \times n$ که تمامی درایه‌های قطری آن‌غیرصفرند و n-برداری $\mathbf{b}$ داده‌شده‌اند. درایه‌های x_n، x_{n-1}، ...، x_1 متعلق به‌جواب $\mathbf{x}$ از دستگاه $A\mathbf{x} = \mathbf{b}$ را می‌توان (به‌همین‌ترتیب) به‌وسیلهٔ

For $k = n, n-1, \ldots, 1$, do

$$x_k := \frac{b_k - \sum_{j=k+1}^{n} a_{kj}x_j}{a_{kk}}$$

1. back-substitution

به‌دست آورد.

در اینجا دونکته قابل‌توجه‌است: وقتی $k=n$، دراین‌صورت حاصلجمع $\sum_{j=k+1}^{n}$ به‌صورت $\sum_{j=n+1}^{n}$ در می‌آید و به‌صورت حاصلجمع روی هیچ عضوی تعبیرمی‌شود و طبق قراردادمقدار آن صفر است. همچنین‌متذکرمی‌شویم که‌با توجه به‌توضیح پسجایگذاری، تقریباً نتیجهٔ زیر واضح‌است.

قضیهٔ ۶.۴ ماتریس بالامثلثی A وارونپذیر است اگر و فقط اگر همهٔ درایه‌های‌قطری آن مخالف صفر باشند.

در واقع روش پسجایگذاری نشان می‌دهدکه وقتی‌تمام درایه‌های قطری A غیرصفر باشند، دستگاه خطی $A\mathbf{x}=\mathbf{b}$ به‌ازای بردار مفروض $\mathbf{b}$، حداکثریك جواب دارد؛ازاین‌رو بنابر قضیهٔ ۴.۴، A می‌باید وارونپذیر باشد. از سوی دیگر به‌ازای هر $j=1,\dots,n$، مقادیری مانند $x_1,\dots,x_n$ که همه برابرصفرنیستند، وجود دارند به‌طوری‌که بنابرقضیهٔ ۲.۴ داریم

$$a_{11}x_1+\dots+a_{1j}x_j=0$$

$$\cdots\cdots\cdots$$

$$a_{j-1,1}x_1+\dots+a_{j-1,j}x_j=0$$

اما اگر $a_{jj}=0$، بردار $\mathbf{y}=[x_1\dots x_j\,0\dots0]^T$، بردارصفر نیست، ولی در دستگاه $A\mathbf{y}=0$ صدق می‌کندکه با توجه به‌قضیهٔ ۴.۴، نتیجه می‌شودکه A وارونپذیر نیست.

بنابراین ما محق بوده‌ایم‌که بردار $\mathbf{x}$ راکه با الگوریتم ۱.۴ محاسبه شده، جواب دستگاه (۱۵.۴) بنامیم.

□ **مثال۱.۴:** دستگاه خطی زیر را درنظر می‌گیریم

$$\begin{aligned}2x_1+3x_2-x_3&=5\\-2x_2-x_3&=-7\\-5x_3&=-15\end{aligned}\qquad(16.4)$$

از معادلهٔ آخرنتیجه می‌گیریم $x_3=b_3/a_{33}=15/5=3$، با این‌جواب، ازمعادلهٔ دوم چنین به‌دست می‌آوریم

$$x_2=(b_2-a_{23}x_3)/a_{22}=(-7+3)/(-2)=2$$

لذا با توجه به‌معادلهٔ اول خواهیم داشت

$$x_1=(b_1-a_{12}x_2-a_{13}x_3)/a_{11}=(5-3\times2+3)/2=1$$ □

اما اگر ماتریس ضرایب دستگاه $A\mathbf{x}=\mathbf{b}$ بالامثلثی نباشد، نخست روش حذف گاوس را براین دستگاه اعمال می‌کنیم. احتمالا دانشجویان درجبرمقدماتی با این روش آشنایی پیداکرده‌اند. هدف از این روش تبدیل دستگاه مفروض به‌دستگاه هم‌ارزی است که ماتریس ضرایب آن هم‌ارز یك ماتریس بالامثلثی باشد. دراین صورت دستگاه اخیر را می‌توان از راه پسجایگذاری حل کرد.

دو دستگاه خطی $A\mathbf{x}=\mathbf{b}$ و $\tilde{A}\mathbf{x}=\tilde{\mathbf{b}}$ را **هم‌ارز** نامیم هرگاه هرجواب یك دستگاه یك جواب دیگری نیز باشد.

قضیهٔ ۷.۴ گیریم دستگاه خطی $A\mathbf{x}=\mathbf{b}$ داده شده باشد، و فرض می‌کنیم که این‌دستگاه را تحت سلسله عملیاتی از نوع زیرقراردهیم.

(i) ضرب یك معادله دریك عدد ثابت مخالف صفر
(ii) جمع مضربی از یك معادله با معادلهٔ دیگر
(iii) تعویض جای دو معادله با هم

اگر این سلسله عملیات دستگاه جدید $\tilde{A}\mathbf{x}=\tilde{\mathbf{b}}$ را پدید آورد، آنگاه دستگاههای $A\mathbf{x}=\mathbf{b}$ و $\tilde{A}\mathbf{x}=\tilde{\mathbf{b}}$ هم‌ارزند. به‌ویژه دراین صورت، A وارونپذیر است اگر وفقط اگر، $\tilde{A}$ وارونپذیر باشد.

برای دیدن یك برهان از این قضیه، به‌مسئلهٔ ۲.۴ـ۱۱ مراجعه کنید.

عمل حذف، برپایهٔ قضیهٔ فوق و ملاحظات زیر انجام می‌گیرد: اگر $A\mathbf{x}=\mathbf{b}$ یك دستگاه خطی باشد و اگر به‌ازای مقداری از k و j داشته باشیم $a_{kj}\neq 0$، آنگاه می‌توان مجهول x_j را ازیك‌معادلهٔ $i\neq k$ باافزودن $-(a_{ij}/a_{kj})$ برابر معادلهٔ kام به معادلهٔ iام، حذف کنیم. دستگاه حاصل $\tilde{A}\mathbf{x}=\tilde{\mathbf{b}}$ با دستگاه اصلی هم‌ارز است.

روش حذف گاوس، درساده‌ترین شکل خود، از یك دستگاه خطی مفروض $A\mathbf{x}=\mathbf{b}$ از مرتبهٔ n، به‌ازای $k=0,\ldots,n-1$، یك سلسله از دستگاههای هم‌ارز $A^{(k)}\mathbf{x}=\mathbf{b}^{(k)}$ را پدید می‌آورد. در اینجا $A^{(0)}\mathbf{x}=\mathbf{b}^{(0)}$ درست همان دستگاه اصلی است. دستگاه $(k-1)$ام به‌شکل زیراست:

$$
\begin{aligned}
a_{11}^{(k-1)}x_1+a_{12}^{(k-1)}x_2+\cdots+a_{1,k-1}^{(k-1)}x_{k-1}+a_{1k}^{(k-1)}x_k+\cdots+a_{1n}^{(k-1)}x_n&=b_1^{(k-1)}\\
a_{22}^{(k-1)}x_2+\cdots+a_{2,k-1}^{(k-1)}x_{k-1}+a_{2k}^{(k-1)}x_k+\cdots+a_{2n}^{(k-1)}x_n&=b_2^{(k-1)}\\
&\cdots\\
a_{k-1,k-1}^{(k-1)}x_{k-1}+a_{k-1,k}^{(k-1)}x_k+\cdots+a_{k-1,n}^{(k-1)}x_n&=b_{k-1}^{(k-1)}\\
a_{kk}^{(k-1)}x_k+\cdots+a_{kn}^{(k-1)}x_n&=b_k^{(k-1)}\\
&\cdots\\
a_{nk}^{(k-1)}x_k+\cdots+a_{nn}^{(k-1)}x_n&=b_n^{(k-1)}
\end{aligned}
$$

به‌بیان لفظی، k معادلهٔ اول هم‌اکنون به‌شکل بالامثلثی درآمده‌اند، درحالی‌که $n-k$ معادلهٔ

آخر تنهـا شامل مجهولات x_k، ... ، x_n هستند. از این‌رو دستگاه kام $A^{(k)}\mathbf{x}=\mathbf{b}^{(k)}$ در طی مرحلهٔ kام روش حذف گاوس، به‌طریق زیر به‌دست آمده است: k معادلهٔ اول بدون تغییر مـانده‌اند، بعـلاوه اگـر ضریب x_k، یعنی $a_{kk}^{(k-1)}$، در معادلهٔ kام صفر نباشد، آنگاه $m_{ik}=a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ برابر معادلهٔ kام از معادلهٔ iام کم شده و از این طریق مجهول x_k از معـادلهٔ iام، $i=k+1, \ldots, n$، حـذف شده‌است. روشن است کـه دستگاه حاصل $A^{(k)}\mathbf{x}=\mathbf{b}^{(k)}$ هم‌ارز دستگاه $A^{(k-1)}\mathbf{x}=\mathbf{b}^{(k-1)}$ است، لذا، بـا توجه به‌استقرا، هم‌ارز دستگاه اصلی‌است. بعلاوه اولین $k+1$ معادلهٔ دستگاه kام به‌شکل بالامثلثی‌است.

بعد از $n-1$ مرحله از این عملیات، دستگاه $A^{(n-1)}\mathbf{x}=\mathbf{b}^{(n-1)}$ به‌دست‌آمده‌است که ماتریس ضرایب آن به‌صورت بالامثلثی‌است و بنابراین می‌توان این دستگاه را سریعاً از راه پسجایگذاری حل کرد.

□ **مثال ۲۰۴:** دستگاه خطی زیر را درنظر می‌گیریم

$$\begin{aligned} 2x_1+3x_2-x_3&=5 && \text{(الف)}\\ 4x_1+4x_2-3x_3&=3 && \text{(ب)} \quad (17.4)\\ -2x_1+3x_2-x_3&=1 && \text{(پ)} \end{aligned}$$

برای حذف x_1 از معادلات (ب) و (پ)، $-4/2=-2$ برابر معادلـهٔ (الف) را بـا معادلهٔ (ب) جمع می‌کنیم تا معادلهٔ جدید زیر به‌دست آید

$$0x_1-2x_2-x_3=-7$$

همچنین $-(-2)/2=1$ برابر معادلهٔ (الف) را با معادلهٔ (پ) جمع می‌کنیم، درنتیجه

$$0x_1+6x_2-2x_3=6$$

که این معادله دستگاه جدید $A^{(1)}\mathbf{x}=\mathbf{b}^{(1)}$ را به‌صورت زیر به‌دست می‌دهد

$$\begin{aligned} 2x_1+3x_2-x_3&=5 && \text{(الف)}\\ -2x_2-x_3&=-7 && \text{(ب)} \quad (18.4)\\ 6x_2-2x_3&=6 && \text{(پ)} \end{aligned}$$

این عملیات، مرحلهٔ اول روش حذف گاوس را برای این دستگاه کامل می‌کند. در مرحلهٔ دوم (وآخرین مرحله برای‌این‌مثال)، x_2 را ازمعادلهٔ(پ) باجمع کردن $-6/(-2)=3$ برابر معادلهٔ (ب) با معادلهٔ (پ) حذف می‌کنیم، داریم

$$0x_2-5x_3=-15$$

بنابراین دستگاه جدید و نهایی

$$\begin{aligned} 2x_1+3x_2-x_3&=5\\ -2x_2-x_3&=-7\\ -5x_3&=-15 \end{aligned} \qquad (۱۹.۴)$$

به‌دست می‌آید که بنابر قضیهٔ ۷.۴ با دستگاه اولیهٔ (۱۷.۴) هم‌ارز است، اما ماتریس ضرایب آن به‌صورت ماتریس بالامثلثی است، از این‌رو می‌توان آن را مانند مثال ۱.۴ سریعاً با استفاده از پسجایگذاری حل کرد. □

در تشریح سادهٔ روش حذف گاوس که در بالا ذکر شد، ما از معادلهٔ kام برای حذف x_k از معادلات $k+1$ام، ...، nام، طی مرحلهٔ kام عملیات، استفاده کردیم. البته این کار تنها درصورتی امکانپذیر است که درشروع مرحلهٔ kام، ضریب x_k یعنی $a_{kk}^{(k-1)}$ در معادلهٔ kام صفرنباشد. متأسفانه به‌آسانی می‌توان دستگاهی خطی طرح کرد که در آن این شرط برقرار نباشد. برای مثال، اگر دستگاه $A\mathbf{x}=\mathbf{b}$ با معادلات

$$\begin{aligned} &x_2+x_3=1 && \text{(الف)}\\ &x_1\quad\ +x_3=1 && \text{(ب)}\\ &x_1+x_2\quad\ =1 && \text{(پ)} \end{aligned} \qquad (۲۰.۴)$$

داده شده باشد، آنگاه غیرممکن‌است بتوان برای حذف x_1 از معادله‌های دیگر، از معادلهٔ (الف) استفاده کرد. به‌منظور غلبه براین مشکل به‌طوری‌که باز یك دستگاه بالامثلثی‌هم‌ارز با دستگاه مفروض نتیجه شود، باید درهرمرحله در انتخاب **معادلهٔ لولایی**[1]، یعنی‌معادله‌ای که برای حذف یك مجهول از یکی از معادلات اختیار می‌شود، آزادی بیشتری داشت.

مثلا در دستگاه (۲۰.۴) می‌توانستیم از معادلهٔ (ب) در مرحلهٔ اول حذف به‌عنوان معادلهٔ لولایی‌استفاده کنیم. برای حفظ شکل‌قبلی‌عملیات، نخست معادلهٔ (ب) را با عوض‌کردن با معادلهٔ (الف) درجای اول قرار می‌دهیم. اکنون دراین‌ترتیب جدید دستگاه ضریب x_1 در معادلهٔ (الف) غیرصفر بوده و مشابه با عملیات قبلی می‌توان پیش رفت تا دستگاه جدید $A^{(1)}\mathbf{x}=\mathbf{b}^{(1)}$ به‌فرم زیربه‌دست آید:

$$\begin{aligned} &x_1\quad\ +x_3=1 && \text{(الف)}\\ &x_2+x_3=1 && \text{(ب)}\\ &x_2-x_3=0 && \text{(پ)} \end{aligned}$$

با این دستگاه، مرحلهٔ دوم (و آخرین مرحله) از روش حذف گاوس، بدون هیچ مشکلی پیش می‌رود و دو دستگاه نهایی بالامثلثی به‌گونهٔ زیرحاصل می‌شود

1. pivotal equation

$$x_1 + x_3 = 1$$
$$x_2 + x_3 = 1$$
$$-2x_3 = -1$$

که با پسجایگذاری جوابهای $x_3 = x_2 = x_1 = 1/2$ بهدست می‌آیند.

این آزادی عمل بیشتر در انتخاب معادلات لولایی نه فقط بهدلیل امکان صفر بودن ضرایب ضروری است، بلکه تجربه نشان دادهاست که این امر برای مقابله با نتایج خطای گردکردن نیز اساسی‌است (بـه‌بخش ۳.۴ نگاه‌کنید). کاراضافی‌که باید انجام‌گیرد خیلی ناچیزاست: درشروع مرحلهٔ kام، بررسی‌درمورد ضریب غیرصفر x_k در معادله‌های kام، $(k+1)$ام، ...، nام انجام می‌گیـرد، واگرچنین ضریبی در حداقل یك معادلـه به‌ازای $j > k$ بهدست آمد، جای معادله‌های jام و kام با هم عوض می‌شود.

در حالتی که ماتریس A وارونپذیر باشد، لزوماً چنین ضریب غیرصفری می‌باید وجود داشته باشد. درغیر این صورت دستگاه خطی بهدست‌آمده شامل $(n-k+1)$ تا معادلهٔ

$$0 \times x_k + a_{i,k+1}^{(k-1)} x_{k+1} + \cdots + a_{in}^{(k-1)} x_n = b_i^{(k-1)} \qquad i = k, \ldots, n \quad (21.4)$$

می‌شودکـه در واقـع دارای $n-k$ مجهول به‌صورت x_{k+1}، ...، x_n است. لـذا بنابر قضیهٔ ۳.۴، دستگاه معادلات (۲۱.۴) به‌ازای بعضی مقادیر سمت راست قـابل حل نیستند؛ ازاین‌روکل دستگاه به‌ازای بعضی ازمقادیرسمت راست قابل حل نخواهد بود و بنابراین به‌موجب قضیهٔ ۴.۴ ماتریس ضرایب دستگـاه معادلات فعلی وارونپذیر نیست. امـا چون دستگاه معادلات فعلی با دستگاه معادلات اولیه هم‌ارز یعنی $A\mathbf{x} = \mathbf{b}$ است، نتیجه می‌شود که ماتریس A وارونپذیر نیست و این همان اثبات ادعای ماست.

وقتی‌که این اعمال به‌کمك‌کامپیوترانجام می‌شود، n معادلهٔ اصلی و تغییرات‌مختلفی که درآنها انجام می‌گیرند، باید به‌طریق مناسب ومنظمی ثبت شوند. بدین‌منظور معمولاً از یك‌آرایهٔ عملی[1] یـا یك ماتریس $n \times (n+1)$ کـه آن‌را W می‌نامند و در ابتدا شامل ضرایب و مقادیر سمت راست n معادلهٔ $A\mathbf{x} = \mathbf{b}$ است استفاده می‌کنند. هرگاه مجهولی‌از یك معادله حذف شود، ضرایب تغییر یافته و سمت راست مربوط به‌این معادله محاسبه، و درآرایهٔ عملی W به‌جای ضرایب و مقادیرسمت راست قبلی ذخیره می‌شوند. به‌دلایلی‌که در زیر روشن خواهد شد، مضرب $m_{ik} = a_{ik}^{(k-1)} / a_{kk}^{(k-1)}$ (که درحذف x_k از معادلهٔ iام به‌کار می‌رود) به‌جای‌عدد $a_{ik}^{(k)}$ درمحل w_{ik} ذخیره می‌شود، زیـرا به‌هرحال عدد $a_{ik}^{(k)}$ صفر می‌شود (یا آنکه فرض آن‌است‌که صفرهست). تعویض سطرها نیز با استفاده ازیك تعداد صحیح آرایهٔ $\mathbf{p}$، ثبت می‌گردد.

---

1. working array

الگوریتم ۲.۴: روش حذف‌گاوس. ماتریس W از مرتبهٔ $n(n+1)$ که n اولین ستون آن متضمن ماتریس مرتبهٔ nام A و ستون آخرش، n-برداری $\mathbf{b}$ است، داده شده‌است.

n-برداری $\mathbf{p}$ را با $p_i = i$، به‌ازای $i = ۱, \dots, n$، پر کنید

For $k = ۱, \dots, n-۱$, do:

کوچکترین مقدار $i \geqslant k$ را چنان پیدا کنید که $w_{ik} \neq ۰$
اگر چنین مقدار i به‌دست نیامد، علامت بدهید که A وارونپذیر نیست و عملیات را متوقف سازید.
در غیر این صورت، محتوای p_k و p_i را با یکدیگر و سطرهای kام و iام را با هم در W عوض نموده، عملیات را ادامه دهید

For $i = k+۱, \dots, n$, do:

$m := w_{ik} := w_{ik}/w_{kk}$

For $j = k+۱, \dots, n+۱$, do:

$w_{ij} := w_{ij} - m w_{kj}$

اگر $w_{nn} = ۰$، علامت دهید که A وارونپذیر نیست و عملیات را متوقف سازید.

درغیر این‌صورت، دستگاه اولیهٔ $A\mathbf{x} = \mathbf{b}$ هم ارز است با دستگاه $U\mathbf{x} = \mathbf{y}$ که در آن U و $\mathbf{y}$ برحسب درایه‌های نهایی W به‌گونهٔ زیر داده شده‌اند

$$u_{ij} = \begin{cases} w_{ij} & i \leqslant j \\ ۰ & i > j \end{cases}, \qquad y_i = w_{i,\,n+۱}, \qquad i = ۱, \dots, n \tag{۲۲.۴}$$

به‌ویژه U یك ماتریس بالامثلثی است که کلیهٔ درایه‌های قطری آن غیرصفرند. پس حالا الگوریتم ۱.۴، را می‌توان برای محاسبهٔ جواب $\mathbf{x}$ به‌کار گرفت.

باتوجه به‌جنبه‌های خاص، مانند تقارن یا تنکی ماتریس ضرایب A، غالباً ممکن‌است عملیات ضروری محاسباتی برای حل $A\mathbf{x} = \mathbf{b}$ راکاهش داد. به‌عنوان مثال، اکنون حل دستگاههای سه‌قطری[1] را مختصراً مورد بحث قرار می‌دهیم.

ماتریس مرتبهٔ nام $A = (a_{ij})$ را **سه قطری** گوییم اگر وقتی که $|i-j| > ۱$، آنگاه تساوی $a_{ij} = ۰$ برقرار باشد.

---

1. tridiagonal

به بیان لفظی، A زمانی سه قطری است، که تنها درایه‌های قطری آن یعنی a_{ii}، $(i=1,\dots,n)$ یادرایه‌های زیرقطری[1] آن یعنی $a_{i,i-1}$، به‌ازای $(i=2,\dots,n)$ ویا درایه‌های زبرقطری[2] آن یعنی $a_{i,i+1}$، $(i=1,\dots,n-1)$ غیر صفر باشند. بنابراین ماتریسهای زیر سه‌قطری هستند

$$\begin{bmatrix} ۳ & ۱ & ۰ & ۰ \\ ۱ & ۳ & ۱ & ۰ \\ ۰ & ۲ & ۲ & ۱ \\ ۰ & ۰ & ۱ & ۶ \end{bmatrix} \quad \begin{bmatrix} ۱ & ۰ & ۰ & ۰ \\ ۰ & ۱ & ۰ & ۰ \\ ۰ & ۰ & ۱ & ۰ \\ ۰ & ۰ & ۱ & ۰ \end{bmatrix} \quad \begin{bmatrix} ۴ & ۷ & ۰ & ۰ \\ ۰ & ۸ & ۱ & ۰ \\ ۰ & ۱ & ۱ & ۰ \\ ۰ & ۰ & ۰ & ۱ \end{bmatrix}$$

فرض کنید که ماتریس ضرایب A از دستگاه خطی $\mathbf{Ax}=\mathbf{b}$ سه قطری باشد، بعلاوه فرض می‌کنیم که به‌ازای هر k بتوان معادلهٔ kام را به‌عنوان معادلهٔ لولایی در مرحلهٔ kام به‌کار گرفت. سپس طی مرحلهٔ kام از الگوریتم ۲.۴، فقط متغیر x_k می‌باید از معادلهٔ $(k+۱)$ام، $k=۱,\dots,n-۱$، حذف شود. بعلاوه طی پسجایگذاری، تنها لازم‌است که $x_{k+۱}$ در معادلهٔ kام قرار داده شود تا بتوان x_k را، $k=۱,\dots,n-۱$، پیدا کرد. سرانجام، احتیاجی به‌ذخیره‌کردن درایه‌های صفرماتریس A نیست، بلکه فقط لازم‌است سه بردار متضمن درایه‌های زیرقطری، قطری، و زبرقطری ماتریس A نگهداری شوند.

اکنون دستگاه سه قطری مرتبهٔ nام زیر را به‌طور دقیقتری بررسی می‌کنیم:

$$\begin{aligned}
d_۱x_۱+c_۱x_۲ &= b_۱ \\
a_۲x_۱+d_۲x_۲+c_۲x_۳ &= b_۲ \\
a_۳x_۲+d_۳x_۳+c_۳x_۴ &= b_۳ \\
&\ddots \\
a_{n-۱}x_{n-۲}+d_{n-۱}x_{n-۱}+c_{n-۱}x_n &= b_{n-۱} \\
a_nx_{n-۱}+d_nx_n &= b_n
\end{aligned}$$

با فرض $d_۱\neq ۰$، $x_۱$ را از معادلهٔ دوم حذف می‌کنیم، معادلهٔ دوم جدید به صورت زیر خواهد شد

$$d'_۲x_۲+c_۲x_۳=b'_۲$$

که در آن

---

1. subdiagonal 2. superdiagonal

$$d'_۲ = d_۲ - \frac{a_۲}{d_۱} c_۱ \qquad b'_۲ = b_۲ - \frac{a_۲}{d_۱} b_۱$$

سپس فرض کنید $d'_۲ \neq ۰$، معادلهٔ فوق را برای حذف $x_۲$ از معادلهٔ سوم به‌کار می‌بریم، معادلهٔ سوم جدید به‌صورت زیر خواهد شد

$$d'_۳ x_۳ + c_۳ x_۴ = b'_۳$$

که در آن

$$d'_۳ = d_۳ - \frac{a_۳}{d'_۲} c_۲ \qquad b'_۳ = b_۳ - \frac{a_۳}{d'_۲} b'_۲$$

به‌همین طریق عمل را ادامه می‌دهیم و طی مرحلهٔ kام، x_k را از معادلهٔ $k+۱$ام حذف می‌کنیم (با فرض اینکه $d'_k \neq ۰$) و معادلهٔ kام جدید به‌صورت زیر خواهد شد

$$d'_{k+۱} x_{k+۱} + c_{k+۱} x_{k+۲} = b'_{k+۱}$$

که در آن به‌ازای $k = ۱, ۲, \dots, n-۱$ داریم

$$d'_{k+۱} = d_{k+۱} - \frac{a_{k+۱}}{d'_k} c_k \qquad b'_{k+۱} = b_{k+۱} - \frac{a_{k+۱}}{d'_k} b'_k$$

هنگام پسجایگذاری، با فرض اینکه $d'_n \neq ۰$، در ابتدا خواهیم داشت

$$x_n = \frac{b'_n}{d'_n}$$

و سپس به‌ازای $k = n-۱, \dots, ۱$ داریم

$$x_k = \frac{b'_k - c_k x_{k+۱}}{d'_k}$$

الگوریتم ۳.۴ روش حذف برای دستگاههای سه قطری. ضرایب a_i، d_i و c_i و قسمت سمت راست b_i از دستگاه سه‌قطری زیر داده شده‌اند

$$a_i x_{i-۱} + d_i x_i + c_i x_{i+۱} = b_i \qquad i = ۱, \dots, n \ (a_۱ = c_n = ۰ \text{ باشد})$$

For $k = ۲, \dots, n$, do:

اگر $d_{k-۱} = ۰$، ناکامی را اعلام کنید و عملیات را متوقف سازید

در غیر این‌صورت، به‌ازای $m := \frac{a_k}{d_{k-۱}}$ عملیات را ادامه دهید

$$\begin{bmatrix} d_k := d_k - m * c_{k-1} \\ b_k := b_k - m * b_{k-1} \end{bmatrix}$$

اگر $d_n = 0$، ناکامی را اعلام کنید و عملیات را متوقف سازید

در غیراین صورت به‌ازای $x_n := \frac{b_n}{d_n}$ عملیات را ادامه دهید

$$\text{For } k = n-1, \cdots, 1, \text{ do:}$$
$$\begin{bmatrix} x_k := \dfrac{b_k - c_k * x_{k+1}}{d_k} \end{bmatrix}$$

□ **مثال ۳.۴:** دستگاه خطی زیر را به‌ازای $n = 10$ حل کنید

$$\begin{aligned} 2x_1 - x_2 &= 1 \\ -x_{i-1} + 2x_i - x_{i+1} &= 0 \qquad i = 2, \cdots, n-1 \\ -x_{n-1} + 2x_n &= 0 \end{aligned}$$

این مسئله با برنامهٔ **فورترن** زیر حل می‌شود. توجه کنید که الگوریتم ۳.۴ به‌صورت زیر برنامهٔ زیر ترجمه شده‌است

```
TRID(SUB, DIAG, SUP, B, N)
```

در این زیربرنامه، SUB، DIAG، SUP، و B، N-بردارهایی هستند که انتظار می‌رود ضرایب و قسمت سمت راست دستگاه سه‌قطری

$$\text{SUB}(i)x_{i-1} + \text{DIAG}(i)x_i + \text{SUP}(i)x_{i+1} = \text{B}(i) \qquad i = 1, \cdots, \text{N}$$

را داشته باشند [که در اینجا از SUB(۱) و SUB(N) صرفنظر شده‌است]. این زیربرنامه محتوای DIAG را تغییر می‌دهد و بردار جواب را در B ذخیره کرده، برگشت می‌دهد. جواب دقیق دستگاه به‌صورت زیر داده شده‌است

$$x_i = \frac{n+1-i}{n+1} \qquad i = 1, \cdots, n$$

بنابراین جوابهای محاسبه شده در رقم ششم پس از ممیز، متضمن خطا هستند. این برنامه روی یک کامپیوتر IBM ۳۶۰ اجرا شده‌است.

```
C  FORTRAN PROGRAM FOR EXAMPLE 4.3
      PARAMETER N=10
      INTEGER I
      REAL A(N),B(N),C(N),D(N)
      DO 10 I=1,N
         A(I) = -1.
         D(I) = 2.
         C(I) = -1.
   10    B(I) = 0.
      B(1) = 1.
      CALL TRID ( A, D, C, B, N )
      PRINT 610, (I,B(I),I=1,N)
  610 FORMAT('1THE SOLUTION IS '/(I5,E15.7))
                                          STOP
      END
      SUBROUTINE TRID ( SUB, DIAG, SUP, B, N )
      INTEGER N,   I
      REAL B(N),DIAG(N),SUB(N),SUP(N)
C  THE TRIDIAGONAL LINEAR SYSTEM
C     SUB(I)*X(I-1) + DIAG(I)*X(I) + SUP(I)*X(I+1) = B(I), I=1,...,N
C  (WITH SUB(1) AND SUP(N) TAKEN TO BE ZERO) IS SOLVED BY FACTORIZATION
C  AND SUBSTITUTION. THE FACTORIZATION IS RETURNED IN  SUB , DIAG , SUP
C  AND THE SOLUTION IS RETURNED IN  B .
      IF (N .LE. 1) THEN
         B(1) = B(1)/DIAG(1)
                                          RETURN
      END IF
      DO 11 I=2,N
         SUB(I) = SUB(I)/DIAG(I-1)
         DIAG(I) = DIAG(I) - SUB(I)*SUP(I-1)
   11    B(I) = B(I) - SUB(I)*B(I-1)
      B(N) = B(N)/DIAG(N)
      DO 12 I=N-1,1,-1
   12    B(I) = (B(I) - SUP(I)*B(I+1))/DIAG(I)
                                          RETURN
      END
```

خروجی

جواب برابراست با

```
 1  0.9090915E  00
 2  0.8181832E  00
 3  0.7272751E  00
 4  0.6363666E  00
 5  0.5454577E  00
 6  0.4545485E  00
 7  0.3636391E  00
 8  0.2727295E  00
 9  0.1818197E  00
10  0.9090990E -01
```

□

تمرین

۱-۲۰۴ یك ملاك كارایی یك الگوریتم، تعداد عملیات حسابی لازم برای به‌دست‌آوردن جواب‌است. نشان دهیدكه اگرالگوریتم ۲۰۴ برای یك دستگاه از مرتبهٔ nام به‌كار رود، $n(n-1)/2$ عمل تقسیم، $(n^3-n)/3$ عمل ضرب، و $(n^3-n)/3$ عمل‌جمع لازم‌دارد.

۲-۲۰۴ نشان دهیدكه الگوریتم پسجایگذاری ۱۰۴، مستلزم n عمل تقسیم، $n(n-1)/2$ عمل ضرب، و $n(n-1)/2$ عمل‌جمع است.

۳-۲۰۴ در برخی ازكامپیوترها عمل تقسیم وقتگیرتر از عمل ضرب است، چگونه ممكن

است الگوریتم ۲.۴، برای چنین کامپیوتری اصلاح شود؟

۴-۲.۴ تعداد عملیات جمع وضرب لازم برای ضرب یك ماتریس $n \times n$ در یك n-برداری را محاسبه کنید.

۵-۲.۴ چند عمل جمع، ضرب و تقسیم برای الگوریتم ۲.۴ لازم است، اگر فقط ماتریس نهایی بالامثلثی U موردنیاز باشد؟

۶-۲.۴ با استفاده از روش حذف نشان دهید که دستگاه زیر جواب ندارد

$$x_1 + 2x_2 + x_3 = 3$$

$$2x_1 + 3x_2 + x_3 = 5$$

$$3x_1 + 5x_2 + 2x_3 = 1$$

۷-۲.۴ زمان اجرای یك برنامهٔ متضمن الگوریتم ۲.۴، تا حد زیادی با زمان مصروفه در داخلی‌ترین حلقه معین می‌شود. بدین‌دلیل لازم است که تا حد ممکن حلقهٔ مذکور کارا شود. درعین‌حال در زبان **فورترن**، ذخیره‌کردن آرایه‌ها درحافظه به‌صورت ستونی انجام می‌گیرد و در بسیاری از ماشینها کارکردن با یك آرایه به‌صورت ستون به ستون سریعتر از سطر به‌سطر انجام می‌پذیرد.

بدین دلیل الگوریتم ۲.۴ را چنان بازسازی کنید که داخلی‌ترین حلقه(ها) بر مبنای نمایهٔ سطرها اجرا شود (شوند)، یعنی بازسازی به‌گونه‌ای باشد که در هر زمان به‌عوض سطر، ستون تغییر کند.

۸-۲.۴ دستگاه زیررا با روش حذف حل کنید. تمام محاسبات‌را تا سه رقم اعشاری گرد کنید.

$$0.21x_1 + 0.32x_2 + 0.12x_3 + 0.31x_4 = 0.96$$

$$0.10x_1 + 0.15x_2 + 0.24x_3 + 0.22x_4 = 0.71$$

$$0.20x_1 + 0.24x_2 + 0.46x_3 + 0.36x_4 = 1.26$$

$$0.61x_1 + 0.40x_2 + 0.32x_3 + 0.20x_4 = 1.53$$

جوابهای خود را با پسجایگذاری دردستگاه اولیه امتحان کنید ومیزان دقت عمل‌را براورد نمایید. جواب‌دقیق: $[1, 1, 1, 1]$ است.

۹-۲.۴ با استفاده از زیر برنامهٔ TRID و به‌ازای $n = 30$ و $h = 0.1$ دستگاه خطی زیر را حل کنید

$$-۲(۱+h^۲)x_۱+x_۲=۱$$
$$x_{i-۱}-۲(۱+h^۲)x_i+x_{i+۱}=۰ \qquad i=۲,\ ۳,\ \ldots,\ n-۱$$
$$x_{n-۱}-۲(۱+h^۲)x_n=۱$$

۱۰-۲.۴ با استفاده از قضیهٔ ۶.۴ و فرع لم ۱.۲، ثابت‌کنیدکه هر بسجمله‌ای از درجهٔ نابزرگتر از n به‌ازای مراکز داده شدهٔ $c_۱$، ...، c_n، تنها به‌یك روش می‌تواند به‌شکل نیوتنی نوشته شود.

(راهنمایی: دستگاه خطی مربوط به‌ضرایب آن شکل نیوتنی از بسجمله‌ای راکه با تابع داده شده در مراکز $c_۱$، ...، c_n، $c_{n+۱}$ یکسان‌است درنظر بگیرید.)

۱۱-۲.۴ قضیهٔ ۷.۴ را ثابت کنید. (راهنمایی: نخست ثابت‌کنیدکه هر جواب معادلهٔ $A\mathbf{x}=\mathbf{b}$، یك جواب $\tilde{A}\mathbf{x}=\tilde{\mathbf{b}}$ نیزهست. سپس نشان دهیدکه هرعمل از نوع مذکور در قضیه، می‌تواند به‌توسط عملی با همان نوع خنثی شود، لذا نشان دهیدکه با یك سلسله از عملیات یاد شده، $A\mathbf{x}=\mathbf{b}$ می‌تواند از $\tilde{A}\mathbf{x}=\tilde{\mathbf{b}}$ به‌دست آید.)

۳.۴ تدبیر لولاگزینی

در الگوریتم حذف۲.۴، که در بخش قبل ارائه شد، اگر تمامی محاسبات با دقت بسیار زیاد[1] انجام شود هردستگاه $A\mathbf{x}=\mathbf{b}$ باکارایی و با اطمینان حل می‌شود. درصورتی‌که، طبق معمول، محاسبات با دقت محدود انجام پذیرند، به‌آسانی می‌توان مثالهایی ارائه داد که برای آنها الگوریتم۲.۴، جوابهایی‌کاملاً غلط به‌دست می‌دهد.

دراین بخش فقط یك منشأ ممکن برای چنین خطایی به‌اختصار ارائه می‌شود و آن تدبیر نادرست لولاگزینی‌است. در اینجا، منظور ما از **تدبیر لولاگزینی**[2] طرح‌کلی‌است برای‌گزینش معادلهٔ لولایی (و شاید هم ستون لولایی) درهرمرحله ازحذف.

□ **مثال۴.۴:** دستگاه

$$۰٫۰۰۰۳x_۱+۱٫۵۶۶x_۲=۱٫۵۶۹$$
$$۰٫۳۴۵۴x_۱-۲٫۴۳۶x_۲=۱٫۰۱۸$$

دارای جوابهای $x_۱=۱۰$ و $x_۲=۱$ است. برای حل این‌دستگاه از روش حذف و محاسبات ممیز شناور با ۴ رقم اعشاری استفاده می‌کنیم و اولین معادله را به‌عنوان معادلهٔ لولایی برای مرحلهٔ اول (وتنها این مرحله) می‌گزینیم. داریم

$$m_{۲۱}=۰٫۳۴۵۴/۰٫۰۰۰۳=۱۱۵۱٫۰$$

---

1. infinite-precision 2. pivoting strategy

بنابراین

$$a_{۲۲}^{(۱)} = -۲٫۴۳۶ - (۱۱۵۱٫۰)(۱٫۵۶۶)$$
$$= -۲٫۴۳۶ - ۱۸۰۲٫۰ = -۱۸۰۴٫۰$$
$$b_{۲}^{(۱)} = ۱٫۰۱۸ - (۱۱۵۱٫۰)(۱٫۵۶۹) = ۱٫۰۱۸ - ۱۸۰۶٫۰$$
$$= -۱۸۰۵٫۰$$

این مقادیر به‌ما می‌دهند

$$x_۲ = -۱۸۰۵٫۰ / -۱۸۰۴٫۰ = ۱٫۰۰۱$$

بنابراین از اولین معادله نتیجه می‌شود که

$$x_۱ = [۱٫۵۶۹ - (۱٫۵۶۶)(۱٫۰۰۱)] / ۰٫۰۰۰۳ = ۳٫۳۳۳ \qquad \square$$

یك توضیح «موجه‌نما»[1] برای این نقص چنین‌است: درایهٔ لولای $a_{۱۱} = ۰٫۰۰۰۳$ «خیلی کوچك»است و از آنجا که اگر $a_{۱۱}$ صفر می بود محاسبات با شکست مواجه می‌شد، لذا شگفت‌آور نیست که در محیطی با دقت عمل محدود، این الگوریتم «درنزدیکی صفر» به‌طریق نامطلوبی عمل کند.

البته در توضیح فوق اصطلاحات تعریف نشده‌ای مانند «خیلی کوچك» و «نزدیك صفر» به‌کار برده شده‌اند که آن را از درجهٔ اعتبار می‌اندازند. درحقیقت اگر طرفین معادلهٔ اول را در توان مناسبی از ده ضرب کنیم، می توانیم $a_{۱۱}$ را هرقدر که بخواهیم بزرگ کنیم بی‌آنکه تغییری در جواب محاسبه شده ایجاد شود. برای مشاهدهٔ این امر، دوباره دستگاه مثال ۴.۴ را درنظر می گیریم که در آن معادلهٔ اول در ۱۰^m، که m عددی صحیح است، ضرب شده‌است

$$۰٫۰۰۰۳ \times ۱۰^m x_۱ + ۱٫۵۶۶ \times ۱۰^m x_۲ = ۱٫۵۶۹ \times ۱۰^m$$
$$۰٫۳۴۵۴\ x_۱ - ۲٫۴۳۶ \qquad x_۲ = ۱٫۰۱۸$$

دوباره از معادلهٔ اول به‌عنوان معادلهٔ لولایی استفاده می کنیم و محاسبات ممیزشناور با ۴رقم اعشاری را به‌کار می گیریم، داریم

$$m_{۲۱} = \frac{۰٫۳۴۵۴}{۰٫۰۰۰۳ \times ۱۰^m} = ۱۱۵۱٫۰ \times ۱۰^{-m}$$

بنابراین

---

1. plausible

$$a_{۲۲}^{(۱)} = -۲٫۴۳۶ - (۱۱۵۱٫۰ \times ۱۰^{-m})(۱٫۵۶۶ \times ۱۰^{m}) = -۱۸۰۴٫۰$$

$$b_{۲}^{(۱)} = ۱٫۰۱۸ - (۱۱۵۱٫۰ \times ۱۰^{-m})(۱٫۵۶۹ \times ۱۰^{m}) = -۱۸۰۵٫۰$$

که همان نتایج قبلی است. بنابراین دوباره خواهیم داشت $x_۲ = ۱٫۰۰۱$ و بالاخره $x_۱ = (۰٫۰۰۱ \times ۱۰^{m})/(۰٫۰۰۰۳ \times ۱۰^{m}) = ۳٫۳۳۳$.

در واقع عیب این مثال این‌است که $|a_{۱۱}|$ در مقایسه با $|a_{۱۲}|$ کوچك‌است. لذا خطای نسبتاً کوچك ناشی از گردکردن درمحاسبهٔ $x_۲$، منجر به‌اختلاف بزرگی بین جواب محاسبه شدهٔ $x_۱$ و جواب صحیح $x_۱$ می‌شود. این امر زمانی تأیید می‌شودکه معادلهٔ دوم را به‌عنوان معادلهٔ لولایی انتخاب کنیم، دراین‌صورت داریم $|a_{۲۲}/a_{۲۱}| \approx ۶$ که در مقایسه با $|a_{۱۲}/a_{۱۱}| \approx ۵۲۲۰$ کوچك‌است. از آنجا

$$m_{۱۱} = \frac{۰٫۰۰۰۳}{۰٫۳۴۵۴} = ۰٫۰۰۰۸۶۸۶$$

و معادلهٔ جدید اولی به‌صورت زیر درمی‌آید

$$۱٫۵۶۸ x_۲ = ۱٫۵۶۸$$

به‌طوری‌که $x_۲ = ۱$ همان جواب صحیح‌است وسرانجام از معادلهٔ دوم به‌دست می‌آوریم $x_۱ = ۱۰$. اما (دراین حالت) حتی اگر خطای گردکردن موجب شودکه داشته باشیم $x_۲ = ۱٫۰۰۱$ (همان‌طوری‌که در مثال ۴.۴ پیش‌آمد) باز معادلهٔ دوم خواهد داد

$$x_۱ = \frac{۱٫۰۱۸ + ۲٫۴۳۸}{۰٫۳۴۵۴} = ۱۰٫۰۱$$

که نتیجهٔ خوبی‌است.

تعیین چگونگی تأثیر تدبیرهای مختلف لولایی بر دقت عمل جواب محاسبه شده، بسیار مشکلتر (اگر ناممکن‌نباشد) است. یك استثنای قابل‌توجه و حائزاهمیت، دستگاههای خطی با ماتریس ضرایب معین و مثبت، یعنی دستگاههایی هستندکه برای ماتریس ضرایب آنها شرایط زیر برقرارند.

$$A = A^{T} \quad \text{و به‌ازای کلیهٔ مقادیر } \mathbf{x} \neq ۰ \text{،} \qquad \mathbf{x}^{T} A \mathbf{x} > ۰$$

برای چنین دستگاهی، می‌توان نشان دادکه خطای جواب محاسبه شده ناشی از خطای گرد کردن طی حذف و پسجایگذاری به‌طور قابل‌قبولی کوچك است [ص۱۲۷؛ ۴۱] به‌شرط آنکه تدبیرپیش‌پا افتادهٔ گزینش لولایی به‌صورت عوض نکردن جای معادله‌ها اختیارشود. (تمرین ۴.۴–۹ الگوریتم کارآمدی برای این حالت است). اما درحال حاضر امکان ارائهٔ «بهترین» تدبیر لولایی برای یك دستگاه خطی کلی امکان‌پذیر نیست، حتی به‌خوبی روشن نیست‌که عبارت «بهترین» تدبیر لولایی چه معنایی می‌تواند داشته باشد.

به‌دلیل اقتصادی باید انتخاب معادلهٔ لولایی برای هرمرحله و در آغـاز آن مرحله براساس وضعیت جاری دستگاه موردنظر یعنی بی‌اطلاع قبلی از تأثیر این گزینش برمراحل بعدی صورت گیرد.

یك تدبیر که درحال حاضر پذیرفته‌شده، **لولاگزینی جزئی مدرج**[1] است. دراین راه‌کار، ابتدا «اندازه»ی d_i مربوط به‌سطر iام از مـاتریس A را به‌ازای $i=1,\ldots,n$ محاسبه می‌کنند. اندازهٔ مناسب برای این اندازه عدد زیراست (به‌بخش ۵.۴ نگاه کنید).

$$d_i=\max_{1\leqslant j\leqslant n}|a_{ij}|$$

سپس در آغاز مرحلهٔ کلی یا مرحلهٔ kام از الگوریتم حذف ۲.۴، از بین $(n-k)$ معادلهٔ موجود، آن معادله‌ای به‌عنوان معادلهٔ لولایـی انتخاب می‌شود که مطلقاً بزرگترین ضریب x_k را نسبت به‌اندازهٔ معادله داشته باشد. به‌موجب الگوریتم ۲.۴ این بدان‌معنی است کـه عدد صحیح j بین k و n (معمولاً کوچکترین) چنان انتخاب شده‌است که

$$\frac{|w_{jk}|}{d_j}\geqslant\frac{|w_{ik}|}{d_i}\qquad i=k,\ldots,n$$ به‌ازای کلیهٔ مقادیر

واضح‌است که لولاگزینی جزئی مـدرج در انتخاب تدبیردرست لولاگزینی برای دستگاه مثال ۴.۴ دخالت می‌کند و با دوباره مدرج کردن معادلات، کنار گذاشته نمی‌شود.

این‌امکان وجود داردکه الگوریتم۲.۴ چنان اصلاح شود که نه‌فقط معادلات لولایی، بلکـه مجهولاتی هم‌که باید حذف شوند در معرض انتخاب قرار گیرند. دراین تغییر دو جایگشت $\mathbf{p}$ و $\mathbf{q}$ انتخاب می‌شوندکه معادلهٔ p_kام را به‌عنوان معادله‌ای‌که باید طی مرحلهٔ kام حذف x_{q_k}، $k=1,\ldots,n-1$، به‌کار رود تعیین‌می‌کند. در **لولاگزینی‌کامل**[2] انتخاب معادلهٔ لولایی و مجهولی‌که می‌باید حذف شود با توجه به‌ضریب مطلقاً بزرگتر هر یك از $n-k$ مجهول درهریك از $n-k$ معادلهٔ موردنظرانجام می‌گیرد. بدیهی‌است‌که این‌تدبیر بسیار گرانتر از لولاگزینی جزئـی مدرج تمام می‌شود، و بنابرایـن چندان به‌کار گرفتـه نمی‌شود، ولو اینکه به‌طور مسلم این تدبیر برتر از لولاگزینی جزئی می‌باشد.

تمرین

۱-۳.۴ الگوریتم ۲.۴ را به‌گونه‌ای اصلاح کنیدکه درآن لولاگزینی‌کامل انجام پذیرد.

۲-۳.۴ مثالی از یك دستگاه خطی 2×2 ارائه دهیدکـه درآن لولاگزینی‌کامل نتیجه‌ای دقیقتر ازلولاگزینی جزئی مدرج بدهد. محاسبات در ممیز شناور را با ۴ رقم اعشاری در نظر بگیرید. (راهنمایی: a_{11} و a_{21} را در مقایسه با a_{12} و a_{22} «کوچك بگیرید»).

---

1. scaled partial pivoting 2. total pivoting

۳-۳۰۴ دستگاه معادلات خطی زیر را با به‌کارگیری محاسبات ممیز شناور و چهار رقم اعشاری حل کنید. برای حل آن یکبار معادلهٔ اول و یکبار معادلهٔ دوم را به‌عنوان معادلهٔ لولایی درنظر بگیرید و نهایتاً حل دستگاه فوق را با روش لولاگزینی کامل نیز انجام دهید وجوابها را درهرسه مورد با جواب دقیق $x_1 = ۱ر۰۰۰$ و $x_2 = ۰ر۲۵۰۰$ مقایسه کنید.

$$۰ر۱۴۱۰ \times ۱۰^{-۲} x_1 + ۰ر۴۰۰۴ \times ۱۰^{-۱} x_2 = ۰ر۱۱۴۲ \times ۱۰^{-۱}$$

$$۰ر۲۰۰۰ \times ۱۰^{۰} x_1 + ۰ر۴۹۱۲ \times ۱۰^{۱} x_2 = ۰ر۱۴۲۸ \times ۱۰^{۱}$$

۴-۳۰۴ دستگاه خطی تمرین ۲.۴-۸ را با استفاده از لولاگزینی جزئی مدرج حل کنید و جوابهای خود را با جوابهای تمرین ۲.۴-۸ مقایسه کنید.

۴.۴ تجزیه به‌عوامل مثلثی

روند حذف درالگوریتم ۲.۴ رامی‌توان به‌عنوان یك **تجزیه**[1] ماتریس‌ضرایب A به سه‌عامل

$$A = PLU$$

تلقی کـرد کـه در آن p ماتریس جایگشتی است معرف تعویض سطرها، و L یك ماتریس واحد پایین مثلثی‌است که شامل مضاربی است که (در جالبترین قسمت آن) به‌کاربرده شده است و بالاخره U یك ماتریس بالامثلثی‌است. درحالتی کـه ماتریس A یك ماتریس معین مثبت و متقـارن باشد، این دیـد نسبت به‌الگوریتم ۲.۴ به‌یك الگوریتم کارایـی (**تجزیه به‌عوامل چولسکی**[1]، تمرین ۴.۴-۹ را ببینید) منجر خواهد شد. همچنین لازم‌است به‌طرح با ارزش به‌اصطلاح **طرحهای فشرده**[2] (منسوب به‌دولیتل و کروت[3]، تمرین ۴.۴-۸ را ببینید) توجه نمودکه مزیت آنها درحل دستگاههای خطی توسط ماشینهای حساب رومیزی (یـا جیبی) است کـه تعداد نتایج محاسبات میانی راکه می بـاید ثبت گردند، کاهش مـی‌دهد. روشهای مذکور را می‌توان برای کاهش اثرات ناشی از خطای گردکردن، به‌هنگام جمع حاصلضربهای عددی (دربعضی ماشینها) در روش‌جمع بادقت مضاعف[4] به‌کاربرد. سرانجام آنکه، توجه به‌روش حذف از دیدگاه تجزیـه به‌عوامل، امکان **تحلیل خطای پسرو**[5] را در روش حذف (همان‌گونه‌که دربخش ۶.۴ انجام خواهدگرفت) آسان می‌سازد.

اولاً فرض می‌کنیم که در طی اجرای این الگوریتم هیچ‌گونه تعویض سطری اتفاق نیفتد وآنچه راکه برای معادلهٔ iام پیش خواهد آمد مورد بررسی قرار می‌دهیم. به‌ازای $k = ۱, ۲, \ldots, i-۱$ در طی مرحلهٔ kام، معادله ازصورت

$$a_{ik}^{(k-۱)} x_k + a_{i,k+۱}^{(k-۱)} x_{k+۱} + \cdots + a_{in}^{(k-۱)} x_n = b_i^{(k-۱)}$$

---

1. Choleski 2. compact schemes 3. Doolittle and Crout
4. double-precision 5. backward error analysis

به‌صورت

$$a_{i,k+1}^{(k)}x_{k+1}+\cdots+a_{in}^{(k)}x_n=b_i^{(k)}$$

درخواهد آمد که در آن

$$a_{ij}^{(k)}=a_{ij}^{(k-1)}-m_{ik}a_{kj}^{(k-1)}$$
$$b_i^{(k)}=b_i^{(k-1)}-m_{ik}b_k^{(k-1)}$$

و ضریب

$$m_{ik}=a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$$

در درایهٔ (i,k)ام آرایه‌ای که کار می‌کنیم ذخیره می‌گردد. در اینجا $a_{kj}^{(k-1)}$ و $b_k^{(k-1)}$ در مرحلهٔ kام، به‌ترتیب ضرایب و سمت راست معادلهٔ لولایی هستند و از این‌روشکل نهایی خود را دارند. این بدان‌معنی است که برحسب خروجی از الگوریتم ۲.۴ یعنی برحسب ماتریس بالامثلثی U و بردار $\mathbf{y}$ حاصله در آن الگوریتم، داریم

$$a_{kj}^{(k-1)}=u_{kj},\ j=k,\ldots,n \quad \text{و} \quad b_k^{(k-1)}=y_k$$

در نتیجه

$$\left.\begin{aligned} m_{ik}&=a_{ik}^{(k-1)}/u_{kk}\\ a_{ij}^{(k)}&=a_{ij}^{(k-1)}-m_{ik}u_{kj},\ j=k,\ldots,n\\ b_i^{(k)}&=b_i^{(k-1)}-m_{ik}y_k \end{aligned}\right\} k=1,\ldots,i-1$$

چون $a_{ij}^{(0)}=a_{ij}$ و $b_i^{(0)}=b_i$، از آنجا نتیجه می‌شود که

$$m_{ik}=(a_{ik}-m_{i1}u_{1k}-m_{i2}u_{2k}-\cdots-m_{i,k-1}u_{k-1,k})/u_{kk}\quad k=1,\ldots,i-1$$
$$u_{ij}=a_{ij}-m_{i1}u_{1j}-m_{i2}u_{2j}-\cdots-m_{i,i-1}u_{i-1,j}\qquad j=i,\ldots,n \quad (۲۳.۴)$$

و

$$y_i=b_i-m_{i1}y_1-m_{i2}y_2-\cdots-m_{i,i-1}y_{i-1}\qquad (۲۴.۴)$$

اکنون دوباره این معادلات را چنان می‌نویسیم که داده‌های اولیه، یعنی A و $\mathbf{b}$، درسمت راست ظاهرشوند. دراین‌صورت داریم

$$m_{i1}u_{1k}+\cdots+m_{ik}u_{kk}=a_{ik}\qquad k=1,\ldots,i-1$$
$$m_{i1}u_{1j}+\cdots+m_{i,i-1}u_{i-1,j}+u_{ij}=a_{ij}\qquad j=i,\ldots,n \qquad (۲۵.۴)$$

و

$$m_{i,1}y_1+\cdots+m_{i,i-1}y_{i-1}+y_i=b_i$$

بنابراین، اگر $L=(l_{ij})$ را ماتریس واحد پایین‌مثلثی بگیریم که برحسب مقادیر نهایی آرایهٔ عملی W به‌صورت

$$l_{ij}=\begin{cases} w_{ij}=m_{ij}, & i>j \\ 1, & i=j \\ 0, & i<j \end{cases} \qquad (۲۶.۴)$$

داده شده‌اند، آنگاه می‌توانیم این معادلات‌را (به‌ازای $i=1,\ldots,n$) به‌شکل ماتریسی سادهٔ زیر بنویسیم

$$LU=A$$

و

$$L\mathbf{y}=\mathbf{b}$$

این رابطه تجزیه به‌عوامل مثلثی را، درصورتی که هیچ تعویضی صورت نگرفته باشد،اثبات می‌کند. ازسوی‌دیگر، اگر تعویضهایی صورت گرفته بود وما اول، این‌تعویضها را انجام داده سپس الگوریتم ۲.۴ را بدون هیچ‌گونه تعویضی به‌کار برده بودیم، درآن صورت محتوای نهایی W تغییرمی‌کرد. علت آن این است که تمامی عملیات موجود در الگوریتم متضمن‌کاستن مضربی از یك سطر ازسایر سطرهاست تا در سطرهای دیگرصفری پدید آید و بدین جهت ترتیب نوشتن سطرها امر مهمی به‌شمار نمی‌آید. تنها چیزی که می‌باید به‌آن توجه داشت این‌است‌که وقتی‌سطری به‌عنوان یك سطر لولایی به‌کار می‌رود، دیگر نباید تغییرکند و به‌همین دلیل سطرهایی‌که به‌عنوان سطرهای لولایی به‌کار نرفته‌اند می‌باید به‌صورت جداگانه نگهداری شوند.

در نتیجه اگر تعویض سطرها در طی اجرای الگوریتم ۲.۴ صورت گیرد، آنگاه ماتریسهای L و Uکه به‌توسط الگوریتم حاصل شده‌اند، به‌ازای ماتریس جایگشتی مناسبی مانند P، در رابطهٔ زیرصدق می‌کنند

$$LU=P^{-1}A$$

آنگاه

$$PLU=A \qquad (۲۷.۴)$$

و همچنین

$$L\mathbf{y}=P^{-1}\mathbf{b} \qquad (۲۸.۴)$$

برحسب بردار $\mathbf{p}$ که در الگوریتم ۲.۴ برای ثبت تعویض سطرها به‌کاربرده شده،

معادلهٔ p_kام در طی مرحلهٔ kام به‌عنوان معادلهٔ لولایی به‌کار رفته است. از این‌رو P^{-1} به ازای جمیع مقادیر k می‌باید سطر p_kام را به‌سطر kام بدل کند. این‌امر بدان معنی است که، اگر واقعاً بخواهیم بدانیم، به‌ازای جمیع مقادیر k تساوی، $P\mathbf{i}_k = \mathbf{i}_{p_k}$ برقرار است [بخش (iii) قضیهٔ ۵.۴ را ببینید]. ولی آنچه که برای ما اهمیت دارد، برحسب خروجی $\mathbf{p}$ در الگوریتم ۲.۴، عبارت زیر است

$$P^{-1}\mathbf{b} = (b_{p_k})_{k=1}^{n}$$

به‌عنوان اولین نمونه از کاربرد دیدگاه تجزیه به‌عوامل، اکنون در مورد امکان خرد کردن روند حل معادلهٔ $A\mathbf{x} = \mathbf{b}$ به‌دومرحله، بررسی به‌عمل می‌آوریم، دومرحلهٔ مذکور عبارتند ازمرحلهٔ تجزیه به‌عوامل[1] که درآن عوامل‌مثلثی L و U (و احتمالا ترتیب متفاوتی از سطرهای $\mathbf{p}$) حاصل می‌شوند و دیگر مرحلهٔ یافتن جواب[2]، که درطی آن، ابتدا دستگاه مثلثی

$$L\mathbf{y} = P^{-1}\mathbf{b} \left(= (b_{p_k})_{1}^{n}\right) \qquad (۲۹.۴)$$

نسبت به $\mathbf{y}$ حل می‌شود وسپس دستگاه مثلثی $U\mathbf{x} = \mathbf{y}$ نسبت به $\mathbf{x}$ با روش پسجایگذاری حل‌می‌شود. بایدتوجه داشت که سمت راست $\mathbf{b}$، تنها درمرحلهٔ دوم وارد می‌شود. ازاین‌رو اگر دستگاه معادلات قرار باشد نسبت به‌مقادیر سمت راست دیگری نیز حل شود. تنها مرحلهٔ دوم‌است که تکرار می‌شود.

به‌موجب رابطهٔ (۲۴.۴)، دستگاه مثلثی(۲۹.۴) درالگوریتم ۲.۴ طی مراحل زیر حل‌می‌شود

For $k = 1, \ldots, n$, do:
$\quad y_k := b_{p_k} - l_{k1}y_1 - \cdots - l_{k,k-1}y_{k-1}$

در واقع این‌امرشبیه به‌الگوریتم پسجایگذاری (۱.۴) است که برای حل $U\mathbf{x} = \mathbf{y}$ نسبت به $\mathbf{x}$ به‌کار می‌رود با این تفاوت که بررسی معادله‌ها، به‌علت پایین-مثلثی بودن ماتریس L، از اولین معادله تا آخرین معادله صورت می‌گیرد، تمامی مراحل حل مسئله به‌گونهٔ زیر ثبت می‌گردد:

الگوریتم ۴.۴: پیشجایگذاری و پسجایگذاری[3]

مندرجات نهایی n ستون اول آرایهٔ مورد عمل W و n-برداری $\mathbf{p}$ از الگوریتم ۲.۴ (که برای دستگاه $A\mathbf{x} = \mathbf{b}$ به‌کارگرفته شده) و نیز سمت راست معادله، یعنی $\mathbf{b}$، داده شده‌اند.

---

1. factorization phase 2. solving phase
3. Forward-and back-substitution

$$\text{For } k = 1, \ldots, n, \text{ do:}$$
$$\quad y_k := b_{p_k} - \sum_{j=1}^{k-1} w_{kj} y_j$$

$$\text{For } k = n, n-1, \ldots, 1, \text{ do:}$$
$$\quad x_k := \left(y_k - \sum_{j=k+1}^{n} w_{kj} x_j \right) / w_{kk}$$

اکنون بردار $\mathbf{x} = (x_i)$ جواب $A\mathbf{x} = \mathbf{b}$ را در بردارد.

باز توجه کنید که هردو حاصلجمع گاهی اوقات تهی هستند.

اهمیت عملی بحث قبلی زمانی آشکار می‌شود که تعداد عملیات (با ممیز شناور) را در الگوریتم‌های ۲.۴ و ۴.۴ بشماریم و به‌موجب تمرین ۲.۴-۲، n عمل تقسیم و $n(n-1)/2$ عمل ضرب، و $n(n-1)/2$ عمل جمع برای انجام حلقهٔ دوم الگوریتم ۴.۴ لازم است. حلقهٔ اول نیز به‌همین تعداد عملیات نیاز دارد. جز اینکه به‌هیچ عمل تقسیم احتیاجی ندارد. از این‌رو الگوریتم ۴.۴ متضمن

$n(n-1)$ عمل جمع و n^2 عمل ضرب یا تقسیم می‌باشد.

با در نظر گرفتن تمرین ۲.۴-۴ این تعداد دقیقاً تعداد عملیاتی است که برای ضرب یك ماتریس $n \times n$ در یك n-برداری لازم است.

برعکس، برای محاسبهٔ محتوای نهایی اولین n ستون ماتریس مورد عمل W به‌وسیلهٔ الگوریتم ۲.۴،

$n^3/3 + n^2/2 - n/6$ عمل جمع و $(n^3 - n)/3$ عمل ضرب یا تقسیم لازم است (به‌تمرین ۲.۴-۵ نگاه کنید). بنابراین عمدهٔ کار لازم در حل $A\mathbf{x} = \mathbf{b}$ از راه روش حذف، به‌دست آوردن محتوای نهایی ماتریس مورد عمل W، یعنی، $\theta[(1/3)n^3]$ عمل جمع و به‌همین تعداد عمل ضرب یا تقسیم به‌ازای nهای بزرگ است. پیشجایگذاری و پسجایگذاری بعدی به‌همین تعداد عملیات با یك درجه کمتر، یعنی $\theta(n^2)$ عمل جمع و همین تعداد عمل ضرب به‌ازای هر مقدار سمت راست معادله لازم دارد. بنابراین، وقتی محتوای نهایی W را دانستیم می‌توانیم در مدت زمانی که برای محاسبهٔ محتوای نهایی W لازم است، دستگاه $A\mathbf{x} = \mathbf{b}$ را برای مقادیر بسیار زیاد سمت راست حل کنیم.

در محاسبهٔ این کارها، روش سنتی را دنبال کرده‌ایم و تنها عملیات با ممیز شناور را به‌حساب آورده‌ایم به‌ویژه، محاسبهٔ زیرنمایه‌ها، تعیین هزینه برای اجرای حلقه‌های DO و سایر هزینه‌های مربوط به ثبت وضبط یا ساماندهی[1] را نادیده گرفته‌ایم، این امر به‌جهت آنست که محاسبات اخیر بسیار سریعتر از عملیات با ممیز شناور انجام می‌گیرند. این نوع حسابرسی روی کامپیوترهای امروزی موردی ندارد و تصویر درستی از مقدار هزینه به‌دست

1. bookkeeping

نمی‌دهد (تمرین۲.۴ـ۷ را ببینید). از سوی دیگر این مطلب کـه تا چه حد انجام یك‌کار (ازلحاظ محاسبهٔ زمان موردنیاز) به‌ساماندهی دریك برنامه بستگی دارد، درکامپیوترهای مختلف به‌شدت فرق می‌کند و بنابراین ارائهٔ یك بحث‌کلی در این موضوع در این‌کتاب دشواراست.

یك زیربرنامهٔ **فورترن** به‌نام SUBST که متضمن الگوریتم جایگذاری۴.۴ است در زیرداده شده‌است.

```
      SUBROUTINE SUBST ( W, IPIVOT, B, N, X )
      INTEGER IPIVOT(N),   I,IP,J
      REAL B(N),W(N,N),X(N),   SUM
C******  I N P U T  ******
C  W, IPIVOT, N  ARE AS ON OUTPUT FROM  F A C T O R , APPLIED TO THE
C     MATRIX  A  OF ORDER  N .
C  B  IS AN N-VECTOR, GIVING THE RIGHT SIDE OF THE SYSTEM TO BE SOLVED.
C******  O U T P U T  ******
C  X  IS THE N-VECTOR SATISFYING  A*X  =  B .
C******  M E T H O D  ******
C  ALGORITHM 4.4 IS USED, I.E., THE FACTORIZATION OF  A  CONTAINED IN
C  W  AND  IPIVOT  (AS GENERATED IN  FACTOR ) IS USED TO SOLVE  A*X = B
C  FOR  X  BY SOLVING TWO TRIANGULAR SYSTEMS.
C
      IF (N .LE. 1) THEN
         X(1) = B(1)/W(1,1)
                                              RETURN
      END IF
      IP = IPIVOT(1)
      X(1) = B(IP)
      DO 15 I=2,N
         SUM = 0.
         DO 14 J=1,I-1
   14       SUM = W(I,J)*X(J) + SUM
         IP = IPIVOT(I)
   15    X(I) = B(IP) - SUM
C
      X(N) = X(N)/W(N,N)
      DO 20 I=N-1,1,-1
         SUM = 0.
         DO 19 J=I+1,N
   19       SUM = W(I,J)*X(J) + SUM
   20    X(I) = (X(I) - SUM)/W(I,I)
                                              RETURN
      END
```

باردیگر، یك زیربرنامهٔ **فورترن** به‌نام FACTOR می‌دهیم که ازالگوریتم حذف ۲.۴ و تدبیرگزینش جزئی لولایی مـدرج به‌کارگرفته شده است و این به‌کارگیری برای محاسبهٔ تجزیه به‌عوامل مثلثی(درصورت امکان) یك ماتریس مفروض A ازمرتبهٔ $N \times N$، ذخیره نمودن عوامل تجزیـهٔ موجود درماتریس W از مرتبهٔ $N \times N$، و تـدبیرگزینش لولایی در یك N-برداری IPIVOT است‌که برای استفاده از زیر برنامهٔ SUBST که قبلا داده شده‌آماده‌است. استفاده‌کننده بـاید یك N-برداری D به‌عنوان فضای کار[1] برای ذخیره‌نمودن«اندازه» سطرهای A، نیزفراهم‌کند. اگربه‌ماتریس A نیازی‌نباشد وجا برای ذخیره‌کم باشد، خود A را می‌توانیم به‌جای W در زمرهٔ شناسه‌های دستورالعملCALL (که این عمل در برخی ازگویشهـای زبان **فورترن** غیرقانونی‌است) به‌کاربریم. در این صورت عوامل تجزیهٔ به‌دست آمده درآرایهٔ A ضبط و جایگزین ماتریس اولیه می‌شوند.

1. working space

```
      SUBROUTINE FACTOR ( W, N, D, IPIVOT, IFLAG )
      INTEGER IFLAG,IPIVOT(N),   I,ISTAR,J,K
      REAL D(N),W(N,N),   AWIKOD,COLMAX,RATIO,ROWMAX,TEMP
C****** I N P U T ******
C  W  ARRAY OF SIZE (N,N) CONTAINING THE MATRIX  A  OF ORDER  N  TO BE
C     FACTORED.
C  N  THE ORDER OF THE MATRIX
C****** W O R K   A R E A ******
C  D  A REAL VECTOR OF LENGTH N, TO HOLD ROW SIZES
C****** O U T P U T ******
C  W  ARRAY OF SIZE (N,N) CONTAINING THE LU FACTORIZATION OF  P*A  FOR
C     SOME PERMUTATION MATRIX  P  SPECIFIED BY  IPIVOT .
C  IPIVOT  INTEGER VECTOR OF LENGTH  N  INDICATING THAT ROW  IPIVOT(K)
C     WAS USED TO ELIMINATE  X(K) , K=1,...,N .
C  IFLAG   AN INTEGER,
C       = 1, IF AN EVEN NUMBER OF INTERCHANGES WAS CARRIED OUT,
C       = -1, IF AN ODD NUMBER OF INTERCHANGES WAS CARRIED OUT,
C       = 0, IF THE UPPER TRIANGULAR FACTOR HAS ONE OR MORE ZERO DIA-
C            GONAL ENTRIES.
C     THUS, DETERMINANT(A) = IFLAG*W(1,1)*...*W(N,N) .
C     IF IFLAG .NE. 0, THEN THE LINEAR SYSTEM A*X = B CAN BE SOLVED FOR
C      X  BY A
C        CALL SUBST (W, IPIVOT, B, N, X )
C****** M E T H O D ******
C  THE PROGRAM FOLLOWS ALGORITHM 4.2, USING SCALED PARTIAL PIVOTING.
C
      IFLAG = 1
C              INITIALIZE IPIVOT, D
      DO 9 I=1,N
         IPIVOT(I) = I
         ROWMAX = 0.
         DO 5 J=1,N
    5       ROWMAX = AMAX1(ROWMAX,ABS(W(I,J)))
         IF (ROWMAX .EQ. 0.) THEN
            IFLAG = 0
            ROWMAX = 1.
         END IF
    9    D(I) = ROWMAX
      IF (N .LE. 1)                          RETURN
C              FACTORIZATION
      DO 20 K=1,N-1
C                 DETERMINE PIVOT ROW, THE ROW  ISTAR .
         COLMAX = ABS(W(K,K))/D(K)
         ISTAR = K
         DO 13 I=K+1,N
            AWIKOD = ABS(W(I,K))/D(I)
            IF (AWIKOD .GT. COLMAX) THEN
               COLMAX = AWIKOD
               ISTAR = I
            END IF
   13    CONTINUE
         IF (COLMAX .EQ. 0.)  THEN
            IFLAG = 0
         ELSE
            IF (ISTAR .GT. K)  THEN
C                 MAKE  K  THE PIVOT ROW  BY INTERCHANGING IT WITH
C                 THE CHOSEN ROW  ISTAR .
               IFLAG = -IFLAG
               I = IPIVOT(ISTAR)
               IPIVOT(ISTAR) = IPIVOT(K)
               IPIVOT(K) = I
               TEMP = D(ISTAR)
               D(ISTAR) = D(K)
               D(K) = TEMP
               DO 15 J=1,N
                  TEMP = W(ISTAR,J)
                  W(ISTAR,J) = W(K,J)
   15             W(K,J) = TEMP
            END IF
C                 ELIMINATE  X(K)  FROM ROWS  K+1,...,N .
   16       DO 19 I=K+1,N
               W(I,K) = W(I,K)/W(K,K)
               RATIO = W(I,K)
               DO 19 J=K+1,N
                  W(I,J) = W(I,J) - RATIO*W(K,J)
   19       CONTINUE
         END IF
```

```
   20 CONTINUE
      IF (W(N,N) .EQ. 0.)    IFLAG = 0
                                        RETURN
      END
```

بحث فوق، روش بسیار کارآمدی‌است برای محاسبهٔ وارون یك ماتریس وارونپذیر مفروضی مانند A ازمرتبهٔ n. همان‌طوری که در بخش ۱.۴ اشاره‌شد به‌ازای $j = 1, \ldots, n$، ستون jام $A^{-1}\mathbf{i}_j$ از ماتریس وارون A^{-1}، جواب دستگاه خطی

$$A\mathbf{x} = \mathbf{i}_j$$

است. بنابراین، برای محاسبهٔ A^{-1}، یکبار از زیربرنامهٔ FACTOR استفاده می‌شود، سپس هریك از n دستگاه $A\mathbf{x} = \mathbf{i}_j$، $j = 1, \ldots, n$، به‌وسیلهٔ الگوریتم ۴.۴ یعنی با استفاده از SUBST حل می‌شود. بنابراین به‌محض اینکه عملیات حذف انجام پذیرد فقط $n \times n^2$ عمل ضرب و تقریباً همین تعداد عمل جمع برای پیداکردن A^{-1} لازم‌است.

در عین‌حالی که این دستورالعمل ساده را، برای محاسبهٔ وارون یك ماتریس‌داده‌ایم بلافاصله باید متذکرشویم که معمولا همیشه دلیل موجهی برای محاسبهٔ ماتریس وارون‌وجود ندارد. گاهی‌اوقات در بعضی مسائل اتفاقاً ممکن است‌که درایه‌های A^{-1} دارای ارزش فیزیکی خاصی باشند. مثلا درروشهای آماری برای برازاندن یك تابع به‌داده‌های تجربی به‌وسیلهٔ روش کوچکترین توانهای[1] دوم، درایه‌های ماتریس A^{-1} اطلاعاتی دربارهٔ نوع و اندازهٔ خطا در داده‌ها دربر دارند. اما هرگاه به A^{-1} فقط به‌منظور محاسبهٔ بردار $A^{-1}\mathbf{b}$ (مثلا درحل $A\mathbf{x} = \mathbf{b}$) و یا حاصلضرب ماتریسی $A^{-1}B$ احتیاج باشد، A^{-1} نمی‌بایستی به‌طور صریح محاسبه شود، بلکه بهتر است الگوریتم جایگزینی ۴.۴ برای محاسبهٔ این حاصلضرب به‌کار رود. دلیل این توصیه چنین‌است: محاسبهٔ بردار $A^{-1}\mathbf{b}$ برای یك $\mathbf{b}$ی داده شده به‌معنی پیداکردن جواب معادلهٔ خطی $A\mathbf{x} = \mathbf{b}$ است. همان‌طوری که قبلا اشاره شد، به‌محض اینکه ماتریس A به‌توسط الگوریتم ۲.۴ به‌عوامل مثلثی تجزیه شود محاسبهٔ $A^{-1}\mathbf{b}$ می‌تواند از راه الگوریتم ۴.۴ به‌آسانی و دقیقاً با همان تعداد عملیات ضرب و جمعی که برای تشکیل حاصلضرب A^{-1} و $\mathbf{b}$ لازم‌است، انجام پذیرد. بنابراین، همین قدر که تجزیه به‌عوامل مثلثی برای A دردست باشند، دانستن A^{-1} به‌طور صریح امتیازی در محاسبهٔ $A^{-1}\mathbf{b}$ ندارد، (زیراکه تشکیل حاصلضرب $A^{-1}B$ شامل ضرب هر ستون B در A^{-1} است، این تذکر برای محاسبهٔ چندین حاصلضرب ماتریسی‌نیز کاربرد دارد.) ازسوی دیگر اولین گام در محاسبهٔ A^{-1} پیداکردن عوامل مثلثی برای A است که سپس با nبار کاربرد الگوریتم جایگذاری دنبال می‌شود، بنابراین محاسبهٔ A^{-1}مستلزم مقدارقابل‌ملاحظه‌ای محاسبات‌اولیه درمقایسه با محاسبهٔ $A^{-1}\mathbf{b}$ است. بعلاوه ماتریسی که این چنین محاسبه شده باشد فقط یك وارون تقریبی است و به‌تعبیری دقت عمل کمتری نسبت به‌عوامل مثلثی‌دارد، زیرا که این ماتریس به‌وسیلهٔ محاسبات اضافی دیگر روی عوامل مثلثی به‌دست می‌آید.

---

1. least squares

بنابراین اگردرمحاسبهٔ یك حاصلضرب ماتریسی متضمن A^{-1}، صریحاً A^{-1} را به‌کار بریم، هیچ‌گونه فایده‌ای نخواهد داشت و حتی فاقد دقت عمل خواهد بود.

در زیریك برنامهٔ **فورترن** برای محاسبهٔ وارون یك ماتریس $N \times N$ مفروض A داده‌ایم. دراین برنامه، از زیر برنامه‌های FACTOR و SUBST که قبلا داده شده‌اند استفاده شده‌است. همچنین، نمونهٔ ورودی ونتایج خروجی‌ذکرشده‌اند. تذکرهای زیربرای فهم برنامه ممکن است مفید واقع شوند. مرتبهٔ ماتریسA، یعنی N، بخشی از ورودی این برنامه‌است، بنابراین مشخص‌کردن مرتبهٔ دقیق A هنگام همگردانی[1] ممکن‌نیست. ازسوی دیگر هردو زیربرنامه FACTOR و SUBST، خواهان ماتریسهای A و (یا) W با مرتبهٔ دقیق $N \times N$ هستند. بنابراین دربرنامهٔ **فورترن** زیر، ماتریس A دریك آرایهٔ یك بعدی ذخیره می‌شودکه از این قرارداد زبان **فورترن** که آرایهٔ (I ,J)ام از یك آرایهٔ دوبعدی (N ,M) با درایهٔ (I+N*(J−۱))ام ازآرایهٔ یك‌بعدی‌هم‌ارزاست، استفاده می‌شود. قرارداد مشابهی برای ذخیره‌کردن درایه‌های ستون Jام A^{-1} در آرایهٔ یك بعدی AINV به‌کار گرفته شده‌است: درایهٔ (I+N*(J−۱)) ازآرایهٔ AINV به‌عنوان اولین‌درایهٔ N-برداری X از زیربرنامهٔ SUBST به‌این زیربرنامه داده می‌شود ودراین بردار جواب دستگاه $A\mathbf{x} = \mathbf{i}_j$ ذخیره می‌شود.

برنامهٔ فورترن برای محاسبهٔ وارون یك ماتریس داده شده

```
C  PROGRAM FOR CALCULATING THE INVERSE OF A GIVEN MATRIX
C  CALLS  F A C T O R ,  S U B S T .
      PARAMETER NMAX=30,NMAXSQ=NMAX*NMAX
      INTEGER I,IBEG,IFLAG,IPIVOT(NMAX),J,N,NSQ
      REAL A(NMAXSQ),AINV(NMAXSQ),B(NMAX)
    1 READ 501, N
  501 FORMAT(I2)
      IF (N .LT. 1 .OR. N .GT. NMAX)    STOP
C               READ IN MATRIX ROW BY ROW
      NSQ = N*N
      DO 10 I=1,N
   10    READ 510, (A(J),J=I,NSQ,N)
  510 FORMAT(5E15.7)
C
      CALL FACTOR ( A, N, B, IPIVOT, IFLAG )
      IF (IFLAG .EQ. 0) THEN
         PRINT 611
  611    FORMAT('1MATRIX IS SINGULAR')
                                          GO TO 1
      END IF
      DO 21 I=1,N
   21    B(I) = 0.
      IBEG = 1
      DO 30 J=1,N
         B(J) = 1.
         CALL SUBST ( A, IPIVOT, B, N, AINV(IBEG) )
         B(J) = 0.
   30    IBEG = IBEG + N
      PRINT 630
  630 FORMAT('1THE COMPUTED INVERSE IS '//)
      DO 31 I=1,N
```

1. compilation

```
   31    PRINT 631, I, (AINV(J),J=I,NSQ,N)
  631 FORMAT('0ROW ',I2,8E15.7/(7X,8E15.7))
                                        GO TO 1
      END
```

نمونه‌ای از ورودی

```
 3
 2.      3.      -1.
 4.      4.      -3.
-2.      3.      -1.
```

نتایج به‌دست‌آمده

وارون محاسبه شده عبارتست

```
ROW  1  0.2500000E 00   0.0              -0.2499999E  00
ROW  2  0.5000000E 00  -0.1999998E 00      0.9999996E -01
ROW  3  0.1000000E 01  -0.6000000E 00     -0.2000000E  00
```

تمرین

۱-۴.۴ برنامهٔ **فورترن** برای محاسبهٔ A^{-1} راکه در متن داده شده‌است طوری اصلاح کنیدکه به‌یك برنامهٔ کلیتری برای‌محاسبهٔ حاصلضرب $C=A^{-1}B$ منجرشود. A یك‌ماتریس $n\times n$ (وارونپذیر) و B یك ماتریس $n\times m$ است‌که هردو داده شده‌اند.

۲-۴.۴ وارون ماتریس ضرایب A از دستگاه معادلات تمرین ۲.۴—۸ را محاسبه‌کنید. سپس دقت‌ماتریس‌وارون محاسبه شده، A^{-1}_{comp} ، را از راه محاسبهٔ $A\,A^{-1}_{\text{comp}}$ و $A^{-1}_{\text{comp}}\,A$ بررسی کنید.

۳-۴.۴ نشان دهیدکه ماتریس

$$A=\begin{bmatrix} ۲ & ۲ & ۱ \\ ۱ & ۱ & ۱ \\ ۳ & ۲ & ۱ \end{bmatrix}$$

وارونپذیراست، اما A را نمی‌توان به‌صورت حاصلضرب یك ماتریس پایین-مثلثی دریك ماتریس بالامثلثی نوشت.

۴-۴.۴ ثابت‌کنیدکه جمع وضرب دوماتریس پایین (بالا-)مثلثی یك ماتریس‌پایین(بالا-) مثلثی‌است و وارون یك ماتریس پایین (بالا-)مثلثی یك ماتریس پایین(بالا-)مثلثی‌است.

۵-۴.۴ ثابت کنیدکه تجزیه به‌عوامل مثلثی به‌تعبیر زیر یکتاست: اگر A وارونپذیر و $L_۱U_۱=L_۲U_۲$، و $L_۱$ و $L_۲$ ماتریسهای واحدپایین-مثلثی و $U_۱$ و $U_۲$ ماتریسهای بالامثلثی

باشند، آنگاه $L_۱=L_۲$ و $U_۱=U_۲$ (راهنمایی: تمرین۱۰۴ـ۸ را برای اثبات وارونپذیری $L_۲$ و $U_۱$ بهکار گیرید و سپس نشان دهید که تساوی $L_۲^{-۱}L_۱=U_۲U_۱^{-۱}$ باید برقرار باشد، که با استفاده از تمرین ۴ـ۴ـ۴ نتیجه میشود که $L_۲^{-۱}L_۱$ باید یك ماتریس قطری باشد، بنابراین چون تمام درایههای قطری $L_۱$ و $L_۲$ عدد ۱ میباشند، $L_۲^{-۱}L_۱=I$).

۶ـ۴ـ۴ با استفاده از نتایج تمرین۴ـ۴ـ۵ نشان دهید که اگر A متقارن باشد ($A=A^T$) و بهعوامل مثلثی $A=LU$ تجزیه شود، آنگاه داریم $U=DL^T$ که D یك ماتریس قطری است که عناصر قطریش همان عناصر قطری U هستند.

۷ـ۴ـ۴ ثابت کنید که اگر ماتریس سه قطری A بتواند بهصورت $A=LU$ تجزیه شود، L یك ماتریس پایین-مثلثی و U یك ماتریس بالامثلثی است ـ آنگاه هم L و هم U، ماتریس سه قطری خواهند بود. الگوریتم ۳۰۴ را بهعنوان روشی برای تجزیهٔ ماتریسهای سه قطری تفسیر کنید.

۸ـ۴ـ۴ طرحهای فشرده عوامل مثلثی ماتریس A یعنی L و U را با بهکارگیری معادلهٔ (۲۳۰۴) بهشکل زیر بنا کنید تا درایههای جالب L و U بهدست آیند.

$$l_{ij}=(a_{ij}-l_{i۱}u_{۱j}-\cdots-l_{i,j-۱}u_{j-۱,j})/u_{jj} \qquad \text{بهازای } i>j$$

$$u_{ij}=a_{ij}-l_{i۱}u_{۱j}-\cdots-l_{i,i-۱}u_{i-۱,j} \qquad \text{بهازای } i\leqslant j$$

در واقع، محتوای نهایی آرایهٔ کار W باید پس از انجام همهٔ تغییرات در یکزمان، برای هردرایه، بهدست آید و بنابراین مانع نوشتن نتایج مختلف میانی شود. البته اینعمل میبایستی بهترتیب منظمی انجام گیرد. مثلا l_{ij} (بهازای $i>j$) نمیتواند بهدست آید مگر آنکه قبلا l_{ir} بهازای $r<j$، و u_{rj} به ازای $r\leqslant j$، معین شده باشند، و باز باید قبلا l_{ir} و u_{rj} بهازای $r<j$ معین باشند تا u_{ij} (برای $i\leqslant j$) محاسبه گردد.

(الف) برای بهدست آوردن L و U از A بهروش فشرده، الگوریتمی طرح کنید.

(ب) الگوریتم مذکور را چنان تغییر دهید که برای گزینش لولایی جزئی مدرج بهکار آید.

(پ) اگر الگوریتم مذکور قبلا براین اساس درست نشدهاست، آن را طوری تغییر دهید که درونیترین حلقه با همهٔ زیرنمایهها را در برگیرد (برای انگیزهٔ آن بهتمرین ۷ـ۲ـ۴ مراجعه کنید).

۹ـ۴ـ۴ روش چولسکی اگر ماتریس A ازمرتبهٔ n، حقیقی، متقارن ($A=A^T$) ومثبت معین (یعنی بهازای تمام n-برداریهای غیرصفر $\mathbf{x}$، $\mathbf{x}^TA\mathbf{x}>۰$) باشد، آنگاه امکانپذیر است که A بهصورت LDL^T تجزیه شود، که در آن L یك ماتریس واحد پایین-مثلثی و حقیقی و $D=(d_{ij})$ یك ماتریس قطری (مثبت) است. بنابراین از رابطهٔ (۲۳۰۴) خواهیم داشت

$$l_{ij}=(a_{ij}-l_{i۱}l_{j۱}-\cdots-l_{i,j-۱}l_{j,j-۱})/d_{jj} \qquad \text{بهازای } i>j$$

در حالی که داریم

$$d_{jj}=a_{jj}-l_{j1}^{2}-\cdots-l_{j,j-1}^{2}$$

براساس این معادله‌ها برای تولید (جالبترین قسمت) L و D، یك زیربرنامهٔ فورترن بنویسید وزمانی که L و D مشخص‌شدند، برای حل $A\mathbf{x}=\mathbf{b}$ برحسب $\mathbf{x}$ از راه‌جایگذاری نیز یك زیربرنامه تهیه کنید.

۱۰-۴.۴ نشان دهیدکه روش چولسکی، هروقت‌که ماتریس A به‌صورت BB^T، و B ماتریس وارونپذیر باشد قابل‌اجراست.

۵.۴ خطا وباقیماندهٔ یك جواب تقریبی؛ نرم‌ها

هرجواب حساب شدهٔ یك دستگاه خطی باید، به‌دلیل خطای گردکردن، **یك** جواب تقریبی تلقی‌شود. دراین‌بخش، دربارهٔ مسئلهٔ دشوار تعیین خطای یك جواب تقریبی (بدون دانستن آن‌جواب) بحث‌می‌شود. دراین‌بحث نرم[1] را واردمی کنیم‌وازآن به‌عنوان ابزارمناسبی‌برای اندازه‌گیری «اندازه»ی بردارها وماتریسها استفاده می‌کنیم.

اگر $\hat{\mathbf{x}}$ یك جواب حساب‌شدهٔ دستگاه خطی $A\mathbf{x}=\mathbf{b}$ باشد، آنگاه **خطای** این جواب برابر با تفاضل زیراست

$$\mathbf{e}=\mathbf{x}-\hat{\mathbf{x}}$$

بدیهی‌است‌که این خطا معمولا برما مجهول‌است (چون درغیراین صورت $\mathbf{x}$ معلوم‌است و بحثهای دیگر غیرضروری‌اند). اما همواره می‌توان **باقیماندهٔ**[2] (خطا) یعنی

$$\mathbf{r}=A\mathbf{x}-A\hat{\mathbf{x}}$$

رامحاسبه‌کرد، زیراکه $A\mathbf{x}$ دقیقاً سمت راست $\mathbf{b}$ است. سپس این باقیماندهٔ $\mathbf{r}$، میزان‌صدق کردن $\hat{\mathbf{x}}$ را در دستگاه خطی $A\mathbf{x}=\mathbf{b}$ معین می‌کند. اگر $\mathbf{r}$ یك بردار صفرباشد، آنگاه $\hat{\mathbf{x}}$ جواب (دقیق) این دستگاه است. یعنی در این صورت $\mathbf{e}$ صفراست. اگر $\hat{\mathbf{x}}$ تقریب خوبی برای جواب $\mathbf{x}$ باشد، می‌توان انتظار داشت‌که کلیهٔ درایه‌های $\mathbf{r}$، حداقل به‌معنای نسبی، کوچك باشند.

□ **مثال۵.۴:** دستگاه خطی سادهٔ زیر راکه ریشهٔ یگانهٔ آن، $\mathbf{x}$، دارای درایه‌های $x_1=x_2=1$ است، درنظر می‌گیریم

$$1.01x_1+0.99x_2=2$$

$$0.99x_1+1.01x_2=2$$

جواب تقریبی $\hat{\mathbf{x}}=\begin{bmatrix}1.01\\1.01\end{bmatrix}$ خطایی مساوی $\mathbf{e}=\begin{bmatrix}-0.01\\-0.01\end{bmatrix}$ و باقیمانده‌ای برابر

---

1. norm 2. residual

$\mathbf{r}=\begin{bmatrix}-0.502\\-0.502\end{bmatrix}$ دارد، لذا در این حالت یك باقیماندهٔ «کوچك» (نسبت به‌سمت‌راست معادله) به‌یك خطای نسبتاً «کوچك» مربوط می‌شود. ازسوی دیگر جواب تقریبی $\hat{\mathbf{x}}=\begin{bmatrix}2\\0\end{bmatrix}$ خطایی مساوی $\mathbf{e}=\begin{bmatrix}-1\\1\end{bmatrix}$، ولی باقیمانده‌ای برابر $\mathbf{r}=\begin{bmatrix}-0.502\\0.502\end{bmatrix}$ دارد که بازهم باقیماندهٔ نسبتاً «کوچك» است درصورتی که خطا اکنون نسبتاً «بزرگ» شده‌است. اگر برای سمت راست مقادیر دیگری اختیار کنیم می‌توانیم نتیجهٔ عکس به‌دست‌آوریم. دستگاه خطی

$$\begin{aligned}1.501x_1+0.99x_2&=2\\0.99x_1+1.501x_2&=-2\end{aligned}\qquad(30.4)$$

دارای جواب منحصر به‌فرد $x_2=-100$ و $x_1=100$ است. جواب تقریبی $\hat{\mathbf{x}}=\begin{bmatrix}101\\-99\end{bmatrix}$ خطایی مساوی $\mathbf{e}=\begin{bmatrix}-1\\-1\end{bmatrix}$ ولی باقیمانده‌ای برابر $\mathbf{r}=\begin{bmatrix}-2\\-2\end{bmatrix}$ دارد که در این حالت باقیمانده اکنون نسبتاً «بزرگ» است، درحالی که خطا نسبتاً «کوچك» است (فقط یك درصد جواب است). □

همان‌طوری که این مثال نشان می‌دهد اندازهٔ باقیماندهٔ $\mathbf{r}=\mathbf{b}-A\hat{\mathbf{x}}$ برای یك جواب تقریبی $\hat{\mathbf{x}}$، همیشه شاخص[1] مطمئنی برای اندازهٔ خطای $\mathbf{e}=\mathbf{x}-\hat{\mathbf{x}}$ در این جواب تقریبی نیست. اینکه آیا یك باقیماندهٔ «کوچك» مستلزم یك خطای «کوچك» هست یا نیست بستگی به «اندازه»ی ضرایب ماتریس و وارون آن دارد که دقیقاً درزیر توضیح داده می‌شود. برای این بحث نیاز به‌وسیله‌ای برای اندازه‌گیری «اندازه»ی n-برداری‌ها و ماتریسهای $n\times n$ داریم.

قدرمطلق وسیله‌ای مناسب برای اندازه‌گیری «اندازه»ی اعداد حقیقی وحتی اعداد همتافت است. یقین نداریم که بتوانیم یك راه قطعی برای اندازهٔ یك n-برداری یا یك ماتریس $n\times n$ تعیین کنیم. مطمئناً یك راه قطعی برای انجام این عمل که در تمام موارد مورد قبول باشد وجود ندارد

برای مثال، معیاری که غالباً برای اندازهٔ یك n-برداری $\mathbf{a}$ به‌کار برده می‌شود، عدد غیرمنفی زیر است

$$\|\mathbf{a}\|_\infty=\max_{1\le i\le n}|a_i|\qquad(31.4)$$

---

1. indicator

اکنون فرض کنید که جواب محاسبه شدهٔ $\hat{\mathbf{x}}$ برای $A\mathbf{x}=\mathbf{b}$، که از این راه اندازه گرفته شده‌است، تا شش رقم اعشاری دقیق باشد، یعنی

$$\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} < ۱۰^{-۶} \qquad (۳۲.۴)$$

که نشان‌دهندهٔ یك جواب محاسبه شدهٔ بسیار رضایتبخش درموردی‌است که مجهولات، مثلا، مقادیر تقریبی یك جواب رضایتبخش یك معادلهٔ دیفرانسیل می‌باشند. امـا اگر یکی از مجهولات اتفاقاً درآمد سالانهٔ شما و دیگری تولید ناخالص ملی باشد، نابرابری (۳۲.۴) هیچ‌گونه اشاره‌ای به‌رضایتبخش‌بودن یانبودن $\mathbf{x}$ (درحد موردنظر) ندارد، زیرا با برقراری (۳۲.۴)، خطا درمحاسبهٔ درآمد سالانهٔ شما (حتی‌اگر فقط برای یکسال باشد) ممکن‌است مستقلا شما را بسیار ثروتمند یا برای تمام عمر بدهکار سازد. معیاری مانند

$$\|\mathbf{a}\| = \max\{۱۰^{۱۰}|a_۱|, \max_{۲\leqslant i\leqslant n}|a_i|\}$$

(با این فرض که درآمد سالانـه مجهول اول باشد) اطلاعات بسیار زیادتری برای اندازه خواهد داد تا معیارهای خاص دیگری که درآنها از چندین عدد (به‌جای به‌کارگرفتن فقط یك عدد غیرمنفی) برای بیان «اندازه» یك n-برداری استفاده می‌شود.

اما دربیشترموارد، کافی‌است که اندازهٔ یك n-برداری با نرم اندازه‌گیری شود. نرم، بعضی‌از ویژگیهای قدرمطلق اعداد را حفظ می‌کند. به‌ویژه یك نرم به‌هر n-برداری $\mathbf{a}$ یك عدد حقیقی $\|\mathbf{a}\|$ با شرایط قابل‌قبول زیر نسبت می‌دهد، که‌آن‌را **نرم** $\mathbf{a}$ می‌نامند، که تابع شرایط مناسب زیراست:

(i) به‌ازای هـر n-برداری $\mathbf{a}$، نامساوی $\|\mathbf{a}\| \geqslant ۰$ و تسـاوی $\|\mathbf{a}\| = ۰$، برقراراست اگر، و فقط اگر، $\mathbf{a}=۰$.

(ii) به‌ازای هر n-برداری $\mathbf{a}$ و هرعدد α داریم، $\|\alpha\mathbf{a}\| = |\alpha|\,\|\mathbf{a}\|$ (۳۳.۴)

(iii) به‌ازای هردو n-برداری $\mathbf{a}$ و $\mathbf{b}$ داریم، $\|\mathbf{a}+\mathbf{b}\| \leqslant \|\mathbf{a}\|+\|\mathbf{b}\|$

اولین‌شرط، همهٔ n-برداری‌ها را به‌استثنای بردارصفرمقید می‌کند که دارای «طول» مثبت باشند، شرط دوم‌بیان می‌کند که، مثلا، $\mathbf{a}$ و منفی آن $-\mathbf{a}$ یك «طول» دارند وطول $۳\mathbf{a}$، سه‌برابرطول $\mathbf{a}$ است. شرط سوم **نابرابری مثلثی**[1] نامیده می‌شود، زیراکه بیان‌می‌کند که مجموع طولهای دوضلع یك مثلث هرگز از طول ضلع سوم آن کوچکتر نیست.

احتمالا دانشجویان با طول **اقلیدسی** و یا **نرم** n-برداری $\mathbf{a}=(a_i)$

$$\|\mathbf{a}\|_۲ = \sqrt{\mathbf{a}^T\mathbf{a}} = \sqrt{|a_۱|^۲+|a_۲|^۲+\cdots+|a_n|^۲}$$

1. triangle inequality

لااقل در موارد $n=۲$ یا $n=۳$ آشنایی دارند. اما به‌دلیلی که در زیر توضیح داده می‌شود، ترجیح می‌دهیم که در مثالهای عددی زیر، از **نرم ماکسیمم** (۳۱.۴) به‌عنوان راهی برای اندازه‌گیری طول n-برداری $\mathbf{a}$ استفاده کنیم. به‌آسانی می‌توان تحقیق کرد که (۳۱.۴) یك تعریف نرم است، یعنی $\|\mathbf{a}\|=\|\mathbf{a}\|_\infty$ سه ویژگی نرم را که در (۳۳.۴) آمده است داراست. برای ویژگی (i)، $\|\mathbf{a}\|_\infty$ عبارت است از ماکسیمم کمیتهای غیرمنفی و بنابراین خود غیرمنفی است، همچنین $\|\mathbf{a}\|_\infty=۰$، اگر و فقط اگر به‌ازای کلیهٔ مقادیر i، $|a_i|=۰$، عیناً مثل این است که گفته شود $\mathbf{a}=۰$، بعلاوه اگر α یك عدد باشد، آنگاه داریم

$$\|\alpha\mathbf{a}\|_\infty=\max_i|\alpha a_i|=\max_i|\alpha||a_i|=|\alpha|\max_i|a_i|=|\alpha|\,\|\mathbf{a}\|_\infty$$

که اثبات (ii) است. بالاخره

$$\|\mathbf{a}+\mathbf{b}\|=\max_i|a_i+b_i|\leqslant\max_i(|a_i|+|b_i|)$$

$$\leqslant\max_i|a_i|+\max_i|b_i|=\|\mathbf{a}\|_\infty+\|\mathbf{b}\|_\infty$$

که اثبات (iii) است.

نرمهای برداری دیگری که غالباً به‌کار می‌روند شامل ۱-نرم

$$\|\mathbf{a}\|=\|\mathbf{a}\|_۱=\sum_{i=۱}^{n}|a_i|$$

و موارد متفاوت p-نرم وزین هستند

$$\|\mathbf{a}\|=\|\mathbf{a}\|_{p,\,w}=\Big(\sum_{i=۱}^{n}|a_i|^p w_i\Big)^{۱/p}$$

در رابطهٔ بالا p عددی است بین ۱ و ∞ و اعداد $w_۱$، $w_۲$، ...، w_n کمیتهای ثابت و مثبت‌اند. در حالت $p=۲$ و $w_i=۱$، (به‌ازای جمیع مقادیر i)، همان نرم آشنای اقلیدسی به‌دست می‌آید.

وقتی یك نرم برداری انتخاب شد، اندازهٔ متناظر ماتریس A از مرتبهٔ $n\times n$ را از مقایسهٔ اندازهٔ $A\mathbf{x}$ با اندازهٔ $\mathbf{x}$ تعیین می‌کنیم. **نرم ماتریسی** و متناظر A را دقیقاً، چنین تعریف می‌کنیم

$$\|A\|=\max\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \qquad (۳۴.۴)$$

که ماکسیمم‌گیری روی تمامی n-برداریهای (غیرصفر) $\mathbf{x}$ انجام می‌شود. می‌توان نشان

داد که این ماکسیمم به‌ازای هر ماتریس A از مرتبهٔ $n \times n$ (وانتخاب هر نرم برداری) وجود دارد. نرم ماتریسی $\|A\|$ با دو ویژگی زیر مشخص می‌شود

$$\left.\begin{array}{ll} \text{به‌ازای هر } n\text{-برداری } \mathbf{x} & \|A\mathbf{x}\| \leqslant \|A\|\,\|\mathbf{x}\| \\ \text{و} & \\ \text{به‌ازای } n\text{-برداری غیرصفری مانند } \mathbf{x} & \|A\mathbf{x}\| \geqslant \|A\|\,\|\mathbf{x}\| \end{array}\right\} \quad (۳۵.۴)$$

البته بلافاصله از (۳۵.۴) نتیجه می‌شود که به‌ازای هر $\mathbf{x}$ با شرط برقراری $\|\mathbf{x}\|\,\|A\| \geqslant \|A\mathbf{x}\|$، $\|A\mathbf{x}\| = \|A\|\,\|\mathbf{x}\|$ برقرارخواهد بود. سپس می‌توان نشان داد که ویژگیهای زیر برای نرم ماتریس(۳۴.۴) برقرارند:

$$\left.\begin{array}{ll} \text{(i)} & \text{برای تمامی ماتریسهایی نظیر ماتریس } A \text{ ازمرتبهٔ } n \times n : \|A\| \geqslant ۰ \text{،} \\ & \text{وتساوی } \|A\| = ۰ \text{ برقراراست اگر وفقط اگر } A = ۰. \\ \text{(ii)} & \text{به‌ازای جمیع ماتریسهای } A \text{ از مرتبهٔ } n \times n \text{ و جمیع اعداد } \alpha \text{،} \\ & \|\alpha A\| = |\alpha|\,\|A\| \\ \text{(iii)} & \text{به‌ازای هردو ماتریس } A \text{ و } B \text{ از مرتبهٔ } n \times n : \end{array}\right\} \quad (۳۶.۴)$$

$$\|A+B\| \leqslant \|A\| + \|B\|$$

بنابراین اصطلاح «نرم» برای عدد $\|A\|$ موجه‌است.

بعلاوه

$$\|AB\| \leqslant \|A\|\,\|B\| \quad \text{(iv)} \quad (۳۷.۴)$$

به‌ازای هردوماتریس A و B از مرتبهٔ $n \times n$ داریم

و بالاخره، اگر ماتریس A وارونپذیر باشد، آنگاه $\mathbf{x} = A^{-۱}(A\mathbf{x})$، بنابراین $\|A\mathbf{x}\|\,\|A^{-۱}\| \geqslant \|\mathbf{x}\|$. از ترکیب این رابطه با (۳۵.۴) خواهیم داشت.

$$\frac{\|\mathbf{x}\|}{\|A^{-۱}\|} \leqslant \|A\mathbf{x}\| \leqslant \|A\|\,\|\mathbf{x}\| \quad (۳۸.۴)$$

به‌ازای همهٔ n-برداریهای $\mathbf{x}$،

و هردو نامساوی نافذند، یعنی با انتخاب بردار مناسب (غیرصفر) $\mathbf{x}$، از هریك می‌توان یك نامساوی ساخت.

چنانکه مشاهده می‌شود، نرم ماتریسی

$$\|A\|_۲ = \max \frac{\|A\mathbf{x}\|_۲}{\|\mathbf{x}\|_۲}$$

که براساس نرم اقلیدسی معمولا بسیار مشکل محاسبه می‌شود. درحالی‌که نرم ماتریسی

$$\|A\|_\infty = \max \frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty}$$

که برپایهٔ نرم ماکسیمم استواراست، به‌آسانی قابل‌محاسبه بوده و با عدد زیر بیان می‌شود

$$\|A\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}| \qquad (۳۹.۲)$$

برای اثبات این، باید نشان دهیم که عدد $\|A\| = \max_i \Sigma_j |a_{ij}|$ در دو حکم (۳۵.۲) صدق می‌کند، یعنی

به‌ازای جمیع n-برداریهای $\mathbf{x}$، $\|A\mathbf{x}\|_\infty \leqslant \left(\max_i \sum_{j=1}^{n} |a_{ij}|\right)\|\mathbf{x}\|_\infty$

و به‌ازای بردارغیرصفری نظیر $\mathbf{x}$ $\|A\mathbf{x}\|_\infty \geqslant \left(\max_i \sum_{j=1}^{n} |a_{ij}|\right)\|\mathbf{x}\|_\infty$

اما به‌ازای یك $\mathbf{x}$ دلخواه داریم،

$$\|A\mathbf{x}\|_\infty = \max_{1\leqslant i\leqslant n} \left|\sum_{j=1}^{n} a_{ij}\, x_j\right| \leqslant \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|\,|x_j|$$

$$\leqslant \max_{1\leqslant i\leqslant n} \left((\max_{1\leqslant j\leqslant n}|x_j|) \sum_{j=1}^{n} |a_{ij}|\right) = \|\mathbf{x}\|_\infty \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|$$

که صحت حکم اول را ثابت می‌کند. برای حکم دوم، گیریم i_0 یك عدد صحیح بین ۱ تا n باشد، به‌طوری‌که

$$\sum_{j=1}^{n} |a_{i_0 j}| = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|$$

فرض می‌کنیم $\mathbf{x}$ یك n-برداری با نرم ماکسیمم ۱ باشد به‌طوری‌که

$$a_{i_0 j} x_j = |a_{i_0 j}| \qquad j = 1, \ldots, n$$

مثلا می‌گیریم

$$x_j = \left\{\begin{array}{ll} 1 & a_{i_0 j} \geqslant 0 \text{ اگر} \\ -1 & a_{i_0 j} < 0 \text{ اگر} \end{array}\right\} \quad j = 1, \ldots, n$$

دراین‌صورت برای این بردار $\mathbf{x}$ که غیرصفر بودن آن واضح‌است، داریم $\|\mathbf{x}\|_\infty = 1$ و

$$\|A\mathbf{x}\|_\infty = \max_{1\leqslant i\leqslant n} \left|\sum_{j=1}^{n} a_{ij} x_j\right| \geqslant \left|\sum_{j=1}^{n} a_{i_0 j} x_j\right|$$

$$= \sum_{j=1}^{n} |a_{i_0 j}| = \|\mathbf{x}\|_\infty \left(\max_{1\leqslant i\leqslant n} \sum_j |a_{ij}|\right)$$

که اثبات حکم دوم‌است.

□ **مثال ۵.۴ (الف):** برای ماتریس ضرایب A درمثال ۵.۴ به‌آسانی می‌توان دریافت که

$$\|A\|_\infty = \max\{|۱٫۰۱| + |۰٫۹۹|, |۰٫۹۹| + |۱٫۰۱|\} = ۲$$

دیدیم که

$$A\begin{bmatrix}۱\\۱\end{bmatrix} = \begin{bmatrix}۲\\۲\end{bmatrix} \quad , \quad A\begin{bmatrix}۱۰۰\\-۱۰۰\end{bmatrix} = \begin{bmatrix}۲\\-۲\end{bmatrix}$$

و از این‌رو

$$A\left(\begin{bmatrix}۱\\۱\end{bmatrix} + \begin{bmatrix}۱۰۰\\-۱۰۰\end{bmatrix}\right) = \begin{bmatrix}۴\\۰\end{bmatrix} \quad , \quad A\left(\begin{bmatrix}۱\\۱\end{bmatrix} - \begin{bmatrix}۱۰۰\\-۱۰۰\end{bmatrix}\right) = \begin{bmatrix}۰\\۴\end{bmatrix}$$

بنابراین

$$A\begin{bmatrix}\frac{۱۰۱}{۴} & \frac{-۹۹}{۴}\\ \frac{-۹۹}{۴} & \frac{۱۰۱}{۴}\end{bmatrix}$$

که نشان می‌دهد

$$A^{-۱} = \begin{bmatrix}۲۵٫۲۵ & -۲۴٫۷۵\\ -۲۴٫۷۵ & ۲۵٫۲۵\end{bmatrix}$$

$$= \begin{bmatrix}۱ & ۰\\ ۰ & ۱\end{bmatrix}$$

که درنتیجه

$$\|A^{-۱}\|_\infty = \max\{|۲۵٫۲۵| + |-۲۴٫۷۵|, |-۲۴٫۷۵| + |۲۵٫۲۵|\} = ۵۰$$

بنابراین برای این مثال رابطهٔ (۳۸.۴) بیان می‌کند که

$$\text{به‌ازای هر ۲-برداری } \mathbf{x}، \quad ۰٫۰۲\|\mathbf{x}\|_\infty \leqslant \|A\mathbf{x}\|_\infty \leqslant ۲\|\mathbf{x}\|_\infty$$

با انتخاب $\mathbf{x} = \begin{bmatrix}۱\\۱\end{bmatrix}$ خواهیم داشت $A\mathbf{x} = \begin{bmatrix}۲\\۲\end{bmatrix}$ و بنابراین $\|\mathbf{x}\|_\infty = ۱$ و $\|A\mathbf{x}\|_\infty = ۲$

و نامساوی دوم به‌تساوی تبدیل می‌شود. با انتخاب $\mathbf{x} = \begin{bmatrix}۱۰۰\\-۱۰۰\end{bmatrix}$ خواهیم داشت $A\mathbf{x} = \begin{bmatrix}۲\\-۲\end{bmatrix}$، بنابراین $\|\mathbf{x}\|_\infty = ۱۰۰$ و $\|A\mathbf{x}\|_\infty = ۲$ و دراین انتخاب $\mathbf{x}$، نامساوی اول به‌تساوی بدل می‌شود.

اکنون به‌بحث در بارهٔ رابطهٔ بین خطای $\mathbf{e}=\mathbf{x}-\hat{\mathbf{x}}$ در جواب تقریبی $\hat{\mathbf{x}}$ از $A\mathbf{x}=\mathbf{b}$ و باقیماندهٔ $\mathbf{r}=\mathbf{b}-A\hat{\mathbf{x}}$ بازمی‌گردیم داریم

$$\mathbf{r}=A\mathbf{x}-A\hat{\mathbf{x}}=A(\mathbf{x}-\hat{\mathbf{x}})=A\mathbf{e}$$

بنابراین $\mathbf{e}=A^{-1}\mathbf{r}$ از این‌رو، با توجه به $A=(A^{-1})^{-1}$ از (۳۸.۴) خواهیم داشت

$$\frac{\|\mathbf{r}\|}{\|A\|}\leqslant\|\mathbf{e}\|=\|A^{-1}\mathbf{r}\|\leqslant\|A^{-1}\|\|\mathbf{r}\| \qquad (۴۰.۴)$$

رابطهٔ فوق در باب **خطای نسبی** $\|\mathbf{e}\|/\|\mathbf{x}\|$ برحسب **باقیماندهٔ نسبی** $\|\mathbf{r}\|/\|\mathbf{b}\|$، یک کران بالا و یک‌کران پایین به‌صورت زیرمی‌دهد

$$\frac{\|\mathbf{b}\|}{\|A\|\,\|\mathbf{x}\|}\cdot\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}\leqslant\frac{\|\mathbf{e}\|}{\|\mathbf{x}\|}\leqslant\frac{\|A^{-1}\|\|\mathbf{b}\|}{\|\mathbf{x}\|}\cdot\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} \qquad (۴۱.۴)$$

در اینجا، از روی جوابی که برای دستگاه $A\mathbf{x}=\mathbf{b}$ به‌دست آمده‌است؛ می‌توان $\|\mathbf{x}\|$ را برآورد کرد. و الا، درحالت خاص $\hat{\mathbf{x}}=\mathbf{o}$، باید (۴۰.۴) را به‌کاربرد، یعنی

$$\frac{\|\mathbf{b}\|}{\|A\|}\leqslant\|\mathbf{x}\|\leqslant\|A^{-1}\|\,\|\mathbf{b}\|$$

و از (۴۱.۴) خواهیم داشت که

$$\frac{1}{\|A\|\,\|A^{-1}\|}\cdot\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}\leqslant\frac{\|\mathbf{e}\|}{\|\mathbf{x}\|}\leqslant\|A^{-1}\|\,\|A\|\,\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} \qquad (۴۲.۴)$$

کرانهای موجود در (۴۱.۴) و (۴۲.۴) به‌تعبیر زیر نافذ هستند. A و $\mathbf{x}\neq\mathbf{o}$ هرچه باشند، می‌توان مقادیر غیرصفری برای $\mathbf{e}$ و $\mathbf{r}$ چنان انتخاب کرد که یکی از دو نامساوی (۴۱.۴) به‌تساوی بدل شود. اگر بخواهیم تساوی را از یکی از نامساویهای (۴۲.۴) به‌دست آوریم، می‌باید $\mathbf{x}$ خاصی را نیز انتخاب کنیم، ولی همیشه امکان این‌گونه انتخابها وجود دارد.

به‌علت اهمیت روابط (۴۱.۴) و (۴۲.۴)، آنها را با کلمات بیان می‌کنیم: اگر برای ماتریس ضرایب وارونه‌پذیر A از دستگاه خطی $A\mathbf{x}=\mathbf{b}$ مقدار $\|A\|\,\|A^{-1}\|$ را با k نشان دهیم، آنگاه خطای نسبی یک جواب تقریبی می‌تواند به‌بزرگی k برابر، یا دقیقتر بگوییم، $\|A^{-1}\|\|\mathbf{b}\|/\|\mathbf{x}\|$ برابر باقیماندهٔ نسبی آن باشد، اما می‌تواند به‌کوچکی $1/k$ برابر، یا دقیقتر بگوییم $\|\mathbf{b}\|/(\|A\|\,\|\mathbf{x}\|)$ برابر باقیماندهٔ نسبی نیز باشد. بنابراین اگر $k\approx 1$، آنگاه خطای نسبی و باقیماندهٔ نسبی همیشه یک اندازه هستند، و باقیماندهٔ نسبی را می‌توان با اطمینان برای برآورد خطای نسبی به‌کاربرد. اما هرچه‌که k بزرگتر باشد، به‌همان نسبت از روی باقیماندهٔ نسبی اطلاعات کمتری می‌توان دربارهٔ خطای نسبی کسب کرد.

عدد $\|A\|\,\|A^{-1}\|$، **عدد شرط**[1] A نامیده می‌شود و گاهی اوقات به‌صورت زیر خلاصه‌نویسی می‌شود

$$\text{cond}\,(A) = \|A\|\,\|A^{-1}\|$$

باید توجه کرد که عدد شرط یعنی $\text{cond}\,(A)$ برای ماتریس A به‌نرم ماتریسی که به‌کار گرفته می‌شود بستگی دارد و می‌تواند برای بعضی ماتریسها با تغییر نرم ماتریسی به‌طور قابل‌ملاحظه‌ای تغییر یابد. از سوی دیگر همیشه عدد شرط حداقل برابر ۱ است، زیرا برای ماتریس واحد I داریم $\|I\| = \max\|\mathbf{x}\|/\|\mathbf{x}\| = ۱$، وطبق رابطهٔ (۳۷.۴) داریم، $\|I\| = \|AA^{-1}\| \leqslant \|A\|\,\|A^{-1}\|$.

□ **مثال ۶.۴:** با توجه به‌محاسبات قبلی می‌بینیم که برای ماتریس ضرایب A درمثال ۵.۴ داریم $\text{cond}\,(A) = \|A\|_\infty\|A^{-1}\|_\infty = ۲ \times ۵۰ = ۱۰۰$. بعلاوه مشاهده شد که درمثال ۵.۴ حقیقتاً خطای نسبی در یک جواب تقریبی می‌تواند به‌بزرگی ۱۰۰ برابر باقیماندهٔ نسبی باشد، اما می‌تواند فقط ۱/۱۰۰ باقیماندهٔ نسبی نیز باشد. □

کرانهای (۴۱.۴) و(۴۲.۴) به‌عدد $\|A^{-1}\|$، که به‌آسانی در دسترس نیست، احتیاج دارند. اما در نمونه‌ای از شرایط، عدد $\|\hat{\mathbf{e}}\|$ تخمین خوبی برای $\|\mathbf{e}\|$ است که $\hat{\mathbf{e}}$ جواب محاسبه شده در دستگاه خطی $A\mathbf{e} = \mathbf{r}$ است. از آنجایی که معمولاً $\hat{\mathbf{x}}$ از راه روش گاوس به‌دست می‌آید و یک تجزیه به‌عوامل A در دسترس است، بنابراین $\hat{\mathbf{e}}$ با محاسبات بسیار کمتری نسبت به‌آنچه که برای به‌دست آوردن $\hat{\mathbf{x}}$ لازم است، می‌تواند (به‌وسیلهٔ SUBST) به‌دست آید. دراین امر فرض آن است که $\mathbf{r}$ براساس دقت مضاعف محاسبه شده است. بردار $\hat{\mathbf{e}}$ که این گونه به‌دست می‌آید نخستین بار دست در اصلاح بارستی (الگوریتم ۵.۴) است که در بخش بعدی مورد بحث قرار خواهد گرفت.

تمرین

۱-۵.۴ تحقیق کنید که

$$\|\mathbf{a}\| = \|\mathbf{a}\|_۱ = \sum_{i=۱}^{n} |a_i|$$

می‌تواند برای تمامی n-بردارهای $\mathbf{a}$ تعریفی برای نرم باشد.

۲-۵.۴ ثابت کنید که نرم ماتریسی $\|A\|_۱$ متناظر با نرم برداری $\|\mathbf{a}\|_۱$ از تمرین ۱-۵.۴ را می‌توان به‌صورت زیر محاسبه کرد

---

1. condition number

$$\|A\|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |a_{ij}|$$

۳–۵.۴ اگر یك ۲–برداری $\mathbf{a}$ را به‌عنوان نقطه‌ای دریك صفحه بـا مختصات $\{a_1$ و $a_2\}$ تعبیر کنیم، آنگاه ۲–نرم آن، $\|\mathbf{a}\|_2$، فاصلهٔ اقلیدسی این نقطه از مرکزخواهد شد. بعلاوه، مجموعهٔ تمامی‌بردارهایی که نرم اقلیدسی‌آنها برابر ۱ است، دایره‌ای به‌شعاع ۱ پیرامون مبدأ مختصات تشکیل می‌دهند. «دایره‌ای به‌شعاع ۱ پیرامون مبدأ» رسم‌کنید کـه فاصلهٔ «نقطه»ی $\mathbf{a}$ به‌وسیلهٔ نرمهای‌زیرمحاسبه شود (الف) با ۱–نرم $\|\mathbf{a}\|_1$، (ب) با نرم $\|\mathbf{a}\|_{3/2}$، (پ) با نرم اقلیدسی $\|\mathbf{a}\|_2$، (ت) با نرم $\|\mathbf{a}\|_4$، (ث) با نرم ماکسیمم $\|\mathbf{a}\|_\infty$.

۴–۵.۴ با همان تعبیر تمرین۵.۴–۳، از ۲–بردارها به‌عنوان نقاط واقع در صفحه نشان دهیدکه برای هردو۲–برداری $\mathbf{a}$ و$\mathbf{b}$، سه نقطهٔ o، $\mathbf{a}$، $\mathbf{a}+\mathbf{b}$ رئوس مثلثی هستند با اضلاعی به‌طول (اقلیدسی) $\|\mathbf{a}\|_2$، $\|\mathbf{b}\|_2$، $\|\mathbf{a}+\mathbf{b}\|_2$ و اصطلاح «نامساوی مثلثی»، برای ویژگی (iii) مربوط به‌نرم‌ها [معادلهٔ (۳۳.۴)] را شرح دهید.

۵–۵.۴ نشان دهید، که برای هر ۲–برداری $\mathbf{a}$ و $\mathbf{b}$ و هرنرم برداری داریم

$$|\,\|\mathbf{a}\| - \|\mathbf{b}\|\,| \leqslant \|\mathbf{a}-\mathbf{b}\|$$

۶–۵.۴ نشان دهیدکه برای هر ۲–برداری $\mathbf{a}$ و $\mathbf{b}$ و هرعدد λ بین ۰ و ۱، داریم

$$\|\lambda\mathbf{a}+(1-\lambda)\mathbf{b}\| \leqslant \max(\|\mathbf{a}\|, \|\mathbf{b}\|)$$

۷–۵.۴ نشان دهیدکه نرم‌ماتریسی $\|A\| = \max(\|A\mathbf{x}\|/\|\mathbf{x}\|)$ را نیزمی‌توان به‌طریق زیرمحاسبه‌کرد

$$\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$$

۸–۵.۴ کلیهٔ احکام (۳۶.۴) را دربارهٔ نرمهای ماتریسی ثابت کنید.

۹–۵.۴ با استفاده از تمرین۵.۴–۷، $\|A\|_2$ را برای

$$A = \begin{bmatrix} 3 & -5 \\ 6 & 1 \end{bmatrix}$$

محاسبه‌کنید. (راهنمایی: یك ۲–برداری $\mathbf{x}$ دارای۲–نرم $\|\mathbf{x}\|_2 = 1$ است اگر، وفقط اگر، به‌ازای مقداری مانند θ داشته باشیم، $x_1 = \cos\theta$ و $x_2 = \sin\theta$).

۱۰–۵.۴ با استفاده ازتمرین۴.۴–۲، عدد شرط ماتریس ضرایب A ازدستگاه خطی‌تمرین ۲.۴–۸ را محاسبه‌کنید، سپس خطای نسبی و باقیماندهٔ نسبی جوابهایی راکـه در تمرین ۲.۴–۸ و ۳.۴–۴ برحسب این عدد شرط به‌دست آمده موردبحث قراردهید. همچنین‌تنها

با استفاده از الگوریتم جایگذاری ۴.۲، مقدار $\hat{\mathbf{e}} = A^{-1}\mathbf{r}$ را به‌ازای جوابهای فوق (با محاسبهٔ $\mathbf{r}$ از راه دقت مضاعف) حساب کنید.

۶.۴ تحلیل خطای پسرو و اصلاح بارستی

در بخش ۵.۴ پیشین عدد شرط ماتریس ضرایب A از یك دستگاه خطی $A\mathbf{x} = \mathbf{b}$، یعنی

$$\text{cond}(A) = \|A\|\,\|A^{-1}\| \qquad (۴۳.۴)$$

را به‌عنوان کمیتی مستقل از $\mathbf{x}$ در برآورد خطای یك جواب تقریبی شناسایی کردیم. به‌طور خلاصه: عدد شرط (۴۳.۴) معیاری است برای اینکه ببینیم، باقیماندهٔ نسبی $\|\mathbf{b} - A\hat{\mathbf{x}}\|/\|\mathbf{b}\|$ از جواب تقریبی $\hat{\mathbf{x}}$ تا چه اندازه خطای نسبی این ریشهٔ تقریبی، یعنی $\|\mathbf{x} - \hat{\mathbf{x}}\|/\|\mathbf{x}\|$ را نشان می‌دهد. بنابراین، عدد شرط معیاری است که به‌توسط آن می‌توان امیدوار بود که بتوان به‌کمك آن یك جواب (تقریبی) «خوب» را از یك جواب تقریبی «بد» از راه نگاه کردن به باقیمانده به‌خوبی تمیز داد.

واضح است که محاسبهٔ عدد شرط برای یك ماتریس داده شده کاملا دشوار است، حتی اگر بتوان نرم ماتریس را نسبتاً آسان محاسبه کرد، زیرا (برای محاسبهٔ $\text{cond}(A)$) می‌باید A^{-1} را در دست داشت. گاهی اوقات عدد شرط $\text{cond}(A)$ را می‌توان به‌کمك قضیهٔ زیر برآورد کرد، که این امر به‌شرح بیشتر معنی عدد شرط کمك می‌کند.

قضیهٔ ۸.۴ برای هر ماتریس وارونپذیر A از مرتبهٔ $n \times n$ و هر نرم ماتریسی، عدد شرط A نشانگر فاصلهٔ نسبی A از نزدیکترین ماتریس وارون‌ناپذیر $n \times n$ است. به‌ویژه داریم:

$$\frac{1}{\text{cond}(A)} = \min\left\{\frac{\|A-B\|}{\|A\|} \;\middle|\; B \text{ وارونپذیر نیست}\right\}$$

اثبات کامل این قضیه خارج از ظرفیت این کتاب است (ولی به‌تمرین ۶.۴ـ۵ نگاه کنید) فقط نشان می‌دهیم که

$$\frac{1}{\text{cond}(A)} \leqslant \inf\left\{\frac{\|A-B\|}{\|A\|} \;\middle|\; B \text{ وارون‌ناپذیر است}\right\}$$

یعنی برای هر ماتریس $n \times n$ و وارون‌ناپذیر B داریم

$$\frac{1}{\|A^{-1}\|} \leqslant \|A-B\| \qquad (۴۴.۴)$$

در واقع، اگر B وارونپذیر نباشد، آنگاه طبق قضیهٔ ۴.۴، یك n-برداری غیرصفر $\mathbf{x}$ وجود دارد، به‌طوری که $B\mathbf{x} = \mathbf{o}$. اما داریم

$$\|A-B\|\,\|\mathbf{x}\| \geqslant \|(A-B)\mathbf{x}\| = \|A\mathbf{x}-B\mathbf{x}\| = \|A\mathbf{x}\| \geqslant \frac{\|\mathbf{x}\|}{\|A^{-1}\|}$$

با به‌کار گرفتن (۳۸.۴)، و توجه به $\mathbf{x} \neq \mathbf{0}$، می‌توانیم رابطهٔ بالا را بر $\|\mathbf{x}\| \neq 0$ تقسیم کنیم تا (۴۴.۴) به‌دست آید.

برهانی که هم‌اکنون آوردیم، اثبات فرع مهم زیراست

فرع: اگر A وارونپذیر و B ماتریسی باشد که

$$\|A-B\| < \frac{1}{\|A^{-1}\|}$$

آنگاه B نیز وارونپذیر است.

مثلا درمثال ۵.۴، برای ماتریس

$$A=\begin{bmatrix} ۱٫۰۱ & ۰٫۹۹ \\ ۰٫۹۹ & ۱٫۰۱ \end{bmatrix}$$

پیدا می‌کنیم که $\|A^{-1}\|_\infty \geqslant ۱/۰٫۰۲ = ۵۰$، زیرا که ماتریس $B=\begin{bmatrix} ۱ & ۱ \\ ۱ & ۱ \end{bmatrix}$ وارونپذیر نیست و

$$A-B=\begin{bmatrix} ۰٫۰۱ & -۰٫۰۱ \\ -۰٫۰۱ & ۰٫۰۱ \end{bmatrix}$$

دارای نرم ماکسیمم $\|A-B\|_\infty = ۰٫۰۲$ می‌باشد. بنابراین، چون $\|A\|_\infty = ۲$، خواهیم داشت $\operatorname{cond}(A) \geqslant ۱۰۰$. ماتریس مثلثی وارون‌ناپذیر مثال دیگری است از آنها. اگر A مثلثی باشد با توجه به‌قضیهٔ ۶.۴ می‌دانیم که درایه‌های قطری A غیرصفرند، وقراردادن صفر به‌جای هریك ازدرایه‌های قطری A، ماتریس A را وارونپذیر می‌سازد. درنتیجه اگر A مثلثی باشد، آنگاه

$$\operatorname{cond}(A) \geqslant \frac{\|A\|_\infty}{\min\limits_i |a_{ii}|}$$

عدد شرط در تحلیل مشکلات بعدی حل دستگاه معادلات خطی نیز نقش مهمی‌دارد اگردستگاه خطی $A\mathbf{x}=\mathbf{b}$ از یك مسئلهٔ عملی نتیجه شده باشد، بایـد انتظار داشته باشیم که ضرایب این‌دستگاه درمعرض خطا واقع شده باشند، خواه بدین‌علت که نتیجهٔ محاسبات دیگر ویااندازه‌گیری‌فیزیکی هستند ویاحتی به‌علت خطای گردکردن ازتبدیل‌اعداد دهدهی به‌دودویی طی خواندن اعداد ناشی شده‌اند. بنابراین موقتاً فرض می‌کنیم که سمت راست

معادله عاری از خطا باشد و درحقیقت می‌خواهیم دستگاه خطی

$$\hat{A}\hat{\mathbf{x}}=\mathbf{b} \qquad (۴۵.۴)$$

را به‌جای $A\mathbf{x}=\mathbf{b}$ حل کنیم، دراینجا $A=\hat{A}+E$ و ماتریس E شامل خطاهای موجود در ضرایب‌است. حتی اگر تمامی محاسبات دقیق انجام‌پذیرند، باز ما فقط جواب $\hat{\mathbf{x}}$ را از رابطهٔ (۴۵.۴) به‌جای جواب $\mathbf{x}$ از $A\mathbf{x}=\mathbf{b}$ محاسبه کرده‌ایم. امـا داریم $\mathbf{x}=A^{-1}\mathbf{b}$. بنابراین با فرض اینکه (۴۵.۴) دارای جواب باشد، داریم

$$\mathbf{x}=A^{-1}\mathbf{b}=A^{-1}\hat{A}\hat{\mathbf{x}}=A^{-1}(A+\hat{A}-A)\hat{\mathbf{x}}=\hat{\mathbf{x}}+A^{-1}(\hat{A}-A)\hat{\mathbf{x}}$$

بنابراین با توجه به $\hat{A}-A=-E$ داریم:

$$\mathbf{x}-\hat{\mathbf{x}}=A^{-1}(-E)\hat{\mathbf{x}}$$

بنابراین

$$\|\mathbf{x}-\hat{\mathbf{x}}\|\leqslant\|A^{-1}\|\,\|E\|\,\|\hat{\mathbf{x}}\|=\|A^{-1}\|\,\|A\|\frac{\|E\|}{\|A\|}\|\hat{\mathbf{x}}\|$$

که نتیجهٔ نهایی زیر را به‌دست می‌دهد

$$\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|}{\|\hat{\mathbf{x}}\|}\leqslant\operatorname{cond}(A)\,\frac{\|E\|}{\|A\|} \qquad (۴۶.۴)$$

به‌بیان لفظی، نسبت تفاضل $\mathbf{x}$ و $\hat{\mathbf{x}}$ بر $\|\hat{\mathbf{x}}\|$ می‌تواند به‌بزرگی $\operatorname{cond}(A)$ ضربدر تغییر نسبی $\|E\|/\|A\|$ درماتریس ضرایب باشد. اگربدانیم که ضرایب دستگاه خطی $A\mathbf{x}=\mathbf{b}$ فقط تـا حدود 10^{-5} (نسبت به‌اندازه A) دقیق‌اند و $\operatorname{cond}(A)\approx 10^{5}$، آنگاه محاسبهٔ جواب با دقت نسبی بیش‌از 10^{5-5} مفید فایده‌ای نخواهد بود.

□ **مثال۴.۷:** باردیگر دستگاه خطی(۳۰.۴) در مثال ۵.۴، را بررسی می‌کنیم. قبلا دیدیم کـه برای ماتریس‌ضرایبش داریم $\operatorname{cond}(A)=100$. طبق رابطـهٔ (۴۶.۴) یك درصد تغییر در ضرایب دستگاه می‌تواند جواب را به‌طور قابل‌ملاحظه‌ای تغییردهد. درحقیقت با یك تغییر یك درصد (درجهت صحیح) دستگاه زیر را پدید می‌آوریم

$$x_1+x_2=2$$

$$x_1+x_2=-2$$

که اصلا جواب ندارد زیراکه ماتریس ضرایب دراین‌حال وارونپذیر نخواهد بود. □

تحلیل پیشین به‌کمك تحلیل خطای پسرو در سنجش‌اثرات‌ناشی ازخطای گردکردن،

که درطی عملیات حذف و پسجایگذاری در دقت جواب محاسبه شده حاصل می‌شود، می‌تواند مفید باشد. در این مورد، از نشانگذاری و اصطلاحات و قراردادهای مربوط به‌بخش ۳.۱، استفاده خواهیم نمود.

قضیهٔ ۴.۹ فرض کنید که برای دستیابی به‌عوامل تجزیه PLU متعلق به‌ماتریس A ازمرتبهٔ nام و ازآنجا، رسیدن به‌حل دستگاه خطی $A\mathbf{x}=\mathbf{b}$، ازالگوریتمهای ۲.۴ و ۴.۴ استفاده کنیم، اما برای به‌دست‌آوردن عوامل حساب شدهٔ P، $\hat{L}$، و $\hat{U}$ و جواب حساب شدهٔ $\hat{\mathbf{x}}$، به‌حساب ممیز شناور همراه باگردکردن واحد[1] $u\leqslant$ ۰٫۰۱ متوسل شویم. دراین‌صورت $\hat{\mathbf{x}}$ دقیقاً در معادلهٔ پریشیده[2]

$$(A+PE)\hat{\mathbf{x}}=\mathbf{b} \qquad (۴۷.۴)$$

که درآن

$$|E|\leqslant u_n|P^{-1}A|+u_n(3+u_n)|\hat{L}|\,|\hat{U}| \qquad (۴۸.۴)$$

و

$$u_n:=n(1.01)u$$

صدق خواهدکرد.

در اینجا ماتریسی راکه از $B=(b_{ij})$ با قراردادن قدرمطلق هر درایه به‌جای خود آن درایه، حاصل می‌شود با $|B|$ نشان می‌دهیم، یعنی

$$|B|=(|b_{ij}|)$$

همچنین برای معرفی دوماتریس B و C درحالتی‌که B و C هم‌مرتبه باشند و داشته باشیم

$$b_{ij}\leqslant c_{ij} \qquad \text{به‌ازای جمیع مقادیر } i \text{ و } j$$

قرارداد زیر را اختیار می‌کنیم

$$B\leqslant C$$

قضیهٔ فوق بیان می‌کندکه هرگاه n «خیلی بزرگ» نباشد و اگراندازهٔ $|\hat{L}|\,|\hat{U}|$ در حدود اندازهٔ $|A|$ باشد، آنگاه می‌توانیم دنبال خطاهای حاصل درجواب را از راه تنظیم معادلاتی بگردیم، که مرتبهٔ بزرگی آنها با مرتبهٔ تغییراتی‌که فقط برای دادن معادلات به‌ماشین‌مجبوریم انجام دهیم، یکی‌باشند. به‌عبارت دیگر، خطای ناشی ازاستفاده ازحساب با ممیز شناور در جواب احتساب شده، بدتر از خطایی نیست‌که ما درآغاز به‌علت اجبار درقبول‌گردکردن درایه‌های A به‌صورت اعدادی با ممیز شناور، داشته‌ایم.

---

1. unit round-off 2. perturbed equation

البته، اگر ماتریس $|\hat{L}||\hat{U}|$، خیلی بزرگتر از $|A|$ باشد، خطاهـا در ریشهٔ محاسبه شدهٔ $\mathbf{x}$، ممکن‌است خیلی بزرگتر از خطاهای ناشی از تبدیل مسئله به‌اعداد ماشینی با ممیز شناور باشند. توجه کنید که وقتی کران حاصل درماتریس پریش E بیش از تحملی باشد که در آن درایه‌های A دقیق شناخته می‌شوند، می‌توان ماتریس $|\hat{L}||\hat{U}|$ را (بـا صرف هزینه) عملا حساب نمود و به‌دقت حسابی خیلی بالایی دست یافت. اما مهمتر از همه آنکه چون مرتبهٔ لولا ممکن‌است تأثیر اساسی براندازهٔ $|\hat{L}||\hat{U}|$ بگذارد، از قضیهٔ ۹.۲ این‌نتیجهٔ مهم را استخراج می‌کنیم که تـدبیـر لولاگزینی می‌باید معطوف به‌کوچك نگهداشتن $|\hat{L}|\,|\hat{U}|$ باشد.

اکنون با استفاده از نمادگـذاری و اصطلاحات بخش ۳.۱، برهان سادهٔ قضیهٔ ۹.۴ را ذکر می‌کنیم. ابتدا به‌تعریفهایی که باکاربرد الگوریتم ۲.۴ (به‌گونه‌ای کـه در ماتریس جایگشت P ذکر شد) بدون تعویض درماتریس $A' := P^{-1}A$ صورت می‌گیرد، می‌پردازیم (همان‌کاری راکه در بخش ۴.۴ کردیم). بنابراین درایه‌های جالب عوامل L و U را برطبق (۲۳.۴) به‌وسیلهٔ روابط زیرمحاسبه می‌کنیم

$$l_{ij} = (a'_{ij} - \sum_{k<j} l_{ik}u_{kj})/u_{jj} \qquad i > j$$

$$u_{ij} = a'_{ij} - \sum_{k<i} l_{ik}u_{kj} \qquad i \leqslant j$$

درنتیجـه بنابربخش ۳.۱، به‌ویـژه از مقایسهٔ (۱۲.۱) با (۱۳.۱)، درایه‌های $\hat{l}_{ij}$ و $\hat{u}_{ij}$ از عوامل $\hat{L}$ و $\hat{U}$ که براساس‌حساب ممیزشناور محاسبه شده‌اند درمعادله‌های پریشیدهٔ زیرصدق می‌کنند

$$\hat{l}_{i1}\hat{u}_{1j}\varepsilon^{j} + \hat{l}_{i2}\hat{u}_{2j}\varepsilon^{j-1} + \cdots + \hat{l}_{ij}\hat{u}_{jj}\varepsilon = a'_{ij}\varepsilon^{j-1} \qquad i > j$$

$$\hat{l}_{i1}\hat{u}_{1j}\varepsilon^{i-1} + \hat{l}_{i2}\hat{u}_{2j}\varepsilon^{i-2} + \cdots + \hat{l}_{ii}\hat{u}_{ij} = a'_{ij}\varepsilon^{i-1} \qquad i \leqslant j$$

دراینجا هر ε بیانگرعددی به‌فرم $(1+\delta)$، به‌ازای $|\delta| \leqslant u$، گردکردن واحد، است. برای ساده‌کردن این معادلات می‌بینیم مادامی که نامساوی $u \leqslant 0.01$ برقراراست، به‌ازای هر عددی مانند ε و برای هر r، δای به‌صورت $|\delta| \leqslant (1.01)u$ وجود دارد، به‌طوری که

$$\varepsilon^{r} = 1 + r\delta$$

این رابطه نشان می‌دهد که

$$\sum_{k} \hat{l}_{ik}\hat{u}_{kj} - a'_{ij}$$

$$= \begin{cases} a'_{ij}(j-1)\delta - \hat{l}_{i1}\hat{u}_{1j}j\delta - \cdots - \hat{l}_{ij}\hat{u}_{jj} \times 1 \times \delta & i > j \\ a'_{ij}(i-1)\delta - \hat{l}_{i1}\hat{u}_{1j}(i-1)\delta - \cdots - \hat{l}_{ii}\hat{u}_{ij} \times 0 \times \delta & i \leqslant j \end{cases}$$

و بنابراین

$$\hat{L}\hat{U}-A'=F \qquad (۴۹.۴)$$

که در آن

$$|F|\leqslant u_n(|A'|+|\hat{L}||\hat{U}|) \qquad (۵۰.۴)$$

و $u_n=n(۱۰۱)u$.

این نشان می‌دهد که عوامل محاسبه شدهٔ $\hat{L}$ و $\hat{U}$ برای A'، درست همان عوامل ماتریس پریشیدهٔ $A'+F$ هستند، با ماتریس خطای F که مرتبهٔ گردکردن آن همان مرتبهٔ گردکردن درایه‌های A است به‌شرط آنکه ماتریس $|\hat{L}||\hat{U}|$ خیلی بزرگتر از $|A|$ نباشد.

مراحل محاسبه‌ای که در الگوریتم ۴.۴، یعنی درمرحلهٔ حل کردن، مورد استفاده قرار گرفتند تا حدی شبیه، به آن مراحلی هستند که در بالا بیان شد. بنابراین می‌توان نشان داد که بردار محاسبه شدهٔ $\hat{\mathbf{y}}$ دقیقاً در دستگاه پایین مثلثی پریشیدهٔ زیرصدق می‌کند

$$|G|\leqslant u_n|\hat{L}| \quad \text{با} \quad (\hat{L}+G)\hat{\mathbf{y}}=\mathbf{b}$$

درحالی که جواب محاسبه شدهٔ $\hat{\mathbf{x}}$ دقیقاً در دستگاه خطی پریشیدهٔ زیرصدق می‌کند

$$|H|\leqslant u_n|\hat{U}| \quad \text{با} \quad (\hat{U}+H)\hat{\mathbf{x}}=\hat{\mathbf{y}}$$

از آنجا نتیجه می‌شود که جواب محاسبه شده در رابطهٔ زیرصدق می‌کند

$$(\hat{L}+G)(\hat{U}+H)\hat{\mathbf{x}}=\mathbf{b}$$

ولی داریم

$$\begin{aligned}(\hat{L}+G)(\hat{U}+H)&=\hat{L}\hat{U}+G\hat{U}+\hat{L}H+GH\\&=A'+F+G\hat{U}+\hat{L}H+GH\\&=A'+E\end{aligned}$$

که

$$\begin{aligned}|E|&\leqslant|F|+|G||\hat{U}|+|\hat{L}||H|+|G||H|\\&\leqslant u_n(|A'|+|\hat{L}||\hat{U}|)+u_n|\hat{L}||\hat{U}|(۱+۱+u_n)\end{aligned}$$

که قضیه را ثابت می‌کند.

کران رابطهٔ (۴۸.۴) محفوظ نگاه داشته شده‌است. اگر از لولاگزینی جزئی[1] استفاده شود، آنگاه کران

$$|E|\leqslant nu|P^{-۱}A| \qquad (۵۰.۴ \text{ الف})$$

---

1. partial pivoting

غالباً واقعگرایانه‌تر خواهد شد. درهرحال، این گونه کرانها دید روشنی از نتیجهٔ دقتی که در محاسبات جواب حساب شده صورت گرفته، به‌ما می‌دهند، زیرا که، برای مثال، باتوجه به (۴۶.۴) و (۵۰.۴) درمی‌یابیم که خطای جواب محاسبه شده نسبت به‌اندازهٔ این جواب معمولاً درمحدودهٔ زیر واقع است:

$$\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|}{\|\hat{\mathbf{x}}\|} \leqslant \operatorname{cond}(A) \cdot n \cdot u \qquad (۵۱.۴)$$

اگردقت راکاملا کنار بگذاریم، دستگاه خطی $A\mathbf{x}=\mathbf{b}$ غالباً **بدشرط** نامیده می‌شود اگر $\operatorname{cnod}(A)$ «بزرگ» باشد. تا حدی نزدیکتر به‌واقعیت، دستگاه خطی را نسبت به‌دقت **عمل مصروفه** «**بدشرط**» گویند اگر $\operatorname{cond}(A)$ تقریباً $1/u$ باشد، زیرا بنابر(۵۱.۴) یك جواب محاسبه شده ممکن‌است اصلا شباهتی به‌جواب (دقیق) نداشته باشد.

□ **مثال ۸.۴:** دستگاه خطی زیر را درنظرمی گیریم.

$$\begin{aligned} ۰\text{ر}۲۴x_۱ + ۰\text{ر}۳۶x_۲ + ۰\text{ر}۱۲x_۳ &= ۰\text{ر}۸۴ \\ ۰\text{ر}۱۲x_۱ + ۰\text{ر}۱۶x_۲ + ۰\text{ر}۲۴x_۳ &= ۰\text{ر}۵۲ \\ ۰\text{ر}۱۵x_۱ + ۰\text{ر}۲۱x_۲ + ۰\text{ر}۲۵x_۳ &= ۰\text{ر}۶۴ \end{aligned} \qquad (۵۲.۴)$$

سعی می‌کنیم این دستگاه را با استفاده از الگوریتم حذفی ۲.۴، حساب ممیزشناور با دو رقم اعشاری و لولا گزینی جـزیـی مدرج، حل کنیم. ترتیب گزینش لولا نتیجه می‌دهد که داشته باشیم $\mathbf{p}^T=[۳\ \ ۲\ \ ۱]$ و محتوای نهایی آرایهٔ‌کاری به‌شرح زیراست

$$\left[\begin{array}{cccc} ۰\text{ر}۲۴ & ۰\text{ر}۳۶ & ۰\text{ر}۱۲ & ۰\text{ر}۸۴ \\ \hline ۰\text{ر}۵۰ & -۰\text{ر}۰۲ & ۰\text{ر}۱۸ & ۰\text{ر}۱۰ \\ ۰\text{ر}۶۳ & ۱\text{ر}۰ & -۰\text{ر}۰۱ & ۰\text{ر}۰۱ \end{array}\right]$$

اگر محاسبات را ادامـه دهیم واز راه پسجایگذاری به‌جواب تقریبی $\hat{\mathbf{x}}=\begin{bmatrix} ۲۵ \\ -۱۴ \\ -۱ \end{bmatrix}$ و باقیمانده $\mathbf{r}=\begin{bmatrix} ۰\text{ر}۰ \\ ۰\text{ر}۰ \\ ۰\text{ر}۰۸ \end{bmatrix}$ می‌رسیم. زیرا $\mathbf{x}=\begin{bmatrix} -۳ \\ ۴ \\ ۱ \end{bmatrix}$ جواب دستگاه‌است، لذا جواب محاسبه شدهٔ $\hat{\mathbf{x}}$ در اولین رقم با معنی اشتباه‌است.

نرم ماکسیمم برای مـاتریس ضرایب Aی مربـوط به‌این دستگاه عبارت است از

$\|A\|_\infty = ۰ر۷۲$. بعلاوه، ماتریس

$$B = \begin{bmatrix} ۰ر۲۵۲ & ۰ر۳۶ & ۰ر۱۲ \\ ۰ر۱۱۲ & ۰ر۱۶ & ۰ر۲۴ \\ ۰ر۱۴۷ & ۰ر۲۱ & ۰ر۲۵ \end{bmatrix}$$

وارون‌ناپذیراست (ستون اول از حاصلضرب ۰ر۷ درستون دوم به‌دست آمده‌است) در حالی‌که $\|A-B\|_\infty = ۰ر۰۱۲$. بنابراین از قضیهٔ ۸.۴ نتیجه می‌گیریم که

$$\operatorname{cond}(A) \geqslant \frac{۰ر۷۲}{۰ر۰۱۲} \geqslant ۶۰$$

بنابراین دستگاه فوق، نسبت به‌دقت عملی‌که به‌کار رفته، بسیار بدشرط‌است و خطای‌بسیار بزرگ موجود درجواب محاسبه شده شگفت‌آور نیست.

سپس،محاسبات‌را با استفاده‌ازحساب ممیز شناور با سه رقم اعشاری‌تکرار می‌کنیم. از آنجاکه $\operatorname{cond}(A) \approx ۶۰$، باز انتظارنداریم‌که جواب حساب شده خیلی‌دقیق باشد. به‌موجب الگوریتم ۲.۴، محتوای ماتریس‌کار به‌شرح زیراست

$$\left[\begin{array}{c|ccc} ۰ر۲۴ & ۰ر۳۶ & ۰ر۱۲ & ۰ر۸۴ \\ \hline ۰ر۵۰ & ۰ر۰۲- & ۰ر۱۸ & ۰ر۱۰ \\ ۰ر۶۲۵ & ۰ر۷۵ & ۰ر۰۴ & ۰ر۰۴ \end{array}\right] \qquad (۵۳.۴)$$

و روش پسجایگذاری جواب محاسبه شدهٔ $\hat{\mathbf{x}}^T = [۱ \quad ۴ \quad -۳]$ را به‌دست می‌دهد؛ یعنی، حتی با اینکه دستگاه هنوز نسبت به‌دقت به‌کار گرفته شده اندکی بدشرط است، اما جواب (دقیق) را می‌دهد. این امر از تبدیل سمت راست (۵۲.۴) به $\mathbf{b} = \begin{bmatrix} ۰ر۸۵۲ \\ ۰ر۶۲۰ \\ ۰ر۷۴۰ \end{bmatrix}$ آشکار می‌شود. با به‌کارگرفتن تجزیه به‌عوامل (۵۳.۴) و با توجه به‌الگوریتم (۴.۴) جواب (تقریبی) $\hat{\mathbf{x}} = \begin{bmatrix} -۳ر۳۰ \\ ۴ر۰۵ \\ ۱ر۵۳ \end{bmatrix}$ را محاسبه می‌کنیم(بازازحساب‌باممیزشناور، باسه‌رقم اعشاری، استفاده می‌کنیم)که $\mathbf{r} = \begin{bmatrix} ۰ر۰۰۲۴ \\ ۰ر۰۰۰۸ \\ ۰ر۰۰۲۰ \end{bmatrix}$ باقیمانده آن است. جواب دقیق

برابر است با $\mathbf{x}=\begin{bmatrix}-۳٫۶\\ ۴٫۲۵\\ ۱٫۵۵\end{bmatrix}$، وجواب محاسبه شده دارای ۱۰ درصد خطاست که با (۵۱.۴) سازگار است. □

چنانچه، مثال فوق نشان می‌دهد، یك عدد با شرط بزرگ نسبت به‌دقتی که به‌کاررفته ممکن‌است به‌خطای نسبتاً بزرگی درجواب محاسبه شده منجرشود، اما همیشه الزاماً چنین نیست.

این مطلب راکه‌آیا یك دستگاه خطی نسبت به‌دقتی که به‌کارگرفته شده بدشرط‌است یا نیست، می‌توان به‌سادگی [حتی بدون اطلاع از $\text{cond}(A)$] درطی **اصلاح بارستی**[1] که درزیر ازآن بحث می‌شود، تعیین کرد. باتوجه به $\mathbf{e}=\mathbf{x}-\hat{\mathbf{x}}^{(1)}$، خطا (مجهول)درجواب تقریبی $\hat{\mathbf{x}}^{(1)}$ برای $A\mathbf{x}=\mathbf{b}$ را دربخش۵.۴ پیداکردیم و دیدیم که

$$A\mathbf{e}=\mathbf{r} \qquad (۵۴.۴)$$

که درآن $\mathbf{r}=\mathbf{b}-A\hat{\mathbf{x}}^{(1)}$ باقیمانده قابل‌محاسبه‌ای برای $\hat{\mathbf{x}}^{(1)}$ است. پس دراینجا یك دستگاه خطی‌داریم که جواب‌آن خطای $\mathbf{e}$ و ماتریس‌ضرایب آن با ماتریس‌ضرایب دستگاه اولیه یکی‌است. اگر $\hat{\mathbf{x}}^{(1)}$ از راه الگوریتم حذف ۲.۴ به‌دست آمده باشد، می‌توانیم (۵۴.۴) را با الگوریتم جایگذاری ۴.۴ نسبتاً سریع حل‌کنیم. گیریم $\hat{\mathbf{e}}^{(1)}$ یك جواب (تقریبی) برای(۵۴.۴) باشدکه بدین‌طریق محاسبه شده‌است. دراین‌صورت $\hat{\mathbf{e}}^{(1)}$ درحالت کلی با $\mathbf{e}$ یکی نخواهدبود. اما $\hat{\mathbf{e}}^{(1)}$، حداقل هیچ نباشد، می‌بایستی نشانه‌ای از اندازهٔ $\mathbf{e}$ به‌دست دهد. اگر $||\hat{\mathbf{e}}^{(1)}||/||\hat{\mathbf{x}}^{(1)}||\approx ۱۰^{-s}$، نتیجه می‌گیریم که احتمالاً اولین s رقم اعشاری $\hat{\mathbf{x}}^{(1)}$ واولین s رقم‌اعشاری$\mathbf{x}$ باهم یکی‌هستند. پس، انتظار داریم‌که تقریب $\hat{\mathbf{e}}^{(1)}$ درهمان اندازهٔ دقت $\mathbf{e}$ باشد. بنابراین انتظار داریم که

$$\hat{\mathbf{x}}^{(۲)}=\hat{\mathbf{x}}^{(1)}+\hat{\mathbf{e}}^{(1)}$$

تقریبی بهتر از $\hat{\mathbf{x}}^{(1)}$ برای $\mathbf{x}$ باشد. اکنون می‌توانیم، درصورت لزوم، باقیماندهٔ جدید $\mathbf{r}=\mathbf{b}-A\hat{\mathbf{x}}^{(۲)}$ راحساب و (۵۴.۴) را دوباره حل‌کنیم، و تصحیح جدید $\hat{\mathbf{e}}^{(۲)}$ وتقریب جدید $\hat{\mathbf{e}}^{(۲)}+\hat{\mathbf{e}}^{(۲)}=\hat{\mathbf{x}}^{(۳)}$را برای $\mathbf{x}$ به‌دست آوریم. تعداد ارقام اعشاری یکسان در تقریبهای متوالی $\hat{\mathbf{x}}^{(1)}$، $\hat{\mathbf{x}}^{(۲)}$، ... و همچنین بررسی باقیمانده‌های متوالی بایستی قرینه‌ای از صحت این جوابهای تقریبی به دست دهند. معمولا این بارستها تا زمان برقراری شرط $||\hat{\mathbf{e}}^{(k)}||/||\hat{\mathbf{x}}^{(k)}||\approx ۱۰^{-t}$، که t تعداد ارقام اعشاریی‌است که‌درمحاسبات به‌کارگرفته‌شده، ادامه می‌یابد. می‌توان نشان دادکه تعداد مراحل بارستی‌لازم برای حصول این نتیجه با $\text{cond}(A)$ افزایش می‌یابد. هرگاه $\text{cond}(A)$ «خیلی بزرگ» باشد تصحیحهای $\hat{\mathbf{e}}^{(1)}$ و

1. iterative improvement

$\hat{e}^{(2)}$ و . . . ممکن‌است هرگز کوچك نشوند، و لذا معرف حد اعلای بدشرطی در دستگاه اولیه باشند.

برای توفیق در روش اصلاح بارستی، شرط اساسی این است که باقیمانده با دقت هرچه بیشترمحاسبه شود. اگر، چنان‌که متداول‌است، حساب با ممیز شناور مورد استفاده قرار گیرد، می‌بایستی باقیمانده از راه حساب با دقت مضاعف، محاسبه شود.

الگوریتم ۵.۴ اصلاح بارستی

دستگاه خطی $A\mathbf{x}=\mathbf{b}$ وجواب تقریبی $\hat{\mathbf{x}}$ داده شده‌اند.

با استفاده از حساب با دقت مضاعف، $\mathbf{r}=\mathbf{b}-A\hat{\mathbf{x}}$ را محاسبه کنید.

با استفاده از الگـوریتم ۲.۴ (یا درصورت امکان فقط از الگوریتم ۴.۴) جواب تقریبی $\hat{\mathbf{e}}$ از دستگاه خطی $A\mathbf{e}=\mathbf{r}$ را محاسبه کنید.

اگر $||\hat{\mathbf{e}}||/||\hat{\mathbf{x}}||$ «به‌اندازهٔ کافی کوچك» باشد محاسبات را متوقف کنید و $\hat{\mathbf{x}}+\hat{\mathbf{e}}$را به‌عنوان جواب بگیرید. درغیراین‌صورت قراردهید $\hat{\mathbf{x}}:=\hat{\mathbf{x}}+\hat{\mathbf{e}}$ وعملیات بالارا تکرار کنید.

هنگامی می‌توان از اصلاح بارستی استفاده‌کردکـه یك جواب تقریبی به‌هر طریقی پیدا شده باشد. اصلاح بارستی را همیشه بایستی پس از محاسبهٔ یك جواب تقریبی ازراه حذف به‌کار گرفت زیرا کـه تصحیحها را می‌توان از راه پیشجایگذاری و پسجایگذاری نسبتاً ارزان محاسبه‌کرد. همچنین میزان همگرایی عملیات (درصورت همگرابودن) نشانهٔ خوبی از شرط دستگاه (نسبت به‌دقت عمل به‌کار گرفته شده)است.

□ **مثال۹.۴:** برای جواب تقریبی دستگاه (۵۲.۴) که در مثال ۸.۴ محاسبه شد، روش اصلاح بارستی را به‌کار می‌گیریم. باقیماندهٔ محاسبه شده،که تا دو رقم اعشاری صحیح و گرد شده، برابراست با $\mathbf{r}=\begin{bmatrix}0.0\\0.0\\0.08\end{bmatrix}$. با اعمال الگوریتم ۴.۴ براین‌سمت راست (بـا استفاده از حساب با ممیز شناور با دو رقم اعشاری) تصحیح $\hat{\mathbf{e}}^{(1)}=\begin{bmatrix}120\\-75\\-8\end{bmatrix}$ به‌دست می‌آیدکه اندازهٔ آن بااندازهٔ جواب محاسبه شده یکی است. بنابراین نتیجه می‌گیریم که دستگاه خطی داده شده نسبت به‌دقتی‌که به‌کار رفته بسیار بدشرط‌است و دقت عمل بیشتری باید به‌کار گرفت تا بتوان جواب (۵۲.۴) را محاسبه‌کرد.

درمثال ۸.۴ یك جواب تقریبی $\hat{\mathbf{x}}=\begin{bmatrix}-۳ر۳۰\\+۴ر۰۵\\۱ر۵۳\end{bmatrix}$ برای دستگاه خطی با همان ماتریس ضرایب، اما با سمت راست متفاوت را با استفاده از حساب با ممیز شناور با سه رقم اعشاری محاسبه کردیم. باقیماندهٔ محاسبه شدهٔ صحیح چنین‌است $\mathbf{r}=\begin{bmatrix}۰ر۰۰۲۴\\۰ر۰۰۰۸\\۰ر۰۰۲۰\end{bmatrix}$.

باکاربرد الگوریتم ۴.۴ برای این سمت راست $\mathbf{r}$ (و با به‌کار بستن همان دقت عمل قبلی) تصحیح $\hat{\mathbf{e}}^{(1)}=\begin{bmatrix}-۰ر۳\\۰ر۲\\۰ر۰۲\end{bmatrix}$ نتیجه می‌شود که فقط ۱۰٪ جواب‌محاسبه شده‌است وجواب تصحیح‌شدهٔ $\hat{\mathbf{x}}^{(2)}=\begin{bmatrix}-۳ر۶\\۴ر۲۵\\۱ر۵۵\end{bmatrix}$ رانتیجه می‌دهد. باقیمانده برای این‌جواب تقریبی ۰ درمی‌آید، لذا درست یك مرحله‌از روش‌اصلاح بارستی، جواب (دقیق)این مثال‌را به‌دست می‌دهد. □

تمرین

۱-۶.۴ با استفاده از قضیهٔ ۸.۴ عدد شرط ماتریس زیر را برآوردکنید.

$$A=\begin{bmatrix}۷ & ۸ & ۹\\۸ & ۹ & ۱۰\\۹ & ۱۰ & ۸\end{bmatrix}$$

۲-۶.۴ از الگوریتم اصلاح بارستی در جواب محاسبه شدهٔ تمرین ۳.۴-۴ استفاده‌کنید.

۳-۶.۴ ماتریس A از مرتبهٔ n را **نافذ قطری**[1] (سطراً مؤکد) گوییم اگر به‌ازای $i=۱,\ldots,n$، نامساوی $|a_{ii}|>\Sigma_{j\neq i}|a_{ij}|$ برقرار باشد. با استفاده از فرع قضیهٔ ۸.۴ ثابت کنیدکه یك ماتریس «نافذ قطری» وارونپذیراست (راهنمایی: فرض کنید $A=DB$، و D یك ماتریس قطری باشدکه درایه‌های قطری‌اش با درایه‌های قطری ماتریس Aمساوی است، سپس نشان دهیدکه $\|I-B\|_\infty<۱$.)

---

1. diagonally dominant

۴-۶.۴ عدد شرط ماتریس تمرین ۶.۴-۱ را ازطریق حل دستگاه خطی $A\mathbf{x}=\mathbf{b}$ با (اولاً) $\mathbf{b}^T=[۲۴, ۲۷, ۲۷]$، (ثانیاً) $\mathbf{b}^T=[۲۶ر۹, ۲۶ر۹, ۲۴ر۱]$ برآورد کنید. ازروش اصلاح بارستی استفاده کنید.

۵-۶.۴ نشان دهید که در نرم ماتریس خاص (۳۹.۴)، برای برقراری رابطهٔ (۴۴.۴) ماتریس‌وارون‌ناپذیر B را می‌توان به‌گونهٔ زیربنا کرد: به‌موجب رابطهٔ (۳۵.۴)، می‌توان $\mathbf{x}$ی بانرم ۱ چنان تعیین نمود که تساوی $\|A^{-1}\mathbf{x}\|=\|A^{-1}\|\,\|\mathbf{x}\|$ برقرار باشد. حال B را برابر با ماتریس $A-\mathbf{x}\mathbf{z}^T$، که در آن $\mathbf{z}=y_m^{-1}\mathbf{i}_m$، $\mathbf{y}=A^{-1}\mathbf{x}$، بگیرید و m را چنان انتخاب کنید که $|y_m|=\|\mathbf{y}\|_\infty$.

۶-۶.۴ نشان دهید که مسئلهٔ ۶.۴-۵ را برای یك نرم کلی می‌توان حل کرد، به‌شرط آنکه بدانیم که به‌ازای یك n-برداری غیرصفر مفروض $\mathbf{y}$، چگونه یك n-برداری $\mathbf{z}$ را انتخاب کنیم، تا به‌ازای جمیع n-بردارهای $\mathbf{u}$، داشته باشیم $\mathbf{z}^T\mathbf{u}\leqslant\|\mathbf{u}\|$، تساوی زمانی برقرار است که $\mathbf{u}=\mathbf{y}$. درحالتی که نرم ازنوع ۱-نرم باشد، $\mathbf{z}$ را چگونه انتخاب می‌کند؟

۷.۴* دترمینانها

گرچه فرض براین است که دانشجویان با مفهوم دترمینانها آشنایی دارند، ولی مااین بخش را به‌تعریف رسمی دترمینانها اختصاص می‌دهیم و بعضی ازخواص اولیهٔ آنها را مورد بحث قرارمی‌دهیم.

به‌هرماتریس مربعی از اعداد مانند A، عددی نسبت داده می‌شود به‌نام دترمینان A که با $\det(A)$ نشان می‌دهند. اگر $A=(a_{ij})$ یك ماتریس $n\times n$ باشد، آنگاه **دترمینان** A چنین تعریف می‌شود

$$\det(A)=\sum_{\mathbf{p}}\sigma_{\mathbf{p}}a_{۱,p_۱}a_{۲,p_۲}\cdots a_{n,p_n} \qquad (۵۵.۴)$$

در رابطهٔ بالا مجموع روی تمامی $n!$ جایگشت $\mathbf{p}$ از مرتبهٔ n درنظر گرفته می‌شود و $\sigma_{\mathbf{p}}$ مساوی $+۱$ یا -۱ است بسته به اینکه $\mathbf{p}$ زوج یا فرد باشد (به‌بخش ۱.۴ نگاه کنید). بنابراین اگر $n=۱$، آنگاه داریم

$$\det(A)=\det[a_{۱۱}]=a_{۱۱}$$

درحالی که اگر $n=۲$، داریم

$$\det(A)=\det\begin{bmatrix}a_{۱۱} & a_{۱۲}\\ a_{۲۱} & a_{۲۲}\end{bmatrix}=a_{۱۱}a_{۲۲}-a_{۱۲}a_{۲۱} \qquad (۵۶.۴)$$

به‌ازای $n=۳$، شش حاصلضرب باید با هم جمع شوند، و به‌ازای $n=۱۰$ بیشتر از سه ملیون حاصلضرب، هریك با ۱۰ عامل، باید محاسبه و با هم جمع شوند تا سمت راست

(۵۵.۴) به‌دست آید. بنابراین، تعریف (۵۵.۴) برای محاسبهٔ دترمینانها چندان مفیدنیست، اما در زیر به‌ذکر فهرستی از قواعد مربوط به‌دترمینانها، که به‌آسانی از تعریف ۵۵.۴ به‌دست می‌آیند، می‌پردازیم. سپس نشان می‌دهیم که چگونه می‌توان با به‌کارگیری این قواعد و الگوریتم ۲.۴، مقدار دترمینان را حدوداً با $\Theta(n^3)$ [به‌جای $\Theta(n!)$] عمل محاسبه کرد.

دترمینان یك ماتریس به‌دلیل قضیهٔ زیر حائز اهمیت است.

قضیهٔ ۱۰.۴ گیریم A یك ماتریس $n \times n$ باشد در این صورت A وارونپذیر است اگر و فقط اگر، $\det(A) \neq 0$.

ما از این قضیه در بخش بعدی که به‌محاسبهٔ ویژه‌مقدارها و ویژه‌بردارها مربوط است استفاده خواهیم کرد.

در برخی از ماتریسها، دترمینان به‌آسانی قابل‌محاسبه است.

قاعدهٔ ۱ اگر $A = (a_{ij})$ یك ماتریس بالا(پایین) مثلثی باشد، آنگاه

$$\det(A) = a_{11}a_{22}\cdots a_{nn}$$

یعنی دترمینان درست برابر است با حاصلضرب درایه‌های قطری A.

زیرا مثلاً، اگر A ماتریس بالامثلثی و $\mathbf{p}$ هرجایگشتی غیرازجایگشت همانی[1] باشد، به‌ازای مقداری از i باید داشته باشیم $p_i < i$، و بنابراین حاصلضرب مربوطهٔ $a_{1,p_1}a_{2,p_2}\cdots a_{n,p_n}$ باید متضمن درایه‌های زیر قطری صفر a_{i,p_i} و در نتیجه، باید صفر باشد. بنابراین اگر A یك ماتریس بالا مثلثی باشد، آنگاه تنها عامل جمع موجود در (۵۵.۴) که صفر بودن آن را نمی‌توان تضمین کرد، جملهٔ $a_{11}a_{12}\cdots a_{nn}$ متناظر با جایگشت همانی (زوج) $\mathbf{p}^T = [1 \ 2 \ \cdots \ n]$ است.

به‌ویژه

$$\det(I) = 1 \qquad (۵۷.۴)$$

به‌طور مشابه می‌توان یك قاعدهٔ دیگر را ثابت کرد.

قاعدهٔ ۲. اگر P ماتریس جایگشتی $n \times n$ باشد که با جایگشت

$$P\mathbf{i}_j = \mathbf{i}_{p_j} \qquad j = 1, \cdots, n$$

داده شده‌است، آنگاه با جایگشتی مانند $\mathbf{p}$ داریم

$$\det(P) = \begin{cases} 1 & \text{اگر } \mathbf{p} \text{ زوج باشد} \\ -1 & \text{اگر } \mathbf{p} \text{ فرد باشد} \end{cases}$$

---

1. identity permutation

قاعدهٔ ۳. اگر ماتریس B از تعویض دوستون (سطر) A نتیجه شده باشد، آنگاه

$$\det(B) = -\det(A)$$

□ مثال:

□ $\det\begin{bmatrix} ۱ & ۴ \\ ۲ & ۳ \end{bmatrix} = ۳ - ۸ = -۵$ ، $\det\begin{bmatrix} ۴ & ۱ \\ ۳ & ۲ \end{bmatrix} = ۸ - ۳ = ۵$

در نتیجه، هرگاه دوستون (سطر) ماتریس A یکی باشند (به‌طوری‌که تعویض آنها با هم در A تغییری به‌وجود نیاورد)، آنگاه

$$\det(A) = ۰$$

قاعدهٔ ۴. اگر ماتریس B از ضرب تمامی درایه‌های یك ستون (سطر) ماتریس A در عدد α به‌دست آمده باشد، آنگاه

$$\det(B) = \alpha \det(A)$$

□ مثال:

□ $\det\begin{bmatrix} ۳\times۱ & ۴ \\ ۳\times۲ & ۳ \end{bmatrix} = ۹ - ۲۴ = -۱۵ = ۳(-۵) = ۳\det\begin{bmatrix} ۱ & ۴ \\ ۲ & ۳ \end{bmatrix}$

قاعدهٔ ۵. فرض کنید که سه ماتریس $A_۱$ و $A_۲$ و $A_۳$ از مرتبهٔ $n \times n$ فقط در یك ستون (سطر) مثلا ستون (سطر) j ام، باهم تفاوت داشته باشند و ستون (سطر) j ام $A_۳$ حاصلجمع برداری ستون (سطر) j ام $A_۱$ و ستون (سطر) j ام $A_۲$ باشد. دراین صورت داریم

$$\det(A_۱) + \det(A_۲) = \det(A_۳)$$

□ مثال:

$$\det\begin{bmatrix} ۱ & ۲ \\ ۲ & ۲ \end{bmatrix} + \det\begin{bmatrix} ۱ & ۲ \\ ۱ & ۱ \end{bmatrix} = (۲-۴) + (۱-۲) = ۳ - ۶ = \det\begin{bmatrix} ۱ & ۲ \\ ۳ & ۳ \end{bmatrix}$$

□

قواعد ۱ تا ۵ قضیه‌های ۱۱.۴ و ۱۲.۴ را ایجاب می‌کنند.

قضیهٔ ۱۱.۴ اگر A و B ماتریسهای $n \times n$ باشند، آنگاه

$$\det(AB) = \det(A)\det(B)$$

قضیهٔ ۱۲.۴ هرگاه A یك ماتریس $n \times n$ و $\mathbf{x}=(x_i)$ و $\mathbf{b}$ n- بردارهایی باشند به‌طوری که

$$A\mathbf{x}=\mathbf{b}$$

آنگاه به‌ازای $j=1,\ldots,n$ داریم

$$\det(A^{(j)})=x_j \det(A) \qquad (۵۸.۴)$$

که $A^{(j)}$ ماتریسی است که از قراردادن $\mathbf{b}$ به‌جای ستون j ام A به‌دست آمده‌است.

اگر A وارونپذیر باشد، یعنی اگر (بنابر قضیهٔ ۱۰.۴) $\det(A)\neq 0$، آنگاه می‌توان (۵۸.۴) را نسبت به x_j حل کرده، به‌دست آورد:

$$x_j=\frac{\det(A^{(j)})}{\det(A)} \qquad j=1,\ldots,n$$

رابطهٔ بالا **قاعدهٔ کرامر** برای درایه‌های جواب $\mathbf{x}$ از دستگاه خطی $A\mathbf{x}=\mathbf{b}$ است. به‌دلیل دشواری محاسبهٔ دترمینانها، درحالت کلی قاعدهٔ کرامر تنها از لحاظ نظری اهمیت دارد.

در حقیقت سریعترین راهی که برای محاسبهٔ $\det(A)$ی یك ماتریس دلخواه A از مرتبهٔ $n \times n$ دردست‌است، استفاده از الگوریتم حذف ۲.۴ برای ماتریس فوق (نادیده گرفتن سمت راست آن) است. درقسمت ۴.۴ دیدیم که این الگوریتم ماتریس A را به‌صورت

$$A=PLU$$

تجزیه می‌کند که P ماتریسی‌است جایگشتی که به‌توسط تدبیر لولاگزینی $\mathbf{P}$ معین می‌شود، L یك ماتریس پایین مثلثی است که تمامی درایه‌های قطری‌اش مساوی ۱ است، و بالاخره ماتریس بالامثلثی‌ضرایب $U=(u_{ij})$ است که تمامی‌درایه‌های لولا را روی قطرخود دارد. بنابر قاعدهٔ ۱، $\det(L)=1$، درحالی‌که بنابر قاعدهٔ ۲ $\det(P)$، بسته به‌اینکه $\mathbf{P}$ زوج‌یا فرد باشد، یعنی بسته به‌اینکه تعداد تعویضها درطی‌حذف زوج یا فرد باشد برابر ۱ یا ۱— است. سرانجام باز بنابرقاعدهٔ ۱ داریم $\det(U)=u_{11}u_{22}\cdots u_{nn}$. لذا

$$\det(A)=(-1)^i u_{11}u_{22}\cdots u_{nn} \qquad (۵۹.۴)$$

که i تعداد تعویضها طی‌اجرای الگوریتم‌حذف‌است. باید توجه داشت‌که برنامهٔ **فورترن** FACTOR (درحالتی‌که وارونپذیری A معلوم باشد) عدد $(-1)^i$ را به IFLAG بدل می‌کند، و از این‌رو محاسبهٔ $\det(A)$ را از طریق (۵۹.۴)، از روی درایه‌های قطری آرایهٔ‌کار W آسان می‌سازد.

البته الگوریتم حذف ۲.۴ (حداقل از لحاظ نظری) موقعی نتیجه بخش‌خواهد بود که A وارونپذیر باشد. اما اگر A وارونپذیر نباشد، الگوریتم فوق‌آن را نشان خواهد داد، دراین‌حال بنابرقضیهٔ ۱۰.۴ خواهیم داشت $\det(A)=0$.

و بالاخره، چنانچه ماتریس A دارای خصوصیات ویژه‌ای باشد، گاهی استفاده از قاعدهٔ زیرمفید واقع می‌شود.

قاعدهٔ ۶: بسط دترمینان به‌توسط کهادها[1] بنا بر تعریف، **کهاد** M_{ij} در یك ماتریس $A=(a_{ij})$ از مرتبهٔ $n\times n$ عبارت‌است از دترمینان ماتریس مرتبهٔ $n-1$ امی، که‌ازحذف سطر iام و ستون jام ماتریس A نتیجه می‌شود.

به‌ازای هر i داریم

$$\det(A)=a_{i1}(-1)^{i+1}M_{i1}+a_{i2}(-1)^{i+2}M_{i2}+\cdots+a_{in}(-1)^{i+n}M_{in}$$

و

به‌ازای هر j داریم

$$\det(A)=a_{1j}(-1)^{1+j}M_{1j}+a_{2j}(-1)^{2+j}M_{2j}+\cdots+a_{nj}(-1)^{n+j}M_{nj}$$

قاعدهٔ ۶ به‌ما این امکان را می‌دهدکه یك دترمینان مرتبهٔ n را به‌صورت مجموع دترمینانهایی ازمرتبهٔ $n-1$ بیان‌کنیم. با استفادهٔ تراجعی[2] ازاین‌قاعده، می‌توانیم‌سرانجام $\det(A)$ رابه‌صورت مجموع دترمینانهایی ازمرتبهٔ ۱ بنویسیم. این قاعده به‌ویژه هنگامی در محاسبهٔ $\det(A)$ مفید است‌که A یك ماتریس تنك باشد، که در این حالت بیشتر حاصلجمعها ازدورخارج می‌شوند. به‌عنوان مثال، ازبسط ماتریس زیر برحسب کهادها نسبت به‌سطر اول داریم:

$$\det\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix}=0\det\begin{bmatrix}0&0\\0&1\end{bmatrix}-1\det\begin{bmatrix}1&0\\0&1\end{bmatrix}+0\det\begin{bmatrix}1&0\\0&0\end{bmatrix}$$

$$=-1\det\begin{bmatrix}1&0\\0&1\end{bmatrix}=-1\times 1=-1$$

تمرین

۱-۷۰۴ با استفاده از قضیهٔ ۱۱.۴ و معادلهٔ (۵۷.۴) ثابت کنید که اگر A وارونپذیر باشد، آنگاه $\det(A)\neq 0$.

۲-۷۰۴ با استفاده از قضایای ۱۲.۴ و ۴.۴ ثابت کنیدکه اگر $\det(A)\neq 0$، آنگاه A وارونپذیراست.

۳-۷۰۴ تعداد عملیات حسابی لازم برای محاسبهٔ جواب یك دستگاه خطی مرتبهٔ ۲ را در

1. minors 2. Recursively

روشهای زیر تعیین کنید. (الف) در روش حذف و پسجایگذاری، (ب) در روش کرامر.

۴-۷.۴ اگر $n=۳$؛ آنگاه محاسبهٔ مستقیم (۵۵.۴) مستلزم ۱۲ عمل ضرب و ۵ عمل جمع خواهد بود. تعداد عملیات ضرب وجمع لازم برای محاسبهٔ یك دترمینان مرتبهٔ ۳ برحسب بسط به کهادها (قاعدهٔ ۶) چقدر خواهد بود؟ تعداد عملیات ضرب، یا تقسیم و جمع برای همان دترمینان با روش حذف چقدراست؟

۵-۷.۴ ثابت کنید: اگرماتریس ضرایب دستگاه خطی $A\mathbf{x}=\mathbf{b}$ وارونپذیرباشد، همیشه می‌توانیم (در صورت لزوم) معادلات را طوری مرتب کنیم که تمامی درایه‌های قطری ماتریس ضرایب دستگاه (هم ارز) حاصل، غیرصفر باشند. [راهنمایی: اگر A وارونپذیر باشد، بنابرقضیهٔ ۱۰.۴ حداقل یکی از عوامل جمع در (۵۵.۴) باید مخالف صفرباشد].

۶-۷.۴ درستی قواعد ۱ تا ۵ را در حالتی که ماتریسهای موردبحث از مرتبهٔ ۲ باشند، تحقیق نمایید. سعی کنید قواعد ۴ و ۵ را برای ماتریسهایی ازهرمرتبهٔ دلخواه ثابت کنید.

۷-۷.۴ قضیهٔ ۱۱.۴ را برای حالتی که A و B ماتریسهای مرتبهٔ ۲ باشند، ثابت کنید.

۸-۷.۴ گیریم A یك ماتریس سه‌قطری ازمرتبهٔ n باشد. به‌ازای $p=۱,۲,\dots,n$، گیریم A_p یك ماتریس $p\times p$، حاصل از حذف سطرهای $p+۱,\dots,n$ام وستونهای $p+۱,\dots,n$ام ماتریس A باشد. بااستفاده‌ازقاعدهٔ ۶ ثابت کنیدکه هرگاه $\det(A_۰)=۱$، آنگاه داریم

$$\det(A_p)=a_{pp}\det(A_{p-۱})-a_{p,\,p-۱}a_{p-۱,\,p}\det(A_{p-۲}),\quad p=۲,۳,\dots,n$$

برنامه‌ای بر اساس این فرمول تراجعی برای محاسبهٔ دترمینان ماتریس سه‌قطری بنویسید.

۸.۴* مسئلهٔ ویژه‌مقدارها

در بسیاری از مسائل فیزیکی، ویژه‌مقدارها اهمیت زیادی دارند. برای مثال، پایداری یك هواپیما باجای ویژه مقدارهای یك ماتریس درصفحهٔ همتافت[1] مشخص می‌گردد. بسامد[2] طبیعی ارتعاشات یك باریکهٔ نوری، در واقع ویژه‌مقدارهای یك ماتریس (متناهی) هستند. ویژه مقدارها به‌طور طبیعی درتحلیل بسیاری ازمسائل ریاضی نیز پدید می‌آیند، زیرا این کمیات به‌ویژه وسیلهٔ بسیار مناسب و روشنگری درنمایش یك ماتریس (در شکل قانونی جردن ودرشکلهای متشابه)هستند. به‌همین دلیل، هردستگاه معادلات دیفرانسیل خطی‌معمولی مرتبهٔ اول با ضرایب ثابت، می‌تواند برحسب ویژه‌مقادیر ماتریس ضرایبش حل شود، و همچنین رفتار دنبالهٔ توانهای A، $A^۲$، $A^۳$، ... در یك ماتریس A، برحسب ویژه‌مقدارهای آن، خیلی آسانتر مورد تحلیل قرار می‌گیرد. این گونه دنباله‌ها در حل بارستی دستگاه

1. complex plane 2. frequency

معادلات خطی (وغیرخطی) پدید می‌آیند.

بدین دلایل‌ودلایل‌دیگردر این بخش مقدمهٔ کوتاهی در جایابی[1] ومحاسبهٔ ویژه‌مقدارها می‌آوریم. متأسفانه، کیفیت فنی بیان آن خارج از محدودهٔ این‌کتاب است. کتاب دائرةالمعارفی ج. ه. ویلکینسن[2] [۲۴] وکتاب مقدماتیتر ج. و. استیوارت[3] [۲۳] منابع اطلاعاتی موجودی هستند که روشهای جدیدی مانند روش QR (با تغییرمکان) و بسیاری از مطالب مشروح حذف شده درصفحات بعد را دربر دارند.

عدد (حقیقی یا همتافت) λ را یك **ویژه‌مقدار ماتریس** B گوییم، هرگاه به‌ازای بردار (حقیقی یا همتافت) غیرصفری مانند $\mathbf{y}$ داشته باشیم

$$B\mathbf{y}=\lambda\mathbf{y} \qquad (۶۰.۴)$$

در این صورت n-برداری $\mathbf{y}$ یك **ویژه‌بردار** متعلق به‌ویژه‌مقدار λ نامیده می‌شود. فرمول (۶۰.۴) را می‌توانیم به‌شکل زیر بنویسیم

$$(B-\lambda I)\mathbf{y}=\mathbf{0} \qquad (۶۱.۴)$$

از آنجاکه $\mathbf{y}$ قراراست یك بردار غیرصفر باشد، ملاحظه می‌کنیم‌که λ یك ویژه‌مقدار B است اگر و فقط اگر، دستگاه همگن(۶۱.۴) دارای جوابهای غیرصفر باشد. بنابراین لم زیرنتیجه‌ای‌است از قضیهٔ ۴.۴.

لم ۴.۴ عدد λ یك ویژه‌مقدار ماتریس B است اگر و فقط اگر، $(B-\lambda I)$ وارونپذیر نباشد.

توجه شود که (۶۰.۴) و (۶۱.۴) ویژه‌بردار متعلق به λ را تنها بی‌مضارب عددی (اسکالر)، معین می‌کند. اگر $\mathbf{y}$ یك ویژه‌بردار متعلق به λ و $\mathbf{z}$ مضربی‌عددی از $\mathbf{y}$، مثلا $\mathbf{z}=\alpha\mathbf{y}$ باشد، آنگاه $\mathbf{z}$ نیز یك ویژه‌بردار متعلق به λ است، زیرا داریم،

$$B\mathbf{z}=B(\alpha\mathbf{y})=\alpha(B\mathbf{y})=\alpha(\lambda\mathbf{y})=\lambda(\alpha\mathbf{y})=\lambda\mathbf{z}$$

□ **مثال:** برای هر بردار $\mathbf{y}$، ماتریس همانی I درمعادلهٔ زیرصدق می‌کند

$$I\mathbf{y}=\mathbf{y}=1\mathbf{y}$$

بنابراین ۱ ویژه‌مقداری‌برای I است وهر بردارغیرصفر، یك‌ویژه‌برداری‌است برای Iمتعلق به ۱. ازآنجاکه یك بردار می‌تواند فقط به‌یك (یا هیچ) ویژه‌مقدار متعلق باشد (یا به‌هیچ ویژه‌برداری متعلق نباشد)، نتیجه می‌شودکه ۱ تنها ویژه‌مقدار I است.

ماتریس صفر فقط و فقط یك ویژه‌مقدار دارد و آن عدد صفراست.

ماتریس

---

1. localization 2. J. H. Wilkinson 3. G. W. Stewart

$$B=\begin{bmatrix}۱ & ۲ & ۰\\ ۲ & ۱ & ۰\\ ۰ & ۰ & -۱\end{bmatrix}$$

دارای ویژه‌مقدار ۱ — است، زیرا $B\mathbf{i}_۳=-\mathbf{i}_۳$، همچنین $B(\mathbf{i}_۱+\mathbf{i}_۲)=۳(\mathbf{i}_۱+\mathbf{i}_۲)$ لذا $\lambda=۳$ نیز یك ویژه‌مقدار B است. و بالاخره $B(\mathbf{i}_۱-\mathbf{i}_۲)=-(\mathbf{i}_۱-\mathbf{i}_۲)$ لذا ویژه‌مقدار ۱ — دارای دو ویژه‌بردار $\mathbf{i}_۳$ و $(\mathbf{i}_۱-\mathbf{i}_۲)$ است که مستقل‌خطی هستند. □

اگر ماتریس $B=(b_{ij})$ بالامثلثی باشد، آنگاه λ یك ویژه‌مقدار B خواهد بود اگر و فقط اگر، به‌ازای مقداری مانند i، داشته باشیم $\lambda=b_{ii}$. زیرا که در این صورت ماتریس $(B-\lambda I)$ نیز بالامثلثی خواهد بود، و بنابراین به‌موجب قضیهٔ ۶.۴، $(B-\lambda I)$ وارونپذیر نیست اگر و فقط اگر، یکی از درایه‌های قطری آن صفر باشد، یعنی اگر وفقط اگر، به‌ازای مقداری مانند i، تساوی $b_{ii}-\lambda=۰$ برقرار باشد. بنابراین مجموعهٔ ویژه‌مقدارهای متعلق به‌یك ماتریس مثلثی همان مجموعهٔ اعدادی است که روی قطر آن قرار دارند.

□ **مثال ۱۰.۴:** بخصوص، تنها ویژه‌مقدار ماتریس

$$B=\begin{bmatrix}۰ & ۰ & ۱\\ ۰ & ۰ & ۰\\ ۰ & ۰ & ۰\end{bmatrix}$$

عدد صفر است و هردو بردار $\mathbf{i}_۱$ و $\mathbf{i}_۲$ ویژه‌بردارهای B و متعلق به‌این ویژه‌مقدار هستند. هر ویژه‌بردار دیگر B باید ترکیبی خطی از این دو ویژه‌بردار باشد. چنانچه فرض کنیم $\mathbf{y}(=y_۱\mathbf{i}_۱+y_۲\mathbf{i}_۲+y_۳\mathbf{i}_۳)$ یك ویژه‌بردار B (که متعلق به‌تنها ویژه مقدار ۰ است) باشد، آنگاه

$$۰=۰\mathbf{y}=B\mathbf{y}=y_۱B\mathbf{i}_۱+y_۲B\mathbf{i}_۲+y_۳B\mathbf{i}_۳$$

$$=۰+۰+y_۳B\mathbf{i}_۳$$

از آنجاکه $B\mathbf{i}_۳=\mathbf{i}_۱\neq ۰$، نتیجه می‌شود که $y_۳=۰$ یعنی $\mathbf{y}=y_۱\mathbf{i}_۱+y_۲\mathbf{i}_۲$، که نشان می‌دهد $\mathbf{y}$ یك ترکیب خطی از ویژه‌بردارهای $\mathbf{i}_۱$ و $\mathbf{i}_۲$ است. □

برای اینکه نشان دهیم که چرا ویژه‌مقدارها مورد توجه هستند، اکنون به‌طور مختصر دنباله‌های برداری به‌شکل

$$\mathbf{z},\ B\mathbf{z},\ B^۲\mathbf{z},\ B^۳\mathbf{z},\ \cdots \qquad (۶۲.۴)$$

را مورد مطالعه قرار می‌دهیم. دنبالهٔ مذکور در کاربردهایی که درشروع این بخش نام‌برده شد، پدید می‌آید. چنین دنباله‌هایی را می‌باید در فصل۵، دربحث روشهای بارستی برای حل دستگاههای معادلات، مورد بحث قراردهیم.

فرض کنیدکه بردار آغازین[1] $\mathbf{z}^{(0)}$ در(۶۲.۴) بتواند به‌صورت مجموع ویژه‌بردارهای B نوشته شود، یعنی

$$\mathbf{z}=\mathbf{y}_1+\mathbf{y}_2+\cdots+\mathbf{y}_r \qquad (۶۳.۴)$$

که داریم

$$B\mathbf{y}_i=\lambda_i\mathbf{y}_i \qquad i=1,\ldots,r$$

دراین‌صورت جملهٔ mام از دنبالهٔ (۶۲.۴)، به‌شکل سادهٔ

$$B^m\mathbf{z}=\lambda_1^m\mathbf{y}_1+\lambda_2^m\mathbf{y}_2+\cdots+\lambda_r^m\mathbf{y}_r \qquad (۶۴.۴)$$

خواهد بود بنابراین رفتار دنبالهٔ برداری[2] (۶۲.۴)کاملا با دنبالهٔ عددی[3] ساده

$$\lambda_i^0,\ \lambda_i^1,\ \lambda_i^2,\ \lambda_i^3,\ \cdots \qquad i=1,\ldots,r$$

معین می‌شود. برای مثال نتیجه‌می‌شودکه

$$\lim_{m\to\infty} B^m\mathbf{z}=0 \qquad (۶۵.۴)$$

به‌ازای تمام iها، اگر $|\lambda_i|<1$، آنگاه

بعلاوه فرض کنیدکه λ_iها از لحاظ قدرمطلق مرتب شده باشند، یعنی

$$|\lambda_1|\geqslant|\lambda_2|\geqslant\cdots\geqslant|\lambda_r|$$

که این‌امرهمیشه با مرتب‌کردن درست $\mathbf{y}_i$ها می‌تواند به‌دست‌آید. بعلاوه فرض‌می کنیم که

$$|\lambda_1|>|\lambda_2| \qquad (۶۶.۴)$$

دراین‌فرض نه‌تنها λ_1 باید غیر ازتمام λ_iهای دیگر باشد [که این امرهمیشه باجمع کردن تمام $\mathbf{y}_i$های متعلق به λ_1 در (۶۳.۴) به‌دست می‌آید، و به‌موجب آن تنها یك ویژه‌مقدار متعلق به λ_1 حاصل می‌شود] بلکه λ_i دیگری به‌همان اندازه λ_1 وجود ندارد و این قسمت است‌که موجب می‌شود (۶۶.۴) یك فرض غیربدیهی باشد.

بنابراین از تقسیم دوطرف (۶۴.۴) بر λ_1^m خواهیم داشت

$$(\lambda_1^{-1}B)^m\mathbf{z}=\mathbf{y}_1+\sum_{i=2}^{r}\left(\frac{\lambda_i}{\lambda_1}\right)^m\mathbf{y}_i \qquad m=0,1,2,\ldots$$

---

1. starting vector 2. vector sequence
3. numerical sequence

بنا بر فرضهای اتخاذ شده، داریم

$$\left|\frac{\lambda_i}{\lambda_1}\right| < 1 \qquad i = 2, \dots, r$$

از این‌رو نتیجه می‌شود که

$$\lim_{m\to\infty} (\lambda_1^{-1} B)^m \mathbf{z} = \mathbf{y}_1 \qquad (۶۷.۴)$$

می‌توانیم بگوییم که اگر $\mathbf{z}$ بتواند به‌شکل (۶۳.۴) برحسب ویژه‌بردارهای B چنان نوشته شود که ویژه‌مقدار λ_1 متناظر با $\mathbf{y}_1$، مطلقاً از تمام ویژه‌مقدارهای دیگر بزرگتر باشد، آنگاه شکل مدرج شدهٔ $B^m\mathbf{z}$ به‌طرف $\mathbf{y}_1$ همگرا می‌شود.

□ **مثال ۱۱.۴:** قبلا دیدیم که ماتریس

$$B = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

دارای ویژه‌بردارهای $\mathbf{z}_1 = \mathbf{i}_1 + \mathbf{i}_2$، $\mathbf{z}_2 = \mathbf{i}_1 - \mathbf{i}_2$، $\mathbf{z}_3 = \mathbf{i}_3$ متناظر با ویژه‌مقدارهای $\lambda_1 = 3$، $\lambda_2 = \lambda_3 = -1$ می‌باشد. این ویژه‌بردارها مستقل خطی هستند (تمرین ۱۰.۴–۱۰ را ببینید) و از این‌رو یك پایه برای تمام ۳- بردارها تشکیل می‌دهند بنابراین نتیجه می‌شود که هر ۳-برداری را می‌توان به‌صورت حاصلجمعی از ویژه‌بردارهای B نوشت. به‌ویژه بردار $\mathbf{z}$ که به‌وسیلهٔ $\mathbf{z}^T = [1\ 2\ 3]$ معین می‌شود، می‌تواند به‌صورت

$$\mathbf{z} = \mathbf{y}_1 + \mathbf{y}_2$$

نوشته شود، که در آن

$$\mathbf{y}_1 = 1.5\,\mathbf{z}_1 \qquad \mathbf{y}_2 = -0.5\,\mathbf{z}_2 + 3\,\mathbf{z}_3$$

در جدول ۱.۴، $B^m\mathbf{z}$ و $\mathbf{z}^{(m)} = \alpha_m B^m \mathbf{z}$ به‌ازای $m = 0, \dots, 5$ فهرست شده‌اند. $\mathbf{z}^{(m)}$ چنان مدرج شده‌است که عنصر اول آن برابر با ۱ باشد. به‌نظر می‌رسد که $\mathbf{z}^{(m)}$ به‌سمت ویژه‌بردار $\mathbf{i}_1 + \mathbf{i}_2$ متعلق به $\lambda_1 = 3$ همگرا می‌شود. □

روش توانی[1] که برای محاسبهٔ بزرگترین ویژه‌مقدار (از نظر قدرمطلق) یك ماتریس معین B به‌کار می‌رود مبتنی بر توضیح زیر است. با انتخاب یك $\mathbf{z}$، مثلا $\mathbf{z} = \mathbf{i}_1$، چند جملهٔ اول دنبالهٔ (۶۲.۴) تولید می‌شود و با ادامهٔ کار نسبتهایی به‌شکل

1. power method

جدول ۱۰۴

m	۰	۱	۲	۳	۴	۵
$B^m\mathbf{z}$	۱	۵	۱۳	۴۱	۱۲۱	۳۶۵
	۲	۴	۱۴	۴۰	۱۲۲	۳۶۴
	۳	۳	۳	۳	۳	۳
$\mathbf{z}^{(m)}$	۱	۱٫۰	۱٫۰۰	۱٫۰۰۰	۱٫۰۰۰	۱٫۰۰۰
	۲	۰٫۸	۱٫۰۸	۰٫۹۷۶	۱٫۰۰۸	۰٫۹۹۷
	۳	۰٫۶	۰٫۲۳	۰٫۰۷۳	۰٫۰۲۵	۰٫۰۰۸

$$\mathbf{u}^T B^{m+1}\mathbf{z}/(\mathbf{u}^T B^m \mathbf{z}) \qquad (۶۸.۴)$$

محاسبه می‌شود، از (۶۴.۴) نتیجه می‌شود

$$\frac{\mathbf{u}^T B^{m+1}\mathbf{z}}{\mathbf{u}^T B^m \mathbf{z}} = \frac{\lambda_1 \mathbf{u}^T\mathbf{y}_1 + (\lambda_2/\lambda_1)^{m+1}\mathbf{u}^T\mathbf{y}_2 + \cdots + (\lambda_r/\lambda_1)^{m+1}\mathbf{u}^T\mathbf{y}_r}{\mathbf{u}^T\mathbf{y}_1 + (\lambda_2/\lambda_1)^m \mathbf{u}^T\mathbf{y}_2 + \cdots + (\lambda_r/\lambda_1)^m \mathbf{u}^T\mathbf{y}_r}$$

بنابراین به‌شرط $\mathbf{u}^T\mathbf{y}_1 \neq 0$ و $|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_r|$ خواهیم داشت

$$(\mathbf{u}^T B^{m+1}\mathbf{z})/(\mathbf{u}^T B^m \mathbf{z}) = \lambda_1 + \mathcal{O}((\lambda_2/\lambda_1)^m)$$

توجه کنید، درحالتی که $B = B^T$، یعنی B متقارن باشد در رابطهٔ (۶۸.۴) بردار $\mathbf{u} = B^m\mathbf{z}$ به‌کار می‌آید. نسبت حاصله یعنی

$$(\mathbf{u}^T B\mathbf{u})/(\mathbf{u}^T\mathbf{u})$$

به‌خارج قسمت ریلای[1] (برای $\mathbf{u}$ و B) موسوم است که می‌توان به‌آسانی دید کـه برابر است با

$$(\mathbf{z}^T B^{2m+1}\mathbf{z})/(\mathbf{z}^T B^{2m}\mathbf{z})$$

و از این‌رو با تعداد عملیاتی تا $\mathcal{O}((\lambda_2/\lambda_1)^{2m})$ با λ_1 برابرخواهد شد.

□ **مثال ۱۲۰۴:** ازدنبالهٔ حاصل‌درمثال ۱۱۰۴، به‌ازای $\mathbf{u} = \mathbf{i}_1$ دنبالهٔ نسبتها را به‌صورت زیر به‌دست می‌آوریم

1. Rayleigh quotient

$$۵, ۲ر۶, ۳ر۱۵۳۸\ldots, ۲ر۹۵۱۲\ldots, ۳ر۰۱۶۵\ldots$$

درحالی‌که به‌ازای $\mathbf{u}=\mathbf{i}_۲$، دنبالهٔ زیر به‌دست می‌آید

$$۲, ۳ر۵, ۲ر۸۵۷۱\ldots, ۳ر۵, ۲ر۹۸۳۶\ldots$$

به‌نظر می‌رسد که هردو دنباله به‌سمت $\lambda_۱=۳$ همگرا می‌شوند، اما به‌ازای $\mathbf{u}=\mathbf{i}_۳$، دنبالهٔ زیر را خواهیم داشت

$$۱, ۱, ۱, ۱, ۱$$

که به‌نظر نمی‌رسد به‌سمت ۳ همگرا شود.

چون B متقارن‌است، دنبالهٔ خارج‌قسمتهای ریلای را نیز محاسبه و نسبتهای زیر را پیدا می‌کنیم

$$۱ر۵۷۱۴\ldots, ۲ر۶, \quad ۲ر۹۴۶۵\ldots, ۲ر۹۹۳۹\ldots, ۲ر۹۹۹۳\ldots$$

دراین دنباله تقریباً درهردور یك رقم به‌دست می‌آوریم که حاکی ازاین امراست که دنبالهٔ فوق باید به‌ازای تعداد عملیاتی تا $\mathcal{O}((۱/۳)^{۲^m})$ به‌سمت $\lambda_۱=۳$ همگرا شود. □

شق جالب دیگرروش توانی روش **بارست معکوس**[1] است. در اینجا، علاوه بر انتخاب بردار آغازین $\mathbf{z}$ که دررابطهٔ (۶۳.۴) صدق می‌کند، عددی مانند p که بایکی از ویژه‌مقدارهای B برابر نباشد، انتخاب و سپس دنبالهٔ زیر را تشکیل می‌دهند

$$\mathbf{z}, \hat{B}\mathbf{z}, \hat{B}^۲\mathbf{z}, \hat{B}^۳\mathbf{z}, \ldots$$

با

$$\hat{B}=(B-pI)^{-۱}$$

باید توجه داشت که در رابطهٔ (۶۳.۴) به‌ازای هر ویژه بردار $\mathbf{y}_i$ از B، داریم

$$(B-pI)\mathbf{y}_i=B\mathbf{y}_i-p\mathbf{y}_i=(\lambda_i-p)\mathbf{y}_i$$

بنابراین خواهیم داشت

$$(\lambda_i-p)^{-۱}\mathbf{y}_i=(B-pI)^{-۱}\mathbf{y}_i$$

رابطهٔ فوق نشان می‌دهد که $\mathbf{z}$، مجموع ویژه‌بردارهای $\hat{B}=(B-pI)^{-۱}$ متناظر با ویژه‌مقدارهای $(\lambda_i-p)^{-۱}$، $i=۱, \ldots, r$، نیز هست. اما اگر p تنها به‌یکی ازویژه‌مقدارهای

---

1. inverse iteration

$\lambda_1, \ldots, \lambda_r$، مثل λ_j، کاملا نزدیک باشد، آنگاه $(\lambda_j - p)^{-1}$ درمقایسه باسایرویژه‌مقدارهای $(\lambda_i - p)^{-1}$، از لحاظ قدرمطلق بسیار بزرگ خواهد بود و بنابراین بحث پیشین ما در روش توانی، به‌ما اجازه می‌دهدکه چنین نتیجه‌گیری کنیم که شکل مدرج و مناسبی‌ازدنبالهٔ $\mathbf{z}, \hat{B}\mathbf{z}, \hat{B}^2\mathbf{z} \ldots$ به‌سرعت به‌سمت ویژه‌بردار $\mathbf{y}_j$ متناظر بـا λ_j، همگرا می‌شود، درحین اینکه نسبتهای متناظر با آن، یعنی

$$\mathbf{u}^T\hat{B}^{m+1}\mathbf{z}/(\mathbf{u}^T\hat{B}^m\mathbf{z})$$

نیز با همان سرعت به‌سمت عدد $(\lambda_j - p)^{-1}$ همگرا می‌شود. این امر موجب می‌شود که بارست معکوس، روش بسیارکارایی‌در وضعیت‌زیرباشد: تقریب خوبی‌قبلا برای ویژه‌بردار B به‌دست آورده‌ایم و می‌خواهیم این تقریب را بهبود بخشیم و (یا) ویژه‌بردار متناظر با آن را محاسبه‌کنیم.

همان‌گونه‌که ذکرکردیم، بارست معکوس‌ابتدا به‌ساختن‌ماتریس $\hat{B} = (B - pI)^{-1}$ نیازدارد. اما همچنان که در بخش ۴.۴ بحث شد، مـا چنین عکسی را صراحتاً نمی‌سازیم. بلکه از

$$\mathbf{z}^{(m)} := \hat{B}^m\mathbf{z} \qquad m = 0, 1, 2, \ldots$$

به‌دست می‌آوریم، و توجه می‌کنیم که

$$(B - pI)\mathbf{z}^{(m)} = \mathbf{z}^{(m-1)}$$

درنتیجه، وقتی‌که یکبارتجزیه به‌عوامل PLU رابرای ماتریس $B - pI$ به‌دست آوردیم از راه الگوریتـم جایگزینی ۴.۴ یعنی قراردادن در $\mathcal{O}(n^2)$، $\mathbf{z}^{(m)}$ را از $\mathbf{z}^{(m-1)}$ به‌دست می‌آوریم. این روش، اگرمجبور به‌انجام‌آن باشیم، گرانتر از محاسبهٔ صریح حاصلضرب $\hat{B}\mathbf{z}^{(m-1)}$ از $\hat{B}$ نخواهد بود.

دراینجایك زیربرنامهٔ **فورترن** برای‌انجام بارست عکس‌آورده‌ایم. درمرحلهٔ mام، $\mathbf{u} = B^m\mathbf{z}$ را انتخاب می‌کنیم، یعنی، درهرمرحله خارج قسمت ریلای را انجام می‌دهیم.

```
      SUBROUTINE INVITR ( B, N, EGUESS, VGUESS, W, D, IPIVOT,
     *                                   EVALUE, VECTOR, IFLAG )
C     CALLS  F A C T O R  ,  S U B S T .
      INTEGER IFLAG,IPIVOT(N),   I,ITER,ITERMX,J
      REAL B(N,N),D(N),EGUESS,EVALUE,VECTOR(N),VGUESS(N),W(N,N)
     *     ,EPSLON,EVNEW,EVOLD,SQNORM
C****** I N P U T ******
C  B  THE MATRIX OF ORDER  N  WHOSE EIGENVALUE/VECTOR IS SOUGHT.
C  N  ORDER OF THE MATRIX  B .
C  EGUESS  A FIRST GUESS FOR THE EIGENVALUE.
C  VGUESS  N-VECTOR CONTAINING A FIRST GUESS FOR THE EIGENVECTOR.
C****** W O R K  A R E A ******
C  W  MATRIX OF ORDER  N
C  D  VECTOR OF LENGTH  N
C  IPIVOT  INTEGER VECTOR OF LENGTH  N
C****** O U T P U T ******
C  EVALUE  COMPUTED APPROXIMATION TO EIGENVALUE
C  VECTOR  COMPUTED APPROXIMATION TO EIGENVECTOR
C  IFLAG  AN INTEGER,
```

```
C          = 1 OR -1 (AS SET IN FACTOR), INDICATES THAT ALL IS WELL,
C          = 0 , INDICATES THAT SOMETHING WENT WRONG. SEE PRINTED ERROR
C               MESSAGE .
C****** M E T H O D ******
C     INVERSE ITERATION, AS DESCRIBED IN THE TEXT, IS USED.
C******
C  THE FOLLOWING  T E R M I N A T I O N  P A R A M E T E R S  ARE SET
C  HERE, A TOLERANCE  E P S L O N  ON THE DIFFERENCE BETWEEN SUCCESSIVE
C  EIGENVALUE ITERATES, AND AN UPPER BOUND  I T E R M X  ON THE NUMBER
C  OF ITERATION STEPS.
      DATA EPSLON,ITERMX /.000001,20/
C
C                          PUT  B - (EGUESS)*IDENTITY   INTO  W
      DO 10 J=1,N
         DO 9 I=1,N
    9       W(I,J) = B(I,J)
   10    W(J,J) = W(J,J) - EGUESS
      CALL FACTOR ( W, N, D, IPIVOT, IFLAG )
      IF (IFLAG .EQ. 0)  THEN
         PRINT 610
  610    FORMAT(' EIGENVALUE GUESS TOO CLOSE.
     *          ,'NO EIGENVECTOR CALCULATED.')
                                          RETURN
      END IF
C                        ITERATION STARTS HERE
      PRINT 619
  619 FORMAT(' ITER  EIGENVALUE       EIGENVECTOR COMPONENTS'/)
      EVOLD = 0.
      DO 50 ITER=1,ITERMX
C                        NORMALIZE CURRENT VECTOR GUESS
         SQNORM = 0.
         DO 20 I=1,N
   20       SQNORM = VGUESS(I)**2 + SQNORM
         SQNORM = SQRT(SQNORM)
         DO 21 I=1,N
   21       VGUESS(I) = VGUESS(I)/SQNORM
C                        GET NEXT VECTOR GUESS
         CALL SUBST ( W, IPIVOT, VGUESS, N, VECTOR )
C                        CALCULATE RAYLEIGH QUOTIENT
         EVNEW = 0.
         DO 30 I=1,N
   30       EVNEW = VGUESS(I)*VECTOR(I) + EVNEW
         EVALUE = EGUESS + 1./EVNEW
C
         PRINT 630,ITER,EVALUE,VECTOR
  630    FORMAT(I3,E15.7,2X,3E14.7/(20X,3E14.7))
C                        STOP ITERATION IF CURRENT GUESS IS CLOSE TO
C                        PREVIOUS GUESS FOR EIGENVALUE
         IF ( ABS(EVNEW-EVOLD) .LE. EPSLON*ABS(EVNEW) )
     *                                    RETURN
         EVOLD = EVNEW
         DO 50 I=1,N
   50       VGUESS(I) = VECTOR(I)
C
      IFLAG = 0
      PRINT 660,EPSLON,ITERMX
  660 FORMAT(' NO CONVERGENCE TO WITHIN',E10.4,' AFTER',I3,' STEPS.')
                                          RETURN
      END
```

□ **مثال ۱۳.۴:** برای ماتریس B در مثال ۱۲.۴ بهازای $\mathbf{z}=[1\ ,1\ ,1]^T$ و $p=3.0165$، که در مثال فوق برای $\lambda_1=3$ از نخستین دنبالهٔ نسبتها بهترین حدس است، برنامهٔ INVITR را بهکار میبریم

بارست	ویژه مقدار	مؤلفههای ویژه بردار		
1	0.2991801 + 01	−0.3499093 + 02	−0.3499093 + 02	−0.1437446 + 00
2	0.3000000 + 01	0.4285478 + 02	0.4285478 + 02	0.7232219 − 03
3	0.3000000 + 01	−0.4285496 + 02	−0.4285496 + 02	−0.2971047 − 05
4	0.3000000 + 01	0.4285496 + 02	0.4285496 + 02	0.1220522 − 07

ویژه مقدار = 0.3000000 + 01

ویژه بردار =

0.4285496 + 02 0.4285496 + 02 0.1220522 − 07

چون B متقارن‌است وخارج‌قسمت ریلای محاسبه شده بود. خروجی، همگرایی بسیارسریع ویژه‌بردار(درهرمرحله ازبارست، درحدود۲ رقم‌اعشاری به‌دست می‌آید) وحتی‌همگرایی بسیارسریع ویژه‌مقدار را نشان می‌دهد.

به‌عنوان مثالی از این امر کـه برخلاف روش توانی، بارست معکوس برای هـر ویژه‌مقداری می‌تواند به‌کار رود، از $\mathbf{z}=[1, 1, 1]^T$ و $p=0$ نیز شروع کرده امیدواریم که بدین‌وسیله به‌کوچکترین ویژه‌مقدار B (ازنظرقدرمطلق) دست یابیم.

بارست	ویژه مقدار	مؤلفه‌های ویژه بردار		
1	−0.9000000 + 01	0.1924501 + 00	0.1924501 + 00	−0.5773503 + 00
2	−0.1320000 + 01	0.1005038 + 00	0.1005038 + 00	0.9045340 + 00
3	−0.1033195 + 01	0.3658808 − 01	0.3658809 − 01	−0.9878783 + 00
4	−0.1003661 + 01	0.1232878 − 01	0.1232877 − 01	0.9986311 + 00
5	−0.1000406 + 01	0.4114594 − 02	0.4114604 − 02	−0.9998476 + 00
6	−0.1000045 + 01	0.1371724 − 02	0.1371714 − 02	0.9999831 + 00
7	−0.1000005 + 01	0.4572417 − 03	0.4572513 − 03	−0.9999981 + 00
8	−0.1000001 + 01	0.1524206 − 03	0.1524109 − 03	0.9999998 + 00
9	−0.1000000 + 01	0.5080043 − 04	0.5081010 − 04	−0.1000000 + 01

ویژه مقدار = −0.1000000 + 01

ویژه بردار =

0.5080043 − 04 0.5081010 − 04 −0.1000000 + 01

دراین‌مورد همگرایی بسیار کندتراست، زیرا ۰، مخصوصاً به‌ویژه‌مقدار ۱ — نزدیک نیست،اما با اختیارویژه‌برداری به‌شکل $[1, 0, 0]^T$ (به‌جای‌شکل کلیترممکن $[a, -a, b]^T$ برای ویژه‌مقدار ۱ — از ماتریس B)، بعد از۹ بارست به‌همگرایی می‌رسیم. □

روش توانی و شکل دیگرآن یعنی بارست معکوس، عموماً کاربردی ندارد. به‌طور کلی اگرقرارباشدکه ویژه‌مقدارها به‌صورت اعداد همتافت به‌دست آیند، درابتدا حساب اعدادهمتافت باید به‌کارگرفته شود. راههای ماهرانهٔ ویژه‌ای دردست‌است‌که به‌کمک‌آنها درحساب اعداد حقیقی می‌توان از یك ماتریس حقیقی B به‌یـك جفت ویژه‌مقدار مزدوج همتافت دست‌یافت. مشکل‌بسیارجدی عبارت از احتمال همگرایی بسیارکنداست‌که زمانی پیش می‌آیدکـه بزرگترین ویژه‌مقدار بعدی از لحاظ قدرمطلق به‌بزرگترین مقدار بسیار نزدیك باشد. درحالی‌که فرایندΔ^2ی اتکین (الگوریتم۷.۳) می‌تواند برای شتاب‌بخشیدن به‌همگرایی (اگرهمگرایی وجود داشته باشد) به‌کار رود، درحالت نهایی یعنی وقتی کـه تساوی$|\lambda_2|=|\lambda_1|$ برقرارباشد و $\lambda_1\neq\lambda_2$، (برای مثال وقتی‌که $\lambda_2=\overline{\lambda}_1$)، همگرایی وجود نخواهد داشت. می‌توان بـا تغییرمکان مناسب، راه‌حلی برای ایـن مسئله پیداکرد، یعنی به‌جـای استفاده از خـود مـاتریس B با ماتریس $B-pI$ کارکـرد، به‌طوری‌کـه $|\lambda_1-p|>|\lambda_2-p|$ (تمرینات ۸.۴—۶ و ۸.۴—۷ را ببینید).

و بالاخره زمانی‌که بردارآغازین$\mathbf{z}$ را نتوان به‌صورت حاصلجمع بردارهای خاص B نوشت، روش توانی ثبات نظری خود را (آن‌گونه‌که دراینجا عنوان‌کردیم) از دست

می‌دهد. از آنجایی که ویژه‌بردارهای B در دست نیستند، تنها وقتی می‌توانیم مطمئن باشیم که $\mathbf{z}$ به‌صورت حاصلجمعی از ویژه‌بردارهای ماتریس B از مرتبهٔ $n \times n$ می‌تواند نوشته شود که بدانیم هر n-برداری را می‌توان به‌صورت حاصلجمعی از ویژه‌بردارهای B نوشت، ولی در واقع، معنی این امر این است که بخواهیم ویژه‌بردارهای B آن‌قدر باشند که بتوانند یك پایه تشکیل دهند. یك پایه از n-برداریها راکه تماماً ویژه‌بردارهای ماتریس B از مرتبهٔ $n \times n$ باشند، یك مجموعهٔ **کامل** ویژه‌بردارهای B نامند. روشن است که هرگاه $\mathbf{z}_1, \ldots, \mathbf{z}_n$، مجموعهٔ کامل ویژه‌بردارهای ماتریس B از مرتبهٔ $n \times n$ و بنابراین پایه‌ای برای تمام n-برداریها باشد، آنگاه هر n-برداری خاص $\mathbf{z}$ می‌تواند به‌صورت ترکیبی خطی از این ویژه‌بردارها به‌ازای ضرایب مناسب $a_1, \ldots, a_n$ به‌صورت

$$\mathbf{z} = a_1\mathbf{z}_1 + a_2\mathbf{z}_2 + \cdots + a_n\mathbf{z}_n$$

نوشته شود. اگر $a_i \neq 0$، آنگاه $\mathbf{y}_i = a_i\mathbf{z}_i$ نیز یك ویژه‌بردار B است، درحالی که اگر $a_i = 0$، بی آنکه مشکلی پیش آید، جملهٔ $a_i\mathbf{z}_i$ می‌تواند از حاصلجمع حذف شود. بدین ترتیب $\mathbf{z}$ به‌صورت حاصلجمعی از بردارهای B به‌دست می‌آید، (بجز در حالت نامطلوب $\mathbf{z} = 0$).

متأسفانه، همان گونه که قبلا درمثال ۱۰.۴ دیدیم، معلوم نیست که هرماتریس مجموعهٔ کاملی از ویژه‌بردارها داشته باشد.

تشابه[1]

این امر که معلوم نیست هرماتریسی یك مجموعهٔ کاملی از ویژه‌بردارها داشته باشد، نشانهٔ پیچیدگی ناشی از نظریهٔ ویژه‌مقدارهاست. این پیچیدگی مربوط به‌این حکم است که هر ماتریس مربعی را نمی‌توان به‌شکل

$$Y\Lambda Y^{-1}$$

که در آن Λ یك ماتریس قطری است، نوشت. زیرا تساوی $BY = Y\Lambda$ باماتریس قطری Λ برقرار است، اگر و فقط اگر ستونهای ماتریس Y شامل ویژه‌بردارهای B باشند. درحالی که چنین ماتریس Yای از مرتبهٔ $n \times n$ وارونپذیر خواهد بود، اگر و فقط اگر، ستونهای آن تشکیل یك پایه برای n-برداریها دهند.

دوماتریس A و B را **مشابه**[2] گوییم، هرگاه به‌ازای ماتریسی (وارونپذیر) مانند C داشته باشیم

$$A = C^{-1}BC$$

در ماتریسهای مشابه، ویژه‌مقدارها و ویژه‌بردارهای متناظر به‌آنها یکی هستند. در واقع اگر به‌ازای بردار غیر صفری مانند $\mathbf{x}$، داشته باشیم $(B - \lambda I)\mathbf{x} = 0$ و $A = CBC^{-1}$،

---

1. Similarity 2. similar

آنگاه $C\mathbf{x}$ نیز غیرصفر بوده و داریم $AC=CB$، از این‌رو خواهیم داشت

$$(A-\lambda I)C\mathbf{x}=C(B-\lambda I)\mathbf{x}=C\mathbf{o}=\mathbf{o}$$

به‌طور خلاصه، متناظر با هر زوج ویژه‌مقدار-ویژه‌بردار $(\lambda, \mathbf{x})$ از B، یك زوج ویژه‌مقدار-ویژه‌بردار $(\lambda, C\mathbf{x})$ از A وجود دارد.

مطالب فوق به عنوان اولین مرحله در محاسبهٔ ویژه مقدارهای B بیانگر یك تبدیل تشابهی از B به‌ماتریس $A=C^{-1}BC$ است که برای آن ویژه مقدارها را تاحدی آسانتر می‌توان محاسبه کرد.

برای مثال، اگر بتوان یك ماتریس بالا-مثلثی مشابه با B یافت، آنگاه تمام ویژه‌مقدارهای B معین می‌شوند، زیرا این ویژه‌مقدارها تماماً روی قطر T قراردارند. در حقیقت می‌توان ثابت کرد که:

قضیهٔ ۱۳۰۴: قضیهٔ شور[1] هر ماتریس مربعی B را می‌توان به‌صورت $U^{-1}TU$ نوشت که در آن، T ماتریس بالا-مثلثی و U ماتریس **یکانی**[2] است، یعنی $U^HU=I$.

این حقیقت که U ماتریس یکانی است، این نتیجهٔ جالب را دربردارد که به‌ازای هر $\mathbf{x}$ تساوی $\|U\mathbf{x}\|_2=\|\mathbf{x}\|_2$ برقرار است. از این‌رو داریم $\|T\|_2=\|B\|_2$، لذا ماتریس بالا-مثلثی T مشابه B بوده و حتی اندازهٔ آن با اندازهٔ B نیز یکی می‌باشد. با وجود این، متأسفانه برای ساختن این ترکیبهای U و T معمولا از روند بارستی استفاده می‌کنند.

اما همواره ممکن است که با تعداد $\mathcal{O}(2n^3/3)$ عملیات ممیز شناور، B را ازراه مشابهت به‌ماتریس $H=(h_{ij})$، که تقریباً مثلثی یا **هسنبرگ**[3] یعنی به‌صورت

$$h_{ij}=0, \qquad i>j-1 \text{ به‌ازای}$$

است تبدیل کرد. بنابراین قسمت پایین-مثلثی H بجز احتمالا اولین قطرفرعی پایین قطر اصلی، برابرصفراست. H از روی B با $n-2$ تبدیل تشابهی ساده به‌دست می‌آید که هرکدام از آنها یك ستون صفر دیگر در زیر اولین قطر پایینی ایجاد می‌کند.

مثلا ممکن‌است که از **تقارنهای محوری هاوس هولدر**[4]، یعنی ماتریسهایی به شکل

$$\alpha\mathbf{y}^T\mathbf{y}=2 \quad \text{با} \quad R(\mathbf{y})=I-\alpha\mathbf{y}\mathbf{y}^T \qquad (۶۹.۴)$$

به‌صورت زیر استفاده کرد: فرض کنید که درستونهای ۱، ۲، ...، $k-1$ در زیرقطرفرعی صفر آمده باشد، مثل حالتی که در آن به‌ازای $k=1$ با $H^{(1)}=B$ پیش می‌آید. پس، می‌خواهیم رابطهٔ

$$H^{(k+1)}=(R(\mathbf{y}))^{-1}H^{(k)}R(\mathbf{y})$$

را چنان تشکیل‌دهیم که نخستین $k-1$ ستون تغییر نیافته بماند، درحالی که اکنون درستون

---

1. Schur 2. unitary 3. Hessenberg 4. Householder

kام ودر زیراولین قطرفرعی نیزصفر داریم. بدین دلیل قبل ازهمه توجه داریم که وارون $R(\mathbf{y})$، خود $R(\mathbf{y})$ است، زیرا داریم

$$(I-\alpha\mathbf{y}\mathbf{y}^T)(I-\alpha\mathbf{y}\mathbf{y}^T)=I-\alpha\mathbf{y}\mathbf{y}^T-\alpha\mathbf{y}\mathbf{y}^T+\alpha^2\mathbf{y}(\mathbf{y}^T\mathbf{y})\mathbf{y}^T$$

و۲ $\alpha\mathbf{y}^T\mathbf{y}=$. ازاین‌رو خواهیم داشت، $H^{(k+1)}=R(\mathbf{y})H^{(k)}R(\mathbf{y})$. به همین ترتیب، محاسبه می‌شود که

$$\|R(\mathbf{y})\mathbf{x}\|_2=\|\mathbf{x}\|_2 \qquad (۷۰.۴)$$

یادآوری می‌کنیم که $\|\mathbf{z}\|_2^2=\mathbf{z}^T\mathbf{z}$). این مطلب، اتفاقاً نام «تقارن محوری» را روشن می‌کند. بعد باید بدانیم که کوتاهترین راه برای به‌دست‌آوردن حاصلضرب ماتریسی $AR(\mathbf{y})$ این است که به‌جای هرسطر $\mathbf{x}^T$ از A بردارسطری $\mathbf{x}^T-\alpha(\mathbf{x}^T\mathbf{y})\mathbf{y}^T$ را بگذاریم. ازاین‌رو با انتخاب

$$y_1=\cdots=y_k=0 \qquad (۷۱.۴)$$

kستون اول ازماتریس $H^{(k)}R(\mathbf{y})$ مشابه با kستون $H^{(k)}$ خواهدبود. بعد، باید بدانیم کوتاهترین راه به‌دست آوردن حاصلضرب ماتریسی $R(\mathbf{y})A$ این است که به‌جای هر ستون $\mathbf{x}$ از A بردارستونی $\mathbf{x}-\alpha(\mathbf{y}^T\mathbf{x})\mathbf{y}$ را بگذاریم. چون $H^{(k)}$ در ستونهای ۱ تا $k-1$ ودرپایین سطر$(k-1)$ام خود صفردارد، این امرنشان می‌دهدکه انتخاب(۷۱.۴) مارامطمئن‌می‌سازد که $(k-1)$ ستون‌اول از $R(\mathbf{y})H^{(k)}R(\mathbf{y})$ مشابه با$(k-1)$ستون$H^{(k)}$ است. این‌مطلب‌ما را با مسئلهٔ انتخاب $y_n,\ldots,y_{k+1}$ به‌گونه‌ای که ستون kام از $R(\mathbf{y})H^{(k)}$ درسطرهای $k+2,\ldots,n$ صفرداشته باشد مواجه می‌سازد. با توجه به (۷۱.۴)، معنی این مطلب‌این است که می‌باید $R(\mathbf{y})$ بردار

$$\hat{\mathbf{x}}=[0,\ldots,0,h_{k+1,k}^{(k)},\ldots,h_{n,k}^{(k)}]^T$$

را به‌ازای مقدار اسکالر(عددوار) مانند β، بر روی بردار $\beta\mathbf{i}_{k+1}$ بنگارد [دراینجا عنصر(i,j)ام $H^{(k)}$ با h_{ij}^k نشان داده شده‌است.] باتوجه به‌رابطهٔ(۷۰.۴). این امر به‌معنای آن‌است که

$$\beta=\pm\|\hat{\mathbf{x}}\|_2=((h_{k+1,k}^{(k)})^2+\cdots+(h_{nk}^{(k)})^2)^{1/2}$$

بعلاوه

$$\hat{\mathbf{x}}-(-\beta)\mathbf{i}_{k+1}=\hat{\mathbf{x}}-R(\mathbf{y})\hat{\mathbf{x}}=\hat{\mathbf{x}}-(\hat{\mathbf{x}}-\alpha(\mathbf{y}^T\hat{\mathbf{x}})\mathbf{y})=\alpha(\mathbf{y}^T\hat{\mathbf{x}})\mathbf{y}$$

نشان می‌دهدکه $\mathbf{y}$ می‌باید یك مضرب عددی (اسکالر) از بردار $\hat{\mathbf{x}}+\beta\mathbf{i}_{k+1}$ باشد. این‌امر مشخص می‌کندکه انتخاب زیر برای $\mathbf{y}$کار را به‌اتمام خواهد رسانید:

$$y_i=\begin{cases}\hat{x}_{k+1}+\text{signum}(\hat{x}_{k+1})\|\hat{\mathbf{x}}\|_2 & i=k+1\\ \hat{x}_i & i>k+1\end{cases} \qquad (۷۲.۴)$$

یعنی $\beta=\text{signum}(\hat{x}_{k+1})\|\hat{\mathbf{x}}\|_2$. در اینجا علامت β چنان انتخاب می‌شود که درمحاسبهٔ y_{k+1}، مانع از بین‌رفتن ارقام بامعنی‌شود. α ی متناظر به‌آن می‌تواند به‌سادگی به‌شکل

$$\alpha=1/(\beta y_{k+1}) \qquad (۷۳.۴)$$

نوشته‌شود. بدین‌ترتیب، بعد از $n-2$ مرحله از این قبیل، ماتریس

$$H=R^{-1}BR$$

به‌دست می‌آیدکه در آن H، ماتریس هسنبرگ‌است و

$$R=R(\mathbf{y}^{(1)})R(\mathbf{y}^{(2)})\cdots R(\mathbf{y}^{(n-2)})$$

حاصلضرب تقارنهای محوری هاوس‌هولدر است، از این‌رو داریم

$$R^{-1}=R(\mathbf{y}^{(n-2)})\cdots R(\mathbf{y}^{(2)})R(\mathbf{y}^{(1)})$$

روشن است که تقارن هاوس‌هولدر، ماتریس متقارن حقیقی است (اگر $\mathbf{y}$ حقیقی باشد)، بنابراین درصورتی که B ماتریس‌متقارن وحقیقی‌باشد، H نیزمتقارن حقیقی خواهد بود. بنابراین درصورتی‌که B متقارن حقیقی باشد، H سه‌قطری و متقارن‌است.

اکنون، برای راحتی، مطالب بالا را به‌صورت رسمی درمی‌آوریم.

الگوریتم ۶.۴: تبدیل‌تشابهی‌به‌شکل‌ماتریس بالا-مثلثی‌هسنبرگ بااستفاده از تقارنهای محوری هاوس‌هولدر

ماتریس A از مرتبهٔ n که در n ستون اول از آرایهٔ‌کاری H ازمرتبهٔ $n\times(n+2)$ حفظ شده، مفروض‌است.

For $k=1,\ldots,n-2$, do:

$\beta := \text{signum}(h_{k+1,k})\left(\sum_{j=k+1}^{n}(h_{jk})^2\right)^{1/2}$

$h_{k+1,k} := h_{k+1,k}+\beta$

$\alpha^{-1} := \beta h_{k+1,k}$

For $j=1,\ldots,n$, do:

$\gamma := \left(\sum_{i=k+1}^{n} h_{ik}h_{ji}\right)/\alpha^{-1}$

For $i = k + 1, \ldots, n$, do:
  $h_{ji} := h_{ji} - \gamma h_{ik}$
For $j = k + 1, \ldots, n$, do:
  $\gamma := \left(\sum_{i=k+1}^{n} h_{ik} h_{ij} \right) / \alpha^{-1}$
  For $i = k + 1, \ldots, n$, do:
    $h_{ij} := h_{ij} - \gamma h_{ik}$
$h_{k+1,n+1} := h_{k+1,k}$, $h_{k+1,n+2} := \alpha^{-1}$
$h_{k+1,k} := -\beta$

بنابراین H، متضمن جزء جالب از یك ماتریس بالا مثلثی هسنبرگ است که باماتریس ورودی A درقسمت بالا مثلثی هسنبرگ دراولین n ستون و n سطرش مشابه است. بعلاوه ماتریس فوق شامل اطلاعات کاملی است دربارهٔ بردارهای $\mathbf{y}$ و عددوارهای α که تقاذنهای محوری هاوس‌هولدر گونا گونی‌را که به‌کار رفته‌اند به‌مامی‌دهند. زمانی که ویژه‌بردارهای ماتریس بالامثلثی هسنبرگ می‌باید مجدداً به‌ویژه‌بردارهای ماتریس اصلی A تبدیل‌شوند، این اطلاعات لازم خواهند بود.

روشی که معمولا برای یافتن تمام ویژه‌مقدارهای یك ماتریس کلی B توصیه می‌شود، روش **QR** است. این‌کار با تبدیل ماتریس فوق به‌شکل ماتریس‌هسنبرگ H، که هم‌اکنون ذکر کردیم، آغاز می‌شود. زمانی که این کار انجام گیرد، ماتریس H نخستین عضودردنبالهٔ زیرخواهد بود

$$A^{(0)} = H,\ A^{(1)},\ A^{(2)},\ \ldots$$

که در آن $A^{(k+1)}$ به‌گونهٔ زیر از $A^{(k)}$ به‌دست می‌آید: $A^{(k)}$ به‌ماتریس یکانی Q وماتریس بالا-مثلثی (یا راست-مثلثی) R به‌صورت $A^{(k)} = QR$ تجزیه وسپس به‌شکل زیر تشکیل می‌شود

$$A^{(k+1)} = RQ = Q^{-1}(QR)Q$$

بنابراین $A^{(k+1)}$ با $A^{(k)}$ مشابه است. بعلاوه چون $A^{(k)}$ یك ماتریس‌هسنبرگ است، $A^{(k+1)}$ نیز به‌شکل یك ماتریس هسنبرگ‌است. این امر تعداد عملیات لازم برای به‌دست آوردن عوامل تجزیهٔ ماتریس فوق را به‌مقدار زیادکاهش‌می‌دهد. اما دربسیاری از موارد، به‌ازای مقدار بسیارزیاد k، $A^{(k)}$ به‌سمت یك ماتریس بالا-مثلثی که درایه‌های قطری‌آن لزوماً تمام ویژه‌مقدارهای B هستند، همگرا می‌شود.

بیان جزئیات و نظریه‌ای که در پشت این محاسبه وجود دارد ظریف و دشوار است، به‌ویژه اینکه، برای تسریع همگرایی به‌جای خود $A^{(k)}$، ماتریس $(A^{(k)}-s_kI)$ با تغییرمکان s_k تجزیه می‌شود. اما خواننده می‌باید از این حقیقت آگاه باشد که این روش و سایر روشها، به‌ویژه آنهایی که برای دسته‌های خاصی از ماتریسهای B مناسب‌اند، به‌صورت بسته‌های پیش‌ساخته‌ای[1] از زیربرنامه‌های فورترن موسوم به EISPACK ترجمه‌شده‌اند که از سوی آزمایشگاه ملی آرگون[2] یا بسیاری از مراکز محاسباتی علمی مسقیماً دردسترس علاقه‌مندان قرارداده می‌شود، توضیح کامل درمورد بسته‌های پیش‌ساختهٔ فوق، که شامل فهرست برنامه‌هاست، درکتاب اسمیت و دیگران[3] [۳۲] یافت می‌شود.

جایابی[4]

گاهی تنها یك براورد تقریبی از یك یا تمام ویژه مقدارهای B موردنظراست. حتی اگر شخص مایل به‌محاسبهٔ ویژه‌مقدارها باشد، ممکن است که لازم آید از اطلاعاتی درموردجای تقریبی ویژه‌مقدارها آغاز کند. چنین اطلاعاتی را قضیهٔ جایابی به‌دست می‌دهد که درصفحهٔ همتافت ناحیه‌هایی راکه ویژه‌مقدارهای B درآنها قرار دارند معین می‌کند.

اگر $B\mathbf{x}=\lambda\mathbf{x}$، آنگاه داریم $|\lambda|\,\|\mathbf{x}\|=\|\lambda\mathbf{x}\|=\|B\mathbf{x}\|\leqslant\|B\|\,\|\mathbf{x}\|$، که ایجاب می‌کند که درحالت $\mathbf{x}\neq 0$ داشته باشیم، $|\lambda|\leqslant\|B\|$. این مطلب، ثابت می‌کند که رابطهٔ

$$|\lambda|\leqslant\|B\| \qquad \text{به‌ازای هر ویژه‌مقدار } \lambda \text{ در } B \qquad (۷۴.۴)$$

و به‌ازای هر نرم ماتریسی برقرار باشد.

یك بیان دقیقتر آن در زیر می‌آید:

قضیهٔ ۱۴.۴: قرصهای گرشگورین[5] هر ویژه‌مقدار λ درماتریس $B=(b_{ij})$ از مرتبهٔ $n\times n$ در رابطهٔ زیر صدق می‌کند

$$|b_{ii}-\lambda|\leqslant\sum_{\substack{j=1\\ j\neq i}}^{n}|b_{ij}| \qquad \text{به‌ازای مقداری از } i$$

به‌عبارت دیگر، همهٔ ویژه‌مقدارهای B درناحیه‌ای که از اجتماع[6] برخی قرصها در صفحهٔ همتافت پدید می‌آید، قراردارند. در واقع اگر داشته باشیم

$$|b_{ii}-\lambda|>\sum_{\substack{j=1\\ j\neq i}}^{n}|b_{ij}| \qquad i=1,\ldots,n \text{ به‌ازای}$$

---

1. package 2. Argonne National Laboratory 3. Smith et al
4. Localization 5. Gershgorin's disks 6. union

آنگاه ماتریس $B-\lambda I$ نافذ قطری (سطراً) مؤکد خواهد بود، ازاین‌رو با توجه به‌تمرین ۶.۴-۳، پس $B-\lambda I$ وارونپذیراست یعنی λ یك ویژه مقدار B نیست.

□ **مثال ۱۴.۴:** به‌موجب رابطهٔ (۷۴.۴)، قدرمطلق هر ویژه‌مقدار ماتریس

$$B=\begin{bmatrix} ۱ & ۲ & ۰ \\ ۲ & ۱ & ۰ \\ ۰ & ۰ & -۱ \end{bmatrix}$$

درمثال ۱۱.۴، می‌باید بزرگتر از $||B||_\infty = ۳$ نباشد. قرصهای گرشگورین اطلاعات مفصلتری دراختیار ما قرارمی‌دهد، بدین‌صورت که هر ویژه‌مقدار λ از ماتریس B می‌باید

$$|۱-\lambda| \leqslant ۲ \quad \text{یا در} \quad |-۱-\lambda| \leqslant ۰ \quad \text{یا در}$$

صدق کند. □

دریك ماتریس هرمیتی، به‌ویژه یك ماتریس متقارن حقیقی، همهٔ ویژه‌مقدارهاحقیقی هستند. این‌شبیه‌به‌یك‌ماتریس‌قطری‌است، یعنی دارای یك مجموعهٔ‌کامل ازویژه‌بردارهاست. این یك نتیجه‌ای‌است که به‌آسانی ازقضیهٔ شور به‌دست می‌آید، تمرین۸.۴-۱۵را ببینید. برای یك ماتریس هرمیتی B، هم

$$\max_{x} \mathbf{x}^H B\mathbf{x}/(\mathbf{x}^H\mathbf{x}) \quad \text{و هم} \quad \min_{x} \mathbf{x}^H B\mathbf{x}/(\mathbf{x}^H\mathbf{x})$$

ویژه‌مقدارهای B هستند و سایر ویژه‌مقدارهای B بین این‌دو قراردارند. به‌یاد بیاورید که این **خارج‌قسمتهای ریلای**[1] قبلا دراین قسمت درهنگام بحث ازروش توانی آمده‌اند.

از ترکیب لم ۴.۴ وقضیهٔ ۱۰.۴ قضیهٔ جایابی دقیق زیر به‌دست می‌آید.

قضیهٔ ۱۵.۴ λ یك ویژه‌مقدار ماتریس B است اگر وفقط اگر λ یك ریشهٔ معادلهٔ مشخصهٔ زیر باشد

$$\det(B-\lambda I)=۰$$

ماتریس $(B-\lambda I)$ با ماتریس B فقط این‌تفاوت را داردکه ازهریك از درایه‌های قطری B مقدار λکسم شده‌است. اگر **نماد کرونکر**[2]، δ_{ij}، را برای نشان دادن آرایهٔ $(i\,.\,j)$ام ماتریس همانی به‌کارگیریم، یعنی

$$\delta_{ij}=\begin{cases} ۱ & \text{اگر } i=j \\ ۰ & \text{اگر } i\neq j \end{cases}$$

آنگاه داریم:

---

1. Rayleigh quotients 2. Kronecker symbol

$$(B-\lambda I)\text{ام، ماتریس}(i,j)\text{ درایهٔ} = b_{ij}-\lambda\delta_{ij}$$

بنابراین

$$\det(B-\lambda I)=\sum_{\mathrm{p}}\sigma_{\mathrm{p}}(b_{1p_1}-\lambda\delta_{1p_1})(b_{2p_2}-\lambda\delta_{2p_2})\cdots(b_{np_n}-\lambda\delta_{np_n})$$

$$=\sum_{\mathrm{p}}\sigma_{\mathrm{p}}\Big(\prod_{i\neq p_i} b_{ip_i}\Big)\Big(\prod_{i=p_i}(b_{ip_i}-\lambda)\Big)$$

رابطهٔ بالا نشان می‌دهدکـه $\det(B-\lambda I)$، مجموع بسجمله‌ای‌هایی با متغیر λ است. از آنجاکـه هر حاصلجمع دارای n عامل است، درنتیجه هرحاصلجمع یك بسجمله‌ای است برحسب λ از درجـهٔ حداکثـر n، در حالی کـه جمعیدهٔ متناظر بـا جایگشت همانـی $\mathbf{p}^T=[1\ 2\ \cdots\ n]$ به‌صورت

$$(b_{11}-\lambda)(b_{22}-\lambda)\cdots(b_{nn}-\lambda)$$

و از این‌رو نسبت به λ دقیقاً از درجهٔ nام است. درنتیجه $\det(B-\lambda I)$ که به‌عنوان تابعی از λ مورد بررسی قرار گرفته‌است، یك بسجمله‌ای است دقیقاً از درجهٔ n برحسب λ

$$p(\lambda)=\det(B-\lambda I)=(-\lambda)^n+n\text{ یك بسجمله‌ای از درجهٔ کمتر از}$$

این بسجمله‌ای، **بسجمله‌ای مشخصهٔ**[1] B نامیده می‌شود.

□ **مثال ۱۵.۴:** اگر

$$B=\begin{bmatrix}1 & 2 & 0\\ 2 & 1 & 0\\ 0 & 0 & -1\end{bmatrix}$$

آنگاه

$$p(\lambda)=\det(B-\lambda I)=\det\begin{bmatrix}1-\lambda & 2 & 0\\ 2 & 1-\lambda & 0\\ 0 & 0 & -1-\lambda\end{bmatrix}$$

و از بسط این دترمینان برحسب آخرین سطر یا آخرین ستون داریم

$$p(\lambda)=(-1-\lambda)\det\begin{bmatrix}1-\lambda & 2\\ 2 & 1-\lambda\end{bmatrix}=-(1+\lambda)[(1-\lambda)^2-4]$$

$$=-(1+\lambda)(1-\lambda+2)(1-\lambda-2)$$

---

1. characteristic polynomial

بنابراین ویژه‌مقدارهای A یعنی ریشه‌های $p(\lambda)$ اعداد ۱ — و ۳ هستند که قبلا با روشهای متفاوت دیگری در آغاز این بخش محاسبه شدند. □

از آنجا که یك بسجمله‌ای از درجهٔ n می‌تواند حداکثر n ریشه متمایز داشته باشد (به‌بخش ۱۰.۲ نگاه کنید)، لذا نتیجه می‌شود که یك ماتریس $n \times n$ حداکثر می‌تواند n ویژه‌مقدار داشته باشد. از سوی دیگر به‌موجب قضیهٔ بنیادی جبر، هر بسجمله‌ای از درجهٔ مثبت حداقل یك ریشه دارد (به‌قضیهٔ ۱۰.۱ نگاه کنید)؛ از این‌رو برای هر ماتریس مربعی حداقل یك ویژه‌مقدار وجود دارد. این ویژه‌مقدارها، اگر B یك ماتریس حقیقی باشد، حتی ممکن است همتافت نیز باشند. به‌موجب قضیهٔ ۱۵.۴، تکنیکهایی که برای یافتن ریشه‌های معادلات (مخصوصاً معادلات بسجمله‌ای که در فصل ۳ بحث شد) به‌کار می‌روند برای یافتن ویژه‌مقدارها نیز به‌کار خواهند رفت.

برای مثال، روش درونیابی درجهٔ دوم (روش مولر) را که در بخش ۷.۳ مورد بحث قرار گرفت، می‌توان برای پیدا کردن یك یا چند ویژه‌مقدار حقیقی یا همتافت از یك ماتریس داده شده به‌کار برد. برای استفاده از این روش فقط باید بتوانیم بسجمله‌ای $p(\lambda)$ را به‌ازای هر مقدار λ محاسبه کنیم. از آنجا که برای یك مقدار داده شدهٔ λ، $p(\lambda)$ اصلا دترمینانی از مرتبهٔ n است، از هر روشی که برای محاسبهٔ دترمینان به‌کار می‌رود می‌توان استفاده کرد. به‌ویژه، این کار را می‌توان با روش حذفی که در بخش ۷.۴ شرح داده شد، انجام داد. اما بهتر است در ابتدا ماتریس فوق را به‌ماتریس هسنبرگ و یا احتمالا به‌شکل ماتریس سه‌قطری که قبلا مورد بحث قرار گرفته تبدیل کرد. زیرا این کار هزینهٔ محاسبهٔ یك دترمینان را از تعداد عملیات $\mathcal{O}(n^3)$ به $\mathcal{O}(n^2)$ کاهش می‌دهد. در هر صورت با اعمال درونیابی مربعی برای یافتن یك ریشهٔ ξ_1 از بسجمله‌ای مشخصهٔ $p(\lambda) = \det(A - \lambda I)$ به‌ترتیب زیر عمل می‌کنیم:

۱. گیریم λ_0، λ_1، λ_2 سه تقریب برای $\xi_1 = \lambda$ باشند (یا اگر اطلاعاتی در دست نیست قرار دهید $\lambda_0 = -1$ و $\lambda_1 = 1$ و $\lambda_2 = 0$).

۲. دترمینان زیر را محاسبه کنید

$$p(\lambda_i) = \det \begin{bmatrix} a_{11} - \lambda_i & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda_i & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda_i \end{bmatrix} \qquad i = 0, 1, 2$$

۳. الگوریتم ۱۱.۳ را به‌کار برید تا همگرایی به‌سمت ریشهٔ ξ_1 حاصل شود.

۴. برای پیدا کردن ریشهٔ بعدی، این روند را تکرار کنید اما به‌جای $p(\lambda)$ تابع اندازهٔ[1]

---

1. deflated function

تقلیل یافته

$$\frac{p(\lambda)}{\lambda-\xi_1}$$

را به‌کار گیرید

۵. به‌طریقی که درقسمت ۷.۳ شرح داده شدکار را ادامه دهید.

روش درونیابی درجهٔ دوم درمقایسه با روشهای پیشرفته‌تر ازنظرکارایی محاسباتی چندان قابل رقابت نیست. ولی به‌کار گرفتن این روش بسیار آسان و کاملا کلی وتقریباً همیشه همگراست و در بیشتر موارد دقت عمل کافی به‌دست می‌دهد. این روش‌را می‌توان برای حل مسئلهٔ کلیتر ویژه مقدار زیر نیز به‌کار گرفت

$$\det(A-\lambda B)=0$$

در رابطهٔ بالا A و B ماتریسهایی از مرتبهٔ n هستند.

□ **مثال۱۶.۴: ارتعاشهای آزاد ساختار[1]های ساده.** در مهندسی راه وساختمان مسئله‌ای که غالباً با آن مواجه می‌شویم، مسئلهٔ تعیین فراوانی‌طبیعی λی ارتعاشهای آزادیك‌ساختار غیرمیرا[2]ست که درقبال چندین جرم و چندین درجهٔ آزادی[3] حاصل می‌شود. این مسئله را می‌توان به‌شکل

$$A\mathbf{x}=\lambda M\mathbf{x} \qquad (۷۵.۴)$$

بیان کردکه درآن

M = ماتریس جرم دستگاه
A = سختی[4] دستگاه
$\mathbf{x}$ = تغییرمکان طبیعی

ازآنجاکه (۷۵.۴) معرف یك دستگاه معادلات همگن‌است، اگر دترمینان ضرایب آن صفر باشد، یعنی اگر تساوی

$$\det(A-\lambda M)=0 \qquad (۷۵.۴ \text{ الف})$$

برقرارباشد، آنگاه یك جواب غیر بدیهی $\mathbf{x}$ خواهد داشت. بنابراین اگرماتریسهای A و M داده شده باشند مقادیری از λ که در(۷۵.۴ الف) صدق می‌کنند فراوانیهای طبیعی هستند. روش مولر را می‌توان مستقیماً برای پیداکردن این ویژه مقدارها به‌کار برد. مثلا در یك دستگاه معین با $M=I$ و ماتریس سختی A به‌صورت

---

1. structure 2. undamped 3. degrees of freedom
4. stiffness

$$A=\begin{bmatrix} ۴ & -۱ & -۱ & -۱ \\ -۱ & ۴ & -۱ & -۱ \\ -۱ & -۱ & ۴ & -۱ \\ -۱ & -۱ & -۱ & ۴ \end{bmatrix}$$

فراوانی‌طبیعی λی این دستگاه را پیدا کنید.

یك برنامهٔ‌کامپیوتری بـا استفاده از حذف‌گاوسی ۲.۴ برای محاسبهٔ دترمینانهـا و الگوریتم ۱۱.۳ مولر به‌عنوان ریشه‌یاب، ویژه‌مقدارها را به‌شرح زیر براورد نموده‌است.

$$\lambda_۱ = ۱ر۰۰۰۰۰۰۰۰$$
$$\lambda_۲ = ۵ر۰۰۰۰۲۳۴$$
$$\lambda_۳ = ۴ر۹۹۹۹۹۵۴$$
$$\lambda_۴ = ۴ر۹۹۹۹۹۷۳$$

به‌آسانی دیده می‌شود که مقدار دقیق ویژه‌مقدارها به‌ترتیب برابر|۱، ۵، ۵، ۵ هستند. بااین مثال‌کارایی روش مولر به‌عنوان یك ریشه‌یاب، نشان داده شده‌است کـه در اینجا یك ریشهٔ سه‌گانه[1] بـا دقت عمل نسبتاً خوبی محاسبه شده‌است.

□ **مثال ۱۷.۴:** درایه‌های یك ماتریس سه‌قطری B به‌صورت زیر تولید شده‌اند

$$b_{i,i+۱} = ۰ر۵ \qquad i = ۱, ۲, \cdots, n-۱$$
$$b_{i+۱,i} = ۰ر۵ \qquad i = ۱, ۲, \cdots, n-۱$$
$$b_{ij} = ۰ \qquad \text{برای تمامی } i \text{ و } j \text{ های دیگر}$$

یك برنامهٔ‌کامپیوتری بنویسیدکه ویژه‌مقدارهای B را به‌ازای $n = ۲۰$ پیدا کند.

به‌ازای $n = ۲۰$، روش مولر درکامپیوتر IBM ۷۰۹۴ نتایج زیررا تولیدکرده‌است:

$\pm$۰ر۰۷۴۷۳۰۰۹۳	$\pm$۰ر۷۳۳۰۵۱۸۷
$\pm$۰ر۲۲۲۵۲۰۹۴	$\pm$۰ر۸۲۶۲۳۸۷۷
$\pm$۰ر۳۶۵۳۴۱۰۲	$\pm$۰ر۹۰۰۹۶۸۸۶
$\pm$۰ر۵۰۰۰۰۰۰۰۱	$\pm$۰ر۹۵۵۵۷۲۸۱
$\pm$۰ر۶۲۳۴۸۹۸۰	$\pm$۰ر۹۸۸۸۳۰۸۳

1. triple root

باید توجه کرد که این ویژه‌مقدارها حقیقی و نسبت به‌مبدأ قرینه‌اند و قدرمطلق[1] همهٔ آنها کمتر از یک‌است. ویژه‌مقدارهای این ماتریس به‌روشنی معلوم (به‌تمرین ۸.۴–۴ نگاه کنید) و به‌شرح زیرند.

$$\lambda_k = \cos\frac{k\pi}{n+1} \qquad k = 1, 2, \ldots, n$$

دقت عمل نتایج حاصله از کامپیوتر را می‌توان با فرمول فوق امتحان کرد. به‌ازای $k=7$ و $n=20$ داریم، $\lambda_7 = \cos(\pi/3) = 1/2$ و نتیجهٔ حاصله ازکامپیوترکه در زیرآن در بالا خط کشیده‌است معرف دقت عملی‌است با هفت رقم با معنی. □

ماتریس‌مثال ۱۷.۴، حقیقی وسه‌قطری‌است. برای چنین ماتریسهایی روشهای‌خاصی در دسترس‌است. این امر بسیار مهم‌است، زیرا قبلا دیدیم که هرماتریس متقارن و حقیقی را می‌توان از راه تشابه به‌ماتریس سه‌قطری متقارن حقیقی تبدیل کرد.

رسم برآن‌است که چنین ماتریسی را به‌صورت زیر بنویسند

$$B = \begin{bmatrix} a_1 & b_1 & & & & & \\ b_1 & a_2 & b_2 & & & & \\ & b_2 & a_3 & b_3 & & & \\ & & b_3 & \ddots & \ddots & & \\ & & & \ddots & \ddots & b_{n-1} & \\ & & & & b_{n-1} & a_n & \end{bmatrix}$$

بسجمله‌ای مشخصهٔ ماتریس‌فوق می‌تواند به‌صورت $\det(B-\lambda I) = p_n(\lambda)$ به‌دست آید که برای آن داریم

$$p_0(\lambda) = 1, \quad p_1(\lambda) = a_1 - \lambda$$

$$p_j(\lambda) = (a_j - \lambda)p_{j-1}(\lambda) - b_{j-1}^2 p_{j-2}(\lambda) \qquad j = 2, 3, \ldots, n \qquad (۷۶.۴)$$

دراینجا $p_j(\lambda)$، دترمینان ماتریسی‌است که از اولین j سطروستون ماتریس $B-\lambda I$ تشکیل شده‌است. با استفاده از قاعدهٔ ۶ وبسط ماتریس به‌صورت سطری یا ستونی که درقسمت ۷.۴

---

1. modulu

بیان گردید (تمرین ۷.۴–۱۱ را نگاه کنید)، (۷۶.۴) ثابت می‌شود. خصوصیت بازگشتی[1] (۷۶.۴) این امکان را فراهم می‌سازد که $p_n(\lambda)$ با تعداد عملیاتی در حدود $3n$ محاسبه شود. بعلاوه به‌راحتی می‌توان از رابطهٔ بازگشتی فوق نسبت به λ مشتق گرفت، و این عمل، محاسبهٔ $p'_n(\lambda)$ را به‌صورت بازگشتی امکانپذیر و بنابراین کاربرد روش نیوتن را میسر می‌سازد.

اگر به‌ازای مقداری مانند i داشته باشیم $b_i = 0$ آنگاه از رابطهٔ بازگشتی (۷۶.۴) می‌فهمیم که بسجمله‌ای $p_{i-1}(\lambda)$ عاملی از $p_n(\lambda) = \det(B - \lambda I)$ است. بنابراین ریشه‌های $p_n(\lambda)$، ریشه‌های دو بسجمله‌ای $p_{i-1}(\lambda)$ و $p_n(\lambda)/p_{i-1}(\lambda)$ با درجهٔ کمتر هستند و می‌توان کار را براین دو بسجمله‌ای متمرکز کرد. درغیراین‌صورت اگربه‌ازای جمیع مقادیر i، $b_i \neq 0$، آنگاه B دارای n ویژه‌مقدار متمایزاست. همچنین، دنبالهٔ $p_0(\lambda)$، $p_1(\lambda)$، ...، $p_n(\lambda)$ از مقادیر محاسبه شده در طی محاسبهٔ $p_n(\lambda)$، اطلاعات اضافی زیر را به‌همراه دارد: تعداد تغییرعلامتها (با قاطعیت) در دنبالهٔ مذکور با تعداد ویژه‌مقدارهای B، که از λ کمتر است برابراست، این امر به علت این واقعیت است که بسجمله‌ایهای $p_0(\lambda)$، $p_1(\lambda)$، ...، $p_n(\lambda)$ یك **دنبالهٔ استورم**[2] تشکیل می‌دهند که به‌سرعت بازه‌هایی متضمن تنها یك ویژه‌مقدار را پدید می‌آورند.

□**مثال ۱۸.۴:** برای ماتریس مثال ۱۷.۴، رابطهٔ بازگشتی (۷۶.۴) به صورت سادهٔ زیر درمی‌آید

$$p_0(\lambda) = 1,\ p_1(\lambda) = -\lambda,\ p_j(\lambda) = -(\lambda p_{j-1}(\lambda) + p_{j-2}(\lambda)/4) \qquad j > 1$$

با انتخاب $n = 10$ و $\lambda = 0$ دنبالهٔ زیر حاصل می‌شود که پنج تغییرعلامت دارد

$$(p_j(\lambda)) = 1,\ 0,\ -\frac{1}{4},\ 0,\ \frac{1}{16},\ 0,\ -4^{-3},\ 0,\ 4^{-4},\ 0,\ -4^{-5}$$

$$= -0.0009765625$$

به‌ازای $\lambda = 0.2$، درعوض دنبالهٔ زیر حاصل می‌شود که شش تغییرعلامت را نشان می‌دهد.

$$(p_j(\lambda)) = 1,\ -0.2,\ -0.21,\ 0.0928,\ 0.0334,\ -0.0029,$$

$$-0.00025,\ 0.00079,\ -0.000095,\ -0.00018,\ 0.000059857$$

[در اینجا جز برای مقدار $p_{10}(\lambda)$، اعداد تا دو رقم اول با معنی داده شده‌اند.] درنتیجه دربازهٔ $[0.2, 0]$ دقیقاً یك ویژه‌مقدار B وجود دارد. الگوریتم قاعدهٔ تصحیح خطای

---

1. recurrence 2. Sturm sequence

اصلاح شده (الگوریتم ۳.۳) که با این بازه آغاز می‌شود در چهار مرحله (در یک هیولت‌ـ پاکارد[1] ۶۷) ویژه‌مقدار ۱۴۲۳۱۴۸۳۷ر۰ را به‌دست می‌دهد کـه بـا ویژه‌مقدار درست $\cos(5\pi/11) = 0.14231483827$ متناظر است (تمرین ۸.۰۴ـ۴ را ببینید). □

تمرین

۸.۰۴ـ۱ گیریم a، b عددوار (اسکالر) و A یك ماتریس مربعی باشد. ثابت کنید که اگر λ یك ویژه‌مقدار A باشد، آنگاه $a\lambda+b$ یك ویژه‌مقدار ماتریس $aA+bI$ خواهد بود. [راهنمایی: $(aA+bI)\mathbf{x}$ را که در آن $\mathbf{x}$ یك ویژه‌بردار متعلق به λ است، در نظر گیرید.]

۸.۰۴ـ۲ ثابت کنید کـه اگر λ یك ویژه‌مقدار ماتریس مربعی A و $p(x)$ یك بسجملـه‌ای غیرمشخصی باشد، آنـگاه $p(\lambda)$ یك ویژه‌مقدار $p(A)$ خواهد بود (تمرین ۱۰۴ـ۱۲ را ببینید).

۸.۰۴ـ۳ گیریم A یك ماتریس سه‌قطری از مرتبهٔ n با درایه‌های قطری صفر باشد و به‌ازای $i = 1, \dots, n-1$، $a_{i,i+1} = a_{i+1,i} = 1$. به‌ازای $j = 1, \dots, n$، گیریـم $\mathbf{x}^{(j)}$، n-برداری باشد که درایهٔ iام آن به‌ازای $i = 1, \dots, n$ به‌صورت

$$x_i^{(j)} = \sin[ij\pi/(n+1)]$$

است. ثابت کنید که

$$A\mathbf{x}^{(j)} = 2\cos\left(\frac{j\pi}{n+1}\right)\mathbf{x}^{(j)} \qquad j = 1, \dots, n$$

۸.۰۴ـ۴ با استفاده از تمرینهای ۸.۰۴ـ۱ و ۸.۰۴ـ۳ ثابت کنید که اگر A یك ماتریس سه‌قطری باشد کـه در آن به‌ازای جمیع مقادیـر i، $a_{ii} = d$ و $a_{i+1,i} = a_{i,i+1} = e$، آنگاه همهٔ ویژه‌مقدارهای A اعدادی به‌شکل زیرخواهند بود

$$d + 2e\cos\frac{j\pi}{n+1} \qquad j = 1, \dots, n$$

۸.۰۴ـ۵ با استفاده از روش توانی ویژه‌مقدار با قدرمطلق ماکسیمم و ویژه‌بردارمتناظر با آن را برای ماتریس سه‌قطری A از مرتبـهٔ ۲۰، با $a_{ii} = 4$ و $a_{i+1,i} = a_{i,i+1} = -1$ به‌ازای جمیع مقادیر i، برآورد کرده، بـا جواب دقیق حاصله از تمرین ۸.۰۴ـ۴ مقایسه کنید.

---

1. Hewlett-Packard 67

۶-۸-۴ سعی کنید که ویژه‌مقدار با قدرمطلق ماکسیمم ماتریس A در مثال ۸-۴-۳ (با فرض $n=۲۱$) را با استفاده از روش توانی برآورد کنید. درمورد مشکلاتی که با آن مواجه می‌شود، توضیح دهید.

۷-۸-۴ اگر ماتریسی دویا چندین ویژه‌مقدار با قدرمطلق ماکسیمم برابرداشته باشد، روش توانی کارا نیست. درمورد اینکه چگونه تمرین ۸-۴-۱ می‌تواند براین مشکل فائق آید، بحث و سعی کنید که این راه رفع مشکل را درمورد مسئلهٔ تمرین ۸-۴-۶ به‌کار برید.

۸-۸-۴ نشان دهید که ماتریس $B=\begin{bmatrix}۱ & ۱\\ ۰ & ۱\end{bmatrix}$ دارای، یك مجموعهٔ کامل از ویژه‌بردارها نیست.

۹-۸-۴ گیریم $\mathbf{x}$ و $\mathbf{y}$ دو ویژه‌بردار ماتریس A، به‌ترتیب متعلق به ویژه‌مقدارهای λ و μ باشند. نشان دهید که اگر $\mu\neq\lambda$، آنگاه $\mathbf{x}$ و $\mathbf{y}$ مستقل خطی هستند.

۱۰-۸-۴ با استفاده از تمرین ۸-۴-۹ نشان دهید که ماتریس $\begin{bmatrix}۱ & ۲\\ ۰ & ۲\end{bmatrix}$ می‌باید مجموعهٔ کاملی از ویژه‌بردارها داشته باشد.

۱۱-۸-۴ کلیهٔ ویژه‌مقدارهای ماتریس

$$A=\begin{bmatrix}۴ & -۱ & -۱ & -۱\\ -۱ & ۴ & -۱ & -۱\\ -۱ & -۱ & ۴ & -۱\\ -۱ & -۱ & -۱ & ۴\end{bmatrix}$$

را با تعیین صریح بسجمله‌ای مشخصهٔ آن محاسبه کنید، و سپس ریشهٔ این بسجمله‌ای را به‌دست آورید.

۱۲-۸-۴ ماتریس A در تمرین ۸-۴-۱۱ را از راه تقارنهای محوری هاوس‌هولدر به‌شکل سه‌قطری B تبدیل کنید (چون بسجمله‌ای مشخصهٔ A یك ریشهٔ سه‌گانه دارد، بر طبق مثال ۶-۴-۱۶، حداقل دوتا از b_iها صفرند)، سپس ویژه مقدارهای B را محاسبه کنید.

۱۳-۸-۴ کلیهٔ ویژه‌مقدارهای ماتریس سه‌قطری B درمثال ۷-۴-۱۷ را با استفاده از رابطهٔ بازگشتی (۶-۴-۷۶)، از ویژگی دنبالهٔ استورم برای جداسودن ویژه مقدارها، و سپس از روش نیوتن برای دستیابی به‌تك‌تك ویژه‌مقدارها، حساب کنید.

۱۴-۸۰۴ پس از حل تمرین ۸۰۴-۱۳، روش توانی معکوس[1] را برای تعیین ویژه‌بردارهای متناظر به‌کار برید.

۱۵-۸۰۴ ثابت کنید که یك ماتریس هرمیتی با یك ماتریس قطری متشابه است و همهٔ ویژه‌مقدارهای آن حقیقی هستند. (راهنمایی: نشان دهید که اگر B یك ماتریس هرمیتی باشد، آنگاه ماتریس بالا-مثلثی که در قضیهٔ شور به‌دست آمد، لزوماً هرمیتی است).

۱۶-۸۰۴ از روش مولر برای تعیین فراوانیهای طبیعی مثال ۱۶۰۴، درمورد

$$A=\begin{bmatrix} -2 & 0 & 1 & 0 \\ 0 & -2 & 0 & 1 \\ 1 & 0 & -2 & 0 \\ 0 & 1 & 0 & -2 \end{bmatrix} \qquad M=\begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

استفاده کنید.

۱۷-۸۰۴ فرض کنید که ماتریس A از مرتبهٔ n، دارای یك مجموعهٔ کامل از ویژه‌بردارهای $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ باشد. ثابت کنید که A با یك ماتریس قطری (که درایه‌های قطری آن باید ویژه مقدارهای A باشند) مشابه است. [راهنمایی: ماتریس $C^{-1}AC$ را که در آن به‌ازای $j=1,\ldots,n$ تساوی $C\mathbf{i}_j=\mathbf{x}^{(j)}$ برقرار است، در نظر بگیرید. چرا C می‌باید وارونپذیر باشد؟].

۱۸-۸۰۴ تقلیل اندازه برای روش توانی فرض کنید که با روش توانی یا روشهای دیگر، یك ویژه‌مقدار λ از ماتریس A از مرتبهٔ n و ویژه‌بردار $\mathbf{x}$ متناظر با آن را محاسبه کرده‌ایم، و فرض کنید که $x_n \neq 0$. گیریم B ماتریسی از مرتبهٔ $n-1$ باشد که از ماتریس $C^{-1}AC$ با حذف آخرین سطر و آخرین ستون به‌دست آمده و در آن، تساویهای $C\mathbf{i}_j=\mathbf{i}_j$ و $C\mathbf{i}_n=\mathbf{x}$، به‌ازای $j=1,\ldots,n-1$، برقرارند. ثابت کنید که کلیهٔ ویژه مقدارهای A، احتمالا به‌استثنای ویژه‌مقدار λ، ویژه‌مقدارهای B نیز هستند.

---

1. inverse power method

۵

دستگاه معادلات و بهینه‌سازی نامقید

یك دستگاه کلی n معادله و n مجهول برحسب x_1، x_2، ... ، x_n را می‌توان همواره به‌شکل

$$f_i(x_1, x_2, \cdots, x_n) = 0 \qquad i = 1, \cdots, n \qquad (1.5)$$

نوشت که در آن f_1، ... ، f_n توابع n متغیره هستند. نمادگذاری برداری مذکور در فصل ۴ را همچنان ادامه می‌دهیم و معادلهٔ (۱.۵) را به‌صورت فشرده[1]تر زیر می‌نویسیم

$$\mathbf{f}(\mathbf{x}) = 0 \qquad (2.5)$$

بنا بر این $\mathbf{f}$ تابعی است برداری مقدار[2]. مقدار این تابع به‌ازای n-برداری $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$ برابر با n-برداری $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_n(\mathbf{x})]^T$ است.

با این نمادگذاری نه‌تنها در نوشتن صرفه‌جویی‌شده‌است بلکه بیانگر این حقیقت‌است که روشهای بارستی که برای حل یك معادلهٔ یك مجهولی که در فصل ۳ مورد بحث قرار گرفت، در اینجا هم ازجهتی باید به‌کار آیند. بخصوص روش بارست نقطهٔ ثابت و روش نیوتن و بعضی از شکلهای مختلف این روش مورد بحث قرار می‌گیرند. ولی نمی‌توانیم تجزیه و تحلیل ریاضی این روشها را عمیقاً مورد بحث قرار دهیم. بحث جامع از کلیهٔ مطالب

1. compact 2. Vector-valued function of a vector

موجود در این زمینه‌را می‌توان درشرح تفصیلی ارتگا و راینهولت[۱][۳۳] براین موضوع، به‌دست آورد. حل‌دستگاه معادلات همچنان یك زمینهٔ فعال تحقیقاتی، بخصوص در ساختن الگوریتمهای کاراست.

یك مثال خاص از دستگاه (۱.۵) دستگاه خطی

$$A\mathbf{x}-\mathbf{b}=\mathbf{0}$$

است که حل مستقیم آن مفصلا درفصل۴ موردبحث قرار گرفت، ولی در اینجا دستگاه کلی (۲.۵) معمولا به‌صورت بارستی می باید حل شود؛ یعنی، بـا حل یك‌رشته از دستگاههای خطی که معمولا با روشهای مستقیم بحث شده درفصل۴، انجـام می‌گیرد. اما در بعضی از روشهای بارستی بخصوص روشهای واهلش[۲]،آن رشته ازدستگاههای خطی‌که می باید حل شوند به‌اندازه‌ای ساده هستندکه صرفه دراین‌است که‌این روشها، برای سیستمهایی‌که خود خطی می‌باشند به‌کارگرفته شوند (وبه‌کارگرفته شده‌اند). در حل دستگاههـای خطی، مـا به‌این‌گونه‌روشهای بارستی توجه خاصی خواهیم‌کرد.

وسرانجام، رابطهٔ نزدیك بین حـل دستگاههای معادلات و جستجو برای فرینه‌هـای یك تابع حقیقی n-متغیره‌را کـه در قسمت اول این فصل بیشتر توضیح داده خواهد شد، مطمح‌نظر قرار می‌دهیم.

۱.۵* بهینه‌سازی و تندترین کاهش[۳]

بهینه‌سازی منبعی‌است همیشگی برای حل دستگاههای معادلات، و پاره‌ای از روشهای حل دستگاهها مستقیماً تحت تأثیر آن قرار دارند. برای یادآوری، اگر بخواهیم مینیمم (یـا ماکسیمم) یك تابع حقیقی n-متغیره $F(\mathbf{x})=F(x_1,\ldots,x_n)$ را پیداکنیم،کافی است که به‌مقادیر آن تابع در نقاط بحرانی[۴] یعنی نقاطی از $\mathbf{x}$ که به‌ازای آنها رابطهٔ

$$\nabla F(\mathbf{x})=\mathbf{0}$$

برقراراست، نگاه‌کنیم. دراینجا ∇F گرادیان[۵] F یعنی بردار

$$\nabla F=\left[\frac{\partial F}{\partial x_1}\ \cdots\ \frac{\partial F}{\partial x_n}\right]^T$$

است که درایه‌های آن اولین مشتقهای جزئی تابع F هستند

$$f_i:=\frac{\partial F}{\partial x_i}\qquad i=1,\ldots,n$$

---

1. Ortega and Rheinholdt 2. relaxation 3. steepest descent
4. critical points 5. gradient

بدین‌علت $\nabla F(\mathbf{x})$ نوشته می‌شود که می‌خواهیم به‌نقطهٔ $\mathbf{x}$ که گرادیان در آن نقطهٔ محاسبه شده تأکید کنیم.

به‌یاد آورید که گرادیان ∇F همان «مشتق اول» تابع n-متغیره $F(\mathbf{x})$ است: بنا بر قضیهٔ ۸.۱ مشتق تابع یك متغیرهٔ

$$g(t)=F(\mathbf{x}+t\mathbf{u})$$

در $t=0$ به‌صورت زیر معین می‌شود

$$g'(0)=\nabla F(\mathbf{x})^T\mathbf{u}=f_1(\mathbf{x})u_1+\ldots+f_n(\mathbf{x})u_n$$

عدد بالا اطلاعاتی دربارهٔ رفتار تابع F، هنگامی که بخواهیم از نقطهٔ $\mathbf{x}$ در جهت $\mathbf{u}$ حرکت کنیم، به‌دست می‌دهد. بنابراین F در کلیهٔ جهتهای $\mathbf{u}$ که زاویه‌اش نسبت به‌بردار گرادیان $\nabla F(\mathbf{x})$ کمتر از 90° باشد با بیشترین سرعت درجهت گرادیان، افزایش می‌یابد. دلیل این امر وجود رابطهٔ زیراست

$$(\nabla F)^T\mathbf{u}=\|\nabla F\|_2\|\mathbf{u}\|_2\cos\theta. \qquad (۳.۵)$$

که در آن θ زاویهٔ بین دو بردار است. درحقیقت این نوعی اتحاد منطقی‌است زیرا اگر از ما خواسته شود که زاویهٔ θ بین دو بردار $\mathbf{u}$ و $\mathbf{v}$ را تعیین کنیم، معمولا به‌جواب زیرمتوسل می‌شویم

$$\theta=\cos^{-1}\left(\frac{\mathbf{v}^T\mathbf{u}}{\|\mathbf{v}\|_2\|\mathbf{u}\|_2}\right)$$

اما نکتهٔ حائزاهمیت این‌است که، به‌ازای $\nabla F\neq 0$، تقریباً نیمی از جهتهای ممکن $\mathbf{u}$، مثلا برای آنهایی که داریم $(\nabla F)^T\mathbf{u}>0$، موجب افزایش مقدار F می‌شوند وحداکثر افزایش فقط و فقط زمانی حاصل‌می‌شود که $\mathbf{u}$ با ∇F موازی باشد. با بحثی مشابه، تقریباً نیمی‌ازتمام جهتهای ممکن $\mathbf{u}$ موجب‌کاهش مقدار F می‌شوند و بیشترین‌کاهش زمانی‌است که $\mathbf{u}$ با $-\nabla F$ موازی باشد. درنتیجه $\mathbf{x}$ نمی‌تواندجزدرحالت $\nabla F=0$، برای Fمینیمم یا ماکسیمم باشد.

□ **مثال ۱۰۵:** تابع $F(\mathbf{x})=x_1^3+x_2^3-2x_1^2+3x_2^2-8$ دارای‌گرادیان

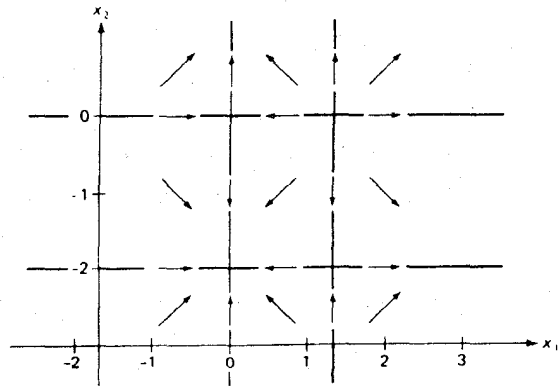
$$\nabla F=[3x_1^2-4x_1 \qquad 3x_2^2+6x_2]^T$$

است. بنابراین‌معادلهٔ $\nabla F(\mathbf{x})=0$ دارای چهارجواب $(0,0)$، $(2,0)$، $(0,-4/3)$ و $(2,-4/3)$ است. برای درك ماهیت این نقاط بحرانی، و برای آنکه تمرینی با گرادیانها داشته باشیم، اکنون نواحی مختلفی راکه در آنها صفحهٔ (x_1,x_2) به‌وسیلهٔ منحنیهای زیر قطع می‌شوند مورد بررسی قرار می‌دهیم

$$\left(\frac{\partial F}{\partial x_1}\right)(\mathbf{x})=0 \quad , \quad \left(\frac{\partial F}{\partial x_2}\right)(\mathbf{x})=0$$

ملاحظه می‌کنیم که $f_1=\partial F/\partial x_1$ روی دوخط مستقیم $x_1=0$ و $x_1=4/3$ صفر می‌شود و بین این خطوط مقدار آن منفی و درجاهای دیگرمثبت‌است. بنابراین اولین‌مؤلفهٔ[1] گرادیان ∇F بین این دوخط منفی و درجاهای دیگرمثبت‌است. همچنین $f_2=\partial F/\partial x_2$، که‌روی دو خط مستقیم $x_2=-2$ و $x_2=0$ صفر می‌شود، بین این خطوط منفی ودرجاهای‌دیگر مثبت‌است. بحث فوق تصویرکیفی شکل۱۰۵ را برای جهت گرادیان در نواحی مختلف که به‌وسیلهٔ خطوط $f_1=0$ و $f_2=0$ تعریف می‌شوند به‌دست می‌دهد. از روی این شکل مشخص‌است که نقطهٔ بحرانی $(-2, 0)$ یك ماکسیمم موضعی‌است (زیرا کلیهٔ گرادیانهای مجاور آن، به‌این نقطه متوجه‌اند) درحالی‌که نقطهٔ بحرانی $(3/4, 0)$ یك مینیمم موضعی است (زیرا کلیهٔ گرادیانهای مجاور آن از این نقطه رو برمی‌گردانند). دو نقطهٔ بحرانی دیگر، نقاط فرین[2] نیستند بلکه نقاط زینی[3] هستند، زیرا درمجاورت آنها گرادیانهایی قرار دارندکه هم به‌آنها متوجه‌اند و هم از آنها رو برمی‌گردانند. □

یك روش اساسی برای پیداکردن نقطهٔ فرین، **روش تندترین‌کاهش (یا افزایش)** است. این روش به‌روش کوشی[4] برمی‌گردد و درآن سعی بر آن‌است‌که مسئلهٔ پیدا کردن مینیمم یك تابع حقیقی n متغیره، از راه چندبار پیداکردن مینیمم یك تابع یك متغیره حل شود. فکر اساسی به‌شرح زیراست: یك تقریب $\mathbf{x}$ برای مقدار مینیمم $\mathbf{x}^*$ از تابع F داده شده‌است. سپس درطول خط مستقیم ماربر $\mathbf{x}$ ودرجهت $-\nabla F(\mathbf{x})$ به‌دنبال آن مینیممی‌از



شکل۱۰۵ نمایش جهتهای‌گرادیان برای تابع $F(x_1, x_2)=x_1^3+x_2^3-2x_1^2+3x_2^2-8$.

1. component 2. extrema 3. saddle
4. Cauchy

از F که به $\mathbf{x}$ ازهمه نزدیکتر است می گردیم. یعنی مینیمم $t^* > 0$ را که نزدیکترین نقطه به‌صفر تابع

$$g(t)=F(\mathbf{x}-t\nabla F(\mathbf{x}))$$

که تابعی است از یک متغیر تصادفی[1] پیدا کنیم، و پس از پیدا کردن آن، تقریب بعدی برای مینیمم $\mathbf{x}^*$ را نقطهٔ $\mathbf{x}-t^*\nabla F(\mathbf{x})$ می گیریم.

الگوریتم ۱۰۵: تندترین کاهش تابع هموار $F(\mathbf{x})$ از n-بردار $\mathbf{x}$ و یک تقریب $\mathbf{x}^{(0)}$ برای مینیمم (موضعی) $\mathbf{x}^*$ از F، داده شده‌اند.

For $m=0, 1, 2, \ldots$, do until satisfied:

$\mathbf{u} := \Delta F(\mathbf{x}^{(m)})$

اگر $\mathbf{u}$ مساوی ۰ شد، آنگاه الگوریتم را **متوقف** سازید.
درغیر این صورت مینیمم $t^* > 0$ را که به ۰ تابع $g(t)=F(\mathbf{x}^{(m)}-t\mathbf{u})$ از همه نزدیکتر باشد تعیین کنید

$\mathbf{x}^{(m+1)} := \mathbf{x}^{(m)}-t^*\mathbf{u}$

□ **مثال ۲۰۵:** براورد $\mathbf{x}^{(0)}=[1, -1]^T$ برای مینیمم موضعی $(0, 4/3)$ از تابع $F(x_1, x_2)=x_1^3+x_2^3-2x_1^2+3x_2^2-8$ درمثال ۱.۵ داده شده‌است، داریم

$$\nabla F(\mathbf{x}^{(0)})=[-1, -3]^T$$

بنابراین در اولین مرحله از روش تندترین کاهش، به‌دنبال یافتن مینیمم تابع

$$g(t)=F(1+t, -1+3t)$$
$$=(1+t)^3+(-1+3t)^3-2(1+t)^2+3(-1+3t)^2-8$$

می‌رویم. با قراردادن $g'(t)=0$ خواهیم داشت

$$0=3(1+t)^2+3(3t-1)^2 3-4(1+t)+3\times 2(3t-1)3$$
$$=84t^2+2t-10$$

که دارای دوجواب $-5/14$ یا $1/3=168/(-2\pm\sqrt{4+3360})=t^*$ است (با استفاده از فرمول حل معادلهٔ درجهٔ دوم). ریشهٔ مثبت $t^*=1/3$ را انتخاب می‌کنیم زیرا می‌خواهیم از $\mathbf{x}^{(0)}$ درجهت گرادیان $-\nabla F(\mathbf{x}^{(0)})$ حرکت کنیم. این انتخاب، خودمینیمم یعنی $\mathbf{x}^{(1)}=[4/3, 0]^T$ را به‌دست می‌دهد. □

1. univariate

واضح است که روش تندترین کاهش، ضامن کاهش مقدار تابع در هر مرحله است، یعنی $F(\mathbf{x}^{(m+1)}) < F(\mathbf{x}^{(m)})$، (مگر در موقعی که $\nabla F(\mathbf{x}^{(m)}) = 0$). این امر را می‌توان پایه‌ای برای اثبات همگرایی (با فرض $\|x^{(m)}\| \geqslant$ یك عدد ثابت، به‌ازای جمیع مقادیر m) روش فوق قرارداد. ولی به‌سادگی می‌توان مثالهایی ارائه داد که نشان دهند روش فوق ممکن است خیلی به‌کندی همگرا شود.

□ **مثال ۳.۵:** تابع $F(\mathbf{x}) = x_1^2 + \alpha x_2^2$ با $\alpha > 0$، دارای یك مینیمم کلی[1] در $\mathbf{x} = 0$ است. گرادیان این تابع، خطی و به‌صورت

$$\nabla F = [2x_1, 2\alpha x_2]^T$$

است بنابراین بلافاصله می‌توان نقطهٔ بحرانی یگانهٔ آن را از دستگاه

$$2x_1 = 0$$

$$2\alpha x_2 = 0$$

به‌دست آورد. اما برای اینکه نکته‌ای را متذکر شویم به‌جای استفاده از روش فوق، از روش تندترین کاهش استفاده می‌کنیم، این امر مستلزم تعیین مینیمم تابع

$$g(t) = F(\mathbf{x} - t\nabla F(\mathbf{x}))$$

$$= F(x_1(1-2t), x_2(1-2\alpha t))$$

است. با قراردادن $g'(t) = 0$، معادلهٔ زیر را به‌دست خواهیم آورد

$$0 = 2(x(1-2t))(-2) + \alpha 2(x(1-2\alpha t))(-2\alpha)$$

که جواب آن عبارت است از $t^* = (1/2)(x_1^2 + \alpha^2 x_2^2)/(x_1^2 + \alpha^3 x_2^2)$. بنابراین اگر حدس معمولی ما $\mathbf{x} = [x_1, x_2]^T$ باشد، آنگاه

$$\frac{x_1 x_2(\alpha - 1)}{x_1^2 + \alpha^3 x_2^2}[\alpha^2 x_2 - x_1]^T$$

حدس بعدی ما خواهد بود.

اکنون $\mathbf{x}$ را به‌شکل خاص $c[\alpha, \pm 1]^T$ می‌گیریم. در این حال حدس بعدی ما به‌صورت

$$c\frac{\alpha - 1}{\alpha + 1}[\alpha, \mp 1]^T$$

---

1. global

خواهد شد. یعنی خطا به‌نسبت $(\alpha-1)/(\alpha+1)$ کاهش یافته است. برای مثال اگر $\alpha=100$ و $\mathbf{x}^{(0)}=[1, 0.01]^T$، آنگاه بعد از ۱۰۰ مرحله استفاده از روش تندترین کاهش، نقطهٔ زیر را خواهیم داشت

$$\mathbf{x}^{(100)}=\left(\frac{\alpha-1}{\alpha+1}\right)^{100}[1, 0.01]^T=[0.135\ldots, 0.00135\ldots]^T$$

که نسبت به‌حدس اولیه، جواب باز به‌مقدار ۷/۸ کمتراست. □

درشکل۲.۵ قسمتی از بارست تندترین‌کاهش برای مثال ۲.۵ نشان داده شده‌است. برای فهمیدن این شکل می‌باید به‌دونکتهٔ زیرتوجه‌کرد:

(i) از آنجاکه $(d/dt)F(\mathbf{x}+t\mathbf{u})=(\nabla F(\mathbf{x}+t\mathbf{u}))^T\mathbf{u}$، بنابرقضیهٔ ۸.۱، گرادیان F درنقطهٔ مینیمم $\mathbf{x}-t^*\nabla F(\mathbf{x})$ ازتابع F، درجهت گرادیان منفی، برامتدادآن عمود[1] است، یعنی

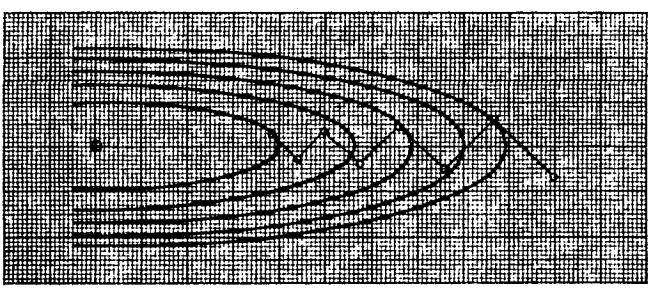
$$\nabla F(\mathbf{x}^{(m+1)})^T\nabla F(\mathbf{x}^{(m)})=0$$

(ii) یك تابع دومتغیرهٔ $F(x_1, x_2)$ درصفحهٔ (x_1, x_2) غالباً به‌وسیلهٔ خطوط تراز یا خطوط مرزی[2] آن یعنی با منحنیهای

$$F(x_1, x_2)=\text{ثابت}$$

توصیف می‌شود. این خطوط درشکل ۲.۵ نشان داده شده‌اند. از آنجاکه الزاماً گرادیان دریك‌نقطه برخط تراز ماربرآن نقطه عمود است (تمرین ۱.۵–۳) بنابراین خطوط مزبور اطلاعاتی دربارهٔ جهت گرادیان به‌دست می‌دهند.

همان‌طوری‌که این‌مثال نشان می‌دهد انتخاب جهت تندترین‌کاهش ممکن‌است تدبیر



شکل۲.۵ زمانی‌که به‌دنبال یك مینیمم دریك محدودهٔ باریك باشیم، روش‌تندترین کاهش ممکن‌است به‌نحو غیرمؤثری موجب پس‌وپیش‌شدن شود.

1. perpendicular 2. contour lines

خوبی به‌حساب آید، اما غالباً طرح دقیق و مناسبی نیست.

امروزه روشهای کاهش پیچیده‌تری به‌کار می‌رود که در آن $\mathbf{x}^{(m+1)}$ از $\mathbf{x}^{(m)}$ به‌گونهٔ زیر به‌دست می‌آید

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + t_m \mathbf{u}^{(m)} \qquad (۴.۵)$$

در اینجا، $\mathbf{u}^{(m)}$ **یك جهت کاهش**[1] است، یعنی $\nabla F(\mathbf{x}^{(m)})^T \mathbf{u}^{(m)} < ۰$، و t_m از راه **خط جویی**[2]، یعنی با مینیمم کردن تقریبی تابع

$$g(t) = F(\mathbf{x}^{(m)} + t\mathbf{u}^{(m)})$$

پیدا می‌شود. اگر گرادیان F در دسترس باشد، آنگاه عمل خطجویی به‌پیداکردن ریشهٔ مناسب تابع

$$g'(t) = \nabla F(\mathbf{x}^{(m)} + t\mathbf{u}^{(m)})^T \mathbf{u}^{(m)}$$

تبدیل می‌شود، و برای این منظور روشهای فصل سه را می‌توان به‌کار گرفت. گرچه باید به‌یاد داشته باشیم که دقتی که همراه این ریشه است به‌درجهٔ نزدیکی ریشهٔ مزبور به‌مینیمم $F(\mathbf{x})$ بستگی دارد.

اگر گرادیان F در دسترس نباشد (یا تصورشود که محاسبهٔ آن پرهزینه‌است) آنگاه پیشنهاد می‌شود که به‌شکلی از **درونیابی درجهٔ دوم** استفاده به‌عمل آید. در زیر نمونه‌ای از آن آمده‌است.

الگوریتم ۲.۵: خطجویی از راه درونیابی درجهٔ دوم تابع $g(t)$ با $g'(۰) < ۰$، و یك عدد مثبت $t_{\max}$ و تحمل (مساهله[3]) مثبت ε داده شده‌اند

۱. $s_۱ := ۰$

۲. $s_۲$ و $s_۳$ را طوری انتخاب کنید که $t_{\max} \geqslant s_۳ \geqslant s_۲ \geqslant ۰$ و $g[s_۱, s_۲] \geqslant ۰$.

۳. **اگر**[4] $s_۲ = s_۳ = t_{\max}$، آنگاه قرار دهید $t_m := t_{\max}$ و از الگوریتم **خارج کنید**[5] **وگرنه**[6] سهمی $p_۲(t)$ را که در نقاط $s_۱$، $s_۲$، $s_۳$ بر $g(t)$ منطبق می‌باشد بررسی کنید.

۴. **اگر** $g[s_۱, s_۲, s_۳] \geqslant ۰$، بنابراین $p_۲(t)$ مینیمم ندارد، و سپس قرار دهید $t_{\max} := s_۳$ و **به‌مرحلهٔ سوم بروید**[7] **وگرنه** مینیمم s را در $p_۲(t)$ محاسبه کنید، یعنی

$$s := (s_۱ + s_۲ - g[s_۱, s_۲]/g[s_۱, s_۲, s_۳])/۲$$

---

1. descent direction 2. line search 3. tolerance
4. IF 5. EXIT 6. ELSE 7. GO TO

۰۵. **اگر** $s > t_{\max}$، آنگاه قراردهید $(s_1, s_2, s_3) := (s_2, s_3, t_{\max})$ **و به مرحلهٔ سوم بروید**

۱۰۵ اگر $|g(s) - \min_i g(s_i)| \leqslant \varepsilon$ یـا $|g(s) - p_2(s)| \leqslant \varepsilon$، آنگاه قرار دهید $t_m := s$ **و از الگوریتم خارج شوید.**

وگرنه یك سری سه نقطه‌ای مرتب‌جدید[1] (s_1, s_2, s_3) ازمجموعهٔ چهار نقطه‌ای $\{s_1, s_2, s_3, s\}$ چنان‌انتخاب کنیدکه یارابطهٔ $g[s_1, s_3] < 0 < g[s_2, s_3]$ برقرار باشد یا، اگر برقراری رابطهٔ فوق امکان‌پذیر نباشد، $\max_i g(s_i)$ تاحد ممکن کوچك شودوسپس **به‌مرحلهٔ ۴ بروید.**

درخروج ازالگوریتم، t_m به‌عنوان تقریبی از مینیمم $g(t)$ در $[0, t_{\max}]$ محسوب می‌شود.

توجه‌کنیدکه **خروج** ازمرحلهٔ ۱۰۵ تضمین‌نمی‌کندکه t_m محاسبه شده به‌مینیمم $g(t)$ «نزدیك» باشد، تمرین۱۰۵–۵ رابینید. وقتی‌که‌الگوریتم ۲۰۵ به‌عنوان قسمتی‌ازالگوریتم مینیمم‌سازی چند متغیره[2] به‌کار گرفته می‌شود، معمولا این الگوریتم با $s_3 = s_2 = 0$ شروع می‌شود [زیرا معمولا $g'(0) = \nabla F(\mathbf{x}^{(m)})^T \mathbf{u}^{(m)}$ دردست‌است] و $s_3 = t_{\max} = 1$، ومرحلهٔ ۱۰۵ به‌شکل «$t_m := s$ قرارداده و از الگوریتم **خارج شوید**» ساده می‌شود. می‌توان نشان دادکه این‌عمل اشکالی ندارد، به‌شرطی‌که جهت جستجو یعنی $\mathbf{u}^{(m)}$ طوری انتخاب شودکه $\mathbf{x}^{(m)} + \mathbf{u}^{(m)}$ یك مینیمم موضعی از یك تابع درجهٔ دوم، که تقریبی از تابع F در حوالی $\mathbf{x}^{(m)}$ است، باشد.

اشاره‌کردیم‌که بهینه‌سازی منجربه‌پیدایش دستگاههای معادلات، یعنی، دستگاههایی به‌صورت خاص

$$\nabla F(\mathbf{x}) = 0$$

می‌شود. برعکس دستگاه دلخواه از n معادلهٔ n مجهولی

$$\mathbf{f}(\mathbf{x}) = 0$$

را می‌توان اساساً به‌وسیلهٔ بهینه‌سازی حل‌کرد، زیرا مثلا هرمینیمم تابع

$$F(\mathbf{x}) := \|\mathbf{f}(\mathbf{x})\|_2^2 = (f_1(\mathbf{x}))^2 + \cdots + (f_n(\mathbf{x}))^2 \qquad (۵.۵)$$

یك جواب معادلهٔ $\mathbf{f}(\mathbf{x}) = 0$ است و بالعکس. برای این تابع خاص F داریم

$$\frac{\partial F}{\partial x_i} = \sum_{j=1}^{n} 2 f_j(\mathbf{x}) \left(\frac{\partial f_j}{\partial x_i}\right)$$

---

1. new ordered three-point sequence
2. multivariate minimization algorithm

یا

$$\nabla F = 2(\mathbf{f}')^T\mathbf{f} \qquad (۶.۵)$$

که در آن

$$\mathbf{f}' := \left(\frac{\partial f_i}{\partial x_j}\right)_{i,j=1}^{n}$$

ماتریس ژاکوبی[1] تابع بردار-مقداری $\mathbf{f}$ است.

تمرین

۱-۱۰۵ نقاط بحرانی تابع

$$F(x_1, x_2) = x_1^3/3 + x_2^2 x_1 + 3$$

را با ترسیم منحنیهای $\partial F/\partial x_1 = 0$ و $\partial F/\partial x_2 = 0$ پیدا کنید. سپس با استفاده از جهتهای گرادیان مجاور آنها، این نقاط را به‌نقاط ماکسیمم، مینیمم و نقاط زینی دسته‌بندی کنید.

۲-۱۰۵ با استفاده از روش تندترین کاهش و افزایش، مینیمم و ماکزیمم تابع مذکور در تمرین ۱۰۵-۱ را تا 10^{-6} رقم صحیح پیدا کنید.

۳-۱۰۵ گیریم $\mathbf{u}$ جهت تماس بر یک خط تراز، ثابت $= F(x_1, x_2)$ در نقطهٔ $\mathbf{x} = [x_1, x_2]^T$ باشد. با استفاده از قضیهٔ ۸.۱ ثابت کنید که $(\nabla F)(\mathbf{x})^T\mathbf{u} = 0$.

۴-۱۰۵ یک زیربرنامهٔ **فورترن** برای اجرای الگوریتم ۲.۵ بنویسید، سپس از آن برای حل تمرین ۱.۵-۲ استفاده کنید (توجه: پیدا کردن ماکسیمم تابع F — مثل پیدا کردن مینیمم تابع F — است).

۵-۱۰۵ (س. ر. رابینسن[2] [۳۴]): گیریم $h(t)$ یک تابع هموار در بازهٔ $[a, b]$ باشد، و $h''(t) > 0$ و $h(a) = h(b)$.

(الف) ثابت کنید که $h(t)$ دارای یک مینیمم منحصر به‌فرد t^* در $[a, b]$ است.

(ب) پیدا کردن t^* را با انتخاب بازهٔ $[\alpha, \beta]$ که متضمن t^* باشد، وسپس با به‌کار گیری الگوریتم ۲.۵ برای ورودی $t_{\max} = \beta - \alpha$ و $g(t) = h(t-\alpha)$، به ازای مقداری مانند $\varepsilon > 0$ و انتخاب اولیهٔ ($t_{\max}$ و $t_{\max}/2$ و ۰) به‌ازای (s_1, s_2, s_3)، در نظر بگیرید. در این صورت براورد حاصلهٔ t_m برای t^* بستگی به‌مقادیر α و β، ε پیدا می‌کند. ثابت کنید: اگر، به‌ازای کلیهٔ این‌گونه مقادیر α و β، داشته باشیم $\lim_{\varepsilon\to 0} t_m = t^*$، آنگاه $h(t)$ می‌باید

---

1. Jacobian matrix 2. S. R. Robinson

یك سهمی باشد.[راهنمایی: α و β را طوری انتخاب کنید که تساوی $h(\alpha)=h(\beta)$ برقرار باشد.]

(ج) نتیجه‌گیری کنید که الگوریتم ۲.۵ ممکن‌است کلا براورد خوبی برای مینیمم g به‌دست ندهد (حتی اگر ε بسیار کوچك باشد)، مگر اینکه g به‌یك سهمی نزدیك باشد.

۱.۵ـ۶: تقریب با روش کوچکترین توانهای دوم یك کار محاسباتی معمولی مستلزم تعیین پارامترهای $a_1, \dots, a_k$ است به‌طوری که مدل $y=R(x;a_1,\dots,a_k)$ با داده‌های (x_i, y_i) به‌ازای $i=1,\dots,N$، به‌خوبی ببرازد[1]، یعنی به‌طوری که

$$R(x_i;a_1,\dots,a_k)=y_i+\varepsilon_i \qquad i=1,\dots,N$$

که N-برداری $\varepsilon=[\varepsilon_1\ \varepsilon_2\ \dots\ \varepsilon_N]^T$ تا حد ممکن کوچك باشد.

(الف) با فرض اینکه R به‌طور هموار[2] به‌بردار پارامتری $\mathbf{a}=[a_1 a_2 \dots a_k]^T$ بستگی داشته باشد، نشان دهید که انتخاب $\mathbf{a}^*$ که موجب مینیمم‌شدن $\|\varepsilon\|_2$ می‌شود، می‌باید در معادلات به‌اصطلاح نرمال[3] زیر صدق کند

$$A[R(x_1;\mathbf{a}^*), R(x_2;\mathbf{a}^*), \dots, R(x_N;\mathbf{a}^*)]^T = Ay$$

که در رابطهٔ بالا، ماتریس A از مرتبهٔ $k\times N$ با رابطهٔ زیر معین می‌شود

$$\left(\frac{\partial R}{\partial a_i}(x_j;a^*)\right)$$

(ب) اعداد خاص a_1، a_2 را درمدل

$$y=a_1 e^{a_2 x}$$

طوری پیدا کنید که با تعبیر فوق، به‌بهترین‌وجه با مشاهدات زیر برازنده باشد.

x_i	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
y_i	۱٫۴۸	۱٫۱۰	۰٫۸۱	۰٫۶۱	۰٫۴۵	۰٫۳۳	۰٫۲۴	۰٫۱۸	۰٫۱۳	۰٫۱۰

۲.۵* روش نیوتن

درفصل۳، هنگام حل یك معادلهٔ یك مجهولی

$$f(\xi)=0$$

1. fits 2. smoothly 3. normal equations

برحسب ξ، روش نیوتن را این گونه به‌دست آوردیم که (اولا) از بسط تیلر

$$f(x+h)=f(x)+f'(x)h+\mathcal{O}(h^2)$$

تابع f در نقطهٔ x استفاده کردیم و سپس (ثانیاً) از جملهٔ بالاترین رتبهٔ $\mathcal{O}(h^2)$ صرفنظر کردیم، معادلهٔ «خطی‌شدهٔ»

$$0=f(x)+f'(x)h$$

را به‌جای معادلهٔ اصلی $0=f(x+h)$، نسبت به h حل کردیم و درنتیجه

$$h=-f(x)/f'(x)$$

و تقریب «بهبود یافتهٔ»

$$x-f(x)/f'(x)$$

را به‌دست آوردیم. اکنون که سعی داریم یك n-بردار ξ راطوری معین کنیم که دردستگاه n معادلهٔ

$$\mathbf{f}(\xi)=0$$

صدق کند، عیناً به‌همان روش عمل می‌کنیم. بنابر قضیهٔ ۹.۱ می‌دانیم که مؤلفهٔ iام تابع f_i از تابع بردار-مقداری $\mathbf{f}$ در رابطهٔ زیرصدق می‌کند

$$f_i(\mathbf{x}+\mathbf{h})=f_i(\mathbf{x})+(\nabla f_i(\mathbf{x}))^T\mathbf{h}+\mathcal{O}(\|\mathbf{h}\|^2)$$

و آن درصورتی‌است که f_i مشتق جزئی و پیوستهٔ مراتب اول و دوم داشته باشد. بنابراین داریم

$$f(\mathbf{x}+\mathbf{h})=\mathbf{f}(\mathbf{x})+\mathbf{f}'(\mathbf{x})\mathbf{h}+\mathcal{O}(\|\mathbf{h}\|^2) \qquad (۷.۵)$$

که در آن ماتریس $\mathbf{f}'$، ماتریس ژاکوبی $\mathbf{f}$ در $\mathbf{x}$ نامیده می‌شود و به‌صورت زیرداده می‌شود

$$\mathbf{f}'(\mathbf{x})=(\partial f_i/\partial x_j)_{i,j=1}^n$$

مجدداً از جملهٔ بالاترین مرتبه $\mathcal{O}(\|\mathbf{h}\|^2)$ صرفنظر و معادلهٔ «خطی‌شدهٔ»

$$0=\mathbf{f}(\mathbf{x})+\mathbf{f}'(\mathbf{x})\mathbf{h}$$

را به‌جای معادلهٔ اصلی $0=\mathbf{f}(\mathbf{x}+\mathbf{h})$ نسبت به‌مقدار تصحیحی $\mathbf{h}$ حل می‌کنیم و در نتیجه جواب

$$\mathbf{h}=-\mathbf{f}'(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$$

را به‌شرط آنکه ماتریس ژاکوبی $\mathbf{f}'(\mathbf{x})$ وارونپذیر باشد، به‌دست می‌آوریم. بدین‌طریق برای

ξ تقریب جدید

$$\mathbf{x}-\mathbf{f}'(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$$

را به‌دست می‌آوریم. این مرحله، مرحلهٔ اساسی روش نیوتن برای یك دستگاه است. البته معادلهٔ نیوتن

$$\mathbf{f}'(\mathbf{x})\mathbf{h}=-\mathbf{f}(\mathbf{x})$$

نسبت به‌مقدار تصحیحی $\mathbf{h}$ برحسب $\mathbf{x}$، یك دستگاه خطی است و باروشهای مستقیم مذکور درفصل۴ حل می‌شود.

الگوریتم ۳.۵: روش نیوتن برای یك دستگاه معادلات. دستگاه n معادلهٔ n مجهولی

$$\mathbf{f}(\xi)=0$$

که درآن $\mathbf{f}$ تابعی‌است مقدار-برداری و دارای مؤلفه‌های هموار همراه بـا حدس اولیـهٔ $\mathbf{x}^{(0)}$ برای جواب ξ ازدستگاه فوق، داده شده‌است

For $m = 0, 1, 2, \ldots,$ until satisfied, do:
$\quad \mathbf{x}^{(m+1)} := \mathbf{x}^{(m)} - \mathbf{f}'(\mathbf{x}^{(m)})^{-1}\mathbf{f}(\mathbf{x}^{(m)})$

می‌توان نشان دادکـه روش نیوتن به‌سمت ξ همگرا می‌شود، به‌شرط آنـکه $\mathbf{x}^{(0)}$ به‌اندازهٔکافی به ξ نزدیك بـاشد، و به‌شرطی کـه ماتریس $\mathbf{f}'$، ژاکوبی $\mathbf{f}$، پیوسته و $\mathbf{f}'(\xi)$ وارونپذیر باشد. بعلاوه اگرمشتقات جزئی مرتبهٔ دوم مؤلفه‌های تابع $\mathbf{f}$ نیز پیوسته باشند، آنگاه رابطهٔ

$$\|\xi-\mathbf{x}^{(m+1)}\| \leqslant c\|\xi-\mathbf{x}^{(m)}\|^2$$

به‌ازای مقدار ثابتی‌مانند c و جمیع مقادیر m به‌اندازهٔکافی بزرگ برقراراست. به‌عبارت دیگر، روش نیوتن به‌صورت عبارت درجهٔ دوم همگرا می‌شود (مثال ۶.۵ را نگاه کنید).

□ **مثال۴.۵:** اعداد $1 > \xi_n < \cdots < \xi_2 < \xi_1 < \xi_0 = 0$ راطوری تعیین‌کنیدکه به‌ازای $\xi_0 = 0$ و $\xi_{n+1} = 1$ و $G(x)=x^3$ داشته باشیم

$$G'(\xi_i)=G[\xi_{i-1}, \xi_{i+1}] \qquad i=1,\ldots,n$$

این‌کار مستلزم حل دستگاه

$$3\xi_i^2=(\xi_{i+1}^3-\xi_{i-1}^3)/(\xi_{i+1}-\xi_{i-1}) \qquad i=1,\ldots,n$$

یا

$$\mathbf{f}(\xi)=0$$

است که در آن $\xi_0 = 0 < \xi_1 < \cdots < \xi_{n+1} = 1$ و

$$f_i(\mathbf{x}) = x_{i-1}^2 + x_{i-1}x_{i+1} + x_{i+1}^2 - 3x_i^2 \qquad i = 1, \cdots, n$$

متناظراً، ماتریس ژاکوبی $\mathbf{f}'(\mathbf{x})$ ماتریس سه‌قطری به‌شکل زیر است

$$f'(x) = \begin{bmatrix} -6x_1 & 2x_2 & & & & \\ 2x_1 + x_3 & -6x_2 & x_1 + 2x_3 & & & \\ & 2x_2 + x_4 & -6x_3 & x_2 + 2x_4 & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & 2x_{n-2} + x_n & -6x_{n-1} & x_{n-2} + 2x_n \\ & & & & 2x_{n-1} + 1 & -6x_n \end{bmatrix}$$

بنابراین درحل معادلهٔ نیوتن

$$\mathbf{f}'(\mathbf{x})\mathbf{h} = -\mathbf{f}(\mathbf{x})$$

نسبت به‌عضو تصحیحی $\mathbf{h}$ برای $\mathbf{x}$، الگوریتم ۳.۴ درحل یك دستگاه خطی با ماتریس ضرایب سه‌قطری به‌کار گرفته می‌شود.

می‌توان نشان داد که این مسئله دقیقاً یك جواب دارد. توجه کنید که ماتریس ژاکوبی $\mathbf{f}'(\xi)$ به‌ازای هر جواب ξ و با شرط $\xi_n < \cdots < \xi_1$ ماتریس نافذ قطری مؤکد است (به‌تمرین ۲.۵–۲ نگاه کنید) بنابراین $\mathbf{f}'(\xi)$ وارونپذیر است. لذا اگر حدس اولیهٔ $\mathbf{x}^{(0)}$ به‌اندازهٔ کافی نزدیك به ξ انتخاب شود، انتظار می‌رود که همگرایی به‌صورت درجهٔ دوم باشد.

به‌ازای $\mathbf{x}^{(0)} = [1 \quad 2 \cdots n]^T/(n+1)$ و $n = 3$، مسئله را حل می‌کنیم و بارسته‌ها و خطاهای زیر را به‌دست می‌آوریم.

m	$\mathbf{x}^{(m)}$			$\|\mathbf{f}(\mathbf{x}^{(m)})\|_1$	$\|\mathbf{h}\|_1$
0	0.2500000	0.5000000	0.7500000	0.188 + 0	0.889 − 1
1	0.3583333	0.6000000	0.8083333	0.340 − 1	0.135 − 1
2	0.3386243	0.5856949	0.8018015	0.109 − 2	0.426 − 3
3	0.3379180	0.5852901	0.8016347	0.157 − 5	0.512 − 6
4	0.3379171	0.5852896	0.8016345	0.284 − 11	0.823 − 12

همگرایی درجهٔ دوم، هم درکاهش مقدار خطای باقیمانده[1] ($F(X^{(x)})$و هم درکاهش مقدار تصحیحی نیوتن یعنی $\mathbf{h}$ برای $\mathbf{x}^{(m)}$، مشخص است. محاسبات لازم با دقت مضاعف (تقریباً ۱۷ رقم اعشاری) روی کامپیوتر UNIVAC ۱۱۱۰ انجام گرفته‌است.

استفاده از روش نیوتن، اشکالاتی به‌همراه داردکه دراین مثال ساده آشکار نیست. عموماً دو اشکال عمده وجود دارد: (۱) عدم همگرایی به‌دلیل حدس نامناسب اولیه و (۲) هزینهٔ ساختن معادلهٔ نیوتن به‌طور صحیح وحل آن نسبت به‌مقدار تصحیحی $\mathbf{h}$. در زیر، هردو اشکال را به‌طور جداگانه مورد بحث قرار می‌دهیم.

دونظرکه در امر تقویت، یا حداقل درتسریع به‌همگرایی تا حدی موفق بوده‌اند عبارت‌اند از تداوم[2] یا نشانیدن[3] و میراندن[4]. در **تداوم**، مسئلهٔ حل $\mathbf{f}(\xi)=\mathbf{o}$ را به‌نحو مناسبی به‌عنوان آخرین مسئله از یك خانواده مسائل یك پارامتری پیوستهٔ

$$\mathbf{g}(\xi,\ t)=\mathbf{o}$$

با شرط

$$\mathbf{g}(\mathbf{x},\ 1)=\mathbf{f}(\mathbf{x})$$

مورد توجه قرار می‌دهند، که $\mathbf{g}(\mathbf{x},\ 0)$ تابعی است که درحل

$$\mathbf{g}(\xi,\ 0)=\mathbf{o}$$

مشکلی ایجاد نخواهد کرد.

پس از اینکه $\xi^{(0)}$طوری پیداشدکه درتساوی $\mathbf{g}(\xi^{(0)},\ 0)=\mathbf{o}$ صادق باشد، دنبالهٔ $t_0=0<t_1<\cdots<t_N=1$ را انتخاب و معادلهٔ

$$\mathbf{g}(\xi^{(i)},t_i)=\mathbf{o}$$

را به‌ازای $i=1,\ 2,\ \ldots,\ N$ و با روش نیوتن و با استفاده از بردار $\xi^{(i-1)}$ به‌عنوان حدس اولیه و یا شاید، حتی بردار برونیابی‌شدهٔ

$$\xi^{(i-1)}+(t_i-t_{i-2})\frac{\Delta\xi^{(i-2)}}{\Delta t_{i-2}}$$

(اگر $i>1$) حل می‌کنند. امید این است‌که مسائل مجاور هم یعنی $\mathbf{g}(\xi,\ t_i)=\mathbf{o}$ و $\mathbf{g}(\xi,\ t_{i-1})=\mathbf{o}$ آنقدر به‌هم نزدیك باشندکه یك جواب خوب از یکی، یك حدس اولیهٔ مناسبی برای جواب دیگری باشد. انتخاب معمولی برای تابع $\mathbf{g}$ عبارت‌است از

$$\mathbf{g}(\mathbf{x},\ t)=t\mathbf{f}(\mathbf{x})+(1-t)(\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{x}^{(0)}))$$

---

1. residual error
2. continuation
3. imbedding
4. damping

$$\mathbf{g}(\mathbf{x}, t)=t\mathbf{f}(\mathbf{x})+(1-t)(\mathbf{x}-\mathbf{x}^{(0)})$$

در روش میرای[1] نیوتن مقدار $\mathbf{x}$ دربارست بعدی یعنی $\mathbf{x}^{(m+1)}=\mathbf{x}^{(m)}+\mathbf{h}$ قابل قبول واقع نمی‌شود، اگر این امر به‌افزایش خطای مانده منجرشود، یعنی اگر $\|\mathbf{f}(\mathbf{x}^{(m+1)})\|_2 > \|\mathbf{f}(\mathbf{x}^{(m)})\|_2$. در چنین موردی بردارهای $\mathbf{x}^{(m)}+\mathbf{h}/2^i$ به‌ازای $i=1, 2, \ldots$ بررسی می‌شوند و $\mathbf{x}^{(m+1)}$ به‌عنوان اولین برداری که به‌ازای آن خطای مانده کمتر از $\|\mathbf{f}(\mathbf{x}^{(m)})\|_2$ است به‌کار گرفته می‌شود.

الگوریتم ۴۰۵: روش میرای نیوتن برای یك دستگاه دستگاه n معادلهٔ n مجهولی $\mathbf{f}(\xi)=0$ که در آن $\mathbf{f}$ تابعی‌است بردار-مقداری و دارای مؤلفه‌های هموار تابعی، به‌همراه یك حدس اولیهٔ $\mathbf{x}^{(0)}$ برای یك جواب ξ از دستگاه فوق داده شده‌است.

For $m = 0, 1, 2, \ldots$ until satisfied, do:
  $\mathbf{h} := -\mathbf{f}'(\mathbf{x}^{(m)})^{-1}\mathbf{f}(\mathbf{x}^{(m)})$
  $* i := \min\{j : 0 \le j, \|\mathbf{f}(\mathbf{x}^{(m)} + \mathbf{h}/2^j)\|_2 < \|\mathbf{f}(\mathbf{x}^{(m)})\|_2\}$
  $\mathbf{x}^{(m+1)} := \mathbf{x}^{(m)} + \mathbf{h}/2^i$

بدون مطالعه روشن نیست که آیا مرحلهٔ * همیشه می‌تواند اجرا شود یا نه. برای اینکه i تعریف شود، لازم وکافی‌است که جهت نیوتن $\mathbf{h}$ در $\mathbf{x}=\mathbf{x}^{(m)}$ برای تابع

$$F(\mathbf{x})=\|\mathbf{f}(\mathbf{x})\|_2^2$$

یك‌جهت کاهشی باشد. چون داریم

$$\nabla F(\mathbf{x})=2\mathbf{f}'(\mathbf{x})^T\mathbf{f}(\mathbf{x})$$

و بنابر (۵.۵) و (۶.۵)، $\mathbf{h}$ یك‌جهت‌کاهشی برای F در $\mathbf{x}$ است اگر و فقط اگر رابطهٔ

$$(\mathbf{f}'(\mathbf{x})^T\mathbf{f}(\mathbf{x}))^T\mathbf{h}<0$$

برقرار باشد. از سوی دیگر داریم $\mathbf{h}=-\mathbf{f}'(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$ و بنابراین

$$\begin{aligned}(\nabla F(\mathbf{x}))^T\mathbf{h} &= 2(\mathbf{f}'(\mathbf{x})^T\mathbf{f}(\mathbf{x}))^T(-\mathbf{f}'(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x}))\\ &= -2\mathbf{f}(\mathbf{x})^T\mathbf{f}(\mathbf{x}) = -2\|\mathbf{f}(\mathbf{x})\|_2^2 < 0\end{aligned}$$

این رابطه نشان می‌دهدکه جهت نیوتن برای $F(\mathbf{x})=\|\mathbf{f}(\mathbf{x})\|_2^2$ حقیقتاً یك جهت‌کاهشی است، بنابراین عدد صحیح i درمرحلهٔ * به‌خوبی تعریف شده‌است.

گرچه درعمل به‌جای مرحلهٔ * مرحلهٔ زیر را می‌گذارند

$$i := \min\{j : 0 \le j \le j_{\max},\ \|\mathbf{f}(\mathbf{x}^{(m)}+\mathbf{h}/2^j)\|_2 < \|\mathbf{f}(\mathbf{x}^{(m)})\|_2\}$$

---

1. damped

اگر i تعریف نشده باشد، **آنگاه خروج** از الگوریتم به‌جهت وجود اشکال. **وگرنه** یا $j_{\max}$ که از قبل تعیین و مثلا برابر ۱۰ انتخاب شده شروع کنید.

□ **مثال ۵.۵:** دستگاه $\mathbf{f}(\xi)=\mathbf{0}$ با

$$f_1(\mathbf{x})=x_1+3\ln|x_1|-x_2^2 \qquad f_2(\mathbf{x})=2x_1^2-x_1x_2-5x_1+1$$

دارای چندین جواب است. به‌این دلیل، به‌منظور تأمین همگرایی به‌سمت یك جواب خاص یا به‌طور کلی برای تأمین همگرایی، تخمین اولیه باید با دقت انتخاب شود.

معادله‌های نیوتن به‌صورت زیر هستند

$$\begin{bmatrix} 1+3/x_1 & -2x_2 \\ 4x_1-x_2-5 & -x_1 \end{bmatrix}\mathbf{h}=-\mathbf{f}(\mathbf{x})$$

اگر با حدس اولیهٔ $\mathbf{x}^{(0)}=[2\ 2]^T$ شروع کنیم، بارستهای مندرج درجدول زیر به دست می‌آیند.

m	$\mathbf{x}^{(m)}$		$\|\mathbf{f}(\mathbf{x}^{(m)})\|_2$	$\|\mathbf{h}\|_2$
0	2.	2.	0.500 + 1	0.238 + 2
1	−18.1588	−10.5794	0.572 + 3	0.112 + 2
2	−8.3710	−5.2287	0.142 + 3	0.543 + 1
3	−3.5525	−2.7191	0.351 + 2	0.266 + 1
4	−1.2015	−1.4728	0.860 + 1	0.198 + 1
5	−0.0004	0.0945	0.234 + 2	0.187 + 4
6	0.0451	−1866.2415	0.348 + 7	0.933 + 3
7	0.0233	−933.1179	0.871 + 6	0.467 + 3
8	0.0108	−466.5520	0.218 + 6	0.233 + 3
		etc.		

روشن‌است که بارستها اساساً همگرا نیستند. اما اکنون با به‌کار گرفتن روش میرای نیوتن و شروع با همان حدس اولیه، جدول زیر به‌دست می‌آید.

m	$\mathbf{x}^{(m)}$		$\|\mathbf{f}(\mathbf{x}^{(m)})\|_2$	$\|\mathbf{h}\|_2$	i
0	2.	2.	0.500 + 1	0.238 + 2	4
1	0.7400698	1.2137849	0.299 + 1	0.160 + 1	1
2	0.5310238	0.4415855	0.205 + 1	0.217 + 1	2
3	0.5178341	−0.1001096	0.178 + 1	0.372 + 1	3
4	0.5584838	−0.5637875	0.173 + 1	0.832 + 1	6
5	0.5847026	−0.6910621	0.172 + 1	0.967 + 1	6
6	0.6215780	−0.8376443	0.171 + 1	0.937 + 1	6
7	0.6657612	−0.9772562	0.171 + 1	0.684 + 1	5
8	0.7448782	−1.1760004	0.169 + 1	0.328 + 1	3
9	0.9489394	−1.5313175	0.163 + 1	0.676	0
10	1.5501608	−1.8410875	0.105 + 1	0.315	0
11	1.3892191	−1.5703845	0.132	0.473 − 1	0
12	1.3735386	−1.5257440	0.249 − 2	0.781 − 3	0
13	1.3734783	−1.5249650	0.608 − 6	0.156 − 6	0
14	1.3734783	−1.5249648	0.843 − 7		

در اینجا برای هر بارست، عدد صحیح i که در مرحلهٔ * از الگوریتم ۴.۵ معین‌شده نیز آورده شده‌است. در ابتدا مقدار $\mathbf{h}$ که در نظر گرفته شده نسبتاً بزرگ است، اما این مقدار تا اندازهٔ $۱/۶۴ = ۱/۲^۶$ میرا می‌شود. همچنین مقدار $\|\mathbf{f}(\mathbf{x}^{(m)})\|_۲$ از خطای مانده نیز به‌زحمت از یك مرحله تا مرحلهٔ دیگر کاهش می‌یابد. اما درنهایت تمام مراحل روش نیوتن اجرا شده و بارستها، همان‌طوری که باید، به‌صورت درجهٔ دوم همگرا می‌شوند. (حقیقتاً مشاهدهٔ این بارستها روی پایانهٔ[1] کامپیوتری تجربهٔ مهیجی است. سرانجام هنگامی که همگرایی درجهٔ دوم برقرار می‌شود، احساس شادی به‌شخص دست می‌دهد.)

محاسبات فوق با دقت ساده[2] روی کامپیوتر UNIVAC ۱۱۱۰ انجام گرفته است. بنابراین خطای $\|\mathbf{f}(\mathbf{x}^{(۱۴)})\|_۲$ درسطح خطاهای عادی است. □

اشکال دوم در استفاده از روش نیوتن مربوط به‌تشکیل معادلهٔ نیوتن و حل آن به‌ازای مقدار تصحیحی $\mathbf{h}$ است. اگر $\mathbf{f}$ تابع نسبتاً پیچیده‌ای باشد، ساختن ماتریس ژاکوبی خود کار دشواری‌است، زیرا در این‌صورت امکان اشتباهات زیادی از جمله در مشتقگیری و یا رمزگذاری[3] عناصر $\mathbf{f}'$ وجود دارد. معمولاً نتیجهٔ این‌نوع اشتباهات سبب ازدست رفتن همگرایی‌درجهٔ دوم، و یا درموارد حاد موجب ازدست رفتن‌همگرایی به‌طور کلی‌می‌باشد. اکنون برخی مراکزمحاسباتی، برنامه‌هایی برای مشتقگیری نمادی از عبارتها و یاحتی از توابعی که به‌وسیلهٔ یك زیر برنامه تولید شده‌اند، فراهم کرده‌اند. استفاده‌ازاین‌گونه‌برنامه‌ها کمك زیادی در ساختن ماتریس ژاکوبی به‌ما می‌کند. اگر چنین برنامه‌هایی در دسترس نباشند، می‌توانیم ماتریس ژاکوبی رمزشدهٔ $\mathbf{f}'(\mathbf{x})$ خود را از مقایسهٔ آن دریك نقطهٔ $\mathbf{x}$ با تقریبهای عددی نه‌چندان دقیق آرایه‌های آن به‌شکل

$$\frac{\partial f_i}{\partial x_j}(\mathbf{x}) \approx [f_i(\mathbf{x}+\varepsilon\mathbf{i}_j)-f_i(\mathbf{x})]/\varepsilon \qquad (۸.۵)$$

یا

$$\frac{\partial f_i}{\partial x_j}(\mathbf{x}) \approx [f_i(\mathbf{x}+\varepsilon\mathbf{i}_j)-f_i(\mathbf{x}-\varepsilon\mathbf{i}_j)]/(۲\varepsilon) \qquad (۹.۵)$$

که باهردو رابطهٔ فوق، درمبحث حسابان (به‌فصل۷ نگاه‌کنید) آشنا شده‌ایم آزمایش کنیم.

یك راه دیگر این‌است‌که فقط، به‌رمزگذاری توابع $f_۱$، ...، f_n و سپس به‌کار گرفتن فرمولهای (۸.۵) یا (۹.۵) برای ساختن یك تقریب مناسب J برای $f'(\mathbf{x})$ اکتفا کنیم. این امر مستلزم انتخاب مناسبی برای اندازهٔ نمو ε[4] است (به‌بخش۱.۷ نگاه کنید). گیریم J_m ماتریس ژاکوبی $\mathbf{f}'(\mathbf{x}^{(m)})$ یا یك تقریب مناسبی برای آن باشد. وقتی

---

1. terminal 2. single precision 3. coding
4. step size

که J_m ساخته شد، می‌باید معادلهٔ

$$J_m\mathbf{h} = -\mathbf{f}(\mathbf{x}^{(m)})$$

را نسبت به‌مقدار تصحیحی $\mathbf{h}$ حل کرد. درحالت کلی J_m یك ماتریس کامل ازمرتبهٔ nاست، لذا تعداد $\mathcal{O}(n^3)$ عمل برای به‌دست آوردن $\mathbf{h}$ لازم است. از سوی دیگر، اگر همگرایی وجود داشته باشد و $\mathbf{f}'(\mathbf{x})$ به‌طور پیوسته به $\mathbf{x}$ بستگی داشته باشد، $\mathbf{f}'(\mathbf{x}^{(m+k)})$ و $\mathbf{f}(\mathbf{x}^{(m)})$ تفاوت بسیار کمی بـا هم خواهنـد داشت. بنابراین برای صرفه‌جویی در عملیات، منطقی خواهد بودکه به‌جای $\mathbf{f}'(\mathbf{x}^{(m+k)})$ از $\mathbf{f}'(\mathbf{x}^{(m)})$ استفاده کنیم. زیرا بـا یکبار تجزیه‌کردن $\mathbf{f}'(\mathbf{x}^{(m)})$ می‌توان دستگاه فوق را به‌ازای مقادیر متعدد سمت راست با هزینه‌ای معادل بـا $\mathcal{O}(n^2)$ عملیـات، حل کـرد. ایـن روش **اصلاح‌شدهٔ نیوتن** است، کـه در آن تسـاوی $J_{m+k}=\mathbf{f}'(\mathbf{x}^{(m)})$ به‌ازای $k=0, 1, 2, \ldots$ برقراراست، تا وقتی که (یامگرآنکه) کندی سرعت همگرایی اعلام کند کـه J_{m+k} به‌عنوان آخرین ماتریس ژاکوبی بـه‌کار گرفته خواهد شد.

یك انحراف خیلی زیادتر از روش نیوتن، در روشهای به‌اصطلاح **باب روزکردن ماتریسی**[1] مطرح شده‌است، دراین روش J_{m+1} از J_m با افزودن یك ماتریس‌ازمرتبهٔ[2] یك یا دوکه بستگی‌به J_m، $\mathbf{x}^{(m)}$، $\mathbf{h}$، $\mathbf{f}(\mathbf{x}^{(m)})$ و $\mathbf{f}(\mathbf{x}^{(m+1)})$ دارد به‌دست می‌آید. هدف اصلی این‌است‌که J_{m+1} طوری انتخاب شودکه با توجه به‌روابط

$$\delta\mathbf{x}=\mathbf{x}^{(m+1)}-\mathbf{x}^{(m)} \quad \text{و} \quad \delta\mathbf{f}=\mathbf{f}(\mathbf{x}^{(m+1)})-\mathbf{f}(\mathbf{x}^{(m)})$$

رابطهٔ

$$J_{m+1}(\delta\mathbf{x})=\delta\mathbf{f}$$

به‌دست آید این،مطلب دور از ذهنی نیست، زیرا درمواردی‌که به‌ازای $\mathbf{x}$ نزدیك به $\mathbf{x}^{(m)}$ تساوی $J_{m+1}=\mathbf{f}'(\mathbf{x})$ برقرار باشد، یك تساوی تقریبی موجوداست.

اگرماتریسی‌که به J_m افزوده می‌شود ازمرتبهٔ یك یا دوباشد، آنگاه این‌امکان‌وجود داردکه تغییر حاصله در $K_m=J_m^{-1}$ به‌صورت افزودن ماتریسی‌که به‌آسانی قابل محاسبه است بیان شود. بنابراین با حفظ رد K_m به‌جای رد J_m، می‌توان از تجزیـهٔ J_m اجتناب کرد. یکی از روشهای معمول از این نوع، **روش برویدن**[3] است. در این روش، ابتـدا $K_0=\mathbf{f}'(\mathbf{x}^{(0)})^{-1}$ محاسبه و سپس K_{m+1} از روی K_m، به‌توسط رابطهٔ

$$K_{m+1}=K_m-\frac{[K_m(\delta\mathbf{f})-\delta\mathbf{x}](\delta\mathbf{x})^TK_m}{(\delta\mathbf{x})^TK_m(\delta\mathbf{f})} \qquad (10.5)$$

و رابطهٔ

---

1. matrix-updating 2. rank 3. Broyden's method

$$\mathbf{x}^{(m+1)}=\mathbf{x}^{(m)}-K_m\mathbf{f}(\mathbf{x}^{(m)}) \qquad (۱۱.۵)$$

محاسبه می‌شود. $J_m=K_m^{-1}$ متناظر به‌آن در رابطهٔ

$$J_{m+1}(\delta\mathbf{x})=\delta\mathbf{f}$$

صدق می‌کند درحالی‌که رابطهٔ $J_{m+1}\mathbf{z}=J_m\mathbf{z}$ به‌ازای‌جمیع مقادیر $\mathbf{z}$ عمود بر $\delta\mathbf{x}$ برقرار است. درعمل، دراین روش بارستی، از روش میراکردن نیز استفاده می‌شود.

تمرین

۱-۲۰۵ با استفاده از روش نیوتن جوابهای دستگاه $\nabla F(\mathbf{x})=0$ راکه درآن F تابع داده شده در تمرین ۱۰۵-۱ است، پیداکنید. نتیجهٔ کار خود را با آنچه‌که در تمرین ۱۰۵-۲ خواسته شده‌است مقایسه کنید.

۲-۲۰۵ ثابت کنید: که اگر $G'(c)=G[a,b]$ و $a<c<b$، و $G''(x)$ و $G'''(x)$ هردو در بازهٔ $[a,b]$ مثبت باشند، آنگاه $c>(a+b)/2$. [راهنمایی: گیریم $\bar{c}=(a+b)/2$ و با بسط کلیهٔ متغیرها به صورت سری تیلر درحول $\bar{c}$ نشان دهیدکه $G'(\bar{c})<G[a,b]$. درغیراین‌صورت مستقیماً از (۸.۷) استفاده کنید].

نتیجه‌گیری کنیدکه ماتریس ژاکوبی $f'(\xi)$ درمثال ۴.۵ به‌طورقطری اکیداً نافذ و بنابراین وارونپذیراست.

۳-۲۰۵ بااستفاده ازروش‌نیوتن‌جواب دستگاه نسبتاً پیچیدهٔ زیرادرفاصلهٔ $0\leqslant x,\ y\leqslant 1$ پیداکنید.

$$\cos\left[\frac{x^2-\sqrt{\sin(xy)+3}}{4+(xy)^2}\right]+\sin(3xy-1)=0.934$$

$$\exp\{\cos[(xy)^3-3]\}+\tan\left[\frac{x}{y}(0.08+\cos x)\right]=1.79$$

(البته شناسه‌های توابع مثلثاتی برحسب رادیان اندازه‌گیری می‌شوند).

اگرهمگرایی‌درجهٔ دوم به‌دست نیاوردید، با استفاده از (۸.۵) یا(۹.۵) رمزگذاری ماتریس ژاکوبی را آزمون کنید.

۴-۲۰۵ با به‌کارگیری روش نیوتنی میرا، جواب مسئله‌ای راکه در مثال ۵.۵ مورد بحث قرارگرفت با نقطهٔ شروع $\mathbf{x}^{(0)}=[1\ \ 2]^T$ پیداکنید.

۵-۲۰۵ سعی کنید مسئلهٔ مثال ۵.۵ را با روش تداوم و نقطهٔ شروع $\mathbf{x}^{(0)}=[1\ \ 2]^T$ و استفاده از $t_0,\ldots,t_N=0,\ 0.1,\ 0.3,\ 0.6,\ 1$ حل کنید. (درمراحل اول، بارستها را

آنقدر ادامه دهید که همگرایی درجهٔ دوم پدیدار شود.

۶-۲۰۵ مسئلهٔ مثال ۴۰۵ را به‌ازای $n=10$ و $G(x)=x^5$ حل کنید.

۵*-۳ بارست نقطهٔ ثابت و روشهای واهلش

روش نیوتن و بعضی از صورتهای متفاوت آن، که در قسمت ۲۰۵ موردبحث قرار گرفتند، همگی مثالهایی از **بارست نقطهٔ ثابت** هستند. دراینجا می‌توان معادلهٔ

$$\mathbf{f}(\boldsymbol{\xi})=\mathbf{0}$$

را به‌صورتی هم‌ارز با آن:

$$\boldsymbol{\xi}=\mathbf{g}(\boldsymbol{\xi})$$

نوشت و سپس با شروع از حدس اولیهٔ $\mathbf{x}^{(0)}$ دنبالهٔ

$$\mathbf{x}^{(m+1)}=\mathbf{g}(\mathbf{x}^{(m)}) \qquad m=0,1,2,\dots$$

را تولید کرد با این امید که دنبالهٔ فوق به‌سمت نقطهٔ ثابت $\boldsymbol{\xi}$ از $\mathbf{g}$ همگرا شود.

برای مثال، روش نیوتن یك روش بارستی نقطهٔ ثابت از این گونه‌است، با **تابع بارستی** $\mathbf{g}$، که به‌صورت رابطهٔ زیرداده شده‌است

$$\mathbf{g}(\mathbf{x})=\mathbf{x}-\mathbf{f}'(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$$

به‌صورت کلیتر، در روشهای **شبه-نیوتنی**[1]، یك تابع بارستی به‌صورت زیر به‌کار برده می‌شود.

$$g(\mathbf{x})=\mathbf{x}-C\mathbf{f}(\mathbf{x}) \qquad (۱۲۰۵)$$

که در آن $C=C(\mathbf{x})$ یك ماتریس‌است. روش واهلش، که بعداً دراین قسمت موردبحث قرار خواهدگرفت، دید دیگری‌است برای تشکیل تابعهای بارستی برای حل $\mathbf{f}(\boldsymbol{\xi})=\mathbf{0}$ به‌وسیلهٔ بارست نقطهٔ ثابت.

تجزیه و تحلیل روش بارست نقطهٔ ثابت برای دستگاه معادلات با آنچه که برای یك معادله درفصل ۳ ارائه شد، یك تفاوت جزئی دارد و آن اینکه دراین‌حال دربارست mام، اندازهٔ خطای $\mathbf{x}^{(m)}-\boldsymbol{\xi}$ به‌جای مقادیر قدرمطلق، توسط نرم معین می‌شود.

قضیهٔ ۱۰۵ فرض کنید که تابع بارست $\mathbf{g}$، مجموعهٔ بسته‌ای مانند S را برخود بنگارد. یعنی اگر $\mathbf{x}$ جزء مجموعهٔ S باشد، $\mathbf{g}(\mathbf{x})$ نیز به S متعلق‌است و بعلاوه فرض کنید که $\mathbf{g}$ روی S

---

1. quasi-Newton

انقباضی[1] باشد، یعنی رابطهٔ

$$\|\mathbf{g}(\mathbf{x})-\mathbf{g}(\mathbf{y})\| \leqslant K\|\mathbf{x}-\mathbf{y}\|$$

به‌ازای جمیع مقادیر $\mathbf{x}$ و $\mathbf{y}$ در S و مقداری مانند $K < ۱$ برقرار باشد، دراین صورت

(i) $\mathbf{g}$ یك نقطهٔ ثابت در S دارد.

(ii) اگر $\mathbf{g}$ یك نقطهٔ ثابت $\mathbf{g}$ در S باشد، آنگاه روش بارست نقطهٔ ثابت که با $\mathbf{x}^{(۰)}$ در S شروع می‌شود به‌سمت $\boldsymbol{\xi}$ همگرا خواهد شد، یعنی برای دنباله‌ای نظیر $\mathbf{x}^{(m+۱)}=\mathbf{g}(\mathbf{x}^{(m)})$، به‌ازای $m=۰,۱,۲,\ldots$، داریم $\lim_{m\to\infty}\|\boldsymbol{\xi}-\mathbf{x}^{(m)}\|=۰$

به‌بیان صریحتر

$$\|\boldsymbol{\xi}-\mathbf{x}^{(m)}\| \leqslant \frac{K}{۱-K}\|\mathbf{x}^{(m)}-\mathbf{x}^{(m-۱)}\| \qquad (۱۳.۵)$$

بنابراین

$$\|\boldsymbol{\xi}-\mathbf{x}^{(m)}\| \leqslant \frac{K^m}{۱-K}\|\mathbf{x}^{(۱)}-\mathbf{x}^{(۰)}\| \qquad (۱۴.۵)$$

فرضیات فوق ناظر براین هستند، که می‌توانیم با هر $\mathbf{x}^{(۰)}$ در $\mathbf{S}$ شروع کنیم و به‌ازای $m=۰,۱,۲,\ldots$، بارست $\mathbf{x}^{(m+۱)}=\mathbf{g}(\mathbf{x}^{(m)})$ را تا بینهایت به‌ازای هر $\mathbf{x}^{(m)}$ در S ادامه دهیم. وانگهی با استدلالی که خارج از ظرفیت این کتاب است (مثلا با استفاده از تمامیت فضای n-بعدی)، قسمت (i) نتیجه می‌شود، وسرانجام برای به‌دست آوردن براورد (۱۳.۵)، و ازآنجا (۱۴.۵)، ملاحظه می‌کنیم که

$$\begin{aligned}\|\boldsymbol{\xi}-\mathbf{x}^{(m)}\| &= \|\mathbf{g}(\boldsymbol{\xi})-\mathbf{g}(\mathbf{x}^{(m-۱)})\| \\ &\leqslant K\|\boldsymbol{\xi}-\mathbf{x}^{(m-۱)}\| \qquad (۱۵.۵)\end{aligned}$$

و از این‌رو چون $\mathbf{g}$ انقباضی است، بنابر نامساوی مثلثی (۴-۳۳، iii) داریم

$$\begin{aligned}\|\boldsymbol{\xi}-\mathbf{x}^{(m-۱)}\| &\leqslant \|\boldsymbol{\xi}-\mathbf{x}^{(m)}\|+\|\mathbf{x}^{(m)}-\mathbf{x}^{(m-۱)}\| \\ &\leqslant K\|\boldsymbol{\xi}-\mathbf{x}^{(m-۱)}\|+\|\mathbf{x}^{(m)}-\mathbf{x}^{(m-۱)}\|\end{aligned}$$

یا

$$(۱-K)\|\boldsymbol{\xi}-\mathbf{x}^{(m-۱)}\| \leqslant \mathbf{x}^{(m)}-\mathbf{x}^{(m-۱)}\|$$

حال از ترکیب نامساوی فوق با (۱۵.۵)، (۱۳.۵) به‌دست می‌آید.

---

1. contractive

□ **مثال ۶.۵:** روش نیوتن یك روش بارست نقطهٔ ثابت با تابع بارست زیراست

$$\mathbf{g}(\mathbf{x})=\mathbf{x}-\mathbf{f}'(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$$

بنابراین

$$\mathbf{f}'(\mathbf{x})[\mathbf{g}(\mathbf{x})-\mathbf{x}]=-\mathbf{f}(\mathbf{x})$$

در حالی‌که بنابررابطهٔ (۷.۵) و بافرض آنکه $\mathbf{f}$ مشتقات جزئی و پیوستهٔ اول ودوم داشته باشد، داریم

$$\begin{aligned}\mathbf{0}&=\mathbf{f}(\xi)=\mathbf{f}(\mathbf{x}+\xi-\mathbf{x})\\&=\mathbf{f}(\mathbf{x})+\mathbf{f}'(\mathbf{x})(\xi-\mathbf{x})+\mathcal{O}(\|\xi-\mathbf{x}\|^2)\end{aligned}$$

بنابراین از قراردادن $-\mathbf{f}'(\mathbf{x})[\mathbf{g}(\mathbf{x})-\mathbf{x}]$ به‌جای $\mathbf{f}(\mathbf{x})$ داریم

$$\mathbf{0}=\mathbf{f}'(\mathbf{x})[-(\mathbf{g}(\mathbf{x})-\mathbf{x})+(\xi-\mathbf{x})]+\mathcal{O}(\|\xi-\mathbf{x}\|^2)$$

یا

$$\mathbf{f}'(\mathbf{x})[\mathbf{g}(\mathbf{x})-\xi]=\mathcal{O}(\|\xi-\mathbf{x}\|^2)$$

عبارت بالا بیانگر این‌است که، به‌ازای مقداری ثابت مانند c داریم

$$\|\mathbf{g}(\mathbf{x})-\xi\|\leqslant\|\mathbf{f}'(\mathbf{x})^{-1}\|\ .\ c\ .\ \|\xi-\mathbf{x}\|^2$$

اکنون اگـر $\mathbf{f}'(\xi)$ وارونپذیر باشد، آنـگاه چون بنابرفرض $\mathbf{f}'(\mathbf{x})$ پیوسته است، می‌توانیم مقدار مثبتی مانند δ ویك M طوری پیداکنیم که $\mathbf{f}'(\mathbf{x})^{-1}$ به‌ازای جمیع مقادیر $\mathbf{x}$ در محدودهٔ δ از ξ وجود داشته ویك نرم ماتریسی نابیشتر از M داشته باشد. امـا در این‌صورت، اگر ε کوچکتر از δ و $(Mc)^{-1}$ انتخاب شود، به‌ازای کلیهٔ مقادیر $\mathbf{x}$ درمجموعهٔ بستهٔ

$$S=\{\mathbf{x}:\|\xi-\mathbf{x}\|\leqslant\varepsilon\}$$

$\mathbf{f}'(\mathbf{x})^{-1}$ وجود دارد (بنابراین $\mathbf{g}(\mathbf{x})$ تعریف شده‌است) و

$$\|\mathbf{g}(\mathbf{x})-\xi\|\leqslant Mc\|\xi-\mathbf{x}\|^2\leqslant(Mc\varepsilon)\|\xi-\mathbf{x}\|\leqslant\|\xi-\mathbf{x}\|$$

لـذا $\mathbf{g}$ مجموعـهٔ بستهٔ S را برخودش می‌نـگارد. بعـلاوه اگر $\varepsilon<\|\xi-\mathbf{x}\|$، آنگاه، $K=Mc\|\xi-\mathbf{x}\|<1$، بنابراین ξ یك نقطهٔ ثابت جاذب[1] از $\mathbf{g}$ است، و بارستی کـه با هر $\mathbf{x}^{(0)}$ کمتر از ε درمحدودهٔ ξ شروع شود، به‌سمت ξ همگرا خواهدبود. □

1. attracting fixed point

حال، به‌عنوان یك مثال دیگر حل دستگاه خطی

$$A\xi=\mathbf{b} \qquad (۱۶.۵)$$

را به‌وسیلهٔ بازست نقطهٔ ثابت درنظر می‌گیریم. این‌گونه روشهای بازستی تماماً می‌توانند برپایهٔ فکر **وارون تقریبی**[1] استوار باشند. منظور ما این‌است که به‌ازای هرماتریس C رابطهٔ

$$\|I-CA\|<۱ \qquad (۱۷.۵)$$

به‌ازای مقداری از نرم ماتریسی برقرار باشد.

لم ۱۰۵ اگر C یك وارون تقریبی ماتریس A باشد، یعنی اگر به‌ازای یك نرم ماتریسی، نامساوی $\|I-CA\|<۱$ برقرار باشد، آنگاه C و A هردو وارونپذیرند.

در حقیقت، اگر C یا A وارونپذیر نباشند، آنگاه ماتریس CA نیز وارونپذیر نخواهد بود (به‌تمرین ۱۰۴–۸ نگاه کنید). دراین‌حال بنا برقضیهٔ ۴.۴، می‌توانیم $\mathbf{x}\neq ۰$ ی پیدا کنیم که داشته باشیم $CA\mathbf{x}=۰$. اما دراین صورت رابطهٔ

$$۰\neq\|\mathbf{x}\|=\|(I-CA)\mathbf{x}\|\leqslant\|I-CA\|\ \|\mathbf{x}\|<\|\mathbf{x}\|$$

به‌دست می‌آید که بی‌معنی است.

بخصوص، اگر A یك وارون تقریبی داشته باشد، (۱۶.۵) دقیقاً دارای یك جواب است.

در رابطه با یك وارون تقریبی C از A، تابع بازست زیر را درنظر می‌گیریم

$$\begin{aligned}\mathbf{g}(\mathbf{x})&=C\mathbf{b}+(I-CA)\mathbf{x}\\&=\mathbf{x}+C(\mathbf{b}-A\mathbf{x})\end{aligned}$$

توجه کنید که تابع بازست فوق ازنوع تابع شبه‌نیوتنی، یعنی به‌شکل $\mathbf{g}(\mathbf{x})=\mathbf{x}-C\mathbf{f}(\mathbf{x})$ است، اگر $\mathbf{f}(\mathbf{x})=A\mathbf{x}-\mathbf{b}$. همچنین

$$\begin{aligned}\mathbf{g}(\mathbf{x})-\mathbf{g}(\mathbf{y})&=C\mathbf{b}+(I-CA)\mathbf{x}-[C\mathbf{b}+(I-CA)\mathbf{y}]\\&=(I-CA)(\mathbf{x}-\mathbf{y})\end{aligned}$$

درنتیجه داریم

$$\|\mathbf{g}(\mathbf{x})-\mathbf{g}(\mathbf{y})\|\leqslant\|I-CA\|\ \|\mathbf{x}-\mathbf{y}\| \qquad (۱۸.۵)$$

که نشان می‌دهد، $\mathbf{g}$ انقباضی بوده و رابطهٔ زیربرقراراست

$$K=\|I-CA\|<۱$$

---

1. approximate inverse

بنابراین بارست نقطهٔ ثابت

$$\mathbf{x}^{(m+1)}=\mathbf{x}^{(m)}+C(\mathbf{b}-A\mathbf{x}^{(m)}) \qquad m=0,1,2,\cdots$$

باشروع ازهرنقطهٔ $\mathbf{x}^{(0)}$، بسمت جواب منحصر به‌فرد ξ از معادلهٔ (۱۶.۵) همگراخواهد بود، و خطا در هر مرحله از بارست، حداقل به‌اندازهٔ یك عامل $K=\|I-CA\|$ کاهش می‌یابد.

□ **مثال ۷.۵:** فرض کنید ماتریس A نافذ قطری سطراً مؤکد باشد یعنی داشته باشیم

$$|a_{ii}|>\sum_{j\neq i}|a_{ij}| \qquad i=1,\cdots,n$$

گیریم $D=\operatorname{diag}(a_{11},a_{22},\cdots,a_{nn})$ قطر A باشد، دراین‌صورت رابطهٔ

$$\|I-D^{-1}A\|_\infty=\max_i\{1-\sum_j|a_{ij}|/|a_{ii}|\}=\max_i\sum_{j\neq i}|a_{ij}|/|a_{ii}|<1$$

جدول ۱.۵

Jacobi				*Gauss-Seidel*		
$x_1^{(m)}$	$x_2^{(m)}$	$x_3^{(m)}$	m	$x_1^{(m)}$	$x_2^{(m)}$	$x_3^{(m)}$
0	0	0	0	0	0	0
1.2	1.2	1.2	1	1.2	1.08	0.972
0.96	0.96	0.96	2	0.9948	1.0033	1.00019
1.008	1.008	1.008	3	0.99965	1.000016	1.000033
0.9984	0.9984	0.9984	4			
1.00032	1.00032	1.00032	5			
0.999936	0.999936	0.999936	6			

نشان می‌دهدکه D یك وارون تقریبی A است. شکل بارستی متناظر با آن یعنی

$$\mathbf{x}^{(m+1)}=\mathbf{x}^{(m)}+D^{-1}(\mathbf{b}-A\mathbf{x}^{(m)}) \qquad m=0,1,2,\cdots \qquad (19.5)$$

همان **بارست ژاکوبی** است. توجه‌کنیدکه می‌توان $\mathbf{x}^{(m+1)}$ را از روی $\mathbf{x}^m$ با حل معادلهٔ iام نسبت به‌مجهول iام به‌ازای هر i و به‌دنبال آن مقادیر فعلی سایرمجهولهارا ازفرمول:

$$x_i^{(m+1)}=(b_i-\sum_{j\neq i}a_{ij}x_j^{(m)})/a_{ii} \qquad i=1,\cdots,n$$

به‌دست آورد. برای دستگاه خطی خاص

$$10x_1+x_2+x_3=12$$
$$x_1+10x_2+x_3=12$$
$$x_1+x_2+10x_3=12$$

روش بارست ژاکوبی که‌از نقطهٔ $\mathbf{x}^{(0)}=\mathbf{0}$ شروع می‌شود بردارهای $\mathbf{x}^{(1)}$، $\mathbf{x}^{(2)}$، ... ، $\mathbf{x}^{(6)}$ را که درجدول ۱۰۵ آورده شده‌اند تولید می‌کند. به‌نظر می‌رسد که دنبالهٔ فوق به‌شکل مطلوبی به‌سمت جواب دستگاه یعنی $[1\ 1\ 1]^T$ همگراست. برای این مثال داریم

$$\|I-D^{-1}A\|_\infty=\max_i\left\{\frac{1}{10}+\frac{1}{10}\right\}=0.2$$

به‌طوری که می‌توانیم انتظار داشته باشیم که خطا درهرمرحله حداقل به‌اندازهٔ عامل ۰٫۲ کاهش می‌یابد، این امر نیز از اعداد جدول ۱۰۵ آشکاراست. □

البته پیدا کردن وارون تقریبی C برای A اساساً آسان‌است. برای مثال رابطهٔ $C=A^{-1}$ این‌کار را خواهدکرد و بارست مربوطه به‌آن دریك مرحله همگرا خواهد شد. اما نکته‌ای که دربه‌کارگرفتن روش بارستی برای حل $A\boldsymbol{\xi}=\mathbf{b}$ حائزاهمیت است، این‌است که در وهلهٔ اول با روش بارستی ممکن‌است یك جواب تقریبی با دقت قابل‌قبولی بسیار سریعتر ازحل مستقیم $A\boldsymbol{\xi}=\mathbf{b}$ به‌دست آورد. از این‌رو، مسئلهٔ مهم، انتخاب C به‌طریقی که بتوانیم به‌ازای هر $\mathbf{r}$ بردار $C\mathbf{r}$ را نسبت به‌بردار $A^{-1}\mathbf{r}$ با محاسبات کمتری به‌دست آوریم. به‌عنوان نمونه، می‌توان C را وارون یك ماتریس قطری (مانند بارست ژاکوبی) و یا وارون یك ماتریس مثلثی (مانند بارست گاوس-زایدل که در زیربحث می‌شود) و یا وارون حاصلضرب دو ماتریس مثلثی (مانند اصلاح بارستی الگوریتم ۵.۴) و یا حتی وارون یك ماتریس سه قطری وغیره گرفت.

الگوریتم ۵.۵: بارست نقطهٔ ثابت برای دستگاههای خطی دستگاه خطی $A\boldsymbol{\xi}=\mathbf{b}$ از مرتبهٔ n داده شده‌است

یك ماتریس C از مرتبهٔ n طوری انتخاب کنید که

(i) به‌ازای یك بردار مفروض $\mathbf{r}$، بردار $C\mathbf{r}$ «به‌آسانی» محاسبه شود

(ii) به‌ازای حداقل یك نرم ماتریسی داشته باشیم، $\|I-CA\|<1$. یك n-بردار $\mathbf{x}^{(0)}$، مثلا $\mathbf{x}^{(0)}=\mathbf{0}$ را انتخاب کنید

For $m = 0, 1, 2, \ldots,$ until satisfied, do:
$\mathbf{x}^{(m+1)} := \mathbf{x}^{(m)} + C(\mathbf{b} - A\mathbf{x}^{(m)})$

در صورت نبودن خطای گردکردن، دنبالهٔ حاصله $\mathbf{x}^{(0)}$، $\mathbf{x}^{(1)}$، $\mathbf{x}^{(2)}$، ... به‌سمت جواب دستگاه خطی داده شده همگرا می‌شود.

مانند فصل۳، دراینجا عبارت «until satisfied» برای تأکید روی ناتمام‌بودن توضیح مفروض به‌کار رفته‌است. برای کامل‌کردن الگوریتم، می‌باید معیارهای دقیقی‌برای توقف بارستها تعریف شود. معیارهای نمونه به‌صورت زیرند:

الگوریتم را متوقف سازید اگر (الف): $\|\mathbf{x}^{(m)}-\mathbf{x}^{(m-1)}\|\leqslant\varepsilon$

یا اگر (ب): $\dfrac{\|\mathbf{x}^{(m)}-\mathbf{x}^{(m-1)}\|}{\|\mathbf{x}^{(m)}\|}<\varepsilon$

یا اگر (ج): به‌ازای یك M داده شده $m>M$

معیار آخر می‌باید همیشه درهر برنامه‌ای که الگوریتم فوق را اجرا می‌کند گنجانیده شود. تذکری راکه برای اولین‌بار در بخش ۶.۱ داده شد، تکرار می‌کنیم: این حقیقت که رابطهٔ

$$\|\mathbf{x}^{(m)}-\mathbf{x}^{(m-1)}\|<\varepsilon$$

برقراراست بدان معنی نیست که داشته باشیم

$$\|\mathbf{x}^{(m)}-\xi\|<\varepsilon$$

اما با توجه به (۱۳.۵) و (۱۸.۵) می‌دانیم که

$$\|\mathbf{x}^{(m)}-\xi\|\leqslant\frac{K}{1-K}\|\mathbf{x}^{(m)}-\mathbf{x}^{(m-1)}\| \qquad (۲۰.۵)$$

که در آن $K=\|I-CA\|$.

مثلا در مثال ۷.۵ برای بارست ژاکوبی به‌دست آوردیم که $\|I-CA\|_\infty\leqslant$ ۰ر۲ و $\|\mathbf{x}^{(۶)}-\mathbf{x}^{(۵)}\|_\infty=$ ۰ر۰۰۰۳۸۴. بنابراین رابطهٔ (۲۰.۵)، براورد زیر را به‌دست می‌دهد

$$\|\xi-\mathbf{x}^{(۶)}\|_\infty\leqslant\frac{\text{۰ر۲}}{\text{۱}-\text{۰ر۲}}\,\text{۰ر۰۰۰۳۸۴}=\text{۰ر۰۰۰۰۹۶}$$

و در واقع $\|\xi-\mathbf{x}^{(۶)}\|=$ ۰ر۰۰۰۰۶۴، به‌طوری که خطا فقط به‌اندازهٔ ۵۰ درصد زیادی براورد شده‌است. متأسفانه، براورد مطلوبی از $\|I-CA\|$ معمولا مشکل به‌دست می‌آید، یا آنکه این براورد آنقدر به ۱ نزدیك است که مخرج $K-$ ۱ در (۲۰.۵) را بیش از حد کوچك می‌سازد و بنابراین کران حاصله در $\|\xi-\mathbf{x}^{(m)}\|$ قابل استفاده نیست.

لازم‌است متذکرشویم که C ممکن‌است وارون تقریبی A باشد حتی اگر به‌ازای یك نرم ماتریسی داشته باشیم

$$\|I-CA\|>\text{۱}$$

تمام آنچه‌که برای یك وارون تقریبی C برای A (وبرای همگرایی بارست نقطهٔ ثابت

متناظر به‌آن) لازم‌است، این است که $I-CA$ یك نرم ماتریسی کمتر از ۱ داشته باشد. مثلا ماتریس $B=\begin{bmatrix} ۰ & ۰٫۹ \\ ۰ & ۰ \end{bmatrix}$ در رابطهٔ زیر صدق می‌کند

$$||B||_\infty = ۰٫۹ < ۱$$

با این‌همه، در برخی نرم‌های ماتریسی داریم $||B|| < ۱$، مثلا $||B||_۱ = ۰٫۹ < ۱$. بنابراین، مهم پیداکردن روشهایی است برای بیان اینکه به‌ازای یك نرم ماتریسی، نامساوی $||B|| < ۱$ برقرار باشد (بی‌آنکه کلیهٔ نرمهای ماتریسی ممکن را بررسی کنیم). قضیهٔ زیر (در اصل) چنین روشی را دراختیار ما قرار می‌دهد.

قضیهٔ ۲۰۵ گیریم $\rho(B)$ شعاع طیفی[1] ماتریس B باشد، یعنی

$$\rho(B) = \max\{|\lambda| : \lambda \text{ یك ویژه‌مقدار ماتریس } B \text{ است}\}$$

دراین حال به‌ازای جمیع مقادیر $\varepsilon > ۰$ یك نرم برداری وجود دارد به‌طوری که نرم ماتریسی مربوط به B در رابطهٔ $||B|| \leqslant \rho(B)+\varepsilon$ صدق می‌کند.

درنتیجه C یك وارون تقریبی برای ماتریس A است اگر و فقط اگر $\rho(I-CA) < ۱$.

بعلاوه هرچه شعاع طیفی ماتریس $I-CA$ کوچکتر باشد همگرایی در روش بارست نقطهٔ ثابت، در پایان سریعتر خواهد بود.

مطلب فوق را می‌توان این‌گونه نشان داد، که درحل دستگاه $A\xi = \mathbf{b}$ خطای

$$\mathbf{e}^{(m)} = \xi - \mathbf{x}^{(m)}$$

درمرحلهٔ mام بارست نقطهٔ ثابت

$$\mathbf{x}^{(m+۱)} = \mathbf{x}^{(m)} + C(\mathbf{b} - A\mathbf{x}^{(m)}), \qquad m = ۰, ۱, ۲, \cdots$$

به‌ازای کلیهٔ مقادیر m در رابطهٔ زیر صدق می‌کند

$$\mathbf{e}^{(m+۱)} = (I-CA)\mathbf{e}^{(m)}$$

بنابراین به‌ازای کلیهٔ مقادیر n و با داشتن $B = I-CA$، خواهیم داشت $\mathbf{e}^{(m)} = B^m\mathbf{e}^{(۰)}$. این امر نشان می‌دهد که دنبالهٔ خطاها یعنی $\mathbf{e}^{(۰)}$، $\mathbf{e}^{(۱)}$، $\mathbf{e}^{(۲)}$، ... به‌شکلی است که درفصل ۴ (به ۶۲۰۴ تا ۶۷۰۴ نگاه کنید) در رابطه با روش توانی موردبحث قرارگرفت. در آنجا بیان کردیم که دنبالهٔ نرمال شدهٔ

$$\mathbf{e}^{(m)}/||\mathbf{e}^{(m)}|| \qquad m = ۰, ۱, ۲, \cdots$$

---

1. spectral radius

معمولا به‌سمت یك ویژه‌بردار $B=I-CA$ متعلق به‌بزرگترین ویژه‌مقدار (از نظر قدرمطلق) B همگرا می‌شود، یعنی

$$\mathbf{e}^{(m+1)}=B\mathbf{e}^{(m)}\approx\lambda\mathbf{e}^{(m)}$$

که در آن $|\lambda|=\rho(B)$. بنابراین درنهایت به‌طورکلی مقدار خطا در هرمرحله از بارست به‌اندازهٔ عامل $\rho(B)$، و نه سریعتر، کاهش می‌یابد.

اکنون مثالهای خاصی از بارست نقطهٔ ثابت برای دستگاههای خطی را موردبحث قرار می‌دهیم. یکی از این‌مثالها روش اصلاح بارستی‌است که درفصل قبل‌موردبحث قرار گرفت. برای یادآوری، دراین روش ابتدا ماندهٔ $\mathbf{r}^{(m)}=\mathbf{b}-A\mathbf{x}^{(m)}$ برای mامین جواب تقریبی $\mathbf{x}^{(m)}$ محاسبه می‌شود، سپس با استفاده از تجزیهٔ مثلثی A کـه هنگام حذف محاسبه می‌شود، جواب (تقریبی) $\mathbf{y}^{(m)}$ برای دستگاه خطی $A\mathbf{y}=\mathbf{r}^{(m)}$ به‌دست می‌آید و با افزودن $\mathbf{y}^{(m)}$ به $\mathbf{x}^{(m)}$ جواب تقریبی‌مناسب (چنین‌امیداست) $\mathbf{x}^{(m+1)}=\mathbf{x}^{(m)}+\mathbf{y}^{(m)}$ حاصل‌می‌شود. درحالت‌کلی بردار $\mathbf{y}^{(m)}$ جواب (دقیق) $A\mathbf{y}=\mathbf{r}^{(m)}$ نیست. علت این امرتا حدی ناشی‌از خطای گردکردن هنگام پیشجایگذاری و پسجایگذاری است. می‌توان نشان دادکه سهم اصلی در ایجاد خطا در $\mathbf{y}^{(m)}$، معمولا ناشی از بی‌دقتی درتجزیهٔ A به‌صورت مثلثی PLU، یعنی‌ناشی ازاین امراست‌که PLU تنها تقریبی‌است از A. اگر ازخطای گردکردن به‌هنگام پیشجایگذاری و پسجایگذاری صرفنظرکنیم، خواهیم داشت

$$\mathbf{y}^{(m)}=(PLU)^{-1}\mathbf{r}^{(m)}=(PLU)^{-1}(\mathbf{b}-A\mathbf{x}^{(m)})$$

بنابراین

$$\mathbf{x}^{(m+1)}=\mathbf{x}^{(m)}+(PLU)^{-1}(\mathbf{b}-A\mathbf{x}^{(m)})$$

مطالب فوق نشان می‌دهند که روش اصلاح بارستی، حالت خاصی از روش بارست نقطهٔ ثابت‌است و C همان تجزیهٔ A به‌صورت مثلثی PLU است. اما برای رده‌هایی‌ازماتریسهای A، ماتریسی مانند C می‌توان پیداکردکه درحالتهای (i) و (ii) الگوریتـم ۵.۵ صادق باشد و حجم محاسباتی آن نسبت به‌محاسباتی کـه برای تجزیهٔ مثلثی A لازم است، بسیار کمتر باشد. برای یك دستگاه خطی با چنین ماتریس ضرایبی، مفیدتر است‌که روش حذفی راکنار بگذاریم و جواب دستگاه را مستقیماً با الگوریتم ۵.۵ محاسبه‌کنیم.

برای بحث درمورد دو روش‌معمول برای انتخاب C، ماتریس ضرایب $A=(a_{ij})$ را به صورت حاصلجمع یك ماتریس اکیداً پایین‌مثلثی $\hat{L}=(\hat{l}_{ij})$ و یـك ماتریس قطری $\hat{D}=(\hat{d}_{ij})$ و یك ماتریس اکیداً بالامثلثی $\hat{U}=(\hat{u}_{ij})$ می‌نویسیم،

$$A=\hat{L}+\hat{D}+\hat{U}$$

که در آن

$$\hat{l}_{ij}=\begin{cases}a_{ij} & i>j\\ 0 & i\leqslant j\end{cases}\qquad \hat{d}_{ij}=\begin{cases}a_{ij} & i=j\\ 0 & i\neq j\end{cases}\qquad \hat{u}_{ij}=\begin{cases}0 & i\geqslant j\\ a_{ij} & i<j\end{cases}$$

بعلاوه، فرض می‌کنیم کـه تمام درایه‌های قطری A غیرصفر، یعنی D وارونپذیر باشد. اگر این شرط برقرار نباشد، ابتدا معادلات را طوری مرتب می‌کنیم که شرط فوق برقرار شود، که این کار هنگامی که A وارونپذیر باشد همیشه شدنی است. (به‌تمرین ۷.۴ـ۵ نگاه کنید).

در **بارست ژاکوبی یا روش تغییر مکانهای همزمان**[1]، همان گونه که در مثال ۷.۵ـ۵ بحث شد، می‌توان $C=D^{-1}$ انتخاب کرد.

اگر بارست ژاکوبی همگرا باشد، قسمت قطری A یعنی $\hat{D}$، تقریب نسبتـاً خوبی برای A است که در نتیجه داریم

$$\|B\|=\|I-\hat{D}^{-1}A\|<1$$

اما در این حالت انتظار می‌رود کـه قسمت پایین‌مثلثی A یعنی $\hat{L}+\hat{D}$ حتی تقریب بهتری برای A باشد، یعنی انتظار می‌رود که داشته باشیم

$$\|I-(\hat{L}+\hat{D})^{-1}A\|\leqslant\|I-\hat{D}^{-1}A\|<1$$

لذا به‌نظر می‌رسد که روش بارست نقطهٔ ثابت با $C^{-1}=\hat{L}+\hat{D}$، نسبت به‌روش ژاکوبی، بارستی با همگرایی سریعتر است. گرچه در حالت کلی مطلب فوق صحیح نیست، اما برای رده‌های مختلفی از ماتریس A صادق است؛ مثلا هرگاه A نافذ قطری سطراً مؤکد باشد یا هرگاه A سه قطری (یا به‌طور کلیتر هرگاه A یك بلوك‌سه‌قطری با بلوکهای قطری قطری باشد) یا هرگاه A دارای عناصر قطری مثبت و عناصر غیر قطری منفی باشد.

بارست نقطهٔ ثابت با $C^{-1}=\hat{L}+\hat{D}$ به **بارست گاوس‌ـزایدل**[2] یا **روش تغییر مکانهای پیاپی**[3] موسوم است. در این روش داریم

$$\mathbf{x}^{(m+1)}=\mathbf{x}^{(m)}+(\hat{L}+\hat{D})^{-1}(\mathbf{b}-A\mathbf{x}^{(m)})$$

یا

$$(\hat{L}+\hat{D})\mathbf{x}^{(m+1)}=(\hat{L}+\hat{D}-A)\mathbf{x}^{(m)}+\mathbf{b}$$

یا

$$\hat{D}\mathbf{x}^{(m+1)}=-\hat{L}\mathbf{x}^{(m+1)}-\hat{U}\mathbf{x}^{(m)}+\mathbf{b}$$

که نتیجهٔ آن فرمول زیر است

$$x_i^{(m+1)}=\frac{-\sum_{j<i}a_{ij}x_j^{(m+1)}-\sum_{j>i}a_{ij}x_j^{(m)}+b_i}{a_{ii}}\qquad i=1,\dots,n$$

---

1. method of simultaneous displacements
2. Gauss-Seidel iteration
3. method of successive displacements

ظاهراً با دانستن $x_1^{(m+1)}, \ldots, x_{i-1}^{(m+1)}$ می‌توانیم درایهٔ نام iام $\mathbf{X}^{(m+1)}$ را محاسبه کنیم.

الگوریتم ۶.۵: بارست‌گاوس‌ـزایدل، دستگاه خطی $A\mathbf{X}=\mathbf{b}$ از مرتبهٔ n، که همهٔ درایه‌های قطری ماتریس ضرایب آن $A=(a_{ij})$ غیرصفرند داده شده‌است.

درایه‌های $B=(b_{ij})$ و $\mathbf{c}=(c_i)$ را به‌وسیلهٔ روابط زیرمحاسبه کنید

$$b_{ij}=\begin{cases} -a_{ij}/a_{ii} & i\neq j \\ 0 & i=j \end{cases}$$

$$c_i=\frac{b_i}{a_{ii}} \qquad \text{به‌ازای جمیع مقادیر } i \text{ و } j$$

یک $\mathbf{X}^{(0)}$، مثلاً $\mathbf{X}^{(0)}=0$ را انتخاب کنید

For $m = 1, 2, \ldots,$ until satisfied, do:
  For $i = 1, \ldots, n$, do:
    $x_i^{(m)} := \Sigma_{j=1}^{i-1} b_{ij}x_j^{(m)} + \Sigma_{j=i+1}^{n} b_{ij}x_j^{(m-1)} + c_i$

اگر یک نرم ماتریسی از $(\hat{L}+\hat{D})^{-1}\hat{U}$ کمتر از یک باشد، آنگاه دنبالهٔ $\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots$ که به‌طریق فوق تولید می‌شود به‌سمت جواب دستگاه داده شده همگراست.

بردارهای $\mathbf{X}^{(1)}$ و $\mathbf{X}^{(2)}$ و $\mathbf{X}^{(3)}$ که از به‌کاربردن روش بارستی گاوس‌ـزایدل برای دستگاه خطی مثال ۷.۵ نتیجه شده‌اند، درجدول ۱.۵ آورده شده‌اند. توجه کنید که در این مثال، بارست گاوس‌ـزایدل خیلی‌سریعتر از بارست ژاکوبی همگرا می‌شود. دقت عمل بعد از سه مرحله بهتر از دقت عملی‌است که بعد از شش مرحله در بارست ژاکوبی به‌دست آمده است.

در بارست ژاکوبی از درایه‌های $\mathbf{X}^{(m)}$ تنها برای محاسبهٔ بارست بعدی $\mathbf{X}^{(m+1)}$ استفاده می‌شود؛ در حالی که در روش بارستی گاوس‌ـزایدل درایهٔ $\mathbf{X}^{(m)}$ برای محاسبهٔ درایه‌های بعدی $\mathbf{X}^{(m)}$ به‌کاربرده می‌شود، وبدین دلیل است که نام تغییرمکانهای همزمان و تغییر مکانهای پیاپی بر آن گذارده شده است. بخصوص بارست ژاکوبی مستلزم به خاطرسپردن مقادیر دو بارست است، درحالی که بارست گاوس‌ـزایدل فقط به‌یک بردار نیازدارد.

می‌توان نشان داد که اگر ماتریس‌ضرایب A نافذ قطری (سطراً) مؤکد باشد، بارست گاوس‌ـزایدل همگراست. این روش همچنین همگراست اگر A معین[1] و مثبت باشد، یعنی اگر A حقیقی و متقارن باشد و به‌ازای جمیع بردارهای $\mathbf{y}$ غیرصفر نامساوی زیر برقرار باشد

$$\mathbf{y}^T A\mathbf{y} > 0$$

---

1. definite

بالاخره از بین تکنیکهای متعدد موجود برای تسریع تقارب بارست نقطهٔ ثابت، ما **به‌بیش واهلش پیاپی**[1] (با علامت اختصاری بوپ) اشاره می‌کنیم که در آن از بارست پیشنهادی گاوس-زایدل دایر برتغییر از $x^{(m)}$ به $x^{(m+1)}$ فراتر می‌روند. مثلا به‌جای گرفتن رابطهٔ

$$x_i^{(m+1)}=x_i^{(m)}-(\sum_{j<i}a_{ij}x_j^{(m+1)}+\sum_{j>i}a_{ij}x_j^{(m)}-b_i)/a_{ii}$$ به‌ازای جمیع مقادیر i

به‌روال گاوس-زایدل دایر برتغییر از $x^{(m)}$ به $x^{(m+1)}$ گامی فراتر می‌نهند و از رابطهٔ

$$x_i^{(m+1)}=x_i^{(m)}-\omega(\sum_{j<i}a_{ij}x_j^{(m+1)}+\sum_{j>i}a_{ij}x_j^{(m)}-b_i)/a_{ii}$$ به‌ازای جمیع مقادیر i

که در آن ω (بزرگتر از ۱) **پارامتر بیش واهلش** نامیده می‌شود، استفاده می‌کنند. بارست حاصل را می‌توان به‌روشنی به‌صورت رابطهٔ زیر، که ولی چندان روشن نیست، نوشت

$$\mathbf{x}^{(m+1)}=\mathbf{x}^{(m)}+C_\omega(\mathbf{b}-A\mathbf{x}^{(m)})$$

ماتریس بارست متناظر، به‌صورت $I-C_\omega A=(\hat{D}+\omega\hat{L})^{-1}(\hat{D}-\omega\hat{U})$ است. از لحاظ نظری، پارامتر بیش‌واهلش ω می‌باید طوری انتخاب شود که $\rho(I-C_\omega A)$ تا حد امکان کوچك باشد. البته در وهلهٔ اول، این کار درمقایسه باحل دستگاه معادلات $A\xi=\mathbf{b}$، مشکلتر است. اما ممکن است بخواهیم چنین دستگاه خطی را برای سمت راستهای متفاوت و فقط تاحدی دقیق حل کنیم، که دراین‌صورت صرفه دراین‌است که از راه آزمایش مقدار «مطلوبی» برای ω تعیین نماییم. همچنین برای ماتریسهای خاص A که درحل عددی معادلات استانداردِ دیفرانسیل با مشتقات جزئی با آنها برخورد می‌کنیم، می‌توان $\rho(I-C_\omega A)$ را برحسب شعاع طیفی ماتریس بارست $\hat{D}^{-1}(\hat{L}+\hat{U})$ — متعلق به‌بارست ژاکوبی نوشت و بنابراین می‌توان اظهارنظری کیفی دربارهٔ بهترین انتخاب بهینهٔ ω نمود. یك نمونهٔ انتخاب برای ω مقداری بین ۱٫۲ و ۱٫۶ است.

چنانچه قبلا اشاره شد، روشهای بارستی معمولا برای دستگاههای خطی بزرگ که دارای ماتریس ضرایب تنك می‌باشند به‌کار گرفته می‌شوند. درماتریسهای تنك تعدادعناصر غیرصفر زیاد نیست و بنابراین درهرمرحله تعداد عملیات حسابی که باید انجام گیرد کم است. بعلاوه روشهای بارستی نسبت به‌رشد[2] خطای گردکردن حساسیت کمتری دارند و فقط اندازهٔ خطای گردکردن حاصل درهرمرحله مهم‌است. از طرف دیگر روشهای بارستی همیشه همگرا نخواهند بود وحتی وقتی هم که همگرا می‌شوند ممکن‌است همگرایی آنقدر کند باشدکه رسیدن به‌جواب مستلزم بارستهای‌زیاد وغیرعملی باشد. دردستگاههای‌بزرگ، برای رسیدن به‌جوابی با چهار یا پنج رقم اعشار، تعداد کل بارستهای لازم ممکن است به‌چند صدتا برسد.

---

1. successive overrelaxation 2. growth

زیربنایی فکری که روش بارست ژاکوبی و گاوس-زایدل بر آن قرار دارد فکر **واهلش** است که این فکر در زمینهٔ دستگاه کلی زیر که یك دستگاه n معادلهٔ n مجهولی غیرخطی است نیز به‌خوبی مصداق دارد

$$\mathbf{f}(\xi) = 0$$

ساده‌ترین صورت دستگاه فوق این است که فرض شود معادلات طوری مرتب شده‌اند که بتوان با حل معادلهٔ i ام یعنی

$$f_i(\xi_1, \cdots, \xi_n) = 0$$

نسبت به‌مجهول i ام، معادلهٔ هم‌ارز زیر را به‌دست آورد

$$\xi_i = g_i(\xi_1, \cdots, \xi_{i-1}, \xi_{i+1}, \cdots, \xi_n)$$

در این صورت برای یك تقریب مفروض $\mathbf{x}$ برای ξ، سعی می‌شود که با تغییر مؤلفهٔ i ام آن به‌صورت

$$g_i(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$$

این مؤلفه بهبود یابد.

واژهٔ «واهلش» برای فرایند فوق منسوب به‌سوت‌ول[1] است. درحقیقت تخمین جاری $\mathbf{x}$ همان جواب دقیق برای دستگاه وابستهٔ[2]

$$x_i = g_i(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots x_n) + r_i \qquad i = 1, \cdots, n$$

است که در آن جمله‌های r_i، که مربوط به‌خطا هستند، بدین جهت به‌حساب آمده‌اند که دستگاه مجبور شود جواب $\mathbf{x}$ را بپذیرد. در روش واهلش، درقبال برداشتن جملهٔ r_i ازمعادلهٔ i ام، برای مؤلفهٔ i ام مقدار جدیدی به‌دست می‌آید و بنابراین مؤلفهٔ فوق اصلاح می‌شود.

واهلش معمولا به‌اسلوب گاوس-زایدل انجام می‌گیرد، یعنی مقدار جدید مؤلفهٔ i ام بلافاصله برای اصلاح مؤلفهٔ دیگر به‌کار برده می‌شود. بعداً به‌اسلوبی منظم، بررسی تمام معادله‌ها تا به‌آخر انجام می‌گیرد و کلیهٔ مؤلفه‌های $\mathbf{x}$ تغییر می‌یابد. اتمام یك بررسی کامل یك **زدایش**[3] نامیده می‌شود.

شکلهای متفاوت و مفیدی برای فکر واهلش وجود دارد. مثلا ممکن است در بعضی موارد سهلتر باشد که به‌جای معادلهٔ i ام شکل هم‌ارز آن

$$\xi_i = g_i(\xi)$$

گذاشته شود، که در این فرمول هم‌ارز نیز، قسمت سمت راست آشکارا به ξ_i بستگی دارد.

---

1. Southwell 2. related system 3. sweep

به‌عنوان مثال دیگر، ممکن‌است با تغییر چندین مؤلفه از حدس جاری، معادلهٔ iام را فوراً واجد شرط کرد. به‌عبارت دیگر، ممکن است تخمین جدید $\mathbf{x}+\alpha\mathbf{y}^{(i)}$ را طوری تعیین کرد که تساوی

$$f_i(\mathbf{x}+\alpha\mathbf{y}^{(i)})=0$$

که در آن $\mathbf{y}^{(i)}$ برداری‌است‌ثابت وبستگی به i دارد، برقرارشود. البته برای واهلش‌معمولی داریم $\mathbf{y}^{(i)}=\mathbf{i}_i$.

□ **مثال۸.۵:** سعی می‌کنیم دستگاه معادلات خطی مثال ۴.۵ یعنی

$$3\xi_i^2=\xi_{i-1}^2+\xi_{i-1}\xi_{i+1}+\xi_{i+1}^2 \qquad i=1,\dots,n$$

را با توجه به $\xi_0=0<\xi_1<\dots<\xi_{n+1}=1$ باروش بارست گاوس-زایدل حل کنیم. بنابراین‌مانند مثال ۴.۵ با شروع از تخمین اولیهٔ $\mathbf{x}=[1\ 2\dots n]^T/(n+1)$ و $n=3$ بارست زیر را اجرا می‌کنیم

For $m = 0, 1, 2, \dots,$ until satisfied, do:
  For $i = 1, \dots, n$, do:
    $x_i := \sqrt{(x_{i-1}^2 + x_{i-1}x_{i+1} + x_{i+1}^2)/3}$

در جدول زیر فهرستی از چند بارست اول‌که بعد از هر زدایش به‌دست آمده ارائه شده است. همگرایی خطی‌است (بنابراین بـا همگرایی روش نیوتن قابل مقایسه نیست)، امـا همگرایی آنقدر منظم‌است‌که شتاب همگرایی را می‌توان پیداکرد. استفاده از روش‌بیش‌واهلش متوالـی به‌ازای $\omega=1.2$ درست در ۱۰ زدایش، ۲۱ بارست فـوق را به‌دست می‌دهد.

m	$\mathbf{x}^{(m)}$			$\|\mathbf{x}^{(m)}-\mathbf{x}^{(m-1)}\|_1$
0	0.2500000	0.5000000	0.7500000	
1	0.2886751	0.5361404	0.7796553	0.104
2	0.3095408	0.5612525	0.7908343	0.572 − 1
3	0.3240393	0.5734927	0.7963221	0.322 − 1
4	0.3311062	0.5794904	0.7990201	0.158 − 1
5	0.3345689	0.5824365	0.8003476	0.774 − 2
	.	. . .		.
10	0.3378204	0.5852071	0.8015973	0.223 − 3
11	0.3378695	0.5852490	0.8016162	0.110 − 3
	.	. . .		.
20	0.3379170	0.5852896	0.8016345	0.190 − 6
21	0.3379171	0.5852896	0.8016345	0.931 − 7

□

تمرین

۱-۳۰۵ دستگاه معادلات

$$x-\sinh y=0$$
$$2y-\cosh x=0$$

را با استفاده از روش بارست نقطهٔ ثابت حل کنید. یك جواب نزدیك به $[0.6\quad 0.6]^T$ وجود دارد.

۲-۳۰۵ با آزمایش، انتخاب مطلوبی برای پارامتر بیش واهلش تعیین کنید، و آن را برای واهلش پیاپی درمثال ۸.۵ به‌کاربرید.

عملیات فوق را برای $n=10$ و سپس برای مسئلهٔ ۲.۵-۶ مربوط به‌آن نیز انجام دهید.

۳-۳۰۵ سعی کنید دستگاه

$$x^2+xy^3=9 \qquad 3x^2y-y^3=4$$

را با روش بارست نقطهٔ ثابت حل کنید.

۴-۳۰۵ نشان دهیدکه روش بارست نقطهٔ ثابت با ماتریس بارست

$$B=\begin{bmatrix} 0.9 & 1000 \\ 0 & 0.9 \end{bmatrix}$$

همگرا خواهد بود، حتی اگر $1000<\|B\|_\infty=\|B\|_1$.

۵-۳۰۵ با استفاده از قضیهٔ شور[1] ثابت کنیدکه برای هر ماتریس مربعی B و هر $\varepsilon>0$، یك نرم برداری وجود داردکه نرم ماتریسی متناظر آن در رابطهٔ $\|B\|\leqslant\rho(B)+\varepsilon$ صدق می‌کند. (راهنمایی: نرم برداری را به‌شکل $\|\mathbf{x}\|:=\|DU\mathbf{x}\|_\infty$ تشکیل دهیدکه در آن U به‌وسیلهٔ قضیهٔ شور چنان معین می‌شودکه $A=U^{-1}BU$ یك ماتریس بالامثلثی باشد و $D=\operatorname{diag}[1,\delta,\delta^2,\ldots,\delta^{n-1}]$ چنان انتخاب می‌شودکه کلیهٔ درایه‌های غیرقطری $D^{-1}AD$ از لحاظ قدرمطلق مقداری کمتر از ε/n داشته باشند.)

۶-۳۰۵ نشان دهیدکه روشهای بارستی ژاکوبی و گاوس-زایدل که برای حل دستگاه معادلات خطی $A\xi=\mathbf{b}$، که در آن A ماتریس بالامثلثی وارونپذیری است، به‌کارمی‌روند در مراحل متناهی زیادی همگرا می‌شوند.

---

1. Schur

۳۰۵-۷ دستگاه معادلات

$$\begin{bmatrix} ۴ & -۱ & ۰ & ۰ \\ -۱ & ۴ & -۱ & ۰ \\ ۰ & -۱ & ۴ & -۱ \\ ۰ & ۰ & -۱ & ۴ \end{bmatrix} \begin{bmatrix} x_۱ \\ x_۲ \\ x_۳ \\ x_۴ \end{bmatrix} = \begin{bmatrix} ۱ \\ ۱ \\ ۱ \\ ۱ \end{bmatrix}$$

را با استفاده از روش بارست ژاکوبی و بارست گاوس-زایدل حل کنید. همچنین با استفاده از الگوریتم ۳.۴؛ تجزیه به‌عوامل ماتریس ضرایب دستگاه فوق را به‌دست آورید. آنگاه با به‌کارگیری روش اصلاح بارستی و همان حدس اولیه، دستگاه فوق را حل کنید. حجم محاسبات (عملیات ممیز شناور) لازم برای به‌دست آوردن جواب تقریبی با دقت مطلقی کمتر از $۱۰^{-۶}$ را درمورد سه روش فوق، براورد کنید.

۳۰۵-۸ ثابت کنید که روش بارست ژاکوبی همگراست، اگر ماتریس ضرایب Aی دستگاه، نافذ قطری ستوناً مؤکد باشد یعنی

$$|a_{jj}| > \sum_{i \neq j} |a_{ij}| \qquad j = ۱, \ldots, n$$

(راهنمایی: از نرم ماتریسی متناظر با نرم برداری، $\|\mathbf{x}\| = \sum_{i=۱}^{n} |a_{ii} x_i|$ استفاده کنید.)

٦

تقریب

دراین فصل مسئلهٔ تقریب زدن یك تابـع کلی را به‌کمك رده‌ای از توابع ساده‌تر بررسی می‌کنیم. برای این‌توابع تقریب دوکاربرد وجود دارد. نخست گذاردن توابع ساده‌تر به‌جای توابع پیچیده تا عملیات معمولی زیادی نظیر مشتقگیری و انتگرالگیری یـا حتی محاسبهٔ توابع بتواند بسیارراحت‌ترانجام گیرد. کاربرد مهم دیگر آن بازیابی یك تابـع از روی اطلاعات جزئی در بـاب آن، مثـلا از یك جدول تقریبی (احتمـالا فقط تقریبی) است. متداولترین ردهٔ توابع تقریب کننده، بسجمله‌ایهای جبری، بسجمله‌ایهای مثلثاتـی و اخیراً هم توابع بسجمله‌ای-تکه‌ای‌بوده‌اند. با بهترین تقریب و تقریب مطلوب به‌وسیلهٔ هریك از این رده‌ها را بررسی می‌کنیم.

۱.۶ تقریب یکنواخت به‌وسیلهٔ بسجمله‌ایها

دراین بخش مسئلهٔ ساختن یك بسجمله‌ای $p(x)$ از درجهٔ نابیشتر از n راکه تابع‌داده شدهٔ $f(x)$ را به‌طوریکنواخت دربازهٔ $a \leqslant x \leqslant b$ تقریب می‌زند مورد بررسی قرارمی‌دهیم. منظور این است که خطا در تقریب $P(x)$ برای تابع $f(x)$ را، به‌وسیلهٔ عدد یـا نرم زیر اندازه می‌گیریم

$$\|f-p\|_\infty = \max_{a \leqslant x \leqslant b} |f(x)-p(x)| \qquad (۱.۶)$$

آرمانی‌صحبت کنیم، می‌خواهیم بهترین تقریب یکنواخت از Π_n را به‌دست‌آوریم،

یعنی می‌خواهیم یك بسجمله‌ای $p_n^*(x)$ از درجهٔ نابیشتر از n به‌دست آوریم که برای آن داشته باشیم

$$\|f-p_n^*\|_\infty=\min_{p\in\Pi_n}\|f-p\|_\infty \qquad (۲.۶)$$

در اینجا نماد $p\in\pi_n$ را به‌عنوان تلخیصی از عبارت «p یك بسجمله‌ای از درجهٔ نابیشتر از n است» به‌کاربرده‌ایم. به‌عبارت دیگر، p_n^* بسجمله‌ای خاصی است از درجهٔ نابیشتر از n که همانقدر به‌تابع f نزدیك است که یك بسجمله‌ای از درجهٔ نابیشتر از n ممکن است به‌آن نزدیك باشد. عدد $\|f-p_n^*\|_\infty$ را با

$$\mathrm{dist}_\infty(f,\pi_n)$$

نشان می‌دهیم و آن را **فاصلهٔ یکنواخت** بسجمله‌ای f از درجهٔ نابیشتر از n در بازهٔ $a\leqslant x\leqslant b$ می‌نامیم.

پیش از آنکه بحثی دربارهٔ ساختن یك بسجمله‌ای مطلوب با بهترین بسجمله‌ای تقریب‌زننده انجام دهیم، اندکی به‌بررسی راههای برآورد $\mathrm{dist}_\infty(f,\pi_n)$ می‌پردازیم. مثلا اگر چنین براوردی نشان دهد که $\mathrm{dist}_\infty(f,\pi_n)<۱۰$، و ما در پی تقریبی باشیم که تا دو رقم آخر بعد ازممیز مطلوب باشد، آنگاه بیهوده وقت و نیرو صرف ساختن p_n^* نخواهیم کرد. برای چنین منظوری به‌ویژه بسیار مهم است که **کرانه‌های پایینی** برای $\mathrm{dist}_\infty(f,\pi_n)$ به‌دست آوریم و یکی از راههای به‌دست آوردن این کرانه‌ها به‌صورت زیر است.

با توجه به‌فصل ۲ خاطرنشان می‌کنیم که

$$g[x_0,\cdots,x_{n+1}]=\sum_{i=0}^{n+1}g(x_i)/w'(x_i)$$

که در آن داریم

$$w(x)=(x-x_0)\cdots(x-x_{n+1})$$

(به‌تمرین ۲.۲–۱ نگاه کنید) و گفتیم که اگر $g(x)$ یك بسجمله‌ای از درجهٔ نابیشتر از n باشد، این $(n+1)$امین عامل تفاضل برابر با صفر خواهد بود (به‌تمرین ۲.۲–۵ نگاه کنید). بنابراین برای هربسجمله‌ای خاص $p\in\pi_n$، داریم

$$\begin{aligned}f[x_0,\cdots,x_{n+1}]&=f[x_0,\cdots,x_{n+1}]-p[x_0,\cdots,x_{n+1}]\\&=(f-p)[x_0,\cdots,x_{n+1}]\\&=\sum_{i=0}^{n+1}(f(x_i)-p(x_i))/w'(x_i)\end{aligned}$$

در نتیجه اگر $x_0, \dots, x_{n+1}$ همگی در بازهٔ $a \leqslant x \leqslant b$ باشند، آنگاه

$$|f[x_0, \dots, x_{n+1}]| \leqslant \|f-p\|_\infty \cdot W(x_0, \dots, x_{n+1})$$

که در آن عدد مثبت $W(x_0, \dots, x_{n+1})$ از رابطهٔ

$$W(x_0, \dots, x_{n+1}) = \sum_{i=0}^{n+1} 1/|w'(x_i)| \qquad (۳.۶)$$

معین می‌شود. اکنون p را چنان انتخاب می‌کنیم که خصوصیت p_n^* را دارا باشد. بنابراین $\|f-p\|_\infty = \text{dist}_\infty(f, \pi_n)$، و کران پایین زیر را بدست می‌آوریم

$$|f[x_0, \dots, x_{n+1}]| / W(x_0, \dots, x_{n+1}) \leqslant \text{dist}_\infty(f, \pi_n) \qquad (۴.۶)$$

□ **مثال ۱.۶:** به‌ازای $n=1$ و $x_0 = -1$، $x_1 = 0$، $x_2 = 1$ داریم

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

$$= f(-1)/2 - f(0) + f(1)/2$$

از این‌رو $W(-1, 0, 1) = 2$، و بنابراین برای $a \leqslant -1$ و $b \leqslant 1$، خواهیم داشت

$$|f[-1, 0, 1]| / 2 \leqslant \text{dist}_\infty(f, \pi_1)$$

مثلا، برای $f(x) = e^x$، داریم $f[-1, 0, 1] = e^{-1}/2 - e^0 + e^1/2 = 0.54308$ و درنتیجه، $\text{dist}_\infty(e^x, \pi_1) \geqslant 0.27154$. □

استفاده‌ازکران پایین(۴.۶) مستلزم‌محاسبهٔ اعداد $w'(x_i) = \Pi_{j \neq i}(x_i - x_j)$ برای تشکیل $W(x_0, \dots, x_{n+1})$ است بـه منظور ملاحظهٔ یك روش کارا برای انجام این امر ۱۰.۶–۱۴ نگاه‌کنید). برای انتخاب بعضی مقادیر x_i، این اعداد شکل سادهٔ خاصی پیـدا می‌کنند. مثلا اگر

$$x_i = \cos \frac{i}{n+1} \pi \qquad i = 0, \dots, n+1 \qquad (۵.۶)$$

آنگاه

$$\frac{1}{w'(x_i)} = \frac{2^{n-1}}{n+1} (-1)^i \begin{cases} 1 & \text{اگر } i \text{ مساوی صفر یا } n+1 \text{ باشد} \\ 2 & \text{ودر غیر این‌صورت} \end{cases} \qquad (۶.۶)$$

از این‌رو، $W(x_0,\dots,x_{n+1})=2^n$ (به‌تمرین ۱.۰۶–۵ نگاه کنید) و بنابراین اگر بازهٔ $a\leqslant x\leqslant b$ شامل هردومقدار ۱ و ۱– باشد، خواهیم داشت

$$\frac{1}{2(n+1)}\left|f(1)-2f\left(\cos\frac{\pi}{n+1}\right)\right.$$

$$\left.+2f\left(\cos\frac{2\pi}{n+1}\right)-\dots-(-1)^n f(-1)\right|\leqslant \mathrm{dist}_\infty(f,\pi_n) \qquad (7.6)$$

برای استفاده از این کران پایین در بازه‌های دیگر، اول باید با یک تغییر خطی متغیرها بازهٔ موردنظر را به بازهٔ $1\leqslant x\leqslant 1-$ منتقل ساخت.

□ **مثال ۲.۰۶:** در بازهٔ استاندارد[1] $1\leqslant x\leqslant 1-$، تابع تقریب π_2 را برای تابع

$$f(x)=\tan \pi/4x$$

در نظر بگیرید. این یک تابع فرد است، یعنی $f(-x)=f(x)$، بنابراین به‌ازای مقادیر فرد n، کران پایین برابر با صفر می‌شود که این امر هیچ کمکی به‌ما نمی‌کند. درعوض تابع تقریب π_4 را در نظر می‌گیریم. دراین صورت با توجه به‌رابطهٔ (۷.۰۶) داریم

$$\frac{1}{10}\left|f(1)-2f\left(\cos\frac{\pi}{5}\right)+2f\left(\cos\frac{2\pi}{5}\right)-2f\left(\cos\frac{3\pi}{5}\right)\right.$$

$$\left.+2f\left(\cos\frac{4\pi}{5}\right)-f(-1)\right|\leqslant \mathrm{dist}_\infty(f,\pi_4)$$

و یا $0.00203\leqslant \mathrm{dist}_\infty(f,\pi_4)$. درحقیقت می‌توان نشان داد که

$$\mathrm{dist}_\infty(f,\pi_4)=0.00419\dots$$

و بنابراین کران پایین کاملا خوب است. □

در رابطه با این کرانه‌های پایین، قضیهٔ زیر وجود دارد که به «دولا واله‌پوسن»[2] نسبت‌داده می‌شود که مارا ازمحاسبهٔ $w'(x_i)$ برحذر می‌دارد، امامستلزم‌ساختن تقریب‌کننده برای $p\in\pi_n$ است.

قضیهٔ ۱.۰۶ فرض کنید که خطای $f(x)-p(x)$ درتقریب بسجمله‌ای $p\in\pi_n$ برای f در نقاط $x_0<x_1<\dots<x_{n+1}$ تغییرعلامت دهد، یعنی

$$(-1)^i[f(x_i)-p(x_i)]\varepsilon\geqslant 0 \qquad i=0,\dots,n+1 \text{ به‌ازای}$$

---

1. standard 2. de la Vallée-Poussin

که در آن $\varepsilon = \text{signum}\,[f(x_0) - p(x_0)]$. اگر به‌ازای جمیع مقادیر i داشته باشیم $a \leqslant x_i \leqslant b$، آنگاه

$$\text{dist}_\infty(f, \pi_n) \geqslant \min_i |f(x_i) - p(x_i)|$$

در واقع، همان‌گونه‌که درقضیه فرض شده‌است، اگر نقاط x_i مرتب شده باشند، آنگاه خواهیم داشت

$$(-1)^{n+1-i} w'(x_i) > 0 \qquad i = 0, \cdots, n+1 \text{ به‌ازای}$$

و بنابراین تمام جمعوند[1]ها درحاصلجمع

$$(f-p)[x_0, \cdots x_{n+1}] = \sum_{i=0}^{n+1} [f(x_i) - p(x_i)]/w'(x_i)$$

دارای یک علامت هستند. اما این بدان معناست که

$$|f[x_0, \cdots, x_{n+1}]| = \sum_{i=0}^{n+1} |(f(x_i) - p(x_i))/w'(x_i)|$$
$$\geqslant \min_i |f(x_i) - p(x_i)| \cdot W(x_0, \cdots, x_{n+1})$$

و این رابطه همراه با رابطهٔ (۴.۶)، قضیه را ثابت می‌کند.

اکنون فرض کنید ترتیبی اتخاذ کنیم که درقضیهٔ ۱.۶، به‌ازای $i = 0, 1, \cdots, n+1$ تساوی $|f(x_i) - p(x_i)| = \|f-p\|_\infty$ نیز برقرار باشد. دراین‌صورت خواهیم داشت

$$\|f-p\|_\infty \geqslant \text{dist}_\infty(f, \pi_n) \geqslant \min_i |f(x_i) - p(x_i)| = \|f-p\|_\infty$$

و چون اولین و آخرین عبارت دراین رشته نامساویها با هم یکی هستند، بنابراین کلا باید تساوی برقرار باشد. بخصوص بسجمله‌ای p می‌باید بهترین تقریب یکنواخت برای f از نوع π_n باشد. این نکته اثبات نیمه‌آسان قضیهٔ زیر منسوب به‌چبیشف‌است.

قضیهٔ ۲.۶ تابع f که در بازهٔ $a \leqslant x \leqslant b$ پیوسته‌است دقیقاً بهترین تقریب یکنواخت از نوع π_n را در $a \leqslant x \leqslant b$ دارد. بسجمله‌ای $p \in \pi_n$ بهترین تقریب یکنواخت برای f در بازهٔ $a \leqslant x \leqslant b$ است اگر، و فقط اگر، $n+2$ نقطهٔ $a \leqslant x_0 < \cdots < x_{n+1} \leqslant b$ چنان وجود داشته باشند که

$$(-1)^i [f(x_i) - p(x_i)] = \varepsilon \|f-p\|_\infty \qquad i = 0, \cdots, n+1 \qquad (۸.۶)$$

که در آن $\varepsilon = \text{signum}\,[f(x_0) - p(x_0)]$. دراینجا درحالتی که $f^{(n+1)}(x)$ در بازهٔ

---

1. summands

$a \leqslant x \leqslant b$ تغییر علامت ندهد، خواهیم داشت $a = x_0$ و $b = x_{n+1}$.

یك برهان این قضیهٔ اساسی را می‌توان درهر کتاب درسی در بـاب نظریهٔ تقریب، مثلا در کتاب رایس[1] [۱۷] یا ریولین[2] [۳۵] پیدا کرد.

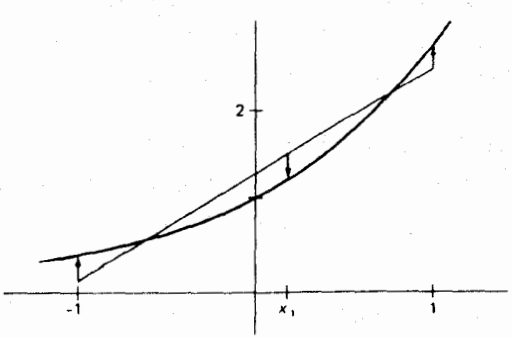
□ **مثال ۳.۶:** باردیگر به‌دست آوردن تقریب برای $f(x) = e^x$ را در بـازهٔ استانداردِ $-1 \leqslant x \leqslant 1$ بررسی می‌کنیم. درمثال ۱.۶ مشاهده گردید که $\text{dist}_\infty(f, \pi_1) \geqslant$ ۰٫۲۷. اکنون $p(x)$ را به‌صورت $p(x) = a + bx$، با $b = (e^1 - e^{-1})/2$ و $a = (e - bx_1)/2$ که در آن $b = p'(x_1) = f'(x_1) = e^{x_1}$ یـا $x_1 = \ln b$، انتخاب می‌کنیم؛ به‌شکل ۱.۶ نگاه کنید. در این صورت می‌توان تحقیق کرد که $f(x) - p(x)$ در شرط تناوب (۸.۶) با $n = 1$ و $x_0 = -1$، $x_2 = 1$ صدق می‌کند، یعنی

$$f(-1) - p(-1) = -[f(x_1) - p(x_1)] = f(1) - p(1)$$
$$= (e^1 + e^{-1})/2 - a = ۰٫۲۷۸۸۰\ldots$$

لذا این خط مستقیم خاص می‌باید بهترین تقریب یکنواخت برای e^x در بازهٔ $-1 \leqslant x \leqslant 1$ از π_1 باشد و $\text{dist}(e^x, \pi_1) =$ ۰٫۲۷۸۸۰... این رابطه نشان می‌دهد که کران پایین به‌دست آمده درمثال ۱.۶ کاملا دقیق است. □

یك مثال به‌ویژه مهم با بهترین تقریب یکنواخت در بازهٔ $-1 \leqslant x \leqslant 1$ از π_n برای تابع $f(x) = x^{n+1}$ فراهم شده است. خطا در این تقریب، به‌طوری که مشاهده خواهد شد، مضربی است از $T_{n+1}(x)$، که بسجمله‌ای چبیشف از درجهٔ $n+1$ است.

بنا بر تعریف، **بسجمله‌ای چبیشف از درجهٔ** k (در بازهٔ $-1 \leqslant x \leqslant 1$) با قاعدهٔ زیر معین می‌شود



شکل ۱.۶ بهترین تقریب یکنواخت خط راست برای e^x در $-1 \leqslant x \leqslant 1$

1. Rice 2. Rivlin

$$T_k(\cos\theta)=\cos k\theta \qquad (۹.۶)$$

بنا بر این

$$T_۰(x)=۱ \qquad T_۱(x)=x \qquad (۱۰.۶)$$

و بموجب فرمول جمع برای توابع مثلثاتی داریم

$$T_{k+۱}(x)=۲xT_k(x)-T_{k-۱}(x) \qquad k=۱,۲,\ldots \qquad (۱۱.۶)$$

از رابطهٔ فوق خواهیم داشت

$$T_۲(x)=۲xT_۱(x)-T_۰(x)=۲x^۲-۱$$

$$T_۳(x)=۲xT_۲(x)-T_۱(x)=۲x(۲x^۲-۱)-x=۴x^۳-۳x$$

وهشت بسجمله‌ای اول از این بسجمله‌ایها در جدول ۱.۶ و نمودار پنج بسجمله‌ای اول آنها درشکلهای ۲.۶ و ۳.۶ داده شده‌اند.

رابطهٔ بازگشتی (۱۱.۶) روشن می‌سازد که $T_k(x)$ که با (۹.۶) تعریف شده درواقع یك بسجمله‌ای است دقیقاً از درجهٔ k با ضریب بزرگترین درجهٔ $۲^{k-۱}$. بعلاوه از تعریف (۹.۶) آشکار است که

$$|T_k(x)|\leqslant ۱ \qquad -۱\leqslant x\leqslant ۱ \qquad (۱۲.۶)$$

بهازای جمیع مقادیر

و $T_k(x)$ متناوباً در $k+۱$ نقطهٔ

جدول ۱.۶

$$T_۰(x)=۱$$

$$T_۱(x)=x$$

$$T_۲(x)=۲x^۲-۱$$

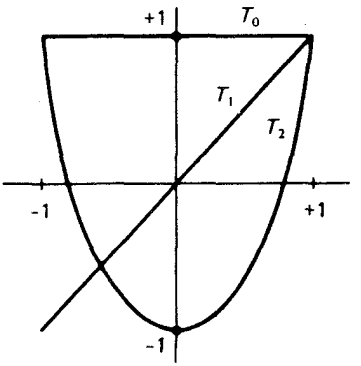
$$T_۳(x)=۴x^۳-۳x$$

$$T_۴(x)=۸x^۴-۸x^۲+۱$$

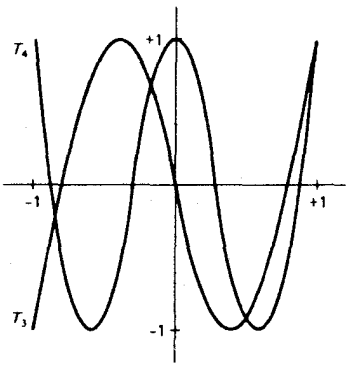
$$T_۵(x)=۱۶x^۵-۲۰x^۳+۵x$$

$$T_۶(x)=۳۲x^۶-۴۸x^۴+۱۸x^۲-۱$$

$$T_۷(x)=۶۴x^۷-۱۱۲x^۵+۵۶x^۳-۷x$$



شکل ۲.۶



شکل ۳.۶

$$x_j=\cos\frac{k-j}{k}\pi \qquad j=0,\ldots,k$$

به‌این‌کران ± 1 می‌رسد. یعنی از قاعدهٔ (۹.۶) به‌دست می‌آوریم

$$T_k(x_j)=T_k\left(\cos\frac{k-j}{k}\pi\right)=\cos k\frac{k-j}{k}\pi=(-1)^{k-j} \qquad j=0,\ldots,k$$

اما این رابطه نشان می‌دهد که مخصوصاً برای بسجمله‌ای $p_n(x)$ از درجهٔ نابیشتر از n، تساوی

$$2^{-n}T_{n+1}(x)=x^{n+1}-p_n(x)$$

برقرار است و این بسجمله‌ای، بنابرقضیهٔ ۲.۶ بهترین تقریب یکنواخت بـرای x^{n+1} در بازهٔ $-1\leqslant x\leqslant 1$ است و نیز

$$\operatorname{dist}_\infty(x^{n+1},\pi_n)=1/2^n \qquad (13.6)$$

درحالت‌کلی به‌دست‌آوردن بهترین تقریب یکنواخت از π_n، کار ساده‌ای نیست. با فرض مشتقپذیر بودن تابع $f(x)$، می‌توان براساس قضیهٔ ۲.۶ و با قید

$$a\leqslant x_0<\cdots<x_{n+1}\leqslant b$$

دستگاه معادلات غیرخطی

$$f(x_i)-p_n^*(x_i)=(-1)^i d \qquad i=0,\ldots,n+1$$

$$\phi(x_i)[f'(x_i)-p_n^{*\prime}(x_i)]=0 \qquad i=0,\ldots,n+1 \qquad (14.6)$$

را برای نقـاط x_0، ...، x_{n+1}، وبرای $n+1$ ضریب از $p_n^*(x)$ و عدد (مثبت یا منفی) $d=\pm\|f-p_n^*\|_\infty$ حل نمود. دراینجا

$$\phi(x)=\begin{cases}0 & \text{اگر } x=b \text{ یا } x=a\\ 1 & \text{در غیراین‌صورت}\end{cases}$$

تابع $\phi(x)$ عاملی‌است برای تمیز بین یك فرینهٔ[1] داخلی خطای $f(x)-p_n^*(x)$ که در آن مشتق اول می‌بایـد صفرشود، و یك فرینهٔ کرانی‌که در آن نیازی به‌صفر شدن مشتق نیست (گرچه می‌باید درنامساوی دیگری که دراینجا ذکر نشده‌است صدق کند). دراالگوریتم رمه[2] و درشکل دیگر آن یعنی شکل مورناهان‌ـرنچ[3] (به‌کتاب رایس[۱۷] نگاه‌کنید) سعی بر آن است‌که این دستگاه معادلات را با روش نیوتن‌که در فصل ۵ بحث شد حل‌کنند. اما برای انجام این کار دستگاه معادلات را بـا ساختارخاصی از دستگاه غیرخطی (۱۴.۶) تطابق می‌دهند. یك تخمین اولیه با یك بسجمله‌ایی درونیاب مناسب $p_n\in\pi_n$ برای $f(x)$ و بـا به‌کارگیری ضرایب $p_n(x)$ و فرینهٔ موضعی $f(x)-p_n(x)$ به‌آسانی به‌دست می‌آید.

در اینجا وقت خوانندگان رابرای بحث دربارهٔ جزئیات ایجاد بهترین بسجمله‌ای تقریب یکنواخت نمی‌گیریم. زیـرا تقریباً بهترین تقریب را می‌توان با سعی‌کمتر و بـا درونیابی مناسب به‌دست آورد.

درحقیقت با توجه به‌قضیهٔ ۲.۶ می‌دانیم‌که خطای $f(x)-p_n^*(x)$ در بهترین‌تقریب یکنواخت‌برای تابع پیوستهٔ $f(x)$ در بازهٔ $a\leqslant x\leqslant b$ می‌باید $n+1$ مرتبه تغییر[4] کند، یعنی می‌باید در رابطهٔ

$$(-1)^i[f(x_i)-p_n^*(x_i)]=\varepsilon\|f-p_n^*\|_\infty \qquad i=0,\dots,n+1$$

باشرایط $\varepsilon=\operatorname{signum}[f(x_0)-p_n^*(x_0)]$ و $a\leqslant x_0<\dots x_{n+1}\leqslant b$ صدق نماید. دراین صورت به‌موجب قضیهٔ ارزش متوسط برای توابع پیوسته (قضیهٔ ۱.۳)، بایدنقاطی مانند $\xi_n<\dots<\xi_0$ باشرط $x_i<\xi_i<x_{i+1}$، به‌ازای جمیع‌مقادیرند، طوری‌وجود داشته باشندکه دراین‌نقاط، خطای $f(x)-p_n^*(x)$ برابر صفر شود، یعنی در این نقاط، در درونیابی $f(x)$ مقدار $p_n^*(x)$ بهترین تقریب باشد. بنابراین‌دراصل تنها اگر بدانیم‌که درکدام نقاط بایددرونیابی‌را انجام دهیم آنگاه می‌توانیم حتی بهترین تقریب را ازراه درونیابی بنا کنیم.

اما خاطرنشان می‌سازیم کـه بهترین تقریب بـرای x^{n+1} از π_n در بـازهٔ استانداردِ $-1\leqslant x\leqslant 1$ مضربی است از بسجمله‌ای چبیشف از درجهٔ $n+1$ یعنی $T_{n+1}(x)$، کـه بنابر خود تعریف (۹.۶) در رابطهٔ $n+1$ نقطهٔ

---

1. extremum 2. Remez 3. Murnaghan_Wrench
4. alternate

$$\xi_{k,n+1}=\cos\frac{2k+1}{2n+2}\pi \qquad k=0,\dots,n \qquad (15.6)$$

صفر می‌شود. این بدان معنی‌است که برای تابع خاص $f(x)=x^{n+1}$ می‌توانیم بهترین تقریب‌سازیکنواخت را از π_n، ازراه درونیابی درنقاط (۱۵.۶)، به‌اصطلاح **نقاط چبیشف** برای بازهٔ استانداردِ $-1\leqslant x\leqslant 1$، به‌دست آوریم. چنانکه معلوم خواهد شد، این شیوهٔ عمل، تقریبهای خوبی (اگر بهترین نباشد) برای توابع پیوسته تولید می‌کند.

برای اینکه ببینیم چرا باید چنین باشد با توجه به‌روابط (۱۶.۲) یا (۳۷.۲) یادآوری می‌کنیم که خطای $f(x)-p_n(x)$ در بسجمله‌ای درونیاب برای $f(x)$ درنقاط $x_0,\dots,x_n$ در رابطهٔ زیرصدق می‌کند

$$f(x)-p_n(x)=f[x_0,\dots,x_n](x-x_0)\cdots(x-x_n)$$

درنتیجه، به‌موجب رابطهٔ (۴.۶) داریم

$$|f(x)-p_n(x)|\leqslant|x-x_0|\cdots|x-x_n|\cdot W(x_0,\dots,x_n,x)\,\mathrm{dist}_\infty(f,\pi_n)$$

به‌شرط آنکه $x_0,\dots,x_n$ و x دربازهٔ مورد نظر قرارگرفته باشند. اکنون قرار می‌دهیم $x=x_{n+1}$. لذا با توجه به‌رابطهٔ (۳.۶) داریم

$$W(x_0,\dots,x_n,x)=\sum_{i=0}^{n+1}1\Big/\prod_{\substack{j=0\\j\neq i}}^{n+1}|x_j-x_i|$$

و بنابراین

$$\begin{aligned}|x-x_0|\cdots|x-x_n|W(x_0,\dots,x_n,x)&=|x_{n+1}-x_0|\cdots|x_{n+1}-x_n|W(x_0,\dots,x_{n+1})\\&=\sum_{i=0}^{n+1}\prod_{\substack{j=0\\j\neq i}}^{n+1}\left|\frac{x_{n+1}-x_j}{x_i-x_j}\right|\\&=1+\sum_{i=0}^{n}|l_i(x_{n+1})|\end{aligned}$$

که درآن

$$l_i(x)=\prod_{\substack{j=0\\j\neq i}}^{n}\frac{x-x_j}{x_i-x_j} \qquad (16.6)$$

iامین بسجمله‌ای لاگرانژ است [به (۵.۲) و (۶.۲) نگاه کنید]. این مطلب اثبات قضیهٔ زیراست

قضیهٔ ۳۰۶ گیریم $p_n(x)$ یك بسجمله‌ای ازدرجهٔ نابیشتر از n باشدکه $f(x)$ را درنقاط $x_0 < x_1 < \cdots < x_n$ در بازهٔ مطلوب $a \leqslant x \leqslant b$ درونیابی کند.دراین صورت داریم

$$\operatorname{dist}_\infty(f, \pi_4) \leqslant \|f - p_n\|_\infty \leqslant (1 + \|\Lambda_n\|_\infty) \operatorname{dist}_\infty(f, \pi_n) \quad (17.6)$$

که درآن Λ_n با تساوی

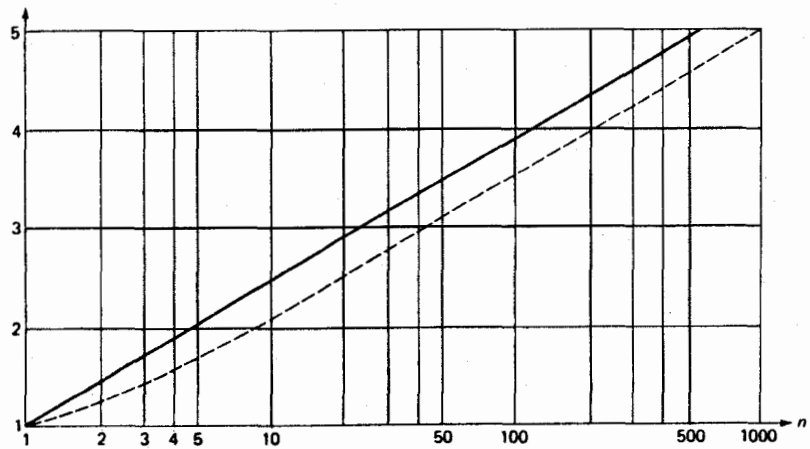
$$\Lambda_n(x) = \sum_{i=0}^{n} |l_i(x)|$$

داده شده‌است و بسجمله‌ای لاگرانژ $l_i(x)$ با رابطهٔ (۱۶.۶).

این امر ایجاب می‌کندکه‌نقاط درونیابی $x_0, \ldots, x_n$ در بازهٔ $a \leqslant x \leqslant b$ چنان انتخاب شودکه نرم‌یکنواخت $\|\Lambda_n\|_\infty$ از تابع لبگ[1] $\Lambda_n(x)$ تا حد ممکن کوچك باشد. انفاقاً این امر تقریباً به‌وسیلهٔ **نقاط چبیشف** (۱۵.۶) که دربازهٔ موردنظر $a \leqslant x \leqslant b$ تنظیم شده‌اند، یعنی با نقاط

$$x_i = \left[a + b + (a - b) \cos \frac{2i + 1}{2n + 2} \pi \right] / 2 \qquad i = 0, \ldots, n \quad (18.6)$$

امکانپذیر می‌شود. در شکل ۴.۶، $\|\Lambda_n\|_\infty$ برحسب این نقـاط به‌عنوان تابعی از n ترسیم شده‌است. اعداد $\|\Lambda_n^e\|_\infty$ که متناظر با نقاط به‌اصطلاح **مبسوط چبیشف**[2]



شکل ۴.۶ عدد $\|\Lambda_n\|_\infty$ برای نقاط چبیشف (خط‌پر) و برای‌نقاط مبسوط‌چبیشف (خط‌چین) به‌عنوان تابعی از n.

1. Lebesgue 2. expanded Chebyshev points

$$x_i^c = \left[a+b+(a-b)\left(\cos\frac{2i+1}{2n+2}\pi\right) \Big/ \left(\cos\frac{\pi}{2n+2}\right)\right] \Big/ 2$$

$$i = 0, \cdots, n \qquad \text{(۱۸.۶ الف)}$$

هستند نیز مشخص شده‌اند. می‌توان نشان داد که $\|\Lambda_n^c\|_\infty$ به‌ازای جمیع مقادیر n درمحدودهٔ ۰٫۲ کوچکترین مقدار ممکنه برای $\|\Lambda_n\|_\infty$ است.

ازشکل ۴.۶ و قضیهٔ ۳.۶ چنین بر می‌آید که به‌ازای $n \leqslant 47$، خطا در بسجمله‌ای درونیاب $f(x)$ درنقاط مبسوط چبیشف (۱۸.۶ الف)، هیچگاه ازچهار برابر بهترین خطای ممکنه بزرگتر نیست ومعمولا کوچکتر از آن است. اگرمثلا درتمام نقاط بازهٔ $a \leqslant x \leqslant b$، $p_n^*(x)$ بهترین تقریب یکنواخت در محدودهٔ 10^{-5} از $f(x)$ باشد، آنگاه بسجمله‌ای درونیاب، دربدترین حالت، درمحدودهٔ 4×10^{-5} از $f(x)$ خواهد بود، که با از دست دادن دقتی به‌اندازهٔ کمتر از نصف یك رقم اعشاری معادل است. ازدست دادن این مقدار دقت معمولا با درونیابی از راه یك بسجمله‌ای از یك یا دو درجه بالاتر امکانپذیر است.

بعکس، اگر Λ_n^u معرف تابع لبگ برای فواصل یکنواخت از نقاط درونیابی، به‌نحوی که هنگام درونیابی از جدول پیش می‌آید باشد، آنگاه خواهیم داشت:

$$\|\Lambda_n^u\| \geqslant e^{n/2} \qquad (۱۹.۶)$$

که با افزایش n به‌سرعت افزایش می‌یابد. (برای به‌دست‌آوردن نتیجه‌ای از این نوع به‌کتاب ریولین [۳۵؛ صفحهٔ ۹۹] مراجعه کنید).

□ **مثال ۴.۶:** درمثال ۲.۶، برای $f(x) = \tan(\pi/4x)$ در بازهٔ متعارف $-1 \leqslant x \leqslant 1$، کران پایین $\text{dist}_\infty(f, \pi_4) \leqslant 0.00052$ به‌دست آمد و همانجا گفته شد که در واقع داریم $\text{dist}_\infty(f, \pi_4) = 0.000410\cdots$. اگر برای این $f(x)$ در پنج نقطهٔ مبسوط چبیشف (۱۸.۶ الف) درونیابی انجام گیرد، بسجمله‌ای $p(x)$ (که به‌دلیل تقارن آرمانی از درجهٔ ۳ است) به‌دست خواهد آمد که برای آن داریم $\|f-p\|_\infty = 0.000576\cdots$ که بزرگی آن تنها به‌اندازهٔ ۱٫۴ برابر کوچکترین مقدار خطای ممکن است. با افزودن تنها یك نقطهٔ درونیابی [که از لحاظ محاسباتی ارزانتر از ساختن $p_4^*(x)$ است] یك بسجمله‌ای ازدرجهٔ ۵ تولید می‌شود که فاصلهٔ آن از $f(x)$ برابراست با $0.000068\cdots$، که بهبود قابل توجهی است نسبت به $\text{dist}_\infty(f, \pi_4) = 0.000410\cdots$.

تمرین

۱-۱۰.۶ با استفاده از (۷.۶) کران پایین $\text{dist}_\infty(e^x, \pi_3)$ را در بازهٔ $-1 \leqslant x \leqslant 1$ براورد کنید. و آن را با فاصلهٔ تابع e^x از بسجمله‌ای $p_3(x)$ از درجهٔ نابیشتر از ۳، که با تابع e^x درچهار نقطهٔ مبسوط چبیشف هماهنگ است مقایسه کنید [به‌رابطهٔ (۱۸.۶ الف) به‌ازای $n = 3$ نگاه کنید].

۲-۱۰۶ تمرین ۱-۱۰۶ را برای بازهٔ $-۱ \leqslant x \leqslant ۰$ تکرار کنید. (راهنمایی: درعوض، تابع $e^{(x+۱)/۲}$ را در بازهٔ $-۱ \leqslant x \leqslant ۱$ بررسی کنید).

۳-۱۰۶ در تمرینهای ۱-۱۰۶ و ۲-۱۰۶، با استفاده از بسجمله‌ای درونیاب $p_۳(x)$ و قضیهٔ ۱۰۶، کران پایین دیگری برای $\operatorname{dist}_\infty(e^x, \pi_۳)$ به‌دست آورید (توجه: برای تعیین بزرگترین کران پایین، باید فرینهٔ $e^x - p_۳(x)$ را، مثلا با روش نیوتن، محاسبه کرد).

۴-۱۰۶ $p_۳^*(x)$ را برای e^x در بازهٔ متعارف $-۱ \leqslant x \leqslant ۱$ محاسبه کنید. [راهنمایی: در این حالت از روش نیوتن برای حل دستگاه معادلات (۱۴.۶) استفاده کنید، به‌عنوان حدس اولیه برای $p_۳^*(x)$ بسجمله‌ای درونیاب $p_۳(x)$ را، که در تمرین ۱-۱۰۶ ساخته شد بگیرید و به‌عنوان حدس اولیه برای نقاط تناوب $x_۴ < \cdots < x_۱ < x_۰$، از فرینهٔ موضعی $e^x - p_۳(x)$ شروع کنید. $۱ = x_۴^{(۰)} < \cdots < x_۱^{(۰)} < x_۰^{(۰)} = -۱$ مربوط به‌خطای توجه نمایید که بنا برقضیهٔ ۲.۰۶، داریم $x_۴ = ۱$ و $x_۰ = -۱$].

۵-۱۰۶ رابطهٔ (۶.۶) را ثابت کنید [راهنمایی: تحقیق کنید که، چون x_iها فرینه‌های موضعی $T_{n+۱}(x)$ هستند، به‌ازای مقداری از x_i که با رابطهٔ (۵.۶) معین می‌شود و به‌ازای یك ثابت مناسب c_n، تساوی $w(x) = c_n(۱ - x^۲)T'_{n+۱}(x)$ برقرار است. معادلهٔ دیفرانسیل $(۱ - x^۲)T''_k(x) = xT'_k(x) - k^۲T_k(x)$ را با مشتقگیری از رابطهٔ (۹.۶) نسبت به θ به‌دست آورید و از آن برای حذف $(۱ - x^۲)T''_{n+۱}(x)$ از عبارت $w'(x)$ حاصل استفاده کنید. و نیز از این معادلهٔ دیفرانسیل برای اثبات تساوی $xT'_{n+۱}(x) = (n+۱)^۲T_{n+۱}(x)$ به‌ازای $x = x_۰$ و $x_{n+۱}$، کمك بگیرید. سرانجام به برقراری تساویهای $T_{n+۱}(x_j) = (-۱)^j$ و $T'_{n+۱}(x_j) = ۰$ به‌ازای $j = ۱, \ldots, n$ احتیاج پیدا می‌کنید].

۶-۱۰۶ ثابت کنید که بهترین تقریب یکنواخت $p_۱^*(x)$ برای تابع کوژ[1] $f(x)$ در بازهٔ $a \leqslant x \leqslant b$، تقریب $p_۱^*(x) = p_۱(x) + (۱/۲)\min_{a \leqslant y \leqslant b}\{f(y) - p_۱(y)\}$ است که در آن $p_۱(x)$ خط مستقیمی است که در نقاط a و b با $f(x)$ منطبق است.

۷-۱۰۶ گیریم $p_n^*(x)$ بهترین تقریب یکنواخت برای $f(x)$ در بازهٔ متعارف $-۱ \leqslant x \leqslant ۱$ باشد. با استفاده از یکتایی[2] $p_n^*(x)$، ثابت کنید که $p_n^*(x)$ تابعی فرد (زوج) است، در صورتی که $f(x)$ فرد (زوج) باشد یعنی، به‌ازای جمیع مقادیر x تساوی

$$f(-x) = -f(x) \quad (f(-x) = f(x))$$

برقرار است.

از آنجا نتیجه گیری کنید که کران پایین حاصل در مثال ۲.۰۶، برای

$$\operatorname{dist}_\infty\left(\tan\frac{\pi}{۴}x, \pi_۴\right)$$

---

1. convex 2. uniqueness

کران پایین برای $\operatorname{dist}_\infty(\tan(\pi/4)\,x, \pi_3)$ نیزهست.

۸-۱۰۶ فرض کنیدکه تابع $\psi(x)$ بر بسجمله‌ایهای ازدرجهٔ نابیشترازn در بازهٔ $a \leqslant x \leqslant b$ متعامداست یعنی، به ازای جمیع مقادیر $p \in \pi_n$، تساوی $\int_a^b \psi(x)P(x)dx = 0$ برقرار است. دراین‌صورت ثابت کنیدکه به‌ازای هر تابع خاص پیوستهٔ $f(x)$ رابطهٔ

$$\left|\int_a^b \psi(x)f(x)dx\right| \Big/ \int_a^b |\psi(x)|dx \leqslant \operatorname{dist}_\infty(f, \pi_n)$$

برقراراست.

۹-۱۰۶ از فرمول جمع کوسینوسها برای اثبات، رابطهٔ (۱۱.۶) استفاده‌کنید.

۱۰-۱۰۶ برای تابع $f(x) = \sqrt{x}$ یك بسجمله‌ای تقریب مناسب ازدرجهٔ n، به‌ازای $n = 1, 2, 3, \ldots, 10$، در بازهٔ $0 \leqslant x \leqslant 1$ پیدا و تحقیق کنیدکه

$$\operatorname{dist}_\infty(\sqrt{x}, \pi_n) \approx \operatorname{const} n^{-1}$$

با توجه به‌نتیجهٔ فوق درجهٔ n لازم برای برقراری $\operatorname{dist}_\infty(\sqrt{x}, \pi_n) \leqslant 10^{-6}$ را براورد کنید.

۱۱-۱۰۶ تمرین ۱۰۶-۱۰ را دربازهٔ $-1 \leqslant x \leqslant 1$ تکرار و فرض کنیدکه

$$\operatorname{dist}_\infty(\sqrt{|x|}, \pi_n) \approx \operatorname{const} n^{-\alpha}$$

مقدار α را تخمین بزنید.

۱۲-۱۰۶ محاسبات مثال ۴.۲ را تکرارکنید ولی این‌بار برای نقاط درونیابی، به‌جای استفاده ازنقاط درونیابی با فواصل‌مساوی، از نقاط مبسوط چبیشف (۱۸.۶ الف) استفاده کنید. نتایج به‌دست آمده را با نتایج مثال ۴.۲ مقایسه و سعی کنید که آنها را برحسب شکلهای ۴.۶ و رابطهٔ (۱۹.۶) بیان کنید.

۱۳-۱۰۶ تمرین ۱۰۶-۱۲، را منتهی برای تابع $f(x) = |x|$ تکرارکنید. (این مثال خوبی برای این امراست‌که درتقریب با بسجمله‌ایها، رفتار نامناسب تابع **در یك ناحیه** به‌تقریب بد **درهمه جا** منجرمی‌شود. به‌کارگرفتن بسجمله‌ای-**تکه‌ای** تقریب‌ساز روش‌خوبی برای اجتناب از این صورت ناخوش‌آیند تقریب با بسجمله‌ایهاست.)

۱۴-۱۰۶ ثابت کنیدکه‌کران پایینی‌که از رابطهٔ (۴.۶) به‌دست آمد می‌تواند به‌صورت $|f[x_0, \ldots, x_{n+1}]/g[x_0, \ldots, x_{n+1}]|$ محاسبه شودباشرط $x_0 < x_1 < \cdots < x_{n+1}$، و در این رابطه $g(x)$ هرتابعی است که برای آن به‌ازای جمیع مقادیر i، تساوی

---

* کلمهٔ انگلیسی const در رابطهٔ فوق، به‌مفهوم مقدار ثابت آمده‌است.ـم.

$g(x_i)=(-1)^i$ برقرار است. آنگاه الگوریتم (۳.۲) را چنان تغییر دهید که محاسبهٔ $g[x_0,\cdots,x_{n+1}]$ همزمان با محاسبهٔ $f[x_0,\cdots,x_{n+1}]$ انجام گیرد.

۲.۶ برازاندن داده‌ها[1]

تاکنون دربارهٔ تقریب‌زدن یك تابع از راه درونیابی در بعضی نقاط بحث کرده‌ایم. در این شیوهٔ عمل قبلا فرض می‌شود که مقادیر $f(x)$ در این نقاط معلوم‌اند. از این‌رو درونیابی در شرایط معمول زیر کمتر به‌کار می‌رود (اگر نگوییم که بکلی خطرناك است): تابع $f(x)$ معرف رابطهٔ بین دو کمیت فیزیکی x و $y=f(x)$ است و اعداد f_n از طریق اندازه‌گیری و یا آزمایش دیگر به‌دست آمده‌اند که تنها مقدار $f(x)$ را در x_n تقریب می‌کنند، یعنی

$$f(x_n)=f_n+\varepsilon_n \qquad n=1,\cdots,N$$

که در اینجا خطاهای تجربی[2] ε_n معلوم نیستند. مسئلهٔ برازاندن داده‌ها عبارت است از بازیابی تابع $f(x)$ از روی داده‌های (تقریبی) مفروض f_n به‌ازای $n=1,\cdots,N$.

دقیق بگوییم، هرگز نمی‌دانیم که در اعداد f_n خطا وجود دارد. بلکه براساس اطلاعات موجود از تابع $f(x)$ یا حتی تنها از روی احساس درمورد چگونگی تابع، اظهار نظر می‌کنیم که $f(x)$ تابعی به‌پیچیدگی f_n نیست و یا با آن سرعتی که اعداد f_n نشان می‌دهند، تغییر نمی‌کند و بنابراین به‌این باور می‌رسیم که می‌باید در اعداد f_n خطا وجود داشته باشد.

برای مثال، داده‌هایی را که در شکل ۵.۶ مشخص شده‌اند در نظر می‌گیریم. در اینجا

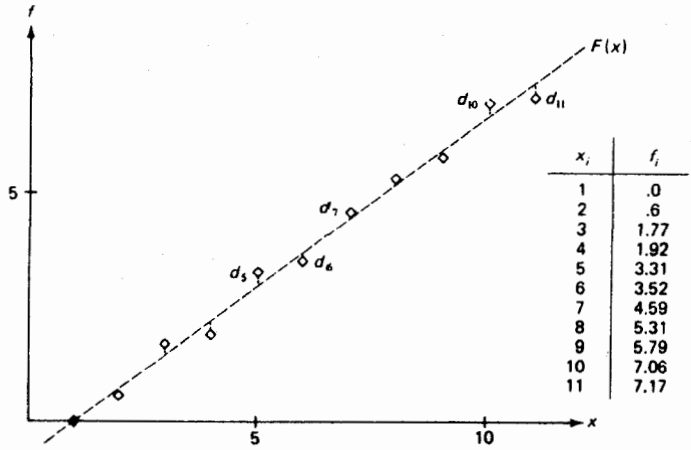
$$x_n=x \qquad n=1,\cdots,11$$

اگر دلیلی برای این باور داشته باشیم که تابع $f(x)$ یك خط راست است، مطمئناً در داده‌های مفروض خطا وجود دارد. اگر فقط بدانیم که $f(x)$ یك تابع کوژ است، باز می‌توانیم نتیجه بگیریم که داده‌ها دارای خطاهستند. حتی اگر چیزی دربارهٔ $f(x)$ ندانیم، باز ممکن است وسوسه شویم و از روی شکل ۵.۶ نتیجه بگیریم که $f(x)$ یك خط راست است، ولو اینکه اکنون زمینهٔ محکمی برای آن نداشته باشیم. اما خواه در بارهٔ $f(x)$ چیزی بدانیم یا ندانیم می‌توانیم از روی داده‌های رسم شده نتیجه‌گیری کنیم که بسیاری از اطلاعات مربوط به $f(x)$ که در داده‌های f_n وجود دارند، می‌توانند به‌خوبی با یك خط راست نشان داده شوند.

خلاصه بگوییم، برازاندن داده‌ها براساس این باور نهاده شده است که داده‌های مفروض f_n شامل مؤلفه‌ای است که به‌کندی تغییر می‌کند و به $f(x)$ نزدیك می‌شود و یا اطلاعاتی دربارهٔ آن می‌دهد و یك مؤلفهٔ دیگری با دامنهٔ نسبتاً کوتاه است که در مقایسه با

---

1. Data Fitting 2. experimental



x_i	f_i
1	.0
2	.6
3	1.77
4	1.92
5	3.31
6	3.52
7	4.59
8	5.31
9	5.79
10	7.06
11	7.17

شکل ۵.۶ تقریب خط راست با روش کوچکترین توانهای دوم.

مؤلفهٔ اول به‌سرعت تغییر می‌کند و معرف **خطا** یا **نوفه**[1] در داده‌هاست. کاری که می‌باید انجام گیرد این‌است که به‌وسیلهٔ تابعی مانند $F^*(x)$ داده‌هایی تقریب‌زده یا برازانده شود که $F^*(x)$ شامل یا معرف بیشترین (اگرنه همهٔ) اطلاعات موجود درداده‌ها درمورد $f(x)$ ومتضمن اندکی خطا یا نوفه (اگر اصلا هست) باشد.

برای برخی داده‌ها این‌کار عملا با انتخاب تابعی مانند

$$F(x)=F(x;\, c_1, \cdots, c_k) \qquad (۲۰.۶)$$

که به‌پارامترهای معین $c_1, \cdots, c_k$ بستگی دارد، انجام می‌گیرد. معمولا سعی می‌شود که تابع $F(x)$ چنان انتخاب شود که به‌طور خطی به‌پارامترها بستگی داشته باشد، به‌طوری که $F(x)$ بتواند به‌صورت

$$F(x)=c_1\phi_1(x)+c_2\phi_2(x)+\cdots+c_k\phi_k(x) \qquad (۲۱.۶)$$

درآید که در آن $\{\phi_i\}$، مجموعه‌ای است از توابع که از قبل انتخاب شده است و $\{c_i\}$ پارامترهایی هستند که می‌باید تعیین شوند. برای مثال، $\{\phi_i\}$ ممکن است به‌صورت مجموعه‌ای از تکجمله‌ایهای[2] $\{x^{i-1}\}$ یا مجموعه‌ای از توابع مثلثاتی $\{\sin \pi i x\}$ باشد. معمولا در مقایسه با تعداد نقاط داده‌ها یعنی N، عدد k کوچک است. انتظار ما این‌است که از یک طرف k، به‌قدر کافی بزرگ باشد تا اطلاعات مربوط به $f(x)$، موجود در داده‌ها، با انتخاب مناسبی برای پارامترهای $c_1, \cdots, c_k$، به‌خوبی نشان داده شود، درحالی که ازطرف دیگر،

1. noise 2. monomials

k آنقدر کوچك باشد که اجازهٔ تولید مجدد خطا یا نوفه را ندهد.

زمانی که افراد مجرب در فن برازاندن داده‌ها تصمیم بگیرند که شکل صحیح (۲۰.۶) را بـرای تابـع تقریب به‌کار برند، می‌بایـد برای پارامترهـای c_i مقادیر خاصی مانند $c_1^*, \ldots, c_k^*$ چنان تعیین کنند که به‌تقریب «مطلوب» $F^*(x) = F(x; c_1^*, \ldots, c_k^*)$ دست یابند. فکر کلی این است که $\{c_i\}$ طوری انتخاب شود که مقادیر انحراف

$$d_n = f_n - F(x_n; c_1, \ldots, c_k) \qquad n = 1, \ldots, N$$

به‌طور همزمان تا حد امکان کوچك شوند. (برای چنین انحرافی در یك مثال نمونه، به شکل ۵.۶ نگاه کنید). برحسب اصطلاحات مذکور در فصل ۴، سعی می‌شود که یك نرم از N-برداری $\mathbf{d} = [d_1\ d_2 \ldots d_N]^T$ بسازند کـه مقدارش تـا حد امکان کوچك باشد، یعنی سعی می‌شود که $\|d\|$ به‌عنوان تابعی از $c_1, \ldots, c_k$، حداقل شود.

انتخاب متداول برای نرم به‌شرح زیر است

(i) ۱-نرمی

$$\|\mathbf{d}\|_1 = \sum_{n=1}^{N} |d_n|$$

این نرم موقعی انتخاب می‌شود که خواسته باشیم انحراف میانگین تا حـد امکان کوچك باشد، یا

(ii) ∞-نرمی

$$\|\mathbf{d}\|_\infty = \max_{1 \le n \le N} |d_n|$$

وقتی که بخواهیم کلیهٔ انحرافات به‌طور یکنواخت کوچك باشند.

اما، اگر سعی کنیم که این مسئلهٔ حداقل‌سازی را درقالب فصل ۵ یا به‌روشهای دیگر حل کنیم، فوراً درخواهیم یافت که تعیین مینیمم‌های $c_1^*, \ldots, c_k^*$، به‌حل یك دستگاه معادلات غیرخطی منجر می‌شود [مثلا، برای مسئلهٔ مربوط به‌تقریب یکنواخت در یك بازه، به‌دستگاه معادلات (۱۴.۶) نگاه کنید]. بنابراین رسم چنین است که آن نرمی را که باید مینیمم شود،

۲-نرمی انتخاب می‌کنند.

$$\|\mathbf{d}\|_2 = \left(\sum_{n=1}^{N} |d_n|^2 \right)^{1/2}$$

زیرا این نرم در پیدا کردن مینیمم مقادیر c_i^* به‌دستگاه معادلات خطی منجر می‌شود. دراین حال تقریب نتیجـه شدهٔ $F(x; c_1^*, \ldots, c_k^*)$ را **تقریب با روش کوچکترین مربعات** برای داده‌های مفروض نامند.

اکنون دستگاه معادلات برای استخراج c_i^*ها را به‌دست می‌آوریم. ازآنجاکه تابع

جذر[1]، یکنواست، مینیمم‌سازی $\|\mathbf{d}\|_2$ مثل مینیمم‌سازی $\|\mathbf{d}\|_2^2$ خواهد بود. البته برای اینکه $\mathbf{c}^*=[c_1 \cdots c_k]^T$ مینیمم تابع

$$E(\mathbf{c})=E(c_1,\ldots,c_k)=\|\mathbf{d}\|_2^2=\sum_{n=1}^{N}[f_n-F(x_n;\mathbf{c})]^2$$

باشد، لازم‌است که گرادیان E در $\mathbf{c}^*$ صفرشود، یعنی

$$\nabla E(\mathbf{c}^*)=0$$

(به‌بخش ۱۰۵ نگاه‌کنید)، بنابراین چون بنابر (۲۱۰۶) داریم

$$(\partial/\partial c_i)[f_n-F(x_n;\mathbf{c})]=-\phi_i(x_n)$$

$\mathbf{c}$ می‌باید در به‌اصطلاح **معادلات نرمال**

$$-2\sum_{n=1}^{N}[f_n-F(x_n;\mathbf{c}^*)]\phi_i(x_n)=0 \qquad i=1,\ldots,k \qquad (22.6)$$

صدق کند.

صفت «نرمال» بدین دلیل به‌این معادلات داده شده‌است که آنها تصریح می‌کنند که بردارخطای $\mathbf{e}=[e_1\ e_2 \cdots e_N]^T$ با $e_n=f_n-F(x_n;\mathbf{c}^*)$، به ازای کلیهٔ مقادیر n، باید برهرکدام از k-بردارهای

$$\boldsymbol{\phi}_i=[\phi_i(x_1)\ \phi_i(x_2)\cdots\phi_i(x_N)]^T \qquad i=1,\ldots,k$$

نرمال یا عمود باشد. زیرا برحسب این N-بردارها، معادلات (۲۲.۶) به‌صورت

$$-2\mathbf{e}^T\boldsymbol{\phi}_i=0 \qquad i=1,\ldots,k$$

در می‌آیند. ازآنجاکه تابع تقریب کلی به‌شکل

$$F(x)=c_1\phi_1(x)+c_2\phi_2(x)+\cdots+c_k\phi_k(x)$$

است، این‌رابطه می‌گوید که بردارخطا باید (با این تعبیر) برهمهٔ توابع تقریبی‌ممکن‌عمود باشد، یعنی

$$\mathbf{e}^T(c_1\boldsymbol{\phi}_1+c_2\boldsymbol{\phi}_2+\cdots+c_k\boldsymbol{\phi}_k)=0 \qquad \text{به‌ازای جمیع مقادیر } c_1,\ldots,c_k$$

که این رابطه بردار $c_1^*\boldsymbol{\phi}_1+c_2^*\boldsymbol{\phi}_2+\cdots+c_k^*\boldsymbol{\phi}_k$ را به‌عنوان تصویرقائم[2] بردار داده‌ها، $\mathbf{f}=[f_1\ f_2\cdots f_N]^T$، روی ابرصفحه‌ای[3] که براثر $\boldsymbol{\phi}_1$، $\boldsymbol{\phi}_2$، ...، $\boldsymbol{\phi}_k$ تنیده شده معرفی می‌کند.

---

1. square-root function
2. orthogonal projection
3. hyperplane

معادلات نرمال را مجدداً به‌شکل

$$\sum_{j=1}^{k} c_j^* \phi_j^T \phi_i = \mathbf{f}^T \phi_i \qquad i = 1, \dots, k \qquad (۲۳.۶)$$

می‌نویسیم تا صریحاً بیان کنیم که این معادلات دستگاهی از k معادلهٔ خطی با k مجهول c_1^*، c_2^*، ...، c_k^* را تشکیل می‌دهند. چنانچه خواهیم دید، این دستگاه همواره دست‌کم دارای یک جواب است [$\phi_i(x)$ هرچه باشد]. بعلاوه هر جوابی از (۲۳.۶) مقدار $E(c_1, \dots, c_k)$ را مینیمم می‌کند.

برای ارائهٔ یک مثال، اکنون به‌یافتن تقریب با روش کوچکترین مربعات برای داده‌هایی که درشکل ۵.۶ با یک خط راست مشخص شده‌اند می‌پردازیم.

دراین مثال داریم $x_n = n$، $n = 1, \dots, 11$، و

$$F(x; c_1, c_2) = c_1 + c_2 x$$

لذا $k = 2$ و $\phi_1(x) = 1$ و $\phi_2(x) = x$. دستگاه خطی (۲۳.۶) به‌شکل

$$11c_1^* + 66c_2^* = 41.04$$

$$66c_1^* + 506c_2^* = 328.05$$

در می‌آید که وقتی با روش حذف گاوس حل شود، جواب منحصر زیر را به‌دست خواهد داد:

$$c_1^* = -0.73140\dots \qquad c_2^* = 0.74370\dots$$

خط راست حاصله نیز درشکل ۵.۶ رسم شده‌است.

در اینجا باید به‌این واقعیت ناخوشایند اشاره شود که اگر ماتریس ضرایب (۲۳.۶) در اکثر موارد چنان بدشرط نمی‌شد که کاربرد مستقیم الگوریتم حذفی ۲.۴ را به‌نتایج غیرقابل اعتمادی منجرسازد، آنگاه همه‌چیز طبق دلخواه روبه‌راه می‌شد. این مطلب بامثال سادهٔ زیر نشان داده شده‌است.

□ **مثال ۵.۶:** مقادیر تقریبی $f_n \approx f(x_n)$ به‌ازای

$$x_n = 10 + \frac{n-1}{5} \qquad n = 1, \dots, 6$$

داده شده‌اند ودلایلی داریم که معتقد باشیم این‌داده‌ها می‌توانند به‌نحو مناسبی با یک سهمی نشان داده شوند. بنابراین مقادیر زیر را انتخاب می‌کنیم

$$\phi_1(x) = 1 \qquad \phi_2(x) = x \qquad \phi_3(x) = x^2$$

دراین حالت ماتریس ضرایب A، حاصل از دستگاه (۲۳.۶) برابر است با

$$A=\begin{bmatrix} ۶ & ۶۳ & ۶۶۲٫۲ \\ ۶۳ & ۶۶۲٫۲ & ۶۹۶۷٫۸ \\ ۶۶۲٫۲ & ۶۹۶۷٫۸ & ۷۳۳۹۳٫۵۶۶۴ \end{bmatrix}$$

و نتیجه می‌شود که $\|A\|_\infty \approx ۸\times ۱۰^۴$. از طرف دیگر به‌ازای

$$\mathbf{x}=\begin{bmatrix} ۱۰٫۰۷ \\ -۲٫۰ \\ ۰٫۰۹۹ \end{bmatrix}$$

خواهیم داشت

$$A\mathbf{x}=\begin{bmatrix} -۰٫۰۲ \\ -۰٫۱۸ \\ -۱٫۲۸ \end{bmatrix}$$

از این‌رو با توجه به‌نامساوی (۳۸.۴) یعنی

$$\|A\mathbf{x}\| \geqslant \frac{\|\mathbf{x}\|}{\|A^{-۱}\|}$$

خواهیم داشت

$$\|A^{-۱}\|_\infty \geqslant ۷٫۸ \quad \text{یا} \quad ۱٫۲۸=\|A\mathbf{x}\|_\infty \geqslant \frac{۱۰٫۰۷}{\|A^{-۱}\|_\infty}$$

بنابراین عدد شرط A برابراست با

$$\operatorname{cond}(A)=\|A\|_\infty\|A^{-۱}\|_\infty \geqslant ۱۰^۵$$

درواقع همان‌گونه‌که نتایج خاص زیر نشان می‌دهند، عدد شرط A بسیار بزرگتر از $۱۰^۵$ است. $f(x)$ را به‌صورت زیر انتخاب می‌کنیم

$$f(x)=۱۰-۲x+\frac{x^۲}{۱۰}$$

و مقادیر دقیق داده‌ها یعنی $f_n=f(x_n)$، $n=۱,\ldots,۶$، را به‌کارمی‌بریم. دراین‌صورت چون $f(x)$ یك بسجمله‌ای از درجهٔ ۲ است، $F^*(x)$ باید خود $f(x)$ باشد، بنابراین‌باید داشته باشیم

$$c_۱^*=۱۰ \qquad c_۲^*=-۲ \qquad c_۳^*=۰٫۱$$

با به‌کاربردن‌الگوریتم حذفی۲.۴، برای حل‌دستگاه (۲۳.۶) روی‌کامپیوتر CDC۶۵۰۰، نتایج‌زیر به‌دست‌می‌آید

$$c_1^* = ۹ر۹۹۹۹۹۹۷۴۳۷۰۰۰ \qquad c_2^* = -۱ر۹۹۹۹۹۹۹۵۱۱۰۰۰$$

$$c_3^* = ۵ر۰۹۹۹۹۹۹۹۷۶۰۰۰$$

لذا محاسبه در حساب ممیز شناور با ۱۴ رقم اعشاری برای این دستگاه ۳×۳، نتیجه‌ای به‌دست می‌دهد که تنها در حدود ۸ رقم اعشاری آن صحیح است. اگر درایهٔ (۳,۳) مربوط به A را به‌صورت ۷۳۳۹۳ر۶ گرد نموده و محاسبه را تکرار کنیم، نتایج به‌دست آمده حیرت‌انگیز است

$$c_1^* = ۶ر۰۳۵۰۰۰ \qquad c_2^* = -۱ر۲۴۳۰۰۰ \qquad c_3^* = ۵ر۰۶۳۹۰۰۰$$

به‌طور مشابه اگر تمام محاسبات به‌صورت ممیز شناور با هفت رقم اعشاری صورت پذیرد، نتایج برابرند با

$$c_1^* = ۸ر۴۹۲۰۰۰ \qquad c_2^* = -۱ر۷۱۲۰۰۰ \qquad c_3^* = ۵ر۰۸۶۳ \qquad \square$$

مثال فوق به‌وضوح نشان می‌دهد که بدون انجام کارهای مقدماتی، اقدام به حل معادلات نرمال می‌تواند خطرناک باشد. اقدام مذکور متضمن انتخاب دقیق $\phi_i(x)$ است.

راه ساده‌ای که برای اجتناب از مسئلهٔ بدشرطی به‌نظر می‌رسد، انتخاب $\phi_i(x)$ است به‌گونه‌ای که در نقاط $x_1, \ldots, x_N$ **متعامد** باشند، یعنی

$$\phi_j^T \phi_i = \sum_{n=1}^{N} \phi_i(x_n)\phi_j(x_n) = ۰ \qquad i \neq j \text{ به‌ازای} \qquad (۲۴.۶)$$

اگر رابطهٔ (۲۴.۶) برقرار باشد، معادلات (۲۳.۶) به‌صورت زیر تبدیل می‌شوند

$$c_i^* \phi_i^T \phi_i = \mathbf{f}^T \phi_i \qquad i = ۱, \ldots, k \qquad (۲۵.۶)$$

که بدیهی است که حل آنها، دیگر مشکل زیادی ندارد.

البته این روش زیبا، جدا از شرط مسئله، تنها مسئله‌ای را جایگزین مسئلهٔ دیگر می‌کند و اکنون می‌باید توابع متعامد را پیدا کنیم. اگر بخواهیم که ϕ_iها نیز به‌صورت بسجمله‌ای باشند، با استفاده از یک رابطهٔ بازگشتی سه‌جمله‌ای که برای دنباله‌های بسجمله‌ایهای متعامد معتبر است، می‌توان چنین توابع بسجمله‌ای متعامدی را به‌نحو کارا درست کرد. این مطلب را در بخشهای ۳.۶ و ۴.۶ مورد بحث قرار می‌دهیم. اگر، همان گونه که در عمل غالباً اتفاق می‌افتد، نتوانیم $f(x)$ را به‌صورت یک بسجمله‌ای فرض کنیم، باید از راههای دیگری برای ساختن توابع متعامد مناسب استفاده کنیم. یک نمونه از چنین تکنیکی الگوریتم اصلاح شدهٔ گرام-اشمیت[1] است که در برخی از متون درسی آمده‌است (برای مثال به کتاب رایس [۱۷] مراجعه کنید). به‌صورت دیگر، باید به‌انتخاب $\phi_1(x), \ldots, \phi_k(x)$ که «تقریباً»

---

1. Gram-Schmidt

متعامد باشند قناعت کرد. معنی این عبارت مهم توضیح این حقیقت است که ماتریس ضرایب دستگاه (۲۳.۶) برای این گونه $\phi_i(x)$ها تقریباً «قطری» مثلا نافذ قطری است. اگر نقاط x_1، ...، x_n در یك بـازهٔ (a, b) تقریبـاً به‌طور یکنواخت توزیع شده باشند، آنـگاه $\phi_1(x)$، ...، $\phi_k(x)$ تقریباً به‌سمت عمـود بودن میل خواهند کرد، اگر هر $\phi_i(x)$ در بازهٔ (a, b) یکبار بیشتر از $\phi_{i-1}(x)$ تغییر علامت دهد (به تمرین ۲.۶ـ۳ رجوع کنید).

تمرین

۱ـ۲.۶ تقریب باروش کوچکترین مربعات را برای داده‌های مرسوم درشکل ۵.۶، به‌وسیلهٔ توابعی به‌صورت

$$F(x)=c_1+c_2x+c_3\sin[۱۲۳(x-۱)]$$

ازحل معادلات نرمال مناسب، محاسبه کنید. آیا حس می کنید که این تقریب معرف تمـام اطلاعات مربوط به $f(x)$ موجود داده‌ها هست؟ چرا؟

۲ـ۲.۶ با دنبال کردن برهان مذکور درمتن، معادله‌های نرمال برای بهترین c_1^*، c_2^* رادر حالت

$$F(x)=F(x;c_1,c_2)=c_1e^{c_2x}$$

به‌دست آورید. آیا این معادله‌های نرمال هنوز خطی هستند؟

۳ـ۲.۶ با استفاده از توابع زیر، تمام محاسبات مثال ۵.۶ را تکرار کنید

$$\phi_1(x)=۱ \qquad \phi_2(x)=x-۱۰٫۵ \qquad \phi_3(x)=(x-۱۰٫۳)(x-۱۰٫۷)$$

با توجه به آخرین بند این بخش، این معادلات نرمال می‌باید حالا خیلی بهشرط‌تر باشند، آیا همینطوراست؟

۳.۶* بسجمله‌ایهای متعامد

دراین بخش، بعضی از ویژگیهای مربوط به‌دنباله‌های بسجمله‌ایهای متعامد ومثالهای خاص آنهـا را به‌اختصار مورد بحث قرار می‌دهیم. اگرچـه انگیزهٔ مستقیم این بحث از مسئلهٔ تقریب به‌روش کوچکترین توانهای دوم با بسجمله‌ایها (که در بخش بعد باید بحث شود) ناشی شده‌است، ولی ما ازآن برای بسجمله‌ایهای متعامد بعـداً در زمینه‌های مختلف، مثلا دربخش ۳.۷ استفاده خواهیم کرد. برای فراهم کردن زمینه برای آن بخش، دراینجا ما از یك مفهوم تعامد توابع استفاده می کنیم که کلیتر ازآن چیزی‌است کـه در بخش ۲.۶ آمده است.

در آنچه که ذیلا می‌آید، گیریم (a, b) یك بازهٔ مفروض و $w(x)$ یك تابع مفروضی تعریف شده (حداقل) در بازهٔ (a, b) ومثبت در آن بازه باشد. بعلاوه **حاصلضرب داخلی** هردو تابع $g(x)$ و $h(x)$ [معین در بازهٔ (a, b)] یعنی

$$\langle g, h\rangle$$

را به‌یکی از دوطریق زیر تعریف می‌کنیم:

$$\langle g, h\rangle = \int_a^b g(x)h(x)w(x)\,dx \qquad (۲۶.۶)$$

یا

$$\langle g, h\rangle = \sum_{n=1}^{N} g(x_n)h(x_n)w(x_n) \qquad (۲۷.۶)$$

درحالت اول فرض می‌کنیم که به‌ازای جمیع توابع $g(x)$ و$h(x)$ موردنظر، این‌انتگرال (حداقل یك انتگرال ناسره[1]) وجود دارد، در حالت دوم فرض می‌کنیم که N نقطهٔ $x_1, \ldots, x_N$، همگی در بازهٔ (a, b) مفروض باشند ودرطی بحث، ثابت در نظرگرفته شوند. توجه‌کنیدکه به‌ازای $w(x)\equiv 1$، رابطهٔ (۲۷.۶) به‌حاصلضرب داخلی دو تابع $\mathbf{g}^T\mathbf{h} = \mathbf{h}^T\mathbf{g}$ تبدیل می‌شودکه در بحث تقریب با روش کوچکترین مربعات در بخش ۲.۶ ظاهرشد.

با معین بودن حاصلضرب داخلی‌دوتابع، دو تابع $g(x)$ و $h(x)$ را **متعامد** (بریکدیگر) گویند هرگاه

$$\langle g, h\rangle = 0$$

برای مثال به‌آسانی تحقیق می‌شودکه توابع $g(x)\equiv 1$، $h(x) = x$ متعامدند اگرحاصلضرب داخلی آنها برابرعبارت زیرباشد

$$\langle g, h\rangle = \int_{-1}^{1} g(x)h(x)\,dx$$

و نیز متعامدند اگرحاصلضرب داخلی آنها برابر

$$\langle g, h\rangle = \sum_{n=-10}^{10} g(n)h(n)$$

یا برابربا

---

1. improper integral

$$\langle g, h\rangle = \int_{-1}^{1} \frac{g(x)h(x)}{(1-x^2)^{1/2}}\,dx$$

باشد. توابع $g(x)=\sin nx$ و $h(x)=\sin mx$ به‌ازای اعداد صحیح m و n، $m \neq n$، متعامد هستند اگر تساوی

$$\langle g, h\rangle = \int_{0}^{2\pi} g(x)h(x)\,dx$$

برقرار باشد، چنانکه این مطلب برای $g(x)=\sin nx$، $h(x)=\cos mx$ صادق‌است.

بعلاوه $P_0(x)$، $P_1(x)$، $P_2(x)$، ... را یک **دنباله** (متناهی یا نامتناهی) از **بسجمله‌ایهای متعامد** نامند به شرط آنکه همهٔ $P_i(x)$ها بریکدیگر عمود باشند و هر $P_i(x)$ یك بسجمله‌ای دقیقاً از درجهٔ i باشد. به‌عبارت دیگر

(i) به‌ازای هر i، (یك بسجمله‌ای از درجهٔ کمتر از i $P_i(x)=(\alpha_i x^i+$ با $\alpha_i \neq 0$ باشد.

(ii) هرگاه $i \neq j$، آنگاه $\langle P_i, P_j\rangle = 0$.

برای مثال توابع

$$P_0(x)\equiv 1 \qquad P_1(x)=x \qquad P_2(x)=3x^2-1$$

یك دنباله از سه بسجمله‌ای متعامد را تشکیل می‌دهند اگر

$$\langle g, h\rangle = \int_{-1}^{1} g(x)h(x)\,dx$$

قبلا اشاره کردیم که $\langle P_0, P_1\rangle = 0$، و نیز داریم

$$\langle P_0, P_2\rangle = \int_{-1}^{1} 1(3x^2-1)dx = x^3-x\Big|_{-1}^{1} = 0$$

و

$$\langle P_1, P_2\rangle = \int_{-1}^{1} x(3x^2-1)\,dx = \frac{3}{4}x^4-\frac{1}{2}x^2\Big|_{-1}^{1} = 0$$

گیریم $P_0(x)$ و $P_1(x)$، ...، $P_k(x)$ یك دنبالهٔ متناهی از بسجمله‌ایهای متعامد باشند. آنگاه ویژگیهای زیر می‌توانند ثابت شوند:

ویژگی ۱ اگر $P(x)$ یك بسجمله‌ای غیرمشخص از درجهٔ نابیشتر از k باشد، آنگاه به‌ازای ضرایب d_0، ...، d_k که به‌وسیلهٔ $p(x)$ به‌طور منحصر به‌فرد تعیین‌شده‌اند، می‌توان $p(x)$ را به‌صورت

$$p(x)=d_0P_0(x)+d_1P_1(x)+\cdots+d_kP_k(x) \qquad (۲۸.۶)$$

نوشت. به‌ویژه اگر

$$p(x)=(a_kx^k+k \text{ از درجهٔ کمتر یك بسجمله‌ای})$$

و اگر ضریب بزرگترین درجهٔ $P_k(x)$ برابر با α_k باشد، آنگاه

$$d_k=\frac{a_k}{\alpha_k}$$

این ویژگی از قسمت (i) مذکور دربالا، و با استقرا روی k نتیجه می‌شود. برای مثال بالا، می‌توان بسجمله‌ای کلی از درجهٔ نابیشتر از ۲ی

$$p_2(x)=a_0+a_1x+a_2x^2$$

را به‌صورت زیرنوشت

$$p_2(x)=\left(a_0+\frac{a_2}{3}\right)P_0(x)+a_1P_1(x)+\frac{a_2}{3}P_2(x)$$

با ترکیب ویژگی ۱ با (ii)، ویژگی۲ به‌دست می‌آید.

ویژگی۲ اگر $p(x)$ یك بسجمله‌ای ازدرجهٔ کمتر از k باشد، آنگاه $p(x)$ و $P_k(x)$ متعامد خواهند بود، یعنی

$$\langle p, P_k\rangle=0$$

اگر درمثال بالا، بگیریم $p(x)=1+x$، خواهیم داشت

$$\langle p, P_2\rangle=\int_{-1}^{1}(1+x)(3x^2-1)\,dx=\frac{3}{4}x^4+x^3-\frac{1}{2}x^2-x\Big|_{-1}^{1}=0$$

این ویژگی به‌اصطلاح «بی‌گزند» دارای چندین نتیجهٔ مهم است.

ویژگی۳ اگرحاصلضرب داخلی با رابطهٔ (۲۶.۶) معین شود، آنگاه $P_k(x)$ دارای k ریشهٔ حقیقی ساده‌است که تمامی آنها در بازهٔ (a, b) قراردارند، یعنی $P_k(x)$ به‌ازای k نقطهٔ متمایز $\xi_{1,k}, \xi_{2,k}, \ldots, \xi_{k,k}$ در بازهٔ (a, b)، به‌شکل زیراست

$$P_k(x)=\alpha_k(x-\xi_{1,k})(x-\xi_{2,k})\cdots(x-\xi_{k,k}) \qquad (۲۹.۶)$$

درمثال خودمان داریم

$$P_0(x)\equiv\alpha_0\equiv 1 \qquad P_1(x)=\alpha_1(x-\xi_{1,1})=1\times(x-0)$$

$$P_2(x)=\alpha_2(x-\xi_{1,2})(x-\xi_{2,2})=3\left(x+\frac{1}{\sqrt{3}}\right)\left(x-\frac{1}{\sqrt{3}}\right)$$

یك دلیل سادهٔ ویژگی۳ به‌قرار زیر است: گیریم $k > 0$ و $\xi_{1,k}, \ldots \xi_{r,k}$ همهٔ نقاطی از بازهٔ (a,b) باشند که در آنها $P_k(x)$ تغییر علامت می‌دهد. آنگاه می‌گوییم

$$r \geqslant k$$

زیرا اگر r کمتر از k باشد، آنگاه به‌ازای $\bar{x} \in (\max_i \xi_{i,k}, b)$ بسجمله‌ای

$$p(x) = P_k(\bar{x})(x-\xi_{1,k})(x-\xi_{2,k}) \cdots (x-\xi_{r,k})$$

یك بسجمله‌ای از درجهٔ کمتر از k خواهد بود که در هر نقطه از بازهٔ (a,b)، علامتی مشابه با $P_k(x)$ دارد. از این‌رو، از یك طرف بنا بر ویژگی۳ داریم

$$\int_a^b p(x)P_k(x)w(x)\,dx = \langle p, P_k \rangle = 0$$

درحالی که از طرف دیگر، به‌ازای جمیع مقادیر $x \in (a,b)$، جز $x = \xi_{1,k}, \ldots, \xi_{r,k}$، خواهیم داشت $p(x)P_k(x)w(x) > 0$ مطمئناً این دو نتیجه متناقض یکدیگرند. در نتیجه باید داشته باشیم $r \geqslant k$، یعنی $P_k(x)$ می‌باید حداقل k مرتبه در بازهٔ (a,b) تغییر علامت دهد. اما چون $P_k(x)$ یك بسجمله‌ای از درجهٔ k و هر $\xi_{i,k}$ یك ریشه از $P_k(x)$ است، لذا r نمی‌تواند از k بزرگتر باشد (بخش۱۰۲ را ببینید)، بنابراین r می‌باید با k برابر باشد یعنی k نقطهٔ متمایز $\xi_{i,k}$، $i = 1, \ldots, k$، دقیقاً ریشه‌های $P_k(x)$ هستند.

به‌طور مشابه ثابت می‌شود که وقتی حاصلضرب داخلی با رابطهٔ (۲۷.۶) معین شده باشد رابطهٔ (۲۹.۶) برقرار خواهد بود به‌شرط آنکه بین x_nها حداقل k نقطهٔ متمایز وجود داشته باشند.

ویژگی۴ بسجمله‌ایهای متعامد در یك **رابطهٔ بازگشتی سه جمله‌ای**[1] صدق می‌کنند. اگر قرار دهیم

$$A_i = \frac{\alpha_i + 1}{\alpha_i} \qquad \text{به‌ازای جمیع مقادیر } i$$

$$P_{-1}(x) \equiv 0$$

و اگر

$$S_i = \langle P_i, P_i \rangle$$

به‌ازای $i = 0, \ldots, k-1$، مخالف صفر باشد، آنگاه این رابطهٔ بازگشتی می‌تواند به‌صورت

$$P_{i+1}(x) = A_i(x-B_i)P_i(x) - C_iP_{i-1}(x) \qquad i = 0, 1, \ldots, k-1 \qquad (۳۰.۶)$$

---

1. three-term recurrence relation

توشته شود، که در آن مقادیر B_i و C_i عبارتند از

$$B_i=\frac{\langle xP_i(x),\, P_i(x)\rangle}{S_i} \qquad i=1,\dots,k-1$$

و

$$C_i=\begin{cases} \text{اختیاری است} & i=0 \\ \dfrac{A_iS_i}{A_{i-1}S_{i-1}} & i>0 \end{cases}$$

این ویژگی می‌تواند برای تولید دنباله‌های کلی بسجمله‌ایهای متعامد (به‌شرط آنکه اعداد S_i و B_i بتوانند محاسبه شوند و S_iها صفر نباشند) به‌کار رود. در این روال معمولا ضرایب پیشرو α_i، یا هم‌ارز با آنها، اعداد A_i، چنان انتخاب می‌شوند که به‌تعبیری دنبالهٔ حاصل بالاخص ساده باشد.

□ **مثال ۶.۶: بسجمله‌ای لژاندر**[1]. اگر حاصلضرب داخلی با

$$\langle g,\, h\rangle=\int_{-1}^{1} g(x)h(x)\,dx$$

داده شده باشد. بسجمله‌ایهای متعامد حاصل، بسجمله‌ایهای لژاندر نامیده می‌شوند با شروع از

$$P_0(x)\equiv 1$$

به‌دست می‌آید

$$S_0=\int_{-1}^{1} 1\,dx=2 \qquad S_0B_0=\int_{-1}^{1} x\cdot 1\,dx=0$$

ازاین‌رو، با توجه به‌ویژگی۴، با انتخاب $A_i=1$، به‌ازای جمیع مقادیر i، خواهیم داشت

$$P_1(x)=x$$

بعلاوه

$$S_1=\int_{-1}^{1} x^2\,dx=\frac{2}{3} \qquad S_1B_1=\int_{-1}^{1} x\cdot x^2\,dx=0 \qquad C_1=\frac{S_1}{S_0}=\frac{1}{3}$$

لذا باز بنابر ویژگی۴

---

1. Legendre

$$P_۲(x)=x^۲-\frac{۱}{۳}$$

مجدداً

$$S_۲=\int_{-۱}^{۱}\left(x^۲-\frac{۱}{۳}\right)^۲dx=\frac{۸}{۴۵} \qquad S_۲B_۲=\int_{-۱}^{۱}x\left(x^۲-\frac{۱}{۳}\right)^۲dx=۰$$

$$C_۲=\frac{۴}{۱۵}$$

بنابراین

$$P_۳(x)=x^۳-\frac{۳}{۵}x$$

معمولا چنین‌است‌که بسجمله‌ایهای لژاندر را به‌صورت نرمال درمی‌آورند تا داشته باشیم

$$P_k(۱)=۱ \qquad \text{به‌ازای جمیع مقادیر } k$$

با این نرمال‌سازی، ضرایب در رابطهٔ بازگشتی به‌صورت زیر درمی‌آیند

$$A_k=\frac{۲k+۱}{k+۱} \qquad B_k=۰ \qquad C_k=\frac{k}{k+۱} \qquad k=۰,۱,۲,\dots$$

لذا

$$P_{k+۱}(x)=\frac{(۲k+۱)xP_k(x)-kP_{k-۱}(x)}{k+۱}$$

جدول ۲.۶، چندتا از نخستین بسجمله‌ایهای لژاندر را به‌ما می‌دهد. □

جدول ۲.۶

k	$P_k(x)$
۰	۱
۱	x
۲	$(۳/۲)(x^۲-۱/۳)$
۳	$(۵/۲)[x^۳-(۳/۵)x]$
۴	$(۳۵/۸)[x^۴-(۶/۷)x^۲+۳/۳۵]$

□ **مثال ۷.۶: بسجمله‌ایهای چبیشف.** اگر حاصلضرب داخلی با

$$\langle g, h\rangle = \int_{-1}^{1} \frac{g(x)h(x)}{(1-x^2)^{1/2}}\,dx$$

داده شده باشد، بسجمله‌ایهای چبیشف $T_k(x)$ که در بخش ۱.۶، معرفی شدند به‌دست خواهند آمد. ما قبلا رابطهٔ بازگشتی آنها را

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) \qquad k = 1, 2, \ldots$$

از روی رابطهٔ معرف آنها

$$T_k(\cos\theta) = \cos k\theta$$

به‌دست آوردیم. □

□ **مثال ۸.۶: بسجمله‌ایهای هرمیت.** زمانی که حاصلضرب داخلی

$$\langle g, h\rangle = \int_{-\infty}^{\infty} g(x)h(x)e^{-x^2}\,dx$$

به‌کار برده شده باشد، $H_k(x)$ به‌دست می‌آید. با اجرای نرمال‌سازی معمولی این بسجمله‌ایها در رابطهٔ بازگشتی زیر صدق می‌کنند

$$H_{k+1}(x) = 2xH_k(x) - 2kH_{k-1}(x) \qquad k = 0, 1, 2, \ldots$$

چندتا از نخستین بسجمله‌ایهای هرمیت درجدول ۳.۶ داده شده‌اند.

جدول ۳.۶

k	$H(x)$
۰	۱
۱	$2x$
۲	$4x^2 - 2$
۳	$8x^3 - 12x$
۴	$16x^4 - 48x^2 + 12$

□

□ **مثال ۹۰۶: بسجمله‌ایهای تعمیم‌یافتهٔ لاگر**[1]. $L_k^{\alpha}(x)$ ها در ارتباط با حاصلضرب داخلی زیر

$$\langle g, h\rangle = \int_{0}^{\infty} g(x)h(x)x^{\alpha}e^{-x}\,dx$$

می‌باشند. ضرایب برای رابطهٔ بازگشتی چنین‌اند

$$A_k = -\frac{1}{k+1} \qquad B_k = 2k+\alpha+1 \qquad C_k = \frac{k+\alpha}{k+1}$$

تولید پنج نخستین بسجمله‌ای لاگر (به‌ازای $\alpha = 0$) به‌عهدهٔ دانشجویان گذاشته شده است (تمرین ۳.۶–۱ را ببینید). □

دو مثال آخر در انتگرالگیری عددی بر بازه‌های نیمه‌نامتناهی یا نامتناهی اهمیت خاصی دارند (بند ۳.۷ را ببینید).

با بحث از الگوریتمی برای ارزیابی بسجمله‌ای مفروضی برحسب بسجمله‌ایهای متعامد، این قسمت را به‌پایان می‌رسانیم. فرض کنید که $P_0(x)$ و $P_1(x)$، ...، $P_k(x)$ یك دنبالهٔ متناهی از بسجمله‌ایهای متعامد باشند، و فرض کنید که یك بسجمله‌ای $P(x)$ از درجهٔ ناکمتر از k برحسب $P_i(x)$ به‌ما داده شده است، یعنی ضرایبی مانند d_0، ...، d_k داریم به‌طوری که تساوی

$$p(x) = d_0P_0(x)+d_1P_1(x)+\cdots+d_kP_k(x) \qquad (۳۱.۶)$$

برقرار است. درمحاسبهٔ $p(x)$ در یك نقطهٔ خاص $\bar{x}$، می‌توانیم از رابطهٔ بازگشتی سه‌جمله‌ای (۳۰.۶) برای $P_i(x)$ به‌صورت زیر استفاده کنیم: به‌موجب رابطهٔ (۳۰.۶) داریم

$$P_k(\bar{x}) = A_{k-1}(\bar{x}-B_{k-1})P_{k-1}(\bar{x}) - C_{k-1}P_{k-2}(\bar{x})$$

و بنابراین

$$p(\bar{x}) = d_0P_0(\bar{x})+\cdots+d_{k-2}P_{k-2}(\bar{x})+(d_{k-2}-d_kC_{k-1})P_{k-2}(\bar{x})$$
$$+[d_{k-1}+d_kA_{k-1}(\bar{x}-B_{k-1})]P_{k-1}(\bar{x})$$

یا با استفاده از قرارداد اختصاری

$$\bar{d}_k = d_k \qquad \bar{d}_{k-1} = d_{k-1}+\bar{d}_kA_{k-1}(\bar{x}-B_{k-1})$$

داریم

---

1. Laguerre

$$p(\bar{x})=d_0P_0(\bar{x})+\cdots+d_{k-2}P_{k-2}(\bar{x})+(d_{k-2}-\bar{d}_kC_{k-1})P_{k-2}(\bar{x})$$
$$+\bar{d}_{k-1}P_{k-1}(\bar{x}) \qquad (۳۲.۶)$$

باز به‌موجب رابطهٔ (۳۰.۶)

$$P_{k-1}(\bar{x})=A_{k-2}(\bar{x}-B_{k-2})P_{k-2}(\bar{x})-C_{k-2}P_{k-3}(\bar{x})$$

و از قراردادن این مقدار در رابطهٔ (۳۲.۶)، خواهیم داشت

$$p(\bar{x})=d_0P_0(\bar{x})+\cdots+(d_{k-3}-\bar{d}_{k-1}C_{k-2})P_{k-3}(\bar{x})+\bar{d}_{k-2}P_{k-2}(\bar{x})$$

که دراینجا از قرارداد اختصاری زیر استفاده کرده‌ایم

$$\bar{d}_{k-2}=d_{k-2}+\bar{d}_{k-1}A_{k-2}(\bar{x}-B_{k-2})-\bar{d}_kC_{k-1}$$

اگر به‌همین صورت ادامه دهیم، متوالیاً روابط

$$\bar{d}_j=d_j+\bar{d}_{j+1}A_j(\bar{x}-B_j)-\bar{d}_{j+2}C_{j+1} \qquad j=k-2,\ldots,0$$

را محاسبه می‌کنیم که سرانجام به‌تساوی $p(\bar{x})=\bar{d}_0P_0(\bar{x})=\bar{d}_0\alpha_0$ خواهیم رسید.

الگوریتم ۱.۶ ضرب تودرتو برای بسجمله‌ایهای متعامد ضرایب

$$A_j, B_j, C_j \text{ ، } j=0,\ldots,k-1$$

برای رابطهٔ بازگشتی‌سه‌جمله‌ای (۳۰.۶)، که بسجمله‌ایهای متعامد $P_0(x)$، ...، $P_k(x)$ در آن صدق می‌کنند، و نیزمقدار ثابت $\alpha_0:=P_0(x)$ و ضرایب d_0، ...، d_k مربوط به $p(x)$ در رابطهٔ (۳۱.۶) و نقطهٔ $\bar{x}$ داده شده‌اند.

$$\bar{d}_k:=d_k$$

اگر $k=0$، آنگاه از الگوریتم **خارج** شوید

$$\bar{d}_{k-1}:=d_{k-1}+\bar{d}_kA_{k-1}(\bar{x}-B_{k-1})$$

اگر $k=1$، آنگاه از الگوریتم **خارج** شوید

For $j=k-2,k-3,\ldots,0$, do:
$\quad\bar{d}_j:=d_j+\bar{d}_{j+1}A_j(\bar{x}-B_j)-\bar{d}_{j+2}C_{j+1}$

هنگام **خروج** از الگوریتم، $p(\bar{x})$ به‌صورت زیرمعین می‌شود

$$p(\bar{x})=\bar{d}_0P_0(x)$$

به‌کاربردن **زبان فورترن** برای این الگوریتم با این اشکال‌کوچك همراه است که

برخی از گویشهای زبان فورترن اجازهٔ استفاده از زیرنمایهٔ صفر را به‌ما نمی‌دهد. و نیز مقدار حافظهٔ موردنیاز و تعداد محاسبات لازم از یك مجموعه از بسجمله‌ایهای متعامد تا مجموعهٔ دیگر تغییرمی‌کند.

□ **مثال ۱۰۰۶:** درحالتی‌که بسجمله‌ایهای متعامد، بسجمله‌ایهای چبیشف باشند، برنامه‌ای به‌زبان فورترن برای الگوریتم ۱۰۶، بنویسید.

در این حالت نیازی به‌نگهداری A_i، B_i، C_i در آرایه‌ها نیست، زیرا به i بستگی ندارند، و نیز محاسبهٔ $\bar{d}_j$ تنها به‌محاسبهٔ $\bar{d}_{j+1}$ و $\bar{d}_{j+2}$ نیازدارد، از این‌رو به‌نگهداری تمام آرایهٔ $\bar{d}_i$، $k, \ldots, 0 = i$، نیازی نیست.

تابع چبیشف به‌زبان فورترن که در زیر آمده‌است، مسئلهٔ داده شده را حل می‌کند. NTERMS عبارت از تعداد جملات در $p(x)$ است، یعنی $p(x)$ ازدرجهٔ نابیشتر از NTERMS − ۱ است. قرار براین‌است که هم NTERMS وهم‌ضرایب $D(i) = d_{i-1}$، $i = 1, \ldots, \text{NTERMS}$، با COMMON POLY نامگذاری شوند.

```
      REAL FUNCTION CHEB (X)
C  RETURNS THE VALUE OF THE POLYNOMIAL OF DEGREE .LT. NTERMS WHOSE
C  CHEBYSHEV COEFFICIENTS ARE CONTAINED IN  D .
      INTEGER NTERMS,   K
      REAL D,X,   PREV,PREV2,TWOX
      COMMON /POLY/ NTERMS,D(30)
      IF ( NTERMS .EQ. 1) THEN
         CHEB = D(1)
                                          RETURN
      END IF
      TWOX = 2.*X
      PREV2 = 0.
      PREV = D(NTERMS)
      IF (NTERMS .GT. 2) THEN
         DO 10 K=NTERMS-1,2,-1
            CHEB = D(K) + TWOX*PREV - PREV2
            PREV2 = PREV
            PREV = CHEB
   10    CONTINUE
      END IF
      CHEB = D(1) + X*PREV - PREV2
                                          RETURN
      END
```

□

تمرین

۱-۳۰۶ با استفاده از رابطهٔ بازگشتی مناسب، پنج نخستین بسجمله‌ای لاگر را (به‌ازای $\alpha = 0$) بسازید

۲-۳۰۶ ریشه‌های بسجمله‌ایهای لژاندر $P_2(x)$، $P_3(x)$، و $P_4(x)$، را بیابید.

۳-۳۰۶ ریشه‌های بسجمله‌ایهای هرمیت $H_2(x)$ و $H_3(x)$ و $H_4(x)$ را به‌دست آورید.

۴-۳۰۶ بسجمله‌ای $p(x) = x^4 + 2x^3 + x^2 + 2x + 1$ را به‌صورت حاصلجمعی از بسجمله‌ایهای لژاندر بیان کنید.

۵-۳۰۶ مستقیماً تحقیق کنید که بسجمله‌ای لژاندر $P_3(x)$ برهر بسجمله‌ای ازدرجهٔ ۲، متعامداست.

۶-۳۰۶ ثابت کنید که اگر $P_k(x)$ بسجمله‌ای لژاندری از درجهٔ k باشد، آنگاه

$$\int_{-1}^{1} [P_k(x)]^2\,dx = \frac{2}{2k+1}$$

از رابطهٔ بازگشتی سه‌جمله‌ای که بسجمله‌ایهای لژاندر در آن صدق می‌کنند استفاده کنید.

۷-۳۰۶ گیریم $P_0(x)$، $P_1(x)$، ... یك دنباله از بسجمله‌ایهای متعامد باشند و x_0، ...، x_k، $k+1$ ریشهٔ متمایز از $P_{k+1}(x)$. ثابت کنید که به‌ازای این نقاط، بسجمله‌ایهای لژاندر $l_i(x)=\Pi_{j\neq i}(x-x_j)/(x_i-x_j)$، $i=0,\ldots,k$ بر یکدیگر عمودند. [راهنمایی: نشان دهید که به‌ازای $i\neq j$، رابطهٔ $l_i(x)l_j(x)=P_{k+1}(x)g(x)$، که در آن $g(x)$ یك بسجمله‌ای است از درجهٔ نابیشتر از k، برقرار است.]

۴.۶ تقریب با روش کوچکترین مربعات به‌وسیلهٔ بسجمله‌ایها

در این بخش استفاده از دنباله‌های بسجمله‌ایهای متعامد برای محاسبهٔ تقریبهایی بسجمله‌ایهای (وزین) از راه کوچکترین مربعات مورد بحث قرار می‌گیرد.

گیریم $f(x)$ تابعی تعریف شده بر بازهٔ (a,b) باشد، و فرض کنید که بخواهیم $f(x)$ را بر بازهٔ (a,b) با یك بسجمله‌ای از درجهٔ نابیشتر از k تقریب کنیم. اگر اختلاف بین $f(x)$ و $p(x)$ به‌وسیلهٔ

$$\langle f(x)-p(x),\, f(x)-p(x)\rangle = \begin{cases} \displaystyle\int_a^b [f(x)-p(x)]^2 w(x)\,dx \\ \qquad\qquad \text{یا} \\ \displaystyle\sum_{n=1}^{N} [f(x_n)-p(x_n)]^2 w(x_n) \end{cases} \tag{۳۳.۶}$$

که در آن حاصلضرب داخلی توسط رابطهٔ (۲۶.۶)، (۲۷.۶) داده شده، اندازه‌گیری شود، آنگاه طبیعی است که باید درجستجوی یك بسجمله‌ای از درجهٔ نابیشتر از k باشیم که به‌ازای آن (۳۳.۶) تاحد ممکن کوچك باشد. این بسجمله‌ای را **تقریب $f(x)$ به‌وسیلهٔ بسجمله‌ایهایی از درجهٔ نابیشتر از k، با روش کوچکترین مربعات (وزین) نامند.**

مسئلهٔ یافتن یك چنین بسجمله‌ای را در بخش ۲.۶ برای حالت خاصی که حاصلضرب داخلی با رابطهٔ (۲۷.۶)، و تابع وزن با $w(x)\equiv 1$ داده شده بود، حل کردیم. در حالت کلی، مراحل انجام کار به‌شرح زیر است: فرض کنید که برای حاصلضرب داخلی، بتوانیم یك دنباله از بسجمله‌ایهای متعامد $P_0(x)$، ...، $P_k(x)$ به‌دست آوریم. به‌موجب ویژگی ۱ مربوط به‌چنین دنباله‌هایی (قسمت ۳.۶ را ببینید)، به‌ازای ضرایب مناسب d_0، ...، d_k

هر بسجمله‌ای $p(x)$ از درجهٔ نابیشتر از k می‌تواند به‌شکل

$$p(x)=d_0P_0(x)+\cdots+d_kP_k(x)$$

نوشته شود. از قراردادن این $p(x)$ در (۳۳.۶)، نتیجه می‌گیریم که می‌خواهیم به‌ازای تمام انتخابهای ممکنه از d_0، ...، d_k

$$E(d_0,\ldots,d_k)=\langle f(x)-d_0P_0(x)-\cdots-d_kP_k(x),$$
$$f(x)-d_0P_0(x)-\cdots-d_kP_k(x)\rangle$$

به‌حداقل برسد. اگر مانند بخش ۲.۶ عمل کنیم، نشان داده می‌شود که «بهترین» ضرایب d_0^*، ...، d_k^* باید در **معادلات نرمال**

$$d_0^*\langle P_0,P_i\rangle+d_1^*\langle P_1,P_i\rangle+\cdots+d_k^*\langle P_k,P_i\rangle=\langle f,P_i\rangle,\ i=0,\ldots,k$$

که به‌علت تعامد $p_j(x)$ها، به

$$d_i^*\langle P_i,P_i\rangle=\langle f,P_i\rangle \qquad i=0,\ldots,k$$

بدل می‌شوند صدق کنند. از این‌رو، اگر

$$S_i=\langle P_i,P_i\rangle \qquad i=0,\ldots,k$$

همه غیرصفر باشند، آنگاه بهترین ضرایب اصلا به‌صورت زیر معین می‌شوند

$$d_i^*=\frac{\langle f,P_i\rangle}{S_i} \qquad i=0,\ldots,k \qquad (۳۴.۶)$$

□ **مثال ۱۱.۶:** از مجموعهٔ تمام بسجمله‌ایهای $p(x)$ از درجهٔ نابیشتر از سه آن یک‌را که انتگرال

$$\int_{-1}^{1}[e^x-p(x)]^2\,dx$$

را مینیمم می‌سازد محاسبه کنید.

در این حالت داریم $f(x)=e^x$ و حاصلضرب داخلی با

$$\langle g,h\rangle=\int_{-1}^{1}g(x)h(x)\,dx$$

معین می‌شود. از مثال ۶.۶ چنین برمی‌آید که بسجمله‌ایهای متعامد برای این حاصلضرب داخلی، بسجمله‌ایهای لژاندر هستند. با استفاده از جدول ۲.۶ مربوط به‌این بسجمله‌ایها،

محاسبات زیر را انجام می‌دهیم

$$\langle f, P_0\rangle = \int_{-1}^{1} e^x\,dx = e - \frac{1}{e}$$

$$\langle f, P_1\rangle = \int_{-1}^{1} e^x x\,dx = \frac{2}{e}$$

$$\langle f, P_2\rangle = \frac{3}{2}\int_{-1}^{1} e^x\left(x^2 - \frac{1}{3}\right)dx = e - \frac{7}{e}$$

$$\langle f, P_3\rangle = \frac{5}{2}\int_{-1}^{1} e^x\left(x^3 - \frac{3}{5}x\right)dx = -5e + \frac{37}{e}$$

می‌توان نشان داد که برای بسجمله‌ایهای لژاندر (تمرین ۳۰.۶ـ۶ را ببینید)، داریم

$$S_i = \langle P_i, P_i\rangle = \frac{2}{2i+1} \qquad \text{به‌ازای جمیع مقادیر } i$$

لذا $S_0 = 2$، $S_1 = 2/3$، $S_2 = 2/5$، $S_3 = 2/7$. با استفاده از (۳۴.۶) برای محاسبهٔ d_i^* و استفاده از $e = 2{,}718281828$، ملاحظه می‌کنیم که تقریب با روش کوچکترین توانهای دوم برای e^x با بسجمله‌ایهای مکعبی بر بازهٔ $(1, -1)$ برابراست با

$$p^*(x) = 1{,}175201194P_0(x) + 1{,}103638324P_1(x)$$
$$+ 0{,}357814435\,06P_2(x) + 0{,}070455633\,67P_3(x)$$

اگر به‌جای $P_i(x)$ها مقادیرهم ارز آنها را برحسب توانهای x بگذاریم وازجدول ۲.۶ استفاده و حاصل را مرتب کنیم، خواهیم داشت

$$p^*(x) = 0{,}999629401\,83 + 0{,}997954873\,0x$$
$$+ 0{,}536721526\,0x^2 + 0{,}176139084\,2x^3$$

ماکسیمم انحراف این بسجمله‌ایها از e^x بر بازهٔ $(1, -1)$ درحدود ۰٫۰۱۱ است. □

اگر نتوان بسجمله‌ایهای متعامد مناسبی درجدولها به‌دست آورد، باید آنها را پدید آورد. این‌کار به‌کمك رابطهٔ بازگشتی سه‌جمله‌ای (۳۰.۶) می‌توانند انجام گیرد. اکنون یك بیان الگوریتمی از این تکنیك برای حالت عملا مهمی که درآن حاصلضرب داخلی برابراست با

$$\langle g, h\rangle = \sum_{n=1}^{N} g(x_n)h(x_n)w(x_n) \qquad (۳۵.۶)$$

می‌دهیم، در اینجا $x_1, \ldots, x_N$ نقاط ثابتی بر بازهٔ (a, b) هستند.

الگوریتم ۲۰۶: تولید بسجمله‌ایهای متعامد، برای سادگی، به‌دست‌آوردن بسجمله‌ایهای متعامدی را درنظر می‌گیریم که ضریب جملهٔ پیشرو در آنها ۱ باشد. لذا

$$A_i = \alpha_i = 1 \qquad \text{به‌ازای جمیع مقادیر } i$$

مرحلهٔ ۰ قراردهید $P_0(x) \equiv 1$. بعلاوه عبارت زیر را محاسبه کنید

$$S_0 = \langle P_0, P_0 \rangle = \sum_{n=1}^{N} w(x_n)$$

اگر $N \geqslant 1$ و $w(x) > 0$، آنگاه S_0 برابر با صفر نیست و می‌توان محاسبهٔ

$$P_1(x) = (x - B_0)P_0(x) = x - B_0$$

را ادامه داد که بنابرویژگی ۴، مربوط به‌بسجمله‌ایهای متعامد (بخش ۳.۶ را ببینید)، داریم

$$B_0 = \frac{\langle xP_0(x), P_0(x) \rangle}{S_0} = \sum_{n=1}^{N} \frac{x_n w(x_n)}{S_0}$$

وقتی که $P_0(x), \ldots, P_j(x)$ راحساب شده تلقی کنیم، مرحلهٔ عمومی یا مرحلهٔ jام به‌گونهٔ زیر انجام می‌گیرد:

مرحلهٔ jام عبارت

$$S_j = \langle P_j, P_j \rangle = \sum_{n=1}^{N} [P_j(x_n)]^2 w(x_n)$$

رامحاسبه می‌کنیم. چون $P_j(x)$ یك بسجمله‌ای است دقیقاً از درجهٔ j، S_j تنها زمانی می‌تواند صفر باشد که بیش از j نقطه از نقاط $x_1, \ldots, x_N$ متمایز نباشند. از این‌رو اگر بیش از j نقطهٔ متمایز بین x_nها وجود داشته باشد، می‌توانیم

$$B_j = \frac{\langle xP_j(x), P_j(x) \rangle}{S_j} = \sum_{n=1}^{N} x_n [P_j(x_n)]^2 \frac{w(x_n)}{S_j}$$

و

$$C_j = \frac{S_j}{S_{j-1}}$$

را محاسبه کنیم و بسجمله‌ای متعامد بعدی را به‌صورت زیر به‌دست آوریم

$$P_{j+1}(x) = (x - B_j)P_j(x) - C_j P_{j-1}(x) \qquad (۳۶.۶)$$

□ **مثال ۱۲.۶:** با استفاده از بسجمله‌ایهای متعامد، مسئلهٔ تقریب‌زدن بـا روش کوچکترین مربعات درمثال ۵.۶ حل کنید.

برای این مثال داریم $f(x)=۱۰-۲x+x^۲/۱۰$

$$x_n=۱۰+\frac{n-۱}{۵}\quad f_n=f(x_n)\qquad n=۱,\dots,۶$$

و درپی آن بسجمله‌ای از درجهٔ نابیشتر از ۲ هستیم که مقدار

$$\sum_{n=۱}^{۶}[f_n-P(x_n)]^۲$$

را مینیمم کند، یعنی حاصلضرب داخلی (۲۷.۶) را بـه‌ازای $w(x)\equiv ۱$ باید مورد بحث قراردهیم. با متابعت از الگوریتم ۲.۶، مقادیر زیر را محاسبه می‌کنیم

$$P_۰(x)\equiv ۱\qquad \text{از این‌رو}\qquad S_۰=\sum_{n=۱}^{۶}۱=۶$$

$$B_۰=\sum_{n=۱}^{۶}\left[۱۰+\frac{n-۱}{۵}\right]\Big/S_۰=\frac{۶۳}{۶}=۱۰\text{ر}۵$$

بنابراین

$$P_۱(x)=x-۱۰\text{ر}۵\qquad S_۱=\sum_{n=۱}^{۶}\left[\frac{n-۱}{۵}-۰\text{ر}۵\right]^۲=۰\text{ر}۷$$

و چون $S_۱\neq ۰$، به‌محاسبهٔ $P_۲(x)$ می‌پردازیم. اگر ازهفت رقم اعشاری استفاده ونتیجه راگردکنیم، خواهیم داشت

$$B_۱=\sum_{n=۱}^{۶}\frac{[۱۰+(n-۱)/۵][(n-۱)/۵-۰\text{ر}۵]^۲}{S_۱}=\frac{۷\text{ر}۳۵}{۰\text{ر}۷}=۱۰\text{ر}۵$$

$$C_۱=\frac{S_۱}{S_۰}=\frac{۰\text{ر}۷}{۶}=۰\text{ر}۱۱۶۶۶۶۷$$

از این نتایج چنین به‌دست خواهیم آورد

$$P_۲(x)=(x-۱۰\text{ر}۵)^۲-۰\text{ر}۱۱۶۶۶۶۷\qquad S_۲=۰\text{ر}۰۵۹۷۳۳۳۲$$

درمرحلهٔ بعد با استفاده از رابطهٔ (۳۴.۶) وحساب ممیزشناور با هفت رقم اعشار، بهترین ضرایب $d_۰^*$، $d_۱^*$ و $d_۲^*$ را برای تقریب بـا روش کوچکترین مربعات محاسبه مـی کنیم، که نتایج این محاسبات چنین‌اند $p^*(x)=d_۰^*p_۰(x)+d_۱^*p_۱(x)+d_۲^*p_۲(x)$

$$d_0^* = \sum_{n=1}^{6} \frac{f_n}{6} = ۰\text{ر}۰۳۶۶۶۶۶۷$$

$$d_1^* = \sum_{n=1}^{6} \frac{f_n P_1(x_n)}{۰\text{ر}۷} = ۰\text{ر}۱$$

$$d_2^* = \sum_{n=1}^{6} \frac{f_n P_2(x_n)}{۰\text{ر}۰۵۹۷۳۳۳۲} = ۰\text{ر}۰۹۹۹۹۹۹ \qquad \square$$

برای مقایسهٔ این نتایج با آنچه‌که درمثال ۵.۶ محاسبه شد، $p^*(x)$ را برحسب x^2، x، ۱ می‌نویسیم و خواهیم داشت

$$p^*(x) = ۰\text{ر}۰۳۶۶۶۶۶۷ + ۰\text{ر}۱(x - ۱۰\text{ر}۵)$$
$$+ ۰\text{ر}۰۹۹۹۹۹۹[(x - ۱۰\text{ر}۵)^2 - ۰\text{ر}۱۱۶۶۶۶۷]$$
$$= ۰\text{ر}۰۳۶۶۶۶۶۷ - ۱\text{ر}۰۵ + ۰\text{ر}۰۹۹۹۹۹۹(۱۱۰\text{ر}۲۵ - ۰\text{ر}۱۱۶۶۶۶۷)$$
$$+ [۰\text{ر}۱ + ۰\text{ر}۰۹۹۹۹۹۹(-۲۱)]x + ۰\text{ر}۰۹۹۹۹۹۹x^2$$

لذا با این روش محاسبه، مقادیر c_i^*های مثال ۵.۶ برابرخواهند بود با

$$c_1^* = ۹\text{ر}۹۹۹۹۹۸\ldots \qquad c_2^* = -۱\text{ر}۹۹۹۹۹۹۸\ldots$$

$$c_3^* = ۰\text{ر}۰۹۹۹۹۹۹\ldots$$

برعکس، درمثال۵.۶، وقتی معادلات نرمال (۲۳.۶) را برحسب c_i^*ها با استفاده ازحساب ممیزشناور با هفت رقم اعشار به‌طور مستقیم حل‌کردیم، مقادیر زیربه‌دست آمدند

$$c_1^* = ۸\text{ر}۲۹۲۰۰۰ \qquad c_2^* = -۱\text{ر}۷۱۲۰۰۰ \qquad c_3^* = ۰\text{ر}۰۸۶۳۰۰۰$$

بنابراین، بـا استفاده از بسجمله‌ایهای متعامد، نتایج بهبـود مؤثری را در این مثال نشان می‌دهند.

درضمن، معمولا زحمت محاسبهٔ $p^*(x)$ برحسب x را، به‌خود نمی‌دهند، بلکـه هر زمان‌که $p^*(x)$ باید محاسبه شود، الگوریتم ۱.۶ را، همراه با d_i^*، که محاسبه شده، به‌کار می‌برند، زیراکه ضرایب B_i و C_iی رابطهٔ بازگشتی دردست هستند.

تولید بسجمله‌ایهای متعامد ومحاسبهٔ بهترین ضرایب d_i^*، برای صرفه‌جویی‌درحافظه، بااجرا به‌زبان **فورترن** بهتر ترکیب و به‌یك عمل بدل می‌شود. برای محاسبهٔ d_j^* و$P_{j+1}(x)$، تنها اعداد

$$n = ۱, \ldots, N \qquad P_{j-1}(x_n), \qquad P_j(x_n)$$

موردنیازند. بنابراین، اگربه‌محض‌دردسترس‌بودن $P_j(x_n)$، $n = ۱, \ldots, N$، d_j^* محاسبه

شود آنگاه، همین که $P_{j+1}(x)$ و $P_{j+2}(x)$ محاسبه شدند $P_j(x_n)$ها، $n=1,\ldots,N$، می‌توانند بدون ایجادهیچ گونه مسئله‌ای نادیده گرفته شوند. باز احتیاجی نیست که $P_j(x)$ رامثلا، برحسب توانهایی ازxبسازیم، زیراکه، فقط بمقادیر آنها در x_n، $n=1,\ldots,N$ نیازداریم.

```
      SUBROUTINE ORTPOL ( X, F, W, NPOINT, PJM1, PJ, ERROR )
C  CONSTRUCTS THE DISCRETE WEIGHTED LEAST SQUARES APPROXIMATION BY POLY-
C  NOMIALS OF DEGREE .LT. NTERMS  TO GIVEN DATA.
C******  I N P U T  ******
C  (X(I), F(I)), I=1,...,NPOINT  GIVES THE ABSCISSAE AND ORDINATES OF
C               THE GIVEN DATA POINTS TO BE FITTED.
C  W  NPOINT-VECTOR CONTAINING THE POSITIVE WEIGHTS TO BE USED.
C  NPOINT  NUMBER OF DATA POINTS.
C******  I N P U T  VIA COMMON BLOCK  P O L Y  ******
C  NTERMS  GIVES THE ORDER (= DEGREE + 1) OF THE POLYNOMIAL APPROXIMANT.
C******  W O R K  A R E A S  ******
C  PJM1, PJ  ARRAYS OF LENGTH  NPOINT  TO CONTAIN THE VALUES AT THE X'S
C               OF THE TWO MOST RECENT ORTHOGONAL POLYNOMIALS.
C******  O U T P U T  ******
C  ERROR  NPOINT-VECTOR CONTAINING THE ERROR AT THE X'S OF THE POLYNOM-
C       IAL APPROXIMANT TO THE GIVEN DATA.
C******  O U T P U T  VIA COMMON BLOCK  P O L Y  ******
C  B, C  ARRAYS CONTAINING THE COEFFICIENTS FOR THE THREE-TERM RECUR-
C       RENCE WHICH GENERATES THE ORTHOGONAL POLYNOMIALS.
C  D  COEFFICIENTS OF THE POLYNOMIAL APPROXIMANT TO THE GIVEN DATA WITH
C       RESPECT TO THE SEQUENCE OF ORTHOGONAL POLYNOMIALS.
C       THE VALUE OF THE APPROXIMANT AT A POINT  Y  MAY BE OBTAINED BY A
C       REFERENCE TO  ORTVAL(Y) .
C******  M E T H O D  ******
C  THE SEQUENCE  P0, P1, ..., PNTERMS-1  OF ORTHOGONAL POLYNOMIALS WITH
C  RESPECT TO THE DISCRETE INNER PRODUCT
C        (P,Q)  =  SUM ( P(X(I))*Q(X(I))*W(I) , I=1,...,NPOINT)
C  IS GENERATED IN TERMS OF THEIR THREE-TERM RECURRENCE
C        PJP1(X) = (X - B(J+1))*PJ(X) - C(J+1)*PJM1(X) ,
C  AND THE COEFFICIENT  D(J)  OF THE WEIGHTED LEAST SQUARES APPROXIMAT-
C  ION TO THE GIVEN DATA IS OBTAINED CONCURRENTLY AS
C        D(J+1) = (F,PJ)/(PJ,PJ) , J=0,...,NTERMS-1 .

C  ACTUALLY, IN ORDER TO REDUCE CANCELLATION,  (F,PJ)  IS CALCULATED AS
C  (ERROR,PJ),  WITH  ERROR = F  INITIALLY, AND, FOR EACH  J , ERROR RE-
C  DUCED BY  D(J+1)*PJ  AS SOON AS  D(J+1)  BECOMES AVAILABLE.
C
      INTEGER NPOINT,NTERMS,   I,J
      REAL B,C,D,ERROR(NPOINT),F(NPOINT),PJ(NPOINT),PJM1(NPOINT),
     *     W(NPOINT),X(NPOINT),   P,S(20)
      COMMON /POLY/ NTERMS,B(20),C(20),D(20)
C
      DO 9 J=1,NTERMS
         B(J) = 0.
         D(J) = 0.
    9    S(J) = 0.
      C(1) = 0.
      DO 10 I=1,NPOINT
         D(1) = D(1) + F(I)*W(I)
         B(1) = B(1) + X(I)*W(I)
   10    S(1) = S(1) + W(I)
      D(1) = D(1)/S(1)
      DO 11 I=1,NPOINT
   11    ERROR(I) = F(I) - D(1)
      IF (NTERMS .EQ. 1)                    RETURN
      B(1) = B(1)/S(1)
      DO 12 I=1,NPOINT
         PJM1(I) = 1.
   12    PJ(I) = X(I) - B(1)
C
      DO 30 J=2,NTERMS
         DO 21 I=1,NPOINT
            P = PJ(I)*W(I)
            D(J) = D(J) + ERROR(I)*P
            P = P*PJ(I)
            B(J) = B(J) + X(I)*P
```

```
21        S(J) = S(J) + P
          D(J) = D(J)/S(J)
          DO 22 I=1,NPOINT
22           ERROR(I) = ERROR(I) - D(J)*PJ(I)
          IF (J .EQ. NTERMS)                    RETURN
          B(J) = B(J)/S(J)
          C(J) = S(J)/S(J-1)
          DO 27 I=1,NPOINT
             P = PJ(I)
             PJ(I) = (X(I) - B(J))*PJ(I) - C(J)*PJM1(I)
27           PJM1(I) = P
30 CONTINUE
                                                RETURN
   END
```

محاسبهٔ $D(j)$، به‌گونه‌ای که در این زیربرنامه انجام گرفت، شاید نیاز به‌مقداری توضیح داشته باشد. چون $D(j)=d^*_{j-1}$، از رابطهٔ (۳۴.۶) نتیجه می‌گیریم

$$D(j)=\sum_{n=1}^{\text{NPOINT}}\frac{f_nP_{j-1}(x_n)w(x_n)}{S_{j-1}} \qquad (۳۷.۶)$$

درحالی‌که دربرنامه، $D(j)$ به‌صورت زیرمحاسبه شده‌است

$$D(j)=\sum_{n=1}^{\text{NPOINT}}\frac{\text{ERROR}(n)P_{j-1}(x_n)w(x_n)}{S_{j-1}} \qquad (۳۸.۶)$$

که در آن

به‌ازای جمیع مقادیر n

$$\text{ERROR}(n)=f_n-\text{D}(1)P_0(x_n)-\cdots-\text{D}(j-1)P_{j-2}(x_n) \qquad (۳۹.۶)$$

اگر(۳۹.۶) را در (۳۸.۶) قراردهیم، چون P_{j-1} بر $P_0(x)$، ...، $P_{j-2}(x)$ متعامد است خواهیم داشت

$$\begin{aligned}
&\text{D}(j)\\
&=\sum_{n=1}^{\text{NPOINT}}\frac{[f_n-\text{D}(1)P_0(x_n)-\cdots-\text{D}(j-1)P_{j-2}(x_n)]P_{j-1}(x_n)w(x_n)}{S_{j-1}}\\
&=\frac{\sum_n f_nP_{j-1}(x_n)w(x_n)-\text{D}(1)\langle P_0,P_{j-1}\rangle-\cdots-\text{D}(j-1)\langle P_{j-2},P_{j-1}\rangle}{S_{j-1}}\\
&=\sum_n\frac{f_nP_{j-1}(x_n)w(x_n)}{S_{j-1}}
\end{aligned}$$

از این‌رو درحساب دقیق یا حساب بـا دقت نامتناهی[1]، هردو رابطهٔ (۳۷.۶) و (۳۸.۶) مقادیر واحدی برای $\text{D}(j)$ به‌دست می‌دهند. اما درحساب با دقت متناهی، می‌توان انتظار داشت‌که رابطهٔ (۳۸.۶) به‌دلیل زیر دقیقتر باشد: چون

1. infinite-precision

$$p_r^*(x)=\mathrm{D}(1)P_0(x)+\cdots+\mathrm{D}(r+1)P_r(x)$$

تقریب با روش کوچکترین توانهای دوم (وزین) برای $f(x)$ با بسجمله‌ایهای از درجهٔ نابیشتر از r است، نتیجهٔ آن که می‌توان انتظار داشت اعداد

$$\mathrm{ERROR}(n)=f_n-P_{j-1}^*(x_n) \qquad n=1,\cdots,\mathrm{NPOINT}$$

از اعداد f_n، $n=1,\cdots,\mathrm{NPOINT}$، کوچکتر باشند. از این‌رو، به‌نظر می‌رسد که محاسبهٔ رابطهٔ(۳۸.۶) درمقایسه بامحاسبهٔ رابطهٔ(۳۷.۶)، به‌علت تفریق مقادیر تقریباً مساوی از یکدیگر، احتمالاً موجب از دست رفتن ارقام با معنا می‌شود (تمرین ۴.۶–۱ را ببینید)

□ **مثال ۱۳.۶:** مقادیر f_n از تابع $f(x)=e^x$ در

$$x_n=(n-1)/10-1,(n=1,\cdots,21)$$

وگرد شده تا دو رقم بعد از ممیز، داده شده‌اند. سعی کنید که اطلاعات مربوط به $f(x)$ راکه در این داده‌ها گنجانیده شده‌اند بازیابی کنید:

ما سعی می‌کنیم این مسئله را با محاسبهٔ آن بسجمله‌ای $p_3^*(x)$ از مجموعهٔ تمام بسجمله‌ایهای $p_3(x)$ ازدرجهٔ نابیشتر از۳ که عبارت

$$\sum_{n=1}^{21}[f_n-p_3(x_n)]^2$$

را مینیمم می‌سازد، حل کنیم. برنامهٔ **فورترن** زیر به‌کمك زیربرنامهٔ ORTPOL، که قبلا اشاره‌شد، $p_3^*(x)$ را برآورد می‌کند وسپس بااستفاده از FUNCTION ORTVAL، که اساس آن الگوریتم ۱.۶ است، $p_3^*(x)$ را در x_n محاسبه می‌کند.

```
C  PROGRAM FOR EXAMPLE 6.13 .
      PARAMETER NPMAX=100
      INTEGER NTERMS,   I,J,NPOINT
      REAL B,C,D,ERROR(NPMAX),F(NPMAX),PJ(NPMAX),PJM1(NPMAX),W(NPMAX)
     *     ,X(NPMAX)
      COMMON /POLY/ NTERMS,B(20),C(20),D(20)
      NPOINT = 21
      DO 1 I=1,NPOINT
         W(I) = 1.
         X(I) = -1. + FLOAT(I-1)/10.
    1    F(I) = FLOAT(IFIX(EXP(X(I))*100. + .5))/100.
      NTERMS = 4
      CALL ORTPOL( X, F, W, NPOINT, PJM1, PJ, ERROR )
      PRINT 601, (J,B(J),C(J),D(J),J=1,NTERMS)
  601 FORMAT(I2,3E16.8)
      DO 60 I=1,NPOINT
         PJM1(I) = EXP(X(I))
   60    PJ(I) = ORTVAL(X(I))
      PRINT 660, (X(I),F(I),PJ(I),ERROR(I),PJM1(I),I=1,NPOINT)
  660 FORMAT(F5.1,F8.3,F10.5,E13.3,F10.5)
                                          STOP
      END


      REAL FUNCTION ORTVAL (X)
C  RETURNS THE VALUE AT  X  OF THE POLYNOMIAL OF DEGREE .LT. NTERMS
C  GIVEN BY
C           D(1)*P0(X) + D(2)*P1(X) + ... + D(NTERMS)*PNTERMS-1(X),
C  WITH THE SEQUENCE  P0, P1, ...  OF ORTHOGONAL POLYNOMIALS GENERATED
C  BY THE THREE-TERM RECURRENCE
C     PJP1(X) = (X - B(J+1))*PJ(X) - C(J+1)*PJM1(X) , ALL J .
C
```

```
      COMMON /POLY/ NTERMS,B(20),C(20),D(20)
      PREV = 0.
      ORTVAL = D(NTERMS)
      IF (NTERMS .EQ. 1)                    RETURN
      DO 10 K=NTERMS-1,1,-1
         PREV2 = PREV
         PREV = ORTVAL
         ORTVAL = D(K) + (X - B(K))*PREV - C(K+1)*PREV2
   10 CONTINUE
                                            RETURN
      END
```

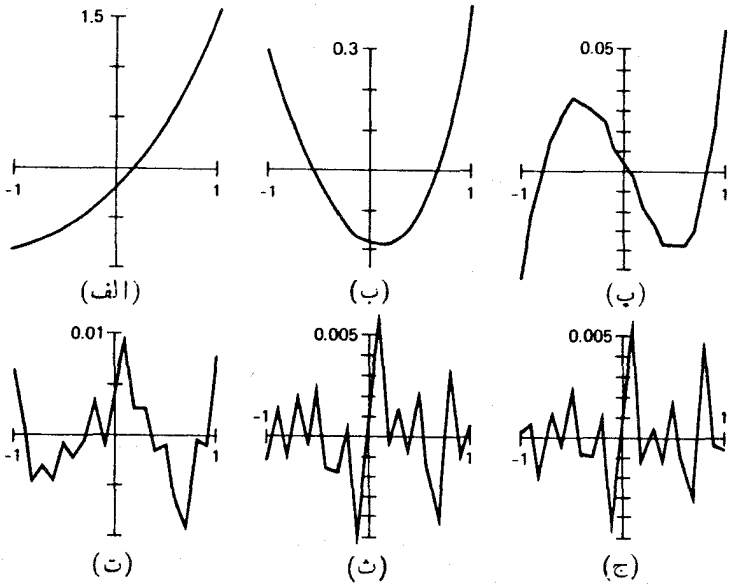
جدول ۴.۶ نتایج کامپیوتری برای مثال ۱۳.۶

x_n	f_n	$p_3^*(x_n)$	$f_n - p_3^*(x_n)$	$p_4^*(x_n)$	$f_n - p_4^*(x_n)$	e^{x_n}
−1.0	0.370	0.36387	6.130E − 03	0.37115	−1.154E − 03	0.36788
−0.9	0.410	0.40874	1.263E − 03	0.40874	1.263E − 03	0.40657
−0.8	0.450	0.45481	−4.806E − 03	0.45097	−9.719E − 04	0.44933
−0.7	0.500	0.50315	−3.148E − 03	0.49804	1.964E − 03	0.49659
−0.6	0.550	0.55484	−4.836E − 03	0.55021	−2.134E − 04	0.54881
−0.5	0.610	0.61094	−9.436E − 04	0.60789	2.108E − 03	0.60653
−0.4	0.670	0.67524	−2.542E − 03	0.67156	−1.565E − 03	0.67032
−0.3	0.740	0.74070	−7.045E − 04	0.74183	−1.832E − 03	0.74082
−0.2	0.820	0.81650	3.497E − 03	0.81940	6.029E − 04	0.81873
−0.1	0.900	0.90101	−1.010E − 03	0.90507	−5.070E − 03	0.90484
0.0	1.000	0.99530	4.710E − 03	0.99976	2.358E − 04	1.00000
0.1	1.110	1.10044	9.558E − 03	1.10450	5.499E − 03	1.10517
0.2	1.220	1.21751	2.490E − 03	1.22040	−4.045E − 04	1.22140
0.3	1.350	1.34758	2.422E − 03	1.34871	1.294E − 03	1.34986
0.4	1.490	1.49172	−1.717E − 03	1.49074	−7.399E − 04	1.49182
0.5	1.650	1.65100	−1.000E − 03	1.64795	2.052E − 03	1.64872
0.6	1.820	1.82650	−6.499E − 03	1.82188	−1.876E − 03	1.82212
0.7	2.010	2.01929	−9.287E − 03	2.01418	−4.176E − 03	2.01375
0.8	2.230	2.23044	−4.368E − 04	2.22660	0.397E − 03	2.22554
0.9	2.460	2.46102	−1.020E − 03	2.46102	−1.020E − 03	2.45960
1.0	2.720	2.71211	7.890E − 03	2.71939	6.061E − 04	2.71828

جدول ۴.۶، نتایج محاسباتی راکه روی کامپیوتر CDC ۶۵۰۰ انجام گرفته به‌دست می‌دهد. خطای $f_n - p_3^*(x_n)$ درشکل ۶.۶ ت، رسم شده‌است و نشان می‌دهد خطا تاحدی به‌صورت منظم عمل می‌کندکه این خودگویای آن‌است‌که $p_3^*(x)$ تمام اطلاعات گنجانده شده در داده‌های مفروض را نشان نمی‌دهد. از این‌رو تقریب با روش کوچکترین مربعات $p_4^*(x)$ برای داده‌های مفروض، به‌وسیلهٔ بسجمله‌ایهایی ازدرجـهٔ نابیشتر از ۴ نیز محاسبه شده‌اند. این نتایج نیز درجدول ۴.۶ داده شده‌اند. خطای $f_n - p_4^*(x)$ در شکل ۶.۶ ث، رسم شده‌است و دیده می‌شودکه خطا کامـلا نامنظم عمل می‌کند. از این‌رو می‌توان فرض کردکه $p_4^*(x)$ تمام اطلاعات گنجانـده شده در داده‌های مفروض f_n را نشان می‌دهـد. افزایش‌هرچه بیشتردرجهٔ بسجمله‌ای تقریب کننده تنهاموجب می‌شودکه به‌تابع تقریب کننده، آزادی بیشتری برای تقریب زدن نوفه در داده‌ها نیز بدهد. □

تمرین

۱-۴.۶ اگر $f(x) = 6000 + x$، آنگاه هر تقریب برای $f(x)$ از راه کوچکترین مربعات به‌وسیلهٔ خطهای راست، خود $f(x)$ است. بسجمله‌ای



شکل۶۰۶ خطا در تقریب‌زدن با روش کوچکترین مربعات برای داده‌های مثال ۳.۶ به‌وسیلهٔ بسجمله‌ایهای درجهٔ (الف) صفر، (ب) یك، (پ) دو، (ت) سه، (ث) چهار، (ج) پنج.

$$p_1^*(x) = d_0^* + d_1^* x$$

که عبارت

$$\sum_{n=-2}^{2} [f(x) - p_1(n)]^2$$

را مینیمم می‌سازد حساب کنید. توجه کنید که ۱ و x متعامدند، لذا تنها باید d_0^* و d_1^* را محاسبه کنید. تفاضل بین (۳۷.۶) و (۳۸.۶) را با محاسبهٔ d_1^* به‌دوطریق و با استفاده از حساب با ممیزشناور با چهار رقم اعشار نشان دهید.

۲-۴۰۶ از مجموعهٔ تمام بسجمله‌ایهای $p(x)$ از درجهٔ نابیشتر از ۲، آن یك را که

$$\int_{-1}^{1} [\sin \pi x - p(x)]^2 \, dx$$

را مینیمم می‌سازد محاسبه کنید. از بسجمله‌یی لژاندر استفاده کنید و تمام محاسبات را با پنج رقم اعشار انجام دهید. (توجه: $\pi = ۳ر۱۴۱۵۹۳$.)

۳-۴۰۶ زیربرنامهٔ ORTPOL را روی کامپیوتر خود پیاده کنید و سپس از آن برای حل مسئلهٔ زیر استفاده کنید. ازجدولی ازمقادیر $f(x) = \sin \pi x$ در $x_n = (n-1)/۱۰ - ۱$

$(n=۱,\ldots,۲۱)$، مقادیر $f_n(x)=\sin \pi x_n$ را گرد شده تا سه رقم اعشاری به‌دست آورید. سپس از مجموعهٔ تمام بسجمله‌ایهای نابیشتر از درجهٔ ۴، $p_۴(x)$، آن بسجمله‌ای $p_۴^*(x)$ را که

$$\sum_{n=۱}^{۲۱} [f_n - p_۴(x_n)]^۲$$

را مینیمم می‌سازد به‌دست آورید.

۵.۶* تقریب به‌وسیلهٔ بسجمله‌ایهای مثلثاتی

بسیاری از پدیده‌های فیزیکی، مانند نور و صوت، مشخصهٔ دوره‌ای[1] دارند. این پدیده‌ها با توابع $f(x)$ که **دوره‌ای** هستند، یعنی توابعی که به‌ازای کلیهٔ مقادیر x و یك عدد ثابت τ، به‌نام **دورهٔ تناوب**[2]، در رابطهٔ

$$f(x+\tau)=f(x)$$

صدق می‌کنند، بیان می‌شوند. از آنجاکه فقط بسجمله‌ایهای دوره‌ای توابع ثابت هستند، لذا باید دسته‌ای دیگر از توابع را برای تقریب‌زدن مؤثر توابع دوره‌ای به‌کار گیریم و برای انجام این‌کار بسجمله‌ایهای مثلثاتی از همه مناسبترند.

بنابر تعریف، **یك بسجمله‌ای مثلثاتی ازمرتبهٔ** n هر تابعی است به‌شکل

$$p(x)=a_۰/۲+\sum_{j=۱}^{n} [a_j \cos jx + b_j \sin jx] \qquad (۴۰.۶)$$

که در آن $a_۰$، ...، a_n، و $b_۱$، ...، b_n ثابتهای حقیقی یاهمتافت باشند. این‌گونه بسجمله‌ای مثلثاتی، یك بسجمله‌ای ۲π-دوره‌ای[3] است. بنابراین وقتی که می‌خواهیم برای تابع τ-دوره‌ای $f(x)$، به‌ازای $\tau \neq ۲\pi$ تقریب پیدا کنیم، لازم‌است که اصلاحاتی انجام دهیم. درچنین مواردی قرار می‌گذاریم که تابع ۲π-دوره‌ای $g(x)=f(\tau x/(۲\pi))$ را مورد توجه قراردهیم. آنگاه پس از ساختن یك بسجمله‌ای مثلثاتی تقریبی $p(x)$ برای $g(x)$، از آن یك تابع تقریبی τ-دوره‌ای برای $f(x)$ به‌شکل $p(۲\pi x/\tau)$ به‌دست می‌آوریم. با درنظرگرفتن مطلب فوق، از این به‌بعد فرض می‌کنیم که تابع $f(x)$ که باید تقریب زده شود قبلا ۲π-دوره‌ای باشد.

چنانچه معلوم خواهد شد، غالباً مناسبتر است که بسجمله‌ایهای مثلثاتی ازمرتبهٔ n را به‌شکل همتافت هم‌ارز آنها، یعنی به‌صورت زیر بنویسیم

$$p(x)=\sum_{j=-n}^{n} c_j e^{ijx} \qquad (۴۱.۶)$$

---

1. periodic 2. period 3. 2π-periodic

در اینجا و در بقیهٔ این بخش و بخشهای بعدی علامت i معرف **واحد انگاری**[1]

$$i=\sqrt{-1}$$

است و ارتباط بین روابط (۴۰.۶) و (۴۱.۶) به‌وسیلهٔ **فرمول اویلر**

$$e^{ix}=\cos x+i\sin x \qquad (۴۲.۶)$$

برقرار می‌شود (که اثبات آن را می‌توان در تمرین ۷.۱–۹ پیدا کرد). با توجه به‌فرمول اویلر خواهیم داشت[با رعایت $\cos(-jx)=\cos jx$ و $\sin(-jx)=-\sin jx$] که

$$\sum_{j=-n}^{n} c_j e^{ijx}=\sum_{j=-n}^{n} c_j[\cos jx+i\sin jx]$$

$$=c_0+\sum_{j=1}^{n}[(c_j+c_{-j})\cos jx+i(c_j-c_{-j})\sin jx]$$

این رابطه نشان می‌دهد که رابطهٔ (۴۱.۶) به‌همان شکل رابطهٔ (۴۰.۶) است، با فرض

$$a_j=c_j+c_{-j} \qquad b_j=i(c_j-c_{-j}) \qquad j=0,\dots,n \qquad (۴۳.۶ \text{ الف})$$

این رابطه را به‌آسانی می‌توان عکس کرد و رابطهٔ (۴۰.۶)، را که به‌شکل رابطهٔ (۴۱.۶) است به‌دست آورد، با شرایط

$$c_j=(a_j-ib_j)/2 \qquad c_{-j}=(a_j+ib_j)/2 \qquad j=0,\dots,n \qquad (۴۳.۶ \text{ ب})$$

باید توجه کرد که (۴۱.۶) معرف یك تابع حقیقی‌است اگر، وفقط اگر خود مزدوج همتافت خود باشد. اما از آنجا که داریم

$$\overline{\sum_{j=-n}^{n} c_j e^{ijx}}=\sum_{j=-n}^{n} \bar{c}_j e^{-ijx}=\sum_{j=-n}^{n} \bar{c}_{-j} e^{ijx}$$

این رابطه گویای این‌است که (۴۱.۶) یك تابع حقیقی‌است اگر، و فقط اگر، تساوی

$$c_j=\bar{c}_{-j} \qquad \text{به‌ازای همهٔ مقادیر } j \qquad (۴۴.۶)$$

برقرار باشد. بنابراین اگر (۴۰.۶) یا (۴۱.۶) تابع حقیقی باشد، آنگاه (۴۳.۶ الف) به‌شکل زیر ساده می‌شود

$$a_j=2\operatorname{Re} c_j \qquad b_j=-2\operatorname{Im} c_j \qquad (۴۵.۶)$$

تقریب‌زدن به‌وسیلهٔ بسجمله‌ایهای مثلثاتی اساساً توسط سری فوریه[2]

---

1. imaginary unit 2. Fourier series

$$f(x) \approx \sum_{j=-\infty}^{\infty} \hat{f}(j) e^{ijx} \qquad (۴۶.۶)$$

انجام می‌پذیرد که در آن **ضرایب فوریه** $\hat{f}(j)$ با انتگرال زیر محاسبه می‌شوند

$$\hat{f}(j) = \frac{۱}{۲\pi} \int_{۰}^{۲\pi} f(x) e^{-ijx}\, dx \qquad (۴۷.۶)$$

درشرایط نسبتاً معتدل، سری فوق به‌سمت $f(x)$ همگرا می‌شود [اما نه برای هر $f(x)$]. برای مثال، اگر $f(x)$ تابعی پیوسته و دارای اولین مشتق پیوسته-تکه‌ای[1] باشد، سری به‌طور یکنواخت همگرا می‌شود.

سری فوریه با توجه به‌دستور مسلم زیر به‌دست می‌آید

$$\int_{۰}^{۲\pi} \overline{e^{ijx}}\, e^{ikx}\, dx = \int_{۰}^{۲\pi} e^{i(k-j)x}\, dx = \begin{cases} \int_{۰}^{۲\pi} ۱\, dx = ۲\pi & \text{اگر } k=j \\ \dfrac{۱}{i(k-j)} e^{i(k-j)x} \Big|_{۰}^{۲\pi} = ۰ & \text{اگر } k \neq j \end{cases}$$

این امر نشان می‌دهد که توابع ۱ و $e^{\pm ix}$ و $e^{\pm i۲x}$ نسبت به ضرب اسکالر یا ضرب داخلی

$$\langle g, h \rangle = \frac{۱}{۲\pi} \int_{۰}^{۲\pi} \overline{h(x)} g(x)\, dx$$

متعامدند. به‌عبارت دیگر

$$\langle e^{ikx}, e^{ijx} \rangle = \begin{cases} ۱ & k=j \\ ۰ & k \neq j \end{cases}$$

که این مطلب اثبات قضیهٔ زیر است.

قضیهٔ ۴.۶ حاصلجمع جزئی[2]

$$\sum_{j=-n}^{n} \hat{f}(j) e^{ijx}$$

سری فوریه برای $f(x)$ نسبت به نرم

$$\|g\| = \|g\|_{۲} = \left[\frac{۱}{۲\pi} \int_{۰}^{۲\pi} |g(x)|^{۲}\, dx \right]^{۱/۲}$$

---

1. piecewise-continuous 2. partial sum

بهترین تقریب برای $f(x)$ به‌وسیلهٔ بسجمله‌ایهای مثلثاتی از مرتبهٔ n است. بعلاوه می‌توان نشان دادکه **رابطهٔ پارسوال**[1]

$$\sum_j |\hat{f}(x)|^2 = \frac{1}{2\pi}\int_0^{2\pi} |f(x)|^2\,dx \qquad (۴۸.۶)$$

برقراراست.

ضرایب فوریه $\hat{f}(j)$ مربوط به‌تابع $f(x)$ برای «درك» این تابع به‌شرح زیربه‌کار گرفته می‌شوند. فرض کنیدکه تابع $f(x)$ حقیقی و 2π-دوره‌ای باشد. اگر $f(x)$ به‌صورت موقعیت جسم درحال حرکتی بریك خط در زمان x درنظرگرفته شود، آنگاه تابع 2π-دوره‌ای $f(x)$ یك حرکت دوره‌ای را بیان می‌کند. حال اگر

$$\hat{f}(j) = |\hat{f}(j)|e^{i\theta_j}$$

درست باشد [**شکل قطبی** عدد مختلط $\hat{f}(j)$]، آنگاه می‌توان سری فوریه برای $f(x)$ را به‌صورت زیر نوشت

$$f(x) \approx 2\sum_{j=0}^{\infty} |\hat{f}(j)|\cos(\theta_j + jx)$$

(به‌تمرین ۵.۶–۷ نگاه کنید). بدین طریق حرکت دوره‌ای ما که با $f(x)$، بیان شده به‌صورت حاصلجمع یا برهمنهی[2] نوسانهای همساز[3] ساده نمایش داده شده‌است. jامین حرکت از این نوع

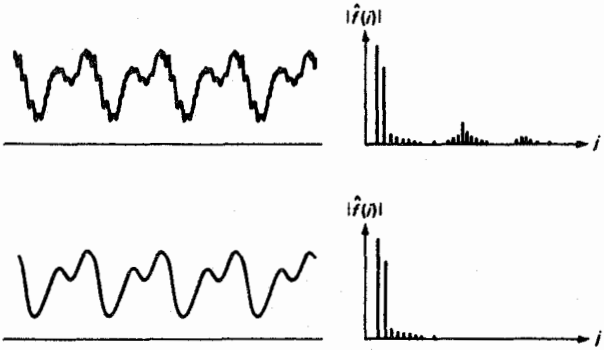
$$2|\hat{f}(j)|\cos(\theta_j + jx)$$

دارای **دامنهٔ** $2|\hat{f}(j)|$، **فراوانی** $j/(2\pi)$، **فراوانی زاویه‌ای** j، دوره‌یاطول موج $2\pi/j$ و **زاویهٔ فاز** θ_j می‌باشد. عدد $|\hat{f}(j)|$ معرف میزان‌حرکت‌همساز ساده بافراوانی‌زاویه‌ای j است‌که درکل حرکت موجود است. تمامی‌دنبالهٔ $|\hat{f}(0)|$ و $|\hat{f}(1)|$، ... (یااحتمالا، دنبالهٔ مربعات آنها) را **طیف توان**[4] یا به‌طور ساده **طیف** $f(x)$ گویند. باید توجه داشت که به‌موجب رابطهٔ پارسوال (۴۸.۶) طیف $f(x)$ باکران $\|f\|_2$ محدود می‌شود، اما بسته به‌اینکه چگونه «کل‌انرژی» $\|f\|_2^2$ روی طیف $|\hat{f}(0)|$ و $|\hat{f}(1)|$ و ... توزیع شده باشد، $f(x)$ ممکن‌است رفتار بسیار متفاوتی داشته باشد. یك تابع «نوفه‌ای» به‌ازای مقادیر بزرگ j، دارای $|\hat{f}(j)|$ی قابل‌ملاحظه‌ای‌است، درصورتی‌که دریك تابع «هموار» طیف همراه با افزایش j به‌سرعت‌کاهش می‌یابد، به‌شکل ۷.۶ نگاه‌کنید.

یك روش موردتوجه ما برای **هموارکردن**، عبارت است از تولید ضرایب فوریهٔ

---

1. Parseval's relation
2. superposition
3. harmonic oscillation
4. power spectrum



شکل ۷.۶ دو تابع 2π-دوره‌ای حقیقی و طیف توان آنها. تابع دومی از کوچک‌کردن اثر فراوانیهای بالا در تابع اولی به‌دست آمده‌است.

تابع داده شده از روی داده‌ها، و سپس **پالایش** این ضرایب یعنی نادیده گرفتن برخی فراوانیها، معمولاً فراوانیهای زیاد، و سپس بازسازی تابع مزبور به‌صورت سری فوریه با استفاده ازاین ضرایب «خالص‌شده» یا «پالایش‌شده». برای مثال به‌شکل ۷.۶ نگاه کنید.

می‌توان نشان دادکه درصورتی‌که $f(x)$ دارای $k-1$ مشتق پیوسته (مانند یك تابع دوره‌ای!) و مشتق kام به‌طور تکه‌ای پیوسته (یا حتی فقط تغییرات کرانی محدود) باشد، رابطهٔ

$$|\hat{f}(j)| = \mathcal{O}(|j|^{-k-1}) \qquad (۴۹.۶)$$

برقرار است. برای مثال، «موج مربعی»

$$f(x) = \text{signum}(\sin x) = \begin{cases} 1 & 0 < x < \pi \\ -1 & \pi < x < 2\pi \end{cases}$$

فقط به‌طور تکه‌ای پیوسته‌است. بنابراین انتظار می‌رودکه وقتی $|j| \to \infty$، $|\hat{f}(j)|$ با سرعتی نه بیشتر از $1/|j|$ به‌سمت صفر میل نماید. این امر با محاسبات مستقیم زیرتأیید می‌شود:

$$\hat{f}(j) = \frac{1}{2\pi}\int_0^{2\pi} \text{signum}(\sin x) e^{-ijx}\,dx$$

$$= \frac{1}{2\pi}\left\{\int_0^{\pi} e^{-ijx}\,dx - \int_{\pi}^{2\pi} e^{-ijx}\,dx\right\}$$

$$= \frac{1}{2\pi}\,\frac{1}{-ij}\left\{2e^{-ij\pi} + 2\right\} = i\,\frac{2}{\pi}\begin{cases} 0 & j \text{ زوج} \\ 1/j & j \text{ فرد} \end{cases}$$

باید توجه داشت که طیف تابع

$$f(x)=x$$

سویعتر از $1/j$ تحلیل نمی‌رود، ولو اینکه تابع مزبور غالباً بی‌نهایت بار مشتقپذیر باشد. علت آن این‌است که تحلیل فوریه (آن‌گونه که در اینجا تشریح شده است) با این تابع به‌صورت یك تابع 2π-دوره‌ای، که مقدارش به‌ازای $0<x<2\pi$ برابر x است، عمل می‌نماید. اما این تابع اخیر در تمام مضارب 2π دارای یك جهش ناپیوسته‌است!

معمولا بعیداست که بتوان ضرایب فوریه (۴۷.۶) را به‌طوردقیق محاسبه کرد، بدین علت‌که یا انتگرال را نمی‌توان به‌شکل بسته محاسبه‌کرد و یا تابع $f(x)$ دقیقاً مشخص نیست. درهرحال از **انتگرالگیری عددی** استفاده می‌شود. در فصل ۷، ضمن یك متن کلی، مدخلی برای این مبحث قدیمی و دامنه‌دار آورده شده‌است. به‌منظورکنونی ما، قاعدهٔ تقریب‌زدن بسیارسادهٔ زیر کفایت می‌کند

$$\int_0^{2\pi} g(x)\,dx \approx \frac{2\pi}{N} \sum_{n=0}^{N-1} g\left(\frac{2\pi n}{N}\right) \qquad (۵۰.۶)$$

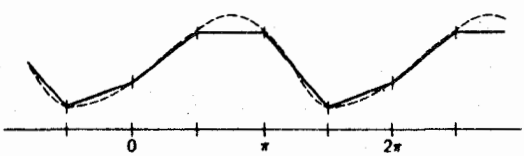
این قانون ذوزنقه‌ای مرکب (۴۹.۷) است که برای انتگرال فوق اعمال شده‌است، البته با درنظرگرفتن اینکه تابع $g(x)$ یك تابع 2π-دوره‌ای‌است، و بنابراین، به‌ویژه، $g(2\pi)=g(0)$. قاعدهٔ فوق را می‌توان‌ازگذاردن یك تابع درونیاب به‌طورتکه‌ای خطی که درنقاط انفصال ومتساوی‌الفاصلهٔ 0 و $\pm 2\pi/N$، $\pm 2\pi 2/N$، $\pm 2\pi 3/N$، ... با $g(x)$ تطابق دارد به‌جای تابع 2π-دوره‌ای $g(x)$ زیر علامت انتگرال، به دست آورد. به‌شکل ۸.۶ نگاه کنید.

تقریب متناظر با $\hat{f}(j)$ را به‌وسیلهٔ $\hat{f}_N(j)$ در زیر مشخص می‌کنیم:

$$\hat{f}_N(j)=\frac{1}{N}\sum_{n=0}^{N-1} f(x_n)e^{-ix_n} \qquad (۵۱.۶)$$

که در آن

$$x_n=2\pi n/N \qquad n=0,\ldots,N-1$$



شکل ۸.۶ یك تابع 2πـ دوره‌ای (نقطه‌چین) و یك تابع درونیاب به‌طور تکه‌ای خطی (تمام خط) در $N=4$ نقطه در هر دوره

این نقاط x_n را **نقاط نمونه‌گیری**[1] و اعداد $f(x_n)$ متناظر را **مقادیر نمونه**[2] گویند. عدد $N/2\pi$ را **بازهٔ نمونه‌گیری** وعدد $N/(2\pi)$ را **فراوانی نمونه‌گیری**[3] گویند.

سؤال این‌است که $\hat{f}_N(j)$ با چه دقتی $\hat{f}(j)$ را تقریب می‌زند؟ برای پاسخ به‌این سؤال، این امر را متذکر می‌شویم که توابع ۱ و $e^{\pm ix}$ و $e^{\pm i2x}$، برخی ویژگیهای تعامدی نسبت به‌نوع دیگری ازضرب اسکالر یا داخلی‌دارندکه‌آن را **ضرب داخلی گسسته** می‌نامند، یعنی

$$\langle g, h\rangle_N = \frac{1}{N}\sum_{n=0}^{N-1}\overline{h(x_n)}g(x_n) \qquad (۵۲.۶)$$

روشنتر بگوییم، داریم

$$\langle e^{ikx}, e^{ijx}\rangle_N = \begin{cases} 1 & \text{اگر } k=j(\bmod N) \\ 0 & \text{اگر } k\neq j(\bmod N) \end{cases} \qquad (۵۳.۶)$$

و اثبات این‌مطلب درحد اثبات محاسبهٔ حاصلجمع یك‌سری هندسی متناهی‌است (به‌تمرین ۵.۶–۸ نگاه‌کنید).

با توجه به‌مطلب فوق، ملاحظه می‌کنیم که

$$\hat{f}_N(j) = \langle f, e^{ijx}\rangle_N$$

بنابراین، با فرض اینکه سری فوریه $\sum\hat{f}(j)e^{ijx}$ به‌سمت $f(x)$ مطلقاً همگرا می‌شود (که این خود مستلزم چیزی جزوجود $\lim\sum_{-\infty}^{\infty}|\hat{f}(j)|$ نیست) از رابطهٔ (۵۳.۶) نتیجه می‌شود که

$$\hat{f}_N(j) = \Big\langle \sum_{k=-\infty}^{\infty}\hat{f}(k)e^{ikx}, e^{ijx}\Big\rangle_N = \sum_k \hat{f}(k)\langle e^{ikx}, e^{ijx}\rangle_N$$

یا

$$\hat{f}_N(j) = \sum_{k=j(\bmod N)}\hat{f}(k) \qquad (۵۴.۶)$$

یعنی: ضریب تقریبی فوریهٔ $\hat{f}_N(j)$، متشکل از همهٔ ضرایب درست فوریهٔ $\hat{f}(k)$ است که تابع مربوطه‌شان، e^{ikx}، را نمی‌توان به‌وسیلهٔ ضرب داخلی (۵۲.۶) از تابع e^{ijx} تمیزداد.

این پدیده **دگرنامی**[4] نامیده شده‌است. اگر $k=j(\bmod N)$، آنگاه تساوی $k=j+mN$ به‌ازای برخی از اعداد صحیح m برقراراست. اما دراین‌صورت، به‌ازای

---

1. sampling points 2. sample values
3. sampling frequency 4. aliasing

هر n داریم

$$e^{ikx_n} = e^{i(j+mN)x_n} = e^{ijx_n}e^{imNx_n}$$

و

$$e^{imNx_n} = e^{(i2\pi)mn} = 1$$

دراین‌حال رابطهٔ فوق بیانگر تساوی زیراست

$$e^{ikx} = e^{ijx} \qquad \text{برای } x = x_n = 2\pi n/N \text{ و به‌ازای همهٔ مقادیر } n$$

یعنی، دوتابع e^{ikx} و e^{ijx} درکلیهٔ نقاط نمونه‌گیری‌که در محاسبهٔ $\hat{f}_N(j)$، یعنی در ضرب داخلی گسستهٔ $\langle\ ,\ \rangle_N$ به‌کار برده شده‌اند، تطابق دارند. اگرتنها مقادیر تابع را در نقاط نمونه‌گیری x_n، به‌ازای جمیع مقادیر n، درنظر گیریم، آنگاه نمی‌توان بین توابع e^{ikx} و e^{ijx} تفاوتی قائل‌شد.

یك مثال جالب از نتیجهٔ فوق، مثال چرخهای واگن درفیلمهاست،که بدون حرکت و یا حتی درحال چرخش در جهت مخالف با حرکت واگن به‌نظر می‌رسند. در اینجا هر ۱/۲۰ ثانیه از یك حرکت دوره‌ای نمونه‌برداری می‌شود. و در دید تماشاگر به‌صورت کندترین حرکت سازگار با شواهد جلوه می‌کند.

به‌همین‌ترتیب معمولا (هنگام نمونه برداری در N نقطه با فاصلهٔ یکنواخت در $[0, 2\pi]$)، تابع e^{ijx}را با $e^{ij'x}$ یکی می‌گیرند، البته دراینجا $j' = j \pmod N$ وفراوانی (زاویه‌ای) $|j'|$کوچکترین مقدارممکن‌است. باید توجه‌کرد‌که بدین‌طریق j' به‌طورمنحصر به‌فردی به‌وسیلهٔ j و N معین می‌شود، جز در حالت زیر: اگر N زوج و j مضرب فردی از $N/2$ باشد، آنگاه هم $N/2$ و $N/2$ — هردو می‌توانند j' گرفته شوند. درمورد اخیر چنین‌متداول شده‌است‌که میانگین دوتابع $e^{i(N/2)x}$ و $e^{-i(N/2)x}$، یعنی‌تابع $\cos(N/2)x$ را به‌عنوان نمایندهٔ ردهٔ خودانتخاب می‌کنند.

به‌گونه‌ای مشابه، گرچه رابطهٔ (۵۱.۶)، به‌ازای‌کلیهٔ مقادیر j، تابع

$$\hat{f}_N(j) = \langle f, e^{ijx}\rangle_N$$

را به‌عنوان تقریبی برای $\hat{f}(j)$ درنظر می‌گیرد ولی معمولا تابع فوق تنها به‌ازای $|j| \leqslant N/2$ تقریبی برای $\hat{f}(j)$ محسوب می‌شود. به‌ویژه وقتی $f(x)$ تابعی هموار و $|j|$ از $N/2$ خیلی‌کوچکتر باشد تعبیر خوبی دارد. در این صورت از ترکیب روابط (۴۹.۶) و (۵۴.۶)، درحالتی‌که $f(x)$ دارای $k-1$ مشتق پیوسته، و kامین مشتق آن به‌طور تکه‌ای پیوسته باشد، خواهیم داشت

$$\hat{f}_N(j) = \hat{f}(j) + \mathcal{O}(N^{-k}) \qquad (۵۵.۶)$$

درواقع، وقتی تابعی در N نقطهٔ‌متساوی‌الفاصله درباز‌هٔ $[0, 2\pi]$نمونه‌برداری‌می‌شود

اثر دگرنامی درفراوانیهای بالاتراز$(N/۲)/(۲\pi)$، مانع مشاهده‌شدن پدیده‌های دوره‌ای در $f(x)$ می‌شود. به‌طور قطعی بگوییم، اگر بخواهیم یك پدیدهٔ دوره‌ای با فراوانی ν را مشاهده‌کنیم، باید حداقل با یك فراوانی به‌بزرگی ۲ν نمونه‌برداری‌کنیم

اکنون به‌اجمال تقریب‌ساز مربوط به‌بسجمله‌ای مثلثاتی

$$p(x)=\sum_{|j|<N/۲}\hat{f}_N(j)e^{ijx}+\mathrm{Re}[\hat{f}_N(N/۲)e^{i(N/۲)x}]$$

را موردبحث قرارمی‌دهیم. دراینجا تنها زمانی آخرین جمله وجود داردکه $N/۲$ یك‌عدد صحیح، یعنی N یك عدد زوج باشد. اما با توجه به‌اینکه جملهٔ اخیر را برای تکمیل بحث آوردیم (به‌تمرین ۵۰۶–۱۱ نگاه‌کنید) اکنون تنها به‌حالتی‌که N فرد باشد،

$$N=۲n+۱$$

می‌پردازیم. دراین‌حال به‌موجب(۵۳۰۶)، تعداد$N=۲n+۱$ تابع ۱ و $e^{\pm ix}$، ...، $e^{\pm inx}$ وجود داردکه نسبت به‌ضرب داخلی گسسته $\langle\ ,\ \rangle_N$ متعامدند، یعنی

$$\langle e^{ikx},e^{ijx}\rangle_N=\begin{cases}۱ & k=j\\ ۰ & k\neq j\end{cases}\qquad k,j=-n,\dots,n\qquad(۵۶۰۶)$$

بنا به‌دلائل مذکور دربخش۲۰۶، قضیهٔ زیرحاصل می‌شود.

قضیهٔ ۵۰۶ به‌ازای هر $m\leqslant n$، بسجمله‌ای مثلثاتی مرتبهٔ mام

$$p_m(x)=\sum_{j=-m}^{m}\hat{f}_N(j)e^{ijx}$$

بهترین تقریب به‌وسیلهٔ بسجمله‌ایهای مثلثاتی‌مرتبهٔ mام نسبت به‌نرم گسستهٔ میانگین‌مربعات زیر برای $f(x)$ است.

$$\|g\|_۲=(\langle g,g\rangle_N)^{۱/۲}=\frac{۱}{N}\left(\sum_{j=۰}^{N-۱}|g(j۲\pi/N)|^۲\right)^{۱/۲}$$

به‌ازای $m=n$، این بدان معنی‌است‌که بسجمله‌ایهای مثلثاتی مرتبهٔ nام

$$p_n(x)\quad\sum_{j=-n}^{n}f_N(j)e^{ijx}$$

تابع $f(x)$ را درنقاط نمونه‌گیری$x_j=۲\pi j/N$، به‌ازای کلیهٔ مقادیرj، درونیابی می‌کند.

اگر $f(x)$ یك تابع حقیقی باشد، آنگاه می‌توانیم به‌موجب رابطهٔ (۴۵۰۶) بسجمله‌ایهای درونیاب را به‌شکل حقیقی زیربنویسیم

$$p_n(x)=a_0/2+\sum_{m=1}^{n}[a_m\cos mx+b_m\sin mx] \qquad (۵۷.۶)$$

که در آن

$$a_m=2\mathrm{Re}\,\hat{f}_N(m)=\frac{2}{N}\sum_{k=0}^{N-1}f(x_k)\cos mx_k \qquad (۵۸.۶\text{ الف})$$

$$b_m=-2\mathrm{Im}\hat{f}_N(m)=\frac{2}{N}\sum_{k=0}^{N-1}f(x_k)\sin mx_k \qquad (۵۸.۶\text{ ب})$$

□ **مثال ۱۴.۶:** برای تابع $f(x)=\sin x$ یك درونیاب مثلثاتی مرتبهٔ ۱ به‌دست می‌آوریم. دراین صورت $N=3$، وکمیتهای مربوطه عبارتند از:

$$\omega=e^{i2\pi/3}=\cos 2\pi/3+i\sin 2\pi/3=\frac{1}{2}+i\sqrt{\frac{3}{4}}$$

این مقادیر حائز اهمیت‌اند زیرا $(\mathbf{w}^{(m)}):=(e^{imx_j})=(\omega^{mj})$. بعلاوه

j	x_j	$f(x_j)$	$\mathbf{w}^{(0)}$	$\mathbf{w}^{(1)}$	$\mathbf{w}^{(-1)}$
0	0	0	1	1	1
1	$2\pi/3$	$\sqrt{\frac{3}{4}}$	1	ω	$\omega^{-1}=\bar{\omega}$
2	$4\pi/3$	$-\sqrt{\frac{3}{4}}$	1	ω^2	$\omega^{-2}=\bar{\omega}^2$

اما داریم $c_j=\hat{f}_N(j)=(\mathbf{w}^{(j)})^{\mathrm{H}}f/M$. بنابراین $\omega^2=\omega^{-1}=\bar{\omega}$، و داریم

$$c_0=\frac{1}{3}\left[0+\sqrt{\frac{3}{4}}-\sqrt{\frac{3}{4}}\right]=0$$

$$c_1=\frac{1}{3}\left[0+\sqrt{\frac{3}{4}}\omega^{-1}-\sqrt{\frac{3}{4}}\omega^{-2}\right]=-\frac{1}{3}\sqrt{\frac{3}{4}}(\omega-\bar{\omega})=-\frac{2}{3}i\left(\sqrt{\frac{3}{4}}\right)^2$$

$$=-i/2$$

$$c_{-1}=\frac{1}{3}\left[0+\sqrt{\frac{3}{4}}\omega-\sqrt{\frac{3}{4}}\omega^2\right]=\bar{c}_1=i/2$$

بنابراین $a_1 = 2\mathrm{Re}(c_1) = c_1 + c_{-1} = 0$ و $b_1 = -2\mathrm{Im}(c_1) = i(c_1 - c_{-1}) = 1$ که نشان می‌دهند

$$p_1(x) = 0 + 0 \cdot \cos 1x + 1 \cdot \sin 1x = \sin x$$

همان‌گونه که انتظار داشتیم. □

در این رهگذر تذکر داده می‌شود که درونیابی به‌وسیلهٔ بسجمله‌ایهای مثلثاتی مرتبهٔ n درهریك از $2n+1$ نقطهٔ متمایز در $[0, 2\pi)$ به‌طور منحصر به‌فردی انجام می‌گیرد. در مورد بسجمله‌ای درونیاب حاصلهٔ $p_n(x)$ برای $f(x)$ می‌توان نشان داد که

$$\|f - p_n\|_\infty \leqslant \mathrm{const}\ \mathrm{dist}_\infty(f, \mathring{\pi}_n) \qquad (59.6)$$

در اینجا نرم ماکسیمم در بازهٔ $[0, 2\pi]$ گرفته شده‌است

$$\|g\|_\infty = \max_{0 \leqslant x \leqslant 2\pi} |g(x)|$$

و

$$\mathrm{dist}_\infty(f, \mathring{\pi}_n) = \min_{p \in \mathring{\pi}_n} \|f - p\|_\infty$$

که در آن نماد $p \in \mathring{\pi}_n$ به‌جای ملخص عبارت «$p(x)$ یك بسجمله‌ای مثلثاتی از مرتبهٔ n است» به‌کار برده شده‌است. می‌توان نشان داد که رابطهٔ (۵۹.۶) به‌نامساوی متناظر (۱۷.۶) برای درونیابی به‌وسیلهٔ بسجمله‌ای بسیار شبیه است. به‌ویژه اینکه عدد ثابت (const) به‌نقاط درونیابی بستگی دارد. در این عبارات نقاط درونیابی با فواصل یکنواخت که در اینجا منحصراً استفاده شده‌اند. بهینه هستند، بدین تعبیر که عدد ثابت (const) رابطهٔ (۵۹.۶) را تا حد امکان کوچك می‌کند، به کتاب دبوروپینکوس[1] [۳۹] نگاه کنید. بهترین مقدار این عدد ثابت توسط الیش و تسلر[2] [۳۸] به‌صورت زیر محاسبه شده‌است

$$\mathrm{const}_{\text{uniform}}\ (\text{ثابت یکنواخت})$$

$$= 1 + \frac{1}{N}\left\{1 + 2\sum_{k=1}^{n} 1/\sin\left(\frac{(2k-1)\pi}{(2n+1)2}\right)\right\} \approx \frac{2}{\pi}\ln N + c \qquad (60.6)$$

بنابراین برای مقادیری از n که جنبهٔ عملی داشته باشند، درونیابی در نقاط با فواصل یکنواخت، تقریبی به‌دست می‌دهد که از بهترین تقریب یکنواخت احتمالی از $\mathring{\pi}_n$ خیلی بدتر نیست. بنابراین در صورتی که بسجمله‌ای درونیاب به‌آسانی به‌دست آید، معمولاً نیازی به‌انجام روند پیچیده برای به‌دست‌آوردن بهترین تقریب یکنواخت نیست. این مسئلهٔ اخیر در بخش

1. de Boor and Pinkus 2. Ehlich and Zeller

بعد موردبحث قرار خواهدگرفت.

دراین مورد، اشاره می‌شودکه در صورتی‌که $f(x)$ دارای kمشتق بـا مشتق kامی به‌طور تکه‌ای پیوسته باشد، رابطهٔ (۴۹.۶) ایجاب می‌کندکه

$$\mathrm{dist}_\infty(f, \mathring{\pi}_n) = \mathcal{O}(n^{-k}) \qquad (۶۱.۶)$$

تمرین

۱-۵.۶ سری فوریه برای تابـع 2π-دوره‌ای $f(x)$ راکه به‌وسیلهٔ $f(x)=x$ داده شده است در بازهٔ $(0, 2\pi]$ محاسبه‌کنید.

۲-۵.۶ تحقیق‌کنیدکه تابع 2π-دوره‌ای $f(x)$که مقادیرش در بازهٔ $(0, 2\pi]$ به‌شرح‌زیر داده شده‌است

$$f(x)=\begin{cases}(x/\pi)^2 - x/\pi & 0 \leqslant x \leqslant \pi \\ (x-\pi)/\pi - ((x-\pi)/\pi)^2 & \pi \leqslant x \leqslant 2\pi\end{cases}$$

پیوسته بوده ومشتق اول پیوسته نیز دارد (مانند یك تابع 2π-دوره‌ای)، امـا درمشتق دوم جهشهایی* دارد. سپس‌طیف $f(x)$ را تشکیل ونشان دهیدکه وقتی $j \to \infty$، مانند j^{-3}، (ونه سریعتر ازآن) به‌سمت صفرمیل می‌کند.

۳-۵.۶ سری فوریه را که درتمرین ۵.۶-۲ به‌دست آمد برحسب sin و cos بنویسید. چرا انتظار داریدکه‌کلیهٔ a_jها برابر با صفرباشند؟

۴-۵.۶ اگر $f(x)$ یك تابع 2π-دوره‌ای باشد، آنگاه به‌ازای هر عدد صحیح m، تابع $gm(x)=f(mx)$ نیز 2π - دوره‌ای‌است. چه رابطه‌ای بین $\hat{f}(j)$ و $\hat{g}_m(j)$ وجوددارد؟

۵-۵.۶ اگرتابع $f(x)$ یك تابـع 2π-دوره‌ای باشد، آنـگاه به‌ازای هرعدد α، تابـع $g_\alpha(x)=f(x-\alpha)$ نیز 2π-دوره‌ای‌است. چه رابطه‌ای بین $\hat{f}(j)$ و $\hat{g}_\alpha(j)$ وجود دارد؟

۶-۵.۶ فرض‌کنیدکه تابع $f(x)$ تابعی بسیار هموار با دورهٔ (تناوب) τ باشد. امـا در تبدیل آن به‌یك تابـع 2π-دوره‌ای $g(x)=f(\tau x/(2\pi))$ اشتباهاً τ' به‌جای τ بـه‌ازای $\tau' \neq \tau$ به‌کار گرفته شده است. اثرات احتمالی این اشتباه روی ضرایب فوریهٔ $\hat{g}(j)$ چه هستند؟

۷-۵.۶ ثابت‌کنیدکه اگر $f(x)$ یك تابع حقیقی باشد، آنگاه رابطهٔ

$$\hat{f}(j)e^{ijx} + \hat{f}(-j)e^{-ijx} = 2|\hat{f}(j)|\cos(\theta_j + jx)$$

---

* توضیح مترجم: جهش به‌معنی ناپیوستگی مقطعی‌است.

به‌ازای یك اختلاف فاز مناسب θ_j برقراراست. (راهنمایی: از این واقعیت استفاده کنید که هرعدد همتافت z را می‌توان به‌ازای یك θی مناسب به‌شکل قطبی $|z|e^{i\theta}$ نوشت.)

۸-۵.۶ قضیهٔ (۵۳.۶)را اثبات کنید (راهنمایی: بیاد آورید که چگونه مجموع یك‌سری هندسی به‌دست می‌آید.)

۹-۵.۶ رابطهٔ (۶.۵) را اثبات کنید.

۱۰-۵.۶ فرمولهای جمع برای $\sin(\alpha+\beta)$ و $\cos(\alpha+\beta)$ را ازفرمول‌اویلر(۴۲.۶) و قانون جمع نماها، $e^{A+B}=e^Ae^B$، به‌دست آورید.

۱۱-۵.۶ ثابت کنیدکه اگر N زوج و $f(x)$ حقیقی باشد، آنگاه تابع

$$p(x)=\sum_{|j|<N/2}\hat{f}_N(j)e^{ijx}+\mathrm{Re}[\hat{f}(N/2)e^{i(N/2)x}]$$

$f(x)$ را درنقاط نمونه‌گیری x_k به‌ازای کلیهٔ مقادیر k درونیابی می‌کند.

۱۲-۵.۶ چگونه می‌توان بسجمله‌ای درونیابی مثلثاتی برای $f(x)$ را درنقاط

$$\alpha+k2\pi/N, k=0,\cdots,N-1,$$

αعدد مثبتی کمتر از $2\pi/N$، به‌دست آورد؟

۶.۶* تبدیلهای سریع فوریه[1]

درآنالیزهمساز گسسته، یاآنالیزفوریه، برای تجزیهٔحرکت 2π-دوره‌ای (که با $f(x)$بیان شده است) به‌همسازهای ساده، اعداد

$$c_j=\hat{f}_N(j)=\frac{1}{N}\sum_{k=0}^{N-1}f(x_k)e^{-ijx_k} \qquad (۶۲.۶)$$

$$x_k=2\pi k/N \text{ باشرط} \qquad (۶۳.۶)$$

را محاسبه می‌کنند. همان‌طوری‌که در بخش۵.۶ دیدیم‌اگر$f(x)$ دارای مشتق kام پیوستهٔ تکه‌ای باشد،آنگاه

$$f(x)\approx\sum_{j=-\infty}^{\infty}\hat{f}(j)e^{ijx}$$

با فرض

$$\hat{f}(j)=\hat{f}_N(j)+\mathcal{O}(N^{-k})$$

ما به‌آن حالتی‌که فراوانیهایی در$f(x)$وجودداشته‌باشند،وبه‌توان این‌فراوانیهاعلاقه‌مندیم.

1. Fast fourier transform

اما به‌علت اثر دگرنامی، به‌ازای $|j| > N/2$، $\hat{f}_N(j)$ به‌عنوان تقریبی برای $\hat{f}(j)$ مفید فایده‌ای نیست و معمولا تقریب خوبی تنها برای موقعی‌است که مقادیر $|j|$ خیلی کمتر از $N/2$ باشد. با توجه بدین امر باید محاسبهٔ $\hat{f}_N(j)$ به‌ازای مقادیر «بزرگ» N، و لذا مسئلهٔ مهم چگونگی محاسبهٔ $\hat{f}_N(j)$ به‌نحوکارا، مورد توجه قرار گیرد.

واضح‌است که محاسبهٔ هر تابع خاص $\hat{f}_N(j)$ مستلزم تعداد $\mathcal{O}(N)$ عمل ضرب وجمع است. بنابراین محاسبهٔ مستقیم N عدد ازاین نوع اعداد (مثلا، اعداد $\hat{f}_N(j)$ به‌ازای $|j| \leqslant N/2$) نیاز به $\mathcal{O}(N^2)$ عمل دارد. ازاین‌رو برای ۱۰۰۰ نقطهٔ نمونه، به‌میلیونها عمل نیاز خواهد بود و تا این اواخر، این امر مانع اصلی برسر راه آنالیز گسستهٔ فوریه بود.

وقتی‌کاملا معلوم شدکه محاسبهٔ همزمان N شمارهٔ متوالی از $\hat{f}_N(j)$ها به‌دلیل روابط قوی بین این‌اعداد تنها به $\mathcal{O}(N \log N)$ عمل حساب نیازدارد، وضعیت به‌نحو مؤثری دگرگونی پیداکرد. کلید حل این مشکل **تبدیل سریع فوریه** (ت س ف) یا **FFT**[1] بودکه این محاسبات رابرای $N \leqslant ۱۰۰۰$ به‌یك امرعادی مبدل‌کرد وحتی این امکان را به‌وجود آوردکه Nهای برابر با ده‌ها هزار نیز به‌کارگرفته شوند.

در اینجا ما می‌توانیم تنها به‌آن طرح اساسی‌که به‌چنین افزایش مؤثری در کارایی منجرشده اشاره‌کنیم. تازه‌ترین مطالب مذکور در این باب را می‌توان در مقالهٔ ۱۹۷۸ س. وینگراد[2] [۳۶] به‌دست آورد. به‌ویژه‌کارهایی‌که قبل و بعد از انتشار مقالهٔ نهادی و بی‌سابقهٔ کولی وتیوکی[3] [۳۷] انجام گرفته، از مدتها قبل روشن کرده‌اند که FFT (ت س ف)های زیادی وجود دارند و برای به‌دست آوردن کارایی بیشتر، لازم (و مقرون به‌صرفه)است که به‌ازای مقادیرمختلفی از N که می‌خواهند به‌کاربرند، برنامه‌های مختلفی بنویسند.

برای تحلیل محاسبات مربوط به‌اعداد $\hat{f}_N(j)$ به‌ازای $|j| \leqslant N/2$ از روی اعداد $f(x_0), \ldots, f(x_{N-1})$، آسانتر است که **تبدیل گسستهٔ فوریه**[4] F_N معرفی شود. این تبدیل N-بردار

$$\mathbf{z} = [z_1 z_2 \cdots z_N]^T$$

را به N-بردار

$$F_N \mathbf{z} = \hat{\mathbf{z}}$$

منتقل می‌کندکه با

$$\hat{z}_j = \sum_{n=1}^{N} z_n \omega_N^{(j-1)(n-1)} \qquad j = 1, \ldots, N \qquad (۶۴.۶)$$

---

1. fast Fourier transform.
2. S. Winograd
3. Cooley and Tukey
4. discrete Fourier transform

معین می‌شود و در آن ω_N، nامین ریشهٔ واحد است

$$\omega_N = e^{-i2\pi/N}$$

ارتباط بین محاسبهٔ $\hat{f}_N(j)$ و این تبدیل گسستهٔ فوریه به‌شرح زیر است. اگر N-بردار خاص زیر را در نظر بگیریم

$$\mathbf{z} = [f(x_0)f(x_1) \cdots f(x_{N-1})]^T$$

آنگاه

$$\hat{f}_N(j) = \frac{1}{N}\begin{cases}\hat{z}_{j+1} & j = 0, 1, 2, \cdots \\ \hat{z}_{j+1+N} & j = -1, -2, \cdots\end{cases} \qquad |j| \leqslant N/2 \qquad (۶۵.۶)$$

بنابراین کافی است که سعی خود را بر محاسبهٔ کارای تبدیل گسستهٔ فوریه متمرکز سازیم.

این بحث را با ملاحظهٔ این امر آغاز می‌کنیم که z_j به‌صورتی که با رابطهٔ (۶۴.۶) داده شده، یک بسجمله‌ای است از درجهٔ کمتر از N بر حسب کمیت ω_N^{j-1}، بنابراین می‌توان آن را طی N عمل با ضرب تودرتو محاسبه کرد. در اینجا یک عمل ضرب بعلاوهٔ یک عمل جمع را به‌عنوان یک عمل[1] تلقی می‌کنیم. بنابراین ارزیابی سرراست رابطهٔ (۶۴.۶)، به‌ازای جمیع مقادیر j، به‌تعداد N^2 عمل نیازمند است.

گسترده‌ترین طرحی که برای یک (ت س ف) داریم، طرحی است که کولی و تیوکی آن را به‌زبان ساده درآورده‌اند و هنگامی قابل استفاده است که N مساوی با حاصلضرب اعداد صحیح باشد. اکنون این طرح را اول برای حالت زیر مورد بحث قرار می‌دهیم

$$N = P \cdot Q$$

N-بردار $\mathbf{z}$ را که به‌شیوهٔ **فورترن** در یک آرایهٔ یک‌بعدی ذخیره شده در نظر می‌گیریم. در این صورت می‌توانیم این آرایه را به‌شیوهٔ **فورترن** به‌صورت یک آرایهٔ دوبعدی Z با ابعاد (P, Q) نشان دهیم. این بدان معناست که داریم

$$Z(p, q) = z_{p+P(q-1)}$$

همچنین حاصلجمع $\sum z_n \omega_N^{(p-1)(n-1)} = \hat{z}_p$ را به‌حاصلجمع مضاعف:

$$\hat{z}_p = \sum_{p=1}^{P}\sum_{q=1}^{Q} Z(p, q)\omega_N^{(p-1)[p-1+P(q-1)]}$$

$$= \sum_{p=1}^{P}\left[\sum_{q=1}^{Q} Z(p, q)\omega_Q^{(p-1)(q-1)}\right]\omega_N^{(p-1)(p-1)}$$

---

1. operation

تجزیه می‌کنیم. در اینجا از تساوی $\omega_N^P=\omega_Q$ استفاده کرده‌ایم، این تساوی این واقعیت بحرانی را روشن می‌سازد که جمع داخلی در آخرین قسمت سمت راست در v باید Q-دوره‌ای باشد، یعنی از قراردادن $v+Q$ به‌جای Q، مقدار آن به‌علت واقعیت $\omega_Q^Q=1$ تغییر نمی‌کند. معنی این عبارت این است که تنها به‌محاسبهٔ این حاصلجمع به‌ازای $v=1,\dots,Q$ (وهر p) نیازداریم. بنابراین به‌ازای هر $p=1,\dots,P$ از Q-بردار $Z(p,\cdot)$، Q-برداری را محاسبه می‌کنیم که مؤلفه‌های آن اعداد

$$\sum_{q=1}^{Q} Z(p,q)\omega_Q^{(v-1)(q-1)} \qquad v=1,\dots,Q$$

باشند. یعنی، تبدیل گسستهٔ فوریهٔ مربوط به Q-بردارهای $Z(p,\cdot)$ را به‌ازای $p=1,\dots,P$، کلا با صرف تعداد $P\cdot Q^2=N\cdot Q$ عمل، محاسبه می‌کنیم.

اما، می‌توانستیم تبدیل $Z(p,\cdot)$ را بر روی $Z(p,\cdot)$ ذخیره کنیم. اما برای جلوگیری از توسعه‌های بیشتر، تبدیل $Z(p,\cdot)$ را در $Z_1(\cdot,p)$ ذخیره می‌کنیم، که Z_1 یك آرایهٔ دوبعدی‌است که اندازهٔ آن به‌جای (P,Q) برابر (Q,P) است.

با توجه به‌این امر، محاسبهٔ $\hat{z}_v$ به‌محاسبهٔ حاصلجمع

$$\hat{z}_v=\sum_{p=1}^{P} Z_1(v_Q,p)\omega_N^{(v-1)(p-1)} \qquad v=1,\dots,N$$

تبدیل می‌شود. در اینجا نماد v_Q معرف عدد صحیحی است بین ۱ و Q که به‌ازای آن $v-v_Q$ بر Q قابل قسمت است. بنابراین به‌ازای برخی از اعداد صحیح v' بین ۱ و P داریم

$$v=v_Q+Q(v'-1)$$

زیرا اگر بردار $\hat{z}$ را به‌شیوهٔ فورتن به‌یك آرایهٔ دوبعدی Z_0 با اندازهٔ (Q,P) نشان‌دهیم، داریم

$$\hat{z}_v=Z_0(v_Q,v') \qquad (۶۶.۶)$$

با توجه به‌این رابطه باید مقدار

$$Z_0(v_Q,v')=\sum_{p=1}^{P} Z_1(v_Q,p)\omega_N^{(v_Q-1+Q(v'-1))(p-1)} \qquad \text{به‌ازای جمیع مقادیر } v_Q \text{ و } v'$$

را محاسبه‌کنیم. در اینجا قسمت‌سمت‌راست، یك بسجمله‌ای است ازدرجهٔ کمتر از P برحسب کمیت $\omega_N^{v_Q-1+Q(v'-1)}=\omega_N^{v-1}$. این کمیت را می‌توان مرحله به‌مرحله به‌ترتیب سادهٔ محاسباتی زیر تولیدکرد.

$$
\begin{array}{l}
x := 1 \\
\text{for } \nu' = 1, \ldots, P, \text{ do:} \\
\quad \text{for } \nu_Q = 1, \ldots, Q, \text{ do:} \\
\qquad Z_0(\nu_Q, \nu') := \sum_{p=1}^{P} Z_1(\nu_Q, p) x^{p-1} \\
\quad x := x \cdot \omega_N
\end{array}
\qquad (۶۷.۶)
$$

البته در داخلی‌ترین حلقه، حاصلجمع را باید به‌وسیلهٔ ضرب‌تودرتو محاسبه نمود. در این صورت تعدادکل عمل مصروف دراین مرحله برابراست با $Q \cdot P^2 = NP$ (در صورتی که از تعداد N عمل ضرب لازم برای تولید بردهای مختلف لازم صرف‌نظر کنیم). بدین‌ترتیب در Z_0، تبدیل گسستهٔ فوریهٔ $\hat{\mathbf{z}}$ مربوط به $\mathbf{z}$ تنها با صرف تعداد $N(P+Q)$ عمل، در مقایسه با N^2 عمل که با روش ساده لازم بود، به‌دست آمده‌است.

حال اگر N، مساوی با حاصلضرب سه یا چند عدد صحیح بزرگتر ازیک، مثلا

$$N = P_1 \cdots P_m$$

باشد، آنگاه با استفاده از مرحلهٔ دوم (۶۷.۶) به طریق اندکی پیچیده‌تر، می‌توانیم تبدیل گسستهٔ فوریهٔ $\mathbf{z}$ راحتی با تعداد عمل کمتری محاسبه کنیم.

برای توضیح مطلب فوق، به‌اندکی ملامتگذاری نیاز داریم تا بتوانیم نشان دهیم که چگونه یك آرایهٔ یك بعدی راکه به‌شیوهٔ فورترن نشان داده شده، می‌توان به‌صورت یك آرایهٔ دو یا سه بعدی نمایش داد. اگر Z آرایه‌ای یك‌بعدی به‌طول N باشد، آنگاه آرایهٔ دوبعدی هم‌ارز آن به‌ابعاد $(A, N/A)$ را با Z^A وآرایهٔ سه‌بعدی هم‌ارز آن به‌ابعاد $(A, B, N/(AB))$ را با $Z^{A,B}$ نشان می‌دهیم. بدین‌ترتیب داریم

$$Z^{A,B}(a,b,c) = Z^A(a, b+B(c-1)) = Z^{AB}(a+A(b-1), c)$$
$$= Z(a+A(b-1+B(c-1)))$$

حال گیریم، مانندقبل، Z یك آرایهٔ یك بعدی مشتمل بر $\mathbf{z}$ و به‌ازای $k = 0, \ldots, m$، Z_k یك آرایهٔ یك‌بعدی مشتمل برتبدیل گسستهٔ فوریهٔ قسمتهای Z به‌شرح زیرباشد:

$$Z_k^A(\ ,c) = F_A Z^{BP}(c, \ ), \qquad c = 1, \ldots, BP \qquad (۶۸.۶)$$

که در آن

$$B := B_k := P_1 \cdots P_{k-1}$$
$$P := P_k \qquad (۶۸.۶ \text{ الف})$$
$$A := A_k = P_m \cdots P_{k+1}$$

باید توجه‌داشت که Z نقش Z_m را به‌خوبی ایفا می‌کند و Z_0 شامل $\hat{\mathbf{z}} = F_N \mathbf{z}$ است. برای

رسیدن از Z_k به Z_{k-1}، ازشکل اندکی بسط داده شدهٔ (۶۷.۶) که در آن B، P و A با رابطهٔ (۶۸.۶ الف) مشخص می‌شوند، استفاده می‌کنیم

$$
\begin{array}{l}
x := 1 \\
\text{for } p = 1, \ldots, P, \text{ do:} \\
\quad \text{for } a = 1, \ldots, A, \text{ do:} \\
\quad\quad \text{for } b = 1, \ldots, B, \text{ do:} \\
\quad\quad\quad Z_{k-1}^{A,P}(a,p,b) := \sum_{\pi=1}^{P} Z_k^{A,B}(a,b,\pi) \cdot x^{\pi-1} \\
\quad\quad x := x \cdot \omega_{AP}
\end{array}
\qquad (۶۹.۶)
$$

درحقیقت الگوریتم فوق مقادیر زیر را تولید می‌کند

$$Z_{k-1}^{A,P}(a,p,b) = \sum_{\pi=1}^{P} Z_k^{A,B}(a,b,\pi)\omega_{AP}^{[a-1+A(p-1)](\pi-1)}$$

از طرف دیگر رابطهٔ (۶۸.۶) مستلزم رابطهٔ

$$Z_k^{A,B}(\,.\,,b,\pi) = F_A Z^{B,P}(b,\pi,\,.\,) = \sum_{\alpha=1}^{A} Z^{B,P}(b,\pi,\alpha)\omega_A^{(\,.\,-1)(\alpha-1)}$$

است. بنابراین

$$Z_{k-1}^{A,P}(a,p,b) = \sum_{\pi=1}^{P}\sum_{\alpha=1}^{A} Z^{B,P}(b,\pi,\alpha)\omega_{AP}^{P(a-1)(\alpha-1)+[a-1+A(P-1)](\pi-1)}$$

اما اکنون به‌موجب تساوی $\omega_{AP}^{AP} = 1$ می‌توان به‌نمای طرف راست هرمضرب صحیحی از AP را افزود و این امر به‌نتیجهٔ زیرمنجرمی‌شود

$$Z_{k-1}^{A,P}(a,p,b) = \sum_{\pi=1}^{P}\sum_{\alpha=1}^{A} Z^{B,P}(b,\pi,\alpha)\omega_{AP}^{[a-1+A(p-1)][\pi-1+P(\alpha-1)]}$$

وبدین‌ترتیب ثابت می‌شود که Z_{k-1}، همان‌طوری که توسط (۶۹.۶) تولید شده، در رابطهٔ (۶۸.۶) (که در آن $k-1$ جایگزین k شده‌است) صدق می‌کند.

به‌ویژه، Z_0 شامل تبدیل گسستهٔ فوریهٔ $\mathbf{z}$ است. با m بار استفاده ازالگوریتم(۶۹.۶) با شروع از $Z_m = Z$ به Z_0 خواهد رسید.

زیربرنامهٔ فورترن زیر الگوریتم مذکور را پیاده می‌کند.

```
      SUBROUTINE FFT ( Z1, Z2, N, INZEE )
CONSTRUCTS THE DISCRETE FOURIER TRANSFORM OF  Z1 (OR Z2) IN THE COOLEY-
C  TUKEY WAY, BUT WITH A TWIST.
      INTEGER INZEE,N,    AFTER,BEFORE,NEXT,NEXTMX,NOW,PRIME(12)
      COMPLEX Z1(N),Z2(N)
C******  I N P U T  ******
C  Z1, Z2  COMPLEX N-VECTORS
C  N  LENGTH OF  Z1  AND  Z2
C  INZEE  INTEGER INDICATING WHETHER  Z1  OR  Z2  IS TO BE TRANSFORMED
C      = 1 , TRANSFORM  Z1
C      = 2 , TRANSFORM  Z2
```

```
C******  W O R K  A R E A S  ******
C  Z1, Z2  ARE BOTH USED AS WORKARRAYS
C******  O U T P U T  ******
C  Z1  OR  Z2  CONTAINS THE DESIRED TRANSFORM (IN THE CORRECT ORDER)
C  INZEE  INTEGER INDICATING WHETHER  Z1 OR Z2  CONTAINS THE TRANSFORM,
C     = 1 , TRANSFORM IS IN  Z1
C     = 2 , TRANSFORM IS IN  Z2
C******  M E T H O D  ******
C     THE INTEGER  N  IS DIVIDED INTO ITS PRIME FACTORS (UP TO A POINT).
C  FOR EACH SUCH FACTOR  P , THE P-TRANSFORM OF APPROPRIATE P-SUBVECTORS
C  OF  Z1 (OR Z2) IS CALCULATED IN  F F T S T P  AND STORED IN A SUIT-
C  ABLE WAY  IN  Z2 (OR Z1).  SEE TEXT FOR DETAILS.
C
      DATA NEXTMX,PRIME / 12, 2,3,5,7,11,13,17,19,23,29,31,37 /
      AFTER = 1
      BEFORE = N
      NEXT = 1
C
   10 IF ((BEFORE/PRIME(NEXT))*PRIME(NEXT) .LT. BEFORE) THEN
         NEXT = NEXT + 1
         IF (NEXT .LE. NEXTMX) THEN
                                        GO TO 10
         ELSE
            NOW = BEFORE
            BEFORE = 1
         END IF
      ELSE
         NOW = PRIME(NEXT)
         BEFORE = BEFORE/PRIME(NEXT)
      END IF
C
      IF (INZEE .EQ. 1)  THEN
         CALL FFTSTP( Z1, AFTER, NOW, BEFORE, Z2 )
      ELSE
         CALL FFTSTP( Z2, AFTER, NOW, BEFORE, Z1 )
      END IF
      INZEE = 3 - INZEE
      IF (BEFORE .EQ. 1)                RETURN
      AFTER = AFTER*NOW
                                        GO TO 10
      END


      SUBROUTINE FFTSTP ( ZIN, AFTER, NOW, BEFORE, ZOUT )
CALLED IN  F F T .
CARRIES OUT ONE STEP OF THE DISCRETE FAST FOURIER TRANSFORM.
      INTEGER AFTER,BEFORE,NOW,    IA,IB,IN,J
      REAL ANGLE,RATIO,TWOPI
      COMPLEX ZIN(AFTER,BEFORE,NOW),ZOUT(AFTER,NOW,BEFORE),    ARG,OMEGA,
     *                                                         VALUE
      DATA TWOPI / 6.2831 85307 17958 64769 /
      ANGLE = TWOPI/FLOAT(NOW*AFTER)
      OMEGA = CMPLX(COS(ANGLE),-SIN(ANGLE))
      ARG = CMPLX(1.,0.)
      DO 100 J=1,NOW
         DO 90 IA=1,AFTER
            DO 80 IB=1,BEFORE
               VALUE = ZIN(IA,IB,NOW)
               DO 70 IN=NOW-1,1,-1
   70             VALUE = VALUE*ARG + ZIN(IA,IB,IN)
   80          ZOUT(IA,J,IB) = VALUE
   90       ARG = ARG*OMEGA
  100 CONTINUE
                                        RETURN
      END
```

اگر N مساوی باحاصلضرب m عدد صحیح باشد،

$$N=P_1P_2\cdots P_m$$

آنگاه برنامه‌ای مانند برنامهٔ فوق امکان محاسبهٔ تبدیل $\hat{\mathbf{z}}=F_N\mathbf{z}$ را با تعداد عملیاتی برابر با

$$W=N(P_1+P_2+\cdots+P_m)$$

(به‌جای N^2) فراهم می‌سازد. از آنجاکه به‌ازای اعداد صحیح Q و R بزرگتر از ۱، بجز $Q=R=2$، تا مساوی $Q+R<QR$ برقرار است، این عدد W زمانی به‌کمترین مقدار می‌رسد که از هر عامل N بجز آن عواملی از ۲ که از ترکیب آنها ۴ حاصل می‌شود، عملا استفاده شود. بعلاوه

$$W/N=P_1+\cdots+P_m \quad \text{و} \quad \log N=\log P_1+\cdots+\log P_m$$

بنابراین

$$\frac{W}{N\log N}=\frac{P_1+\cdots+P_m}{\log P_1+\cdots+\log P_m}=\left(\sum_{j=1}^{m}\frac{P_j}{\log P_j}\log P_j\right)\Big/\sum_{j=1}^{m}\log P_j$$

این رابطه نشان می‌دهد که عدد $W/(N\log N)$ یك میانگین وزین اعداد $p_j/\log P_j$، $j=1,\ldots,m$، است. به‌آسانی دیده می‌شود که $P/\log_2 P$ به‌عنوان تابعی از عدد صحیح P در $P=3$ دارای مقدار حداقل ۱٫۸۸۹... و در $P=2$ و $P=4$، مقدار ۲ و در $P=10$ تنها برابر با ۳٫۰۱۰۰... است. بنابراین نامساوی

$$1.889\,N\log_2 N\leqslant W$$

برقرار است در حالی که حتی برای عوامل P_j به‌بزرگی ۱۰، مقدار W بزرگتر از $3.02\,N\log_2 N$ نیست.

درمواردی که بردار داده شدهٔ $\mathbf{z}$ حقیقی باشد، صرفه‌جویی بیشتری امکانپذیر است. زیرا تساوی

$$\hat{z}_{N+1-j}=\hat{z}_{j+1} \qquad (7.0.6)$$

برقرار است. به‌روابط (۶۴.۶) و (۶۵.۶) نگاه کنید.

زمانی که N عددی اول و یا N مساوی حاصلضرب اعداد صحیحی باشد که دو به‌دو نسبت به‌هم اول هستند، (ت س ف)های دیگری‌هم وجود دارند. به‌مقالهٔ وینگراد[۳۶] نگاه کنید.

تمرین

۱-۶.۶ مستقیماً با توجه به‌تعریف(۶۴.۶) ثابت کنید که (۷۰.۶) برای حالتی که $\mathbf{z}$ حقیقی باشد صادق است.

۲-۶.۶ با استفاده از برنامهٔ (ت س ف) (مثلا به‌ازای $N=81$) جوابهای خود را برای تمرینهای ۵.۶-۱ و ۵.۶-۲ امتحان کنید. [این کار شما را وادار خواهد ساخت که به‌تمام جزئیات موجود در رابطهٔ (۶۵.۶) دقیقاً توجه کنید!]

۶.۶-۳ با استفاده از برنامهٔ (ت س ف) ضرایب فوریهٔ $\hat{f}(j)$ را برای توابع زیر (به‌طور تقریبی) محاسبه کنید

الف. $f(x)=\sin ۳x$ ب. $f(x)=\sin(\pi x)$

برای مثال N را برابر با ۸۱ یا ۳۲۴ یا هرعدد دیگری که می‌خواهید بگیرید. چرا با افزایش $|j|$ ضرایب فوریه برای $f(x)=\sin(\pi x)$ به‌سرعت کاهش نمی‌یابد؟

۶.۶-۴ برنامهٔ (ت س ف) تیلر[1] برای حالت خاص ۳ر۴ $=N$ را درز بگیرید و تاآنجاکه می‌توانید درمحاسبات و مقدار حافظه صرفه‌جویی کنید.

۶.۶-۵ باافزودن عبارتهای مخصوص به‌جای دامنهٔ حلقهٔ DO روی IB، مثلا درموردی که NOW = ۲, ۳ یا ۴، زیربرنامهٔ FFTSTP را بهتر کنید.

۶.۶-۶ درباب استفاده از (ت س ف)، برای محاسبهٔ حاصلجمع مثلثاتی

$$t(x)=\sum_{j=-n}^{n} a(j)e^{ijx}$$

درنقاط $x_j=\alpha+۲\pi j/(۲n+۱)$، $j=۰,\ldots,n$، به‌ازای یك مقدار ثابت α دربازهٔ $[۰, ۲\pi/(۲n+۱)]$ بحث کنید.

۶.۶-۷ با استفاده از (ت س ف) بسجمله‌ای درونیاب مثلثاتی را در $N=۲n+۱$ نقطهٔ، $x_j=۲\pi j/N, j=۰,\ldots,N-۱$ برای موج‌مربعی $f(x)=\text{signum}(\sin x)$ و با استفاده از $N=۳۵$ بسازید. سپس باردیگر از (ت س ف) برای محاسبهٔ بسجمله‌ای درونیاب در ۱۰۵ نقطهٔ، $y_j=۲\pi j/۱۰۵$ و $j=۰,\ldots,۱۰۴$ استفاده کنید. (راهنمایی: ازتمرین ۶.۶-۶ استفاده کنید).

۶.۶-۸ با استفاده از (ت س ف) تقریبی بسازید برای طیف یك تابع $f(x)$ که‌مقادیرش درنقاط $x_j=۲\pi_j/N$، $j=۰,\ldots,N-۱$ و با مثلا $N=۱۲۸$، ازیك مولد عددی (شبه) تصادفی[2] که اعدادی با توزیع یکنواخت بین ۰و۱ تولید می‌کند، به‌دست آمده‌باشند. نتیجه را با طیف تابعی که درتمرین ۶.۵-۲ بررسی کردیم مقایسه کنید.

۶.۶-۹ با استفاده‌ازتمرین ۶.۶-۸ درچگونگی استفاده از(ت س ف) برای بازیابی‌مقادیر $f(x_j)$، $j=۰,\ldots,N-۱$، ازیك تابع «هموار» ۲π-دوره‌ای $f(x)$ ازداده‌های‌مفروض $f(x_j)+\varepsilon_j$، به‌ازای کلیهٔ مقادیر j، که درآن ε_j نوفه‌ای است با توزیع یکنواخت، بحث کنید.

۶.۶-۱۰ نشان دهیدکه $F_N^{-۱}\mathbf{z}=۱/N(\overline{F_N\overline{\mathbf{z}}})$. (یعنی N-بردار $\mathbf{z}$ را ازتبدیل‌مجردفوریهٔ

---

1. Tailor 2. (pseudo) random

آن $\hat{\mathbf{z}}=F_N\mathbf{z}$ به دست آورید، بدین‌ترتیب که (الف) با تبدیل کلیهٔ درایه‌های $\hat{\mathbf{z}}$ به‌فرد و زوجهای همتافتشان و سپس (ب) با تشکیل تبدیل گسستهٔ فوریهٔ بردار حاصلهٔ $\hat{\mathbf{z}}$ و آنگاه (پ) با تقسیم هردرایهٔ بردارحاصل بر N.)

۱۱-۶.۶ چگونگی استفاده از (ت س ف) را برای تشکیل بسجمله‌ای درونیابی از درجهٔ تابیشتر از n در نقاط چبیشف (۱۸.۶) برای داده‌های مفروض شرح دهید (راهنمایی: بسجمله‌ای درونیاب را ازیك ترکیب خطی از $n+1$ بسجمله‌ای چبیشف، T_0، ...، T_n، بااستفاده‌از(۶.۹)بسازید.ارزیابی‌بعدی، البته، ازراه **تابع چبیشف** صورت خواهد گرفت).

۷.۶ تقریب به‌وسیلهٔ بسجمله‌ای ـ تکه‌ای

یك مثال ساده و آشنا از تقریب‌زدن به‌وسیلهٔ بسجمله‌ای‌ـتکه‌ای، درونیابی خطی درجدولی است از مقادیر $f(x_i)$، $i=1,\dots,N+1$، که در آن

$$a=x_1<x_2<\cdots<x_{N+1}=b$$

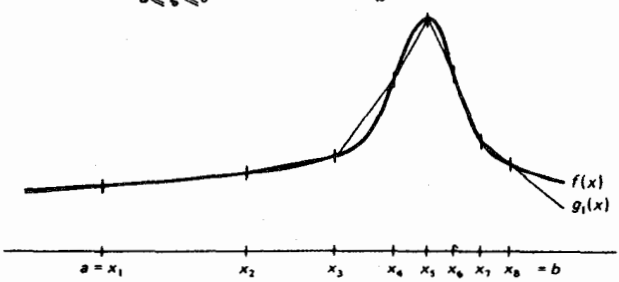
دراینجا $f(x)$ دریك نقطهٔ $\bar{x}$، با تعیین بازهٔ $[x_k, x_{k+1}]$ که شامل $\bar{x}$ است و گرفتن

$$p_1(\bar{x})=f(x_k)+f[x_k, x_{k+1}](\bar{x}-x_k)$$

به‌عنوان تقریبی برای $f(\bar{x})$، تقریب‌زده می‌شود. درنتیجه، $f(x)$ توسط تابع «خط‌شکسته» یا تابع خطی‌ـتکه‌ای[1] $g_1(x)$ روی $[a, b]$ (به‌شکل۹.۶ نگاه‌کنید) با **نقاط انفصال**[2] $x_2, \dots, x_N$، که $f(x)$ را در نقاط $x_1, \dots, x_{N+1}$ درونیابی می‌کند، تقریب زده شده است. ازمثال۲.۶که‌برای هریك از زیر بازه‌های $[x_k, x_{k+1}]$، $k=1,\dots,N$، به‌کار برده شده نتیجه می‌شودکه به‌ازای جمیع مقادیر $x\in[a, b]$ داریم

$$|f(x)-g_1(x)|\leqslant\frac{1}{8}\max_k\left\{(\Delta x_k)^2\max_{x_k\leqslant\xi\leqslant x_{k+1}}|f''(\xi)|\right\}$$

$$\leqslant\max_{a\leqslant\xi\leqslant b}|f''(\xi)|\frac{1}{8}(\max_k\Delta x_k)^2 \qquad (۷۱.۶)$$



شکل۹.۶ درونیابی به‌وسیلهٔ خط شکسته

1. piecewise-linear 2. breakpoints

به‌شرطی که $f(x)$ در بازهٔ $[a, b]$ دو بار مشتقپذیر باشد. باید توجه‌داشت که با کوچك نمودن Δx_k، به‌ازای جمیع مقادیر k، می‌توان خطای درونیابی را تا حد ممکن تقلیل داد. بعدهم توجه می کنید که این گونه افزایش در نقاط درونیابی، کار کردن بعدی با $g_1(x)$ را پیچیده‌تر نمی کند، زیرا $g_1(x)$ «به‌طور محلی» تابعی است بسیار ساده.

با استفاده از یك تابع بسجمله‌ای-تکه‌ای $g_2(x)$ از درجهٔ بزرگتر از یك، به‌جای تابع خطی-تکه‌ای $g_1(x)$، می‌توان برای تابع $f(x)$ تقریبی تولید کرد که جملهٔ خطای مربوط به $\max_k \Delta x_k$ در آن شامل توان $(r+1)$ام باشد؛ از این‌رو با کوچك شدن $\max \Delta x_k$، این جملهٔ خطا نسبت به‌خطای (۷۱.۶) برای درونیابی خطی-تکه‌ای، سریعتر به‌سمت صفر میل می‌کند. مخصوصاً تقریب درجهٔ سوم-تکه‌ای[1] متداولتر است. اکنون چندین نمونه از درونیابی درجهٔ سوم-تکه‌ای را مورد بحث قرار می‌دهیم.

گیریم $f(x)$ تابعی حقیقی مقدار و معین در یك بازهٔ $[a, b]$ باشد. می‌خواهیم یك تابع (بسجمله‌ای) درجهٔ سوم-تکه‌ای $g_3(x)$ بسازیم که $f(x)$ را در نقاط $x_1, \ldots, x_{N+1}$ با

$$a = x_1 < x_2 < \cdots < x_{N+1} = b \qquad (۷۲.۶)$$

درونیابی کند؛ همان گونه که با درونیابی تکه‌ای - خطی عمل کردیم، نقاط درونیابی داخل $x_2, \ldots, x_N$ را در نقاط انفصالی $g_3(x)$ انتخاب می‌کنیم، یعنی در هر بازهٔ $[x_i, x_{i+1}]$، $g_3(x)$ را مانند یك بسجمله‌ای درجهٔ سوم $P_i(x)$، $i = 1, \ldots, N$، می‌سازیم.

برای تسهیل استفاده از $g_3(x)$ در محاسبات بعدی هر قطعهٔ درجهٔ سوم $P_i(x)$ از $g_3(x)$ را به‌صورت زیر می‌نویسیم

$$P_i(x) = c_{1,i} + c_{2,i}(x - x_i) + c_{3,i}(x - x_i)^2 + c_{4,i}(x - x_i)^3 \qquad (۷۳.۶)$$

زمانی که ضرایب $c_{j,i}$، $i = 1, \ldots, N$ و $j = 1, \ldots, 4$، معلوم باشند، آنگاه برنامهٔ تابع فورترن موسوم به PCUBIC $g_3(x)$ را به‌ازای هر نقطهٔ خاص $x = \bar{x}$ به‌نحو چشمگیری محاسبه می‌نماید.

```
      REAL FUNCTION PCUBIC ( XBAR, XI, C, N )
C  RETURNS THE VALUE AT  XBAR  OF THE PIECEWISE CUBIC FUNCTION ON  N
C  INTERVALS WITH BREAKPOINT SEQUENCE  XI  AND  COEFFICIENTS  C .
      INTEGER N,   I,J
      REAL C(4,N),XBAR,XI(N+1),   DX
      DATA I /1/
      IF (XBAR .GE. XI(I)) THEN
         DO 10 J=I,N
            IF (XBAR .LT. XI(J+1))       GO TO 30
   10    CONTINUE
         J = N
      ELSE
         DO 20 J=I-1,1,-1
            IF (XBAR .GE. XI(J))         GO TO 30
   20    CONTINUE
         J = 1
      END IF
   30 I = J
      DX = XBAR - XI(I)
      PCUBIC = C(1,I) + DX*(C(2,I) + DX*(C(3,I) + DX*C(4,I)))
                                          RETURN
      END
```

1. piecewise-cubic approximation

اکنون به‌تعیین تابع درونیابی درجهٔ سوم-تکه‌ای $g_۳(x)$ می‌پردازیم. از آنجایی که می‌خواهیم

$$g_۳(x_i)=f(x_i) \qquad i=۱,\dots,N+۱$$

باید داشته باشیم

$$P_i(x_i)=f(x_i) \qquad P_i(x_{i+۱})=f(x_{i+۱}) \qquad i=۱,\dots,N \qquad (۷۴.۶)$$

باید توجه کرد که روابط (۷۴.۶) ایجاب می‌کنند که داشته باشیم

$$P_{i-۱}(x_i)=P_i(x_i) \qquad i=۲,\dots,N$$

لذا پیوسته بودن $g_۳(x)$ در بازهٔ $[a,b]$ تضمین می‌شود.

با توجه به‌قضیهٔ ۱.۲ یا ۴.۲ خاطرنشان می‌سازیم که همیشه می‌توانیم تابع مفروضی را در چهار نقطه توسط یك بسجمله‌ای درجهٔ سوم درونیابی کنیم. تا اینجا لازم بوده که هر یك‌ازقطعه‌های‌درجهٔ سوم $P_i(x)$، تابع $f(x)$ را تنها در دونقطه درونیابی کند وازاین‌رو در انتخاب $P_i(x)$ مقداری آزادی عمل داشته‌ایم و تفاوت روشهای مختلف درونیابی تنها درچگونگی استفاده از این آزادی عمل است.

در درونیابی **درجهٔ سوم-تکه‌ای هرمیت**، $P_i(x)$ چنان تعیین می‌شود که $f(x)$ در نقاط x_i، x_i، $x_{i+۱}$، و $x_{i+۱}$ درونیابی شود، یعنی چنان‌که روابط زیرنیز برقرار باشند

$$P_i'(x_i)=f'(x_i) \qquad P_i'(x_{i+۱})=f'(x_{i+۱}) \qquad i=۱,\dots,N \qquad (۷۵.۶)$$

دراین حال از فرمول نیوتن (۳۲.۲) نتیجه می‌شود که به‌ازای $i=۱,\dots,N$

$$P_i(x)=f(x_i)+f[x_i,x_i](x-x_i)+f[x_i,x_i,x_{i+۱}](x-x_i)^۲$$
$$+f[x_i,x_i,x_{i+۱},x_{i+۱}](x-x_i)^۲(x-x_{i+۱})$$

چون داریم $(x-x_{i+۱})=(x-x_i)+(x_i-x_{i+۱})$ نتیجه می‌شود که

$$P_i(x)=f(x_i)+f'(x_i)(x-x_i)+(f[x_i,x_i,x_{i+۱}]-f[x_i,x_i,x_{i+۱},x_{i+۱}]\Delta x_i)$$
$$\times(x-x_i)^۲+f[x_i,x_i,x_{i+۱},x_{i+۱}](x-x_i)^۳$$

که در آن $\Delta x_i=x_{i+۱}-x_i$، و از روی آن ضرایب $c_{۱,i}$، $c_{۲,i}$، $c_{۳,i}$ و $c_{۴,i}$ را می‌توان به‌طور مستقیم به‌دست آورد. با استفاده از فرمولهای اختصاری

$$f_i=f(x_i) \qquad s_i=f'(x_i) \qquad i=۱,\dots,N+۱ \qquad (۷۶.۶)$$

داریم

$$c_{۱,i}=f_i \qquad c_{۲,i}=s_i$$

$$c_{۳,i}=f[x_i, x_i, x_{i+۱}]-f[x_i, x_i, x_{i+۱}, x_{i+۱}]\Delta x_i$$

$$=\frac{f[x_i, x_{i+۱}]-s_i}{\Delta x_i}-c_{۴,i}\ \Delta x_i \qquad (۷۷.۶)$$

$$c_{۴,i}=\frac{f[x_i, x_{i+۱}, x_{i+۱}]-f[x_i, x_i, x_{i+۱}]}{\Delta x_i}$$

$$=\frac{s_{i+۱}+s_i-۲f[x_i, x_{i+۱}]}{(\Delta x_i)^۲}$$

با ذخیره شدن f_i در $c_{۱,i}$ و s_i در $c_{۲,i}$ به‌ازای $i=۱,\ldots,N+۱$، در زیر برنامهٔ فورترن زیر برای محاسبهٔ $c_{۳,i}$ و $c_{۴,i}$، $i=۱,\ldots,N$ از رابطهٔ (۷۷.۶) استفاده شده.

```
      SUBROUTINE CALCCF ( XI, C, N )
      INTEGER N,   I
      REAL C(4,N+1),XI(N+1),   DIVDF1,DIVDF3,DX
C******  I N P U T  ******
C  XI(1), ..., XI(N+1)  STRICTLY INCREASING SEQUENCE OF BREAKPOINTS.
C  C(1,I), C(2,I), VALUE AND FIRST DERIVATIVE AT  XI(I), I=1,...,N+1,
C     OF THE PIECEWISE CUBIC FUNCTION.
C******  O U T P U T  ******
C  C(1,I), C(2,I), C(3,I), C(4,I)  POLYNOMIAL COEFFICIENTS OF THE FUNC-
C        TION ON THE INTERVAL  (XI(I), XI(I+1))    I=1,...,N .
C
      DO 10 I=1,N
         DX = XI(I+1) - XI(I)
         DIVDF1 = (C(1,I+1) - C(1,I))/DX
         DIVDF3 = C(2,I) + C(2,I+1) - 2.*DIVDF1
         C(3,I) = (DIVDF1 - C(2,I) - DIVDF3)/DX
   10    C(4,I) = DIVDF3/(DX*DX)
                                        RETURN
      END
```

□ **مثال ۱۵.۶:** مسئلهٔ درونیابی مثال ۴.۲ را با استفاده از درونیابی درجهٔ سوم‌تکه‌ای هرمیت حل کنید، یعنی به‌ازای $N=۲, ۴, \ldots, ۱۶$ بگیرید

$$x_i=\frac{(i-۱)۱۰}{N}-۵ \qquad i=۱,\ldots,N+۱$$

و

$$f(x)=(۱+x^۲)^{-۱}$$

را دراین نقاط درونیابی کنید و مانند قبل ماکسیمم خطای درونیابی‌را در بازهٔ $[-۵, ۵]$ براورد کنید.

برنامهٔ فورترن زیر این مسئله را حل می‌کند:

```
C  PROGRAM FOR EXAMPLE 6.15 .
      INTEGER I,J,K,N
      REAL C(4,1~),ERRMAX,H,X(1~),Y
C     PIECEWISE CUBIC HERMITE INTERPOLATION AT EQUALLY SPACED POINTS
C     TO THE FUNCTION
            F(Y) = 1./(1. + Y*Y)
C
      PRINT 600
  600 FORMAT('1   N',5X,'MAXIMUM ERROR')
      DO 40 N=2,16,2
         H = 10./FLOAT(N)
         DO 10 I=1,N+1
            X(I) = FLOAT(I-1)*H - 5.
            C(1,I) = F(X(I))
C           C(2,I) = F'(X(I))
   10       C(2,I) = -2.*X(I)*C(1,I)**2
         CALL CALCCF ( X, C, N )
C                 ESTIMATE MAXIMUM INTERPOLATION ERROR ON (-5,5).
         ERRMAX = 0.
         DO 30 I=1,101
         Y =.1*I - 5.
            ERRMAX = MAX(ERRMAX, ABS(F(Y)-PCUBIC(Y,X,C,N)))
   30    CONTINUE
   40    PRINT 640, N,ERRMAX
  640 FORMAT(I5,E18.7)
                                        STOP
      END
```

نتایج کامپیوتری برای مثال ۱۵.۶

N	MAXIMUM ERROR
2	4.9188219E − 01
4	2.1947326E − 01
6	9.1281965E − 02
8	3.5128250E − 02
10	1.2705882E − 02
12	4.0849234E − 03
14	1.6011164E − 03
16	1.6953134E − 03

درمقایسه با درونیابی به‌وسیلهٔ بسجمله‌ای (مثال ۴.۲ را ببینید)، دراین‌حال باافزایش N، ماکسیمم خطا به‌نحو محسوسی کاهش می‌یابد. □

براورد خطا در درونیابی درجهٔ سوم-تکه‌ای هرمیت به‌آسانی صورت می‌گیرد. چون به‌ازای $x \in [x_i, x_{i+1}]$، داریم $g_3(x) = P_i(x)$، که دراینجا $P_i(x)$ تابع $f(x)$ را در نقاط x_i، x_{i+1}، x_{i+1}، x_{i+1}، درونیابی می‌کند، از رابطهٔ (۳۷.۲) نتیجه می‌شود که به‌ازای $x \in [x_i, x_{i+1}]$ داریم

$$f(x) - g_3(x) = f[x_i, x_i, x_{i+1}, x_{i+1}, x](x-x_i)^2(x-x_{i+1})^2$$

$$= \frac{1}{4!} f^{(4)}(\xi_x)(x-x_i)^2(x-x_{i+1})^2 \qquad \xi_x \in (x_i, x_{i+1})$$ به‌ازای برخی از

به‌شرط آنکه $f(x)$ چهاربار به‌طور پیوسته مشتقپذیر باشد، بعلاوه

$$\max_{x \in [x_i, x_{i+1}]} |(x-x_i)^2(x-x_{i+1})^2| = \left(\frac{1}{2}\Delta x_i\right)^4 \leqslant \frac{(\max_j \Delta x_j)^4}{16}$$

بنابراین به‌ازای $a \leqslant x \leqslant b$:

$$|f(x)-g_3(x)| \leqslant \frac{1}{384} \max_k \left\{ (\Delta x_k)^4 \max_{x_k \leqslant \xi \leqslant x_{k+1}} |f^{(4)}(\xi)| \right\}$$

$$\leqslant \max_{\xi \in [a,b]} |f^{(4)}(\xi)| \frac{(\max_i \Delta x_i)^4}{384} \qquad (۷۸.۶)$$

درونیابی‌درجهٔ سوم‌ـتکه‌ای هرمیت، به‌اطلاعاتی درباب $f'(x)$ نیاز دارد. درعمل، اغلب به‌دست‌آوردن اعداد موردنیاز $f'(x_i)$، $i=1,\dots,N+1$، اگر غیرممکن‌نباشد دشوار است. در چنین حالتی برای محاسبهٔ s_i، تقریب مناسبی برای $f'(x_i)$، $i=1,\dots,N+1$، درنظر می‌گیرند. بنابراین در درونیابی درجهٔ **سوم‌ـتکه‌ای بسل**، به‌جای $s_i=f'(x_i)$ از رابطهٔ

$$s_i=\frac{\Delta x_{i-1} f[x_i, x_{i+1}]+\Delta x_i f[x_{i-1}, x_i]}{\Delta x_{i-1}+\Delta x_i} \qquad (۷۹.۶)$$

استفاده می‌کنند، اما مانند قبل، با تعیین ضرایب $c_{j,i}$ برای تکه‌های درجهٔ سوم طبق رابطهٔ (۷۷.۶) به‌گونهٔ دیگرعمل می‌کنند. باید توجه‌کردکه در به‌دست آوردن اعدادی برای مشتقهای کرانه‌ای $g_3(x)$، s_1 و s_{n+1}، تساوی (۷۹.۶) به‌دونقطهٔ اضافی x_0 و x_{N+2} نیاز دارد. این نقاط به‌گونه‌ای، مثلا به‌صورت

$$x_0=x_2 \qquad x_{N+2}=x_{N-1}$$

انتخاب می‌شوند. یا متناظر با انتخاب $x_0=a$ و $x_{N+2}=b$، اعداد

$$s_1=f'(a) \qquad s_{N+1}=f'(b) \qquad (۸۰.۶)$$

را به‌کارمی‌برند به‌شرط آنکه این اعداد در دسترس باشند. بازیك امکان دیگرآن‌است که s_1 و s_{N+1} به‌طریقی انتخاب شوندکه شرایط «پایانی‌آزاد»[1]

$$g_3''(a)=g_3''(b)=0 \qquad (۸۱.۶)$$

برقرار باشند.

اگر استفاده از $f_i=f(x_i)$، $i=1,\dots,N+1$، رادررابطهٔ(۷۷.۶) ادامه‌دهیم، آنگاه اعداد خاص‌انتخابی s_i، $i=1,\dots,N+1$، هرچه باشند تابع‌درجهٔ سوم‌ـتکه‌ای $g_3(x)$، تابع $f(x)$ را درنقاط $x_1,\dots,x_{N+1}$ درونیابی می‌کند. بعلاوه $g_3(x)$ گذشته از پیوسته بودن در بازهٔ $[a,b]$، به‌طور پیوسته مشتقپذیر نیز هست، زیرا (۷۷.۶) ایجاب می‌کندکه داشته باشیم

$$P'_{i-1}(x_i)=s_i=P'_i(x_i) \qquad i=2,\dots,N$$

---

1. free_end

همان گونه که اکنون نشان خواهیم داد، همواره ممکن است اعداد $s_1, \ldots, s_{N+1}$ را چنان تعیین کرد که $g_3(x)$ حاصل، حتی دوبار به‌طور پیوسته مشتقپذیر باشد. این روش تعیین $g_3(x)$ به **درونیابی درجهٔ سوم قلمی**[1] معروف است. در این حالت اطلاق نام «قلمی» به‌درونیاب $g_3(x)$ از این جهت است که نمودار آن، تقریب‌زدن وضعیتی است که «قلم ظریف و انعطاف‌پذیر» یک نقشه‌کش، درصورت اجبار به گذشتن از نقاط $\{x_i, f_i\}$، $i = 1, \ldots, N+1$، اختیار می‌کند.

شرط اینکه $g_3(x)$ دوبار به‌طور پیوسته مشتقپذیر باشد، با شرط

$$P''_{i-1}(x_i) = P''_i(x_i) \qquad i = 2, \ldots, N$$

یا با توجه به‌رابطهٔ (۷۳.۶) با شرط

$$2c_{3,i-1} + 6c_{4,i-1}\Delta x_{i-1} = 2c_{3,i} \qquad i = 2, \ldots, N$$

هم‌ارز است. از این‌رو با داشتن (۷۷.۶) می‌خواهیم داشته باشیم

$$\frac{2(f[x_{i-1}, x_i] - s_{i-1})}{\Delta x_{i-1}} + 4c_{4,i-1}\Delta x_{i-1} = \frac{2(f[x_i, x_{i+1}] - s_i)}{\Delta x_i} - 2c_{4,i}\Delta x_i$$
$$i = 2, \ldots, N$$

اگر برای بیان $c_{4,i-1}$ و $c_{4,i}$ برحسب f_jها و s_jها، از رابطهٔ (۷۷.۶) استفاده، و آن‌را ساده کنیم خواهیم داشت

$$(\Delta x_i)s_{i-1} + 2(\Delta x_{i-1} + \Delta x_i)s_i + (\Delta x_{i-1})s_{i+1}$$
$$= 3(f[x_{i-1}, x_i]\Delta x_i + f[x_i, x_{i+1}]\Delta x_{i-1}) \qquad i = 2, \ldots, N$$
(۸۲.۶)

این، یک دستگاه $N-1$ معادلهٔ خطی است با $N+1$ مجهول $s_1, \ldots, s_{N+1}$. اگر s_1 و s_{N+1} را به‌طریقی، مثلا به‌کمک رابطهٔ (۷۹.۶) یا رابطهٔ (۸۰.۶)، انتخاب کنیم می‌توانیم دستگاه (۸۲.۶) را نسبت به $s_2, \ldots, s_N$ با روش حذف گاوس (فصل ۴ را ببینید) حل کنیم. دراین‌صورت ماتریس ضرایب دستگاه (۸۲.۶) نافذ قطری (سطراً مؤکداً) و بنابراین (تمرین ۶.۴–۳ را ببینید) وارونپذیر خواهد بود، و لذا دستگاه (۸۲.۶) یک‌جواب منحصر‌به‌فرد خواهد داشت. وقتی $s_2, \ldots, s_N$، یعنی جوابهای دستگاه خطی (۸۲.۶) را به‌دست آوردیم، برای ساختن ضرایب بسجمله‌ای محلی تابع درجهٔ سوم قلمی درونیاب، از آنها و شیبهای کرانی s_1 و s_{N+1} دربرنامهٔ CALCCF استفاده می‌کنیم.

می‌توان نشان داد (مثلا دبور[۴۰ ;V(۶)] را ببینید) که خطا در درونیاب درجهٔ

---

1. cubic spline interpolation

سوم‌قلمی، به‌ازای $a \leqslant x \leqslant b$، در رابطهٔ

$$|f(x)-g_3(x)| \leqslant \max_{\xi \in [a,b]} |f^{(4)}(\xi)| \frac{5(\max_i \Delta x_i)^4}{384} \qquad (۸۳.۶)$$

صدق می‌کند. این خطای کرانی تنها پنج برابر خطای کرانی (۷۸.۶) برای درونیابی درجهٔ سوم هرمیت‌است، اگرچه در درونیابی درجهٔ سوم هرمیت از دوبرابر اطلاعات در بارهٔ تابع $f(x)$، یعنی مقادیر $f'(x_i)$، $i = 2, \ldots, N$، و نیز مقادیر تابع استفاده می‌شود. این بدان معنی‌است‌که شیبهای $g_3'(x_i)$ مربوط به‌تابع درونیاب قلمی باید برای شیبهای متناظر $f'(x_i)$ مربوط به $f(x)$ تقریبهای خوبی باشند. می‌توان نشان داد (برای مثال کتاب دبور [۴۰؛ V(۱۱) - (۱۲)] را ببینید) که به‌ازای $a \leqslant x \leqslant b$:

$$|f'(x)-g_3'(x)| \leqslant \max_{\xi \in [a,b]} |f^{(4)}(\xi)| \frac{(\max_i \Delta x_i)^3}{24} \qquad (۸۴.۶)$$

درمورد یك دنبالهٔ یکنواخت از نقاط یعنی $x_i = x_0 + ih$، به‌ازای جمیع مقادیر i، حتی داریم
به‌ازای $i = 2, \ldots, N$:

$$|f'(x_i)-g_3'(x_i)| \leqslant \max_{\xi \in [a,b]} |f^{(5)}(\xi)| h^4/60 \qquad (۸۵.۶)$$

کـه این مسئله، موجب تعمیم درونیابی تابـع درجـهٔ سوم قلمی به‌عنوان وسیله‌ای برای **مشتقگیری عددی** (فصل۷ را ببینید) شده‌است.

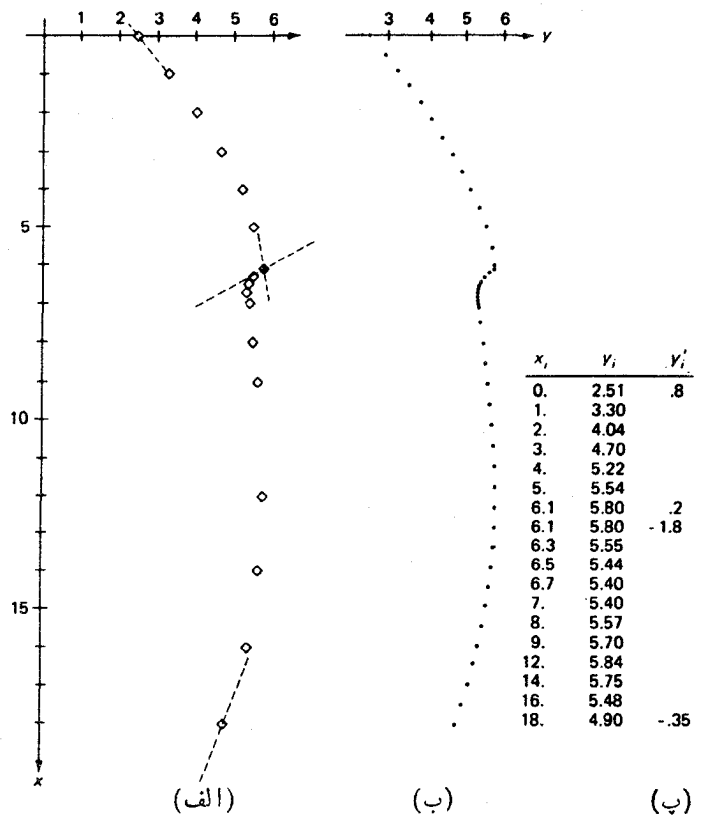
در زیر برنامهٔ **فورترن** SPLINE (قلمی) زیر روش حذف گاوس پذیرفتـه شده است تـا از امتیاز مشخصهٔ سه قطری بودن ماتریس‌ضرایب (۸۲.۶) (الگوریتـم ۳.۴ را ببینید) برای محاسبهٔ $c_{2,i} = s_i'$، $i = 2, \ldots, N$، به‌عنوان جواب دستگاه (۸۲.۶) استفاده کننـد، بـا فـرض اینـکـه اعـداد، $c_{1,i} = f_i$، $i = 1, \ldots, N+1$، و $c_{2,1} = s_1'$ $c_{2,N+1} = s_{N+1}'$ داده شده باشند.

```
      SUBROUTINE SPLINE ( XI, C, N )
      PARAMETER NP1MAX=50
      INTEGER N,   M
      REAL C(4,N+1),XI(N+1),   D(NP1MAX),DIAG(NP1MAX),G
C****** I N P U T ******
C  XI(1), ..., XI(N+1)  STRICTLY INCREASING SEQUENCE OF BREAKPOINTS
C  C(1,I), C(2,I), VALUE AND FIRST DERIVATIVE AT  XI(I), I=1,...,N+1,
C     OF THE CUBIC SPLINE.
C****** O U T P U T ******
C  C(1,I), C(2,I), C(3,I), C(4,I)  POLYNOMIAL COEFFICIENTS OF THE SPLINE
C        ON THE INTERVAL  (XI(I), XI(I+1)) ,  I=1,...,N .
      DATA DIAG(1),D(1) /1.,0./
      DO 10 M=2,N+1
         D(M) = XI(M) - XI(M-1)
   10    DIAG(M) = (C(1,M) - C(1,M-1))/D(M)
      DO 20 M=2,N
         C(2,M) = 3.*(D(M)*DIAG(M+1) + D(M+1)*DIAG(M))
   20    DIAG(M) = 2.*(D(M) + D(M+1))
      DO 30 M=2,N
         G = -D(M+1)/DIAG(M-1)
```

```
      DIAG(M) = DIAG(M) + G*D(M-1)
30    C(2,M) = C(2,M) + G*C(2,M-1)
   DO 40 M=N,2,-1
40    C(2,M) = (C(2,M) - D(M)*C(2,M+1))/DIAG(M)
                                    RETURN
   END
```

□ **مثال ۱۶.۶: تقریب‌زدن یك منحنی طرح توسط یك تابع درجهٔ سوم قلمی** یك منحنی طرح، مثلاقسمتی ازمقطع عرضی یك درب ماشین، به گونه‌ای که درشکل۱۰.۶ (الف) آمده، داده شده‌است. این منحنی در $x = 6.1$ یك ناپیوستگی شیبی دارد. همان‌گونه که در شکلهای ۱۰.۶ (الف) و (ج) نشان داده شده است، اندازه‌گیریهایی انجام گرفته و شیبهای انتهایی به‌طور نموداری براورد شده‌اند. مسئلهٔ ما پیداکردن یك تابع $s(x)$ است که به‌داده‌های مفروض ببرازد و «هموارجلوه‌کند».

با درونیابی درجهٔ سوم قلمی برای داده‌های مفروض واستفاده از دوتابع‌درجهٔ سوم قلمی که به‌طور پیوسته به‌هم وصل‌شده‌اند ولی در $x = 6.1$ شیبهای متفاوتی‌دارند،جواب این مسئله به‌آسانی به‌دست می‌آید. این‌کار در برنامهٔ فورترن زیر با استفاده از زیربرنامه‌های SPLINE و CALCCF که قبلا موردبحث قرار گرفت، انجام می‌گیرد.



x_i	y_i	y_i'
0.	2.51	.8
1.	3.30	
2.	4.04	
3.	4.70	
4.	5.22	
5.	5.54	
6.1	5.80	.2
6.1	5.80	-1.8
6.3	5.55	
6.5	5.44	
6.7	5.40	
7.	5.40	
8.	5.57	
9.	5.70	
12.	5.84	
14.	5.75	
16.	5.48	
18.	4.90	-.35

شکل۱۰.۶ تقریب درجهٔ سوم قلمی برای یك منحنی طرح.

این برنامه داده‌های تا $x=۶٫۱$، به‌انضمام دوشیب انتهایی مفروض، را می‌دهد وضرایب بسجمله‌ای محاسبه شده مربوط به‌شش تکهٔ نخست بسجمله‌ای را در

$$C(J, I), J=۱, \cdots, ۴ \qquad I=۱, \cdots, ۶$$

ذخیره می‌کند. سپس داده‌ها از $x=۶٫۱$ تا $x=۱۸$ همراه بـا دوشیب انتهایی خوانده می‌شوند و با استفادهٔ مجدد از برنامه‌های SPLINE وCALCCF، ضرایب

$$C(J, I), J=۱, \cdots, ۴ \qquad I=۷, \cdots, ۱۶$$

متعلق به ۱۰ تکهٔ بسجمله‌ای باقیمانده به‌دست می‌آیند. سرانجام تابع درجـهٔ سوم-تکه‌ای محاسبه شدهٔ $s(x)$ بـا به‌کارگیری برنامهٔ PCUBIC، به‌ازای مقادیر مختلف x ارزیابی می‌شود. برخی از این مقادیر درشکل۱۰.۶ (ب)، مشخص شده‌اند. حتی بدون ناپیوستگی شیبی، درونیابی بسجمله‌ای برای این داده‌ها تقریبی «ناهموار» یعنی تقریبی نوسانی تولید می‌کند، زیرا این ناحیهٔ با انحنای نسبتاً زیاد در نزدیك نقطهٔ $x=۶٫۱$، توسط یك قسمت مبهم ونسبتاً صاف دنبال می‌شود (تمرین۶.۷–۲ را ببینید).

برنامهٔ فورترن برای درونیابی قلمی درجهٔ سوم مثال (۱۶.۶)

```
C  PROGRAM FOR EXAMPLE 6.16
      PARAMETER NP1MAX = 50
      INTEGER I,IEND,N,N1,N2
      REAL C(4,NP1MAX),FX,X,XI(NP1MAX)
      READ 500, N1
  500 FORMAT(I2)
      READ 501, (XI(I),C(1,I),I=1,N1),C(2,1),C(2,N1)
  501 FORMAT(2E10.3)
      N = N1 - 1
      CALL SPLINE(XI,C,N)
      CALL CALCCF(XI,C,N)
C
      READ 500, N2
      IEND = N + N2
      READ 501, (XI(I),C(1,I),I=N1,IEND),C(2,N1),C(2,IEND)
      N = N2 - 1
      CALL SPLINE(XI(N1),C(1,N1),N)
      CALL CALCCF(XI(N1),C(1,N1),N)
C
      N = IEND - 1
      X = XI(1)
      DO 10 I=1,40
         FX = PCUBIC(X,XI,C,N)
         PRINT 600, I,X,FX
  600    FORMAT(I5,F10.1,E20.9)
   10    X = X + .5
                                        STOP
      END
```

□

دراینجا ما تنها مقدمهٔ کوتاهی برای تقریب به‌وسیلهٔ بسجمله‌ای-تکه‌ای داده‌ایم برای توضیحات بیشتر، مثلا، به‌کتاب دبور [۴۵] مراجعه کنید.

تقریب به‌وسیلهٔ بسجمله‌ای و تقریب درجهٔ سوم-تکه‌ای ازجهات مختلف و مهمی که هنگام درونیابی آشکار می‌شوند، بـا هم متفاوت‌اند. اگر داده‌ها در نقاط متساوی‌الفاصله

داده شده باشند، آنگاه همان گونه که در مثال ۴.۲ نشان داده شده، با افزایش تعداد این نقاط، درونیابی به‌وسیلهٔ بسجمله‌ای، قدرت خود را از دست می‌دهد. این گونه مشکلات حتی در درونیابی درجهٔ سوم قلمی اصلا وجود ندارد. (توجه می کنید که در درونیابی به‌وسیلهٔ بسجمله‌ای مثلثاتی نیز هیچ مشکلی وجود ندارد.) و نیز با افزایش تعداد نقاط، بسجمله‌ای (و بسجمله‌ای مثلثاتی) پیچیده‌تر، یعنی ارزیابی آن پرهزینه‌تر می‌شود. همچنین وقتی درجهٔ بسجمله‌ای بیشتر از ۱۰ باشد، به‌علت بد شرطی شکل توانی، یا باید از دقت مضاعف استفاده کرد، یا بسجمله‌ای را به‌شکلهای دیگری، مثلا برحسب بسجمله‌ایهای چبیشف، نوشت. چنین مشکلاتی در درونیابی درجهٔ سوم ـ تکه‌ای وجود ندارد. زیرا تعداد نقاط درونیابی هرقدر زیاد باشد، درونیاب همیشه به‌طور محلی یك تابع خیلی ساده یعنی یك بسجمله‌ای درجهٔ سوم است. و سرانجام اگر تابعی که می‌باید تقریب زده شود در بعضی جاها رفتار نامناسب داشته باشد، آنگاه بهترین بسجمله‌ای تقریب، در همه‌جا تقریب ناتوانی خواهد بود (به‌تمرین ۱۰۶ـ۱۰ و ۱۰۶ـ۱۱ نگاه کنید). در تقریب بسجمله‌ای ـ تکه‌ای این امکان وجود دارد که با انتخاب مناسب نقاط ناپیوستگی، این اثرات به‌یك بازه درمحدودهٔ نقاطی که رفتار تابع نامطلوب است محدود شوند، و در بقیهٔ جاها تقریب خوبی به‌دست آید.

تمرین

۱ـ۷۰۶ به‌موجب نمادگذاری در این قسمت، معادله‌ای بیابید که f_1، f_2، s_1، s_2 باید در آن صدق کنند تا شرط «پایانی آزاد»

$$g_3''(a) = 0$$

برقرار شود.

۲ـ۷۰۶ یك بسجمله‌ای از درجهٔ مناسب پیدا کنید که منحنی طرح مثال ۱۶.۶ را در تمام نقاط داده‌های مفروض از ۱ر۶ تا ۱۸ (به‌انضمام شیبها) درونیابی کند، و آن را با تقریب قلمی که درمثال ۱۶.۶ حساب شده، مقایسه کنید.

۳ـ۷۰۶ داده‌های مثال ۱۶.۶ را توسط درونیابی درجهٔ سوم بسل، درونیابی کنید و نتایج حاصله را مقایسه کنید.

۴ـ۷۰۶ درونیابی درجهٔ سوم بسل **محلی** است، بدین‌معنی که مقدار تابع درونیاب $g_3(x)$ درهر نقطهٔ $\bar{x}$، تنها به‌چهار مقدار تابع مفروض درنزدیکترین نقاط به‌نقطهٔ $\bar{x}$ بستگی دارد. برعکس، درونیابی درجهٔ سوم قلمی **کلی**[1] است، یعنی مقدار $g_3(x)$ در هر نقطهٔ مفروض به‌کلیهٔ اطلاعات داده شده در بازهٔ $f(x)$ بستگی دارد. این دوحکم را ثابت کنید.

۵ـ۷۰۶ سعی کنید ازدرونیابی یك تابع مفروض توسط یك تابع سهمی ـ تکه‌ای $g_2(x)$ الگوی مناسبی بسازید. آیا می‌توانید تابع $g_2(x)$ را به‌طور پیوسته مشتقپذیر سازید؟

---

1. global

۷

مشتقگیری و انتگرالگیری

درفصل۲، تکنیکهایی را برای تقریب‌زدن یك تابع مفروض به‌وسیلهٔ یك بسجمله‌ای، نوعاً به‌روش درونیابی، بسط دادیم. دراین فصل، یك استفادهٔ مهم از این بسجمله‌ایهای تقریبی یعنی **جایگزینی تحلیلی** را مورد مطالعه قرار می‌دهیم. در اینجا، سروکار، با گذاردن یك بسجمله‌ای تقریبی به‌جای یك تابع پیچیده یا فقط یك تابع جدولبندی شده است به‌طوری که اعمال بنیادی دیفرانسیل وانتگرال بتوانند آسانتر انجام گیرند، یا اساساً بتوانندانجام گیرند. این اعمال شامل عملهای

$$I(f)=\int_b^b f(x)\,dx \qquad D(f)=f'(a)$$

$$S_n(f)=\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx$$

و حتی

$$Z(f)=\lim_{h\to 0} f(h)$$

هستند. به‌گونه‌ای مجرد، اگر L معرف یکی ازاعمال فوق (یا عملی مشابه) روی توابع باشد، و به‌ازای یك $f(x)$ مفروض، $p(x)$ تقریبی برای $f(x)$ باشد، آنگاه عدد $L(f)$ را با عدد $L(p)$ تقریب می‌زنیم. امید ما این‌است‌که عمل L بتواند به‌آسانی روی $p(x)$

انجام گیرد و اگر $p(x)$ یك بسجمله‌ای و L یکی از اعمال فوق‌الذکر باشد، این امید قابل توجیه باشد.

خطی بودن عمل L (همان‌گونه که اعمال مذکور در بالا هستند) در برآورد خطای $L(f)-L(p)$، تاحدی مفید است. معنی خطی بودن این است که

$$L(f(x)+g(x))=L(f)+L(g)$$

$$L(af(x))=aL(f)$$

که در اینجا a یك عدد است و $f(x)$ و $g(x)$ دو تابع. خطی بودن ایجاب می‌کند که داشته باشیم

$$L(f)-L(p)=L(e)$$

که $e(x)$ عبارت است از خطای موجود در تابع تقریب $p(x)$ از تابع $f(x)$، یعنی

$$f(x)=p(x)+e(x)$$

ما معمولا $p(x)$ را به‌عنوان یك بسجمله‌ای درونیاب انتخاب می‌کنیم. مثلا $p(x)$ یك بسجمله‌ای از درجهٔ نابیشتر از k است که $f(x)$ را در نقاط $x_0, \ldots, x_k$ درونیابی می‌کند. اگر این نقاط متمایز باشند، آنگاه بنابر (۷۰۲) داریم

$$p(x)=\sum_{i=0}^{k} f(x_i)l_i(x)$$

که در آن $l_i(x)$ها بسجمله‌ایهای لاگرانژ به‌ازای نقاط $x_0, \ldots, x_k$ هستند. اما اگر عمل L خطی باشد، نتیجه می‌شود که

$$L(p)=\sum_{i=0}^{k} f(x_i)w_i$$

که در آن اعداد w_i با روابط

$$w_i=L(l_i) \qquad i=0,\ldots,k$$

مشخص می‌شوند و به $f(x)$ بستگی ندارند؛ و بنابراین می‌توانند (به‌ازای هر مجموعه از نقاط خاص $x_0, \ldots, x_k$) یکبار برای همیشه محاسبه شوند. در این شکل نمایش، معمولا تقریب $L(p)$ یك **قاعده**[1] [برای تقریب $L(f)$] نامیده می‌شود، و اعداد w_i **اوزان** یا ضرایب آن و نقاط $x_0, \ldots, x_k$ **رئوس**[2] آن نام دارند. در این قاعده، با اعمال عمل L بر تابع خطای بسجمله‌ای درونیابی، همان‌گونه که در (۱۸۰۲) و (۳۷۰۲) آمده است، و با

---

1. rule 2. nodes

استفاده از این حقیقت که تفاضل منقسم تابعی است خوش تعریف از شناسه‌هایش، عبارت زیر را برای خطا به‌دست می‌آوریم

$$E(f)=L(f)-L(p)$$

۱.۷ مشتقگیری عددی

در وهلهٔ اول بعضی از تکنیکهای عددی برای تقریب‌زدن مشتق یك تابع داده شده، یعنی $f'(x)$ را بررسی می‌کنیم. قواعد حاصل درحل‌عددی معادلات دیفرانسیل ازاهمیت درجهٔ اولی برخوردارند، و علت اساسی شرح آنها دراینجا همین امر است. این قواعد برای به‌دست‌آوردن تقریبهای عددی یك مشتق از روی مقادیر تابع نیز به‌کار می‌روند. اما باید متذکر شویم که مشتقگیری عددی برمبنای بسجمله‌ای درونیاب، اساساً فرایندی‌است ناپایدار و نمی‌توانیم دقت خوبی را انتظار داشته باشیم حتی اگر داده‌های اولیه دقیق باشند. به‌طوری که خواهیم‌دید، خطای $f'(x)-p'(x)$ ممکن‌است بسیار بزرگ باشد، بخصوص زمانی که مقادیر $f(x)$ در نقاط درونیابی «نوفه‌ای[1]» باشند. این مطالب در زیر دقیقتر توضیح داده خواهد شد.

گیریم $f(x)$ تابعی باشد در بازهٔ $[c, d]$ به‌طور پیوسته‌مشتقپذیر. اگردر بازهٔ $[c, d]$، نقاط $x_0, \dots, x_k$ نقاطی متمایز باشند، می‌توانیم $f(x)$ را به‌موجب (۳۷.۲) به‌صورت زیر بنویسیم

$$f(x)=p_k(x)+f[x_0, \dots, x_k, x]\psi_k(x) \qquad (۱.۷)$$

که در آن $p_k(x)$ بسجمله‌ای‌است از درجهٔ نابیشتر از k که $f(x)$ درنقاط $x_0, \dots, x_k$ درونیابی می‌کند و

$$\psi_k(x)=\prod_{j=0}^{k}(x-x_j)$$

بنابر(۳۸.۲)، اگر $f(x)$ به‌اندازهٔ‌کافی هموار باشد، داریم

$$\frac{d}{dx}f[x_0, \dots, x_k, x]=f[x_0, \dots, x_k, x, x]$$

لذا، دریك چنین‌حالتی، می‌توانیم از (۱.۷) مشتق بگیریم و چنین به‌دست آوریم

$$f'(x)=p_k'(x)+f[x_0, \dots, x_k, x, x]\psi_k(x)+f[x_0, \dots, x_k, x]\psi_k'(x) \qquad (۲.۷)$$

عملگر[2] D را درنقطهٔ غیرمشخص a روی $[c, d]$ به‌شرح زیر تعریف می‌کنیم

---

1. noisy 2. operator

$$D(f)=f'(a)$$

اگر $D(f)$ را با $D(p_k)$ تقریب زنیم، آنگاه به‌موجب (۲.۷)، خطا در این تقریب به‌شرح زیر است

$$\begin{aligned}E(f)&=D(f)-D(p_k)\\&=f[x_0,\cdots,x_k,a,a]\psi_k(a)+f[x_0,\cdots,x_k,a]\psi_k'(a)\end{aligned}$$

یا به‌ازای مقادیری مانند $\xi,\eta\in(c,d)$

$$E(f)=\frac{f^{(k+2)}(\xi)\psi_k(a)}{(k+2)!}+\frac{f^{(k+1)}(\eta)\psi_k'(a)}{(k+1)!}\qquad(3.7)$$

در مشتقگیری عددی، عبارت (۳.۷) درحالت کلی برای خطای $E(f)$ اطلاعات بسیار کمی از خطای واقعی به‌دست می‌دهد، زیرا به‌ندرت مشتقهای $f^{(k+1)}$ و $f^{(k+2)}$ موجود در $E(f)$ را دردست داریم وتقریباً هیچ‌وقت شناسه‌های ξ و η برما معلوم نیستند. دربرخی موارد این عبارت خطا یا با انتخاب نقطهٔ a که در آن مشتق باید محاسبه شود و یا با گزینش مناسب نقاط درونیابی $x_0,\cdots,x_k$ می‌توانند بسیارساده شود.

درابتدا حالتی راکه a یکی ازنقاط درونیابی باشد بررسی می‌کنیم. گیریم به‌ازای مقداری مانند i، داشته باشیم $a=x_i$. دراین صورت چون $\psi_k(x)$ متضمن عامل $(x-x_i)$ است، نتیجه می‌شود که $\psi_k(a)=0$، واولین‌جمله در رابطهٔ خطای (۳.۷) حذف می‌شود. بعلاوه $\psi_k'(a)=q(a)$، که در آن

$$q(x)=\frac{\psi_k(x)}{x-x_i}=(x-x_0)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_k)$$

بنابراین، اگر به‌ازای مقداری مانند i، $a=x_i$ را انتخاب کنیم، آنگاه معادلهٔ (۳.۷) به‌معادلهٔ زیر بدل می‌شود

$$E(f)=\frac{1}{(k+1)!}f^{(k+1)}(\eta)\prod_{\substack{j=0\\j\neq i}}^{k}(x_i-x_j)\qquad \eta\in(c,d)$$

به‌ازای نقطه‌ای مانند

(۴.۷)

راه دیگر برای ساده‌کردن عبارت خطای (۳.۷) انتخاب a است به‌طوری که $\psi_k'(a)=0$، زیرا دراین صورت جملهٔ دوم رابطهٔ (۳.۷) صفرمی‌شود. اگر k عدد فردی باشد، می‌توانیم این عمل را با قراردادن x_iها به‌طور متقارن پیرامون a، یعنی به‌طوری که داشته باشیم

$$x_{k-j}-a=a-x_j\qquad j=0,\ldots,\frac{k-1}{2}\qquad(5.7)$$

انجام دهیم. زیرا

$$(x-x_j)(x-x_{k-j})=(x-a+a-x_j)(x-a+a-x_{k-j})$$
$$=(x-a)^2-(a-x_j)^2 \qquad j=0,\dots,\frac{k-1}{2}$$

بنابراین خواهیم داشت

$$\psi_k(x)=\prod_{j=0}^{(k-1)/2}[(x-a)^2-(a-x_j)^2]$$

زیرا به‌ازای کلیهٔ مقادیر j

$$\frac{d}{dx}[(x-a)^2-(a-x_j)^2]|_{x=a}=2(x-a)|_{x=a}=0$$

بنابراین نتیجه می‌شود که $\psi_k'(a)=0$. خلاصه، اگر رابطهٔ (۵.۷) برقرار باشد، آنگاه (۳.۷) به‌معادلهٔ زیر بدل می‌شود

$$E(f)=\frac{1}{(k+2)!}f^{(k+2)}(\xi)\prod_{j=0}^{(k-1)/2}[-(a-x_j)^2] \qquad (۶.۷)$$

باید توجه کرد که درجهٔ مشتق $f(x)$ در رابطهٔ (۶.۷) یك مرتبه بالاتر از درجهٔ مشتق در رابطهٔ (۴.۷) است.

اکنون مثالهای خاصی را مورد بررسی قرار می‌دهیم. اگر $k=0$، آنگاه داریم $D(p_k)=0$ که نکتهٔ اطمینان بخشی است، اما (معمولا) تقریب خیلی خوبی برای $D(f)=f'(a)$ نیست. بنابراین $k\geqslant 1$ را انتخاب می‌کنیم. به‌ازای $k=1$، داریم

$$p_k(x)=f(x_0)+f[x_0,x_1](x-x_0)$$

از این‌رو بدون درنظرگرفتن a، خواهیم داشت

$$D(p_k)=f[x_0,x_1]$$

اگر $a=x_0$، آنگاه به‌ازای $h=x_1-x_0$ روابط (۲.۷) و (۴.۷) فرمول تفاضل‌ـپیشرو[1] زیر را به‌دست می‌دهد

$$f'(a)\approx f[a,a+h]=\frac{f(a+h)-f(a)}{h} \qquad E(f)=-\frac{1}{2}hf''(\eta)$$

(۷.۷)

1. forward_difference

از سوی دیگر، اگر بگیریم $a=(۱/۲)(x_۰+x_۱)$، آنگاه $x_۱$ و $x_۰$ پیرامون a متقارن بوده و به‌ازای $x_۰=a-h$ و $x_۱=a+h$ و $h=(۱/۲)(x_۱-x_۰)$ رابطهٔ (۶.۷) خود فرمول معمولی تفاضل مرکزی[1] زیر را به‌دست می‌دهد

$$f'(a)\approx f[a-h,a+h]=\frac{f(a+h)-f(a-h)}{۲h} \qquad (۸.۷)$$

$$E(f)=-\frac{h^۲}{۶}f'''(\eta)$$

بنابراین، اگر $x_۱$ و $x_۰$ «نزدیك به‌هم باشند» آنگاه تقریب برای $f'(a)$ در نقطهٔ میانی $a=(۱/۲)(x_۰+x_۱)$ بهتر از تقریب دریکی ازدوسر $a=x_۰$ یا $a=x_۱$ است. این‌امر چندان شگفت‌آور نیست زیرا به‌موجب قضیهٔ مقدار میانگین برای مشتقات (به‌قسمت ۷.۱ نگاه کنید) داریم

$$f[x_۰,x_۱]=f'(a) \qquad \text{به‌ازای نقطه‌ای مانند } a \text{ بین } x_۰ \text{ و } x_۱$$

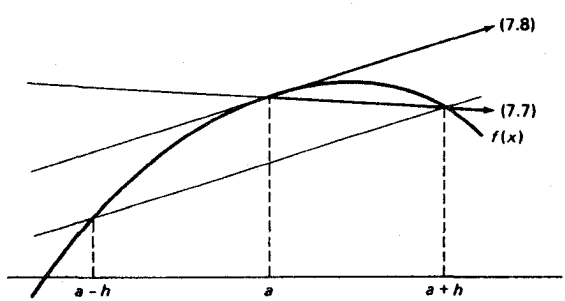
این نکته درشکل ۱.۷ نیز نشان داده شده‌است.

سپس استفاده از سه نقطهٔ درونیابی را مورد بررسی قرار می‌دهیم به‌طوری که $k=۲$. بنابراین

$$p_k(x)=f(x_۰)+f[x_۰,x_۱](x-x_۰)+f[x_۰,x_۱,x_۲](x-x_۰)(x-x_۱)$$

لذا

$$p_k'(x)=f[x_۰,x_۱]+f[x_۰,x_۱,x_۲](۲x-x_۰-x_۱)$$



شکل ۱.۷ مشتقگیری عددی

1. central-difference

بنابراین اگر $a=x_0$، آنگاه از روابط (۲.۷) و (۴.۷) به‌دست می‌آوریم

$$f'(a)=f[a,x_1]+f[a,x_1,x_2](a-x_1)+\frac{1}{6}(a-x_1)(a-x_2)f'''(\eta) \qquad (۹.۷)$$

اکنون درحالت خاص، گیریم $x_1=a+h$ و $x_2=a+2h$، بنابراین (۹.۷) به‌صورت زیر بدل می‌شود

$$\boxed{f'(a)\approx\frac{-3f(a)+4f(a+h)-f(a+2h)}{2h}} \qquad (۱۰.۷)$$

$$\boxed{E(f)=\frac{h^2}{3}f'''(\xi) \qquad \text{به‌ازای نقطه‌ای مانند } \xi \text{ بین } a \text{ و } a+2h}$$

از طرف دیگر، اگر $x_1=a-h$ و $x_2=a+h$، آنگاه داریم

$$f'(a)\approx\frac{f(a+h)-f(a-h)}{2h}$$

$$E(f)=-\frac{h^2}{6}f'''(\xi) \qquad |\xi-a|<|h| \text{ به‌ازای} \qquad (۱۱.۷)$$

که عیناً همان معادلهٔ (۸.۷) است.

فرمولهایی برای تقریب‌زدن مشتقهای مراتب بالاتر $f(x)$ را می‌توان به‌طریقی مشابه به‌دست آورد. بنابراین با دوبار مشتق گرفتن از (۱۰.۷)، خواهیم داشت

$$\begin{aligned}f''(x)=p_k''(x)&+2f[x_0,\cdots,x_k,x,x,x]\psi_k(x)\\&+2f[x_0,\cdots,x_k,x,x]\psi_k'(x)\\&+f[x_0,\cdots,x_k,x]\psi_k''(x)\end{aligned} \qquad (۱۲.۷)$$

به‌ازای $k=2$ و $a=x_0$ از این رابطه نتیجه می‌شود

$$\begin{aligned}f''(a)=2f[a,x_1,x_2]&+2f[a,x_1,x_2,a,a](a-x_1)(a-x_2)\\&+f[a,x_1,x_2,a]2(a-x_1+a-x_2)\end{aligned}$$

از این‌رو به‌ازای $x_1=a+h$ و $x_2=a+2h$، داریم

$$f''(a) \approx \frac{f(a)-2f(a+h)+f(a+2h)}{h^2}$$

$$E(f)=\frac{h^2}{6}f^{iv}(\xi)-hf'''(\eta) \qquad (۱۳.۷)$$

درعوض، با انتخاب $x_1=a-h$ و $x_2=a+h$، به‌گونه‌ای‌که نقاط درونیابی پیرامون a متقارن باشند، خواهیم داشت

$$f''(a) \approx \frac{f(a-h)-2f(a)+f(a+h)}{h^2}$$

$$E(f)=-\frac{h^2}{12}f^{iv}(\xi) \qquad |\xi-a|<|h| \text{ به ازای} \qquad (۱۴.۷)$$

ملاحظه می‌کنیم که قراردادن نقاط درونیابی به‌طور متقارن پیرامون a، مجدداً فرمولی از مرتبهٔ بالاتر را نتیجه داده‌است.

و سرانجام از (۱۷.۲) نتیجه می‌گیریم که

$$k!f[x_0,\cdots,x_k]$$

تقریب «خوبی» برای $f^{(k)}(a)$ است، به‌شرط آنکه کلیهٔ x_iها «به‌اندازهٔ کافی نزدیک» به a باشند. فرمولهای (۷.۷)، (۸.۷)، (۱۰.۷) همگی به‌صورت کلی زیر هستند.

$$D(f)=D(p_k)+\text{const}\,h^r f^{(r+1)}(\xi) \qquad (۱۵.۷)$$

که در آن $h^r f^{(r+1)}(\xi)$ مقداری‌است ثابت، $D(f)=f'(a)$، و h فواصل نقاطی است که برای درونیابی به‌کارمی‌روند. بعلاوه، عدد $D(p_k)$ تنها در تعدادی متناهی[1] از نقاط مجزا[2] شامل مقادیر $f(x)$ است. روند قراردادن $D(f)$ به‌جای $D(p_k)$ را عمل **جداسازی**[3] نامند و ثابت جملهٔ خطا یعنی $h^r f^{(r+1)}(\xi)$ را **خطای جداسازی** گویند.

از رابطهٔ (۱۵.۷) نتیجه می‌شود که تنها با محاسبهٔ $D(p_k)$ به‌ازای h به‌اندازهٔ کافی کوچك، باید بتوانیم بـا هردقت مطلوبی، $D(f)$ را محاسبه کنیم. ولی محدودیت طول کلمه‌ها در کامپیوترها و از دست رفتن ارقام با معنی هنگام تفریق کمیتهای تقریباً مساوی، توأماً حصول دقت از مرتبهٔ بالا را دشوار می‌سازند. زیـرا برای کامپیوتری با طول کلمهٔ

1. finite 2. discrete 3. discretization

ثابت و برای یك تابع مفروض، مقدار بهینه‌ای از h وجود دارد که کمتر از آن، تقریب را بدتر می‌سازد. برای مثال، مقادیر جدول ۱.۷ را درنظر می‌گیریم. این مقادیر به‌صورت محاسبات با دقت ساده درحساب با ممیز شناور و با استفاده از IBM مدل ۷۰۹۴ انجام گرفته‌اند.

در این جدول، ستونی که با عنوان D_h مشخص شده‌است مقادیر $f'(a)$ را که توسط (۸.۷) براورد می‌شود به‌دست می‌دهد وستون D_h^2 مقادیر $f''(a)$ را، که با (۱۴.۷) براورد شده‌است. تابع $f(x)$ برابر e^x اختیار شده، و واضح است که به‌ازای $a=0$ مقادیر دقیق $f'(a)$ و $f''(a)$ برابر یك است. ازجدول فوق چنین برمی‌آید که با کوچك‌شدن h تامقدار $h=0.01$، مقادیر D_h و D_h^2 مرتباً رو به بهترشدن می‌روند. پس‌ازاین، نتایج بدترمی‌شوند. به‌ازای $h=0.0001$ در D_h چهار رقم با معنی و در D_h^2 هفت رقم با معنی از دست می‌روند. تنها راه چاره برای از دست نرفتن این ارقام با معنی، وقتی h کوچکتر می‌شود، افزودن تعدادی ارقام با معنی به‌ارقام حساب شدهٔ $f(x)$ است. معمولا این عمل‌روی بیشتر کامپیوترها غیرممکن‌است. بعلاوه $f(x)$ خود، نتیجهٔ محاسبات دیگری است که خطاهای دیگری را وارد می‌کنند.

برای تحلیل این مسئله، مطالعهٔ فرمول (۱۱.۷) رابطهٔ زیر را به‌دست می‌دهد

$$f'(a)=\frac{f(a+h)-f(a-h)}{2h}-\frac{h^2f'''(\xi)}{6}$$

درحقیقت به‌دلیل گردکردن، اعداد $f(a+h)+E_+$ و $f(a-h)+E_-$ به‌جای اعداد $f(a+h)$ و $f(a-h)$ درمحاسبات به‌کار می‌روند، بنابراین محاسبات زیر را انجام می‌دهیم

$$f'_{\text{comp}}=\frac{f(a+h)+E_+-f(a-h)-E_-}{2h}$$

$$=\frac{f(a+h)-f(a-h)}{2h}+\frac{E_+-E_-}{2h}$$

جدول ۱.۷

h	EXP(h)	EXP($-h$)	D_h	D_h^2
1.0	0.27182817E 01	0.36787944E 00	0.11752012E 01	0.10861612E 01
0.1	0.11051708E 01	0.90483743E 00	0.10016673E 01	0.10008334E 01
0.01	0.10100501E 01	0.99004984E 00	0.10000161E 01	0.10000169E 01
0.001	0.10010005E 01	0.99900050E 00	0.99999458E 00	0.99837783E 00
0.0001	0.10000999E 01	0.99990001E 00	0.99994244E 00	0.14901161E 01

لذا، با توجه به‌رابطهٔ (۱۱۰۷) خواهیم داشت

$$f'(a)=f'_{\text{comp}}-\frac{E_+-E_-}{2h}-\frac{h^2f'''(\xi)}{6} \qquad (۱۶۰۷)$$

ازاین‌رومشاهده می‌شودکه خطا درتقریب محاسبهٔ f'_{comp} برای $f'(a)$ از دوقسمت تشکیل شده‌است، یك قسمت براثر گردکردن وقسمت دیگربراثر جداسازی. اگر$f'''(x)$کراندار باشد، آنگاه به‌ازای $h \to 0$ خطای جداسازی به‌سمت صفر میل می‌کند، امـا اگر فرض کنیم (همان‌گونه‌که درعمل باید انجام داد)که E_+-E_-کم نمی‌شود، آنگاه خطای گرد کردن زیاد می‌شود (اما به‌تمرین۱۰۷ـ۵ نگاه‌کنید).

مقدار بهینهٔ h به‌موجب‌تعریف، مقداری‌است‌که به‌ازای‌آن، حاصلجمع مقادیرخطای گردکردن و خطای جداسازی حداقل می‌شود. برای آنکه روش پیداکردن مقدار بهینهٔ h را نشان دهیم، مسئلهٔ فوق را در رابطه بـا محاسبهٔ $f'(0)$ درحالت $f(x)=e^x$ بررسی می‌کنیـم. فرض‌کنیدکـه در محاسبهٔ e^x خطای حاصل برابر بـا $\pm 1\times10^{-8}$ و مقدار E_+-E_- محدود و تقریباً برابربا $\pm 2\times10^{-8}$ باشد. دراین‌صورت با توجه به‌رابطهٔ (۱۶۰۷) خطای گردکردن R به‌طور تقریبی برابراست با

$$R=\pm\frac{2\times10^{-8}}{2h}$$

چون $f'''(\xi)$ تقریباً برابر یك‌است، خطای جداسازی T تقریباً برابربا

$$T=-\frac{1}{6}h^2$$

خواهد بود. برای پیداکردن مقدار بهینهٔ h می‌باید رابطهٔ زیر را مینیمم سازیم

$$|R|+|T|=\frac{10^{-8}}{h}+\frac{1}{6}h^2=g(h)$$

برای یافتن‌مقداری از hکه به‌ازای آن مقدار $g(h)$ مینیمم باشد، از رابطـهٔ $g(h)$ نسبت به h مشتق می‌گیریم و ریشهٔ آن را به‌دست می‌آوریم. بنابراین خواهیم داشت:

$$g'(h)=\frac{-10^{-8}}{h^2}+\frac{h}{3}=0$$

که جواب مثبت آن برابراست با

$$h^3=3\times10^{-8}$$

یا

$$h = ۱۰^{-۳}\sqrt[۳]{۳۰} \approx ۰٫۰۰۳$$

که این مقدار بهینهٔ h است. دانشجویان می‌توانند با بررسی جدول ۱.۷ تحقیق کنند که بهترین مقدار برای h بین ۰٫۰۱ و ۰٫۰۰۱ قرار دارد.

فرمولهایی که در این قسمت درمشتقگیری عددی به‌دست آمد درمطالعهٔ روشهای حل عددی معادلات دیفرانسیل بسیار مفیدند (به‌فصل ۸ و ۹ نگاه کنید). اما تحلیلهای فوق‌نشان می‌دهند که برای محاسبهٔ تقریبی مشتقها، این فرمولها از مطلوبیت محدودی برخوردارند. این تحلیلها نشان می‌دهند که می‌توانیم با به‌کار بستن محاسبات با دقت عمل «به‌اندازهٔ کافی» بالا، با اثرخطای گرد کردن مقابله کنیم. اما وقتی $f(x)$ را تنها به‌طور تقریبی درچندین نقطه داشته باشیم، این امرغیرممکن است.

اگرمحاسبهٔ عددی مشتقها غیرقابل اجتناب باشد، معمولا ترجیح می‌دهند که $D(f)$ را از روی $D(p_k)$ برآورد کنند، که در آن $p_k(x)$ تقریب $f(x)$ با کوچکترین توانهای دوم به‌کمك بسجمله‌هایی از درجهٔ پایین (به‌بخش ۴.۶ نگاه کنید) است. روش مناسب دیگر برای تقریب‌زدن $D(f)$، استفاده از $D(g_۳)$ است، که در آن $g_۳(x)$ یك درونیاب درجهٔ سوم قلمی است که $f(x)$ را در تعدادی از نقاط درونیابی می‌کند، یا به‌بهترین نحو، $f(x)$ را به‌تعبیر کوچکترین توانهای دوم تقریب می‌زند.

تمرین

۱-۱۰.۷ با استفاده ازجدول زیر و روابط (۷.۷)، (۸.۷) و (۱۰.۷)، $f'(۱٫۴)$ را تعیین کنید. همچنین با استفاده از رابطهٔ (۱۴.۷)، $f''(۱٫۴)$ را به‌دست آورید. نتایج به‌دست آمده‌را با $f'(۱٫۴) = \cosh ۱٫۴ = ۲٫۱۵۰۹$ و $f''(۱٫۴) = \sinh ۱٫۴ = ۱٫۹۰۴۳$ که تا تعداد ارقام داده شده دقیق هستند، مقایسه نمایید.

x	$f(x)$
۱٫۲	۱٫۵۰۹۵
۱٫۳	۱٫۶۹۸۴
۱٫۴	۱٫۹۰۴۳
۱٫۵	۲٫۱۲۹۳
۱٫۶	۲٫۳۷۵۶

۲-۱۰.۷ با استفاده ازجدول مقادیر زیر که به‌تابع $f(x) = \sinh x$ مربوط است واستفاده

از رابطهٔ(۸.۷) به‌ازای $h=$ ۰٫۰۰۱ و $h=$ ۰٫۰۰۲، f'(۰٫۴۰۰) را تعیین کنید، کدامیک از مقادیر به‌دست آمده دقیق هستند؟ جواب درست برابر است با

$$f'(۰٫۴)=\cosh ۰٫۴ = ۱٫۰۸۱۰۷۲۰$$

x	$f(x)$
۰٫۳۹۸	۰٫۴۰۸۵۹۱
۰٫۳۹۹	۰٫۴۰۹۶۷۱
۰٫۴۰۰	۰٫۴۱۰۷۵۲
۰٫۴۰۱	۰٫۴۱۱۸۳۴
۰٫۴۰۲	۰٫۴۱۲۹۱۵

۳-۱۰۷ در معادلهٔ (۱۶.۷)، گیریم $f(x)=\sinh x$، و فرض می‌کنیم که درطی محاسبهٔ $\sinh x$، میزان خطای گردکردن ثابت باقی بماند به‌طوری‌که داشته باشیم

$$E_+ - E_- = ۰٫۵ \times ۱۰^{-۷}.$$

اگر فرمول (۸.۷) برای محاسبهٔ $f'(۰)$ به‌کار رود، مقدار بهینهٔ h را که می‌باید در فرمول فوق به‌کار رود معین کنید.

۴-۱۰۷ با سه بار مشتقگیری از رابطهٔ (۱.۷) و با انتخاب $k=۳$، به‌ازای $a=x_۰$، $x_۱=a-h$، $x_۲=a+h$، $x_۳=a+۲h$ فرمولی برای $f'''(a)$ به‌دست آورید. همچنین جملهٔ خطا را برای این فرمول به‌دست آورید.

۵-۱۰۷ به‌کمک کامپیوتر دنبالهٔ اعداد

$$a_n = f[۲-۲^{-n},\ ۲+۲^{-n}] \qquad n=۱, ۲, ۳, \cdots$$

را که در آن $f(x)=\ln x$ است، محاسبه کنید. بدون درنظرگرفتن اثرات خطای گردکردن داریم

$$\lim_{n\to\infty} a_n = f'(۲) = ۰٫۵$$

و با بحثی که در این قسمت آمد، به‌دلیل خطای گردکردن، داریم

$$\lim_{n\to\infty} |a_n| = \infty$$

آیا این امر واقعاً اتفاق می‌افتد؟ اگر نمی‌افتد، چرا؟ این امر بحثی را که در متن انجام گرفته از اعتبار می‌اندازد.

۶-۱۰۷ با بسط $f(a+h)$ و $f(a-h)$ به‌سری تیلر پیرامون نقطهٔ a، فرمول (۸.۷) را بررسی کنید.

۷-۱۰۷ با استفاده از بسط سری تیلر، فرمول (۱۴.۷) را برای $f''(a)$ به‌دست آورید.

۲.۷ انتگرالگیری عددی: برخی قواعد اساسی

مسئلهٔ انتگرالگیری عددی یا تربیع[1]، عبارت‌است از براوردکردن عدد

$$I(f)=\int_a^b f(x)\,dx \qquad (۱۷.۷)$$

این مسئله زمانی پیش می‌آیدکه نتوان انتگرالگیری را به‌طور دقیق انجام داد، یاهنگامی که $f(x)$ تنها در تعداد متناهی از نقاط، داده شده باشد.

بدین دلیل، خط مشیی راکه در آغاز این فصل بیان‌کردیم پیش می‌گیریم. $I(f)$ را به‌وسیلهٔ $I(p_k)$ تقریب می‌زنیم، که دراینجا $p_k(x)$ یك بسجمله‌ای‌است ازدرجهٔ نابیشتر از k که درنقاط x_0، ...، x_k بر $f(x)$ منطبق است. این تقریب معمولا به‌شکل یك قاعده یعنی به‌صورت حاصلجمع وزین

$$I(p_k)=A_0 f(x_0)+A_1 f(x_1)+\cdots+A_k f(x_k)$$

ازمقادیر تابع $f(x_0)$، ...، $f(x_k)$ نوشته‌می‌شود. وزنهامی‌توانندبه‌صورت $A_i=I(l_i)$، که در آن $l_i(x)$، iامین بسجمله‌ای لاگرانژاست محاسبه شوند.

اکنون فرض کنیدکه عامل زیرعلامت انتگرال $f(x)$ دریك بازهٔ $[c, d]$ که متضمن نقاط a و b است، به‌اندازهٔ کافی هموار باشد به‌طوری‌که بتوانیم آن را مانند (۳۷.۲) به‌شکل

$$f(x)=p_k(x)+f[x_0, \cdots, x_k, x]\psi_k(x)$$

بنویسیم، که $\psi_k(x)$ درتساوی زیرصدق می کند

$$\psi_k(x)=\prod_{j=0}^{k}(x-x_j)$$

دراین‌صورت مقدار خطا دربراورد $I(p_k)$ برای $I(f)$ به‌شرح زیراست

---

1. quadrature

$$E(f)=I(f)-I(p_k)=\int_a^b f[x_0,\cdots,x_k,x]\psi_k(x)\,dx \qquad (۱۸.۷)$$

که در آن $f[x_0,\cdots,x_k,x]$ پیوسته و از این‌رو بنابر قضیهٔ ۵.۲ به‌صورت تابعی از x انتگرال‌پذیر است.

گاهی این عبارت خطا می‌تواند ساده شود. برای مثال اگر $\psi_k(x)$ روی (a,b) دارای یك علامت باشد، آنگاه بنابر قضیهٔ مقدار میانگین برای انتگرالها (به‌قسمت ۷.۱ نگاه کنید) داریم،

به‌ازای مقداری مانند $\xi\in(a,b)$

$$\int_a^b f[x_0,\cdots,x_k,x]\psi_k(x)\,dx=f[x_0,\cdots,x_k,\xi]\int_a^b \psi_k(x)\,dx \qquad (۱۹.۷)$$

بعلاوه اگر $f(x)$ به‌طور پیوسته $(k+1)$مرتبه بر (c,d) مشتقپذیر باشد، با توجه به (۱۸.۷) و (۱۹.۷) خواهیم داشت

$$E(f)=\frac{1}{(x+1)!}f^{(k+1)}(\eta)\int_a^b \psi_k(x)\,dx \qquad \eta\in(c,d)$$

به‌ازای مقداری مانند

(۲۰.۷)

حتی اگر $\psi_k(x)$ دارای یك علامت نباشد، تا حدودی ساده‌کردن عبارت خطای (۱۸.۷) امکان‌پذیر است. یك مورد خاص مطلوب از این نوع وقتی‌است‌که داشته باشیم

$$\int_a^b \psi_k(x)\,dx=0 \qquad (۲۱.۷)$$

درچنین حالتی می‌توانیم ازهمانی[1]

$$f[x_0,\cdots,x_k,x]=f[x_0,\cdots,x_k,x_{k+1}]+f[x_0,\cdots,x_{k+1},x](x-x_{k+1})$$

که به‌ازای مقدار دلخواهی مانند x_{k+1} معتبر است، برای یافتن معادلهٔ

$$\begin{aligned}E(f)&=\int_a^b f[x_0,\cdots,x_{k+1}]\psi_k(x)\,dx\\&\quad+\int_a^b f[x_0,\cdots,x_{k+1},x](x-x_{k+1})\psi_k(x)\,dx\\&=\int_a^b f[x_0,\cdots,x_{k+1},x]\psi_{k+1}(x)\,dx\end{aligned}$$

1. identity

استفاده کنیم، زیرا طرف دوم معادلهٔ اخیر به‌صورت زیر تبدیل می‌شود

$$\int_a^b f[x_0, \cdots, x_{k+1}]\psi_k(x)\,dx = f[x_0 \cdots, x_{k+1}] \int_a^b \psi_k(x)\,dx = 0$$

اکنون اگر بتوانیم x_{k+1} را به‌طریقی انتخاب کنیم که تابع $\psi_{k+1}(x) = (x - x_{k+1})\psi_k(x)$ در بازهٔ (a, b) دارای یك علامت و $f(x)$ تابعی $(k+2)$ مرتبه مشتقپذیر باشد، آنگاه (مانند قبل) نتیجه می‌شود که

$$E(f) = \frac{1}{(k+2)!} f^{(k+2)}(\eta) \int_a^b \psi_{k+1}(x)\,dx \qquad \eta \in (c, b)$$

به‌ازای مقداری مانند

(۲۲.۷)

باید توجه کرد که درجهٔ مشتق $f(x)$ که در (۲۲.۷) ظاهر می‌شود، یك درجه بالاتر از درجهٔ مشتقی است که در (۲۰.۷) آمده بود. این امر مانند مشتقگیری عددی معرف آن است که مرتبهٔ (۲۲.۷) از مرتبهٔ (۲۰.۷) بالاتر است.

اکنون مثالهای خاصی را مورد بررسی قرار می‌دهیم. گیریم $k = 0$، آنگاه

$$f(x) = f(x_0) + f[x_0, x](x - x_0)$$

بنابراین

$$I(p_k) = (b - a)f(x_0)$$

اگر $x_0 = a$، آنگاه این تقریب به‌صورت زیر درمی‌آید

$$\boxed{I(f) \approx R = (b - d)f(a)} \qquad (۲۳.۷)$$

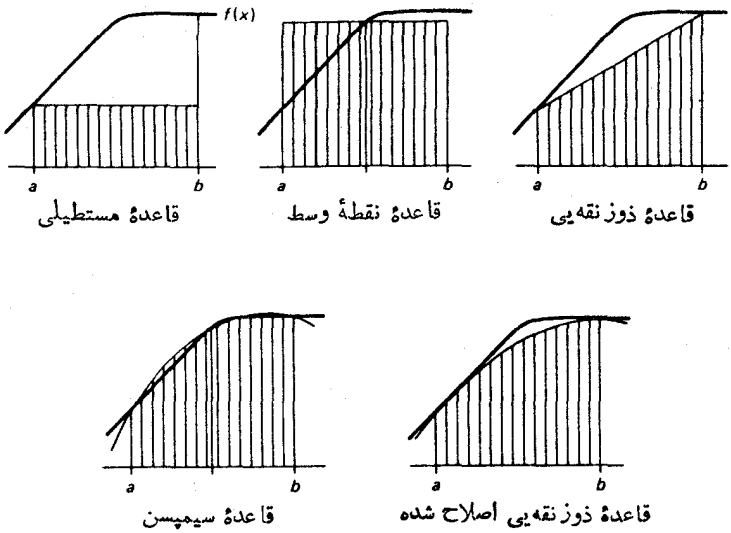
که به‌**قاعدهٔ مستطیلی**[1] معروف است (به‌شکل ۲.۷ نگاه کنید). چون در این حالت تابع $\psi_0(x) = x - a$ روی (a, b) دارای یك علامت است، خطای E^R در قاعدهٔ مستطیلی می‌تواند از رابطهٔ (۲۰.۷) محاسبه شود. لذا داریم

$$\boxed{E^R = f'(\eta) \int_a^b (x - a)\,dx = \frac{f'(\eta)(b - a)^2}{2}} \qquad (۲۴.۷)$$

اگر $x_0 = (a+b)/2$، آنگاه $\psi_0(x)$ دارای یك علامت نخواهد بود. ولی در این صورت داریم

---

1. rectangle rule



شکل ۲.۷ انتگرال عددی

$$\int_a^b (x-x_0)\,dx = 0$$

درحالی‌که $(x-x_0)^2$ دارای یك علامت‌است. بنابراین، دراین حالت مقدار خطا در $I(p_k)$ را می‌توان از (۲۲.۷) به‌ازای $x_1 = x_0$ محاسبه‌کرد. که خواهیم داشت

$$I(f) \approx M = (b-a) f\left(\frac{a+b}{2}\right)$$

$$E^M = \frac{f''(\eta)(b-a)^3}{24}$$

به‌ازای مقداری مانند $\eta \in (a,b)$

(۲۵.۷)

که به‌**قاعدهٔ نقطهٔ وسط**[1] موسوم‌است.

سپس گیریم $k = 1$، آنگاه

$$f(x) = f(x_0) + f[x_0, x_1](x-x_0) + f[x_0, x_1, x]\psi_1(x)$$

برای به‌دست‌آوردن $\psi_1(x) = (x-x_0)(x-x_1)$ که روی (a,b) دارای یك‌علامت

1. midpoint rule

باشد، مقادیر $x_0 = a$ و $x_1 = b$ را انتخاب می‌کنیم. در این صورت، با توجه به رابطهٔ (۲۰.۷) خواهیم داشت

$$I(f) = \int_a^b \{f(a) + f[a, b](x-a)\}\, dx + \frac{1}{2} f''(\eta) \int_a^b (x-a)(x-b)\, dx$$

یا

$$\boxed{I(f) \approx T = \frac{1}{2}(b-a)[f(a)+f(b)]}$$

$$\boxed{E^T = -\frac{f''(\eta)(b-a)^3}{12} \quad \text{به‌ازای مقداری مانند } \eta \in (a, b)}$$

(۲۶.۷)

که به‌**قاعدهٔ ذوزنقه‌ای**[1] موسوم است (به‌شکل ۲.۷ نگاه کنید).

اکنون، گیریم $k = 2$، در این صورت

$$f(x) = p_2(x) + f[x_0, x_1, x_2, x]\psi_2(x)$$

باید توجه داشت که برای نقاط متمایز x_0، x_1، x_2 در بازهٔ (a, b)، تابع

$$\psi_2(x) = (x-x_0)(x-x_1)(x-x_2)$$

روی (a, b) دارای یك علامت نیست. اما اگر $x_0 = a$ و $x_1 = (a+b)/2$ و $x_2 = b$ انتخاب کنیم، آنگاه با انتگرالگیری مستقیم یا استدلالهای متقارن، می‌توان نشان داد که

$$\int_a^b \psi_2(x)\, dx = \int_a^b (x-a)(x-(a+b)/2)(x-b)\, dx = 0$$

که در این حالت خطا به‌شکل رابطهٔ (۲۲.۷) خواهد بود. اگر $x_2 = x_1 = (a+b)/2$ انتخاب کنیم، آنگاه داریم

$$\psi_2(x) = (x-a)\left(x-\frac{a+b}{2}\right)^2 (x-b)$$

که روی (a, b) دارای یك علامت است. بنابراین از روابط (۱۸.۷) و (۲۲.۷) نتیجه می‌شود که

$$I(f) = I(p_2) + \frac{1}{4!} f^{iv}(\eta) \int_a^b \psi_2(x)\, dx$$

---

1. trapezoid (al) rule

رابطهٔ زیر مستقیماً محاسبه می‌شود

$$\int_a^b \psi_2(x)\,dx=\int_a^b (x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)\,dx=-\frac{4}{15}\left(\frac{b-a}{2}\right)^5$$

لذا مقدار خطا دراین فرمول به‌صورت زیر درخواهد آمد

$$E^S(f)=\frac{-1}{90}\left(\frac{b-a}{2}\right)^5 f^{iv}(\eta)\qquad \eta\in[a,b]$$

اکنون برای به‌دست آوردن فرمولی متناظر با حالت $k=2$، با انتخاب نقـاط درونیاب $x_0=a$، $x_1=(a+b)/2$، و $x_2=b$، تابع $I(p_2)$ رامستقیماً محاسبه می‌کنیم. بهتر است بسجمله‌ای درونیاب را به‌شکل زیر بنویسیم

$$p_2(x)=f(a)+f[a,b](x-a)+f\left[a,b,\frac{a+b}{2}\right](x-a)(x+b)$$

بنابراین داریم

$$\int_a^b p_2(x)\,dx=f(a)(b-a)+f[a,b](b-a)^2/2$$
$$+f\left[a,b,\frac{a+b}{2}\right]\int_a^b (x-a)(x-b)\,dx$$

اما همان‌گونه‌که هنگام به‌دست‌آوردن رابطهٔ (۲۶.۷)، پیدا کردیم،

$$\int_a^b (x-a)(x-b)\,dx=-(b-a)^3/6$$

بنابراین

$$\int_a^b p_2(x)\,dx=f(a)(b-a)+f[a,b](b-a)^2/2$$
$$-f\left[a,\frac{a+b}{2},b\right](b-a)^3/6 \qquad (۲۷.۷)$$

که در رابطهٔ بالا از تساوی

$$f\left[a,b,\frac{a+b}{2}\right]=f\left[a,\frac{a+b}{2},b\right]$$

که به‌علت وجود تقارن در تفاضل منقسم برقراراست، استفاده شده است. اما اکنون داریم

$$f[a,b](b-a)=f(b)-f(a)$$

در حالی که

$$f\left[a,\frac{a+b}{۲},b\right](b-a)^{۲}=\left(f\left[\frac{a+b}{۲},b\right]-f\left[a,\frac{a+b}{۲}\right]\right)(b-a)$$

$$=۲\left(f(b)-f\left(\frac{a+b}{۲}\right)-\left(f\left(\frac{a+b}{۲}\right)-f(a)\right)\right)$$

از گذاردن این عبارتها در (۲۷.۷) خواهیم داشت

$$\int_a^b p_{۲}(x)\,dx=(b-a)\left\{f(a)+(f(b)-f(a))/۲\right.$$

$$\left.-۲\left(f(b)-۲f\left(\frac{a+b}{۲}\right)+f(a)\right)/۶\right\}$$

$$=\frac{b-a}{۶}\left\{f(a)+۴f\left(\frac{a+b}{۲}\right)+f(b)\right\}$$

بنابراین درست به **قاعدهٔ سیمپسن** رسیدیم که مقدار خطای متناظر آن چنین‌است

$$\boxed{I(f)\approx S=\frac{b-a}{۶}\left\{f(a)+۴f\left(\frac{a+b}{۲}\right)+f(b)\right\}}$$

$$\boxed{E^S=-\frac{f^{iv}(\eta)[b-a)/۲]^{۵}}{۹۰}} \qquad (۲۸.۷)$$

و سرانجام می‌گیریم $k=۳$. دراین صورت

$$f(x)=p_{۳}(x)+f[x_{۰},x_{۱},x_{۲},x_{۳},x]\psi_{۳}(x)$$

با انتخاب $x_{۰}=x_{۱}=a$، $x_{۲}=x_{۳}=b$، می‌توان اطمینان داشت که تابع

$$\psi_{۳}(x)=(x-a)^{۲}(x-b)^{۲}$$

در بازهٔ (a,b) یك علامت دارد و از این‌رو ازرابطهٔ (۲۰.۷)، نتیجه می‌گیریم کـه خطا را می‌توان به‌صورت زیربیان کرد

$$E(f)=\frac{۱}{۴!}f^{iv}(\eta)\int_a^b(x-a)^{۲}(x-b)^{۲}\,dx=\frac{f^{iv}(\eta)(b-a)^{۵}}{۷۲۰}$$

برای به‌دست آوردن یك فرمول انتگرالگیری متناظر با انتخاب نقاط $x_0 = x_1 = a$، $x_2 = x_3 = b$، در ابتدا مشاهده می‌کنیم که

$$p_3(x) = f[a] + f[a, a](x-a) + f[a, a, b](x-a)^2 + f[a, a, b, b](x-a)^2(x-b)$$

لذا

$$\int_a^b p_3(x)\,dx = f(a)(b-a) + f[a, a]\frac{(b-a)^2}{2} + f[a, a, b]\frac{(b-a)^3}{3} + f[a, a, b, b]\left\{\frac{(b-a)^4}{4} - \frac{(b-a)^4}{3}\right\} \qquad (۲۹.۷ \text{ الف})$$

براساس درونیابی بوسانی بخش۷.۲، داریم

$$f[a, a] = f'(a)$$

$$f[a, a, b] = \{f[a, b] - f'(a)\}/(b-a)$$

$$f[a, a, b, b] = \{f'(b) - 2f[a, b] + f'(a)\}/(b-a)^2$$

با گذاردن روابط بالا در (۲۹.۷ الف) و ساده‌کردن آن خواهیم داشت

$$\int_a^b p_3(x)\,dx = f(a)(b-a) + f'(a)\frac{(b-a)^2}{2} + \{f[a, b] - f'(a)\}\frac{(b-a)^2}{3} - \{f'(b) - 2f[a, b] + f'(a)\}\frac{(b-a)^2}{12}$$

و بالاخره ازقراردادن $(f(b)-f(a))/(b-a)$ به‌جای $f[a, b]$ و مرتب کردن رابطهٔ حاصله برحسب توانهای $(b-a)$ به‌فرمول زیر می‌رسیم

$$I(f) \approx CT = \frac{(b-a)}{2}[f(a)+f(b)] + \frac{(b-a)^2}{12}[f'(a) - f'(b)] \qquad (۲۹.۷ \text{ ب})$$

که به‌دلایلی کـه محتاج به‌توضیح نیست، به‌قاعدهٔ ذوزنقه‌یی اصلاح شده معروف است. خطای قاعدهٔ ذوزنقه‌یی تصحیح شده چنین‌است

$$E^{CT}=\frac{f^{iv}(\eta)(b-a)^{۵}}{۷۲۰}$$

البته اگر قواعد فوق‌الذکر در مورد انتگرالگیری عددی، تقریب رضایتبخشی برای $I(f)$ به‌دست ندهند، می‌توان درجهٔ kای بسجمله‌ای درونیاب به‌کار رفته را افزایش داده خطرات انجام چنین عملی در بخش ۷.۶ موردبحث قرارگرفت و در آنجا به‌کارگیری درونیابی بسجمله‌ای تکه‌ای[1] پیشنهاد شدکه روش بسیار منطقی ومناسبی برای دستیابی به‌دقت بالاست. بنابراین $I(f)$ را به‌کمك $I(g_k)$ تقریب می‌زنیم که در آن $g_k(x)$ یك تابع بسجمله‌ای تکه‌ای از درجهٔ «پایین» k است که $f(x)$ را درونیابی می‌کند. قواعد انتگرالگیری حاصل، که معمولا **قواعد مرکب**[2] نامیده می‌شوند، در قسمت ۴.۷ مورد بحث قرارخواهندگرفت.

در این بخش پنج قاعدهٔ اصلی انتگرالگیری را به‌دست آورده‌ایم. این قاعده‌ها عبارت‌اند ازقاعدهٔ مستطیلی(۲۴.۷)، قاعدهٔ نقطهٔ وسط(۲۵.۷)، قاعدهٔ ذوزنقه‌یی (۲۶.۷)، قاعدهٔ سیمپسن (۲۸.۷) و قاعدهٔ ذوزنقه‌یی اصلاح شده (۲۹.۷). از میان قواعد مذکور تنها، قاعدهٔ ذوزنقه‌یی اصلاح شده‌است که به‌مشتق $f(x)$ نیاز دارد و این امر یك عیب آشکار براین روش خاص‌است. عبارتهای خطای مربوط به‌این قواعد نشان می‌دهندکه هرگاه تابع $f(x)$ به‌اندازهٔ کافی هموار باشد، استفاده از قاعدهٔ سیمپسن و یا قاعدهٔ ذوزنقه‌یی اصلاح شده، می‌باید ارجح باشد. با وجود این، توابعی وجود دارندکه فرمولهای پایین‌مرتبه‌نسبت به‌فرمولهای بالامرتبه، نتیجهٔ بهتری‌را به‌دست می‌دهند. [به‌تمرین ۲.۷–۲ نگاه‌کنید].

□ **مثال ۱.۷:** برای پیداکردن برآوردی برای انتگرال زیر، ازهریك از پنج قاعدهٔ مذکور دربالا استفاده‌کنید

$$I=\int_{۰}^{۱}e^{-x^{۲}}dx$$

مقادیر $a=۰$، $b=۱$، $(a+b)/۲=۱/۲$ را برمی‌گزینیم و با توجه به‌جدول مقادیر خواهیم داشت

$$f(۰)=۱ \qquad f(۱)=e^{-۱}=۰٫۳۶۷۸۸ \qquad f\left(\frac{۱}{۲}\right)=e^{-۱/۴}=۰٫۷۷۸۸۰$$

در محاسبات به‌مقادیر زیرنیز احتیاج داریم

$$f'(۰)=۰ \qquad f'(۱)=-۲e^{-۱}=-۰٫۷۳۵۷۶$$

1. piecewise-polynomial interpolation 2. composite rule

دراین‌صورت می‌توانیم مقادیر زیر را از فرمولهای مناسب، محاسبه کنیم

$$R=۱\times e^{۰}=۱$$

$$M=۱\times e^{-۱/۴}=۰٫۷۷۸۸۰$$

$$T=\frac{۱}{۲}[e^{۰}+e^{-۱}]=۰٫۶۸۳۹۴$$

$$S=\frac{۱}{۶}[e^{۰}+۲e^{-۱/۴}+e^{-۱}]=۰٫۷۴۷۱۸$$

$$CT=\frac{۱}{۲}[e^{۰}+e^{-۱}]+\frac{۱}{۱۲}[۰+۲e^{-۱}]=۰٫۷۴۵۲۵$$

مقدار انتگرال تا پنج رقم صحیح اعشاری برابر با $I=۰٫۷۴۶۸۲$ است. چنانچه از بررسی عبارتهای خطا برمی‌آید، واین حقیقت که چند مشتق اول تابع $e^{-x^{۲}}$ ازنظر مقدار خیلی تغییر نمی‌کنند، بدیهی است که قاعدهٔ ذوزنقه‌یی اصلاح شده (CT) و قاعدهٔ سیمپسن (S) بهترین نتایج را به‌دست می‌دهند. □

تمرین

۱-۲۰۷ اگر تابع

$$\psi_{۲}(x)=(x-a)(x-(a+b)/۲)(x-b)$$

مفروض باشد، با روش انتگرالگیری مستقیم تحقیق کنید $\int_{a}^{b}\psi_{۲}(x)\,dx=۰$.

۲-۲۰۷ با استفاده ازهریك از پنج قاعده‌ای کـه در این بخش بیان کردیم، تقریبی برای $I=\int_{۰}^{۱}x\sin x\,dx$ به‌دست آورید. نتایج حاصله را با مقدار صحیح

$$I=\sin ۱-\cos ۱=۰٫۳۰۱۱۶۹$$

مقایسه کنید.

۳-۲۰۷ تابع $f(x)$ بربازهٔ $[۰,۱]$ به‌شرح زیرتعریف شده‌است

$$f(x)=\begin{cases}x & ۰\leqslant x\leqslant ۱/۲\\ ۱-x & ۱/۲\leqslant x\leqslant ۱\end{cases}$$

با استفاده از قواعد زیر، مقدار $\int_{۰}^{۱}f(x)\,dx$ را محاسبه کنید:
(الف) قاعدهٔ ذوزنقه‌یی بربازهٔ $[۰,۱]$.

(ب) قاعدهٔ ذوزنقه‌یی ابتدا بر بازهٔ $[۰, ۱/۲]$ و سپس بر بازهٔ $[۱/۲, ۱]$.
(پ) قاعدهٔ سیمپسن بر بازهٔ $[۰, ۱]$.
(ت) قاعدهٔ ذوزنقه‌یی اصلاح‌شده بر بازهٔ $[۰, ۱]$.
علت وجود تفاوت در نتایج حاصله را بیان کنید.

۴-۲.۷ با ملاحظهٔ اینکه $p_۳(x)$ یك بسجمله‌ای از درجهٔ ۳ است و $p_۳^{iv}(x)=۰$، و نیز اینکه قاعدهٔ سیمپسن (۲۸.۷) می‌تواند دقیقاً برای محاسبهٔ $I(p_۳)$ به‌کار رود، در نتیجه قاعدهٔ ذوزنقه‌یی اصلاح شده را می‌توان بسیار ساده‌تر به‌دست آورد. بنابراین

$$I(p_۳)=\frac{b-a}{۶}\left\{p_۳(a)+۴p_۳\left(\frac{a+b}{۲}\right)+p_۳(b)\right\}$$

چون $p_۳(x)$، درونیاب $f(x)$ در نقاط a، a، b، b است، باید داشته باشیم

$$p_۳(a)=f(a),\ p_۳(b)=f(b)$$

با استفاده از نتایج حاصله از بخش۷.۲ درمورد درونیابی بوسانی، نشان دهیدکه

$$p_۳\left(\frac{a+b}{۲}\right)=\frac{۱}{۲}[f(a)+f(b)]+\frac{b-a}{۸}[f'(a)-f'(b)]$$

سپس‌رابطهٔ بالا را دررابطهٔ $I(p_۳)$ قراردهید تا قاعدهٔ ذوزنقه‌یی اصلاح شدهٔ (۲۹.۷ ب) به‌دست آید.

۵-۲.۷ برای براورد مقدار انتگرال

$$I=\int_۰^۱(۱-x^۲)^{۳/۲}\,dx$$

از قاعدهٔ سیمپسن استفاده‌کنید.

۶-۲.۷ برای براورد مقدار انتگرال $I=\int_۰^۱ xe^{-x^۲}\,dx$، قاعدهٔ ذوزنقه‌یی‌را به‌کار برید. کرانی را درخطای حاصل از قاعدهٔ ذوزنقه‌یی (۲۶.۷) به‌دست آورید و آن را باخطای واقعی مقایسه‌کنید.

۳.۷ انتگرالگیری عددی: قواعدگاوسی

همهٔ قواعدی راکه دربخش۲.۷ استخراج کردیم، به‌استثنای قانون ذوزنقه‌یی اصلاح شده، می‌توان به‌شکل زیرنوشت

$$I(g)\approx A_۰g(x_۰)+A_۱g(x_۱)+\cdots+A_kg(x_k) \qquad (۳۰.۷)$$

که در آن وزنهای A_0، ...، A_k به‌تابع خاص $g(x)$ بستگی ندارند. تا به‌حال به‌نحوی رأسهای[1] x_0، ...، x_k را انتخاب کرده‌ایم، مثلاً آنها را درجدولی با فواصل مساوی در نظر گرفتیم و سپس به‌ازای جمیع مقادیر i، وزنهای A_i را به‌صورت $I(l_i)$ محاسبه کردیم. این عمل تضمین می‌کند که قاعدهٔ مزبور برای بسجمله‌هایی نابیشتر از درجهٔ k دقیق باشد. اما این امکان وجود دارد که با انتخاب رأسهای مناسب، این قاعده را برای بسجمله‌هایی نابیشتر از درجهٔ $2k-1$ نیز دقیق درآورد. این مسئله زیربنای فکری قواعد گاوس است.

قواعد حاصله پیچیده‌تر از قواعد مستخرجه در بخش ۲.۷ به‌نظر می‌رسند. در این قواعد، هم رأسها و هم وزنها، درحالت کلی، اعداد گنگ[2] هستند. این امر موجب می‌شود که وقتی بنا باشد افراد محاسبات را به‌وسیلهٔ دست انجام دهند، عموماً از به‌کاربردن این قواعد خودداری کنند. اما با یك کامپیوتر، معمولا تفاوتی ندارد که محاسبهٔ یك تابع در $x=\pm 1/\sqrt{3} \approx \pm 0.57735\ 02691$ انجام گیرد یا $x=3$. هنگامی که رأسها و وزنها درچنین قواعدی به‌شکل خاصی (مثلا به‌صورت، زیربرنامهٔ LGNDRE که در زیرآمده است) ذخیره شده باشند، این قواعد به‌همان آسانی قاعدهٔ ذوزنقه‌یی یا قاعدهٔ سیمپسن به‌کار برده می‌شوند. درعین‌حال، براساس تعداد مقادیری که برای تابع به‌کاربرده شده، معمولا قواعد گاوسی درمقایسه با قواعد بخش ۲.۷ بسیار دقیقترند.

ما این قواعد گاوسی را در قالب کلیتر یك انتگرال $\int_a^b f(x)\,dx$ که در آن عامل زیرعلامت انتگرال، $f(x)$، به‌اندازهٔ کافی برای توجیه‌کاربرد قواعد بخش ۲.۷ مشتقپذیر نیست، موردبحث قرارمی‌دهیم. مثلا $f(x)$ به‌ازای مقداری مانند $\alpha > -1$ ودرنزدیکی نقطهٔ a، ممکن است به‌صورت $(x-a)^\alpha$ عمل کند یا آنکه a و(یا) b ممکن است نامتناهی باشند. غالباً دراین‌گونه موارد می‌توان انتگرال را به‌صورت

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

نوشت، که درآن $w(x)$ تابعی است نامنفی انتگرالپذیر و

$$g(x) = \frac{f(x)}{w(x)}$$

تابعی است هموار. درمثال بالا، $w(x)=(x-a)^\alpha$، همین وضع را دارد. انتخابهای دیگری که برای $w(x)$ صورت می‌گیرد در زیرموردبحث قرارمی‌گیرد. حالتی که تابع زیرعلامت انتگرال تابع بی‌دردسری است نیز در اینجا باانتخاب سادهٔ $w(x)\equiv 1$، موردبحث قرار می‌گیرد.

حال محاسبهٔ تقریبی انتگرال وزین

$$I(g)=\int_a^b g(x)w(x)\,dx \qquad (۳۱.۷)$$

---

1. nodes 2. irrational

را با استفاده از قاعده‌ای به‌شکل (۳۰.۷) در نظر می‌گیریم. گوییم که قاعدهٔ (۳۰.۷) برای تابع خاص $p(x)$ صحیح است، اگر با قرار دادن $p(x)$ به‌جای $g(x)$ در (۳۰.۷)، این رابطه به‌یك تساوی بدل شود. برای مثال، قاعدهٔ ذوزنقه‌یی

$$\int_a^b g(x)\,dx \approx \frac{b-a}{۲} g(a) + \frac{b-a}{۲} g(b)$$

برای همهٔ بسجمله‌های ازدرجهٔ نابیشتر از یك، صحیح‌است. برای بررسی این امر فقط کافی است به‌عبارت خطای این قاعده، یعنی

$$E^T = \frac{g''(\eta)(b-a)^۳}{۱۲}$$

نگاه‌کنیم. ازآنجاکه این عبارت خطا متضمن مشتق دوم $g(x)$ است، و چون مشتق دوم هربسجمله‌ای ازدرجهٔ نابیشتر از یك برابر باصفر است، لذانتیجه می‌شودکه هرگاه $g(x)$ یك بسجمله‌ای از درجهٔ نابیشتر ازیك باشد، خطا برابر باصفر است. به‌صورت کلیتر، اگر عبارت خطای (۳۰.۷) به‌شکل

$$E = g^{(r+۱)}(\eta) \times \text{ثابت}(\text{تابعی از } x_۰, \ldots, x_k) \qquad (۳۲.۷)$$

باشد، آنگاه قاعدهٔ (۳۰.۷) باید برای تمامی بسجمله‌ایهای نابیشتر از درجهٔ r، صحیح باشد.

بنابراین، اگر بخواهیم قاعده‌ای به‌شکل (۳۰.۷) بسازیم که، با kی ثابت برای بسجمله‌ایهای از درجات بالاتر صحیح باشد، می‌باید قاعده‌ای بسازیم که عبارت خطای آن به‌شکل (۳۲.۷)، یعنی به‌صورت یك عدد صحیح تاحد امکان بزرگ، باشد. این‌کار را می‌توانیم با ترفندی‌که دربخش ۲.۷ به‌کاربردیم انجام دهیم.

مانند بخش (۲.۷)، از جایگذاری تحلیلی استفاده می‌کنیم یعنی نقاط $x_۰, \ldots, x_k$ را دربازهٔ (a, b) انتخاب می‌کنیم و می‌نویسیم

$$g(x) = p_k(x) + g[x_۰, \ldots, x_k, x]\psi_k(x)$$

که درآن $p_k(x)$ یك بسجمله‌ای است از درجهٔ نابیشتر از kکه درونیاب $g(x)$ در نقاط $x_۰, \ldots, x_k$ است و داریم

$$\psi_k(x) = (x-x_۰)\cdots(x-x_k)$$

که نتیجه می‌دهد

$$I(g) = I(p_k) + \int_a^b g[x_۰, \ldots, x_k, x]\psi_k(x)w(x)\,dx$$

واضح است که تقریب $I(p_k)$ برای $I(g)$ به‌شکل (۳۰.۷) است. زیرا اگر $p_k(x)$ را به‌شکل لاگرانژ بنویسیم (به بخش ۲.۲ نگاه کنید)

$$p_k(x)=g(x_0)l_0(x)+g(x_1)l_1(x)+\cdots+g(x_k)l_k(x)$$

که در آن داریم

$$l_i(x)=\prod_{\substack{j=0\\ j\neq i}}^{k}\frac{x-x_j}{x_i-x_j} \qquad i=0,\ldots,k$$

در این صورت

$$I(p_k)=\int_a^b p_k(x)w(x)\,dx$$

$$=g(x_0)\int_a^b l_0(x)w(x)\,dx+\cdots+g(x_k)\int_a^b l_k(x)w(x)\,dx$$

بنابراین

$$I(p_k)=A_0g(x_0)+A_1g(x_1)+\cdots+A_kg(x_k) \qquad (۳۳.۷)$$

که در آن

$$A_i=\int_a^b l_i(x)w(x)\,dx \qquad i=0,\ldots,k \qquad (۳۴.۷)$$

سپس خطای

$$I(g)-I(p_k)=\int_a^b g[x_0,\ldots,x_k,x]\psi_k(x)w(x)\,dx$$

را در نظر می‌گیریم. فرض کنید که

$$\int_a^b \psi_k(x)w(x)\,dx=0$$

سپس همان‌گونه که در بخش ۲.۷ بحث شد، به‌ازای هر انتخابی از x_{k+1} داریم

$$I(g)-I(p_k)=\int_a^b g[x_0,\ldots,x_k,x_{k+1},x]\psi_{k+1}(x)w(x)\,dx$$

اکنون اگر تساوی

$$\int_a^b \psi_{k+1}(x)w(x)\,dx=0$$

نیز برقرار باشد، آنگاه به‌همان دلیل داریم

$$I(g)-I(p_k)=\int_a^b g[x_0,\cdots,x_{k+2},x]\psi_{k+2}(x)w(x)\,dx$$

بنابراین، درحالت کلی، اگر برای نقاطی مانند x_0، ...، x_{k+m} داشته باشیم

$$\int_a^b \psi_k(x)(x-x_{k+1})\cdots(x-x_{k+1+i})w(x)\,dx=0$$

$$i=0,\cdots,m-1 \qquad (۳۵.۷)$$

آنگاه برای هر انتخابی از x_{k+m+1} داریم

$$I(g)-I(p_k)=\int_a^b g[x_0,\cdots,x_{k+m+1},x]\psi_{k+m+1}(x)w(x)\,dx \qquad (۳۶.۷)$$

اکنون با توجه به‌بخش ۶.۶ یادآوری می‌کنیم که برای بسیاری از $w(x)$ها می‌توانیم یك بسجمله‌ای $p_{k+1}(x)$ چنان پیدا کنیم که به‌ازای تمام بسجمله‌ایهای $q(x)$ نابیشتر از درجهٔ k (به‌ویژگی ۳ی بسجمله‌ایهای متعامد در بخش ۶.۶ مراجعه کنید) داشته باشیم

$$\int_a^b p_{k+1}(x)q(x)w(x)\,dx=0 \qquad (۳۷.۷)$$

بعلاوه، بنابر ویژگی ۲ی بسجمله‌ایهای متعامد، داریم

$$p_{k+1},x)=\alpha_{k+1}(x-\xi_0)(x-\xi_1)\cdots(x-\xi_k)$$

که در آن ξ_0، ...، ξ_k، $k+1$ نقطهٔ متمایز در بازهٔ (a,b) هستند که در آنها بسجمله‌ای p_{k+1} صفر می‌شود. لذا اگر بگیریم

$$x_j=\xi_j \qquad j=0,\cdots,k \qquad (۳۸.۷)$$

و به‌ازای $j=1,\cdots,k+1$، x_{k+j} را نقاط انتخابی در بازهٔ (a,b) برگزینیم، آنگاه به‌ازای $m=k$، رابطهٔ (۳۵.۷) ودرنتیجه رابطهٔ (۳۶.۷) برقرارخواهد بود. به‌ازای تمام بسجمله‌ایهای نابیشتر از درجهٔ k، رابطهٔ (۳۵.۷) به‌شکل (۳۷.۷) خواهد بود که در آن به‌ازای $m\leqslant k$، داریم

$$q(x)=\frac{(x-x_{k+1})\cdots(x-x_{k+1+i})}{\alpha_{k+1}} \qquad i=0,\cdots,m-1$$

بنابراین

$$I(g)-I(p_k)=\int_a^b g[x_0,\cdots,x_{2k+1},x]\psi_{2k+1}(x)w(x)\,dx \qquad (۳۹.۷)$$

برای به‌دست‌آوردن این خطا به‌شکل(۳۲.۷)، x_{k+j}ها را به‌صورت زیر انتخاب می‌کنیم

$$x_{k+j}=\xi_{j-1} \qquad j=1,\dots,k+1$$

دراین‌صورت خواهیم داشت

$$\begin{aligned}\psi_{2k+1}(x)&=(x-x_0)\dots(x-x_{2k+1})\\&=(x-\xi_0)\dots(x-\xi_k)(x-\xi_0)\dots(x-\xi_k)\\&=\left[\frac{P_{k+1}(x)}{\alpha_{k+1}}\right]^2\end{aligned}$$

به‌طوری‌که $\psi_{2k+1}(x)w(x)$ دارای یك علامت، یعنی‌علامت نامنفی‌در بازهٔ (a,b) خواهد بود. بنابراین می‌توان قضیهٔ مقدار میانگین برای انتگرال را به‌کاربرد (به‌بخش ۷۰۱ نگاه کنید) و رابطهٔ

$$I(g)-I(p_k)=g[x_0,\dots,x_{2k+1},\eta]\int_a^b\left[\frac{1}{\alpha_{k+1}}P_{k+1}(x)\right]^2w(x)\,dx$$

را به‌دست آورد. و سرانجام اگر $g(x)$ به‌طور پیوسته، $2k+2$مرتبه مشتقپذیر باشد، می‌توانیم قضیهٔ ۵.۲ را به‌کارگیریم و خطا را به‌شکل زیربیان کنیم

$$I(g)-I(p_k)=\frac{1}{(2k+2)!}g^{(2k+2)}(\xi)\frac{S_{k+1}}{\alpha_{k+1}^2} \qquad (۴۰.۷)$$

که در آن داریم

$$S_{k+1}=\int_a^b[P_{k+1}(x)]^2w(x)\,dx$$

به‌طورخلاصه، نشان دادیم که اگرنقاط $x_0,\dots,x_k$ را در رابطهٔ(۳۳.۷) به‌عنوان ریشه‌های بسجمله‌ای $P_{k+1}(x)$، از درجهٔ $(k+1)$ام، برگزینیم و این بسجمله‌ای نسبت به‌تابع وزین $w(x)$ در بازهٔ (a,b) برهر بسجمله‌ای نابیشتر ازدرجهٔ k متعامد باشد و اگر ضرایب $A_i(i=0,\dots,k)$ در رابطهٔ (۳۳.۷) براساس رابطهٔ (۳۴.۷) انتخاب شوند، آنگاه‌فرمول گاوس حاصل، (۳۳.۷)، برای‌کلیهٔ بسجمله‌ایهای نابیشتر از درجهٔ $2k+1$ صحیح‌است. قواعد انتگرالگیری از این نوع را، به‌تعبیر ما ودر شرایط مذکور در بالا، «بهترین قاعدهٔ ممکن» گویند.

اکنون مثالهایی را بررسی می‌کنیم. در ابتدا می‌گیریم $w(x)=1$. اگر(a,b) بازهٔ متناهی باشد،آنگاه برای تغییرحدود انتگرالگیری از (a,b) تا $(-1,1)$، از تغییرخطی متغیرها به‌صورت $x=[(b-a)t+(b+a)]/2$ استفاده می‌کنیم. با انجام این‌کار

خواهیم داشت

$$\int_a^b f(x)\,dx = \int_{-1}^{1} f(x(t))x'(t)\,dt = \int_{-1}^{1} f(x(t))\frac{b-a}{2}\,dt \quad (۴۱.۷)$$

با فرض اینکه این تغییر قبلا انجام گرفته باشد، انتگرال (۳۱.۷) را که به شکل زیر است درنظر می‌گیریم

$$I(g)=\int_{-1}^{1} g(x)w(x)\,dx$$

ازآنجایی که داریم $w(x)=1$، بسجمله‌ایهایی متعامد مناسب بسجمله‌ایهای لژاندر هستند (به‌مثال ۶.۶ نگاه کنید). دراین‌حالت

$$P_1(x)=x \qquad \xi_0=0$$

$$P_2(x)=\frac{3}{2}\left(x^2-\frac{1}{3}\right) \qquad \xi_0=-\frac{1}{\sqrt{3}},\ \xi_1=\frac{1}{\sqrt{3}}$$

$$P_3(x)=\frac{5}{2}\left(x^3-\frac{3}{5}x\right) \qquad \xi_0=-\sqrt{\frac{3}{5}},\ \xi_1=0,\ \xi_2=\sqrt{\frac{3}{5}}$$

و غیره. اگر $k=1$ انتخاب شود، آنگاه $x_0=\xi_0=-1/\sqrt{3}$، $x_1=\xi_1=1/\sqrt{3}$ و با قراردادن این مقادیر در روابط (۳۳.۷) و (۴۰.۷) خواهیم داشت

$$\int_{-1}^{1} g(x)\,dx \approx A_0 g\left(\frac{-1}{\sqrt{3}}\right)+A_1 g\left(\frac{1}{\sqrt{3}}\right) \quad (۴۲.۷)$$

$$E=Cg^{iv}(\eta)$$

که در آن

$$A_1=\int_{-1}^{1}\frac{x-(-1/\sqrt{3})}{1/\sqrt{3}-(-1/\sqrt{3})}\,dx=1$$

$$A_0=\int_{-1}^{1}\frac{x-1/\sqrt{3}}{(-1/\sqrt{3})-1/\sqrt{3}}\,dx=1$$

$$C=\frac{1}{4!}\frac{S_2}{\alpha_2^2}=\frac{1}{24}\times\frac{2}{5}\times\frac{4}{9}=\frac{1}{135}$$

چون

$$S_۲=\int_{-۱}^{۱}[P_۲(x)]^۲dx=\frac{۲}{۵}\qquad \alpha_۲=\frac{۳}{۲}$$

با گذاردن این مقادیر ثابت در (۴۲.۷)، فرمول انتگرال دونقطه‌ای گاوس

$$\int_{-۱}^{۱} g(x)\,dx\approx g\left(\frac{-۱}{\sqrt{۳}}\right)+g\left(\frac{۱}{\sqrt{۳}}\right) \qquad (۴۳.۷)$$

را با خطای آن

$$E=\frac{۱}{۱۳۵}g^{iv}(\eta) \qquad (۴۴.۷)$$

به‌دست خواهیم آورد.

به‌ازای $k>۱$، هم نقاط x_i و هم وزنهای A_i، هردو، اعداد گنگ می‌شوند. ولی محاسبهٔ آنها ساده است. این رأسها و وزنها به‌ازای $k=۰,\ldots,۵$ در زیربرنامهٔ فرعی فورترن به‌نام LGNDRE ذخیره می‌شوند. توجه کنید که محدودیتهای(قیلی) فورترن ما را وادار نموده که رأسها و وزنها را به‌جای از ۰ تا k، از ۱ تا $NP=k+۱$ شماره‌گذاری کنیم. بنابراین پارامتر ورودی NP به‌جای مشخص‌نمودن درجهٔ بسجمله‌ای مورد بررسی، به‌مشخص کردن تعداد نقاط می‌پردازد.

```
      SUBROUTINE LGNDRE ( NP , POINT, WEIGHT )
C  SUPPLIES POINTS AND WEIGHTS FOR GAUSS-LEGENDRE QUADRATURE
C  INTEGRAL(F(X), -1 .LE. X .LE. 1)  IS APPROXIMATELY EQUAL TO
C           SUM(F(POINT(I))*WEIGHT(I), I=1,...,NP) .
      INTEGER NP,  I
      REAL POINT(NP),WEIGHT(NP)
      IF (NP .GT. 6) THEN
         PRINT 600,NP
  600    FORMAT(' THE GIVEN NUMBER NP =',I2,' IS GREATER THAN 6.'
     *        /' EXECUTION STOPPED IN SUBROUTINE  L G N D R E .')
                                          STOP
      END IF
                                          GO TO (1,2,3,4,5,6),NP
    1 POINT(1) = 0.
      WEIGHT(1) = 2.
                                          GO TO 99
    2 POINT(2) = .57735 02691 89626 D0
      WEIGHT(2) = 1.
                                          GO TO 95
    3 POINT(2) = 0.
      POINT(3) = .77459 66692 41483 D0
      WEIGHT(2) = .88888 88888 88888 9 D0
      WEIGHT(3) = .55555 55555 55555 6 D0
                                          GO TO 95
    4 POINT(3) = .33998 10435 84856 D0
      POINT(4) = .86113 63115 94053 D0
      WEIGHT(3) = .65214 51548 62546 D0
      WEIGHT(4) = .34785 48451 37454 D0
                                          GO TO 95
    5 POINT(3) = 0.
      POINT(4) = .53846 93101 05683 D0
      POINT(5) = .90617 98459 38664 D0
      WEIGHT(3) = .56888 88888 88888 9 D0
      WEIGHT(4) = .47862 86704 99366 D0
      WEIGHT(5) = .23692 68850 56189 D0
                                          GO TO 95
```

```
      6 POINT(4) = .23861 91860 83197 D0
        POINT(5) = .66120 93864 66265 D0
        POINT(6) = .93246 95142 03152 D0
        WEIGHT(4) = .46791 39345 72691 D0
        WEIGHT(5) = .36076 15730 48139 D0
        WEIGHT(6) = .17132 44923 79170 D0
C
     95 DO 96 I=1,NP/2
           POINT(I) = -POINT(NP+1-I)
     96    WEIGHT(I) = WEIGHT(NP+1-I)
     99                                    RETURN
        END
```

□ **مثال ۲.۷:** به‌منظور مقایسه، می‌خواهیم دوباره مقدار $I=\int_0^1 \exp(-x^2)\,dx$ را، ولی این‌بار با استفاده از فرمول پنج‌نقطه‌ای گاوس، محاسبه کنیم $(k=4)$.

در اینجا تغییر متغیرهای لازم (۴۱.۷)، به‌صورت $x=(t+1)/2$ است، بنابراین

$$I=\int_0^1 e^{-x^2}\,dx=\int_{-1}^{1}\frac{1}{2}e^{-(t+1)^2/4}\,dt=\sum_{i=0}^{4}A_i\exp\left(-(\xi_i+1)^2/4\right)/2$$

طبیعتاً برای انجام محاسبات از برنامه‌ای به‌صورت زیر استفاده خواهیم کرد.

```
      EXAMPLE 7.2 GAUSSIAN INTEGRATION
      REAL INTGRL,P(5),WEIGHT(5)
      F(T) = EXP(-(1.+T)**2/4.)/2.
      CALL LGNDRE ( 5, P, WEIGHT )
      INTGRL = WEIGHT(1)*(F(P(1))+F(P(5))) + WEIGHT(2)*(F(P(2))+F(P(4)))
     *       + WEIGHT(3)*F(P(3))
      PRINT 600,INTGRL
  600 FORMAT(' EXAMPLE 7.2, GAUSS QUADRATURE'/' INTEGRAL = ',1PE14.7)
                                          STOP
      END
```

خروجی برنامهٔ فوق به‌گونهٔ زیر است

$$\text{INTEGRAL}=۷۴۶۸۲۴۱۳\text{-}۰۰۱$$

برای دسترسی به‌دقتی که نسبت به‌قانون ذوزنقه‌یی قابل‌قیاس باشد به ۲۸۰۰ قسمت فرعی[1] احتیاج‌است، درحالی‌که قانون سیمپسن به‌تقریباً ۲۰ قسمت فرعی‌نیازدارد. □

□ **مثال ۳.۷:** بـا استفاده از قاعدهٔ انتگرالگیری گاوس به‌ازای $k=3$، برای انتگرال زیرتقریبی بیابید

$$I=\int_1^3\frac{(\sin x)^2}{x}\,dx$$

(مقدار درست $I=۰٫۷۹۴۸۲۵۱۸\ldots$).

دوباره محدودهٔ انتگرال را به‌بازهٔ $[-1,1]$ تغییرمی‌دهیم و این‌کار با تغییر متغیر $x=t+2$ انجام می‌گیرد. این امرموجب می‌شودکه داشته باشیم

$$I=\int_{-1}^{1}\frac{(\sin(t+2))^2}{t+2}\,dt$$

---

1. subdivision

بعد از آنکه برنامهٔ اصلی[1] درمثال۲.۷ را به‌صورت زیر تغییر دادیم

```
F(T) = SIN(T + 2.)**2/ (T + 2.)
CALL LGNDRE ( 4, P, WEIGHT )
INTGRL = WEIGHT (1)*(F(P(1)) + F(P(4))) + WEIGHT(2)*(F(P(2)) + F(P(3)))
```

خروجی به‌گونهٔ زیر خواهد شد

□ INTEGRAL = ۷٫۹۴۸۲۸۳۳-۰۰۱

هنگامی که با انتگرالهای تکین[2] سروکار داریم، به‌ویژه فرمولهایی از نوع گاوس بسیار مفیدند. اگر برای مثال، محاسبهٔ $\int_a^b f(x)\,dx$، که در آن $f(x)$ در a و (یا) b دارای تکینی[3] جبری می‌باشد موردنظر باشد، آنگاه انتگرال را به‌انتگرال زیر تبدیل می‌کنند

$$\int_{-1}^{1} g(x)w(x)\,dx$$

که در آن $w(x)$ به‌ازای توانهای مناسب α و β چنین است:

$$w(x)=(1-x)^{\alpha}(1+x)^{\beta}$$

در این حالت ξ_iها، ریشه‌های بسجمله‌ای مناسب ژاکوبی هستند. در حالت خاص $\alpha=\beta=-(1/2)$، بسجمله‌ایهای حاصل، درست همان بسجمله‌ایهای چبیشف می‌شوند که درمثال۷.۶ معرفی شدند و درقسمت ۱.۶ مورد بحث قرار گرفتند. برای این حالت‌خاص، قاعدهٔ جالب زیر به‌دست می‌آید

$$\int_{-1}^{1} \frac{g(x)}{(1-x^2)^{1/2}}\,dx \approx \frac{\pi}{k+1}\sum_{i=0}^{k} g(\xi_i) \qquad (۴۵۰۷)$$

که در آن همهٔ وزنهای A_i برهم منطبق و ξ_iها نقاط چبیشف هستند [رابطهٔ (۱۸.۶) را ببینید].

$$\xi_i=\cos\left(\frac{2i+1}{k+1}\times\frac{\pi}{2}\right) \qquad i=0,\ldots,k \qquad (۴۶۰۷)$$

اگر بازهٔ انتگرالگیری نیمه‌نامتناهی[4] باشد، گاهی بهتر است که انتگرال‌را به‌شکل زیر تبدیل کنیم

$$\int_{0}^{\infty} g(x)w(x)\,dx$$

---

1. body 2. singular 3. singularity
4. semi_infinite

که در آن

$$w(x)=x^{\alpha}e^{-x}$$

دراین‌حالت ξ_iها ریشه‌های بسجمله‌ای خاص**لاگر**[1] هستند؛ به‌مثال۶.۹ نگاه کنید. وسرانجام انتگرالهایی به‌شکل

$$\int_{-\infty}^{\infty} g(x)e^{-x^2}\,dx$$

را می‌توان اغلب با استفاده از ریشه‌های بسجمله‌ای مناسب هرمیت به خوبی براورد نمود (به‌مثال۸.۶ نگاه کنید).

برای کلیهٔ این مثالها (و مثالهای دیگر) جدولهایی هم برای ξ_iها و هم برای A_iها وجود دارند و احتمالا جدیدترین و وسیعترین آنهـا «فرمولهای انتگـرالگیری گاوسی» استرود و سیکرست[2] هستند، [۲۰] و نیز [۲۷] را ببینید.

تمرین

۱-۳۰۷ قاعدهٔ سیمپسن برای چه بسجمله‌هایی صحیح است؟

۲-۳۰۷ قاعده‌ای به‌شکل

$$I(f)=\int_{-1}^{1} f(x)\,dx \approx A_0 f\left(-\frac{1}{2}\right)+A_1 f(0)+A_2 f\left(\frac{1}{2}\right)$$

بسازیدکه برای تمام بسجمله‌های نابیشتر از درجهٔ ۲ صحیح باشد.

۳-۳۰۷ مقدار انتگرال زیر را تا چهار رقم اعشاری دقیق محاسبه کنید

$$\int_{0}^{1} \frac{\sin \pi x}{[x(1-x)]^{3/2}}\,dx$$

[راهنمایی: انتگرال‌را به‌نحو مناسبی تغییر دهید و از رابطـهٔ (۴۵.۷) و (۴۶.۷) استفاده کنید].

۴-۳۰۷ براوردی برای انتگرال $\int_0^{\infty} e^{-x^2}\,dx$ پیداکنید.

۵-۳۰۷ به‌ازای $k=3$ و با استفاده از ریشه‌های ξ_i که در زیر برنامهٔ LGNDRE داده شده‌اند وزنهای A_i درفرمول گاوسی را بیابید.

---

1. Laguerre 2. Stroud and Secrest

۶-۳۰۷ با استفاده از فرمول پنج نقطه‌ای گاوسی، برای انتگرالهایی که در تمرینهای ۳-۴۰۷ و ۴-۴۰۷ داده شده‌اند براوردی تعیین کنید.

۷-۳۰۷ از تمرین ۷-۳۰۶ استفاده کنید تا نشان دهید که می‌توان رابطهٔ (۳۴۰۷) را به‌ازای $i = ۰, \dots, k$ به‌صورت $A_i = \int_a^b [l_i(x)]^۲ w(x)\,dx$ نیز نوشت. از اینجا نتیجه گیری کنید که وزنهای فرمول گاوسی همواره مثبت‌اند.

۸-۳۰۷ قاعدهٔ لوباتو[1] برای انتگرالگیری $I = \int_{-۱}^{۱} f(x)\,dx$ یك فرمول گاوسی است، به‌استثنای آنکه این قاعده شامل ± ۱ به‌عنوان دو طول[2] ثابت است. این قاعده به‌شکل زیر است [رابطهٔ (۳۱۰۷) را ببینید]

$$I(f) \approx A_۰ f(-۱) + A_۱ f(x_۱) + \cdots + A_{k-۱} f(x_{k-۱}) + A_k f(۱)$$

قاعدهٔ لوباتو را برای حالت $k = ۲$، به‌دست آورید و نشان دهید که این قاعده برای تمام بسجمله‌ایهای نابیشتر از درجهٔ ۳، صحیح است.

۹-۳۰۷ برنامهٔ فرعی LGNDRE را با به‌کاربردن آن درمحاسبهٔ

$$\int_{-۱}^{۱} x^n\,dx = \frac{۱}{(n+۱)}\left(۱ - (-۱)^{n+۱}\right) \qquad n = ۰, ۱, ۲, \cdots$$

امتحان کنید. به‌ازای چه مقادیری از n، قاعدهٔ گاوس-لژاندر روی نقاطی به‌تعداد NP، مقدار انتگرال را دقیقاً به‌دست می‌دهد؟

۴۰۷ انتگرالگیری عددی: قاعده‌های مرکب[3]

قواعد سادهٔ محاسبهٔ انتگرال که در بخشهای پیشین برای براورد

$$I = \int_a^b f(x)\,dx$$

پرورانیده شده بود، بخصوص وقتی بازهٔ $[a, b]$ نسبتاً بزرگ باشد، معمولاً براوردی به‌قدر کافی دقیق به‌دست نمی‌دهد. در عمل مرسوم است که بازهٔ مفروض $[a, b]$ را به N بازهٔ کوچکتر تقسیم می‌کنند و قواعد سادهٔ محاسبهٔ انتگرال را برای هریك از این زیربازه‌ها به‌کار می‌برند. بنابراین بازهٔ $[a, b]$ را به‌بازه‌های کوچکتری چنان تقسیم می‌کنیم که

$$a = x_۰ < x_۱ < x_۲ < \cdots < x_N = b$$

و یك تابع بسجمله‌ای-تکه‌ای (به‌بخش ۷.۶ مراجعه کنید) در نقاط انفصال $\{x_i\}$ را،

---

1. Lobatto 2. abscissa 3. composite rules

$(i=1,\dots,N)P_{i,k}(x)$، با $(i=1,\dots,N-1)$ $g_k(x)$ نشان می‌دهیم. بعلاوه، گیریم معرف یك بسجمله‌ای نابیشتر ازدرجهٔ k باشدکه با $g_k(x)$ دربازهٔ (x_{i-1},x_i) تطابق‌دارد. بنابر قواعد انتگرالگیری می‌دانیم که

$$I(f)=\int_a^b f(x)\,dx=\sum_{i=1}^{N}\int_{x_{i-1}}^{x_i} f(x)\,dx$$

و

$$I(g_k)=\sum_{i=1}^{N}\int_{x_{i-1}}^{x_i} g_k(x)\,dx=\sum_{i=1}^{N}\int_{x_{i-1}}^{x_i} P_{i,k}(x)\,dx$$

بنابراین تقریب‌زدن $I(f)$ با $I(g_k)$، مانند تقریب‌زدن

$$\int_{x_{i-1}}^{x_i} P_{i,k}(x)\,dx \text{ به‌وسیلهٔ } \int_{x_{i-1}}^{x_i} f(x)\,dx \qquad i=1,\dots,N$$

و جمع‌کردن نتایج‌است. بدیهی است‌که درهر زیربازهٔ (x_{i-1},x_i) دقیقاً مانند بخشهای ۲.۷ و۳.۷ عمل‌می‌کنیم. بخصوص‌اینکه می‌توانیم درهر زیربازه باگذاردن یك‌بسجمله‌ای به‌جای تابع زیرعلامت انتگرال، هریك از قواعد حاصله ازبخشهای ۲.۷ و ۳.۷ را به‌کار بریم و سپس نتایج حاصله را جمع‌کنیم.

x_iها را با فواصل مساوی به‌صورت زیر انتخاب می‌کنیم مگرآنکه برای انتخاب دیگر دلیل موجهی وجود داشته باشد

$$x_i=a+ih \qquad i=0,\dots,N$$

که $h=\dfrac{b-a}{N}$.

همچنین مانند بخش۶.۲ علامت اختصاری زیر را به‌کار می‌بریم

$$f_s=f(a+sh)$$

لذا $f_i=f(x_i)$ به‌ازای $i=0,\dots,N$.

اکنون مثالهای خاصی را درنظرمی‌گیریم. اگر قاعدهٔ مستطیلی (۲۳.۷) را در هر زیربازه به‌کاربریم، آنگاه برای زیربازهٔ (x_{i-1},x_i) خواهیم داشت

$$\int_{x_{i-1}}^{x_i} f(x)\,dx=(x_i-x_{i-1})f(x_{i-1})+\frac{f'(\eta_i)(x_i-x_{i-1})^2}{2}$$

$$=hf(x_{i-1})+\frac{f'(\eta_i)h^2}{2}$$

از محاسبهٔ حاصلجمع این مقادیرخواهیم داشت

$$I(f) \approx R_N = h \sum_{i=1}^{N} f_{i-1} \qquad \text{(۴۷.۷ الف)}$$

که رابطهٔ فوق، قاعدهٔ مستطیلی مرکب (بر N زیربازه است). خطای آن نیز درست برابر با مجموع خطاهایی‌است که درهر زیربازه ایجاد می‌شود

$$E_N^R = \sum_{i=1}^{N} \frac{f'(\eta_i)h^2}{2}$$

که در آن $\eta_i \in (x_{i-1} - x_i)$ می‌باشد. اگر $f'(x)$ پیوسته باشد (همان‌گونه که فرض کرده‌ایم)، رابطهٔ فوق را می‌توان با استفاده از قضیهٔ ۲.۱ از بخش۷.۱، به‌صورت زیر ساده کرد:

$$\sum_{i=1}^{N} \frac{f'(\eta_i)h^2}{2} = f'(\eta) \sum_{i=1}^{N} \frac{h^2}{2} = \frac{f'(\eta)Nh^2}{2}$$

لذا به‌ازای $Nh = b - a$، خواهیم داشت

$$E_N^R = \frac{f'(\eta)(b-a)h}{2} \qquad \text{(۴۷.۷ ب)}$$

به‌ازای مقداری مانند $\eta \in (a, b)$

سپس قاعدهٔ مرکب سیمپسن را به دست می‌آوریم. اگر در رابطهٔ (۲۸.۷) بگیریم $a = x_{i-1}$ و $b = x_i$ و $h = x_i - x_{i-1}$، آنگاه برای یك زیربازهٔ تنها خواهیم داشت

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = \frac{h}{6}[f_{i-1} + 4f_{i-1/2} + f_i] - \frac{f^{iv}(\eta_i)(h/2)^5}{90}$$

$$x_{i-1} < \eta_i < x_i$$

با محاسبهٔ حاصلجمع به‌ازای $i = 1, \ldots, N$، رابطهٔ زیرحاصل می‌شود

$$\begin{aligned} I(f) &= \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} f(x)\,dx \\ &= \frac{h}{6} \sum_{i=1}^{N} [f_{i-1} + 4f_{i-1/2} + f_i] - \sum_{i=1}^{N} \frac{f^{iv}(\eta_i)(h/2)^5}{90} \\ &= S_N + E_N^S \end{aligned}$$

تقریب مرکب سیمپسن S_N را می‌توان ساده‌کرد و نتیجهٔ زیر را به‌دست‌آورد

$$S_N=\frac{h}{۶}\left[f_۰+f_N+۲\sum_{i=۱}^{N-۱}f_i+۴\sum_{i=۱}^{N}f_{i-۱/۲}\right] \qquad (۴۸.۷\text{ الف})$$

درعین‌حال عبارت خطا را نیز می‌توان با استفاده از قضیهٔ ۲.۱ از بخش۷.۱، به‌صورت زیرساده‌کرد:

$$E_N^S=-\frac{f^{iv}(\xi)(h/۲)^۴(b-a)}{۱۸۰} \qquad a<\xi<b \qquad (۴۸.۷\text{ ب})$$

باید توجه داشت که در قاعدهٔ سیمپسن باید بتوانیم تابع را در نقاط میانی $x_{i-۱/۲}$، $(i=۱,\dots,N)$، ونیزدرنقاط انفصال x_i، $(i=۰,۱,\dots,N)$، محاسبه‌کنیم. این امر مخصوصاً ایجاب می‌کندکه همواره به‌تعداد فردی از نقاط متساوی‌الفاصله، که در آنها مقدار تابع زیرعلامت انتگرال را می‌دانیم، نیاز داشته باشیم.

به‌همین‌ترتیب برای قاعدهٔ نقطهٔ وسط مرکب چنین به‌دست می‌آوریم:

$$I(f)\approx M_N=h\sum_{i=۱}^{N}f_{i-۱/۲} \qquad E_N^M=\frac{f''(\xi)h^۲(b-a)}{۲۴} \qquad (۴۹.۷\text{ الف})$$

با توجه به‌قاعدهٔ نقطهٔ وسط (۲۵.۷)، و قاعدهٔ ذوزنقه‌یی مرکب، و (۲۶.۷) داریم

$$I(f)\approx T_N=h\sum_{i=۱}^{N-۱}f_i+\frac{h}{۲}(f_۰+f_N)$$

$$E_N^T=-\frac{f''(\eta)h^۲(b-a)}{۱۲} \qquad (۴۹.۷\text{ ب})$$

از قاعدهٔ ذوزنقه‌یی اصلاح‌شدهٔ (۲۹.۷)، رابطهٔ زیربه‌دست می‌آید

$$I(f)\approx CT_N=h\sum_{i=۱}^{N-۱}f_i=\frac{h}{۲}(f_۰+f_N)+\frac{h^۲}{۱۲}[f'(a)-f'(b)] \qquad (۵۰.۷)$$

$$E_N^{CT}=\frac{f^{iv}(\eta)h^۴(b-a)}{۷۲۰}$$

توجه داشته باشید که کلیهٔ مشتقهای درونی[1] $f'(x_i)$، $i=1,2,\ldots,N-1$، هنگام جمع نتایج حاصل از قاعدهٔ ذوزنقه‌یی اصلاح شده بر هر زیربازه، از یکدیگر حذف می‌شوند. بنابراین قاعدهٔ ذوزنقه‌یی اصلاح‌شدهٔ مرکب، درحقیقت یك قاعدهٔ ذوزنقه‌یی مرکب اصلاح شده است، یعنی

$$CT_N = T_N + \frac{h^2[f'(a)-f'(b)]}{12} \qquad (۵۱۰۷)$$

البته قاعدهٔ ذوزنقه‌یی اصلاح شده دارای این اشکال است که مشتق $f(x)$ باید دردست و یا قابل‌محاسبه باشد [به‌استثنای حالتی که در آن $f(x)$ یك تابع $(b-a)$-دوره‌ای است].

اگر قرار باشد که یکی از این قواعد مرکب به‌کار روند، در آغاز باید مقدار مناسبی مانند N، و یا معادل با آن، مقداری مانند $h=(b-a)/N$ تعیین شود. اگر اطلاعی از مقدار مشتقی که در عبارت خطا ظاهر می‌شود دردست باشد، اصلا می‌توان h یا N را طوری معین کرد که مقدار خطای حاصل از تحمل (تولرانس) مقرر کمتر باشد.

□ **مثال ۴۰۷:** N را طوری تعیین کنید که قاعدهٔ ذوزنقه‌یی مرکب (۳۳.۵ ب)، مقدار انتگرال

$$\int_0^1 e^{-x^2}\,dx$$

را (با فرض اینکه بتوان e^{-x^2} را دقیقاً محاسبه کرد) تا شش رقم اعشار صحیح به‌دست دهد و سپس، تقریب این انتگرال را نیز محاسبه کنید.

در این مثال داریم $f(x)=e^{-x^2}$، $a=0$، $b=1$ و $h=1/N$. بنابراین به‌ازای مقداری مانند $\eta\in(a,b)$ مقدار خطا در قاعدهٔ ذوزنقه‌یی مرکب برابر با $-f''(\eta)N^{-2}/12$ است. از آنجا که η را نداریم، بهترین فرضی که می‌توانیم بکنیم این است که بگوییم مقدار خطا از لحاظ قدرمطلق نباید بزرگتر از مقدار

$$\max_{0\leqslant\eta\leqslant1}\frac{|f''(\eta)|N^{-2}}{12}$$

باشد. محاسبهٔ زیر را انجام می‌دهیم

$$f''(x)=e^{-x^2}(4x^2-2)$$

بعلاوه $f'''(x)=e^{-x^2}4x(3-2x^2)$ در نقطه‌های $x=0$ و $x=\pm\sqrt{2/5}$ برابر با صفر است. لذا $\max|f''(x)|$ در بازهٔ $[0,1]$ باید در نقطهٔ $x=0$ یا در نقاط انتهایی $x=0,1$ موجود باشد. از این‌رو

---

1. jnterior derivatives

$$\max_{0\leqslant\eta\leqslant 1}|f''(\eta)| = \max\{|f''(0)|, |f''(1)|\} = \max\{2, 2e^{-1}\} = 2$$

بنابراین اگر N را چنان انتخاب کنیم که

$$\frac{2N^{-2}}{12} < 5\times 10^{-7}$$

و یا

$$N^2 > \frac{10^6}{3} = \frac{10^7}{6\times 5}$$

یا

$$N > \frac{10^3}{\sqrt{3}} \approx 578$$

دقت تا شش رقم اعشاری (بعد از ممیز) برای ما تضمین شده‌است. نتایج کامپیوتری زیر نشان می‌دهدکه N اندکی بیش از اندازه برآورد شده‌است.

محاسبات روی IBM مدل ۷۰۹۴ با دقت مضاعف (DP) ودقت سادهٔ (SP) انجام گرفته و نتایج حاصل برای مقادیر مختلف N به‌شرح زیراست.

N	$I(SP)$	$I(DP)$	ERROR(SP)	ERROR(DP)
50	7.4679947E-01	7.4670061D-01	2.466E-05	2.452D-05
100	7.4681776E-01	7.4681800D-01	6.37 E-06	6.13 D-06
200	7.4682212E-01	7.4682260D-01	2.01 E-06	1.53 D-06
400	7.4682275E-01	7.4682375D-01	1.56 E-06	3.8 D-07
800	7.4682207E-01	7.4682404D-01	2.06 E-06	9. D-08

مقدارصحیح I تا هشت رقم با معنی برابراست با $I = 0.74682413$. بنابراین، چنین به‌نظر می‌رسدکه تعداد تقسیمها هرچه اختیار شوند ظاهراً در محاسبه با دقت ساده نمی‌توانیم به‌شش رقم اعشاری دقیق دست یابیم. درحقیقت نتایج برای $N = 800$ نسبت به $N = 400$ بدتر است. این امرنشان می‌دهد که خطای ناشی از گردکردن روی سه رقم آخر اثر گذاشته‌است. چنانچه ازقبل تا حدی پیشبینی شده بود نتایج با دقت مضاعف نشان می‌دهندکه به‌ازای $N = 400$ دقت تا شش‌رقم اعشاری به‌دست می‌آید.

برنامهٔ فورترن محاسبات فوق به‌صورت زیراست:

برنامهٔ فورترن برای مثال ۴.۷ (دقت ساده)

```
C  EXAMPLE 7.4 . TRAPEZOID RULE .
      INTEGER I,N
      REAL A,B,H,T
      F(X) = EXP(-X*X)
    1 PRINT 601
```

```
601 FORMAT(' EXAMPLE 7.4 TRAPEZOIDAL INTEGRATION')
    READ 501,  A,B,N
501 FORMAT(2E20.0,I5)
    IF (N .LT. 2)                          STOP
    T = F(A)/2.
    H = (B - A)/FLOAT(N)
    DO 2 I=1,N-1
  2    T = F(A + FLOAT(I)*H) + T
    T = (F(B)/2. + T)*H
    PRINT 602, A,B,N,T
602 FORMAT(' INTEGRAL FROM A = ',1PE14.7,' TO B = ',E14.7,
   *       ' FOR N = ',I5,' IS ',E14.7)
                                           GO TO 1
    END
```

اگر از قاعدهٔ ذوزنقه‌یی اصلاح شده (۵۰.۷) استفاده کنیم، N مورد نیاز به‌طور قابل‌ملاحظه‌ای کاهش می‌یابد. اما خطا به‌وسیلهٔ

$$\max_{0\leqslant\eta\leqslant 1}\frac{|f^{iv}(\eta)|N^{-4}}{720}$$

کراندار شده‌است. مشتق مرتبهٔ چهارم $f^{iv}(x)=4e^{-x^2}(3-12x^2+4x^4)$ را محاسبه می‌کنیم. بنابراین

$$\max_{0\leqslant\eta\leqslant 1}|f^{iv}(\eta)|=|f^{iv}(0)|=12$$

لذا برای دستیابی به‌دقت با شش رقم، کافی است که داشته باشیم

$$\frac{12N^{-4}}{720}<5\times 10^{-7}$$

یا

$$N^4>(10/3)10^4=\frac{10^7}{300}$$

یا

$$N>13.5\approx\sqrt[4]{(10/3)(10)}$$

به‌طوری که درمقایسه با ۵۷۸ زیربازهٔ موردنیاز برای قانون ذوزنقه‌یی مرکب بدون تصحیح دیفرانسیل نهایی، تنها به ۱۴ زیربازه نیازاست. □

همان‌طوری که مثال فوق نشان می‌دهد اگر مشتقهای مرتبهٔ بالاتر تابع زیر علامت انتگرال یا مشتقهای مرتبهٔ پایینتر آن تقریباً هم‌اندازه باشند، آنگاه فرمولهای مرتبهٔ بالاتر می‌توانند تعداد تابع لازم برای محاسبه را نسبت به‌فرمولهای مرتبهٔ پایینتر به‌طور قابل ملاحظه‌ای کاهش دهند. بالاخص قواعد گاوسی در این مورد می‌توانند بسیار مؤثر باشند.

اگر دربارهٔ اندازهٔ مشتق مناسب $f(x)$ اطلاعی دردست نباشد، قواعد مرکب را تنها

می‌توان برای مقادیرمختلف N به‌کار برد، و بدین‌ترتیب یك دنباله از تقریبات I_N برای $I(f)$، به‌دست می‌آیندکه اگر $f(x)$ به‌اندازهٔ‌کافی هموار باشد این مقادیر ازلحاظ نظری وقتی $N \longrightarrow \infty$، به‌سمت $I(f)$ همگرا می‌شوند. این فرایند هنگامی به‌پایان می‌رسدکه تفاوت بین براوردهای متوالی «به‌اندازهٔ کافی‌کوچك» شود. خطرات این شیوهٔ عمل در بخش ۶.۱ موردبحث قرار گرفت. اشکال دیگری‌که دراین حالت پیش می‌آید آثار ناشی از گردکردن‌است‌که با زیادشدن N افزایش می‌یابد. نتایج کامپیوتری درمثال ۴.۷ این‌امر را به‌خوبی نشان می‌دهند.

□ **مثال ۵.۷:** برنامه‌ای برای قاعدهٔ ذوزنقه‌یی اصلاح شده بنویسید و مثال ۴.۷ را با استفاده از این برنامه حل‌کنید.

برنامهٔ فورترن

```
C     EXAMPLE 7.5 . CORRECTED TRAPEZOID RULE
      INTEGER I,N
      REAL A,B,CORTRP,H,TRAP
      F(X) = EXP(-X*X)
      FPRIME(X) = -2.*X*F(X)
      DATA A,B /0., 1. /
      PRINT 600
  600 FORMAT(9X,'N',7X,'TRAPEZOID SUM',7X,'CORR.TRAP.SUM')
      DO 10 N = 10,15
         H = (B - A)/FLOAT(N)
         TRAP = (F(A) + F(B))/2.
         DO 1 I=1,N-1
    1       TRAP = TRAP + F(A + FLOAT(I)*H)
         TRAP = H*TRAP
         CORTRP = TRAP + H*H*(FPRIME(A) - FPRIME(B))/12.
   10    PRINT 610, N,TRAP,CORTRP
  610 FORMAT(I10,2E20.7)
                                        STOP
      END
```

نتایج با دقت ساده

N	TRAPEZOID SUM	CORR.TRAP.SUM
10	0.7462108E 00	0.7468239E 00
11	0.7463173E 00	0.7468240E 00
12	0.7463983E 00	0.7468240E 00
13	0.7464612E 00	0.7468240E 00
14	0.7465112E 00	0.7468240E 00
15	0.7465516E 00	0.7468241E 00

نتایج با دقت مضاعف

N	TRAPEZOID SUM	CORR.TRAP.SUM
10	7.4621080E-01	7.4682393E-01
11	7.4631727E-01	7.4682399E-01
12	7.4639825E-01	7.4682403E-01
13	7.4646126E-01	7.4682406E-01
14	7.4651126E-01	7.4682408E-01
15	7.4655159E-01	7.4682409E-01

□

□ **مثال ۶.۷:** برنامه‌ای برای قاعدهٔ سیمپسن بنویسید و مثال ۴.۷ را با این برنامه، هم با دقت ساده و هم با دقت مضاعف حل کنید.

برنامهٔ فورترن و نتایج به‌دست آمده از IBM ۷۰۹۴ در زیر به‌ازای تعداد زیرتقسیمهایی برابر با ۱۰۰ و ۵۰ و $N = ۲۵$ داده شده‌اند. توجه کنید که نتایج با دقت سادهٔ ۱۰۰ و $N = ۵۰$ نسبت به $N = ۲۵$، بدتر است که این امر حاکی از اثرات خطای ناشی از گرد کردن است. نتایج با دقت مضاعف، تماماً با تعداد ارقام داده شده صحیح‌اند. از مقایسهٔ این نتایج با نتایجی که از مثال ۴.۷ و ۵.۷ به‌دست آمد، مشاهده می‌شود که هردو قاعدهٔ سیمپسن و قاعدهٔ ذوزنقه‌یی اصلاح‌شده به‌مراتب کاراتر از قاعدهٔ ذوزنقه‌یی هستند.

```
C  PROGRAM FOR EXAMPLE 7.6 . SIMPSON'S RULE .
      INTEGER I,N
      REAL A,B,H,HALF,HOVER2,S,X
      F(X) = EXP(-X*X)
      PRINT 600
  600 FORMAT(' EXAMPLE 7.6 SIMPSON''S RULE'/)
    1 READ 501, A,B,N
  501 FORMAT(2E20.0,I5)
      IF (N .LT. 2)                              STOP
      H = (B - A)/FLOAT(N)
      HOVER2 = H/2.
      S = 0.
      HALF = F(A + HOVER2)
      DO 2 I=1,N-1
         X = A + FLOAT(I)*H
         S = S + F(X)
    2    HALF = HALF + F(X+HOVER2)
      S = (H/6.)*(F(A) + 4.*HALF + 2.*S + F(B))
      PRINT 602, A,B,N,S
  602 FORMAT(' INTEGRAL FROM A = ',1PE14.7,' TO B = ',E14.7,
     *       ' FOR N = ',I5,' IS ',E14.7)
                                                 GO TO 1
    4 FORMAT(2E20.0,I5)
      END
```

نتایج کامپیوتری برای مثال ۶.۷

N	$I(SP)$	ERROR (SP)	$I(DP)$	ERROR (DP)
25	7.4682406E-01	7. E-07	7.4682413D-01	0.
50	7.4682400E-01	1.3E-06	7.4682413D-01	0.
100	7.4682392E-01	2.1E-06	7.4682413D-01	0.

□

سرانجام، قواعد مرکبی را که برفرمولهای گاوس استوارند نیز می‌توان به‌دست آورد. برای آنکه این قواعد با قواعد مرکبی که تاکنون بحث شد سازگار باشند، ما خود را به‌انتگرالهای معین به‌شکل زیرمحدود خواهیم ساخت

$$I = \int_a^b f(x)\,dx$$

بار دیگر بازهٔ (a, b) را به N زیر بازهٔ مساوی تقسیم می‌کنیم به طوری که

به‌ازای $h=(b-a)/N$

$$x_i=a+ih \qquad i=0,1,\dots,N$$

می‌خواهیم انتگرالگیری گاوسی را دربازهٔ iام برای این انتگرال یعنی برای انتگرال

$$I_i=\int_{x_{i-1}}^{x_i} f(x)\,dx \qquad (۵۲.۷)$$

به‌کار بریم. وزنها و نقاط گاوسی که براساس بسجمله‌ایهای لژاندر استوار و درقسمت۳.۷ داده شده‌اند. حدود انتگرالگیری را از -1 تا $+1$ می‌پذیرند. از این‌رو نخست تبدیل متغیرهای خطی به‌صورت زیرانجام می‌دهیم

$$x=\frac{h}{2}t+x_{i-1/2}$$

که در آن

$$x_{i-1/2}=(x_i+x_{i-1})/2$$

از قراردادن این مقادیر در (۵۲.۷) رابطهٔ زیر به‌دست می‌آید

$$I_i=\frac{h}{2}\int_{-1}^{1} f\left(\frac{ht}{2}+x_{i-1/2}\right)dt=\int_{-1}^{1} g_i(t)\,dt$$

که در آن

$$g_i(t)=\frac{h}{2}\,f\left(\frac{ht}{2}+x_{i-1/2}\right)$$

حال در $k+1$ نقطه، انتگرال I_i را با فرمول‌گاوسی‌تقریب می‌زنیم و رابطهٔ زیررا به‌دست می‌آوریم

$$I_i\approx A_0g_i(\xi_0)+A_1g_i(\xi_1)+\dots+A_kg_i(\xi_k) \qquad (۵۳.۷)$$

که درآن وزنها و طولها[1] از زیربرنامهٔ LGNDRE درقسمت ۳.۷ به‌دست می‌آیند.

سرانجام بـا به‌دست‌آوردن حاصلجمع این مقادیر تقریبی در N زیربازه، خواهیم داشت

$$I\approx\sum_{i=1}^{N}\int_{x_{i-1}}^{x_i} f(x)\,dx=\sum_{i=1}^{N}\int_{-1}^{1} g_i(t)\,dt=\sum_{i=1}^{N} I_i$$

که با توجه به (۵۳.۷)، این مقدار تقریبی برابراست با

1. abscissas

$$I \approx \sum_{i=1}^{N} \{A_0 g_i(\xi_0) + A_1 g_i(\xi_1) + \cdots + A_k g_i(\xi_k)\}$$

$$= \frac{h}{2} \sum_{i=1}^{N} \left\{ A_0 f\left(\frac{h}{2}\xi_0 + x_{i-1/2}\right) + \cdots + A_k f\left(\frac{h}{2}\xi_k + x_{i-1/2}\right) \right\}$$

(۵۴۰۷ الف)

توجه دارید که وزنها، مستقل از i هستند.

به‌موجب معادلهٔ خطای (۴۰۷)، در یک بازهٔ (x_{i-1}, x_i)، خطا به‌فرم زیر قابل بیان است.

$$E_i = C_k g_i^{(2k+2)}(\eta_i) \qquad [-1, 1] \text{ در } \eta_i \text{ مانند مقداری به‌ازای}$$

اما این بدان‌معنی است که

$$E_i = C_k \left(\frac{h}{2}\right)^{2k+3} f^{(2k+2)}(\eta_i') \qquad x_{i-1} < \eta_i' < x_i$$

بنابراین خطا در بازهٔ (a, b) می‌تواند به‌صورت زیر بیان شود

$$E_N^G = \frac{1}{2} C_k \left(\frac{h}{2}\right)^{2k+2} f^{(2k+2)}(\eta) \qquad \text{(۵۴۰۷ ب)}$$

□ **مثال ۷۰۷:** با استفاده از انتگرالگیری گاوسی به‌ازای $k = 3$ و به‌تعداد $N = 2$ زیربازه از بازهٔ $[1, 3]$، انتگرال $I = \int_1^3 [(\sin x)^2 / x] dx$ را براورد کنید. مثال ۳۰۷ را ببینید.

```
C   PROGRAM FOR EXAMPLE 7.7. COMPOSITE FOUR-POINT GAUSS-LEGENDRE .
      INTEGER I,N
      REAL A,B,H,HOVER2,P1,P2,POINT(2),S,S1,S2,WEIGHT(2),X
      DATA POINT,WEIGHT / .33998 10436, .86113 63116,
     *                    .65214 51549, .34785 48451 /
      F(X) = SIN(X)**2/X
      PRINT 600
  600 FORMAT(' EXAMPLE 7.7  FOUR-POINT GAUSS-LEGENDRE'/)
    1 READ 501, A,B,N
  501 FORMAT(2E20.8,I5)
      IF (N .LT. 1)                          STOP
      H = (B - A)/FLOAT(N)
      HOVER2 = H/2.
      P1 = POINT(1)*HOVER2
      P2 = POINT(2)*HOVER2
      S1 = 0.
      S2 = 0.
      DO 2 I=1,N
         X = A + FLOAT(I)*H - HOVER2
         S1 = S1 + F(-P1+X) + F(P1+X)
    2    S2 = S2 + F(-P2+X) + F(P2+X)
      S = HOVER2*(WEIGHT(1)*S1 + WEIGHT(2)*S2)
      PRINT 602, A,B,N,S
  602 FORMAT(' INTEGRAL FROM A = ',1PE14.7,' TO B = ',E14.7,
     *       ' FOR N = ',I3,' IS ',E14.7)
                                             GO TO 1
      END
```

جواب حاصل ازمحاسبات با دقت ساده‌که روی‌کامپیوتر UNIVAC ۱۱۱۰ انجام گرفته برابراست با ۰٫۷۹۴۸۲۵۱۷ که میزان خطا درآخرین رقم آن کمتر از ۳ واحد* است. □

تمرین

۱-۴۰۷ قاعدهٔ ذوزنقه‌یی مرکب T_N (۴۹۰۷) و قاعدهٔ نقطهٔ میانی‌مرکب M_N (۴۸۰۷) را به‌دست‌آورید.

۲-۴۰۷ قاعدهٔ دوزنقه‌یی اصلاح شدهٔ مرکب CT_N (۵۰۰۷) را به‌دست‌آورده و تحقیق کنیدکه مشتقهای درونی $f'(x_i)$، $(i = 1, \ldots, N-1)$، درحاصلجمع حذف می‌شوند.

۳-۴۰۷ برنامه‌ای برای قاعدهٔ مرکب سیمپسن بنویسید. ورودیهای این برنامه می‌باید $f(x)$ باشد، بازهٔ‌آن $[a, b]$، وتعداد زیرفاصله‌ها N. از این برنامه برای محاسبهٔ‌انتگرال زیر با $N = ۱۰$ و $N = ۲۰$ زیرفاصله استفاده‌کنید

$$I = \int_{۱}^{۲} x \ln x \, dx$$

۴-۴۰۷ برنامهٔ مربوط به‌قاعدهٔ سیمپسن را درمحاسبهٔ تقریبی انتگرالهای زیرکه تا شش رقم‌اعشاری صحیح باشند به‌کاربرید

$$I = \int_{۰}^{۱} xe^{-x} \, dx \qquad I = \int_{۰}^{۱} x \cos x \, dx \qquad I = \int_{۰}^{۱} (۱+x^{۲})^{۳/۲} \, dx$$

این عمل‌را با $N = ۱۰$ شروع وهر دفعه N را دوبرابرکنید تا به‌دقتی‌که می‌خواهیددست یابید.

۵-۴۰۷ برنامه‌ای برای قاعدهٔ ذوزنقه‌یی اصلاح شده بنویسید. دراین‌حالت، ورودی‌شامل $f(x)$، $f'(x)$، $[a, b]$ و N خواهد بود. این برنامه را برای انتگرال تمرین ۴۰۷-۳ به‌کاربرید و نتایج حاصله را با نتایج به‌دست‌آمده از قاعدهٔ سیمپسن مقایسه‌کنید.

۶-۴۰۷ به‌ازای $k = ۳$، برنامه‌ای برای قاعدهٔ مرکب گاوسی(۵۰۷ الف) بنویسید. برنامهٔ فوق رابرای محاسبهٔ انتگرال‌تمرین۴۰۷-۳ به‌کارگیرید. این‌محاسبات را در ابتدا به‌ازای $N = ۲$ زیربازه و سپس به‌ازای $N = ۴$ زیربازه انجام دهید. مقدار محاسبات و دقت به‌دست‌آمده را با موارد مشابه درقاعدهٔ سیمپسن مقایسه‌کنید.

۷-۴۰۷ تابع خطای $\operatorname{erf}(x)$ به‌صورت زیرتعریف می‌شود

---

* توضیح مترجم ۰٫۰۰۰۰۰۰۰۰۳ ± ۰٫۷۹۴۸۲۵۱۷.

$$\operatorname{erf}(x)=\frac{۲}{\sqrt{\pi}}\int_{۰}^{x} e^{-t^{۲}}\,dt$$

به‌ازای $N=۲$ و $N=۴$ زیربازه و $k=۳$، برای محاسبهٔ $\operatorname{erf}(۰٫۵)$ ازقاعدهٔ مرکب گاوسی استفاده‌کنید. دقت نتیجهٔ حاصله را براورد و با مقدار دقیق

$$\operatorname{erf}(۰٫۵)=۰٫۵۲۰۴۹۹۸۷۶$$

مقایسه‌کنید.

۸-۴.۷ برای تعیین میزان چگالش یك‌بخار خالص درخارج ازیك لولهٔ افقی سرد شده، به‌محاسبهٔ میانگین ضریب انتقال دمای Q نیاز داریم. برای محاسبهٔ این ضریب، همراه با سایر پارامترها، محاسبهٔ انتگرال زیرلازم می‌آید

$$I=\int_{۰}^{\pi}(\sin x)^{۱/۳}\,dx$$

به‌ازای $N=۵, ۱۰, ۱۵, ۲۰$ زیرفاصله و با به‌کارگیری قانون سیمپسن، مقدار این‌انتگرال را محاسبه‌کنید.

جواب: برای $N=۵$، $I\approx ۲٫۵۲۸۶۹۴۹$.

۵.۷ انتگرال تطبیقی[1]

قواعد مرکبی‌که تا به‌حال مورد بحث قرارگرفتند، همه برپایهٔ N زیربازهٔ مساوی نهاده شده‌اند. انتخاب این‌گونه زیربازه‌های مساوی، هرگاه تابع زیرعلامت انتگرال در یك رشته از نقاط بافواصل‌مساوی معلوم باشد، مثلا اگر $f(x)$ تنها به‌صورت جدولی‌ازمقادیر تابع داده شده باشد،کاملا طبیعی و حتی اغلب لازم‌است. اما اگر $f(x)$ درهرنقطه ازبازهٔ انتگرالگیری به‌راحتی‌محاسبه شود، معمولا به‌صرفه نزدیکتر است‌که اززیربازه‌هایی‌استفاده کنیم‌که طول‌آنها با توجه به‌رفتار محلی تابع زیر علامت انتگرال، معین شود. به‌عبارت دیگر، اگر به‌جای اصرار در متساوی‌الفاصله بودن زیربازه‌ها، آنها نامساوی ولی مناسب انتخاب شوند، آنگاه می‌توان $I(f)$ را با دقتی از پیش تعیین‌شده و محاسباتی کمتر به‌دست‌آورد.

برای مثال، قاعدهٔ‌کلی ذوزنقه‌یی مرکب

$$I(f)=\sum_{i=۱}^{N}\frac{x_i-x_{i-۱}}{۲}\left[f(x_{i-۱})+f(x_i)\right]-\sum_{i=۱}^{N}\frac{f''(\eta_i)(x_i-x_{i-۱})^{۳}}{۱۲}$$

را درنظربگیرید،که درآن لازم نیست، نقاط‌انفصال $a=x_۰<x_۱<\cdots<x_N=b$

1. Adaptive

به‌فواصل مساوی قرار گیرند، در بازهٔ (x_{i-1}, x_i) به‌ازای نقطه‌ای مانند $\eta_i \in (x_{i-1}, x_i)$، سهم مقدار

$$-\frac{f''(\eta_i)(x_i - x_{i-1})^3}{12}$$

در رابطه با کل خطا، هم به‌مقدار $f''(x)$ در بازهٔ (x_{i-1}, x_i) بستگی دارد و هم به‌اندازهٔ زیربازهٔ $|x_i, x_{i-1}|$. بنابراین اگر بخواهیم سهمی که هرزیربازه در کل خطا دارد مساوی باشد، می‌توانیم در قسمتهایی از بازهٔ انتگرالگیری (a, b) که مقدار $|f''(x)|$ «کوچک» است، زیربازه‌ها را «بزرگ» بگیریم و در ناحیه‌هایی که $|f''(x)|$ «بزرگ» است، زیربازه‌ها را «کوچک» می‌توان نشان داد که اگر منظور به‌حداقل رساندن تعداد زیربازه‌ها و در پی آن به‌حداقل رساندن تعداد محاسباتی از تابع باشد که برای به‌دست آوردن $I(f)$ تا دقت معینی لازم است، آنگاه روش مزبور بهترین روش است.

آن دسته از روشهای انتگرالگیری که طول زیربازه‌ها را با رفتار محلی تابع زیر علامت انتگرال وفق می‌دهند، روشهای انتگرالگیری **تطبیقی** نام دارند. اشکال مهمی که در این گونه روشها وجود دارد بی‌اطلاعی از مشتقی است که در عبارت خطا ظاهر می‌شود. و این بدان معنی است که در چنین روشهایی، رفتار محلی تابع زیر علامت انتگرال را باید از روی مقادیر آن در چند نقطه حدس زد.

در اینجا یک روش الگوی انتگرالگیری تطبیقی را که بر استفاده از قاعدهٔ سیمپسن به‌عنوان یک فرمول انتگرال پایه متکی است، به‌اجمال شرح می‌دهیم. فرض آن است که تابع $f(x)$، بازهٔ $[a, b]$ و یک معیار خطای ε داده شده‌اند. هدف، محاسبهٔ تقریبی P ی انتگرال $I = \int_a^b f(x)\,dx$ با حداقل تعداد محاسبات برای تابع است به‌طوری که

$$|P - I| \leqslant \varepsilon \qquad (۵۵۰۷)$$

برای این کار، عمل را تا آنجا که ممکن است، با استفاده از عدهٔ کمی از مقادیر تابع انجام می‌دهیم.

عملیات را با تقسیم بازهٔ $[a, b]$ به N زیربازه که معمولاً، ولی نه الزاماً، برابرند آغاز می‌کنیم. گیریم x_i و x_{i+1} دو سر یک چنین زیربازه‌ای باشند و $h = x_{i+1} - x_i$. اکنون برای انتگرال

$$I_i = \int_{x_i}^{x_{i+1}} f(x)\,dx$$

دو تقریب با قاعدهٔ سیمپسن به‌دست می‌آوریم. یکی از این دو تقریب را که با S نشان می‌دهیم بر اساس استفاده از دو قطعه[1] مبتنی شده است، و دیگری را که با $\bar{S}$ نشان می‌دهیم بر اساس

1. panel

استفاده از ۴ قطعه پی‌ریزی شده‌است. به‌موجب فرمول (۲۸.۷) این تقریبها با روابط زیر معین می‌شوند

$$S_i=\frac{h}{۶}\left\{f(x_i)+۴f\left(x_i+\frac{h}{۲}\right)+f\left(x_{i+۱}\right)\right\} \qquad (۵۶.۷\text{ الف})$$

$$\bar{S}_i=\frac{h}{۱۲}\left\{f(x_i)+۴f\left(x_i+\frac{h}{۴}\right)+۲f\left(x_i+\frac{h}{۲}\right)+۴f\left(x_i+\frac{۳h}{۴}\right)\right.$$

$$\left.+f(x_{i+۱})\right\} \qquad (۵۶.۷\text{ ب})$$

از این دوتقریب، می‌توانیم میزان خطـا را درتقریب دقیقتر $\bar{S}_i$ به‌گونهٔ زیر محاسبه‌کنیم. به‌موجب عبارت خطا درقاعدهٔ سیمپسن (۲۸.۷)، داریم

$$I_i-S_i=-\frac{f^{iv}(\eta)}{۹۰}\left(\frac{h}{۲}\right)^۵ \qquad (۵۷.۷\text{ الف})$$

$$I_i-\bar{S}_i=-\frac{۲f^{iv}(\eta)}{۹۰}\left(\frac{h}{۴}\right)^۵ \qquad (۵۷.۷\text{ ب})$$

ظاهرشدن ضریب۲ در رابطهٔ (۵۷.۷ ب) ناشی ازآن حقیقت است‌که عمل انتگرالگیری روی دو زیربـازه انجام می‌گیرد که هریك پهنایی برابر بـا $h/۲$ دارد. بـا فرض اینکه مشتق $f^{iv}(x)$ دربازهٔ $[x_i, x_{i+۱}]$ تقریبـاً ثابت‌است، می‌توان رابطـهٔ (۵۷.۷ ب) را از (۵۷.۷ الف)کم‌کرد و بعد از ساده‌کردن نتیجه خواهیم داشت

$$\bar{S}_i-S_i=\frac{f^{iv}\cdot h^۵}{۲^۵\cdot ۹۰}\left(\frac{۱-۲^۴}{۲^۴}\right)$$

که از این رابطه، نتیجه می‌گیریم

$$\frac{f^{iv}\cdot h^۵}{۲^۵\cdot ۹۰}=\frac{۲^۴(\bar{S}_i-S_i)}{۱-۲^۴} \qquad (۵۸.۷)$$

باگذاردن (۵۸.۷) درسمت راست (۵۷.۷ ب)، مقدار براورد خطا برابر می‌شود با

$$I_i-\bar{S}_i=\frac{\bar{S}_i-S_i}{۲^۴-۱}=\frac{۱}{۱۵}(\bar{S}_i-S_i) \qquad (۵۹.۷)$$

یابه‌صورت گفتاری، میزان خطا درتقریب دقیقتر $\bar{S}_i$ تقریباً برابر است با۱/۱۵ تفاضل بین دوتقریب $\bar{S}_i$ و S_i، که کمیتی‌است به‌آسانی قابل‌محاسبه.

اگر بازهٔ $[a, b]$ با N زیربازه پوشانده شده باشد، واگردرهریك از این‌زیربازه‌ها،

براورد خطا در رابطهٔ

$$E_i=\frac{1}{15}|\bar{S}_i-S_i|\leqslant\frac{h}{b-a}\varepsilon \qquad (۶۰.۷)$$

صدق کند، آنگاه می‌توان نشان داد که تقریب حاصل برای انتگرال I از حاصلجمع

$$P=\sum_{i=1}^{N}\bar{S}_i$$

در کل بازهٔ $[a, b]$ درملاك خطای موردنیاز(۵۵.۷) صدق خواهد کرد. در رابطهٔ (۶۰.۷) توجه به‌این‌امر که مقدار $h=x_{i+1}-x_i$ با تغییر پهنای زیر بازه‌تغییرخواهد کرد، ضروری است.

انتگرالگیری تطبیقی اساساً، عبارت است از به‌کاربردن فرمولهای (۵۶.۷ الف) و (۵۶.۷ ب) برای هر زیربازه‌ای کـه $[a, b]$ را طی می‌کند تا نامساوی (۶۰.۷) برقرار شود. اگرنامساوی (۶۰.۷) دریك یا چند زیربازه صدق نکند، آنگاه زیربازه‌های مزبور به‌بازه‌های کوچکتر تقسیم وکل فرایند تکرار می‌شود.

هرزیربرنامه‌ای‌که بر اساس انتگرالگیری تطبیقی نوشته می‌شود باید باریکهٔ[1] همهٔ زیربازه‌ها را حفظ‌کند تا اطمینان حاصل‌شودکه بازهٔ $[a, b]$ طی شده‌است، و باید پهنای زیربازه‌ها، یعنی h موردنیـاز در فرمولهای (۵۶.۷ الف)، (۵۶.۷ ب)، (۶۰.۷) به‌طور مناسب برگزیده شود. پیچیدگی زیر برنامه‌های انتگرالگیری تطبیقی ناشی از ساماندهی وسیع لازم‌برای حفظ باریکهٔ زیربازه‌های تودرتو، و نیاز به‌جریانهای متفاوت عمل‌هنگام مواجهه بامشکلات‌است. با توجه به‌اینکه درفرمولهای (۵۶.۷ الف) و(۵۶.۷ ب) نقاطی که $f(x)$ درآنها در(۵۶.۷ الف) محاسبه شده‌اند درهمان نقاطی‌هستندکه در(۵۶.۷ ب) وجود دارند، زیر برنامه‌های تطبیقی‌که برقاعدهٔ سیمپسن متکی هستند نیز کاراتر می‌شوند. بنابراین مقادیر $f(x)$ دراین‌نقاط‌را می‌توان‌کم‌کرد. مثال زیر روشی راکه در اینجا بیان شد، روشن خواهد ساخت.

□ **مثال ۸.۷:** به‌کارگیری روش‌انتگرال تطبیقی بر اساس قاعدهٔ سیمپسن، تقریبی برای انتگرال

$$I=\int_0^1\sqrt{x}\,dx$$

دقیق تا خطایی برابر با $\varepsilon=0.00005$ بیابید.

جواب دقیق برابربا $I=2/3$ به‌آسانی محاسبه می‌شود. ولی روشن‌است‌که برای محاسبهٔ انتگرال فوق حتی‌الامکان بـاید از روش قاعدهٔ تطبیقی سیمپسن استفاده کرد. بـا

1. track

رسم نمودار تابع $f(x)=\sqrt{x}$، مشاهده می‌شود که در نزدیکی مبدأ [در واقع $f'(0)=\infty$] منحنی‌دارای شیب خیلی تندی‌است، درحالی که وقتی $x \to 1$، منحنی نسبتاً هموار[1] می‌شود. ازاین‌رو دریك بازهٔ نزدیك به‌مبدأ درمقایسه با یك بازهٔ نزدیك به $x=1$، اشکالات زیادتری در امر انتگرالگیری وجود دارد.

عملیات را با تقسیم بازهٔ $[0, 1]$ به‌دو زیربازهٔ $[0, 1/2]$ و $[1/2, 1]$ آغاز می‌کنیم. در ابتدا فرمولهای (۵۶.۷ الف) و (۵۶.۷ ب) را در بازهٔ $[1/2, 1]$ به‌کار می‌بریم. دراینجا $h=1/2$، و از این‌رو خواهیم داشت

$$S\left[\frac{1}{2}, 1\right]=\frac{1}{12}\{\sqrt{1/2}+4\sqrt{3/4}+\sqrt{1}\}=0.۴۳۰۹۳۴۰۳$$

$$\bar{S}\left[\frac{1}{2}, 1\right]=\frac{1}{24}\{\sqrt{1/2}+4\sqrt{5/8}+2\sqrt{3/4}+4\sqrt{7/8}+\sqrt{1}\}$$

$$=0.۴۳۰۹۶۲۱۹$$

در اینجا برای مشخص ساختن زیربازهٔ موردنظر، تقریباً نشانه‌گذاری متفاوتی به‌کاررفته است. به‌موجب فرمول خطای (۶۰.۷)، داریم

$$E\left[\frac{1}{2}, 1\right]=\frac{1}{15}(\bar{S}-S)=0.۰۰۰۰۰۱۸۷۷۵ < \frac{1/2}{1}(0.۰۰۰۵)$$

$$=0.۰۰۰۰۲۵$$

ازآنجایی‌که ملاك خطا برقرار شده‌است، مقدار $\bar{S}[1/2, 1]$ رامی‌پذیریم وآن را دریك ثبات[2] SUM کنارمی‌گذاریم. سپس فرمولهای (۵۶.۷ الف) و (۵۶.۷ ب) را برای‌بازهٔ $[0, 1/2]$ به‌کار می‌گیریم. به‌ازای $h=1/2$، خواهیم داشت

$$S\left[0, \frac{1}{2}\right]=\frac{1}{12}\{0+4\sqrt{1/4}+\sqrt{1/2}\}=0.۲۲۵۵۹۲۲۳$$

$$\bar{S}\left[0, \frac{1}{2}\right]=\frac{1}{24}\{0+4\sqrt{1/8}+2\sqrt{1/4}+4\sqrt{3/8}+\sqrt{1/2}\}=0.۲۳۲۱۱۷۰۹$$

و

$$E\left[0, \frac{1}{2}\right]=0.۰۰۰۰۴۳۴۹۹ < 0.۰۰۰۰۲۵$$

در اینجا آزمون[3] خطا برقرار نیست وبنابراین بازهٔ $[0, 1/2]$ تقسیم می‌شود. بانصف کردن این‌بازه، دو زیربازهٔ $[0, 1/4]$ و $[1/4, 1/2]$ به‌دست می‌آید. به‌ازای $h=1/4$ و به‌کارگیری فرمولهای (۵۶.۷ الف) و (۵۶.۷ ب)، خواهیم داشت

---

1. flat 2. register 3. test

$$S\left[\frac{۱}{۴},\frac{۱}{۲}\right]=۰ر۱۵۲۳۵۸۱۹$$

$$\bar{S}\left[\frac{۱}{۴},\frac{۱}{۲}\right]=۰ر۱۵۲۳۶۸۱۴$$

$$E\left[\frac{۱}{۴},\frac{۱}{۲}\right]=۰ر۶۶۴\times ۱۰^{-۶}<\frac{۱/۴}{۱}(۰ر۰۰۰۵)=۰ر۰۰۰۱۲۵$$

آشکار است که ملاك خطا برقراراست، از این‌رو مقدار $\bar{S}[۱/۴, ۱/۲]$ را به‌محتوای ثبات SUM می‌افزاییم تا تقریب جزئی زیر را به‌دست آوریم

$$\text{SUM}\left[\frac{۱}{۴},۱\right]=۰ر۴۳۰۹۶۲۱۹+۰ر۱۵۲۳۶۸۱۴=۰ر۵۸۳۳۳۰۳۳$$

با به‌کارگیری مجدد فرمول پایهٔ (۵۶.۷) دربازهٔ $[۰, ۱/۴]$ و به‌ازای $h=۱/۴$، خواهیم داشت

$$S\left[۰,\frac{۱}{۴}\right]=۰ر۰۷۹۷۵۸۹۰$$

$$\bar{S}\left[۰,\frac{۱}{۴}\right]=۰ر۰۸۲۰۶۵۷۸$$

$$D\left[۰,\frac{۱}{۴}\right]=(۰ر۰۰۰۱۵۳۷۹۲۲)\not<۰ر۰۰۰۱۲۵$$

آزمون خطا برقرار نبوده از این‌رو بازهٔ $[۰, ۱/۴]$ را به دو زیربازهٔ $[۰, ۱/۸]$ و $[۱/۸, ۱/۴]$ تقسیم می‌کنیم. با انجام عملیاتی مشابه و به‌ازای $h=۱/۸$، خواهیم داشت

$$S\left[\frac{۱}{۸},\frac{۱}{۴}\right]=۰ر۰۵۳۸۶۶۷۵$$

$$\bar{S}\left[\frac{۱}{۸},\frac{۱}{۴}\right]=۰ر۰۵۳۸۷۰۲۷$$

$$E\left[\frac{۱}{۸},\frac{۱}{۴}\right]=۰ر۰۰۰۰۰۰۲۳۴۶<۱/۸(۰ر۰۰۰۵)=۰ر۰۰۰۰۶۲۵$$

$$S\left[0, \frac{1}{8}\right] = 0.02819903$$

$$\bar{S}\left[0, \frac{1}{8}\right] = 0.02901464$$

$$E\left[0, \frac{1}{8}\right] = 0.00005437 < 0.00006250$$

چون آزمون خطا درهر دوبازه برقرار است، می‌توانیم این مقادیر را به‌محتوای ثبات SUM اضافه‌کنیم، که با انجام این‌کارخواهیم داشت

$$P = \text{SUM}[0, 1] = 0.58331033 + 0.05387027 + 0.02901464$$

$$= 0.66621524$$

از آنجایی‌که مقدار دقیق I برابر 0.66666666 است، دیده می‌شودکه در تمام بازهٔ $[0, 1]$ تقریب P برای I درملاك خطای مطلوب صدق می‌کند، یعنی

$$|P - I| = 0.00045142 < 0.00005 \qquad \square$$

همان‌گونه‌که این‌مثال‌نشان می‌دهد، درجاهایی که‌منحنی $f(x)$ به‌آرامی تغییر می‌کند، روشهای انتگرالگیری تطبیقی بازه‌های بزرگتری را می‌طلبد، اما درجاهایی نظیر قله‌های تیز یانزدیك به‌نقاط تکین، که منحنی $f(x)$ سریعاً تغییر می‌کند، باید بازه‌های انتگرالگیری خیلی کوچکتر را به‌کارگرفت تا دقت عمل لازم به‌دست آید.

ما در اینجا زیربرنامه‌ای براساس روش انتگرالگیری تطبیقی عرضه نمی‌کنیم. همان‌گونه‌که قبلا بیان شد، اگر قرار باشدکه رده‌های بسیاری از توابع بااین برنامه‌اجرا شوند، چنین زیربرنامه‌ای مطمئناً بسیار پیچیده خواهد بود. برای بسیاری ازکامپیوترهای جدید، روشهای انتگرالگیری تطبیقی عالی تهیه‌کرده‌اند.

تمرین

۱-۵.۷ با استفاده از یك ماشین حساب جیبی نتایج به‌آمده از مثال ۸.۷ را برای $S[1/4, 1/2]$، $\bar{S}[1/4, 1/2]$، $S[1/8, 1/4]$، و $\bar{S}[0, 1/8]$ تحقیق کنید.

۲-۵.۷ درمثال ۸.۷، ملاك خطا را به $\varepsilon = 0.0001$ بدل‌کنید. کدامیك از بازه‌هایی‌که تا به‌حال تعیین شده‌است ملاك خطای مطلوب را برقرار می‌کند وکدامیك نمی‌کند؟ مادامی که ملاك خطای جدید برقرارنشده‌است، بازهٔ $[0, 1/8]$ را تقسیم و مانند مثال مذکور، انتگرال را محاسبه‌کنید.

۳-۵۰۷ با استفاده از روش انتگرالگیری تطبیقی براساس قاعدهٔ سیمپسن، تقریبی برای انتگرال

$$I=\int_{0}^{1}(1-x^{2})^{3/2}\,dx$$

با دقتی سه رقم اعشار بیابید. نخست منحنی $f(x)$ را رسم و سپس سعی کنیدکه محلی را که انتظار می‌رود با اشکالاتی مواجه شوید، معین نمایید.

۴-۵۰۷ با استفاده از روش انتگرالگیری تطبیقی، تقریبی برای انتگرال زیر، با دقتی تا شش‌رقم اعشاری بیابید

$$I=\int_{0}^{1}\frac{\sin x}{x^{3/2}}\,dx$$

۵-۵۰۷ برنامه‌ای برای انتگرالگیری تطبیقی براساس قاعدهٔ سیمپسن بنویسید و در این برنامه محدودیتهای زیر را درنظر بگیرید.

۱. ورودی شامل تابع $f(x)$، یك بازهٔ متناهی $[a,b]$، و ملاك خطای مطلق ε باشد.

۲. برنامه می‌باید بازهٔ $[a,b]$ را به‌دوقسمت مساوی تقسیم کند و فرمولهای (۵۶۰۷ الف)، (۵۶۰۷ ب) و (۶۰۷) را برای محاسبهٔ S و $\bar{S}$ و E به‌ازای هر قسمت به‌کار برید.

۳. اگر E درشرایط خطای مطلوب دریك زیربازه صدق می‌کند، $\bar{S}$ ذخیره شود؛ درغیر این‌صورت بازهٔ مزبور نصف و مرحلهٔ ۲ تکرارشود.

۴. اگرتقسیم بازه لازم باشد این‌کار راحداکثر۴ مرحله دریك زیربازه انجام‌دهید.

۵. خروجی این برنامه به‌صورت زیرباشد.

(i) اگر آزمون خطا درمجموعه‌ای از بازه‌هایی که $[a,b]$ را طی می‌کند، برقرارشود یك متغیر به‌صورت عدد صحیح ۱ = IFLAG درخروجی داشته باشیم، و اگر آزمون خطا به‌ازای یك یا چند زیربازه برقرارنشود، متغیر ۲ = IFLAG درخروجی ظاهر شود.

(ii) اگر ۱ = IFLAG، $P=\Sigma\bar{S}_i$ چاپ شود.
اگر ۲ = IFLAG، درآن فواصلی که آزمون خطا برقرار می‌شود حاصلجمع جزئی $PP=\Sigma\bar{S}_i$ چاپ شود، و فهرستی از بازه‌های $[x_i, x_{i+1}]$ که درآنها آزمون خطا برقرارنیست نیزچاپ شود.

۶-۵۰۷ این حکم موجود درمتن را تحقیق کنید: اگر رابطهٔ خطای (۶۰۷) در هریك از N زیربازه‌ای که بازهٔ $[a,b]$ را طی می‌کند برقرارباشد، آنگاه $P=\sum_{i=1}^{N}\bar{S}_i$ در شرط خطای مطلوب (۵۵۰۷) درکل بازهٔ $[a,b]$ صادق خواهد بود.

۶.۷* برونیابی به‌سمت حد

درقسمتهای قبل، تلاش زیادی برای به‌دست آوردن عبارتهایی برای خطای قاعده‌های مختلف انتگرالگیری و مشتقگیری تقریبی به‌عمل آوردیم. خلاصه بگوییم: با $L(f)$ که انتگرال $f(x)$ دریك بازهٔ $[a,b]$ بود، یا با مقدار مشتق $f(x)$ در نقطهٔ a، تقریب $L_h(f)$ را برای $L(f)$ که بستگی به‌پارامتر h دارد و در رابطهٔ

$$\lim_{h\to 0} L_h(f)=L(f)$$

صدق می‌کند، به‌دست می‌آوریم. روشنتر بگوییم، مطابق معمول، رابطهٔ

$$L(f)=L_h(f)+ch^r f^{(s)}(\xi)$$

راکه در آن c یك عدد ثابت و r و s اعداد صحیح مثبت و $\xi=\xi(h)$ یك نقطهٔ نامعلوم از بازهٔ موردبحث بودند، ثابت و اشاره کردیم که حد مستقیم برای مقدار عبارت خطا مستلزم دانستن مقدار $|f^{(s)}(\xi)|$ است، که این خود غالباً نمی‌تواند به‌اندازهٔ کافی دقیق به‌دست آید (اگر نخواهیم بگوییم که اصلا دقیق به‌دست نمی‌آید).

با وجود این، یك‌چنین عبارت خطایی میزان[1] نزدیکی $L_h(f)$ به $L(f)$ را (وقتی $h\to 0$) بیان می‌کند. این اطلاعات را می‌توان گاهی برای براورد خطا از روی مقادیر متوالی $L_h(f)$ به‌کار برد. امکان یك‌چنین براوردی درقسمت ۶.۱ به‌طور مختصر ذکرشد و در بخش ۴.۳ مثال خاصی، فرایند Δ^2ی ایتکن، موردبحث قرارگرفت و مثال دیگری در بخش قبلی ۵.۷ آورده شد.

به‌عنوان یك مثال ساده، تقریب

$$D_h(f)=\frac{f(a+h)-f(a-h)}{2h}$$

برای مقدار

$$D(f)=f'(a)$$

راکه مربوط به‌مشتق اول $f(x)$ در نقطهٔ $x=a$ است درنظر می‌گیریم. اگر $f(x)$ دارای سه‌مشتق پیوسته باشد، آنگاه به‌موجب (۸.۷) یا (۱۱.۷)، داریم

$$D(f)=D_h(f)-\frac{1}{6}h^2 f'''(\xi) \qquad |\xi-a|<|h|$$

به‌ازای مقداری مانند ξ با شرط

چون وقتی $h\to 0$، خواهیم داشت $\xi(h)\to a$، و چون $f'''(x)$ پیوسته است، وقتی $h\to 0$، داریم

---

1. rate

$$f''''(\xi) \longrightarrow f''''(a)$$

از این‌رو

$$\frac{[f''''(\xi)-f''''(a)]h^2}{6}$$

سریعتر از h^2 به‌سمت صفر میل می‌کند. بنابراین با استفاده از علامت گذاریهای مرتبه‌ای که در بخش ۶.۱ بیان شد، خواهیم داشت

$$D(f)=D_h(f)+C_1h^2+o(h^2) \qquad (۶۱.۷)$$

که در آن ثابت $C_1=-f''''(a)/6$ به h بستگی ندارد.

یك مثال عددی می‌تواند به‌روشن کردن اهمیت معادلهٔ (۶۱.۷) کمك کند. به‌ازای $f(x)=\sin x$ و $a=1$ خواهیم داشت

$$D(f)=0.540402$$

$$C_1=0.090050$$

در جدول ۲.۷ مقادیر $D_h(f)$، خطای $E_h(f)=-h^2f''''(\xi)/6$، و دو مؤلفهٔ آن، یعنی C_1h^2 و $o(h^2)$ به‌ازای مقادیر مختلف h داده شده‌اند. (برای جلوگیری از اشکالهای جزئی ناشی از خطای گرد کردن، کلیهٔ عناصر این جدول با دقت مضاعف محاسبه و سپس گرد شده‌اند.) چنانچه این جدول نشان می‌دهد، C_1h^2 سریعاً به مؤلفهٔ نافذ[1] خطا تبدیل می‌شود زیرا اگرچه C_1h^2 (همراه با h) به‌سمت صفر میل می‌کند، ولی مؤلفهٔ $o(h^2)$ سریعتر از آن به‌سمت صفر نزدیك می‌شود. اما این امر بدین‌معناست که می‌توانیم برآورد خوبی در مورد خطا برای مؤلفهٔ اصلی یعنی C_1h^2، به‌صورت زیر داشته باشیم: در رابطهٔ (۶۱.۷) به‌جای h مقدار $2h$ را می‌گذاریم و داریم

جدول ۲.۷

h	$D_h(f)$	$E_h(f)$	C_1h^2	$o(h^2)$	$(D_h-D_{2h})/3$	R_h
6.4	0.009839	0.530463	3.688464	−3.158001		
3.2	−0.009856	0.550158	0.922116	−0.371957	−0.065652	−0.57
1.6	0.337545	0.202757	0.230529	−0.027772	0.115800	2.37
0.8	0.484486	0.055816	0.057632	−0.001816	0.048980	3.54
0.4	0.526009	0.014293	0.014408	−0.000115	0.013841	3.88
0.2	0.536707	0.003594	0.003602	−0.000007	0.003566	3.97
0.1	0.539402	0.000900	0.000901	−0.0000005	0.000898	

۱. dominant (مؤلفهٔ نافذ = آن مؤلفه‌ای که تأثیر بیشتری روی خطا می‌گذارد). ـم.

$$D(f)=D_{۲h}(f)+۴C_۱h^۲+o(h^۲)$$

ازکم‌کردن این معادله از (۶۱.۷) خواهیم داشت

$$۰=D_h(f)-D_{۲h}(f)-۳C_۱h^۲+o(h^۲)$$

یا

$$C_۱h^۲=\frac{D_h(f)-D_{۲h}(f)}{۳}+o(h^۲) \qquad (۶۲.۷)$$

معادلهٔ اخیر بیان می‌داردکه، به‌ازای h به‌قدرکافی کوچك، عدد قابل‌محاسبهٔ

$$\frac{D_h(f)-D_{۲h}(f)}{۳} \qquad (۶۳.۷)$$

براورد خوبی‌است برای مؤلفهٔ خطای نافذ $C_۱h^۲$که معمولا مجهول‌است. این امردرجدول ۲.۷ که درآن اعداد به‌دست‌آمده از رابطهٔ (۶۳.۷) نیز ثبت شده‌اند، به خوبی مشاهده می‌شود.

البته نکته‌ای‌که در بررسیهای‌فوق‌وجود دارد، عبارت «به‌ازای h به‌قدرکافی‌کوچك» است. درحقیقت از جدول۲.۷ مشاهده می‌شودکه در مثال عددی فوق وقتی $h=۰٫۱$، $(D_h-D_{۲h})/۳$ تنها براوردی برای حدود دامنهٔ $C_۱h^۲$ است. و هنگامی‌که $h=۰٫۲$، $(D_h-D_{۲h})/۳$ حتی‌درحدود مقدار دامنه هم نیست. بنابراین عدد (۶۳.۷) رانباید بدون بررسی به‌عنوان براوردی ازخطا قبول‌کرد. بلکه، می‌باید عملیات خود را دربرابر اشتباه بزرگترحفظ‌کردکه این امر بایك بررسی ساده بر اساس بحث زیر امکان‌پذیر است: اگر $C_۱h^۲$ مؤلفهٔ خطای نافذ باشد، یعنی اگر $o(h^۲)$ درمقایسه با $C_۱h^۲$ «کوچك» باشد، آنگاه به‌موجب (۶۲.۷) داریم

$$C_۱h^۲\approx\frac{D_h(f)-D_{۲h}(f)}{۳}$$

وهمچنین

$$C_۱\left(\frac{h}{۲}\right)^۲\approx\frac{D_{h/۲}(f)-D_h(f)}{۳}$$

بنابراین خواهیم داشت

$$\frac{D_h(f)-D_{۲h}(f)}{D_{h/۲}(f)-D_h(f)}\approx\frac{C_۱h^۲}{C_۱h^۲/۴}=۴$$

به‌گونهٔ توضیحی، اگر $C_۱h^۲$ مؤلفهٔ خطا باشد، آنگاه نسبت تفاضلهای قابل‌محاسبه یعنی

$$R_h = \frac{D_h(f) - D_{۲h}(f)}{D_{h/۲}(f) - D_h(f)} \qquad (۶۴.۷)$$

می‌باید تقریباً برابر با ۴ باشد. این امر برای مثال عددی ما درجدول۲.۷، که نسبتهای R_h را نیز داده‌ایم کاملا آشکار است.

به‌محض رسیدن به‌این باور که (۶۳.۷) براورد خوبی برای خطا در $D_h(f)$ است و اطمینان مجدد به‌اینکه $R_h \approx ۴$، آنگاه می‌توان انتظارداشت که

$$D_h^{\wedge}(f) = D_h(f) + \frac{D_h(f) - D_{۲h}(f)}{۳} \qquad (۶۵.۷)$$

درمقایسه با $D_h(f)$، تقریب خیلی بهتری برای $D(f)$ باشد. بخصوص در این صورت، این باورحاصل می‌شود که

$$|D(f) = D_h^{\wedge}(f)| < \frac{|D_h(f) - D_{۲h}(f)|}{۳} \qquad (۶۶.۷)$$

برای اینکه ببینیم تا چه حد تقریب $D_h^{\wedge}(f)$ بهتر است، اکنون بیان مشروحتری از عبارت خطای

$$E_h(f) = -\frac{۱}{۶}h^۲ f'''(\xi)$$

برای $D_h(f)$ به‌دست می‌آوریم. به‌منظور تنوع، به‌جای تفاضلهای منقسم، سری تیلر را به‌کار می‌گیریم. اگر $f(x)$ دارای پنج مشتق پیوسته باشد، آنگاه با بسط $f(a+h)$ و $f(a-h)$ به‌سری تیلر درحول $x=a$ خواهیم داشت

$$f(a+h) = f(a) + f'(a)h + \frac{f''(a)h^۲}{۲} + \frac{f'''(a)h^۳}{۶} + \frac{f^{iv}(a)h^۴}{۲۴} + \frac{f^{v}(a)h^۵}{۱۲۰} + o(h^۵)$$

$$f(a-h) = f(a) - f'(a)h + \frac{f''(a)h^۲}{۲} - \frac{f'''(a)h^۳}{۶} + \frac{f^{iv}(a)h^۴}{۲۴} - \frac{f^{v}(a)h^۵}{۱۲۰} + o(h^۵)$$

معادلهٔ دوم را ازمعادلهٔ اول کم، وسپس نتیجه را بر $۲h$ تقسیم می‌کنیم، که درنتیجه خواهیم داشت

$$D_h(f)=f'(a)+\frac{f'''(a)h^2}{6}+\frac{f^{(5)}(a)h^4}{120}+o(h^4)$$

بنابراین

$$D(f)=D_h(f)+C_1h^2+C_2h^4+o(h^4) \qquad (67.7)$$

که در آن ثابتهای

$$C_1=\frac{-f'''(a)}{6} \qquad C_2=\frac{-f^{(5)}(a)}{120}$$

به h بستگی ندارند. بنابراین با گذاردن $2h$ به‌جای h در (۶۷.۷) داریم

$$D(f)=D_{2h}(f)+4C_1h^2+16C_2h^4+o(h^4) \qquad (68.7)$$

از کم کردن $\frac{1}{3}$ معادلهٔ (۶۸.۷) از $\frac{3}{4}$ معادلهٔ (۶۷.۷)، خواهیم داشت

$$D(f)=D_h^1(f)+C_2^1h^4+o(h^4) \qquad (69.7)$$

که در آن داریم

$$C_2^1=-4C_2=\frac{f^{(5)}(a)}{30}$$

و چون به‌موجب (۶۵.۷) داریم

$$D_h^1(f)=\frac{4D_h(f)-D_{2h}(f)}{3}$$

یك مقایسهٔ (۶۹.۷) با (۶۷.۷) نشان می‌دهد که $D_h^1(f)$ در مقایسه با $D_h(f)$ تقریبی از مرتبهٔ بالاتر برای $D(f)$ است: اگر $C_1\neq 0$، آنگاه $D(f)-D_h(f)$ (همراه با h) تنها بـا همان سرعت h^2 به‌سمت صفر میل مـی‌کند، در حالی کـه $D(f)-D_h^1(f)$ حداقل باسرعت h^4 به‌سمت صفرمیل خواهدکرد.

این‌روند به‌دست آوردن یك تقریب مرتبهٔ بالاتر را ازدوتقریب مرتبهٔ پایینتر، معمولا **برونیابی به‌سمت حد** یا **برونیابی به‌سمت صفر اندازه**[1] نامند. (برای توضیح این‌اصطلاح به‌تمرین ۶.۷ـ۳ نگاه کنید).

برونیابی به‌سمت حد به‌هیچ‌وجه به‌تقریبهای خطای $\mathcal{O}(h^2)$ محـدود نمی‌شود، برای مثال، از رابطهٔ (۶۹.۷) با قراردادن $2h$ به‌جای h، خواهیم داشت

$$D(f)=D_{2h}^1(f)+16C_2^1h^4+o(h^4)$$

بنابراین، ازکم‌کردن این رابطه از (۶۹.۷) و مرتب‌کردن نتیجه داریم

1. zero-grid size

$$C_2^1 h^4 = \frac{D_h^1(f) - D_{2h}^1(f)}{15} + o(h^4)$$

بنابراین، اگر قرار دهیم

$$D_h^2(f) = D_h(f) + \frac{D_h^1(f) - D_{2h}^1(f)}{15}$$

خواهیم داشت

$$D(f) = D_h^2(f) + o(h^4)$$

که نشان می‌دهد، $D_h^2(f)$ در مقایسه با $D_h^1(f)$ برای $D(f)$ حتی تقریبی از مرتبهٔ بالاتر است. روشنتر بگوییم، می‌توان نشان داد که اگر $f(x)$ به‌قدر کافی هموار باشد، آنگاه

$$D(f) = D_h^2(f) + C_2^2 h^6 + o(h^6) \qquad (۷۰.۷)$$

اما باید توجه کرد که به‌ازای هر مقدار خاص h، نمی‌توان انتظار داشت که $D_h^2(f)$ در مقایسه با $D_h^1(f)$ تقریب بهتری برای $D(f)$ باشد، مگر اینکه

$$\frac{D_h^1(f) - D_{2h}^1(f)}{15}$$

برآورد خوبی برای خطا در $D_h^1(f)$ باشد، یعنی مگر اینکه $C_2^1 h^4$ قسمت نافذ خطا در $D_h^1(f)$ باشد. این مطلب فقط زمانی صحیح است که تساوی زیر برقرار باشد

$$R_h^1 = \frac{D_h^1(f) - D_{2h}^1(f)}{D_{h/2}^1(f) - D_h^1(f)} \approx \frac{C_2^1 h^4}{C_2^1 (h/2)^4} = 16$$

بنابراین، پیش از رسیدن به‌این باور که رابطهٔ

$$|D(f) - D_h^2(f)| < \frac{|D_h^1(f) - D_{2h}^1(f)|}{15}$$

برقرار است، این شرط باید بررسی شود. در جدول ۳.۷ نتایج دوبار به‌کاربردن برونیابی به‌سمت حد را برای دنبالهٔ $D_h^1(f)$، که در جدول ۲.۷ محاسبه شده بود، فهرست کرده‌ایم. همچنین مقادیر مختلف R_h^1 را نیز داده‌ایم. کلیهٔ محاسبات با گرد کردن تا شش رقم اعشاری انجام گرفته‌اند.

سرانجام، هیچ اشکالی برای به‌کار گرفتن عدد ۲، که در بالا به‌کار بردیم، در کلیهٔ برونیابیها وجود ندارد. برای مثال، اگر q در واقع عددی ثابت باشد، آنگاه از رابطهٔ (۶۷.۷) نتیجه می‌گیریم

$$D(f) = D_{qh}(f) + q^2 C_1 h^2 + q^4 C_2 h^4 + o(h^4)$$

از کم کردن این معادله از (۶۷.۷) و مرتب کردن نتیجه، خواهیم داشت

جدول ۳.۷

h	$D_h(f)$	R_h	$D_h^1(f)$	R_h^1	$D_h^2(f)$
6.4	0.009839				
3.2	−0.009856	−0.57	−0.075508		
1.6	0.337545	2.37	0.453345	6.1	0.488602
0.8	0.484486	3.54	0.533466	12.5	0.538807
0.4	0.526009	3.88	0.539850	15.1	0.540276
0.2	0.536707	3.97	0.540273	15.7	0.540301
0.1	0.539402		0.540300		0.540302

$$C_1 h^2 = \frac{D_h(f) - D_{qh}(f)}{q^2 - 1} - (1 + q^2) C_2 h^4 + o(h^4)$$

بنابراین، به‌ازای

$$D_{h,q}(f) = D_h(f) + \frac{D_h(f) - D_{qh}(f)}{q^2 - 1}$$

خواهیم داشت

$$D(f) = D_{h,q}(f) - q^2 C_2 h^4 + o(h^4)$$

که نشان می‌دهد $D_{h,q}(f)$ یك تقریب $\mathcal{O}(h^4)$ برای $D(f)$ است. برای مثال ازجدول ۲.۷ محاسبه می‌کنیم که

$$D_{0.1,4}(f) = 0.5394402 + \frac{0.5394402 - 0.5426009}{16 - 1} = 0.5402295$$

که خطای آن تنها برابر با هفت واحد در آخرین رقم است.

نکات اصلی بحثهای فوق در الگوریتم زیر گنجانیده شده‌است.

الگوریتم ۱.۷: برونیابی به سمت حد. برای محاسبهٔ تقریب $L_h(f)$ برای عدد $L(f)$ به‌ازای هر $h > 0$ راه‌حلهایی‌داده شده‌اند، که در آنها $L_h(f)$ در رابطهٔ زیرصدق می‌کند

$$L(f) = L_h(f) + Ch^r + o(h^r) \qquad \text{به‌ازای کلیهٔ } h > 0$$

در این رابطه C عددی‌است ثابت مستقل از h و r عددی است مثبت.

یك h و یك عدد $q > 1$ (برای مثال $q = 2$) را انتخاب می‌کنیم و محاسبهٔ زیررا با دوعدد $L_h(f)$ و $L_{qh}(f)$ انجام می‌دهیم

$$L_{h,q}(f) = L_h(f) + \frac{L_h(f) - L_{qh}(f)}{q^r - 1}$$

آنگاه

$$L(f)=L_{h,q}(f)+o(h^r)$$

لذا برای h به‌قدر کافی کوچك داریم

$$|L(f)-L_{h,q}(f)|<|L_{h,q}(f)-L_h(f)| \qquad (۷۱.۷)$$

پیش از اینکه اطمینان کامل از رابطهٔ (۷۱.۷) به‌دست آید، می‌باید تحقیق کنیم که رابطهٔ

$$\frac{L_h(f)-L_{qh}(f)}{L_{h/p}(f)-L_h(f)}\approx\frac{Ch^r(q^r-۱)}{Ch^r(۱-p^{-r})}=p^r\frac{q^r-۱}{p^r-۱}$$

به‌ازای مقداری مانند $p>۱$ (مثلا، $p=q$) برقرار است.

تمرین

۱-۶.۷ با $f(x)=x+x^۲+x^۵$ و $a=۰$، مقادیر $D_h(f)$ و $D_h^{\downarrow}(f)$ را به‌ازای مقادیر مختلف h محاسبه کنید. چرا $D_h^{\downarrow}(f)$ درمقایسه با $D_h(f)$ همواره تقریب بدی برای $D(f)=f'(۰)$ است؟ (ازحساب با دقت بسیار زیاد استفاده کنید تا خطای گردکردن را به‌علت نارسا بودن رد کنید یا یك عبارت روشنی برای D_h و $D_h^{\downarrow}$ برحسب h به دست آورید.)

۲-۶.۷ با استفاده از روش برونیابی به‌سمت حد، مقدار $f'(۰ر۴)$ را برای داده‌های زیر پیدا کنید

x	$\sinh x=f(x)$
۰ر۳۹۸	۰ر۴۰۸۵۹۱
۰ر۳۹۹	۰ر۴۰۹۶۷۱
۰ر۴۰۰	۰ر۴۱۰۷۵۲
۰ر۴۰۱	۰ر۴۱۱۸۳۴
۰ر۴۰۲	۰ر۴۱۲۹۱۵

در این حالت مقدار برونیابی شده تقریب خوبی نیست. شرح دهید که چرا چنین است. [توجه: مقدارصحیح $f'(۰ر۴)$ برابراست با ۱ر۰۸۱۰۷۲].

۳-۶.۷ نشان دهیدکه برونیابی به‌سمت حد می‌تواند اساس جایگزینی تحلیلی گرفته شود

بخصوص با قراردادهای الگوریتم۱۰.۷ نشان دهیدکه

$$L_{h,q}(f)=\lim_{x\to 0} p(x)\approx \lim_{x\to 0} L_x(f)=L(f)$$

گـه در آن تقریب $p(x)$ برای $g(x)=L_x(f)$ از راه پیداکردن A و B، به‌طوری‌کـه رابطهٔ

$$p(x)=A+Bx^2$$

در $x=h$ و $x=qh$ با $g(x)$ تطابق داشته باشد، به‌دست‌آمده‌است. چگونه این عمل نام «برونیابی به‌سمت حد» را توجیه می‌کند؟

۷.۷* انتگرالگیری رامبرگ

برونیابی به‌سمت حد احتمالا بیشتر به‌علت‌کاربردش بـا قاعدهٔ ذوزنقه‌ای مرکب، کـه نام **انتگرالگیری رامبرگ** به‌خودگرفته، بیشتر معروف‌است. با تقریب قاعدهٔ ذوزنقه‌ای مرکب (به‌قسمت ۴.۷ نگاه‌کنید)

$$T_N=T_N(f)=h\sum_{i=1}^{N-1} f_i+\frac{h(f_0+f_N)}{2} \qquad (۷۲.۷)$$

برای عدد

$$I=I(f)=\int_a^b f(x)\,dx$$

شروع می‌کنیم. در اینجا N عددی‌است صحیح و مثبت‌که با رابطهٔ زیر به h بستگی دارد

$$h=\frac{b-a}{N}$$

و

$$f_i=f_{i,N}=f(a+ih) \qquad i=0,\ldots,N$$

اگر $f(x)$ چهار مرتبه به‌طور پیوسته مشتقپذیر باشد، از روابط (۵۰.۷) و (۵۱.۷) نتیجه می‌گیریم

$$I(f)=T_N(f)+C_1h^2+\mathcal{O}(h^4) \qquad (۷۳.۷)$$

که در آن ثابت $C_1=[f'(a)-f'(b)]/12$ مستقل از hاست. بنابراین برونیابی به‌سمت حد، را می‌توان به‌کار برد وخواهیم داشت

$$T_{N,q}(f)=T_N+\frac{T_N(f)-T_{N/q}(f)}{q^2-1}$$

که یك تقریب $\mathcal{O}(h^4)$ برای $I(f)$ است، در حالی‌که درحالت کلی $T_N(f)$ تنها دارای خطایی از مرتبهٔ $\mathcal{O}(h^2)$ است.

باید توجه کرد که انتخاب q یا N مقید به‌این شده‌است که N/q عدد صحیح باشد. معمولاً $q=2$ انتخاب می‌شود (لذا N باید زوج باشد). این انتخاب q ازنظرمحاسباتی این مزیت را داردکه کلیهٔ مقادیر تابع، که برای محاسبهٔ $T_{N/q}$ به‌کار گرفته می‌شود، برای محاسبهٔ T_N نیز می‌تواند به‌کار رود. بخصوص ثابت می‌کنیم که هرگاه N عددی زوج باشد، داریم

$$T_N(f)=\frac{T_{N/2}(f)}{2}+h\sum_{i=1}^{N/2} f(a+(2i-1)h) \qquad (۷۴۰۷)$$

چون به‌موجب (۷۲۰۷) داریم

$$T_N(f)=h\sum_{i=1}^{N-1} f(a+ih)+\frac{h(f(a)+f(b))}{2}$$

$$=h\sum_{i=1}^{N/2} f(a+(2i-1)h)+h\sum_{i=1}^{N/2-1} f(a+2ih)+\frac{h(f(a)+f(b))}{2}$$

دراینجا، مجموعیابی‌اول برنقاط «فرد» و مجموعیابی دوم برنقاط «زوج» تعمیم داده شده است. دوجملهٔ آخر را می‌توان به‌صورت زیرنوشت

$$\left[2h\sum_{j=1}^{N/2-1} f(a+j(2h))+\frac{2h(f(a)+f(b))}{2}\right]/2$$

بنابراین، از آنجایی‌که داریم

$$2h=\frac{2(b-a)}{N}=\frac{b-a}{N/2}$$

این دوجملهٔ آخر به $T_{N/2}(f)/2$ اضافه می‌شود. وهمین امر رابطهٔ (۷۴۰۷) را ثابت می‌کند. باید توجه داشت‌که (۷۴۰۷) را می‌توان به‌شکل ساده‌تری به‌صورت زیرنوشت

$$T_N(f)=\frac{T_{N/2}(f)+M_{N/2}(f)}{2}$$

که درآن M معرف قاعدهٔ مرکب نقطهٔ میانی (۴۹۰۷الف) است.

اگرتابع زیرعلامت انتگرال‌دارای $2k+2$ بار مشتق‌پیوسته‌باشد، می‌توان به‌طریقی

روشنتر از (۷۳.۷) نشان داد

$$I(f)=T_N(f)+C_1h^2+C_2h^4+C_3h^6+\cdots+C_kh^{2k}+\mathcal{O}(h^{2k+2})$$

که در آن ثابتهای C_1، ... ، C_k مستقل از h اند. از این‌رو با توجه به

$$T_N^1(f)=T_N(f)+\frac{T_N(f)-T_{N/2}(f)}{3}$$

خواهیم داشت

$$I(f)=T_N^1(f)+C_2^1h^4+C_3^1h^6+\cdots+C_k^1h^{2k}+\mathcal{O}(h^{2k+2})$$

که در آن ثابتهای C_2^1، ... ، C_k^1 مستقل از h هستند. بنابراین برونیابی بعدی معنی پیدا می‌کند. با قراردادن

$$T_N^2(f)=T_N^1(f)+\frac{T_N^1(f)-T_{N/2}^1(f)}{15}$$

داریم

$$I(f)=T_N^2(f)+C_3^2h^6+\cdots+C_k^2h_k^{2k}+\mathcal{O}(h^{2k+2})$$

به‌شکل کلیتر، دیده می‌شود که به‌ازای $m=1,\ldots,k$

$$T_N^m(f)=T_N^{m-1}(f)+\frac{T_N^{m-1}(f)-T_{N/2}^{m-1}(f)}{4^m-1}$$

یك $\mathcal{O}(h^{2m+2})$ تقریبی برای $I(f)$ است.

توجه داریم که محاسبهٔ T_N^m مستلزم محاسبهٔ $T_{N/2}^{m-1}$ و T_N^{m-1} و بنابراین مستلزم محاسبهٔ T_N^{m-2}, $T_{N/2}^{m-2}$, $T_{N/4}^{m-2}$, ... و سرانجام $T_{N/2^m}$، ... ، $T_{N/2}$ و T_N است. بنابراین $N/2^m$ باید عدد صحیحی مثلا به‌صورت

$$\frac{N}{2^m}=M$$

برای T_N^m تعریف شده باشد. ساده‌تر این است که این تقریبهای مختلف برای $I(f)$ را به‌صورت عناصر یك آرایهٔ مثلثی، موسوم به‌جدول T مجسم کنیم

$$\begin{array}{lllll}
T_M^0 & & & & \\
T_{2M}^0 & T_{2M}^1 & & & \\
T_{4M}^0 & T_{4M}^1 & T_{4M}^2 & & \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
T_{2^mM}^0 & T_{2^mM}^1 & T_{2^mM}^2 & \cdots & T_{2^mM}^m
\end{array}$$

دراینجا به‌جای T_N، T_N^0 نوشته شده‌است.

الگوریتم ۲.۷: انتگرالگیری رامبرگ تابع $f(x)$ که در بازهٔ $[a, b]$ تعریف شده‌وعدد صحیح مثبت M (معمولا $M = ۱$) داده شده‌اند.

$h := (b - a)/M$
Calculate $T_M^0 = h\Sigma_{i=1}^{M-1} f(a + ih) + h(f(a) + f(b))/2$
For $k = 1, 2, 3, \ldots,$ do:
  $h := h/2$
  Calculate $T_{2^kM}^0 = \frac{1}{2} T_{2^{k-1}M}^0 + h\Sigma_{i=1}^{2^{k-1}M} f(a + (2i - 1)h)$
  For $m = 1, \ldots, k,$ do:
    Calculate $T_{2^kM}^m = T_{2^kM}^{m-1} + (T_{2^kM}^{m-1} - T_{2^{k-1}M}^{m-1})/(4^m - 1)$

اگر $f(x)$ دارای $۲m+۲$ بارمشتق پیوسته باشد، آنگاه

$$I(f) = \int_a^b f(x)\,dx = T_{۲^kM}^m + \mathcal{O}\left(\left[\frac{b-a}{۲^kM}\right]^{۲m+۲}\right) \qquad k = m, m+۱, \ldots$$

همچنین، اگر k به‌اندازهٔ کافی بزرگ باشد، آنگاه

$$|I(f) - T_{۲^kM}^m| < |T_{۲^kM}^m - T_{۲^kM}^{m-۱}|$$

اما پیش از اینکه اطمینان کاملی‌ازاین نامساوی داشته باشیم، بررسی می‌کنیم که دست‌کم رابطهٔ زیر برقرار باشد

$$R_k^{m-۱} = \frac{T_{۲^{k-۱}M}^{m-۱} - T_{۲^{k-۲}M}^{m-۱}}{T_{۲^kM}^{m-۱} - T_{۲^{k-۱}M}^{m-۱}} \approx ۴^m$$

□ **مثال ۹.۷:** انتگرالگیری رامبرگ را برای مثال۱.۷ به‌کارگیرید. انتگرال موردنظر به‌صورت زیر است

$$I(f) = \int_۰^۱ e^{-x^۲}\,dx$$

برنامهٔ فورترنی که در زیر داده شده، شش ردیف اول جدول T و جدول نسبتهای R_k^m متناظر با آن را به‌صورت زیرپدید می‌آورد:

جدول T رامبرگ

0.7313700E 00				
0.7429838E 00	0.7468551E 00			
0.7458653E 00	0.7468258E 00	0.7468238E 00		
0.7465842E 00	0.7468238E 00	0.7468237E 00	0.7468237E 00	
0.7467639E 00	0.7468237E 00	0.7468237E 00	0.7468237E 00	0.7468237E 00
0.7468069E 00	0.7468212E 00	0.7468210E 00	0.7468210E 00	0.7468209E 00

جدول نسبتها

4.03			
4.01	14.88		
4.00	16.50	0.0	
4.17	0.05	0.0	0.0

مقدار M برابر با ۲ انتخاب شده بود، لـذا اولین عنصرجدول T، $T_{۲}(f)$ است. توجه‌کنیدکه اولین‌ستون نسبتها به‌نحومطلوبی به‌سمت ۴ همگرا می‌شود، ولی ازآن به‌بعد شروع می‌کند به‌دورشدن از ۴. این نتیجه حتی درستون دوم نسبتها کـه به‌سمت ۱۶ میل می‌کند (چنانکه باید باشد) مشخصتر است وسپس، آن‌گونه‌که آخرین عنصر نشان می‌دهد غیرقابل‌پیشبینی‌می‌شود. نتیجه اینکه: خطا در درایه‌های آخرین‌سطرجدول T، عمدتاً بر اثر گردکردن حاصل‌شده نه بر اثرتجرید، بنابراین به‌نظر می‌رسدکه

$$۰٫۷۴۶۸۲۳۷$$

بهترین براوردی برای $I(f)$ است‌که از این محاسبات می‌توانسته به‌دست‌آید. چون

$$R_{۴}^{۱}=۱۶٫۵\approx ۱۶=۴^{۲}$$

و

$$T_{۲,M}^{۱}=T_{۲,M}^{۲}=۰٫۷۴۶۸۲۳۷$$

که این تساویها تا تعداد ارقام نشان داده شده برقرارند، و می‌توان نتیجه گرفت‌کـه این براورد تا تعداد ارقام نشان داده شده صحیح‌است. ولی به‌طور دقیق داریم

$$\int_{۰}^{۱} e^{-x^{۲}}\,dx=۰٫۷۴۶۸۲۴۱۳۲۷۹\ldots$$

که اختلاف بین این دوعدد و براورد «صحیح» انجام شده ناشی از این حقیقت‌است‌کـه درمحاسبات با تابع زیرعلامت انتگرال

$$f(x)=e^{-x^{۲}}$$

سروکار نداریم بلکه با $f(x)$گرد شده، یعنی با تابع

$$\mathrm{F(X)=EXP(-X*X)}$$

مواجه هستیم.کلیهٔ محاسبات بادقت ساده روی IBM ۳۶۰که مخصوصاً مشخصهٔ نامطلوبی درگردکردن دارد، انجام گرفته‌است.

برنامهٔ فورترن برای مثال۹.۷

```
      REAL T(100)
      EXTERNAL FERR
      CALL RMBERG( FERR, 0., 1., 2, T, 6 )
                                          STOP
      END
      SUBROUTINE RMBERG ( F, A, B, MSTART, T, NROW )
C  CONSTRUCTS AND PRINTS OUT THE FIRST  NROW  ROWS OF THE ROMBERG T-
C  TABLE FOR THE INTEGRAL OF  F(X)  FROM  A  TO  B  , STARTING WITH THE
C  TRAPEZOIDAL SUM ON  MSTART  INTERVALS.
      INTEGER MSTART,NROW,   I,K,M
      REAL A,B,T(NROW,NROW),   H,SUM
      M = MSTART
      H = (B-A)/M
      SUM = (F(A) + F(B))/2.
      IF (M .GT. 1) THEN
         DO 10 I=1,M-1
   10       SUM = SUM + F(A+FLOAT(I)*H)
      END IF
      T(1,1) = SUM*H
      PRINT 610
  610 FORMAT('1',10X,'ROMBERG T-TABLE'//)
      PRINT 611, T(1,1)
  611 FORMAT(7E15.7)
      IF (NROW .LT. 2)                    RETURN
C
      DO 20 K=2,NROW
         H = H/2.
         M = M*2
         SUM = 0.
         DO 11 I=1,M,2
   11       SUM = SUM + F(A+FLOAT(I)*H)
         T(K,1) = T(K-1,1)/2. + SUM*H
         DO 12 J=1,K-1
C              SAVE DIFFERENCES FOR LATER CALC. OF RATIOS
            T(K-1,J) = T(K,J) - T(K-1,J)
   12       T(K,J+1) = T(K,J) + T(K-1,J)/(4.**J - 1.)
   20    PRINT 611, (T(K,J),J=1,K)
      IF (NROW .LT. 3)                    RETURN
C                CALCULATE RATIOS
      PRINT 620
  620 FORMAT(///11X,'TABLE OF RATIOS'//)
      DO 30 K=1,NROW-2
         DO 25 J=1,K
            IF (T(K+1,J) .EQ. 0.) THEN
               RATIO= 0.
            ELSE
               RATIO = T(K,J)/T(K+1,J)
            END IF
   25       T(K,J) = RATIO
   30    PRINT 630, (T(K,J),J=1,K)
  630 FORMAT(8F10.2)
                                          RETURN
      END
      REAL FUNCTION FERR(X)
      REAL X
      FERR = EXP(-X*X)
                                          RETURN
      END
```

□

تمرین

۱-۷.۷ ثابت کنید که در انتگرالگیری رامبرگ تساوی $T^{1}_{2M} = S_M$ برقرار است که در آن S_M، قاعدهٔ مرکب سیمپسن است، ر. ک. (۴۸.۷).

۲-۷.۷ سعی کنید مقدار $I(f) = \int_a^b f(x)\,dx$ را تا تقریب 10^{-6} برآورد کنید. برای

انجام این‌کار، انتگرالگیری رامبرگ را برای هریك از موارد زیر به‌کار گیرید:

(الف) $f(x)=x^{۲}$ $a=۰$، $b=۱$، M اختیاری

(ب) $f(x)=\sin ۱۰۱\pi x$ $a=۰$، $b=۱$، $M=۱$

(پ) $f(x)=۱+\sin ۱۰\pi x$ $a=۰$، $b=۱$، $M=۱$

(ت) $f(x)=\left|x-\frac{۱}{۳}\right|$ $a=۰$، $b=۱$، $M=۱$ و $M=۳$

(ث) $f(x)=\sqrt{x}$ $a=۰$، $b=۱$، M اختیاری

(ج) $f(x)=\begin{cases}\dfrac{\sin x}{x} & x\neq ۰\\ ۱ & x=۰\end{cases}$ $a=۰$، $b=۱$

۳-۷.۷ با استفاده از انتگرالگیری رامبرگ و به‌ازای داده‌های زیر، انتگرال $\int_{۱}^{۱٫۸} f(x)\,dx$ را تا آنجاکه ممکن است با دقت محاسبه کنید. با شروع از $M=۱$، جدول T را پر کنید.

x	$f(x)$
۱٫۰	۰٫۳۶۷۸۷۹۴۴
۱٫۱	۰٫۳۶۶۱۵۸۱۹
۱٫۲	۰٫۳۶۱۴۳۳۰۵
۱٫۳	۰٫۳۵۴۲۹۱۳۳
۱٫۴	۰٫۳۴۵۲۳۵۷۴
۱٫۵	۰٫۳۳۴۶۹۵۲۴
۱٫۶	۰٫۳۲۳۰۳۴۴۳
۱٫۷	۰٫۳۱۰۵۶۱۹۹
۱٫۸	۰٫۲۹۷۵۳۸۰۰

۴-۷.۷ با برونیابی از قاعدهٔ نقطهٔ میانی و قاعدهٔ ذوزنقه‌ای، برای

$$I_h(f)=\int_{-h}^{h} f(x)\,dx$$

قاعدهٔ سیمپسن را به‌دست آورید. (راهنمایی: ترکیب خطی مناسبی از دو معادلهٔ

$$I_h(f)=T(f)+C_T h^2+\mathcal{O}(h^4) \qquad I_h(f)=M(f)+C_M h^2+\mathcal{O}(h^4)$$

برای حذف جملات h^2 تشکیل دهید، برای انجام این‌کار، باید مقادیر ثابتهای C_T و C_M را پیدا کنید).

۸

حل معادلات دیفرانسیل

بسیاری از مسائل مهندسی وعلوم را می‌توان برحسب معادلات دیفرانسیل بیان کرد. قسمت اعظم انگیزهٔ ساختن کامپیوترهـای ابتدایـی ناشی از نیاز به‌محاسبات دقیق وسریع مسیر پرتابه‌ها بوده‌است. امروزه کامپیوترها درحل معادلات مربوط به‌پرتاب موشکها ونظریهٔ اقمار مصنوعی ونیز درمورد نظریهٔ شبکه‌های الکتریکی، خمش پرتوهای نوری، پایداری هواپیما، نظریهٔ ارتعاش و نظریه‌های دیگر به‌طور وسیعی به‌کار گرفته می‌شوند.

فرض بر آن‌است که دانشجویان با نظریهٔ مقدماتی معادلات دیفرانسیل آشنایی دارند. در ریاضیات اول، تکنیکهای مختلفی برای حل دستهٔ مشخصی از معادلات دیفرانسیل به‌یك شکل معین آموزش داده می‌شود. ولی اکثر معادلاتی راکه درعمل با آنها مواجه می‌شویم نمی‌توانیم با روشهای تحلیلی حل کنیم و بناچار لازم می‌شود از روشهای عددی استفاده کنیم. خوشبختانه برای حل معادلات دیفرانسیل بـا کامپیوتر، روشهای بسیار خوبی وجود دارند. دراین فصل، چند رده ازاین روشها را استخراج و آنها را ازجنبهٔ کارایی محاسباتی ارزیابی خواهیم کرد.

۱۰۸ مقدمات ریاضی

در اینجا مرور برخی تعاریف و مفاهیم مقدماتی نظریهٔ معادلات دیفرانسیل را بسیار مفید می‌دانیم. معادله‌ای راکه شامل رابطه‌ای بین یك تابع مجهول ویك یا چند مشتق آن باشد

معادلهٔ دیفرانسیل نامند. ما همواره فرض‌خواهیم کرد که می‌توان این معادله‌را نسبت به‌مشتق بالاترین مرتبه‌اش حل کرد. یك معادلهٔ دیفرانسیل معمولی ازمرتبهٔ n به‌شکل زیرخواهد بود

$$y^{(n)}(x)=f(x, y(x), y'(x), \cdots, y^{(n-1)}(x)) \qquad (۱.۸)$$

منظور ما ازجواب معادلهٔ (۱.۸) تابعی‌است به‌صورت $\phi(x)$ که به‌طور پیوسته دریك بازهٔ معلوم، nبار مشتقپذیر باشد ودر معادلـهٔ (۱.۸) صدق کند، یعنی $\phi(x)$ باید در رابطـه‌ای به‌صورت

$$\phi^{(n)}(x)=f(x, \phi(x), \phi'(x), \phi''(x), \cdots, \phi^{(n-1)}(x))$$

صدق نماید. معمولا جواب عمومی معادلهٔ (۱.۸) دارای n ثابت اختیاری‌است و بنابراین یك خانوادهٔ n پارامتری ازجوابها وجود دارد. اگر $y(x_0)$، $y'(x_0)$، ...، و $y^{n-1}(x_0)$ در نقطـهٔ $x=x_0$ تعریف شده باشند، آنگاه یك **مسئله با مقدار اولیه** خواهیـم داشت. همواره فرض ما براین‌است که تابع f درشرایط کافی برای تضمین یك جواب یگانه برای این مسئله با مقدار اولیه، صدق می‌کند. $y'=y$ یك مثال ساده از معادلهٔ مرتبهٔ یك است، جواب عمومی آن $y(x)=Ce^x$ است که در آن C یك ثابت اختیاری‌است. در صورتی‌که شرط اولیهٔ $y(x_0)=y_0$ معلوم باشد، جواب را می‌توان به‌شکل $y(x)=y_0e^{(x-x_0)}$ نوشت.

بعلاوه معادلات دیفرانسیل به‌دو ردهٔ **خطی** و **غیرخطی** تقسیم‌بندی شده‌اند. یك‌معادله را زمانی خطی گویندکه تابع f درمعادلهٔ (۱.۸) به‌طور خطی شامل y ومشتقهای آن باشد. معادلات دیفرانسیل خطی‌این ویژگی مهم را دارندکه اگر $y_1(x)$، $y_2(x)$، ...، $y_m(x)$ جوابهایی از (۱.۸) باشند، آنگاه $C_1y_1(x)+C_2y_2(x)+\cdots+C_my_m(x)$ به‌ازای ثابتهای‌اختیاری C_i نیز یك جواب خواهدبود. $y''=y$ یك معادلهٔ ساده ازمرتبهٔ دو است. می‌توان به‌آسانی تحقیق کردکه e^x و e^{-x} جوابهای این معادله هستند و بنابراین به‌موجب ویژگی خطی‌بودن، حاصلجمع زیرنیز یك جواب آن‌است:

$$y(x)=C_1e^x+C_2e^{-x} \qquad (۲.۸)$$

دوجواب y_1 و y_2 از معادلهٔ دیفرانسیل مرتبـهٔ دو را **مستقل خطی** گویند، اگـر دترمینان ورونیسکی[1] این‌جوابها مخالف‌صفر باشد، که بنابر تعریف، **دترمینان و رونیسکی** به‌صورت زیراست

$$W(y_1, y_2)=y_1y_2'-y_2y_1'=\begin{vmatrix} y_1 & y_1' \\ y_2 & y_2' \end{vmatrix} \qquad (۳.۸)$$

مفهوم استقلال خطی را می‌توان برای جوابهای معادلات از مرتبهٔ بالاتر نیز تعمیم‌داد. اگر $y_1(x)$، $y_2(x)$، ...، $y_n(x)$، nجواب به‌طور خطی مستقل ازیك‌معادلهٔ دیفرانسیل همگن

1. Wronskian

درجهٔ n باشند، آنگاه

$$y(x)=C_1y_1(x)+C_2y_2(x)+\cdots+C_ny_n(x)$$

جواب عمومی نامیده می‌شود.

دربین معادلات خطی، آنهایی که ضرایب ثابت دارند به‌ویژه مفیدند، از این جهت که راه‌حل آنها ساده‌تر است. معادلهٔ دیفرانسیل مرتبهٔ nام با ضرایب ثابت به‌صورت زیر نوشته می‌شود

$$Ly=y^{(n)}+a_{n-1}y^{(n-1)}+\cdots+a_0y^{(0)}=0 \qquad (۴.۸)$$

که در آن a_iها حقیقی فرض شده‌اند. اگر به‌دنبال جوابهایی از (۴.۸) که به‌صورت $e^{\beta x}$ هستند باشیم، از قراردادن مستقیم این مقدار درمعادله می‌بینیم که β باید در بسجمله‌ای‌زیر صدق کند

$$\beta^n+a_{n-1}\beta^{n-1}+\cdots+a_0=0 \qquad (۵.۸)$$

این معادله را **معادلهٔ مشخصهٔ** معادلهٔ دیفرانسیل مرتبهٔ nام (۴.۸) نامند. اگر معادلهٔ (۵.۸) n ریشه متمایز $\beta_i(i=1,\ldots,n)$ داشته باشد، آنگاه می‌توان نشان داد که عبارت

$$y(x)=C_1e^{\beta_1x}+C_2e^{\beta_2x}+\cdots+C_ne^{\beta_nx} \qquad (۶.۸)$$

جواب عمومی معادلهٔ (۴.۸) است، که در آن C_iها مقادیر ثابت اختیاری هستند. اگر $\beta_1=\alpha+i\beta$ یك ریشهٔ همتافت از معادلهٔ (۵.۸) باشد، مزدوج آن یعنی $\beta_2=\alpha-i\beta$ نیز یك ریشه‌است. متناظر با این ریشه‌های همتافت مزدوج، دوجواب $y_1=e^{\alpha x}\cos\beta x$ و $y_2=e^{\alpha x}\sin\beta x$ وجود دارد که مستقل‌خطی هستند. هنگامی که رابطهٔ (۵.۸) دارای ریشه‌های مکرر باشد، تکنیکهای خاصی برای به‌دست‌آوردن جوابهای مستقل‌خطی وجود دارند. درحالت خاص، اگر β_1 یك ریشهٔ مکرر رابطهٔ (۵.۸) باشد، آنگاه $y_1=e^{\beta_1x}$ و $y_2=xe^{\beta_1x}$ جوابهای مستقل خطی (۴.۸) می‌شوند. برای معادلهٔ خاص $y''+a^2y=0$، معادلهٔ مشخصه به‌صورت $\beta^2=-a^2$ می‌باشد که ریشه‌های آن عبارت‌اند از $\beta_{1,2}=\pm ia$ و جواب عمومی آن به‌صورت $y(x)=C_1\cos ax+C_2\sin ax$ است.

و سرانجام، اگر معادلهٔ (۱.۸) خطی ولی غیرهمگن باشد، یعنی اگر

$$Ly=g(x) \qquad (۷.۸)$$

و اگر $\zeta(x)$ یك جواب خصوصی(۷.۸) باشد، یعنی اگر

$$L\zeta=g(x)$$

آنگاه با فرض آنکه ریشه‌های (۵.۸) متمایزند، جواب عمومی (۷.۸) به‌صورت زیر درمی‌آید

$$y=\zeta(x)+C_1e^{\beta_1 x}+C_2e^{\beta_2 x}+\ldots+C_ne^{\beta_n x} \qquad (۸.۸)$$

□ **مثال:** جواب معادلهٔ

$$y''-۴y'+۳y=x \quad \text{(الف)}$$

را که درشرایط اولیهٔ

$$y(۰)=\frac{۴}{۹},\ y'(۰)=\frac{۷}{۳} \quad \text{(ب)}$$

صدق کند پیدا کنید.

۱. برای پیداکردن یک جواب خاص $\zeta(x)$ برای قسمت (الف) تابع $\zeta(x)=ax+b$ را به‌عنوان جواب بررسی می‌کنیم، زیرا سمت راست تساوی یک بسجمله‌ای ازدرجهٔ نابیشتر از یک است، وسمت چپ نیز یک چنین بسجمله‌ای است هرگاه $y=y(x)$. با گذاردن $\zeta(x)$ در (الف) مقادیر $a=۱/۳$، $b=۴/۹$ را به دست خواهیم آورد. بنابراین

$$\zeta(x)=\frac{۱}{۳}x+\frac{۴}{۹}$$

۲. برای پیداکردن جوابهای معادلهٔ همگن

$$y''-۴y'+۳y=۰$$

معادلهٔ مشخصهٔ زیر را بررسی می‌کنیم

$$\beta^۲-۴\beta+۳=۰$$

ریشه‌های این معادله عبارت‌اند از $\beta_1=۳$ و $\beta_2=۱$. بنابراین دوجواب مستقل خطی از دستگاه همگن، به‌گونهٔ زیرند

$$y_1(x)=e^{۳x} \qquad y_2(x)=e^{x}$$

۳. جواب عمومی معادلهٔ (الف) به‌صورت زیراست

$$y(x)=\frac{x}{۳}+\frac{۴}{۹}+C_1e^{۳x}+C_2e^{x}$$

۴. برای پیداکردن جوابهایی که درشرایط (ب) صدق کنند، باید داشته باشیم

$$y(۰)=\frac{۴}{۹}+C_1+C_2=\frac{۴}{۹}$$

$$y'(0)=\frac{1}{3}+3C_1+C_2=\frac{7}{3}$$

جواب این دستگاه برابر است با $C_1=1$، $C_2=-1$. بنابراین جواب مورد نظر برابر است با

$$y(x)=\frac{x}{3}+\frac{4}{9}+e^{3x}-e^{x} \qquad \square$$

تمرین

۱-۱۰۸ جواب عمومی معادلات زیر را پیدا کنید

(الف) $y'=-2y$ (ب) $y''-4y'+4y=0$

(پ) $y'''-2y''-y'+2y=0$ (ت) $y'-ay=x$

(ث) $y'-xy=e^{x}$ (ج) $y''-2y'+2y=0$

۲-۱۰۸ جوابهای مسائل با مقدار اولیهٔ زیر را پیدا کنید:

(الف) $y'+2y=1$ $y(0)=1$

(ب) $y''-a^2y=0$ $y(0)=0$، $y'(0)=1$

(پ) $y''-4y'+4y=x$ $y(0)=0$، $y'(0)=1$

۲.۸ معادلات تفاضلی ساده

برای تحلیل روشهای عددی در حل معادلات دیفرانسیل، لازم است که با نظریهٔ سادهٔ معادلات تفاضلی مختصری آشنایی پیدا کنیم. یك معادلهٔ تفاضلی از مرتبهٔ N رابطه‌ای است بین تفاضلهای $y_n=\Delta^0 y_n, \Delta^1 y_n, \Delta^2 y_n, \dots, \Delta^N y_n$ از یك دنبالهٔ عددی، یعنی

$$\Delta^N y_n=f(n, y_n, \Delta y_n, \dots, \Delta^{N-1} y_n) \qquad (۹.۸)$$

یك جواب این معادلهٔ تفاضلی به‌صورت دنباله‌ای است از اعداد y_m، y_{m+1}، y_{m+2}، ... به‌طوری که به‌ازای $n=m, m+1, m+2, \dots$ رابطهٔ (۹.۸) برقرار باشد. بنابراین درصورتی که یك معادلهٔ دیفرانسیل متضمن توابعی باشد که خود و مشتقاتشان بر بازه‌ای از اعداد حقیقی تعریف شده باشند، یك معادلهٔ تفاضلی متضمن توابعی است که خود و تفاضلاتشان بر «بازه»ای از اعداد صحیح تعریف شده باشند.

اگر (۹.۸) یك معادلهٔ تفاضلی **خطی** باشد که سمت راست آن به‌طور خطی به

y_n، ...، $\Delta^{N-1}y_n$ بستگی داشته باشد، آنگاه مطابق معمول می‌توان آن‌را صریحاً برحسب y_jها بصورت زیر نوشت

$$y_{n+N}+a_{n,N-1}y_{n+N-1}+a_{n,N-2}y_{n+N-2}+\cdots+a_{n,0}y_n=b_n$$

بدیهی است که یك معادلهٔ تفاضلی خطی از مرتبهٔ N را می‌توان به‌عنوان یك دستگاه معادلات خطی (متناهی و یا نامتناهی) درنظر گرفت که ماتریس ضرایب آن ماتریسی است نواری* که پهنای نوار آن $N+1$ است.

مثالهای زیر مثالهای ساده‌ای از معادلات تفاضلی خطی هستند.

(۱۰.۸ الف) $y_{n+1}-y_n=1$ به‌ازای جمیع مقادیر n

(۱۰.۸ ب) $y_{n+1}-y_n=n$ $n\geqslant 0$ به‌ازای جمیع مقادیر

(۱۰.۸ پ) $y_{n+1}-(n+1)y_n=0$ $n\geqslant 0$ به‌ازای جمیع مقادیر

(۱۰.۸ ت) $y_{n+2}-(2\cos\gamma)y_{n+1}+y_n=0$ به‌ازای جمیع مقادیر n

با قراردادن مستقیم مقادیر می‌توان نشان داد که معادلات بالا دارای جوابهای زیر هستند

(۱۱.۸ الف) $y_n=n+c$ به‌ازای جمیع مقادیر n

(۱۱.۸ ب) $y_n=\dfrac{n(n-1)}{2}+c$ $n\geqslant 0$ به‌ازای جمیع مقادیر

(۱۱.۸ پ) $y_n=cn!$ $n\geqslant 0$ به‌ازای جمیع مقادیر

(۱۱.۸ ت) $y_n=c\cos\gamma n$ به‌ازای جمیع مقادیر n

که در آنها c یك ثابت اختیاری است.

---

* توضیح مترجم: ماتریس نواری، ماتریسی است به‌شکل زیر

$$A=\begin{bmatrix} d_1 & c_1 & e_1 & & & \\ a_2 & d_2 & c_2 & e_2 & & \\ & a_3 & d_3 & c_3 & e_3 & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & a_{N-2} & d_{N-2} & c_{N-2} & e_{N-2} \\ & & & a_{N-1} & d_{N-1} & c_{N-1} \end{bmatrix}$$

معادلهٔ تفاضلی خطی همگن از مرتبهٔ N با ضرایب ثابت

$$y_{n+N}+a_{N-1}y_{n+N-1}+\cdots+a_0y_n=0 \qquad (۱۲.۸)$$

را به‌تفصیل مورد بررسی قرار می‌دهیم. همان‌گونه‌که درمعادلات دیفرانسیل خطی همگن با ضرایب ثابت عمل می‌کردیم، به‌دنبال جوابهایی می‌گردیم کـه به‌شکل $y_n=\beta^n$ (به‌ازای جمیع مقادیر) باشند. ازگذاردن این مقدار در (۱۲.۸)، نتیجهٔ زیر به‌دست می‌آید

$$\beta^{n+N}+a_{N-1}\beta^{n+N-1}+\cdots+a_0\beta^n=0$$

که از تقسیم آن بر β^n، معادلهٔ مشخصهٔ

$$p(\beta)=\beta^N+a_{N-1}\beta^{N-1}+\cdots+a_0=0 \qquad (۱۳.۸)$$

به‌دست می‌آیدکه **بسجمله‌ای مشخصه‌ای** است ازدرجهٔ N. درابتدافرض‌می‌کنیم‌که ریشه‌های آن یعنی β_1، β_2، ...، β_N متمایز باشند، سپس β_1^n، β_2^n، ...، β_N^n همگی‌جوابهای(۱۲.۸) هستند و بنابر ویژگی خطی‌بودن، نتیجه می‌شودکه مقدار

$$y_n=c_1\beta_1^n+c_2\beta_2^n+\cdots+c_N\beta_N^n \qquad (۱۴.۸)$$

به‌ازای جمیع مقادیر n

که c_iها درآن ثابتهای اختیاری هستند، یك جواب معادلهٔ (۱۲.۸) است. بعلاوه، در این مورد می‌توان نشان دادکه (۱۴.۸) جواب عمومی معادلهٔ (۱۲.۸) است.

به‌عنوان مثال، معادلهٔ تفاضلی

$$y_{n+3}-2y_{n+2}-y_{n+1}+2y_n=0 \qquad (۱۵.۸)$$

معادله‌ای‌است از مرتبهٔ سوم و دارای معادلهٔ مشخصهٔ زیر

$$\beta^3-2\beta^2-\beta+2=0$$

ریشه‌های این معادلهٔ بسجمله‌ای برابراست با $+1$، -1 و 2 و جواب عمومی معادلـهٔ (۱۵.۸) چنین‌است

$$y_n=c_1(1)^n+c_2(-1)^n+c_3(2)^n$$
$$=c_1+(-1)^nc_2+2^nc_3 \qquad (۱۶.۸)$$

اگر اولین $N-1$ مقدار y_n داده شده باشند، معادلهٔ تفاضلی بـا **مقدار اولیهٔ** حاصل را می‌توان آشکارا به‌ازای جمیع مقادیری‌کـه به n داده می‌شود حل کرد. بنابراین اگر در (۱۵.۸)، مقادیراولیهٔ $y_0=0$ و $y_1=1$ و $y_2=1$ داده شده باشند، مقدار y_3 که از(۱۵.۸) محاسبه می‌شود برابراست با

$$y_3=2(1)+1-0=3$$

اگر استفاده از (۱۵.۸) را همچنان ادامه دهیم، خواهیم داشت $y_۴=۵$، $y_۵=۱۱$ و قس‌علیهذا. اما این عمل یك فرمول ثابتی برای y_n به‌ما نمی‌دهد. ولی با استفاده از رابطهٔ (۱۶.۸) و قرار دادن شرایط اولیهٔ $n=۰,۱,۲$، دستگاه معادلات زیر را برای $c_۱$، $c_۲$ و $c_۳$ به‌دست می‌آوریم

$$۰=c_۱+c_۲+c_۳$$

$$۱=c_۱-c_۲+۲c_۳$$

$$۱=c_۱+c_۲+۴c_۳$$

که جواب آن برابراست با $c_۱=۰$، $c_۲=-۱/۳$ و $c_۳=۱/۳$، لذا جواب ثابت[1] برای این مسئله با مقدار اولیه به‌صورت زیر است

$$y_n=-\frac{۱}{۳}(-۱)^n+\frac{۲^n}{۳}$$

اگر بسجمله‌ای مشخصهٔ (۱۳.۸) دارای یك‌جفت ریشهٔ مزدوج همتافت باشد، بازهم جواب می‌تواند به‌شکل حقیقی بیان شود. بنابراین اگر داشته باشیم $\beta_۱=\alpha+i\beta$ و $\beta_۲=\alpha-i\beta$، در ابتدا $\beta_{۱,۲}$ را به‌شکل قطبی[2] بیان می‌کنیم

$$\beta_۱=re^{i\theta}$$

$$\beta_۲=re^{-i\theta}$$

که در آن $r=\sqrt{\alpha^۲+\beta^۲}$ و $\theta=\arctan(\beta/\alpha)$. دراین حال جواب (۱۲.۸) متناظر با این ریشه‌های مزدوج به‌صورت زیرخواهد شد

$$\begin{aligned}c_۱\beta_۱^n+c_۲\beta_۲^n&=c_۱r^ne^{in\theta}+c_۲r^ne^{-in\theta}\\&=r^n[c_۱(\cos n\theta+i\sin n\theta)+c_۲(\cos n\theta-i\sin n\theta)]\\&=r^n(C_۱\cos n\theta+C_۲\sin n\theta)\end{aligned}$$

که در آن $C_۱=c_۱+c_۲$ و $C_۲=i(c_۱-c_۲)$. به‌عنوان یك مثال ساده، معادلهٔ تفاضلی‌زیر را درنظرمی‌گیریم

$$y_{n+۲}-۲y_{n+۱}+۲y_n=۰ \qquad (۱۷.۸)$$

معادلهٔ مشخصهٔ آن $\beta^۲-۲\beta+۲=۰$، و ریشه‌های این معادله برابر است با $\beta_{۱,۲}=۱\pm i$. بنابراین‌داریم $r=\sqrt{۲}$ و $\theta=\pi/۴$، لذا جواب عمومی معادلهٔ(۱۷.۸)

1. closed-form solution 2. polar

به‌صورت زیر است

$$y_n=(\sqrt{۲})^n\left(C_۱\cos\frac{n\pi}{۴}+C_۲\sin\frac{n\pi}{۴}\right)$$

اگر $\beta_۱$ یك ریشهٔ مضاعف معادلهٔ مشخصهٔ (۱۳۰۸) باشد، آنگاه جواب دوم معادلهٔ (۱۳۰۸) برابراست با $n\beta_۱^n$. برای تحقیق این امر، ملاحظه می‌کنیم که اگر $\beta_۱$ یك ریشهٔ مضاعف $p(\beta)$ باشد، آنگاه $p(\beta_۱)=۰$ و نیز $p'(\beta_۱)=۰$. حال با قراردادن $y_n=n\beta_۱^n$ درمعادلهٔ (۱۲۰۸) و مرتب کردن آن و با توجه به‌تساوی $p(\beta_۱)=p'(\beta_۱)=۰$، خواهیم داشت

$$\begin{aligned}(n+N)\beta_۱^{n+N}+a_{n-۱}(n+N-۱)\beta_۱^{n+N-۱}+\cdots+a_۰n\beta_۱^n\\ =\beta_۱^n\{n(\beta_۱^n+a_{N-۱}\beta_۱^{N-۱}+\cdots+a_۰)\\ +\beta_۱[N\beta_۱^{N-۱}+a_{N-۱}(N-۱)\beta_۱^{N-۲}+\cdots+a_۱]\}\\ =\beta_۱^n[np(\beta_۱)+\beta_۱p'(\beta_۱)]=۰\end{aligned}$$

بعلاوه می‌توان نشان دادکه دوجواب $\beta_۱^n$ و $n\beta_۱^n$ مستقل خطی هستند.

به‌عنوان مثال، معادلهٔ تفاضلی زیر را درنظر می‌گیریم

$$y_{n+۳}-۵y_{n+۲}+۸y_{n+۱}-۴y_n=۰$$

ریشه‌های معادلهٔ مشخصهٔ آن عبارت‌اند از ۲، ۲، ۱، و جواب عمومی آن به‌صورت زیر است

$$y_n=۲^n(c_۱+nc_۲)+c_۳$$

و بالاخره جواب معادلهٔ تفاضلی‌خطی غیرهمگن با ضرایب ثابت را مورد بررسی قرار می‌دهیم. جواب عمومی معادلهٔ

$$y_{n+N}+a_{N-۱}y_{n+N-۱}+\cdots+a_۰y_n=b_n \qquad (۱۸۰۸)$$

را می‌توان به‌شکل زیر نوشت

$$y_n=y_n^G+y_n^P$$

که در آن y_n^G جواب عمومی دستگاه همگن (۱۲۰۸) و y_n^P جواب خصوصی (۱۸۰۸) است. درحالت خاص، وقتی که $b_n=b$ یك عدد ثابت باشد، جواب خصوصی را می‌توان با قراردادن $y_n^P=A$ (یك عدد ثابت) درمعادلهٔ (۱۸۰۸) به‌آسانی به‌دست آورد. گذاردن $y_n=A$ درمعادلهٔ (۱۸۰۸)، به‌شرطی که مجموع ضرایب آن صفر نباشد، منجر به تعیین مقدار زیر خواهد شد

$$A=\frac{b}{۱+a_{N-۱}+\cdots+a_۰}$$

مثلا جواب عمومی معادلهٔ غیرهمگن

$$y_{n+۲}-۲y_{n+۱}+۲y_n=۱$$

به‌صورت زیراست

$$y_n=(\sqrt{۲})^n\left(C_۱\cos\frac{n\pi}{۴}+C_۲\sin\frac{n\pi}{۴}\right)+۱$$

خواص سادهٔ معادلات تفاضلی که تا اینجا بررسی شدند، برای استفاده در بقیهٔ این فصل کافی هستند.

□ **مثال:** نشان دهید که جواب عمومی معادلهٔ تفاضلی

(الف) $$y_{n+۲}-(۲+h^۲)y_{n+۱}+y_n=h^۲ \qquad h>۰$$

را می‌توان به‌شکل زیر بیان کرد

(ب) $$y_n=c_۱\left[۱+h+\frac{h^۲}{۲}+\mathcal{O}(h^۳)\right]^n+c_۲\left[۱-h+\frac{h^۲}{۲}+\mathcal{O}(h^۳)\right]^n-۱$$

حل:

۱. جواب خصوصی معادلهٔ (الف) ازقراردادن $y_n^P=C$ در (الف) به‌دست می‌آید که برابراست با $y_n^P=-۱$.

۲. معادلهٔ مشخصهٔ مربوط به‌معادلهٔ غیرهمگن (الف) به‌صورت زیر است

$$\beta^۲-(۲+h^۲)\beta+۱=۰$$

به‌موجب فرمول ریشه‌های معادلهٔ درجهٔ دوم، داریم

$$\beta_{۱,۲}=\frac{۲+h^۲\pm\sqrt{۴h^۲+h^۴}}{۲}$$

$$=\frac{۲+h^۲\pm ۲h\sqrt{۱+h^۲/۴}}{۲}$$

$$=۱+h^۲/۲\pm h\left(۱+\frac{h^۲}{۴}\right)^{۱/۲}$$

از بسط $(۱+t)^{۱/۲}$ به‌سری تیلر پیرامون $t=۰$ و قراردادن $h^۲/۴$ به‌جای t خواهیم داشت

$$\beta_{۱,۲}=۱+\frac{h^۲}{۲}\pm h\left[۱+\frac{h^۲}{۸}+\mathcal{O}(h^۴)\right]$$

$$\beta_۱=۱+h+\frac{h^۲}{۲}+\mathcal{O}(h^۳)$$

$$\beta_۲=۱-h+\frac{h^۲}{۲}+\mathcal{O}(h^۳)$$

بنابراین جواب عمومی دستگاه همگن به‌صورت زیراست

$$y_n^C=c_۱\beta_۱^n+c_۲\beta_۲^n$$

۳. درنتیجه جواب معادلهٔ (الف) عبارت‌است از

$$y_n=y_n^P+y_n^C$$

که بودن جواب به‌شکل (ب) را ثابت می‌کند. □

تمرین

۱-۲۰۸ جواب عمومی معادلات تفاضلی زیر را پیداکنید

(الف) $y_{n+۱}-۳y_n=۵$

(ب) $y_{n+۲}-۴y_{n+۱}+۴y_n=n$

(راهنمایی: برای پیداکردن جواب خصوصی، $y_n^P=an+b$ را آزمایش کنید).

(پ) $y_{n+۲}+۲y_{n+۱}+۲y_n=۰$

(ت) $y_{n+۳}-y_{n+۲}+۲y_{n+۱}-۲y_n=۰$

(ث) $y_{n+۲}-y_{n+۱}-y_n=۰$

۲-۲۰۸ جواب معادلات تفاضلی با مقدار اولیهٔ زیر را پیداکنید

(الف) $y_{n+۲}-۴y_{n+۱}+۳y_n=۲^n \quad y_۰=۰ \quad y_۱=۱$

(ب) $y_{n+۲}-y_{n+۱}-y_n=۰ \quad y_۰=۰ \quad y_۱=۱$

[(راهنمایی: برای پیدا کردن یك جواب خصوصی معادلهٔ (الف)، $y_n^P=A۲^n$ را امتحان کنید)].

۳-۲۰۸ نشان دهیدکه جواب عمومی معادلهٔ تفاضلی

$$y_{n+2}+4hy_{n+1}-y_n=2h$$

راکه در آن h عدد ثابت و مثبتی‌است، می‌توان به‌شکل زیر نوشت

$$y_n=c_1[1-2h+\mathcal{O}(h^2)]^n+c_2(-1)^n[1+2h+\mathcal{O}(h^2)]^n+\frac{1}{2}$$

۴-۲۰۸ نشان دهیدکه اگر $y_0=1$، $y_1=x$، آنگاه جملهٔ nام، $y_n=y_n(x)$، ازجواب معادلهٔ

$$y_{n+2}-2xy_{n+1}+y_n=0$$

یك بسجمله‌ای است از درجهٔ n برحسب x کـه ضریب پیشرو آن 2^{n-1}است. [توجه: $y_n(x)$ها بسجمله‌ایهای چبیشف هستندکه در بخش ۱.۶ دیدیم.]

۳.۸ انتگرالگیری عددی به‌وسیلهٔ سری تیلر

اکنون آماده‌ایم که روشهای عددی انتگرالگیری معادلات دیفرانسیل را بررسی کنیم. ابتدا یك معادلهٔ دیفرانسیل مرتبهٔ اول با مقدار اولیه به‌شکل

$$y'=f(x,y) \qquad y(x_0)=y_0 \tag{۱۹.۸}$$

را درنظر می‌گیریم. تابع f ممکن‌است خطی یا غیرخطی باشد، اما فرض بر آن‌است کـه f به‌اندازهٔ‌کافی هم نسبت به x و هم نسبت به y مشتقپذیراست. می‌دانیم که اگر $\partial f/\partial y$ درحوزهٔ موردنظر پیوسته باشد، معادلهٔ(۱۹.۸) دارای جواب یگانه‌است. اگر $y(x)$جواب دقیق (۱۹.۸) باشد، می‌توانیم $y(x)$ را پیرامون نقطهٔ $x=x_0$ به‌سری تیلر بسط دهیم:

$$y(x)=y_0+(x-x_0)y'(x_0)+\frac{(x-x_0)^2}{2!}y''(x_0)+\cdots \tag{۲۰.۸}$$

روشن‌است که دراین‌بسط مشتقها معلوم‌نیستند زیراکه‌جواب نامعلوم‌است. ولی‌اگر f به‌اندازهٔ کافی‌مشتقپذیرباشد، می‌توان مشتقات مذکور را با گرفتن‌مشتق کلی از معادلهٔ (۱۹.۸) نسبت به x، با توجه به‌اینکه y خود تابعی از x است، به‌دست آورد (به‌بخش ۷.۱ نگاه کنید). بنابراین برای چند مشتق اولیه داریم

$$\begin{aligned}
y'&=f(x,y)\\
y''&=f'=f_x+f_yy'=f_x+f_yf\\
y'''&=f''=f_{xx}+f_{xy}f+f_{yx}f+f_{yy}f^2+f_yf_x+f_y^2f\\
&=f_{xx}+2f_{xy}f+f_{yy}f^2+f_xf_y+f_y^2f
\end{aligned} \tag{۲۱.۸}$$

به‌همین ترتیب، می‌توانیم هر مشتقی از y را برحسب $f(x, y)$ و مشتقهای جزئی آن بیان کنیم. اما کاملا آشکار است که مادامی که $f(x, y)$ تابع خیلی ساده‌ای نباشد، مشتقهای کلی از درجات بالاتر بسیار پیچیده خواهند شد. لذا به‌دلایل عملی، می‌باید تعداد جملات را در بسط (۲۰.۸) به‌تعداد مناسبی محدود سازیم، که این محدودیت به‌محدودیت مقدار x در رابطهٔ (۲۰.۸) منجر می‌شود که می‌تواند تقریب موجهی برای (۲۰.۸) باشد. اگر فرض کنیم که سری ناتمام[1] (۲۰.۸) تقریب خوبی برای نمو h، یعنی به ازای $x - x_0 = h$، به‌دست می‌دهد، آنگاه می‌توانیم y را در $x_0 + h$ محاسبه و مجدداً مقادیر y' و y'' وغیره را در $x = x_0 + h$ حساب کنیم و سپس (۲۰.۸) را برای مرحلهٔ بعدی به‌کار بریم. اگر عمل را به‌همین طریق ادامه دهیم، مجموعهٔ گسسته‌ای از مقادیر y_n خواهیم داشت که تقریبهایی برای جواب حقیقی در نقاط $x_n = x_0 + nh (n = 0, 1, 2, \dots)$ هستند. در این فصل همه‌جا مقدار جواب دقیق در نقطهٔ x_n را با $y(x_n)$ و جواب تقریبی را با y_n نمایش می‌دهیم.

برای آنکه روش فوق را به‌صورت دستوری بیان کنیم، ابتدا عملگر[2] زیر را وارد می‌نماییم

$$T_k(x, y) = f(x, y) + \frac{h}{2!} f'(x, y) + \dots + \frac{h^{k-1}}{k!} f^{(k-1)}(x, y)$$

$$k = 1, 2, \dots \qquad (۲۲.۸)$$

که در آن فرض بر آن است که از یك نمو ثابت h استفاده شده و $f^{(j)}$ معرف مشتق کلی مرتبهٔ j ام تابع $f(x, y(x))$ نسبت به x است. حال می‌توانیم الگوریتم ۱.۸ را بیان کنیم.

الگوریتم ۱.۸: الگوریتم تیلر از مرتبهٔ k برای پیدا کردن یك جواب تقریبی معادلهٔ دیفرانسیل

$$y' = f(x, y)$$

$$y(a) = y_0$$

بر یك بازهٔ $[a, b]$:

۱. نمو $h = (b - a)/N$ را انتخاب وفرض می‌کنیم

$$x_n = a + nh \qquad n = 0, 1, \dots, N$$

۲. تقریبهای y_n را برای $y(x_n)$ با استفاده از فرمول بازگشتی

$$y_{n+1} = y_n + hT_k(x_n, y_n) \qquad n = 0, 1, \dots, N - 1$$

---

1. truncated 2. operator

که در آن $T_k(x, y)$ به‌وسیلهٔ رابطهٔ (۲۲.۸) تعریف شده است، پدید می‌آوریم.

غالباً الگوریتم تیلر و روشهای دیگری را که براساس این الگوریتم مبتنی هستند ودر آنها برای محاسبهٔ y در نقطهٔ $x = x_{n+1}$ تنها از اطلاعات مربوط به y و y' در نقطهٔ $x = x_n$ استفاده می‌کنند، روشهای **یك مرحله‌ای**[1] نامند.

قضیهٔ تیلر باباقیماندهٔ(قضیهٔ ۷.۱)،نشان می‌دهد که **خطای موضعی*** الگوریتم تیلر از مرتبهٔ k به‌شرح زیراست

$$E = \frac{h^{k+1} f^{(k)}(\xi, y(\xi))}{(k+1)!} \qquad x_n < \xi < x_n + h$$

$$= \frac{h^{k+1}}{(k+1)!} y^{(k+1)}(\xi)$$

الگوریتم تیلر را از مرتبهٔ k نامند اگر خطای موضعی E،که در بالا تعریف شد، برابر با $\theta(h^{k+1})$ باشد.

با قراردادن $k = 1$ در الگوریتم ۱.۸، **روش اویلر** و خطای موضعی آن به‌دست می‌آید

$$\boxed{y_{n+1} = y_n + hf(x_n, y_n)} \qquad \boxed{E = \frac{h^2}{2} y''(\xi)} \qquad (۲۳.۸)$$

برای روشن‌ساختن روش اویلر، مسئلهٔ با مقدار اولیهٔ

$$y' = y \qquad y(0) = 1$$

را درنظر می‌گیریم. با به‌کارگرفتن فرمول (۲۳.۸) به‌ازای $h = 0.01$ و حفظ شش رقم اعشار، خواهیم داشت

$$y(0.01) \approx y_1 = 1 + 0.01 = 1.01$$

$$y(0.02) \approx y_2 = 1.01 + 0.01(1.01) = 1.020201$$

$$y(0.03) \approx y_3 = 1.020201 + 0.01(1.020201) = 1.030301$$

$$y(0.04) \approx y_4 = 1.030301 + 0.01(1.030301) = 1.040606$$

---

1. one-step

* خطای موضعی، خطایی است که در هر مرحله به‌ازای $n = 0, 1, 2, \ldots$ وجود دارد. ـم.

از آنجایی که $y = e^x$ جواب دقیق این معادله است، مقدار صحیح آن در $x = ۰٫۰۴$ برابر با ۱٫۰۴۰۸ خواهد شد. واضح است که برای به‌دست آوردن دقت بیشتر در روش اویلر، باید مقدار h را به‌طور قابل‌ملاحظه‌ای کوچك بگیریم.

اگر بگیریم $h = ۰٫۰۰۵$، خواهیم داشت

$$y(۰٫۰۰۵) \approx y_۱ = ۱٫۰۰۵۰$$

$$y(۰٫۰۱۰) \approx y_۲ = ۱٫۰۱۰۰$$

$$y(۰٫۰۱۵) \approx y_۳ = ۱٫۰۱۵۱$$

$$y(۰٫۰۲۰) \approx y_۴ = ۱٫۰۲۰۲$$

$$y(۰٫۰۲۵) \approx y_۵ = ۱٫۰۲۵۳$$

$$y(۰٫۰۳۰) \approx y_۶ = ۱٫۰۳۰۴$$

$$y(۰٫۰۳۵) \approx y_۷ = ۱٫۰۳۵۶$$

$$y(۰٫۰۴۰) \approx y_۸ = ۱٫۰۴۰۸$$

این نتایج تا چهار رقم اعشار صحیح‌اند.

از آنجا که h می‌باید نسبتاً کوچك باشد، معمولا برای انتگرالگیری معادلات دیفرانسیل از روش اویلر استفاده نمی‌شود.

البته برای به‌دست آوردن دقت بیشتر می‌توانستیم الگوریتم تیلر از مرتبهٔ بالاتر را به‌کار بریم و درحالت کلی باید انتظار داشته باشیم که برای یك نمو معین، هرقدر مرتبهٔ الگوریتم بالاتر باشد، به‌همان اندازه دقت بیشتری به‌دست آید. اگر $f(x, y)$ تابع نسبتاً ساده‌ای از x و y باشد، می‌توان مشتقهای لازم را غالباً با هزینهٔ نسبتاً کمتری در کامپیوتر به‌دست آورد. این عمل با به‌کاربردن مشتقگیری نمادی* یا با بهره‌گیری از هریك از ویژگیهای تابع $f(x, y)$ که ممکن باشد، می‌تواند انجام پذیرد (به‌تمرین ۳۰۸-۴ نگاه کنید). ولی لزوم محاسبهٔ مشتقهای مرتبهٔ بالاتر، الگوریتم تیلر را برای انتگرالگیری کلی درکامپیوترهای سریع کاملا نامناسب می‌سازد. مع‌هذا، الگوریتم تیلر از لحاظ نظری مورد توجه زیاداست زیرا در بسیاری از روشهای عملی سعی براین‌است که دقت عمل مشابهی مانند الگوریتم تیلر به‌دست آورند، بی‌آنکه با وضع نامساعد محاسبهٔ مشتقهای مرتبهٔ بالا مواجه باشند. اگرچه الگوریتم کلی تیلر به‌ندرت برای مقاصد عملی به‌کار برده می‌شود، ولی حالت خاص آن یعنی روش اویلر، با شرح بیشتری برای نشان دادن اهمیت نظری آن موردبررسی قرارمی‌گیرد.

□ **مثال ۱۰۸:** با استفاده از سری تیلر جواب معادلهٔ دیفرانسیل

---

* مشتقگیری نمادی به‌معنای مشتقگیری با استفاده از روشهای غیرعددی‌است.ـــم.

$$xy' = x - y \qquad y(۲) = ۲$$

را در $x = ۲ر۱$ تا پنج رقم‌اعشاری دقیق به‌دست آورید.
چند مشتق اول و مقادیر آنها در $x = ۲$ و $y = ۲$ به‌شرح زیرند:

$$y' = ۱ - \frac{y}{x} \qquad y'_۰ = ۰$$

$$y'' = \frac{-y'}{x} + \frac{y}{x^۲} \qquad y''_۰ = \frac{۱}{۲}$$

$$y''' = \frac{-y''}{x} + \frac{۲y'}{x^۲} - \frac{۲y}{x^۳} \qquad y'''_۰ = -\frac{۳}{۴}$$

$$y^{iv} = \frac{-y'''}{x} + \frac{۳y''}{x^۲} - \frac{۶y'}{x^۳} + \frac{۶y}{x^۴} \qquad y^{iv}_۰ = \frac{۳}{۲}$$

بسط سری تیلر پیرامون نقطهٔ $x_۰ = ۲$ چنین‌است:

$$y(x) = y_۰ + (x-۲)y'_۰ + \frac{۱}{۲}(x-۲)^۲ y''_۰ + \frac{۱}{۶}(x-۲)^۳ y'''_۰$$

$$+ \frac{۱}{۲۴}(x-۲)^۴ y^{iv}_۰ + \cdots$$

$$= ۲ + (x-۲)۰ + \frac{۱}{۴}(x-۲)^۲ - \frac{۱}{۸}(x-۲)^۳ + \frac{۱}{۱۶}(x-۲)^۴ + \cdots$$

در $x = ۲ر۱$ خواهیم داشت

$$y(۲ر۱) = ۲ + ۰ر۰۰۲۵ - ۰ر۰۰۰۱۲۵ + ۰ر۰۰۰۰۰۶۲ - \cdots$$

$$\approx ۲ر۰۰۲۳۸$$

چون جمله‌ها دراین سری تیلر ازنظرکمیت‌کاهش می‌یابند و از لحاظ علامت متناوباً تغییر می‌کنند (به‌تمرین۸.۳۰–۴ نگاه‌کنید)، این نتیجه تا پنج رقم اعشاری صحیح است. اکنون اگربخواهیم $y(۲ر۲)$ را با همان دقت به‌دست آوریم، می‌باید دوجملهٔ دیگر ازسری‌فوق را نیز به‌دست آوریم. یا اینکه می‌توانیم بسط جدیدی پیرامون $x = ۲ر۱$ بدهیم و چهار مشتق اول را مجدداً در $x = ۲ر۱$ تعیین و سپس $y(۲ر۲)$ را محاسبه‌کنیم. □

□ **مثال۸.۲۰:** معادلهٔ

$$y' = \frac{1}{x^2} - \frac{y}{x} - y^2$$

$$y(1) = -1$$

را از $x = 1$ تا $x = 2$ حل کنید. از الگوریتم مرتبهٔ ۲ی تیلر استفاده می‌کنیم. مسئله را به‌ازای $h = 1/16, 1/32, 1/64, 1/128$ حل و دقت نتایج را براورد می‌کنیم

حل: چون

$$f(x, y) = \frac{1}{x^2} - \frac{y}{x} - y^2$$

$$f'(x, y) = -\frac{2}{x^3} - \frac{y'}{x} + \frac{y}{x^2} - 2yy'$$

لذا

$$T_2(x, y) = f + \frac{h}{2} f'$$

و

$$y_{n+1} = y_n + h\left[f(x_n, y_n) + \frac{h}{2} f'(x_n, y_n) \right]$$

نتایج به‌دست‌آمده با کامپیوتر IBM ۷۰۹۰ در زیر داده شده‌اند. نمو h درستون اول و مقادیر $y'(2{,}0)$, $y(2{,}0)$, $y'(1{,}5)$ و $y(1{,}5)$ به‌ترتیب در چهارستون بعدی داده شده‌اند. جواب دقیق این معادله $y = -1/x$ است، لذا مقدار دقیق $y(1{,}5)$ برابر $-2/3$ و مقدار دقیق $y(2{,}0)$ برابر $-1/2$ می‌شود، می‌توانیم خطای کلی ناشی از گسسته‌سازی را به‌طریق زیر براورد کنیم: خطای موضعی الگوریتم مرتبهٔ ۲ی تیلر برابر $y'''(h^3/6)$ است. چون داریم $y''' = 6/x^4$، مقدار ماکزیمم آن در بازهٔ $[1, 2]$ برابر ۶ است و بنابراین در هر مرحله، خطای موضعی حداکثر برابر با h^3 است. به‌ازای $h = 1/128$، ۱۲۸ مرحله انتگرالگیری داریم به‌طوری که خطای انباشته شده، حداکثر برابر با $128h^3 = (1/128)^2 \approx 0{,}000006$ است. درتوافق نزدیک با این مقدار براورد، ظاهراً خطای حقیقی در $x = 2{,}0$، برابر با $0{,}000003$ است. در حالت کلی، جواب واقعی را دردست نداریم تا صحت آن را آزمایش کنیم. حتی با نداشتن جواب، به‌ازای $h \to 0$ می‌توانیم دقت جواب را از روی تعداد ارقام موافق براورد کنیم. از آنجایی که با هربار نصف کردن h، تقریباً یك رقم دیگر به‌دقت افزوده می‌شود، ظاهراً در نبود خطای ناشی از گردکردن، نمو $1/1024$ می‌باید حداقل هفت رقم دقیق تولیدنماید.

نتایج کامپیوتری برای مثال ۲.۸

روش ۱ـ روش مرتبهٔ دوم بسط تیلر

H	Y(1.5)	YPRM(1.5)	Y(2.)	YPRM(2.)
0.62500000E-01	−0.66787238E 00	0.44363917E-00	−0.50187737E 00	0.24905779E-00
0.31250000E-01	−0.66696430E 00	0.44424593E-00	−0.50046334E 00	0.24976812E-00
0.15625000E-01	−0.66674034E 00	0.44439532E-00	−0.50011456E 00	0.24994271E-00
0.78125000E-02	−0.66668454E 00	0.44443253E-00	−0.50002744E 00	0.24998628E-00

روش ۲ـ روش ساده شدهٔ مرتبهٔ دوم رونگه‌ـ کوتا

H	Y(1.5)	YPRM(1.5)	Y(2.)	YPRM(2.)
0.62500000E-01	−0.66552725E 00	0.44520275E-00	−0.49822412E-00	0.25088478E-00
0.31250000E-01	−0.66637699E 00	0.44463748E-00	−0.49954852E-00	0.25022554E-00
0.15625000E-01	−0.66659356E 00	0.44449317E-00	−0.49988601E-00	0.25005698E-00
0.78125000E-02	−0.66664808E 00	0.44445683E-00	−0.49997083E-00	0.25001458E-00

روش ۳ـ روش کلاسیك مرتبهٔ چهارم رونگه‌ـ کوتا

H	Y(1.5)	YPRM(1.5)	Y(2.)	YPRM(2.)
0.62500000E-01	−0.66666625E 00	0.44444472E-00	−0.49999941E-00	0.25000029E-00
0.31250000E-01	−0.66666664E 00	0.44444446E-00	−0.49999997E-00	0.25000001E-00
0.15625000E-01	−0.66666666E 00	0.44444444E-00	−0.50000000E 00	0.25000000E-00
0.78125000E-02	−0.66666667E 00	0.44444444E-00	−0.50000001E 00	0.24999999E-00

همین مسئله بعداً با استفاده از دوروش دیگر حل خواهد شد. به‌منظورمقایسه، نتایج به‌دست آمده از هرسه روش دراینجا داده شده‌اند. □

تمرین

۱-۳.۸ برای معادلهٔ

$$y' = -xy + \frac{1}{y^2} \qquad y(1) = 1$$

معادلهٔ تفاضلی متناظر با الگوریتم مرتبهٔ ۳ی تیلر را به‌دست آورید. یك مرحله از انتگرالگیری را به‌ازای $h = 0.1$ با دست انجام دهید. برنامه‌ای برای حل این مسئله بنویسید و انتگرالگیری را از $x = 1$ تا $x = 2$ به‌ازای $h = 1/64$ و $h = 1/128$ انجام دهید.

۲-۳.۸ برای معادلهٔ

$$y' = 2y \qquad y(0) = 1$$

جواب دقیق معادلهٔ تفاضلی حاصل از روش اویلر را تعیین کنید. مقدار «به‌اندازهٔ کافی کوچك» h راچنان برآورد کنیدکه این مقدار، جوابی با دقت چهار رقم رادر بازهٔ $[0, 1]$ تضمین کند. به‌ازای مقدار مناسبی از h، حل این مسئله را برای ۱۰ مرحله انجام دهید.

۳-۸-۸ اگر $y(x)$ در رابطهٔ زیرصدق کند، از سری تیلرمربوط به $y(x)$، مقدار $y(0.1)$ را تا شش رقم اعشاری دقیق به‌دست آورید

$$y' = xy + 1 \qquad y(0) = 1$$

۴-۳-۸ ثابت کنیدکه برای تابع $f(x, y) = 1 - y/x$ ازمثال ۱.۸، $y'' = (1 - 2y')/x$، و $y^{(k)} = -ky^{(k-1)}/x$، $k = 3, 4, \ldots$. براین اساس با استفاده از الگوریتم ۱.۸، یك برنامهٔ فورترن بنویسیدکه مقدار $y(3)$ را تا دقت 10^{-6} از روی جواب $y(x)$ در مثال ۱.۸ به‌دست دهد.

۴.۸ براوردهای خطا و همگرایی روش اویلر

برای حل معادلهٔ دیفرانسیل $y' = f(x, y)$، $y(x_0) = y_0$ با روش اویلر، نموثابت h را انتخاب و از فرمول

$$y_{n+1} = y_n + hf(x_n, y_n) \qquad n = 0, 1, \ldots \qquad (۲۳.۸)$$

که در آن $x_n = x_0 + nh$، استفاده می‌کنیم. جواب دقیق معادلهٔ دیفرانسیل در $x = x_n$ را با $y(x_n)$ و جواب تقریبی حاصل ازکاربرد فرمول (۲۳.۸) را با y_n مشخص می‌سازیم. می‌خواهیم مقدار خطای **گسسته‌سازی** e_n راکه به‌صورت

$$e_n = y(x_n) - y_n \qquad (۲۴.۸)$$

تعریف می‌شود براورد کنیم. یادآوری می‌کنیم که اگر y_0، همان‌گونه‌که فرض خواهیم کرد، دقیق باشد آنگاه مقدار e_0 مساوی ۰ خواهد شد. با فرض وجود داشتن مشتقهای مناسب، می‌توانیم با استفاده از«قضیهٔ تیلر با باقیمانده»، $y(x_{n+1})$ را پیرامون $x = x_n$ بسط دهیم

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2} y''(\xi_n) \qquad x_n \leqslant \xi_n \leqslant x_{n+1} \qquad (۲۵.۸)$$

کمیت $(h^2/2)y''(\xi_n)$ را خطای **گسسته‌سازی موضعی**[1] نامند، یعنی خطایی‌که تنها در یك مرحلهٔ عبور از x_n به x_{n+1}، با فرض معلوم بودن y و y' در $x = x_n$ به‌طوردقیق، حاصل می‌شود. در یك کامپیوتر نیزهنگام محاسبهٔ y_{n+1} با استفاده ازفرمول (۲۳.۸)، خطای ناشی از گردکردن وجود دارد. اما این نوع خطاها دراین بخش نادیده گرفته خواهند شد.

---

1. local discretization

از کسر کردن (۲۳.۸) از (۲۵.۸) و استفاده از (۲۴.۸) خواهیم داشت

$$e_{n+1}=e_n+h[f(x_n, y(x_n))-f(x_n, y_n)]+\frac{h^2}{2}y''(\xi_n) \qquad (۲۶.۸)$$

بنا بر قضیهٔ مقدار میانگین حساب دیفرانسیل، داریم

$$f(x_n, y(x_n))-f(x_n, y_n)=f_y(x_n, \bar{y}_n)(y(x_n)-y_n)$$
$$=f_y(x_n, \bar{y}_n)e_n$$

که در آن $\bar{y}_n$ بین y_n و $y(x_n)$ واقع است. بنابراین رابطهٔ (۲۶.۸) به‌صورت زیر درخواهد آمد

$$e_{n+1}=e_n+hf_y(x_n, \bar{y}_n)e_n+\frac{h^2}{2}y''(\xi_n) \qquad (۲۷.۸)$$

اکنون فرض می کنیم که در بازهٔ مورد نظر داریم

$$|f_y(x, y)|<L \qquad |y''(x)|<Y$$

در اینجا L و Y اعداد ثابت و مثبتی هستند. با قدر مطلق گیری از رابطهٔ (۲۷.۸) خواهیم داشت

$$|e_{n+1}|\leqslant|e_n|+hL|e_n|+\frac{h^2}{2}Y=(1+hL)|e_n|+\frac{h^2}{2}Y \qquad (۲۸.۸)$$

اکنون با روش استقرا نشان خواهیم داد که جواب معادلهٔ تفاضلی

$$\xi_{n+1}=(1+hL)\xi_n+\frac{h^2}{2}Y \qquad (۲۹.۸)$$

به‌ازای $\xi_0=0$ نافذ برجواب (۲۷.۸) است، یعنی نشان خواهیم داد که

$$\xi_n\geqslant|e_n| \qquad n=0, 1, \ldots \qquad (۳۰.۸)$$

از آنجایی که $e_0=\xi_0=0$، رابطهٔ (۳۰.۸) مطمئناً به‌ازای $n=0$ برقرار است. به‌فرض اینکه رابطهٔ (۳۰.۸) به‌ازای عدد صحیح n برقرار باشد، چون $\xi_n\geqslant|e_n|$ و $1+hL>1$، آنگاه از رابطهٔ (۲۹.۸) نتیجه می‌شود که

$$\xi_{n+1}\geqslant|e_{n+1}| \qquad n=0, 1, \ldots$$

که استقرای ما را کامل می کند.

بنابراین جواب ξ_n از معادلهٔ تفاضلی غیرهمگن (۲۹.۸)، کران بالایی برای خطای

گسسته‌سازی e_n به‌دست می‌دهد. از نظریهٔ مربوط به‌معادلهٔ تفاضلی که در بخش ۲۰۸ بیان شد، جواب معادلهٔ (۲۹.۸) به‌شرح زیر به‌دست می‌آید

$$\xi_n = c(1+hL)^n - B \qquad (۳۱.۸)$$

که در آن c یك ثابت اختیاری است و

$$B = \frac{hY}{2L}$$

برای برقراری شرط $\xi_0 = 0$، ملاحظه می‌کنیم که باید $c = +B$ انتخاب شود، که با این انتخاب رابطهٔ (۳۱.۸)، به‌صورت زیر درخواهد آمد

$$\xi_n = B(1+hL)^n - B$$

با توجه به بخش ۷۰۱ نتیجه می‌گیریم که $e^x = 1 + x + e^{\xi} x^2/2$؛ از این‌رو به‌ازای کلیهٔ مقادیر x رابطهٔ $e^x \geqslant 1 + x$ برقرار است، که از این موضوع نتیجه می‌شود که $1 + hL \leqslant e^{hL}$، و بنابراین $(1+hL)^n \leqslant e^{nhL}$. اگر ازاین مطلب در رابطهٔ (۳۱.۸) استفاده کنیم می‌توانیم بگوییم که

$$\xi_n \leqslant B(e^{nhL} - 1)$$

$$= \frac{hY}{2L}(e^{nhL} - 1)$$

$$= \frac{hY}{2L}(e^{(x_n - x_0)L} - 1)$$

که در آنها از تساوی $nh = x_n - x_0$، استفاده کرده‌ایم. چون $|e_n| \leqslant \xi_n$، لذا قضیهٔ زیر را اثبات نموده‌ایم.

قضیهٔ ۲۰۸ گیریم y_n جواب تقریبی معادلهٔ (۱۹.۸) باشد که به‌وسیلهٔ روش اویلر (۲۳.۸) به‌دست آمده‌است. اگر جواب دقیق $y(x)$ از معادلهٔ (۱۹.۸) دارای مشتق دوم پیوسته در بازهٔ $[x_0, b]$ باشد، و اگر دراین بازه نامساویهای

$$|f_y(x, y)| \leqslant L \qquad |y''(x)| < Y$$

به‌ازای مقادیر ثابت مثبت L و Y برقرار باشند، خطای $e_n = y(x_n) - y_n$ ناشی از روش اویلر در نقطهٔ $x_n = x_0 + nh$ به‌صورت زیر کراندار می‌شود

$$|e_n| \leqslant \frac{hY}{2L}(e^{(x_n - x_0)L} - 1) \qquad (۳۲.۸)$$

این قضیه نشان می‌دهد که خطا برابر با $\mathcal{O}(h)$ است؛ یعنی به‌ازای $h \to 0$، خطا به‌سمت صفر میل می‌کند، همانند ch که به‌ازای مقدار ثابتی مانند c وقتی $x = x_n$ ثابت نگه‌داشته شود به‌سمت صفر میل می‌کند. می‌باید تأکید شود که براورد (۳۲.۸)، به‌جای کران واقعی یك کران بالایی به‌دست می‌دهد. اهمیت اساسی این رابطه بیشتر در اثبات همگرایی این روش است، نه در فراهم کردن یك براورد واقعی از خطایی که از قبل تعیین شده است.

□ **مثال ۳۰۸:** کران بالایی را برای خطای گسسته‌سازی روش اویلر در حل معادلهٔ $y' = y$، $y(0) = 1$ از $x = 0$ تا $x = 1$ تعیین کنید.

حل: در اینجا داریم $f(x, y) = y$ و $\partial f / \partial y = 1$. بنابراین می‌توانیم L را برابر با یك بگیریم. وهمچنین چون $y = e^x$، بنابراین $y'' = e^x$، و به‌ازای $0 \leqslant x \leqslant 1$ داریم، $|y''(x)| \leqslant e$. برای پیدا کردن کرانی برای خطا در $x = 1$ داریم $x_n - x_0 = 1$ و $y = e^1$ و از رابطهٔ (۳۲.۸) به‌دست می‌آوریم

$$|e(1)| \leqslant \frac{he}{2}(e-1)$$

$$< 2.4\, h$$

بنابراین خطای $e(1)$ در $x = 1$ با h ۲٫۴ کراندار شده است. برای اینکه ببینیم این کران چقدر به‌مقدار واقعی نزدیك است، از روش اویلر جواب دقیقی برای این مسئله به‌دست می‌آوریم. لذا

$$y_{n+1} = y_n + hf(x_n, y_n)$$

$$= (1+h)y_n$$

جواب این معادلهٔ تفاضلی که $y(0) = 1$ در آن صدق می‌کند به‌صورت زیر است

$$y_n = (1+h)^n$$

حال اگر $h = 0.1$ و $n = 10$ اختیار شوند با بسط $(1.1)^{10}$ درمی‌یابیم که روش اویلر مقدار $y_{10} \approx y(1) = 2.59374$ را به‌دست می‌دهد. از کم کردن این مقدار، از جواب دقیق $y(1) = e = 2.718282$ درمی‌یابیم که درمقایسه با کران ۰٫۲۴ که با استفاده از رابطهٔ (۳۲.۸) به‌دست آمده، مقدار خطا برابر با ۰٫۱۲۴۶ است. □

تمرین

۱-۴۰۸ درمعادلهٔ $y' = -2y$، $y(0) = 1$ و $0 \leqslant x \leqslant 1$

(الف) کران بالایی برای خطا برحسب نمو h با استفاده از رابطهٔ (۳۲.۸) در $x=1$ پیدا کنید.

(ب) معادلهٔ تفاضلیی را که از روش اویلر نتیجه می‌شود حل کنید.

(پ) کران به‌دست‌آمده از (الف) را با خطای حقیقی به‌دست‌آمده از (ب) به‌ازای $h=0.1$ و $h=0.05$، در $x=1$ مقایسه کنید.

(ت) با به‌کار گرفتن روش اویلر، نمو h به‌چه اندازه می‌باید کوچك باشد تا شش رقم دقت در $x=1$ به‌دست آید (فرض کنید که خطای گرد کردن وجود ندارد).

۲-۴.۸ خطای e_n از یك روش انتگرالگیری در تامساوی تفاضلی

$$|e_{n+2}| \leqslant a_1|e_{n+1}|+a_2|e_n|+A$$

که در آن a_1، a_2 و A ثابتهای مثبت هستند و $e_1=e_0=0$، صدق می‌کند. گیریم ξ_n یك جواب معادلهٔ تفاضلی

$$\xi_{n+2}=a_1\xi_{n+1}+a_2\xi_n+A$$

باشد و $\xi_1=\xi_0=0$. به‌استقرا نشان دهید که

$$|e_n| \leqslant \xi_n \qquad \text{به‌ازای جمیع مقادیر } n$$

۵.۸ روشهای رونگه ـ کوتا[1]

چنانکه قبلا ذکر شد، روش اویلر در مسائل عملی چندان مفید نیست، زیرا برای دستیابی به‌یك دقت مناسب به‌نمو بسیار کوچکی نیاز دارد. الگوریتم مرتبهٔ بالاتر تیلر نیز به‌عنوان شیوه‌ای با هدف کلی غیرقابل‌قبول است، زیرا به‌مشتقهای کلی از درجهٔ بالاتر $y(x)$ احتیاج دارد. در روش رونگه ـ کوتا سعی بر این است که به‌دقت بیشتری دست یافته شود و همزمان با آن از محاسبهٔ مشتقهای مرتبهٔ بالاتر اجتناب شود، و این کار با محاسبهٔ تابع $f(x, y)$ در نقاط انتخاب شده در هر زیربازه انجام می‌گیرد. در اینجا ساده‌ترین روش رونگه ـ کوتا را مورد بحث قرار می‌دهیم. ما در اینجا به‌دنبال فرمولی به‌شکل زیر هستیم

$$y_{n+1}=y_n+ak_1+bk_2 \qquad (۳۳.۸)$$

که در آن

$$k_1=hf(x_n, y_n)$$

$$k_2=hf(x_n+\alpha h, y_n+\beta k_1)$$

---

1. Runge–Kutta methods

و a، b، α، β ثابتهایی هستند که باید چنان تعیین شوند که فرمول (۳۳.۸) با الگوریتم تیلر از مرتبهٔ هرچه بزرگتر توافق داشته باشد. با بسط $y(x_{n+1})$ به‌سری تیلر تا جملهٔ h^3 خواهیم داشت

$$y(x_{n+1})=y(x_n)+hy'(x_n)+\frac{h^2}{2}y''(x_n)+\frac{h^3}{6}y'''(x_n)+\cdots$$

$$=y(x_n)+hf(x_n,y_n)+\frac{h^2}{2}(f_x+ff_y)_n$$

$$+\frac{h^3}{6}(f_{xx}+2ff_{xy}+f_{yy}f^2+f_xf_y+f_y^2f)_n+\mathcal{O}(h^4)$$

(۳۴.۸)

که در آن از بسط (۲۱.۸) استفاده کرده‌ایم، ولی زیرنمایهٔ n به‌معنای آن است که کلیهٔ توابع موردنظر باید در $\{x_n, y_n\}$ ارزیابی شوند.

از طرف دیگر با استفاده از بسط تیلر برای توابع دومتغیره (به‌بخش ۷.۱ نگاه کنید)، خواهیم داشت

$$\frac{k_2}{h}=f(x_n+\alpha h, y_n+\beta k_1)=f(x_n,y_n)+\alpha hf_x+\beta k_1f_y$$

$$+\frac{\alpha^2h^2}{2}f_{xx}+\alpha h\beta k_1f_{xy}+\frac{\beta^2k_1^2}{2}f_{yy}+\mathcal{O}(h^3)$$

که در آن کلیهٔ مشتقها در $\{x_n, y_n\}$ ارزیابی شده‌اند.

اکنون اگر این عبارت را در فرمول (۳۳.۸) به‌جای k_2 بگذاریم و توجه داشته باشیم که $k_1=hf(x_n,y_n)$، آنگاه با مرتب کردن نتیجهٔ حاصله برحسب توانهای h خواهیم داشت

$$y_{n+1}=y_n+(a+b)hf+bh^2(\alpha f_x+\beta ff_y)$$

$$+bh^3\left(\frac{\alpha^2}{2}f_{xx}+\alpha\beta ff_{xy}+\frac{\beta^2}{2}f^2f_{yy}\right)+\mathcal{O}(h^4) \quad \text{(۳۴.۸ الف)}$$

از مقایسهٔ این رابطه با (۳۴.۸)، ملاحظه می‌کنیم که برای اینکه توانهای h و h^2 در دو رابطه با هم سازگار باشند، باید داشته باشیم

$$a+b=1$$

$$b\alpha=b\beta=\frac{1}{2} \qquad (۳۵.۸)$$

اگرچه چهارمجهول داریم، ولی تنها سه معادله در دست داریم، بنابراین هنوز یك درجهٔ آزادی درجواب (۳۵۰۸) داریم. می‌توانیم این درجهٔ آزادی را برای به‌دست آوردن هماهنگی درضرایب جملات h^3 به‌کاربریم. ولی آشکار است که‌این امربرای کلیهٔ توابع $f(x, y)$ شدنی نیست.

برای (۳۵۰۸) جوابهای زیادی وجود دارند،که شاید ساده‌ترین آنها به‌صورت‌زیر باشد

$$a=b=\frac{1}{2} \qquad \alpha=\beta=1$$

الگوریتم ۲۰۸: روش مرتبهٔ دوم رونگه‌ـکوتا برای معادلهٔ

$$y'=f(x, y) \qquad y(x_0)=y_0$$

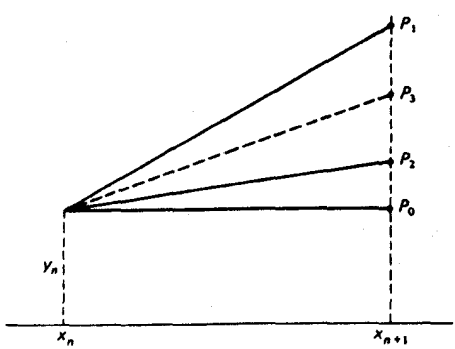
تقریبهای y_n را برای $y(x_0+nh)$ به‌ازای h ثابت و $n=0, 1, \ldots$، با استفاده از فرمولهای بازگشتی زیرپدیدآورید

$$\boxed{\begin{aligned} &y_{n+1}=y_n+\frac{1}{2}(k_1+k_2) \qquad k_1=hf(x_n, y_n) \quad \text{به‌ازای} \\ &\qquad\qquad\qquad\qquad\qquad\qquad k_2=hf(x_n+h, y_n+k_1) \end{aligned}} \qquad (۳۶۰۸)$$

الگوریتم ۲۰۸ را می‌توان از لحاظ هندسی به‌صورت شکل۱۰۸ تجسم نمود. روش اویلر برای y_n، نمو $P_1P_0=hf(x_n, y_n)$ را تولید می‌کند و

$$P_2P_0=hf(x_n+h, y_n+hf(x_n, y_n))$$

نمودیگری برپایهٔشیب به‌دست آمده در x_{n+1}است. با به‌دست آوردن‌مقدارمتوسط این‌دونمو،



شکل۱۰۸

فرمول (۳۶.۸) به‌دست خواهد آمد.

خطای موضعی فرمول (۳۶.۸) به‌شکل زیر است

$$y(x_{n+1})-y_{n+1}=\frac{h^3}{12}(f_{xx}+2ff_{xy}+f^2f_{yy}-2f_xf_y-2ff_y^2)+\theta(h^4)$$

پیچیدگی ضرایب در این عبارت خطا، مشخصهٔ کلیهٔ روشهای رونگه‌ـکوتا است، این امر یکی از جنبه‌های نامطلوب این روشها به شمار می‌آید، زیرا بر اورد خطای موضعی بسیار دشوار است. خطای موضعی فرمول (۳۶.۸) از مرتبهٔ h^3 است، درحالی که این خطا در روش اویلر از مرتبهٔ h^2 است. بنابراین می‌توانیم انتظار داشته باشیم که بتوانیم نمو بزرگتری را با (۳۶.۸) به‌کار بریم. زیانی که ما در این مسئله متحمل می‌شویم این است که تابع $f(x,y)$ را در هر مرحلهٔ انتگرالگیری دوبار محاسبه می‌کنیم. با روش فوق می‌توان فرمولهایی از نوع فرمول رونگه‌ـکوتا از هر مرتبه‌ای را به‌دست آورد. ولی، مشتقات حاصله در آنها بسیار پیچیده می‌شود. عادیترین و متداولترین فرمول از این نوع در قالب الگوریتم ۳.۸، آمده است.

الگوریتم ۳.۸: روش مرتبهٔ ۴ رونگه‌ـکوتا برای معادلهٔ $y'=f(x,y)$، $y(x_0)=y_0$، تقریبهای y_n برای $y(x_0+nh)$ را به‌ازای مقدار ثابت h، $n=0,1,2,\dots$، با استفاده از فرمول بازگشتی زیر تولید کنید

$$\boxed{y_{n+1}=y_n+\frac{1}{6}(k_1+2k_2+2k_3+k_4)} \qquad (۳۷.۸)$$

که در آن

$$k_1=hf(x_n,y_n)$$

$$k_2=hf\left(x_n+\frac{h}{2},\ y_n+\frac{1}{2}k_1\right)$$

$$k_3=hf\left(x_n+\frac{h}{2},\ y_n+\frac{1}{2}k_2\right)$$

$$k_4=hf(x_n+h,\ y_n+k_3)$$

خطای گسسته‌سازی موضعی در الگوریتم ۳.۸ برابر با $\theta(h^5)$ است. و باز ضرری که ما برای این امر متحمل می‌شویم آن است که در هر مرحله به چهار بار ارزیابی تابع نیاز پیدا می‌کنیم. در مسائلی که تابع $f(x,y)$ پیچیده است، این ضرر ممکن است از نظر وقت کامپیوتر قابل ملاحظه باشد. روشهای رونگه‌ـکوتا معایب دیگری نیز دارند که بعداً مورد بحث قرار خواهند گرفت. در عمل، فرمول (۳۷.۸) به‌نحوی گسترده با موفقیتی بسیار زیاد

به‌کار برده می‌شود. امتیاز مهم آن خود-آغازگری[1] آن است: یعنی برای یافتن y و y' در $x = x_n + 1$، تنها به‌مقدار y در نقطهٔ $x = x_n$ نیازدارد.

در زیر یك برنامهٔ فورترن باهدف کلی بر اساس الگوریتم ۲.۸ برای معادلهٔ دیفرانسیل داده شده‌است. برای استفاده ازاین برنامه، باید زیربرنامه‌ای برای محاسبهٔ تابع $f(x, y)$ به‌آن افزوده شود و مقدار اولیهٔ $y(x_0) = y_0$، نقطهٔ انتهایی x_{NSTEP}، و تعدادکل مراحل NSTEPS مشخص شوند.

برنامهٔ فورترن برای الگوریتم ۲.۸

```
C  FORTRAN PROGRAM TO SOLVE THE FIRST ORDER DIFFERENTIAL EQUATION
C        Y'(X) = F(X,Y)
C  WITH INITIAL CONDITION OF
C        Y(XBEGIN) = YBEGIN
C  TO THE POINT  XEND , USING THE SECOND ORDER RUNGE-KUTTA METHOD.
C  A FUNCTION SUBPROGRAM CALLED 'F' MUST BE SUPPLIED.
      INTEGER I,N,NSTEPS
      REAL DERIV,H,K1,K2,XBEGIN,XN,XEND,YBEGIN,YN
    1 READ 501, XBEGIN,YBEGIN,XEND,NSTEPS
  501 FORMAT(3F10.5,I3)
      IF (NSTEPS .LT. 1)                    STOP
      H = (XEND - XBEGIN)/NSTEPS
      XN = XBEGIN
      YN = YBEGIN
      DERIV = F(XN,YN)
      N = 0
      PRINT 601, N,XN,YN,DERIV
  601 FORMAT(1X,I3,3E21.9)
      DO 10 N=1,NSTEPS
         K1 = H*F(XN,YN)
         K2 = H*F(XN+H,YN+K1)
         YN = YN + .5*(K1+K2)
         XN = XBEGIN + N*H
         DERIV = F(XN,YN)
   10    PRINT 601, N,XN,YN,DERIV
                                          GO TO 1
      END
      REAL FUNCTION F(X,Y)
      REAL X,Y
      F = (1./X - Y)/X - Y*Y
                                          RETURN
      END
```

□ **مثال۴.۸:** مسئلهٔ مثال ۲.۸ را با روش مرتبهٔ دوم رونگه-کوتا (۳۶.۸) و با روش مرتبهٔ چهارم رونگه-کوتا (۳۷.۸) حل کنید.

درنتایج کامپیوتری که در بخش۳.۸ داده شد، فرمول (۳۶.۸) را روش۲ و فرمول (۳۷.۸) را روش۳ نامیده‌ایم. مشاهده می‌شودکه روش مرتبهٔ دوم رونگه-کوتا نتایجی به‌دست می‌دهدکه با الگوریتم مرتبهٔ دوم تیلر (روش۱) کاملا قابل مقایسه‌است. ولی روش مرتبهٔ چهارم رونگه-کوتانتایج خیلی بهبود یافته‌ای به‌دست می‌دهد،که به‌ازای $h = 1/16$ تا شش رقم اعشاری و به‌ازای سایرمقادیر h تا هفت یا هشت رقم اعشاری دقت دارند.

---

1. self-starting

کارایی محاسباتی روشهای ۲ و ۳ را می‌توان با بررسی تعداد دفعاتی که تابع می‌باید برای هریك محاسبه شود، مقایسه نمود. روش ۲، درهرمرحله نیاز به‌دوبارمحاسبهٔ تابع دارد و به‌ازای $h=1/128$، جمعاً ۲۵۶ بار محاسبهٔ تابع لازم‌است. روش ۳، درهر مرحله نیاز به‌چهاربار محاسبهٔ تابع دارد و به‌ازای $h=1/16$ جمعاً ۶۴ بار محاسبهٔ تابع لازم‌است ونتایج دقیقتری به‌دست می‌دهد. آشکار است که روش مرتبهٔ چهار رونگه-کوتا، برای این مسئله روش کارامدتری‌است و این مطلب از لحاظ کلی درست‌است. □

تمرین

۱-۵-۸ برای معادلهٔ $y'=x+y$، $y(0)=1$، خطای موضعی روش (۳۶.۸) رامحاسبه کنید. نتایج خود را با خطای الگوریتم مرتبهٔ دوم تیلر مقایسه نمایید. انتظار داریدکه کدامیك از این روشها دربازهٔ $[0,1]$ نتایج بهتری بدهد؟

۲-۵-۸ با استفاده از فرمول (۳۶.۸) و نمو $h=0.01$، چند مرحله از انتگرالگیری $y'=x+y$، $y(0)=1$، را انجام دهید و سپس برنامه‌ای بنویسیدکه این مسئله را از $x=0$ تا $x=1$ حل کند.

۳-۵-۸ به‌معادلهٔ (۳۵.۸) شرطی اضافه‌کنیدکه ضرایب f_{xx} در (۳۴.۸) و (۳۴.۸الف) با هم سازگار شوند. دستگاه معادلات حاصل را نسبت به a, b, α, β حل کنید، عبارت خطا برای روش مرتبهٔ دوم رونگه-کوتاکه با این مقادیر a, b, α, β به‌دست می‌آید تعیین کنید.

۴-۵-۸ می‌توان نشان دادکه خطای روش مرتبهٔ چهارم رونگه-کوتا به‌ازای نمو h، وقتی h به‌سمت صفر میل می‌کند، رابطه‌ای به‌شکل

$$y_n(h)-y(b)=A(b)h^4+\theta(h^5)$$

پیدا می‌کند، که در آن $b=x_0+nh$، و درنتیجه $n(h)=(b-x_0)/h$، و ثابت $A(b)$ به h بستگی ندارد. مانند انتگرالگیری رامبرگ، با استفاده از روش برونیابی، تقریبی برای $y(b)$ به‌دست‌آوریدکه خطای آن برابر با $\theta(h^5)$ باشد.

۶.۸ کنترل اندازهٔ h در روشهای رونگه-کوتا

در بخش ۵.۸، دو روش رونگه-کوتا (رك) یکی ازمرتبهٔ ۲ و دیگری ازمرتبهٔ ۴ موردبررسی قرار گرفت. اگرچه به‌دست آوردن فرمولهای روشهای رونگه-کوتا بسیار پیچیده می‌شود، می‌توان فرمولهای مذکور را از هر مرتبه‌ای به دست آورد. نکتهٔ مهم در استفاده از روشهای یك مرحله‌ای نوع رونگه-کوتا آن‌است که می‌توان خطای موضعی را براورد و درنتیجه مرحلهٔ مناسبی را برای حصول دقت لازم انتخاب کرد. هیچ دلیلی وجود نداردکه

الزاماً اندازهٔ نمو h، به‌گونه‌ای که درمثال ۴.۸ عمل کردیم، درتمام فاصلهٔ موردبحث ثابت بماند. به‌کاربردن نموهای ثابت متفاوت برای براورد دقت، به‌گونه‌ای کـه درمثال ۴.۸ انجام دادیم، ممکن است خیلی کارا نباشد. دراین بخش روشهایی را برای براورد خطای موضعی و برای تغییرنمو برحسب معیارخطایی مورد بررسی قرارمی‌دهیم.

روش‌اول برپایهٔ نصف‌کردن بازه استوار است. فرض می‌کنیم که روش (رك) از مرتبهٔ p را به‌کاربرده‌ایم و بـه‌ازای $h=x_n-x_{n+1}$ به‌نقطـهٔ x_n رسیده‌ایم. اکنون عمل انتگرالگیری از x_n تا $x_{n+1}=x_n+h$ را دوبار، یك‌بار به‌ازای نمو فعلی h و باردیگر با استفاده از نمو به‌طول $h/2$ انجام می‌دهیم. بنابراین در $x=x_{n+1}$ برای مقدار $y(x)$ به‌دوبراورد $y_h(x_{n+1})$ و $y_{h/2}(x_{n+1})$ دست می‌یابیم، ومقایسهٔ‌این‌دو، براوردی ازخطارا به‌دست خواهد داد. برای تعیین این براورد نخست توجه می‌کنیم که روش رونگه‌ـکوتا از مرتبهٔ p، بسط مجانبی خطای موضعی به‌شکل زیر را دارد

$$y_h(x_n+mh)=y(x_n+mh)+C(x_n+mh)h^p+\mathcal{O}(h^{p+1}) \qquad (۳۸.۸)$$

دراینجـا، $y_h(x_n+mh)$ معرف تقریبی‌است برای جواب $y(x)$ درنقطهٔ $x=x_n+mh$ که از روش رونگه‌ـکوتا با نمو h و بـا شروع از مقدار دقیق $y_n=y(x_n)$، بعد از m نمو به‌دست آمده‌است. بعلاوه، اگرچه ثابت $C(x_n+mh)$ به‌نقطهٔ $x=x_n+mh$ و بـه $f(x,y)$ بستگی دارد ولی مستقل از h است. بنابراین

$$y_h(x_{n+1})=y(x_{n+1})+C(x_{n+1})h^p+\mathcal{O}(h^{p+1}) \qquad (۳۹.۸ \text{ الف})$$

$$y_{h/2}(x_{n+1})=y(x_{n+1})+C(x_{n+1})(h/2)^p+\mathcal{O}(h^{p+1}) \qquad (۳۹.۸ \text{ ب})$$

باکم‌کردن رابطهٔ (۳۹.۸ الف) از (۳۹.۸ ب)، ملاحظه می‌کنیم که قسمت اصلی خطا در (۳۹.۸ ب) می‌تواند به‌صورت زیر براورد شود

$$C_n\left(\frac{h}{2}\right)^p \approx \frac{y_{h/2}(x_{n+1})-y_h(x_{n+1})}{1-2^p}$$

کمیت

$$D_n=\frac{|y_{h/2}(x_{n+1})-y_h(x_{n+1})|}{2^p-1} \qquad (۴۰.۸)$$

براوردقابل‌محاسبه‌ای ازخطا را درتقریب $y_{h/2}(x_{n+1})$ برای‌ما فراهم می‌سازد ومی‌تواند برای تصمیم‌گیری دراین باب‌که آیا نمو به‌کار رفتهٔ h دقیقاً مناسب است یـاآنکه خیلی بزرگ یا خیلی‌کوچك‌است، به‌کار رود.

اکنون فرض می‌کنیم کـه یك تحمل خطای موضعی ε داده شده‌است و می‌خواهیـم خطای براورد شدهٔ D_n راکمتر از تحمل خطای موضعی **درهر واحد مرحله** نگه داریـم،

یعنی می‌خواهیم داشته باشیم

$$D_n \leqslant \varepsilon h \qquad (۴۱۰۸)$$

فرض‌می کنیم که $y_h(x_{n+1})$، $y_{h/2}(x_{n+1})$ و D_n رامحاسبه کرده‌ایم. اکنون می‌باید تصمیم بگیریم که آیا مقدار $y_{h/2}(x_{n+1})$ مورد قبول‌است یا نیست، و برای انتگرالگیری بعدی، نمو h چه مقدار باید باشد. به‌روشی که بعداً توضیح داده خواهد شد از روی تحمل‌خطای مفروض ε، کران پایینی برای خطای $\varepsilon' < \varepsilon$ به‌دست می‌آوریم. حالات ممکنهٔ زیر وجود دارند:

$$\varepsilon' < \frac{D_n}{h} < \varepsilon \quad \text{(i)}$$

دراین‌حالت، مقدار $y_{h/2}(x_{n+1})$ را می‌پذیریم، وعمل انتگرالگیری را با همان‌نمو h، از x_{n+1} ادامه می‌دهیم.

$$\frac{D_n}{h} > \varepsilon \quad \text{(ii)}$$

در این حالت، مقدار خطا خیلی بزرگ‌است و ازاین‌رو باید h را ـ مثلا به $h/2$ ـ تبدیل وعمل انتگرالگیری را مجدداً ازنقطهٔ $x = x_n$ شروع کنیم.

$$\frac{D_n}{h} < \varepsilon' \quad \text{(iii)}$$

دراین‌حالت، دقت بیش‌ازمقدار خواسته شده‌است. مقدار $y_{h/2}(x_{n+1})$ رامی‌پذیریم و $2h$ را جایگزین h می‌کنیم و نقطهٔ x_{n+1} انتگرالگیری را از سر می‌گیریم.

اگر اندازهٔ نمو را درهربازه به‌نصف‌شدن یا دوبرابرشدن محدود کنیم، آنگاه برای یك روش ازمرتبهٔ p، کران پایین ε' می‌تواند برابر با مقدار زیرگذاشته شود

$$\varepsilon' = \varepsilon / 2^{p+1}$$

و این‌امر بدان‌جهت‌است که نصف کردن h، خطا را تقریباً به‌اندازهٔ $1/2^{p+1}$ تقلیل‌می‌دهد. برای روش مرتبهٔ ۴ رونگه‌ـکوتا، داریم $p = 4$، از این‌رو $\varepsilon' = \varepsilon/32$. در واقع بیشتر اوقات‌توصیه نمی‌شودکه‌نمورا تغییردهند وبرای‌اطمینان می‌توان $\varepsilon' = \varepsilon/50$ را به‌کاربرد.

شکل‌بسیار پیچیدهٔ دیگری‌ازکنترل نمو که h را به‌دوبرابرشدن یا نصف‌شدن محدود نمی‌کند به‌شکل زیراست. با توجه به‌رابطهٔ (۴۰۸) داریم

$$D_n \approx 2C_n\left(\frac{h}{2}\right)^{p+1} \qquad (۴۲۰۸\text{الف})$$

هدف ما آن‌است که برای مرحلهٔ بعدی، نموی برابر $\bar{h}$ انتخاب کنیم. ازآنجایی که قسمت اصلی خطا درمرحلهٔ بعدی برابر با $۲C_n(\bar{h}/۲)^{p+۱}$ است، h باید چنان انتخاب شودکه رابطهٔ تحمل خطای (۴۱.۸) برقرارشود، از این‌رو باید داشته باشیم

$$۲C_n\left(\frac{\bar{h}}{۲}\right)^{p+۱} \leqslant \varepsilon\bar{h} \qquad (۴۲.۲\text{ب})$$

و باز با فرض‌آنکه C_n خیلی تغییر نمی‌کند، می‌توانیم C_n را از روابط (۴۲.۸ الف) و (۴۲.۸ ب) به‌گونهٔ زیرحذف کنیم: با توجه به‌رابطهٔ (۴۲.۸ ب) داریم

$$۲C_n\left(\frac{h}{۲}\right)^{p+۱}\left(\frac{\bar{h}}{۲}\right)^{p+۱} \leqslant \varepsilon\bar{h}\left(\frac{h}{۲}\right)^{p+۱}$$

$$D_n\frac{(\bar{h})^p}{۲^{p+۱}} \leqslant \frac{\varepsilon h \cdot h^p}{۲^{p+۱}}$$

$$\bar{h}^p \leqslant h^p\varepsilon h/D_n$$

$$\bar{h} \approx h(\varepsilon h/D_n)^{۱/p} \qquad (۴۲.۸\text{ پ})$$

بنابراین اگر قبلا به‌ازای نمو h، انتگرالگیری را با موفقیت انجام داده باشیم، نمو انتگرالگیری بعدی را باید برابر با $\bar{h}$ و یا شاید برای اطمینان بیشتر، کمتر ازآن بگیریم. برای مثال فرض کنیدکه روشی به‌کار برده‌ایم که با $p=۴$ داریم $\varepsilon=۱۰^{-۶}$ و $h=۰ر۱$ و D_n طبق محاسبه باید برابر با $۱۰^{-۵}$ باشد. پس

$$\bar{h} \approx ۰ر۱(۰ر۰۱)^{۱/۴} = ۰ر۱(۰ر۳۲) = ۰ر۰۳۲$$

لذا این شرایط به‌مقدار خیلی کوچکتر h نیاز دارد. از طرف دیگر اگر باز $p=۴$ و $h=۰ر۱$ و $\varepsilon=۱۰^{-۶}$، و $D_n=۱۰^{-۸}$ را حساب کنیم، آنگاه داریم

$$\bar{h} \approx ۰ر۱(۱۰)^{۱/۴} \approx ۰ر۱(۱ر۸) = ۰ر۱۸$$

لذا نمو می‌تواند تقریباً دوبرابر شود. استفاده از نموهای متغیر به‌طور قابل‌ملاحظه‌ای به‌پیچیدگی یك برنامه می‌افزاید و درمجموعه‌ای از نقاط که به‌فواصل غیریکنواخت قرار گرفته‌اند منجر به‌نتایجی‌می‌شودکه برای استفاده‌کننده ناخوشایند است. نصف کردن و دوبرابر کردن بازه‌هاعموماً برای استفاده‌کننده‌قابل‌قبول‌تراست. ازسوی دیگر،در برنامه‌هایی که درآنها کنترل نمو به‌طور خودکار انجام می‌گیرد براوردهای خیلی خوبی ازدقت‌فراهم شده‌است و روی هم‌رفته‌کاملا کارا هستند.

عیب اصلی‌این‌روش برای کنترل خطا، کاراضافه‌ای‌است که‌اساساً باید انجام گیرد. در سالهای اخیرشکلهای متفاوتی‌ازروش‌رونگه‌ـکوتاکه برای کنترل نمو مناسب هستندپیشنهاد

شده‌اندکه چندتا از این شکلهای تازهٔ متفاوت با نامهای مرسن[1]، ورنر[2]، وفهلبرگ[3] همراه هستند، روشی راکه توسط فهلبرگ ارائه گردیده‌است به‌اجمال بررسی می‌کنیم وآن را با رکف ۴۵ [۲۸] نشان می‌دهیم. در این روش، درهرمرحله شش‌بار محاسبهٔ تابع موردنیاز است ولی براورد خطای آن به‌صورت خودکار انجام می‌گیرد، و درعین‌حال دقت آن از روش مرتبهٔ چهارم استانده بیشتراست. فهلبرگ نشان دادکه ترکیب مقادیر چهارتا ازاین توابع با مجموعه‌ای از ضرایب می‌تواند روش مرتبهٔ چهارم را به‌دست دهد، در حالی‌که ترکیب هرشش مقدار تابع با مجموعهٔ دیگری از ضرایب می‌تواند برای به‌وجود آوردن روش مرتبهٔ پنجم به‌کار رود. سپس مقایسهٔ مقادیر به دست آمده از روشهای مرتبه‌های چهارم وپنجم منجر به‌براورد خطایی می‌شودکه می‌تواند برای کنترل h به‌کار رود.

راهی‌را که فهلبرگ اختیارکرده خیلی به‌اختصار توضیح می‌دهیم. فرض می‌کنیم که از معادلهٔ $y'=f(x,y)$ تا یك نقطهٔ x_n به‌ازای نمو h_n انتگرال گرفته‌ایم و اکنون می‌خواهیم در $x=x_{n+1}$ براوردی برای $y(x)$ به‌دست آوریم. براورد اول به ازای ضرایبی مانند c_i بافرمول زیرمعین می‌شود

$$y_{n+1}=y_n+\sum_{i=1}^{6} c_i k_i \qquad (۴۳.۸ \text{ الف})$$

و براورد دوم به‌ازای مجموعهٔ دیگری از ضرایب c_i^* به‌صورت زیر تعیین می‌شود

$$y_{n+1}^*=y_n+\sum_{i=1}^{6} c_i^* k_i \qquad (۴۳.۸ \text{ ب})$$

لذا براورد خطا برای کنترل h با فرمول زیرمحاسبه می‌شود:

$$D_n=\sum_{i=1}^{6}(c_i-c_i^*)k_i$$

همچنان‌که قبلا توضیح داده شد، این رابطه برای براورد مناسبی از h در انتگرالگیری بعدی می‌تواند به‌کار رود. توابع k_i درهردو فرمول یکی هستند و می‌توانند به‌صورت‌زیر بیان شوند

$$k_i=h_n f(x_n+\alpha_i h_n,\ y_n+\sum_{j=1}^{i-1}\beta_{ij}k_i) \qquad i=1,\dots,6$$

درمورد ضرایب α_i و β_{ij} انتخابهای ممکنهٔ زیادی وجود دارندکه به‌روش مرتبهٔ ۵ رونگه‌کوتا منجر می‌شوند. فهلبرگ، مجموعهٔ خاصی از ضرایب را پیشنهاد نموده‌است‌که ما در اینجا آنها را عرضه نخواهیم کرد. دانشجویان علاقه‌مند، برای اطلاع بیشتر از جزئیات این روش می‌توانند به‌مرجع [۲۸] آخر این‌کتاب مراجعه‌کنند.

---

1. Merson 2. Verner 3. Fehlberg

روش دیگر رونگه-کوتا باکنترل اندازهٔ نمو که توسط ورنر ارائه گردیده، اساس بسیاری از زیربرنامه‌های موفق درحل معادلات دیفرانسیل است که به DVERK موسوم ودرکتابهای زیربرنامه‌ای به‌تعداد زیاد در دسترس هستند. درروش ورنرکه با رنگ و ۵۶ نشان داده می‌شود، درهرمرحله به‌هشت بار محاسبهٔ تابع احتیاج است و ازاین‌توابع برای $y(x)$ دو براورد به‌دست می‌آورند که مبنای یکی تقریب مرتبهٔ پنجم‌است و مبنای دیگری تقریب مرتبهٔ ششم. سپس مقایسهٔ این دو براورد، مبنایی برای انتخاب نمو فراهم می‌سازد. چندآزمون اولیه در مورد این روش در دانشگاه تورنتو[1] [۲۹] به‌عمل آمده‌است. روش مذکور سپس به‌صورت زیربرنامهٔ DVERK درآمده ودر IMSL، هوستون[2]، تکزاس[3]، منتشرشده‌است. IMSL که معرف **کتابهای بین‌المللی ریاضی وآمار**[4] است، مجموعه‌ای از زیربرنامه‌های کاملا آزمایش‌شده برای بسیاری‌ازمسائل‌متنوع آمار وریاضی را دربردارد. دستیابی به‌این مجموعهٔ کتاب براساس‌پرداخت حق‌اشتراك[5] صورت می‌گیرد وتقریباً برای تمام‌کامپیوترهای متوسط و بزرگ‌ازجمله IBM، CDC، UNIVAC، Burroughs و Honeywell موجود است. چون‌بسیاری ازمراکزکامپیوتری حق اشتراك برای‌مجموعهٔ IMSL پرداخت می‌کنند، دیگر ما دراینجا زیربرنامهٔ DVERK را نخواهیم داد. چون دراین‌فصل برای حل‌چندین‌مسئله ازاین زیربرنامه استفاده می‌کنیم، پارامترهای موجود در حکم فراخوانی[6] و انتخابهای[7] مختلف را به‌طور اجمال توضیح می‌دهیم.

در استفادهٔ معمولی در صورت نبود انتخـاب[8] و بعد از تعیین مقادیر اولیه، قسمت اصلی‌برنامهٔ مربوط به‌حل معادلهٔ دیفرانسیل‌مرتبهٔ اول $y' = f(x, y)$ از $x = \text{XBEGIN}$ تا $x = \text{XM}$ شامل یك حلقهٔ DO به‌صورت زیر است:

```
      X = XBEGIN
      Y = YBEGIN
      DO 10 K=1,M
         XEND = XBEGIN + FLOAT(K)*(XM - XBEGIN)/FLOAT(M)
         CALL DVERK ( N, FCN1, X, Y, XEND, TOL, IND, C, NW, W, IER )
         PRINT 600, XEND, Y(1), C(24)
600      FORMAT(F19.6,E21.8,F15.0)
 10   CONTINUE
```

معانی پارامترهای موجود در این زیربرنامه به‌قرار زیرند:

N ≡ تعداد معادلاتی که باید حل‌شوند (دراینجا $N = ۱$).

FCNI ≡ نام زیربرنامهٔ مربوط به $f(x, y)$، که استفاده‌کننده باید به‌صورت‌یك برنامهٔ خارجی تهیه‌کند.

X ≡ مقدار اولیهٔ متغیرمستقل.

---

1. Toronto 2. Houston 3. Texas
4. International Mathematical and Statistical Library
5. subscription 6. statement 7. options
8. default options

Y ≡ مقدار اولیهٔ متغیر وابسته.

XEND ≡ مقداری از x که به‌ازای آن جواب می‌باید بیرون داده شود.

TOL ≡ میزان تحمل کنترل خطا؛ در عین اینکه خصوصیات تحمل خطا ممکن است انواع مختلف داشته باشند در انتخاب قراردادی این تحمل سعی بر این است که خطای کلی[1] نسبی از تحمل برای کنترل خطا کمتر باشد.

IND ≡ ۱ موجب نبودن همهٔ انتخابهایی است که باید استفاده شوند.

≡ ۲ اجازه می‌دهد که انتخابها صورت گیرند.

C ≡ بردارهای ارتباطی به‌طول ۲۴؛ اگر IND روی ۲ گذاشته شده باشد، برخی از این بردارها را استفاده‌کننده می‌تواند تعیین کند؛ این انتخابها انواع مختلف کنترل خطا، مینیمم یا ماکسیمم نموها، حدود تعداد ارزیابیهای تابع و غیره را مدنظر قرار می‌دهند.

NW ≡ اولین بعد[2] (تعداد سطرهای) ماتریس فضای کار W باید حداقل به‌اندازهٔ N باشد.

W ≡ ماتریس فضای کار که اولین بعد (تعداد سطرهای) آن برابر با NW است و دومین بعد (تعداد ستونهای) آن باید ناکوچکتر از ۹ باشد.

IER ≡ پرچم[3] خطا برای نشان دادن انواع مختلف خطاهایی که با آنها مواجه می‌شویم.

در حلقهٔ DO، مذکور در بالا، نقاط XEND مقادیری از x هستند که به‌ازای آنها جواب بیرون داده می‌شود. در این حالت جواب در M نقطهٔ متساوی‌الفاصله $XBEGIN+k\Delta X$ بیرون داده خواهد شد که در آن $\Delta X=(XM-XBEGIN)/M$.

در خود زیربرنامه، DVERK نمونه‌های مناسبی را برای دقت موردنیاز انجام خواهد داد. همچنان که مراحل انتگرالگیری پیش می‌رود، نمو به‌طور طبیعی تغییر می‌کند. زیربرنامهٔ مذکور تعداد محاسبات تابع را هم که برای یافتن جواب در نقطهٔ XEND لازم است حفظ خواهد کرد. DVERK زیربرنامه‌ای است با دقت بسیار بالا که در هر مرحله از انتگرالگیری، حداقل به‌هشت بار محاسبهٔ تابع احتیاج دارد. تعداد محاسبات تابع که عملا به‌کار برده می‌شود، در C(۲۴) ذخیره می‌گردد و همان گونه که در بالا انجام دادیم می‌تواند به‌عنوان یك خروجی بیرون داده شود.

روش مزبور برای معادلهٔ دیفرانسیل

$$y'(x)=\frac{1}{x^2}-\frac{y(x)}{x}-y^2(x)$$

$$y(1)=-1$$

که در مثال ۲.۸ مورد بررسی قرار گرفت، به‌کار رفته و برنامهٔ کامل و نتایج حاصل از آن در زیر داده شده است

---

1. global 2. dimension 3. flag

```
C  USE OF  D V E R K  TO SOLVE EXAMPLE 8.2 .
      INTEGER IER,IND,K,N,NW
      REAL C(24),TOL,W(1,9),X,XEND,Y(1)
      DATA N , X ,Y(1), TOL ,IND,NW
     *   / 1 , 1.,-1. ,1.E-7, 1 , 1 /
      EXTERNAL FCN1
      DO 10 K=1,4
         XEND = 1. + FLOAT(K)/4.
         CALL DVERK ( N, FCN1, X, Y, XEND, TOL, IND, C, NW, W, IER )
         PRINT 600, XEND,Y(1),C(24)
  600    FORMAT(11X,F8.6,5X,E16.8,5X,F11.8)
   10 CONTINUE
                                          STOP
      END
      SUBROUTINE FCN1 ( N, X, Y, YPRIME )
      REAL X,Y(1),YPRIME(1)
      YPRIME(1) = (1./X - Y(1))/X - Y(1)*Y(1)
                                          RETURN
      END
```

برونداد

X	Y(1)	FCN EVALS
1.25	−0.79999999	16.
1.50	−0.66666664	24.
1.75	−0.57142854	32.
2.00	−0.49999996	40.

نتایج به‌دست‌آمده از لحاظ دقت با جوابهای حاصل از به‌کاربردن روش کلاسیک[1] مرتبهٔ چهارم با نمو ثابت $h = ۱/۳۲$، قابل‌مقایسه هستند. از آنجایی‌که روشهای کلاسیک مرتبهٔ چهارم درهرمرحله به‌چهاربار محاسبهٔ تابع احتیاج دارند، کلا ۱۲۸ بارمحاسبهٔ تابع لازم خواهد بود تا به‌دقتی درحدود هفت رقم‌اعشاری برسیم. درحالی‌که، برای رسیدن به‌همان دقت، در زیربرنامهٔ DVERK، تنها به ۴۰ بارمحاسبهٔ تابع‌نیاز است. لازم به‌تذکر است که برای داشتن خروجی در $x = ۱٫۲۵$ به ۱۶ بار محاسبهٔ تابع نیاز بودکه این امر حاکی‌ازبزرگی نمو $h = ۱/۴$ بود، و ظاهراً برای رسیدن به‌دقت $۱٫۱۰^{-۷}$ می‌بایستی این نمو نصف می‌شد.

دربخش۱۲.۸ استفاده از زیربرنامهٔ DVERK را برای حل یک دستگاه معادلات دیفرانسیل مرتبهٔ اول نشان خواهیم داد.

تمرین

۱-۶.۸ فرض‌کنید برای حل‌یک معادلهٔ دیفرانسیل،از روش‌مرتبهٔ دوم رونگه-کوتااستفاده کرده‌ایم وکنترل اندازهٔ نمو نیز براساس نصف‌کردن فاصله صورت‌گرفته‌است. اگر از نمو $h = ۰٫۱$ و معیار خطای $\varepsilon = ۱۰^{-۶}$ استفاده‌کنیم و در نقطهٔ $x = x_n$، $D_n = ۱۰^{-۴}$ به‌دست‌آید، نمو $\bar{h}$ برای مرحلهٔ بعدی انتگرالگیری چه مقدار باید باشد؟

۲-۶.۸ برنامه‌ای‌برای روش‌مرتبهٔ دوم رونگه-کوتاکه درآن‌کنترل‌اندازهٔ نمو به‌دوبرابر-

1. classical

شدن و نصف شدن محدود می‌شود، بنویسید. این برنامه را برای حل معادلهٔ مثال ۲.۸ به‌ازای $\varepsilon = ۱۰^{-۶}$ به‌کار برید.

۳-۶.۸ ببینید که آیا مرکز محاسباتی شما مجموعهٔ زیربرنامه‌های IMSL را انجام می‌دهند یا نه. این صورت زیربرنامهٔ DVERK را برای حل معادلات دیفرانسیل زیر به‌کار برید. درهرحالت TOL را برابر با $۱۰^{-۷}$ قرار دهید و خروجی را درمقادیر XEND به‌صورت زیر به‌دست آورید

$$\text{XEND} = \text{XO} + \text{K}(\text{XM}-\text{XO})/۱۰ \qquad \text{K} = ۱, ۲, \ldots, ۱۰$$

(الف) $y' = x - ۱ + y/x$

$$\text{XO} = ۱, \text{XM} = ۲, y(\text{XO}) = ۲$$

(ب) $y' = xy^{۲}$

$$\text{XO} = ۱, \text{XM} = ۴, y(\text{XO}) = ۱$$

۷.۸ فرمولهای چندمرحله‌ای

الگوریتم مرتبهٔ kی تیلر و روشهای رونگه-کوتا هردو مثالهایی از روشهای یك‌مرحله‌ای هستند. این روشها به‌اطلاعاتی دربارهٔ جواب دریك نقطهٔ $x = x_n$ نیاز دارند تا از روی این جواب، به‌تعیین y درنقطهٔ بعدی $x = x_{n+۱}$، بپردازند. در روشهای **چندمرحله‌ای**، از اطلاعات پیرامون جواب در بیش از یك نقطه استفاده می‌کنند. فرض کنید که برای y' و y درچند نقطه با فواصل مساوی مثل $x_۰$، $x_۱$، $\ldots$، x_n، تقریبهایی ازقبل به‌دست آورده باشیم. یك‌دسته از روشهای چندمرحله‌ای برپایهٔ اصل انتگرالگیری عددی استوار شده‌اند. اگر از معادلهٔ دیفرانسیل $y' = f(x, y)$ از x_n تا $x_{n+۱}$ انتگرال بگیریم، خواهیم داشت

$$\int_{x_n}^{x_{n+۱}} y'\,dx = \int_{x_n}^{x_{n+۱}} f(x, y(x))\,dx$$

یا

$$y_{n+۱} = y_n + \int_{x_n}^{x_{n+۱}} f(x, y(x))dx \qquad (۴۴.۸)$$

برای انجام انتگرالگیری در رابطهٔ (۴۴.۸)، $f(x, y(x))$ را با یك بسجمله‌ای، که $f(x, y(x))$ را در $(m+۱)$ نقطهٔ x_n، $x_{n-۱}$، $x_{n-۲}$، $\ldots$، x_{n-m} درونیابی کند، تقریب می‌زنیم. برای این منظور اگر از قرارداد

$$f(x_k, y(x_k)) = f_k$$

استفاده کنیم، می‌توانیم فرمول پسرو نیوتن (تمرین ۶.۲–۸ را ببینید) از مرتبهٔ m را به‌کار گیریم:

$$p_m(x)=\sum_{k=0}^{m}(-1)^k\binom{-s}{k}\Delta^k f_{n-k} \qquad s=\frac{x-x_n}{h}$$

با قراردادن رابطهٔ فوق در (۴۴.۸) و با توجه به‌تساوی $dx=h\,ds$، داریم

$$y_{n+1}=y_n+h\int_0^1\sum_{k=0}^{m}(-1)^k\binom{-s}{k}\Delta^k f_{n-k}\,ds$$

$$=y_n+h\{\gamma_0 f_n+\gamma_1\Delta f_{n-1}+\cdots+\gamma_m\Delta^m f_{n-m}\} \qquad (۴۵.۸)$$

که در آن

$$\gamma_k=(-1)^k\int_0^1\binom{-s}{k}ds \qquad (۴۵.۸\text{ الف})$$

با توجه به‌تعریف تابع دوجمله‌ای که در فصل ۲ آمده بود، می‌توانیم مقادیر γ_k را به‌آسانی محاسبه کنیم، که چندتای اول آن به‌قرار زیرند:

$$\gamma_0=1 \quad \gamma_1=\frac{1}{2} \quad \gamma_2=\frac{5}{12} \quad \gamma_3=\frac{3}{8} \quad \gamma_4=\frac{251}{720}$$

فرمول (۴۵.۸) به‌روش **ادمز–بشفورث**[1] معروف است. ساده‌ترین حالت آن، با قراردادن $m=0$ در فرمول (۴۵.۸)، باردیگر روش اویلر را به‌دست می‌دهد. به‌طور کلی استفاده از فرمول (۴۵.۸) مستلزم داشتن مقدار $y'=f$ در $m+1$ نقطهٔ x_n، x_{n-1}، ...، x_{n-m} است. از روی این مقادیر می‌توانیم تفاضلهای

$$\Delta f_{n-1},\Delta^2 f_{n-2},\cdots,\Delta^m f_{n-m}$$

را تشکیل دهیم و از (۴۵.۸) می‌توانیم y_{n+1} را محاسبه کنیم؛ و از معادلهٔ دیفرانسیل می‌توانیم $f_{n+1}=f(x_{n+1},y_{n+1})$ را به‌دست آوریم. اکنون نقطهٔ x_{n+1} را نقطهٔ x_n می‌گیریم و خط جدیدی از تفاضلها تشکیل می‌دهیم و این روند را تکرار می‌کنیم. به‌ازای $m=3$، که معمولاً در عمل به‌کار برده می‌شود، جدول تفاضلها به‌صورت زیرخواهد بود

---

1. Adams–Bashforth

$$
\begin{array}{llllll}
x_{n-3} & y_{n-3} & f_{n-3} & & & \\
 & & & \Delta f_{n-3} & & \\
x_{n-2} & y_{n-2} & f_{n-2} & & \Delta^2 f_{n-2} & \\
 & & & \Delta f_{n-2} & & \Delta^3 f_{n-3} \\
x_{n-1} & y_{n-1} & f_{n-1} & & \Delta^2 f_{n-2} & \\
 & & & \Delta f_{n-1} & & \\
x_n & y_n & f_n & & &
\end{array}
$$

و فرمول (۴۵.۸) به‌حالت خاص زیر بدل می‌شود

$$y_{n+1} = y_n + h\left(f_n + \frac{1}{2}\Delta f_{n-1} + \frac{5}{12}\Delta^2 f_{n-2} + \frac{3}{8}\Delta^3 f_{n-3}\right) \qquad (۴۶.۸)$$

درعمل از لحاظ محاسباتی آسانتر است که به‌جای کارکردن با تفاضلها با عرضها* کار کنیم. ازتعریف عملگر تفاضل پیشرو Δ، خواهیم داشت

$$\Delta f_{n-1} = f_n - f_{n-1}$$

$$\Delta^2 f_{n-2} = f_n - 2f_{n-1} + f_{n-2}$$

$$\Delta^3 f_{n-3} = f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3}$$

از قراردادن این روابط در (۴۶.۸) و مرتب‌کردن نتیجهٔ حاصله، خواهیم داشت

$$y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \qquad (۴۷.۸)$$

خطای موضعی رابطهٔ (۴۶.۸) را می‌توان به‌صورت زیر به‌دست آورد: باتوجه به تمرین۶.۲–۸ می‌دانیم که خطا در فرمول پسرو نیوتن به‌ازای $n = 3$ و $k = 0$ به‌صورت زیر است

$$h^4 f^{(4)}(\eta)\binom{-s}{4}$$

* عرضها منظور مقادیر تابع‌است و نه‌تفاضلهای آنها.—م.

بنابراین خطا درفرمول (۴۶.۸) به‌کمك رابطهٔ

$$E_{AB}=h\int_{0}^{1}h^{4}f^{(4)}(\eta)\binom{-s}{4}ds$$

تعیین می‌شود. از آنجاکه $\binom{-s}{4}ds$ دربازهٔ $[0,1]$ تغییرعلامت‌نمی‌دهد، نقطه‌ای مانند ξ بین x_{n-3} و x_{n+1} وجود دارد به‌طوری‌که

$$E_{AB}=h^{5}f^{(4)}(\xi)\int_{0}^{1}\binom{-s}{4}ds$$

$$\boxed{E_{AB}=h^{5}y^{v}(\xi)\frac{251}{720}} \qquad (۴۸.۸)$$

برای استفاده از رابطهٔ (۴۷.۸) باید ۴ مقدار آغازین را داشته باشیم. این مقادیر آغازین باید از یك منبع مستقل به‌دست آمده باشند. برای نشان دادن چگونگی استفاده از رابطهٔ (۴۷.۸) چند مرحلهٔ اول انتگرالگیری معادلهٔ زیر را به‌ازای $h=0.1$ انجام می‌دهیم

$$y'=-y^{2}$$
$$y(1)=1$$

جواب دقیق این مسئله برابراست با $y=1/x$. درجدول زیر چهارمقدارآغازین اول از جواب دقیق به‌دست آمده‌اند و درایه‌های دیگر از رابطهٔ (۴۷.۸).

x_n	y_n	$f_n=-y_n^2$	$y(x_n)=1/x_n$
1.0	1.000000000	−1.000000000	
1.1	0.909090091	−0.826446281	0.909090091
1.2	0.833333333	−0.694444444	0.833333333
1.3	0.769230770	−0.591715976	0.769230770
1.4	0.714285632	−0.510204082	0.714285714
1.5	0.666686030	−0.444702660	0.666666667
1.6	0.625244613	−0.390932722	0.625000000

مشاهده می‌شود که مقادیر y_n که توسط فرمول (۴۷.۸) محاسبه شده‌اند در رقم چهارم خود تقریباً دو واحد خطا دارند. با استفاده از فرمول برآورد خطای موضعی (۴۳.۶) و این امر که داریم

$$|y^{(۷)}(x)| = \left|\frac{۵!}{x^۶}\right| \leqslant ۱۲۰ \qquad ۱ < x < ۲$$

کران خطا را به‌دست می‌آوریم

$$|E_{AB}| \leqslant \frac{۲۵۱}{۷۲۰}(۱۲۰)(۱۰^{-۵}) \approx ۰٫۰۰۰۴$$

این کران در حدود دو برابر خطایی است که از رفتن از یك مرحله به‌مرحلهٔ دیگر ایجاد می‌شود.

به‌همین طریق با استفاده از انتگرالگیری عددی، فرمولهای دیگری از نوع چندمرحله‌ای می‌توان به‌دست آورد. در رابطهٔ (۴۳.۸) به‌جای انتگرالگیری از $f(x, y)$ از x_n تا $x_{n+۱}$، می‌توانیم مثلا از x_{n-p} تا $x_{n+۱}$ به‌ازای یك عدد صحیح $p \geqslant ۰$ انتگرالگیری کنیم. اگر با فرمول پسرو نیوتن در $m+۱$ نقطهٔ x_n، $x_{n-۱}$، ...، x_{n-m} مجدداً درونیابی کنیم، خواهیم داشت

$$y_{n+۱} = y_{n-p} + h\int_{-p}^{۱} \sum_{k=۰}^{m} (-۱)^k \binom{-s}{k} \Delta^k f_{n-k}\, ds \qquad (۴۹.۸)$$

درحالت $p = ۰$، فرمول (۴۴.۸) ادمز-بشفورث به‌دست می‌آید. متناظر با $m = ۱$، $p = ۱$ و $m = ۳$، $p = ۳$ فرمولهای خاص و جالبی از این نوع حاصل می‌شوند. این فرمولها همراه با عبارت خطای موضعی آنها به‌صورت زیرند

$$\boxed{y_{n+۱} = y_{n-۱} + ۲hf_n} \qquad \boxed{E = \frac{h^۳}{۳}y'''(\xi)} \qquad (۵۰.۸)$$

$$\boxed{y_{n+۱} = y_{n-۳} + \frac{۴h}{۳}(۲f_n - f_{n-۱} + ۲f_{n-۲})} \qquad \boxed{E = \frac{۱۴}{۴۵}h^۵y^{(۷)}(\xi)}$$

(۵۱.۸)

فرمول (۵۰.۸)، که از لحاظ سادگی با روش اویلر قابل مقایسه‌است، خطای گسسته‌سازی مطلوبتری دارد. همچنین فرمول (۵۱.۸) که مستلزم داشتن $f(x, y)$ تنها در سه نقطه‌است، خطای گسسته‌سازی قابل‌مقایسه‌ای با روش ادمز-بشفورت (۴۷.۸) دارد. می‌توان نشان داد که به‌ازای عدد فرد m و $m = p$، هر فرمولی به‌صورت (۴۹.۸) دارای این ویژگی است

که ضریب تفاضل mام در آن برابر با صفر می‌شود، و بنابراین فرمولی حاصل می‌شود که دقت آن بیش از مقداری است که انتظار می‌رود. از سوی دیگر این فرمولها دستخوش ناپایداری[1] بیشتری قرار می‌گیرند، که مفهوم ناپایداری بعداً مورد بحث واقع خواهد شد.

یك اشکال اصلی فرمولهای چندمرحله‌ای این است که آنها «خود-آغازگر» نیستند. بنابراین در روش ادمز-بشفورت (۴۷۰۸)، و قبل از به‌کار بستن این فرمول باید چهار مقدار متوالی $f(x, y)$ را در نقاط با فواصل مساوی داشته باشیم. این مقادیر آغازین را باید به‌کمك روش مستقلی به‌دست آورد. برای مثال می‌توانیم الگوریتم تیلر یا یکی از روشهای رونگه-کوتا را برای به‌دست آوردن این مقادیر آغازین به‌کار بریم. و نیز برای به‌دست آوردن دقت مطلوب باید اطمینان حاصل کنیم که این مقادیر آغازین خود به‌قدر لازم دقیق هستند. اشکال دوم روش ادمز-بشفورت این است که با اینکه خطای موضعی مربوط به گسسته‌سازی برابر با $\mathcal{O}(h^5)$ است، ضریب جملهٔ خطا اندکی بیشتر از همان ضریب برای فرمولهای رونگه-کوتا از مرتبهٔ مشابه است. به‌همین دلیل روش رونگه-کوتا معمولاً، ولی نه همیشه، از دقت بیشتری برخوردار است. از سوی دیگر فرمولهای چندمرحله‌ای در مقایسه با چهار بار محاسبه کردن در هر مرحله با روش رونگه-کوتا، در هر مرحله تنها به‌یك بار محاسبهٔ مشتق احتیاج دارند، و بنابراین این روش به‌نحو قابل‌ملاحظه‌ای سریعتر است، و به کارهای محاسباتی کمتری نیاز دارد.

□ مثال ۵۰۸: معادلهٔ

$$y' = x + y \qquad y(0) = 0$$

را از $x = 0$ تا $x = 1$ با استفاده از روش ادمز-بشفورت حل کنید.

برنامهٔ فورترن و نتایج حاصله از آن برای این مسئله در زیر داده شده‌است. جواب دقیق این مسئله چنین‌است $y = e^x - x - 1$. با استفاده از این جواب، چهار مقدار آغازین محاسبه شده‌اند. اولین ستون نتایج مقادیر x_n را به‌ازای $h = 1/32$ می‌دهد، ستون دوم مقادیر y_n را که به‌کمك فرمول (۴۷۰۸) محاسبه شده‌اند معین می‌کند، و ستون سوم مقدار $y(x_n)$ را که از روی جواب حساب شده نشان می‌دهد و ستون چهارم مقدار خطای $e_n = y_n - y(x_n)$ را.

نتایج تقریباً تا شش رقم با معنی دقیق‌اند، که همان چیزی‌است که از فرمول خطای (۴۸۰۸) تقریباً انتظار می‌رفت. از آنجایی که خطای انباشته شدهٔ مربوط به گسسته سازی، برابر با $\mathcal{O}(h^4)$ است، انتظار می‌رود که اگر اندازهٔ نمو h نصف شود، این خطا به‌اندازهٔ $1/16$ تقلیل یابد.

---

1. instability

برنامۀ فورترن برای مثال ۵.۸

```
C  ADAMS-BASHFORTH METHOD
       INTEGER I,N,NSTEPS
       REAL ERROR,F(4),H,XBEGIN,XN,YBEGIN,YN
C
       SOLN(X) = EXP(X) - 1. - X
C
C   ** INITIALIZE
       PRINT 600
  600  FORMAT('1ADAMS-BASHFORTH METHOD'/
      * '0',4X,'N',13X,'XN',15X,'YN',13X,'Y(XN)',12X,'ERROR'/)
       NSTEPS = 32
       H = 1./NSTEPS
       YBEGIN = 0.
       XBEGIN = 0.
C
C   ** COMPUTE FIRST FOUR POINTS USING EXACT SOLUTION
       F(1) = XBEGIN + YBEGIN
       N = 0
       ERROR = 0.
       PRINT 601, N,XBEGIN,YBEGIN,YBEGIN,ERROR
  601  FORMAT('  ',I3,4X,4E17.8)
       DO 20 N=1,3
          XN = XBEGIN + N*H
          YN = SOLN(XN)
          F(N+1) = XN + YN
          PRINT 601, N,XN,YN,YN,ERROR
   20  CONTINUE
C
C   ** BEGIN ITERATION
       DO 50 N=4,NSTEPS
          YN = YN + (H/24.)*(55.*F(4)-59.*F(3)+37.*F(2)-9.*F(1))
          XN = XBEGIN + N*H
          F(1) = F(2)
          F(2) = F(3)
          F(3) = F(4)
          F(4) = XN + YN
          YOFXN = SOLN(XN)
          ERROR = YN - YOFXN
          PRINT 601, N,XN,YN,YOFXN,ERROR
   50  CONTINUE
                                          STOP
       END
```

نتایج کامپیوتری برای مثال۵.۸

N	XN	YN	Y(XN)	ERROR
0	0.	0.	0.	0.
1	0.31250000E-01	0.49340725E-03	0.49340725E-03	0.
2	0.62500000E-01	0.19944459E-02	0.19944459E-02	0.
3	0.93750000E-01	0.45351386E-02	0.45351386E-02	0.
4	0.12500000E-00	0.81484411E-02	0.81484467E-02	−0.55879354E-08
5	0.15625000E-00	0.12868421E-01	0.12868434E-01	−0.12922101E-07
6	0.18750000E-00	0.18730211E-01	0.18730238E-01	−0.26309863E-07
7	0.21875000E-00	0.25770056E-01	0.25770098E-01	−0.41676685E-07
8	0.25000000E-00	0.34025350E-01	0.34025416E-01	−0.65192580E-07
9	0.28125000E-00	0.43534677E-01	0.43534756E-01	−0.78696758E-07
10	0.31250000E-00	0.54337843E-01	0.54337934E-01	−0.90803951E-07
11	0.34375000E-00	0.66475919E-01	0.66476032E-01	−0.11269003E-06
12	0.37500000E-00	0.79991280E-01	0.79991400E-01	−0.12014061E-06
13	0.40625000E-00	0.94927646E-01	0.94927788E-01	−0.14156103E-06

نتایج کامپیوتری برای مثال ۵۰۸(ادامه)

N	XN	YN	Y(XN)	ERROR
14	0.43750000E-00	0.11133012E-00	0.11133029E-00	−0.16111881E-06
15	0.46875000E-00	0.12924525E-00	0.12924545E-00	−0.19185245E-06
16	0.50000000E 00	0.14872105E-00	0.14872126E-00	−0.21234155E-06
17	0.53125000E 00	0.16980705E-00	0.16980730E-00	−0.24400651E-06
18	0.56250000E 00	0.19255438E-00	0.19255446E-00	−0.26822090E-06
19	0.59375000E 00	0.21701577E-00	0.21701607E-00	−0.29988587E-06
20	0.62500000E 00	0.24324562E-00	0.24324594E-00	−0.31664968E-06
21	0.65625000E 00	0.27130008E-00	0.27130044E-00	−0.34645200E-06
22	0.68750000E 00	0.30123707E-00	0.30123746E-00	−0.39115548E-06
23	0.71875000E 00	0.33311634E-00	0.33311677E-00	−0.42840838E-06
24	0.75000000E 00	0.36699954E-00	0.36700001E-00	−0.46566129E-06
25	0.78125000E 00	0.40295030E-00	0.40295079E-00	−0.49173832E-06
26	0.81250000E 00	0.44103424E-00	0.44103476E-00	−0.52526593E-06
27	0.84375000E 00	0.48131907E-00	0.48131964E-00	−0.56624413E-06
28	0.87500000E 00	0.52387466E 00	0.52387527E 00	−0.61094761E-06
29	0.90625000E 00	0.56877308E 00	0.56877375E 00	−0.66310167E-06
30	0.93750000E 00	0.61608872E 00	0.61608934E 00	−0.71525574E-06
31	0.96875000E 00	0.66589829E 00	0.66589907E 00	−0.77486038E-06
32	0.09999999E 01	0.71828098E 00	0.71828181E 00	−0.82701445E-06

تمرین

۱-۷۰۸ بـا استفاده از فرمـول (۴۵۰۸ الف)، ضرایب γ_k، $(k=1,\ldots,4)$ γ_k را در فرمول ادمز-بشفورت (۴۵۰۸) به‌دست آورید.

۲-۷۰۸ درفرمول (۴۵۰۸)، m را برابر با ۴ قرار دهید و مانند فرمول (۴۷۰۸)، فرمول ادمز-بشفورت متناظر با این مقدار m را برحسب عرضها به‌دست آورید.

۳-۷۰۸ فرمول میلن[1] (۵۱۰۸) و جملهٔ خطای متناظر آن را به‌دست آورید.

۴-۷۰۸ برنامه‌ای بنویسیدکه درآن فرمول میلن برای انتگرالگیری یك معادلهٔ دیفرانسیل به‌ازای نقاط با فواصل‌مساوی استفاده شده باشد. فرض کنیدکه سه مقدارآغازین‌اول‌معلوم باشند.

۵-۷۰۸ معادلهٔ مثال۵۰۸ را با استفاده ازبرنامهٔ میلن، وبه‌ازای $h=1/32$، حل، و نتایج به‌دست‌آمده را با نتایج داده شده درمثال ۵۰۸ مقایسه کنید.

۶-۷۰۸ با استفاده از روش ادمز-بشفورت (۴۷۰۸) معادلهٔ $xy'=x-y$، $y(2)=2$را از $x=2$ تا $x=3$ به‌ازای $h=0.05$ حل کنید. مقادیرآغازین را از جواب دقیق زیر به‌دست آورید

1. Milne

$$y(x)=\frac{x}{۲}+\frac{۲}{x}$$

۷-۷.۸ با استفاده از روش ادمز-بشفورت (۴۷.۸) به‌ازای $h=۱/۶۴$ و $h=۱/۱۲۸$، معادلهٔ $y'+y=e^{-x}$ را از $x=۰$ تا $x=۱$ حل کنید. دقت نتایج حاصل را براورد نمایید. مقادیر آغازین ازجواب دقیق $y=xe^{-x}$ به‌دست می‌آیند.

۸-۷.۸ با استفاده ازفرمول (۴۹.۸) به‌ازای $m=۲$ (نه $m=۳$) وفرمول خطا در درونیابی بسجمله‌ای (۱۸.۲)، فرمولهای (۵۱.۸) را به‌دست آورید. برای انجام این کار، بحث انجام شده در آغاز بخش ۲.۷ مفید خواهد بود.

۹-۷.۸ با بسط $y_{n+۱}$ و $y_{n-۱}$ تا جملات مرتبهٔ سوم پیرامون $x=x_n$، فرمول (۵۰.۸) را تحقیق و فرض کنید که مقادیر آغازین دقیق هستند.

۸.۸ روشهای پیشگو-تصحیحی[1]

روشهای چندمرحله‌ای قسمت ۷.۸ با استفاده از بسجمله‌ایهایی به‌دست آمده بودند که در نقطهٔ x_n و نقاط ماقبل x_n درونیابی شده بودند. این فرمولها اغلب به‌فرمولهای نوع باز[2] معروفند. فرمولهای نوع بسته[3] را می‌توان براساس درونیابی بسجمله‌ای درنقطهٔ $x_{n+۱}$ و نیز x_n و نقاط ماقبل x_n به‌دست آورد. ساده‌ترین فرمول این نوع از تقریب‌زدن انتگرال موجود در رابطهٔ (۴۳.۸) به‌وسیلهٔ فرمول ذوزنقه‌ای (۲۶.۷) به‌دست می‌آید. با انجام این عمل فرمول زیرحاصل می‌شود

$$y_{n+۱}=y_n+\frac{h}{۲}[f(x_n, y_n)+f(x_{n+۱}, y_{n+۱})] \qquad n=۰, ۱, \cdots \qquad (۵۲.۸)$$

خطای این فرمول برابراست با $-(h^۳/۱۲)y'''$، و لذا بیانگر بهبودی است در روش اویلر. ولی (۵۲.۸) یك معادلهٔ ضمنی برحسب $y_{n+۱}$ است، زیرا $y_{n+۱}$ به‌عنوان یك شناسه در سمت راست ظاهر می‌شود.

درحالت کلی اگر $f(x, y)$ تابعی غیرخطی باشد، نمی‌توانیم معادلهٔ (۵۲.۸) را نسبت به $y_{n+۱}$ دقیقاً حل کنیم. ولی می‌توانیم سعی کنیم که $y_{n+۱}$ را از راه بارستی به‌دست آوریم. بنابراین باثابت نگه‌داشتن x_n، نخستین تقریب $y^{(۰)}_{n+۱}$ را به‌وسیلهٔ فرمول اویلر

$$y^{(۰)}_{n+۱}=y_n+hf(x_n, y_n) \qquad (۵۳.۸)$$

برای $y_{n+۱}$ به‌دست می‌آوریم، سپس $f(x_{n+۱}, y^{(۰)}_{n+۱})$ را محاسبه می‌کنیم و درسمت راست

1. predictor-corrector 2. open type 3. closed type

معادلهٔ (۵۲۰۸) قرار می‌دهیم تا تقریب زیر را به‌دست آوریم

$$y_{n+1}^{(1)} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(0)})]$$

بعد $f(x_{n+1}, y_{n+1}^{(1)})$ را محاسبه می‌کنیم و با استفادهٔ مجدد از معادلهٔ (۵۲۰۸)، تقریب دیگری به‌دست می‌آوریم. در حالت کلی، بارست با رابطهٔ

$$y_{n+1}^{(k)} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(k-1)})] \qquad k = 1, 2, \cdots \quad (۵۴۰۸)$$

تعریف می‌شود. بارست، زمانی به‌پایان می‌رسد که دو بارست متوالی، با دقت مطلوب متوافق باشند. این بارست، که برای به‌دست آوردن مقادیر اصلاح شدهٔ y_{n+1} در نقطهٔ ثابت x_{n+1} به‌کار می‌رود، اغلب **بارست داخلی**[1] نامیده می‌شود تا با معادلهٔ (۵۲۰۸)، که برای تولید مقادیر y_n به‌ازای $n = 0, 1, \cdots$ به‌کار می‌رود، تفاوت داشته باشد. این شیوهٔ عمل را در الگوریتم ۴۰۸ خلاصه می‌کنیم.

الگوریتم ۴۰۸: یک روش پیشگو- تصحیحی مرتبه‌دوم برای معادلهٔ دیفرانسیل $y' = f(x, y)$، $y(x_0) = y_0$ با h داده شده و $x_n = x_0 + nh$ به‌ازای جمیع مقادیر ثابت $n = 0, 1, \cdots$:

۱. با استفاده از فرمول (۵۳۰۸)، $y_{n+1}^{(0)}$ را محاسبه کنید.

۲. با استفاده از فرمول (۵۴۰۸)، $y_{n+1}^{(k)}$، $k = 1, 2, \cdots$، را محاسبه کنید، و این عمل را برمبنای k بارست انجام دهید تا به‌ازای مقدار معینی مانند ε رابطهٔ زیر

$$\frac{|y_{n+1}^{(k)} - y_{n+1}^{(k-1)}|}{|y_{n+1}^{(k)}|} < \varepsilon$$

برقرار شود.

برای مشخص کردن ε در الگوریتم ۴۰۸، باید به‌خاطر داشته باشیم که دقت موردنظر در هر مرحله به‌وسیلهٔ میزان خطای موجود در فرمول اساسی (۵۲۰۸) و نمو h محدود می‌شود.

برای تطبیق این الگوریتم با حل یک مسئلهٔ خاص، باید نکات زیر را مشخص کنیم، (الف) تعداد مراحل لازم، N. (ب) حداکثر تعداد بارستهای داخلی، K. (پ) چه باید کرد، درصورت بیشتر شدن k از K.

چنین متداول است که یک فرمول صریح مانند فرمول اویلر را فرمول بازگویند، و

---

1. inner iteration

یك فرمول ضمنی مانند (۵۲.۸) را فرمول بسته. هنگامی که این فرمولها با هم به‌صورت یك‌جفت فرمول به‌کار برده شوند فرمول نوع باز، **پیشگو** نیز نامیده می‌شود، در حالی که فرمول نوع بسته یك **تصحیحی** خوانده می‌شود. معمولا یك فرمول تصحیحی دقیقتر از یك فرمول پیشگو است، حتی زمانی که مرتبهٔ خطای گسسته‌سازی درهردو یکی باشد و این امر عمدتاً برای این‌است که ضریب عبارت خطا کوچکتر است. در رابطه بافرمولهای تصحیحی، طبیعتاً دوسؤال مطرح می‌شود. اول اینکه «درچه شرایطی بارستهای داخلی روی k، همگرا خواهند شد؟»، دوم اینکه «چند بارست برای به‌دست آوردن دقت مطلوب، لازم است؟» جواب سؤال دوم به‌چندین‌عامل بستگی دارد. ولی اگرفرمولهای پیشگو و تصحیحی از مرتبهٔ واحدی باشند، تجربه نشان داده‌است که اگر نمو h به‌طور مناسب انتخاب شده باشد، یك یا دوبار استفاده از فرمول تصحیحی کافی‌است. اگرمعلوم شودکه این تعداد استفاده‌از فرمول تصحیحی‌کافی نیست، بهتر است که به‌جای ادامهٔ بارستها، نمو h کاهش داده شود. پاسخ سؤال اول درقضیهٔ ۲.۸ مستتر است.

قضیهٔ ۲.۸ اگر $f(x,y)$ و $\partial f/\partial y$ برحسب x و y دربازهٔ بسته $[a,b]$ پیوسته باشند، در صورتی که h چنان‌کوچك انتخاب شده باشدکه به‌ازای $x=x_n$ وکلیهٔ yها با شرط $|y-y_{n+1}|\leqslant|y_{n+1}^{(0)}-y_{n+1}|$، رابطهٔ

$$\left|\frac{\partial f}{\partial y}\right|h<2 \qquad (۵۵.۸)$$

برقرارباشد، آنگاه‌بارستهای داخلی که به‌وسیلهٔ(۵۴.۸) تعریف شده‌اند همگرا خواهندبود.

برای اثبات این‌مطلب، اولا ملاحظه می‌کنیم که درفرمول‌بارست (۵۴.۸)، x_n ثابت است. بنابراین، اگرقراردهیم $Y^{(k)}:=y_{n+1}^{(k)}$ آنگاه، می‌توانیم‌فرمول(۵۴.۸)را به‌صورت

$$Y^{(k)}=F(Y^{(k-1)})$$

بنویسیم که درآن داریم

$$F(Y)=\frac{h}{2}f(x_{n+1},Y)+C$$

و در این رابطه C به n بستگی دارد، نه به Y. می‌توان این مطلب را به‌مثابهٔ موردی از بارست نقطهٔ ثابت‌که درقسمت ۳.۳ بررسی شد درنظرگرفت. در فرع قضیهٔ ۱.۳، ثابت کردیم که چنین بارستی‌همگرا خواهد بود، به‌شرط آنکه $F'(Y)$ پیوسته باشد و به‌ازای جمیع مقادیر Y باشرط $|Y-y_{n+1}|\leqslant|Y^{(0)}-y_{n+1}|$، ($y_{n+1}$ نقطهٔ‌ثابت $F(Y)$ است)، در رابطهٔ

$$|F'(Y)|<1$$

صدق کند. از آنجاکه داریم $F'(Y)=(h/2)\,\partial f/\partial y$ و چون $\partial f/\partial y$ کراندار و بنابر فرض ناصفر است، فرمول بارست (۵۴.۸) همگرا خواهد شد اگر

$$F'(Y)=\left|\frac{h}{2}\frac{\partial f}{\partial y}\right|<1$$

یعنی اگر

$$h<\frac{2}{|\partial f/\partial y|}$$

و چون تساوی $F'(Y)=(h/2)\partial f/\partial y$ برقرار است، قضیه ثابت می‌شود.

□ **مثال ۶.۸:** معادلهٔ

$$y'=x-\frac{1}{y}\qquad y(0)=1$$

را از $x=0$ تا $x=0.2$، با استفاده از الگوریتم ۴.۸ و به‌ازای $h=0.1$ حل کنید

از آنجاکه خطا در فرمول (۵۴.۸) برابر است با $-(h^3/12)y'''$ و چون با مشتقگیری از معادلهٔ بالا خواهیم داشت $y'''(0)\approx -2$، بنابراین خطا تقریباً برابر با 0.0002 خواهد بود. بنابراین نمی‌توانیم درنتایج، دقتی بیش از سه رقم اعشار انتظارداشته باشیم

مرحلهٔ ۱

باروش اویلر: $y_1^{(0)}=0.9$

با فرمول (۵۴.۸): $y_1^{(1)}=0.8994$

$y_1^{(2)}=0.8994$

وچون $y_1^{(1)}$ و $y_1^{(2)}$ تا چهار رقم اعشار متوافق‌اند، این جواب را می‌پذیریم ومقدار y_1' را برابر با $y_1'=f(x_1,y_1)=-1.0118$، به‌دست می‌آوریم

مرحلهٔ ۲، با روش اویلرداریم

$$y_2^{(0)}=0.8994+0.1(-1.0118)=0.7982$$

و با فرمول (۵۴.۸) داریم

$$y_2^{(1)}=0.8994+0.05\left[-1.1118+\left(0.2-\frac{1}{0.7982}\right)\right]=0.7962$$

$$y_2^{(2)} = ۰٫۸۹۹۴ + ۰٫۰۵\left[-۱٫۰۱۱۸ + \left(۰٫۲ - \frac{۱}{۰٫۷۹۶۲}\right)\right] = ۰٫۷۹۶۰$$

$$y_2^{(3)} = ۰٫۷۹۶۰$$

جواب $y_2 = ۰٫۷۹۶۰$ را می‌پذیریم و y_2' را محاسبه نموده و به‌مرحلهٔ بعدی می‌پردازیم.

درجریان محاسبات، انتظار داریم که دقت به‌تدریج کاهش یابد. ظاهراً در اینجا به‌ازای $h = ۰٫۱$ به‌دو یا سه‌بار استفاده از فرمول «تصحیحی» نیاز داریم. علت آن عمدتاً این‌است که مرتبهٔ پیشگویی‌ای که مورد استفاده واقع شده از مرتبهٔ تصحیحی کمتر است.

برای تحقیق همگرایی بارستهای داخلی به‌ازای $h = ۰٫۱$ در این مثال $\partial f/\partial y = ۱/y^۲$ را محاسبه می‌کنیم، و به‌موجب قضیهٔ ۲.۸ می‌خواهیم که h کمتر از $۲y^۲$ باشد. جواب y را نداریم، اما از مراحل بالا آشکاراست که در بازهٔ $[۰٫۲, ۰]$، داریم $y > ۰٫۷$. بنابراین اگر داشته باشیم $h < ۲(۰٫۷)^۲ = ۰٫۹۸$، بارستهای داخلی همگرا می‌شوند. □

تمرین

۱-۸.۸ برای معادلهٔ خاص $y' = Ay$ و $y(۰) = ۱$، نشان دهید که اگر داشته باشیم $|Ah/۲| < ۱$، فرمول تصحیحی ذوزنقه‌ای (۵۲.۸) به‌معادلهٔ تفاضلی منجر می‌شود که جواب آن برابراست با

$$y_n = \left[\frac{(۱ + Ah/۲)}{(۱ - Ah/۲)}\right]^n$$

۲-۸.۸ برای جوابی که در تمرین ۸.۸-۱ به‌دست آمد، نشان دهید که به‌ازای مقدار ثابت $x = x_n = nh$، داریم

$$\lim_{h \to ۰} y_n = e^{Ax_n}$$

۳-۸.۸ معادلهٔ $y' = x^۲ + y$ و $y(۰) = ۱$ را از $x = ۰$ تا $x = ۰٫۵$ با استفاده از روش اویلر به‌عنوان پیشگو و فرمول (۵۴.۸) به‌عنوان تصحیحی، حل کنید. نمو h را طوری تعیین کنید که در $x = ۰٫۵$ دقتی تا چهار رقم اعشار به‌دست آید. عملیات را با نمو $h = ۰٫۰۵$ آغاز کنید.

۹.۸ روش ادمز-مولتن[1]

فرمول تصحیحی از مرتبهٔ بالاتر را می‌توان با استفاده از یك بسجمله‌ای که عمل درونیابی

1. Adams-Moulton

را درنقاط x_{n+1}، x_n، ... ، x_{n-m} (به‌ازای عدد صحیح $m > 0$) انجام می‌دهد، به‌دست آورد. فرمول پسرونیوتن که عمل‌درونیابی را در $m+2$ نقطه برحسب $s=(x-x_n)/h$ انجام می‌دهد چنین‌است

$$p_{m+1}(s)=\sum_{k=0}^{m+1}(-1)^k\binom{-s}{k}\Delta^k f_{n+1-k} \qquad (۵۶.۸)$$

این‌تفاضلها بر اساس مقادیر f_{n+1}، f_n، ... ، f_{n-m} به دست آمده‌اند. اگر از فرمول (۵۶.۸) از x_n تا x_{n+1} انتگرال بگیریم و فرمول (۴۳.۸) را به‌کار بریم، خواهیم داشت

$$y_{n+1}=y_n+h(\gamma'_0 f_{n+1}+\gamma'_1\Delta f_n+\cdots+\gamma'_{m+1}\Delta^{m+1}f_{n-m}) \qquad (۵۷.۸)$$

که درآن

$$\gamma'_k=(-1)^k\int_0^1\binom{-s}{k}ds \qquad k=0,1,\cdots,m+1$$

چند مقدار اول γ'_k عبارت‌اند از

$$\gamma'_0=1 \qquad \gamma'_1=-\frac{1}{2} \qquad \gamma'_2=-\frac{1}{12} \qquad \gamma'_3=-\frac{1}{24} \qquad \gamma'_4=-\frac{19}{720}$$

خطا در فرمول (۵۷.۸) که از خطا دربسجمله‌ای درونیاب ناشی‌شده چنین‌است

$$E=\gamma'_{m+2}h^{m+3}y^{(m+3)}(\xi) \qquad (۵۸.۸)$$

حالت $m=2$ حالتی‌است که اغلب به‌کاربرده می‌شود. اگر در (۵۷.۸) به‌ازای $m=2$، تفاضلها برحسب عرضها بیان شوند، خواهیم داشت

$$\boxed{y_{n+1}=y_n+\frac{h}{24}(9f_{n+1}+19f_n-5f_{n-1}+f_{n-2})} \qquad (۵۹.۸)$$

و خطا برابراست با

$$\boxed{E_{AM}=-\frac{19}{720}h^5y^{(5)}(\xi)} \qquad (۶۰.۸)$$

فرمول (۵۷.۸) به‌**فرمول ادمز-مولتن** معروف‌است. واضح‌است‌که فرمول ادمز-مولتن از مرتبهٔ‌چهارم(۵۹.۸)، یك فرمول مصحح‌ازنوع بسته‌است، زیرا $f_{n+1}=f(x_{n+1},y_{n+1})$ متضمن کمیت مجهول y_{n+1} است. می‌توان نشان دادکه اگر h به‌اندازهٔ‌کافی کوچك وشرط

$|\partial f/\partial y| < ۹h/۲۴$ برقرار باشد، آنگاه فرایند بارست که براساس فرمول (۵۹.۸) به‌دست آمده، همگرا خواهد شد. یك فرمول پیشگوی ساده، فرمول مرتبهٔ چهارم ادمز-بشفورت (۴۷.۸) است که برای استفاده با این فرمول مصحح مناسب خواهد بود. در این حالت، پیشگو ومصحح هم‌مرتبه هستند. اگر h به‌طور مناسب انتخاب شود، آنگاه با یکبار استفاده از فرمول مصحح، بهبودی مهمی در دقت حاصل می‌شود.

مشخصات یك روش مرتبهٔ چهارم پیشگو-تصحیحی در الگوریتم ۵.۸ داده‌شده‌است.

الگوریتم ۵.۸: روش پیشگو- تصحیحی ادمز-مولتن برای معادلهٔ دیفرانسیل $y' = f(x, y)$ با h ثابت و $x_n = x_0 + nh$ و مفروضات (y_0, f_0)، (y_1, f_1)، (y_2, f_2)، (y_3, f_3) به‌ازای مقادیر ثابت $n = ۳, ۴, \ldots$

۱. با استفاده از فرمول زیر، $y^{(0)}_{n+1}$ را محاسبه می‌کنیم

$$y^{(0)}_{n+1} = y_n + \frac{h}{۲۴}(۵۵f_n - ۵۹f_{n-1} + ۳۷f_{n-2} - ۹f_{n-3})$$

۲. $f^{(0)}_{n+1} = f(x_{n+1}, y^{(0)}_{n+1})$ را محاسبه می‌کنیم.

۳. مقدار

$$y^{(k)}_{n+1} = y_n + \frac{h}{۲۴}[۹f(x_{n+1}, y^{(k-1)}_{n+1}) + ۱۹f_n - ۵f_{n-1} + f_{n-2}]$$

$k = ۱, ۲, \ldots$

را محاسبه می‌کنیم.

۴. عمل بارستن روی k را تا حصول رابطهٔ

$$\frac{|y^{(k)}_{n+1} - y^{(k-1)}_{n+1}|}{|y^{(k)}_{n+1}|} < \varepsilon$$

به‌ازای ε مفروض ادامه می‌دهیم.

هنوز این الگوریتم کامل نیست مگر آنکه مشخص نماییم که درصورت نبود همگرایی درمرحلهٔ چهارم چه باید بکنیم. زیربرنامه‌ای مانند DVERK از مشخصهٔ کاملتری برای زیربرنامه‌ای که درهمه‌جا منظور مارا درحل معادلات دیفرانسیل برآورده سازد، برخوردار است.

فرمول تصحیحی علاوه برایجاد دقت بهتر، کار مفید دیگری نیز انجام می‌دهد. این فرمول براوردی از خطای گسسته‌سازی محلی به‌دست می‌دهد که برای تصمیم‌گیری درمورد اینکه آیا نمو h برای ایجاد دقت خواسته شده کافی هست یا نه، به‌کار می‌رود. برای آنکه

این روش براورد خطا را برای روش پیشگو-تصحیحی که مشتمل برفرمولهای ادمز-بشفورت و ادمز-مولتن است آزمایش کنیم، براورد خطای موضعی را برای هریك ازآنها می نویسیم

$$E_{AB}=\frac{۲۵۱}{۷۲۰}h^{۵}y^{۷}(\xi_{۱})$$

$$E_{AM}=-\frac{۱۹}{۷۲۰}h^{۵}y^{۷}(\xi_{۲}) \qquad (۶۱.۸)$$

گیریم $y^{(۰)}_{n+۱}$ معرف مقدار $y_{n+۱}$ حاصل از فرمول (۴۷.۸)، و $y^{(۱)}_{n+۱}$ مقدارحاصل ازیکبار استفاده از الگوریتم ۵.۸ باشد. اگرفرض شودکه مقادیر f درتمام نقاط تا x_n ومنجمله خود آن دقیق باشند و اگر $y(x_{n+۱})$ معرف مقدار دقیق y در نقطهٔ $x_{n+۱}$ باشد، آنگاه از رابطهٔ (۶۱.۸)، براوردهای خطا را بدست خواهیم آورد:

$$y(x_{n+۱})-y^{(۰)}_{n+۱}=\frac{۲۵۱}{۷۲۰}h^{۵}y^{۷}(\xi_{۱}) \qquad (۶۲.۸\text{الف})$$

$$y(x_{n+۱})-y^{(۱)}_{n+۱}=-\frac{۱۹}{۷۲۰}h^{۵}y^{۷}(\xi_{۲}) \qquad (۶۲.۸\text{ب})$$

درحالت کلی، $\xi_{۲}\neq\xi_{۱}$. ولی اگر دربازهٔ موردنظر، فرض شودکه $y^{۷}(x)$ تقریباً ثابت است، آنگاه با تفریق (۶۲.۸ب) از (۶۲.۸الف)، براورد زیررا برای $y^{۷}$ بدست خواهیم آورد:

$$h^{۵}y^{۷}=\frac{۷۲۰}{۲۷۰}(y^{(۱)}_{n+۱}-y^{(۰)}_{n+۱})$$

با گذاردن این مقدار در رابطهٔ (۶۲.۸ ب)، خواهیم داشت

$$y(x_{n+۱})-y^{(۱)}_{n+۱}=-\frac{۱۹}{۲۷۰}(y^{(۱)}_{n+۱}-y^{(۰)}_{n+۱})$$

$$\approx-\frac{۱}{۱۴}(y^{(۱)}_{n+۱}-y^{(۰)}_{n+۱})=D_{n+۱} \qquad (۶۳.۸)$$

بنابراین خطا درمقدارتصحیح شده، تقریباً برابر با $-\frac{۱}{۱۴}$ تفاضل بین مقدار تصحیح شده و مقدار پیشگو شده‌است.

همان گونه‌که قبلااشاره شد، بهتراست‌که فرمول تصحیحی را تنها یکبار به‌کاربرد. اگر میزان دقت حاصل از رابطهٔ (۶۳.۸) کافی نباشد، بهتراست که به‌جای استفاده ازفرمول مذکور برای بیش از یکبار، نمو راکاهش‌داد.

دریك برنامهٔ همه منظوره برای حل معادلات دیفرانسیل، از براورد خطا به‌طریق‌زیر استفاده می‌شود: فرض کنید که بخواهیم مانند رابطهٔ (۴۱.۸)، خطای موضعی دریك مرحلهٔ واحد* کراندار باشد به‌طوری که داشته باشیم

$$E_1 \leqslant \frac{|D_{n+1}|}{h} \leqslant E_2$$

و مقادیر آغازین نیز معین شده باشند. مراحل انجام کار چنین‌است.

۱. با استفاده از فرمول (۴۷.۸)، $y^{(0)}_{n+1}$ را به‌دست می‌آوریم. $f^{(0)}_{n+1}$ را محاسبه می‌کنیم.

۲. با استفاده از فرمول (۵۹.۸)، $y^{(1)}_{n+1}$ را به‌دست می‌آوریم. $f^{(1)}_{n+1}$ را محاسبه می‌کنیم.

۳. از روی رابطهٔ (۶۳.۸)، $|D_{n+1}|$ را محاسبه می‌کنیم.

۴. اگر $E_1 \leqslant |D_{n+1}|/h \leqslant E_2$، با استفاده از همان مقدار h، مرحلهٔ انتگرالگیری بعدی را انجام می‌دهیم.

۵. اگر $|D_{n+1}|/h > E_2$، آنگاه نمو h بسیار بزرگ‌است و می‌باید کاهش‌یابد. دراین حالت، معمول‌است که $h/2$ را به‌جای h می‌گذارند و چهارمقدار آغازین‌رامجدداً محاسبه می‌کنند و سپس به‌مرحلهٔ ۱ بازمی‌گردند.

۶. اگر $|D_{n+1}|/h < E_1$، آنگاه دقت به‌دست‌آمده بیش از اندازهٔ لازم است. ازاین‌رو باقراردادن $2h$ به‌جای h در وقت مصرفی کامپیوترمی‌توانیم صرفه‌جویی کنیم ودر بازه‌ای به‌طول $2h$، چهارمقدار آغازین جدید را مجدداً محاسبه کنیم وبه‌مرحلهٔ ۱ برگردیم.

چنانچه دربالا ذکر شد در استفاده از روشهای پیشگو-تصحیحی با نمو متغیر، لازم است که اولا، در ابتدای‌کار روشی برای به‌دست‌آوردن مقادیر آغازین لازم داشته باشیم؛ ثانیاً، روشی برای به‌دست‌آوردن‌مقادیرلازم yدرمراحل‌نصف را، وقتی بازه نصف‌می‌شود، داشته باشیم؛ ثالثاً، روشی‌برای به‌دست‌آوردن مقادیرلازمy برای هنگامی‌که بازه دوبرابر می‌شود داشته باشیم. برای هریك از این سه وضعیت، فرمولهای خاصی می‌توانند به‌دست آیند. این فرمولها پیچیدگی یك برنامه را سخت افزایش می‌دهند. اما یك ترکیب نسبتاً آرمانی از این فرمولها، به کارگیری روش مرتبهٔ چهارم، رونگه-کوتا (۳۷.۸)، همراه بایك زوج پیشگو-تصحیحی مرتبهٔ چهارم، از قبیل فرمولهای (۴۷.۸) و (۵۹.۸) است. بنابراین روش رونگه-کوتا می‌تواند در ابتدای‌کار برای شروع مسئله، هنگام نصف کردن بازه و یا دوبرابرکردن آن به‌کار رود، در حالی‌که وقتی نمو ثابت است، زوج پیشگو-تصحیحی می‌تواند برای ادامهٔ طبیعی‌آن به‌کار رود.

پیش از اتمام این قسمت، باید یادآور شویم که فرمولهای پیشگو-تصحیحی زیاد

---

* منظور از یك مرحلهٔ واحد، یك مرحلهٔ کامل به‌ازای یك h معین‌است.ـم.

دیگری نیز وجود دارند، و به‌ویژه فرمولهای زیر که منسوب به‌میلن[1] هستند، غالباً به‌کار برده می‌شوند

$$y_{n+1}^{(0)} = y_{n-3} + \frac{4h}{3}(2f_n - f_{n-1} + 2f_{n-2}) \qquad E_M^0 = \frac{28}{90}h^5 y^{(5)}(\xi_1)$$

(۶۴.۸الف)

$$y_{n+1}^{(1)} = y_{n-1} + \frac{h}{3}(f_{n+1}^{(0)} + 4f_n + f_{n-1}) \qquad E_M^1 = -\frac{1}{90}h^5 y^{(5)}(\xi_2)$$

(۶۴.۸ب)

معادلهٔ (۶۴.۸ الف)، درقسمت ۶.۸ به‌دست آمده بود و معادلهٔ (۶۴.۸ ب) براساس قاعدهٔ سیمپسن برای انتگرالگیری عددی مبتنی است. همان گونه که برای فرمولهای ادمز-مولتن عمل کردیم، می‌توانیم نشان دهیم که براورد خطای موضعی با رابطهٔ زیر معین می‌شود

$$D_{n+1} = -\frac{1}{29}(y_{n+1}^{(1)} - y_{n+1}^{(0)}) \qquad (۶۵.۸)$$

به‌نظر می‌رسد که براورد خطا در روش میلن تا اندازه‌ای مطلوبتر از روش ادمز-مولتن باشد، ولی چنانچه خواهیم دید دربرخی موارد رابطهٔ(۶۴.۸ب) دستخوش ناپایداری قرار می‌گیرد.

درحالی که برای انتگرالگیری از معادلات دیفرانسیل، درکتابهای مختلف روشهای زیادی ذکرشده‌اند، روش مرتبهٔ چهارم رونگه-کوتا و روشهای پیشگو-تصحیحی نظیر ادمز-مولتن یا میلن(۶۴.۸)، متداولترین روشها درایالات متحده هستند. اگرچه برای همهٔ مسائل، هیچ روشی کلا بهتر از روش دیگر نیست، ولی بجاست که درحالت کلی به‌مزایا و معایب هریك از این روشها اشاره کنیم.

امتیازمهم روشهای رونگه-کوتادرخود-آغازگر بودن آنهاست. علاوه بر آن پایدارند، دقت خوبی به‌دست می‌دهند، و به‌عنوان یك برنامهٔ کامپیوتری جای نسبتاً کمی از حافظه را اشغال می‌کنند. روشهای RK (رونگه -کوتا)ی استانده، خطاهای موضعی را براورد نمی‌کنند، لذا استفاده‌کننده راهی برای درك مناسب بودن یانبودن نمو h به‌کار رفته، ندارد. البته‌می‌توان روشهای کنترل نمو را، که درقسمت ۶.۸ بیان شد، به‌کاربرد ولی این‌امر ازنظر وقت‌کامپیوتری پرهزینه خواهد بود. اشکال عمدهٔ دوم آن‌است که روش رونگه-کوتا در مقایسه باروشهای پیشگو-تصحیحی‌مرتبهٔ چهارم که تنها به‌دوبارمحاسبهٔ تابع احتیاج دارند، درهرمرحله از انتگرالگیری به‌چهار بار محاسبهٔ تابع نیازدارد. دربعضی از مسائل زمان محاسبه برای روشهای رونگه-کوتا تقریباً دوبرابر است.

---

1. Milne

روشهای پیشگو-تصحیحی درهرمرحله به‌طورخودکار براورد خطا را به‌ما می‌دهند، و بدین‌طریق به برنامه اجازه می‌دهند که برای یك دقت مطلوب مقدار بهینه‌ای از h را انتخاب‌نماید. ضمناً روشهای مزبور سریع نیزهستند، زیرا درهرمرحله تنها به‌دوبارمحاسبهٔ تابع احتیاج دارند. از سوی دیگر، نوشتن زیربرنامه‌های پیشگو-تصحیحی بسیار پیچیده است و برای شروع و دوبرابرکردن و نصف‌کردن نمو h به‌تکنیکهای خاصی نیاز دارند و این روشها ممکن‌است دستخوش ناپایداری عددی قرارگیرند (قسمت ۱۱.۸ را ببینید).

سالهای زیادی بودکه روشهای رونگه-کوتا برای انجام کارهای همه‌منظوره تقریباً به‌طور انحصاری درامریکا به‌کار می‌رفت، و اخیراً روشهای پیشگو-تصحیحی مقبولیت یافته‌اند.

درچند سال گذشته، روشهای همه‌منظورهٔ پیچیده‌ای‌که درآنها هم مرتبه‌های متغیر و هم نموهای متغیر به‌کاربرده می‌شوند توسعه پیداکرده‌اند. روشهای ادمز، که قبلا شرح دادیم، از جملهٔ متداولترین روشهای مرتبه-متغیر و نمو-متغیر هستند. هدف از این روشها انتخاب مرتبهٔ مناسب ونمو مناسب به‌طور خودکار است، واین انتخابها مقدارکارلازم‌برای دستیابی به‌دقتی مشخص برای مسئله‌ای معین را حداقل می‌سازند. مزیت مهم دیگر این روشها خود-آغازگر بودن آنهاست، زیرا در شروع‌کار می‌توانند یك روش پایین-مرتبه را به‌کاربرند و زمانی‌که نمو تغییرمی‌کند برای جبران ازدست رفتن مقادیر، می‌توانند به‌آسانی‌اصلاح شوند. توضیح کامل درمورد زیربرنامهٔ DIFSUB که براساس روش‌ادمز از نوع مرتبه-متغیر ونمو-متغیر تنظیم‌شده‌درکتاب گیر[1] [۳۰،ص‌ص۱۶۷-۱۵۸] آمده‌است. زیربرنامه‌ای به‌نام DVOGER وجود داردکه آن نیز براساس روش گیر در برنامه‌های IMSL تنظیم شده وبرای اجرا روی جدیدترین‌کامپیوترها مورد استفاده قرارگرفته‌است.

□ **مثال۷.۸:** با استفاده‌ازفرمولهای پیشگو-تصحیحی ادمز-مولتن وبه‌ازای $h = 1/32$، مثال ۵.۸ را حل‌کنید، نتایج به‌دست‌آمده را با نتایج حاصله درآن مثال مقایسه‌کنید.

برنامه و نتایج به‌دست‌آمده از کامپیوتر در زیرداده شده‌اند. در این حالت، (مقدار تصحیح شدهٔ) x_n و y_n، براورد خطای موضعی D_n، و خطای واقعی e_n را مشخص کرده‌ایم. از مقایسهٔ این نتایج با نتایج به‌دست‌آمده درمثال ۵.۸، بهبودی قطعی دردقت را ملاحظه می‌کنیم، به‌ویژه زمانی‌که x به‌سمت ۱ نزدیك می‌شود، که دراین حالت نتایج تا هفت یا هشت رقم بامعنی صحیح‌اند. به‌نظر می‌رسدکه براورد خطای موضعی نسبتاً ثابت و به‌طور کلی از خطای واقعی e_n اندکی کمتراست. ولی با بررسی دقیقتر درمی‌یابیم‌که درمراحل ۵ تا ۱۳، نتایج تنها تا شش‌رقم با معنا صحیح‌اند. توضیح این امرچنین‌است‌که برای‌این مراحل، مقادیر y_n از لحاظ مرتبهٔ مقدارکوچکتر ازمقادیرآن درحالت $x \to 1$ می‌باشند. چون D_n یك آزمون خطای مطلق است، تعداد ارقام با معنای دقیق درنتیجه را مشخص نمی‌سازد. وقتی‌که محاسبات درحساب با ممیز شناورانجام می‌گیرند،این‌امری عادی‌است.

---

1. Gear

زمانی‌که با اعداد خیلی بزرگ یا خیلی کوچك در مقایسه بـا ۱ کار می‌شود، برای تعیین تعداد ارقام با معنی صحیح، آزمون نسبی نسبت به‌آزمون مطلق، شاخص بهتری‌است. برای مثال، آزمون خطای نسبی برای فرمـول ادمز-مولتن به‌جای فرمول (۶۳.۸) فرمول زیر خواهد بود

$$\bar{D}_{n+1} = \frac{1}{14} \frac{|y_{n+1}^{(0)} - y_{n+1}^{(1)}|}{|y_{n+1}^{(1)}|}$$

برنامهٔ فورترن برای مثال ۷.۸

```
C  ADAMS-MOULTON METHOD
       INTEGER I,N,NSTEPS
       REAL ERROR,F(4),H,XBEGIN,XN,YBEGIN,YN
C
       SOLN(X) = EXP(X) - 1. - X
C
C  ** INITIALIZE
       PRINT 600
  600 FORMAT('1ADAMS-MOULTON METHOD'/
     *'0',3X,'N',14X,'XN',15X,'YN',9X,'DN = YN - YNP',8X,'ERROR'/)
       NSTEPS = 32
       H = 1./NSTEPS
       YBEGIN = 0.
       XBEGIN = 0.
C
C  ** COMPUTE FIRST FOUR POINTS USING EXACT SOLUTION
       F(1) = XBEGIN + YBEGIN
       N = 0
       ERROR = 0.
       DIFF = 0.
       PRINT 601,N,XBEGIN,YBEGIN,DIFF,ERROR
  601 FORMAT('  ',I3,4X,4E17.8)
       DO 20 N=1,3
          XN = XBEGIN + N*H
          YN = SOLN(XN)
          F(N+1) = XN + YN
          PRINT 601, N,XN,YN,DIFF,ERROR
   20 CONTINUE
C
C  ** BEGIN ITERATION
       DO 50 N=4,NSTEPS
C            PREDICT USING ADAMS-BASHFORTH FORMULA
          YNPRED = YN + (H/24.)*(55.*F(4)-59.*F(3)+37.*F(2)-9.*F(1))
          XN = XBEGIN + N*H
          FNPRED = XN + YNPRED
C            CORRECT USING ADAMS-MOULTON FORMULA
          YN = YN + (H/24.)*(9.*FNPRED + 19.*F(4) - 5.*F(3) + F(2))
          DIFF = (YN - YNPRED)/14.
          F(1) = F(2)
          F(2) = F(3)
          F(3) = F(4)
          F(4) = XN + YN
          YOFXN = SOLN(XN)
          ERROR = YN - YOFXN
          PRINT 601, N,XN,YN,DIFF,ERROR
   50 CONTINUE
                                          STOP
       END
```

نتایج کامپیوتری برای مثال ۷.۸

N	XN	YN	DN	ERROR
0	0.	0.	0.	0.
1	0.31250000E-01	0.49340725E-03	0.	0.
2	0.62500000E-01	0.19944459E-02	0.	0.
3	0.93750000E-01	0.45351386E-02	0.	0.
4	0.12500000E-00	0.81484520E-02	0.78164571E-09	0.53551048E-08
5	0.15625000E-00	0.12868445E-01	0.90637643E-09	0.11408702E-07
6	0.18750000E-00	0.18730249E-01	0.88143028E-09	0.11175871E-07
7	0.21875000E-00	0.25770108E-01	0.91469178E-09	0.10011718E-07
8	0.25000000E-00	0.34025417E-01	0.93132257E-09	0.13969839E-08
9	0.28125000E-00	0.43534759E-01	0.96458407E-09	0.37252903E-08
10	0.31250000E-00	0.54337924E-01	0.99784564E-09	0.83819032E-08
11	0.34375000E-00	0.66476036E-01	0.99784564E-09	0.37252903E-08
12	0.37500000E-00	0.79991416E-01	0.10643686E-08	0.15832484E-07
13	0.40625000E-00	0.94927801E-01	0.11308917E-08	0.13969839E-07
14	0.43750000E-00	0.11133030E-00	0.11308917E-08	0.14901161E-07
15	0.46875000E-00	0.12924545E-00	0.10643686E-08	0.55879354E-08
16	0.50000000E 00	0.14872127E-00	0.11974147E-08	0.74505806E-08
17	0.53125000E 00	0.16980730E-00	0.13304609E-08	0.18626451E-08
18	0.56250000E 00	0.19255465E-00	0.13304609E-08	0.37252903E-08
19	0.59375000E 00	0.21701607E-00	0.13304609E-08	0.
20	0.62500000E 00	0.24324595E-00	0.13304609E-08	0.11175871E-07
21	0.65625000E 00	0.27130044E-00	0.13304609E-08	0.11175871E-07
22	0.68750000E 00	0.30123746E-00	0.13304609E-08	0.
23	0.71875000E 00	0.33311676E-00	0.13304609E-08	−0.37252903E-08
24	0.75000000E 00	0.36700001E-00	0.15965530E-08	−0.74505806E-08
25	0.78125000E 00	0.40295079E-00	0.15965530E-08	0.37252903E-08
26	0.81250000E 00	0.44103477E-00	0.15965530E-08	0.74505806E-08
27	0.84375000E 00	0.48131964E-00	0.18626451E-08	0.74505806E-08
28	0.87500000E 00	0.52387527E 00	0.15965530E-08	0.74505806E-08
29	0.90625000E 00	0.56877375E 00	0.21287372E-08	0.
30	0.93750000E 00	0.61608943E 00	0.21287372E-08	0.
31	0.96875000E 00	0.66589906E 00	0.21287372E-08	−0.74505806E-08
32	0.09999999E 01	0.71828180E 00	0.21287372E-08	−0.74505806E-08

□

تمرین

۱-۹.۸ نشان دهید که بازستی که با فرمول

$$y_{n+1}^{(k)} = y_n + \frac{h}{۲۴}[۹f(x_{n+1}, y_{n+1}^{(k-1)}) + ۱۹f_n - ۵f_{n-1} + f_{n-2}]$$

$k = ۱, ۲, \cdots$

x_n ثابت

تعریف می‌شود همگرا خواهد بود، به‌شرط آنکه داشته باشیم $|(۹h/۲۴)(\partial f/\partial y)| < ۱$.
(قسمت ۸.۸ را ببینید).

۲-۹۰۸ با استفاده از روابط (۵۷۰۸) و (۵۸۰۸)، فرمول خطا را برای روش ادمز-مولتن (۶۰۸) به‌دست آورید.

۳-۹۰۸ فرمول براورد خطای محلی (۶۵۰۸) را برای فرمولهای پیشگو-تصحیحی میلن (۶۴۰۸) به‌دست آورید.

۴-۹۰۸ با استفاده از فرمولهای پیشگو-تصحیحی معادلهٔ $y'=y+x^2$، $y(0)=1$، را از $x=0$ تا $x=2$ با $h=0.1$ حل کنید. مقادیر آغازین صحیح تا ۶ رقم اعشاری به‌قرار زیرند

$$y(0)=1.000000$$
$$y(0.1)=1.105513$$
$$y(0.2)=1.242428$$
$$y(0.3)=1.359576$$

D_{n+1} را حساب و خطا را در $x=2$ براورد کنید.

۱۰.۸* پایداری روشهای عددی

نخستین باری که کامپیوترها برای حل معادلات دیفرانسیل به‌طور وسیع به‌کار گرفته شدند، ملاحظه شد که خطاهای ناشی از برخی از فرمولهای متداول برای انتگرالگیری، مانند فرمولهای میلن (۶۴۰۸) درجواب، درمقایسه با خطایی که تنها از خطای گسسته‌سازی انتظار می‌رود، خیلی بیشترند. بعلاوه، زمانی که نمو کوچکتر می‌شد، این خطاها به‌ازای مقدار ثابتی از x، به‌جای آنکه کوچکترشوند، عملا بزرگتر می‌شدند. برای تجسم این رفتار، روشی را که درقسمت ۷.۸ به‌دست آورده بودیم، بررسی می کنیم

$$y_{n+1}=y_{n-1}+2hf_n \qquad (۶۶۰۸)$$

که درآن خطای گسسته‌سازی برابر با $\frac{1}{6}h^3y'''(\xi)$ است. انتظار می‌رفت که به ازای h ثابت، این‌روش نسبت به‌روش اویلر که دارای خطایی برابر با $O(h^2)$ است، نتایج دقیقتری به‌دست دهد. ولی مسئلهٔ سادهٔ

$$y'=-2y+1 \qquad y(0)=1 \qquad (۶۷۰۸)$$

را که جواب دقیق آن برابر است با $y=\frac{1}{2}+\frac{1}{2}e^{-2x}$ درنظر می گیریم.

نتایجی که در جدول ۱.۸ آورده شده‌اند توسط کامپیوتر و با به‌کارگیری نمو $h=1/32$ به‌دست آمده‌اند. ستون شامل مقادیر انتخابی است از x که دراین مقادیرجوابها چاپ می‌شوند، $Y(N)$ معرف جواب دقیق‌است، $Y1(N)$ معرف جوابی‌است که با روش

جدول ۱.۸

X(N)	Y(N)	Y1(N)	Y2(N)	E1(N)	E2(N)
0.	1.0000000	1.0000000	1.0000000	0.	0.
0.0312500	0.9697065	0.9687500	0.9697065	−0.0009565	−0.0000000
0.5000000	0.6839397	0.6780370	0.6840817	−0.0059027	0.0001420
1.0000000	0.5676676	0.5633943	0.5678247	−0.0042733	0.0001571
1.5000000	0.5248935	0.5225730	0.5251328	−0.0023205	0.0002392
2.0000000	0.5091578	0.5080376	0.5097007	−0.0011202	0.0005429
2.2500000	0.5055545	0.5047962	0.5064264	−0.0007583	0.0008719
2.5000000	0.5033690	0.5028620	0.5047904	−0.0005070	0.0014214
3.0000000	0.5012394	0.5010190	0.5050759	−0.0002203	0.0038365
3.5000000	0.5004559	0.5003628	0.5108669	−0.0000931	0.0104110
3.7500000	0.5002765	0.5002165	0.5174337	−0.0000601	0.0171571
3.7812500	0.5002598	0.5002029	0.4819995	−0.0000568	−0.0182603
3.8125000	0.5002440	0.5001903	0.5196837	−0.0000538	0.0194397
3.8437500	0.5002293	0.5001784	0.4795391	−0.0000509	−0.0206902
3.8750000	0.5002154	0.5001672	0.5222413	−0.0000482	0.0220260
3.9062500	0.5002023	0.5001568	0.4767589	−0.0000456	−0.0234434
3.9375000	0.5001901	0.5001470	0.5251465	−0.0000431	0.0249564
3.9687500	0.5001785	0.5001378	0.4736156	−0.0000408	−0.0265630
4.0000000	0.5001677	0.5001292	0.5284445	−0.0000386	0.0282768

اویلر به‌دست آمده‌است، Y۲(N) معرف جوابی ناشی از فرمول (۶۶.۸) و E۱(N)، E۲(N) خطاهای متناظر آنها. روش (۶۶.۸) به‌مقادیر آغازین y_0 و y_1 احتیاج دارد. مقدار دقیقی را که از جواب دقیق یعنی $y=\frac{1}{2}e^{-2x}+\frac{1}{2}$ به دست می‌آید، برای تعیین y_1 به‌کار می‌بریم.

ستونهای خطا نشان می‌دهند که E۲(N) در چند مرحلهٔ اول به‌طور قابل ملاحظه‌ای از E۱(N) کوچکتر است، اما بعد سریعاً افزایش می‌یابد، به‌طوری که در $x=2.25$، E۲(N) از E۱(N) بزرگتر می‌شود. به‌ازای $x \to 4$، جواب به‌طرف حالت ماندگار[1] $y=1/2$ نزدیک می‌شود. روش اویلر در عمل با خطای یکنوا نزولی به‌این جواب حالت ماندگار نزدیک می‌شود. در حالی که در روش (۶۶.۸) خطا به‌طور نمایی بزرگ می‌شود. بعلاوه، به‌طوری که چند مرحلهٔ آخر (که نتایج در هر مرحله از انتگرالگیری چاپ شده‌اند) نشان می‌دهند، خطاهای E۲(N) علامت نوسانی* دارند. بعد از $x=4$، Y۲(N) دارای ارقام دقیق بامعنای صحیحی نخواهد بود. پدیده‌ای که در این مثال نشان داده شد به **ناپایداری عددی** معروف است.

برای کمک به درک این رفتار، معادلهٔ تفاضلی (۶۶.۸) را عمیقتر بررسی می‌کنیم. در مورد مثالی که بررسی شد، داریم $f_n=-2y_n+1$، و از این رو معادلهٔ (۶۶.۸) به‌صورت زیر درمی‌آید

1. steady-state

* منظور آن است که به‌ترتیب یکبار مثبت و یکبار منفی می‌شوند.ـم.

$$y_{n+1}+4hy_n-y_{n-1}=2h \qquad y_0=1 \qquad (۶۸.۸)$$

می‌توانیم با استفاده از روشهای مذکور در قسمت ۲.۸، این معادلهٔ تفاضلی‌را صریحاً حل کنیم. جواب عمومی معادلهٔ (۶۸.۸) برابر است با

$$y_n=C_1\beta_1^n+C_2\beta_2^n+\frac{1}{2} \qquad (۶۹.۸)$$

که درآن β_1 و β_2 ریشه‌های معادلهٔ مشخصهٔ

$$\beta^2+4h\beta-1=0$$

هستند، این ریشه‌ها برابرند با

$$\beta_{1,2}=-2h\pm\sqrt{1+4h^2}$$

اگر $\sqrt{1+4h^2}$ را به‌سری تیلر تـا جملات خطی بسط دهیم، این ریشه‌ها را می‌توانیم به‌شکل زیر بیان کنیم

$$\beta_1=1-2h+\mathcal{O}(h^2)$$

$$\beta_2=-(1+2h)+\mathcal{O}(h^2)$$

با قراردادن این مقادیر درمعادلهٔ (۶۹.۸) خواهیم داشت

$$y_n=C_1(1-2h+\mathcal{O}(h^2))^n+C_2(-1)^n(1+2h+\mathcal{O}(h^2))^n+\frac{1}{2} \qquad (۷۰.۸)$$

درمبحث حساب دیفرانسیل و انتگرال نشان داده‌ایم که

$$\lim_{\varepsilon\to 0}(1+\varepsilon)^{1/\varepsilon}=e$$

با استفاده از این‌حد و رابطهٔ $n=x_n/h$، به‌ازای x_n ثابت، نتیجه می‌شود

$$\lim_{h\to 0}(1+2h)^n=\lim_{h\to 0}(1+2h)^{(1/2h)(2x_n)}=e^{2x_n}$$

و نیزخواهیم داشت

$$\lim_{h\to 0}(1-2h)^n=e^{-2x_n}$$

از این‌رو به‌ازای $h\to 0$، جواب (۷۰.۸) به‌جواب

$$y_n=\left(C_1e^{-2x_n}+\frac{1}{2}\right)+C_2(-1)^ne^{2x_n} \qquad (۷۱.۸)$$

نزدیك می‌شود. بنابراین جملهٔ اول به‌سمت جواب حقیقی معادلهٔ دیفرانسیل میل می‌کند. جملهٔ دوم اضافی(غیرمربوط)است و تنها ناشی ازاین امراست که یك معادلهٔ تفاضلی مرتبهٔ دوم را به‌جای معادلهٔ تفاضلی مرتبهٔ اول قرارداده‌ایم. اگر تمام عملیات حسابی‌دقیق باشند، با استفاده از شرایط اولیه نتیجه می‌گیریم $C_2 = 0$، ولذا جواب دقیق از معادلهٔ (۷۱.۸) به‌دست می‌آید. ولی درعمل مقداری خطا وارد می‌شودکه عمدتاً به‌علت گردکردن یا به‌علت مقادیرآغازین غیردقیق‌است، و از این‌رو C_2 دقیقاً صفرنخواهد شد. بنابراین درهرمرحله از انتگرالگیری، خطای کوچکی داخل می‌شود، که مقدار آن بر اثر ضرب شدن در عامل $e^{2x_n}(1-)^n$ که به‌طور نمایی افزایش می‌یابد، بزرگ می‌شود. چون قسمت اصلی جواب حقیقی به‌طور نمایی نزول می‌کند، خطای حاصل‌ازجواب اضافی، سرانجام برجواب‌حقیقی مسلط شده منجر به‌نتایج کاملا غیردقیقی می‌شود.

با بیانی خالی ازدقت می‌توانیم روشی را ناپایدار گوییم، که خطاهای وارد درآن هنگام محاسبات، به‌نسبت نمایی افزایش یابند.

روشهای یك‌مرحله‌ای مانند روشهایی از نوع رونگه-کوتا به‌ازای h به‌اندازهٔ‌کافی کوچك، ناپایداری عددی نشان نمی‌دهند. روشهای چندمرحله‌ای در بعضی حالات ممکن است به‌ازای تمام مقادیر h و دربرخی دیگر به‌ازای یك رشته از مقادیر h ناپایدار باشند. برای‌اینکه ببینیم آیا یك روش چندمرحله‌ای پایدار است یـا نیست، می‌توانیم چنین‌عمل کنیم: اگر روش چندمرحله‌ای به‌یك معادلهٔ تفاضلی از مرتبهٔ k منجرشود، ریشه‌های معادلهٔ مشخصهٔ متناظر با معادلهٔ تفاضلی‌همگن‌را به‌دست می‌آوریم وآنها را β_i، $(i = 1, \dots, k)$، می‌نامیم. دراین صورت جواب عمومی معادلهٔ تفاضلی همگن چنین خواهد شد

$$y_n = c_1\beta_1^n + c_2\beta_2^n + \dots + c_k\beta_k^n \qquad (۷۲.۸)$$

یکی از این‌جوابها، مثلا β_1^n، هنگامی‌که $h \to 0$، به‌سمت جواب دقیق معادلهٔ تفاضلی‌میل می‌کند. همهٔ جوابهای دیگرخارجی هستند. یك روش چندمرحله‌ای را **پایدار قوی**[1] گویند اگر $h \to 0$، ریشه‌های اضافی درشرط

$$|\beta_i| < 1 \qquad i = 2, 3, \dots, k$$

صدق کنند. دراین شرایط هرخطایی‌که درمحاسبه وارد شود، هنگام‌افزایشn ازبین‌می‌رود، درحالی‌که اگر یکی ازریشه‌های خارجی β_i از لحاظ قدرمطلق بیشتر ازیك باشد، خطاها به‌طورنمایی بزرگ می‌شوند.

درمعادلهٔ تفاضلی‌کلی $y' = f(x, y)$، به‌دست‌آوردن ریشه‌های معادلهٔ مشخصه، β_i، غیرممکن خواهد بود. ولی برای‌آنکه معیاری ازناپایداری یك روش به‌دست‌آید، معمولا بررسی معادلهٔ خاص $y' = \lambda y$، که درآن λ یك عدد ثابت‌است، کفایت می‌کند.

در ابتدا روش‌مرتبهٔ چهارم ادمز-بشفورت رابررسی می‌کنیم. اگردرمعادلهٔ(۷۰.۸)

1. strongly stable

قراردهیم $f(x, y)=\lambda y$، خواهیم داشت

$$y_{n+1}-y_n-\frac{h\lambda}{۲۴}(۵۵y_n-۵۹y_{n-۱}+۳۷y_{n-۲}-۹y_{n-۳})=۰ \qquad (۷۳.۸)$$

معادلهٔ مشخصهٔ این معادلهٔ تفاضلی چنین‌است:

$$\beta^۴-\beta^۳-\frac{h\lambda}{۲۴}(۵۵\beta^۳-۵۹\beta^۲+۳۷\beta-۹)=۰$$

البته ریشه‌های این معادله توابعی از $h\lambda$ هستند. رسم براین‌است که معادلهٔ مشخصه‌را به‌شکل

$$\rho(\beta)+h\lambda\sigma(\beta)=۰ \qquad (۷۴.۸)$$

می‌نویسند، که در آن $\rho(\beta)$ و $\sigma(\beta)$ بسجمله‌ایهایی هستند که به‌صورت زیر تعریف می‌شوند

$$\rho(\beta)=\beta^۴-\beta^۳$$

$$\sigma(\beta)=-\frac{۱}{۲۴}(۵۵\beta^۳-۵۹\beta^۲+۳۷\beta-۹)$$

دیده می‌شود که به‌ازای $h \to ۰$، معادلهٔ (۷۴.۸) به $\rho(\beta)=۰$ تبدیل می‌شود که ریشه‌های آن برابرند با $\beta_۱=۱$، $\beta_۲=\beta_۳=\beta_۴=۰$. بـه‌ازای $h\neq ۰$، جواب عمومی معادلـهٔ (۷۳.۸) به‌شکل

$$y_n=c_۱\beta_۱^n+c_۲\beta_۲^n+c_۳\beta_۳^n+c_۴\beta_۴^n$$

است که در آن β_iها جوابهای معادلهٔ (۷۴.۸) هستند. می‌توان نشان داد که وقتی $h \to ۰$، $\beta_۱^n$ به‌جواب مطلوب معادلهٔ $y'=\lambda y$ نزدیک می‌شود، درحالی‌که سایر ریشه‌ها متناظر بـا جوابهای خارجی هستند. چون ریشه‌های معادلهٔ (۷۴.۸) توابع پیوسته‌ای از h هستند، از آنجا نتیجه می‌شود که وقتی h به‌اندازهٔ کافی کوچك باشد و بـه‌ازای $i=۲, ۳, ۴$، شرط $|\beta_i|<۱$ برقرار است. وازاین‌رو بنابرتعریف پایداری، روش ادمز-بشفورت قویاً پایدار است. همهٔ روشهای چندمرحله‌ای به‌یك معادلهٔ مشخصه به‌شکل (۷۴.۸) منجر می‌شوند که گاهی سمت چپ آنها را **بسجمله‌ای پایداری** نامند. تعریف پایداری می‌تواند براساس بسجمله‌ای پایداری مطرح شود. روشی را **پایدار قوی** گویند، که در آن همهٔ ریشه‌های $\rho(\beta)=۰$ بجز ریشهٔ سادهٔ $\beta=۱$، از لحاظ قدرمطلق کمتر از یك باشند.

حال خواص پایداری روش میلن (۶۴.۸ب) را که به‌صورت

$$y_{n+1}=y_{n-۱}+\frac{h}{۳}(f_{n+1}+۴f_n+f_{n-۱}) \qquad (۷۵.۸)$$

بیان شده‌است، بررسی می‌کنیم. بازهم ازقراردادن $f(x, y)=\lambda y$، خواهیم داشت

$$y_{n+1}-y_{n-1}-\frac{h\lambda}{3}(y_{n+1}+4y_n+y_{n-1})=0$$

که معادلهٔ مشخصهٔ آن چنین می‌شود:

$$\rho(\beta)+h\lambda\sigma(\beta)=0 \qquad (۷۶۰۸)$$

که در آن

$$\rho(\beta)=\beta^2-1$$

$$\sigma(\beta)=\beta^2+4\beta+1$$

در اینجا معادلهٔ $\rho(\beta)=0$ ریشه‌هایی به‌صورت $\beta_1=1$ و $\beta_2=-1$ دارد و از این‌رو بنا بر تعریف بالا، روش میلن پایدار قوی نیست. برای آنکه مفهوم ضمنی این مطلب را درك کنیم، ریشه‌های بسجمله‌ای پایداری (۷۶۰۸) را محاسبه می‌کنیم. به‌ازای مقدار بسیار کوچك h داریم

$$\beta_1=1+\lambda h+\mathcal{O}(h^2)$$

$$\beta_2=-(1-\lambda h/3)+\mathcal{O}(h^2) \qquad (۷۷۰۸)$$

از این‌رو جواب عمومی معادلهٔ (۷۵۰۸) چنین‌است

$$y_n=c_1(1+\lambda h+\mathcal{O}(h^2))^n+c_2(-1)^n(1-\lambda h/3+\mathcal{O}(h^2))^n$$

اگر n را مساوی x_n/h بگذاریم، و h را به‌سمت صفرمیل دهیم این جواب به‌جواب

$$y_n=c_1e^{\lambda x_n}+c_2(-1)^ne^{-\lambda x_n/3} \qquad (۷۸۰۸)$$

نزدیك می‌شود. در این حالت پایداری به‌علامت λ بستگی پیدا می‌کند. اگر $\lambda>0$ و لذا جواب مطلوب به‌طور نمایی صعودی باشد، واضح است که جواب خارجی به‌طور نمایی نزولی خواهد بود و لذا روش میلن پایدار می‌شود. ازسوی دیگر اگر $\lambda<0$، آنگاه روش میلن ناپایدار خواهد بود زیرا جواب اضافی به‌طور نمایی افزایش می‌یابد و سرانجام جواب مطلوب را محو می‌کند. روشهایی از این نوع را که پایداریشان برای معادلهٔ آزمون $y'=\lambda y$ بستگی به‌علامت λ دارد، روشهای **پایدار ضعیف** نامند، برای معادلهٔ کلیتر $y'=f(x,y)$، هرگاه در بازهٔ انتگرالگیری $\partial f/\partial y<0$، از روش‌میلن می‌توان پایداری ضعیف‌را انتظار داشت.

درعمل همهٔ روشهای چندمرحله‌ای در رشته‌ای از مقادیر برای نمو h، ناپایداری نشان می‌دهند مثلا، روش مرتبهٔ دوم ادمز-بشفورت راکه به‌صورت

$$y_{n+1}=y_n+\frac{h}{2}\{3f_n-f_{n-1}\}$$

تعریف می‌شود در نظر می‌گیریم. اگر این روش برای معادلهٔ آزمون $y'=\lambda y$ به‌کار گرفته شود، معادلهٔ تفاضلی

$$y_{n+1}-y_n-\frac{h\lambda}{2}\{3y_n-y_{n-1}\}=0$$

به‌دست خواهد آمد، و از روی آن بسجمله‌ای پایداری

$$\beta^2-\beta-\frac{h\lambda}{2}\{3\beta-1\}$$

یا معادلهٔ

$$\beta^2-\left(1+\frac{3h\lambda}{2}\right)\beta+\frac{h\lambda}{2}=0$$

به‌دست می‌آید. اگر $\lambda<0$، هردو ریشهٔ این معادلهٔ درجهٔ دوم از لحاظ قدرمطلق کوچکتر از یك هستند به‌شرط آنکه $-1<h\lambda<0$. دراین حالت چون خطاهای ناشی ازجواب خارجی بزرگ نخواهند شد، پایداری مطلق به‌دست خواهیم آورد. ولی اگر $|h\lambda|>1$، آنگاه یکی ازاین ریشه‌ها ازلحاظ قدرمطلق از یك بزرگتر می‌شود و تاحدی به‌ناپایداری بر می‌خوریم. شرط $-1<h\lambda<0$، نمو h راکه برای این روش می‌تواند به‌کار رود به طور مؤثری محدود می‌کند. برای مثال اگر $\lambda=-100$ آنگاه، برای اطمینان از پایداری باید $h<0.01$ انتخاب شود. یك روش چندمرحله‌ای را به‌ازای مقادیری از $h\lambda$که به‌ازای آنها ریشه‌های بسجمله‌ای پایداری آن (۷۲.۸) از لحاظ قدرمطلق کمتر از یك باشند، روش **پایدارمطلق** نامند. روشهای مختلف، نواحی مختلفی برای پایداری مطلق دارند. عموماً ما آن روشهایی‌را ترجیح می‌دهیم که بزرگترین ناحیه را برای پایداری مطلق دارند. برای مثال، می‌توان نشان داد که نواحی پایداری در روشهای ضمنی ادمز-مولتن، ۱۰ برابر بزرگتر از نواحی پایداری روشهای ادمز-بشفورت با همان مرتبه‌است. مخصوصاً روش مرتبهٔ دوم ادمز-مولتن که با رابطهٔ

$$y_{n+1}=y_n+h\left(f_{n+1}-\frac{1}{2}f_n+\frac{1}{2}f_{n-1}\right)$$

معین می‌شود، برای معادلهٔ آزمون $y'=\lambda y$ و به‌ازای $\lambda<0$ و $-\infty<h\lambda<0$، پایدار مطلق‌است.

برای معادلاتی به‌شکل $y'=\lambda y$ و با $\lambda>0$، جواب مطلوب مانند $e^{h\lambda}$، به‌طور نمایی بزرگ خواهد شد. هر روش چندمرحله‌ای باید یك ریشه، ریشهٔ اصلی، داشته باشد که این ریشه تقریبی‌است برای جواب مطلوب. سپس همهٔ ریشه‌های خارجی دیگرباید از لحاظ قدرمطلق کمتر ازاین‌ریشهٔ اصلی باشند. روشی با این خصوصیت راکه همهٔ ریشه‌های

خارجی بسجمله‌ای پایداری‌آن از لحاظ قدرمطلق کمتر از ریشهٔ اصلی باشند روش **پایدار نسبی**[1] نامند. نواحی‌پایداری برای روشهای چندمرحله‌ای در کتاب گیر[۳۵] به‌طورمبسوط موردبحث قرار گرفته‌اند.

تمرین

۱۰۰۸-۱ نشان دهیدکه فرمول‌تصحیحی بر اساس‌قاعدهٔ ذوزنقه‌ای (۵۲۰۸)، برای‌معادلاتی به‌شکل $y'=\lambda y$ پایدار است (تمرین۸۰۸-۱ را ببینید).

۱۰۰۸-۲ نشان‌دهیدکه ریشه‌های معادلهٔ مشخصهٔ(۷۶۰۸)، وقتی $h \to 0$، می‌تواند به‌شکل (۷۷۰۸) بیان شود، ونشان دهیدکه جواب معادلهٔ تفاضلی(۷۵۰۸) وقتی$h \to 0$، به‌جواب (۷۸۰۸) نزدیك می‌شود.

۱۰۰۸-۳ یك برنامهٔ‌کامپیوتری بنویسیدکه ریشه‌های معادلهٔ مشخصهٔ(۷۳۰۸) را برای‌فرمول ادمز-بشفورت به‌دست دهد. برای این برنامه قراردهید $\lambda=-1$ و $\bar{h}=0(0/1)h$. یك مقدارتقریبی $\bar{h}$ را،که بیشتر از آن یك یا چند ریشهٔ این معادله از لحاظ قدرمطلق بزرگتر از یك می‌شوند، تعیین نمایید. و بدین‌طریق‌کران بالایی برای h به‌دست‌آورید‌که به‌ازای مقدار h بیشترازآن‌کران، روش ادمز-بشفورت ناپایدار شود.

۱۰۰۸-۴ به‌ازای $h=1/2$ و از $x=0$ تا $x=6$، معادلهٔ (۶۷۰۸) را با روش میلن (۶۴۰۸) حل‌کنید. مقادیر آغازین را ازجدول۱۰۸ به‌دست‌آورید. به‌نتیجهٔ ناپایداری‌روی جواب توجه نمایید.

۱۱۰۸* کنترل و پخش خطای گردکردن

درقسمت ۴۰۸، خطای گسسته‌سازی را به‌صورت

$$e_n=y(x_n)-y_n$$

تعریف‌کردیم‌که در آن $y(x_n)$ جواب‌حقیقی‌معادلهٔ دیفرانسیل‌است و y_n جواب دقیق‌معادلهٔ تفاضلی‌تقریبی‌برای معادلهٔ دیفرانسیل. درعمل چون درکامپیوترها طول‌کلمه[2] محدوداست، یك مقدار $\tilde{y}_n$ به‌دست می‌آوریم‌که، به‌علت خطاهای گردکردن، با y_n تفاوت دارد. این تفاوت‌که با

$$r_n=y_n-\tilde{y}_n$$

نشان داده می‌شود خطای‌گردکردن انباشته[3]نام دارد، یعنی اختلاف بین‌جواب دقیق معادلهٔ

1. relatively stable 2. word lengths
3. accumulated round-off-error

تفاضلی وجوابی که ازکامپیوتر در نقطهٔ $x = x_n$ به‌دست می‌آید. درهر مرحله از انتگرالگیری یك خطای گردکردن حاصل می‌شود که آن را خطای **گردکردن موضعی** می‌نامیم و با ε_n مشخص می‌کنیم. برای مثال، در روش اویلر، ε_n به‌صورت زیر تعریف می‌شود

$$\tilde{y}_{n+1} = \tilde{y}_n + hf(x_n, \tilde{y}_n) + \varepsilon_n$$

خطای گردکردن انباشته، اصلا برابر با مجموع خطاهای گردکردن موضعی نیست، زیرا هریك از خطاهای موضعی پخش شده و ممکن است همچنان که محاسبات پیش می‌روند، هریك ازاین خطاها افزایش یاکاهش یابد. درحالت کلی، موضوع پخش خطای گردکردن خوب توجیه نشده‌است ونتایج نظری بسیارکمی دربارهٔ این موضوع دردست‌است. خطای گردکردن انباشته به‌عوامل زیادی بستگی دارد، ازجمله، (۱) نوع حسابی که در کامپیوتر به‌کاربرده می‌شود، یعنی حساب با ممیز ثابت یا با ممیزشناور، (۲) روشی که‌کامپیوتر برای گردکردن به‌کار می‌برد، (۳) ترتیبی که عملیات حسابی انجام می‌گیرند، (۴) شیوهٔ عددی که به‌کاربرده می‌شود.

چنان که دربخش۸.۱۰، دربحث ناپایداری عددی نشان داده شد،اثر پخش گردکردن می‌تواند خطرناك باشد. اما حتی در روشهای پایدار، از دست دادن دقت براثرخطاهای گردکردن اجتناب‌ناپذیراست. این امر درفصل۷ که قانون ذوزنقه‌ای برای محاسبهٔ انتگرال به‌کاربرده شده بود ملاحظه شد. ازدست دادن دقت دربازهٔ بزرگترممکن‌است آنقدر جدی باشدکه نتایج را به‌طورکلی بی‌اعتبار سازد.

از خطای گردکردن انباشته، با برخی فرضیات آماری دربارهٔ توزیع خطاهای گردکردن موضعی، می‌توان براوردهایی به‌دست آورد. دراینجا این گونه مسائل‌دنبال نخواهند شد، بلکه به‌دنبال یك شیوهٔ ساده ومؤثری هستیم که هنگام حل معادلات دیفرانسیل، ازدست دادن دقت بر اثر خطاهای گردکردن راکاهش دهد.

بسیاری ازفرمولهای موردبحث برای حل‌معادلات دیفرانسیل دراین فصل‌رامی‌توان به‌شکل

$$y_{n+1} = y_n + h\,\Delta y_n$$

نوشت،که درآن $h\,\Delta y_n$ معرف نموی‌است‌متضمن ترکیبات $f(x, y)$ درنقاط انتخابی. این‌نمودرمقایسه با خود y_n، معمولا نموی کوچك‌است. درتشکیل‌حاصلجمع $y_n + h\,\Delta y_n$ درحساب با ممیزشناور،کامپیوتر $h\,\Delta y_n$ را به‌طرف راست منتقل می‌کند تاتوان $h\Delta y_n$ با توان y_n توافق نماید و با انجام این‌کار، بیتهایی[1] از سمت راست حذف می‌شوند. سپس عمل جمع صورت می‌گیرد، ولی بر اثر حذف بیتها، خطای گردکردن پدیدمی‌آید. برای اینکه این مطلب را روشنتر ببینیم، بیاییم دوعدد با ممیزشناور $(۱۰^۴)$ (۵۴۷۲ر۰) و $(۱۰^۲)$(۳۸۵۶ر۰) را با هم جمع‌کنیم. با این فرض‌که طول‌کلمه را چهار رقم اعشاری

1. bits

بگیریم. اگر عدد دوم را دو رقم به‌سمت راست انتقال دهیم، دو رقم آخر آن حذف می‌شوند و با افزودن آن به‌اولین عدد، مقدار (۱۰۴)(۰٫۵۵۱۰) حاصل می‌شود، در حالی که اگر گردکردن را به‌طور صحیح انجام می‌دادیم نتیجه می‌باید برابر با (۱۰۴)(۰٫۵۵۱۱) می‌شد. البته این مثال مثالی است بیش‌ازحد واقع، زیرا کامپیوترها با بیتهای دودویی و طول کلمهٔ بزرگتر کار می‌کنند، اما با این وجود اثرات انباشتگی[1] می‌توانند جدی باشند.

اکنون شیوهٔ ساده‌ای را ذکر می‌کنیم که خطاهایی از این نوع را به‌طور قابل توجهی کاهش می‌دهد. در ابتدا هر مقدار محاسبه شده‌ای از y_n به‌صورت دقت مضاعف ذخیره، و سپس $h\,\Delta y_n$ با دقت ساده محاسبه می‌شود، و برای این کار تنها آن قسمت از y_n که متضمن دقت ساده و برای تشکیل $h\,\Delta y_n$ لازم است به‌کار می‌رود، مجموع $y_n = h\Delta y_n$ به‌صورت **دقت مضاعف** به‌دست می‌آید، و $y_{n+1} = y_n + h\Delta y_n$ با دقت مضاعف ذخیره می‌شود. این شیوه را **انباشتگی دقت مضاعف جزئی**[2] نامند. در بعضی* از کامپیوترها، انجام محاسبات با دقت مضاعف به‌صورت یك دستورالعمل[3] موجود است، ولی حتی اگر این نیز در کامپیوتر موجود نباشد، در هر مرحله از انتگرالگیری، تنها یك عمل جمع با دقت مضاعف می‌باید انجام گیرد. قسمت اصلی محاسبه، عبارت است از تعیین $h\Delta y_n$، و این کار با دقت ساده انجام می‌پذیرد. کار اضافی و نیز حافظهٔ اضافی در رابطه با این امر بسیار اندك است. از طرف دیگر با انجام این کار به‌دست آوردن دقت ممکن، می‌تواند بسیار قابل توجه باشد، مخصوصاً زمانی که دقت زیاد در یك بازهٔ بزرگ مطلوب باشد. در واقع این شیوهٔ عمل در امر کاهش انباشتگی خطای گردکردن آنقدر مؤثر است که هیچ برنامه‌ای از برنامه‌های همه منظورهٔ پیش‌ساخته برای حل معادلات دیفرانسیل نیست که برای ایجاد شکلی از این انباشتگی دقت مضاعف جزیی، امکاناتی فراهم نکرده باشد.

در اینجا بی‌مناسبت نیست که چند کلمه‌ای به‌عنوان توضیح بیاوریم. دقت یك انتگرالگیری عددی به‌خطای گسسته‌سازی و خطای گردکردن انباشته بستگی دارد. برای آنکه خطای گسسته‌سازی کم باشد، طبیعتاً نمو h کوچك انتخاب می‌شود. از طرف دیگر، هرچه h کوچکتر گرفته شود، مراحل انتگرالگیری بیشتری باید پیموده شود که در پی آن احتمالاً خطای گردکردن بیشتر می‌شود. بنابراین مقدار بهینه‌ای از نمو h وجود دارد که این مقدار بهینه برای یك کامپیوتر مفروض و یك مسئلهٔ مفروض، بهترین دقت را به‌دست می‌دهد. پیداکردن این مقدار بهینه درعمل بدون اینکه مقدار زیادی از وقت کامپیوتر را بگیرد بسیار مشکل است. ولی وجود این مقدار بهینه نشان می‌دهد که اگر نمو را خیلی کوچك بگیریم، خطراتی را درپی خواهد داشت.

---

1. cumulative 2. partial double-precision accumulation

* درمتن اصلی کتاب، کلمهٔ «بعضی» از کامپیوترها آمده‌است، ولی انجام محاسبات با دقت مضاعف در بیشتر کامپیوترهای همه‌منظوره موجود است.ـم.

3. instruction

□ **مثال ۸۰۸:** معادلهٔ

$$y' = \frac{1}{x^2} - \frac{y}{x} - y^2 \quad y(1) = -1$$

را با استفاده از روش ادمز-بشفورت، به‌ازای $h = 1/256$ از $x = 1$ تا $x = 3$ حل کنید. این کار را یک‌بار با انباشتگی دقت مضاعف جزئی و یک‌بار بدون آن انجام دهید.

نتایج به‌دست آمده از کامپیوتر در زیر داده شده‌اند. نمو عمداً آنقدر کوچک انتخاب شده که خطای گسسته‌سازی قابل صرفنظر باشد. نتایج بعد از هر ۱۶ مرحله چاپ شده‌اند. جواب دقیق این مسئله برابر با $y = -1/x$ است. بنابراین دقت می‌توانند به‌آسانی

نتایج کامپیوتری برای مثال ۸۰۸

X	SINGLE PRECISION	PARTIAL DOUBLE PRECISION
0.99999999	−0.99999999	−0.99999999
1.06250000	−0.94117642	−0.94117647
1.12500000	−0.88888878	−0.88888889
1.18750000	−0.84210509	−0.84210526
1.24999990	−0.79999977	−0.80000000
1.31249990	−0.76190444	−0.76190476
1.37500000	−0.72727232	−0.72727273
1.42750000	−0.69565168	−0.69565218
1.50000000	−0.66666608	−0.66666667
1.56249990	−0.63999934	−0.64000001
1.62499990	−0.61538386	−0.61538462
1.68750000	−0.59259175	−0.59259260
1.75000000	−0.57142763	−0.57142858
1.81250000	−0.55172310	−0.55172415
1.87499990	−0.53333220	−0.53333335
1.93749990	−0.51612781	−0.51612905
2.00000000	−0.49999869	−0.50000001
2.06250000	−0.48484711	−0.48484850
2.12500000	−0.47058678	−0.47058825
2.18749990	−0.45714134	−0.45714287
2.24999990	−0.44444284	−0.44444446
2.31250000	−0.43243076	−0.43243245
2.37500000	−0.42105088	−0.42105265
2.43750000	−0.41025458	−0.41024643
2.49999990	−0.39999810	−0.40000002
2.56249990	−0.39024193	−0.39024393
2.62500000	−0.38095033	−0.38095240
2.68750000	−0.37209089	−0.37209304
2.75000000	−0.36363416	−0.36363639
2.81249990	−0.35555328	−0.35555558
2.87499990	−0.34782372	−0.34782612
2.92750000	−0.34042308	−0.34042556
3.00000000	−0.33333080	−0.33333336

□

آزمایش شود. در $x=۳$ نتایج با دقت مضاعف جزیی، تا سه واحد در رقم هشتم دقیق هستند و با دقت ساده تا ۲۵۳ واحد در رقم هشتم. از آنجایی که این خطا از گردکردن ناشی‌شده‌است، این‌مثال به‌وضوح مؤثربودن دقت مضاعف جزیی را در امرکاهش‌انباشتگی خطای گردکردن نشان می‌دهد.

تمرین

۱-۱۱۰۸ برنامه‌ای براساس روش ادمز-بشفورت بسازیدکه هم از انباشتگی‌دقت‌مضاعف جزیی و هم از دقت ساده درآن استفاده شود.

۲-۱۱۰۸ برنامهٔ تمرین ۱-۱۱۰۸ را برای حل معادلهٔ زیر از $x=۰$ تا $x=۲$ بااستفاده از نمو ثابت $h=۰٫۰۱$ به‌کاربرید

$$y'=-۲y \qquad y(۰)=۱$$

مقادیر آغازین می‌توانند ازجواب دقیق $h=e^{-۲x}$ به‌دست آیند. خطای ناشی‌ازگردکردن چقدراست؟

۳-۱۱۰۸ برنامه‌ای برای روش‌کلاسیك مرتبهٔ چهارم رونگه-کوتا بنویسیدکه هم از انباشتگی دقت مضاعف و هم ازدقت ساده درآن استفاده شود. این برنامه را برای حل معادلهٔ تمرین ۲-۱۱۰۸ به‌ازای همان مقدار از نمو به‌کارگیرید.

۱۲.۸* دستگاههای معادلات دیفرانسیل

دربسیاری‌از زیربرنامه‌های معادلات دیفرانسیل‌همه‌منظوره، فرض برآن‌است‌که یك معادلهٔ دیفرانسیل مرتبهٔ N به‌صورت یك دستگاه از N معادلهٔ مرتبهٔ اول بیان شده‌است. برای‌یك معادلهٔ مرتبهٔ N که به‌شکل

$$y^{(N)}=f(x, y(x), y'(x), \cdots, y^{(N-۱)}(x)) \qquad (۷۹.۸)$$

می‌باشد، این تبدیل همیشه به‌صورت زیر انجام می‌پذیرد: با انتخاب $y=y_۱$، داریم

$$\begin{aligned} y_۱' &= y_۲ \\ y_۲' &= y_۳ \\ y_۳' &= y_۴ \\ &\cdots\cdots \\ y_{N-۱}' &= y_N \\ y_N' &= f(x, y_۱, y_۲, \cdots, y_N) \end{aligned} \qquad (۸۰.۸)$$

دستگاه (۸۰.۸) بـا معادلهٔ (۷۹.۸) هم‌ارزاست. هردستگاه معادلات را نمی‌توان به‌شکل سادهٔ (۸۰.۸) بیان نمود. بسیاری ازاوقات، دستگاه N معادلهٔ مرتبهٔ اول به‌شکل‌زیر است

$$\begin{aligned} y_1' &= f_1(x, y_1, y_2, \cdots, y_N) \\ y_2' &= f_2(x, y_1, y_2, \cdots, y_N) \\ &\cdots\cdots\cdots\cdots \\ y_N' &= f_N(x, y_1, y_2, \cdots y_N) \end{aligned} \qquad (۸۱.۸)$$

همهٔ روشهای عددیی راکه در این فصل بررسی شدند می‌توان با دستگاه (۸۱.۸) تطبیق داد. دستگاه (۸۱.۸) را می‌توان به‌صورت فشرده‌تر به‌شکل‌برداری

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$$

بیان نمودکه در آن $\mathbf{y}'$، $\mathbf{f}$ و $\mathbf{y}$ بردارهایی هستند با Nمؤلفه.

طریقهٔ انجام‌کار را در روش‌رونگه‌ـکوتا برای دومعادله،که به‌صورت زیرمی‌نویسیم، نشان می‌دهیم

$$y' = f(x, y, z)$$

$$z' = g(x, y, z)$$

فرمولهای رونگه‌ـکوتا متناظر با فرمول (۳۷.۸) به‌شکل زیردرمی‌آیند

$$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$z_{n+1} = z_n + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4) \qquad (۸۳.۸)$$

که در آنها

$$k_1 = hf(x_n, y_n, z_n)$$

$$l_1 = hg(x_n, y_n, z_n)$$

$$k_2 = hf\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}, z_n + \frac{l_1}{2}\right)$$

$$l_2 = hg\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}, z_n + \frac{l_1}{2}\right)$$

$$k_۳ = hf\left(x_n+\frac{h}{۲},\ y_n+\frac{k_۲}{۲},\ z_n+\frac{l_۲}{۲}\right)$$

$$l_۳ = hg\left(x_n+\frac{h}{۲},\ y_n+\frac{k_۲}{۲},\ z_n+\frac{l_۲}{۲}\right)$$

$$k_۴ = hf(x_n+h,\ y_n+k_۳,\ z_n+l_۳)$$

$$l_۴ = hg(x_n+h,\ y_n+k_۳,\ z_n+l_۳)$$

بدیهی است که این کار را برای یك دستگاه معادلات نیز می‌توان تعمیم داد. باید توجه داشت که همهٔ نموها با زیرنمایه‌های کوچکتر را می‌باید قبل از نموهای بعدی با زیرنمایه‌های بالاتر محاسبه نمود.

فرمولهای ادمز-مولتن که برای معادله‌های (۸۲.۸) به‌کار رفته‌اند به‌صورت زیر عمل می‌کنند:

$$y^{(۰)}_{n+۱} = y_n+\frac{h}{۲۴}[۵۵f(x_n,y_n,z_n)-۵۹f(x_{n-۱},y_{n-۱},z_{n-۱})$$
$$+۳۷f(x_{n-۲},y_{n-۲},z_{n-۲})-۹f(x_{n-۳},y_{n-۳},z_{n-۳})]$$

$$z^{(۰)}_{n+۱} = z_n+\frac{h}{۲۴}[۵۵g(x_n,y_n,z_n)-۵۹g(x_{n-۱},y_{n-۱},z_{n-۱})$$
$$+۳۷g(x_{n-۲},y_{n-۲},z_{n-۲})-۹g(x_{n-۳},y_{n-۳},z_{n-۳})]$$

$$y^{(۱)}_{n+۱} = y_n+\frac{h}{۲۴}[۹f(x_{n+۱},y^{(۰)}_{n+۱},z^{(۰)}_{n+۱})+۱۹f(x_n,y_n,z_n)$$
$$-۵f(x_{n-۱},y_{n-۱},z_{n-۱})+f(x_{n-۲},y_{n-۲},z_{n-۲})]$$

$$z^{(۱)}_{n+۱} = z_n+\frac{h}{۲۴}[۹g(x_{n+۱},y^{(۰)}_{n+۱},z^{(۰)}_{n+۱})+۱۹g(x_n,y_n,z_n)$$
$$-۵g(x_{n-۱},y_{n-۱},z_{n-۱})+g(x_{n-۲},y_{n-۲},z_{n-۲})] \qquad (۸۴.۸)$$

در بخش ۶.۸، برای حل یك معادلهٔ دیفرانسیل، از زیربرنامه‌ای به‌نام DVERK از مجموعهٔ برنامه‌های IMSL استفاده کردیم. در اینجا، از این زیربرنامه برای حل یك دستگاه معادلات دیفرانسیل مرتبهٔ اول استفاده می‌کنیم. در زیربرنامهٔ DVERK، X معرف متغیرمستقل است، درحالی که Y(K)، K = ۱, ۲, ..., N، برای تعیین برداری از متغیرهای غیرمستقل به‌طول N به‌کار می‌رود و فرض بر این است که N معادلهٔ مرتبهٔ اول به‌شکل (۸۱.۸)

وجود دارد. YPRIME(K) باشرط $K=1,2,\ldots,N$، برای تعیین بردار تابعهای $f_1,\ldots,f_N$ در سمت راست (۸۱۰۸) به‌کار می‌رود. زیربرنامهٔ FCN برای تعریف YPRIME(K) به‌کار رفته‌است. استفاده از زیربرنامهٔ DVERK برای یک دستگاه معادلات، از جهاتی مشابه با به‌کاربردن آن برای یک معادلهٔ تنهاست. مثال زیر این مورد استفاده را روشن می‌سازد.

□ **مثال ۹۰۸:** دستگاه معادلات زیر را به‌صورت یک دستگاه معادلات مرتبهٔ اول بیان کنید و آن را با استفاده از زیربرنامهٔ DVERK از $x=0$ تا $x=1$ حل نمایید:

$$\frac{d^2z}{dx^2}=z^2-y+e^x$$

$$\frac{d^2y}{dx^2}=z-y^2-e^x$$

$$z(0)=z'(0)=0 \qquad y(0)=1 \qquad y'(0)=-2 \qquad (۸۵۰۸)$$

در این مثال x متغیر مستقل است، در حالی که $z(x)$ و $y(x)$ متغیرهای وابسته می‌باشند. برای آنکه این دستگاه را به‌صورت یک دستگاه مرتبهٔ اول بیان کنیم، قرار می‌دهیم $z(x)=y_1(x)$، $y(x)=y_2(x)$ و در این‌صورت، این دستگاه مرتبهٔ اول همراه با شرایط اولیه‌اش به‌صورت زیر درمی‌آید

$$y_1'(x)=y_3(x) \qquad y_1(0)=0.5$$

$$y_2'(x)=y_4(x) \qquad y_2(0)=1.5$$

$$y_3'(x)=y_1^2(x)-y_2(x)+e^x \qquad y_3(0)=0.5$$

$$y_4'(x)=y_1(x)-y_2^2(x)-e^x \qquad y_4(0)=-2.5 \qquad (۸۶۰۸)$$

برنامهٔ **فورترن** و نتایج جزیی آن در زیر داده شده‌اند. مقادیر به‌دست‌آمده حداقل تا هشت رقم با معنا صحیح‌اند. به‌نظر می‌رسد که برای دستیابی به این دقت به‌ازای هر خروجی، ۱۶ بار محاسبهٔ تابع مورد نیاز باشد. این امر مستلزم این است که یک مرحلهٔ داخلی یا نمو تقریبی $h=0.05$ به‌کار گرفته شود.

```
C   PROGRAM TO SOLVE EXAMPLE 8.9 USING  D V E R K  (IMSL) .
      INTEGER IER,IND,K,N,NW
      REAL C(24),TOL,W(5,9),X,XEND,Y(4)
      DATA N , X ,       Y              , TOL ,IND,NW
     *   / 4 , 0.,   0.,1.,0.,-2.  ,1.E-9, 1 , 5 /
      EXTERNAL FCN2
      DO 12 K=1,10
         XEND = FLOAT(K)/10.
         CALL DVERK ( N, FCN2, X, Y, XEND, TOL, IND, C, NW, W, IER )
         PRINT 600, XEND,Y(1),Y(2),C(24)
```

```
600    FORMAT(3X,F3.1,3X,2(E16.8,3X),F4.0)
 12 CONTINUE
                                        STOP
    END
    SUBROUTINE FCN2 ( N, X, Y, YPRIME )
    INTEGER N
    REAL X, Y(N), YPRIME(N)
    YPRIME(1) = Y(3)
    YPRIME(2) = Y(4)
    YPRIME(3) = Y(1)**2 - Y(2) + EXP(X)
    YPRIME(4) = Y(1) - Y(2)**2 - EXP(X)
                                        RETURN
    END
```

نتایج کامپیوتری برای مثال ۹۰۸

X	Y(1)	Y(2)	FCN EVALS
.1	5.12342280E − 04	7.90476884E − 01	16
.2	4.19528369E − 03	5.63595308E − 01	32
.3	1.44796017E − 02	3.21283135E − 01	48
.4	3.50756908E − 02	6.44861308E − 02	64
.5	6.99842327E − 02	−2.07035152E − 01	80
.6	1.23532042E − 01	−4.94906488E − 01	96
.7	2.00446026E − 01	−8.02372169E − 01	112
.8	3.05983760E − 01	−1.13460479E + 00	128
.9	4.46147292E − 01	−1.49915828E + 00	144
1.0	6.28019076E − 01	−1.90666076E + 00	160

□

همان گونه‌که این مثال نشان می‌دهد، DVERK زیر برنامه‌ای‌است‌که به‌کاربردنس خیلی ساده‌است و هنگامی‌که به‌دقت بالایی نیازباشد،کارایی‌آن خیلی زیاداست.

تمرین

۱-۱۲۰۸ معادلهٔ مرتبهٔ دوم

$$y''(x)=۲(e^{۲x}-y^{۲})^{۱/۲}$$

$$y(۰)=۰ \qquad y'(۰)=۱$$

را به‌صورت دستگاهی از معادلات مرتبهٔ اول بنویسید و با استفاده از روش کلاسیك مرتبهٔ چهارم رونگه‌ـکوتا به‌ازای نموهای ثابت $h=۱/۶۴$ و $h=۱/۱۲۸$ آن را از $x=۰$ تا $x=۱$ حل نمایید. دقت نتایج حاصل را براورد کنید.

۲-۱۲۰۸ معادلهٔ مرتبهٔ دوم زیر را با استفاده از فرمولهای ادمز-مولتن (۸۴۰۸)، به‌ازای نمو ثابت $h=۰/۱$، از $x=۱$ تا $x=۲$ حل کنید:

$$y''(x)=۲y^{۳}$$

$$y(۱)=۱ \qquad y'(۱)=-۱$$

چهار مقدار آغازین برای $y(x)$ و $f(x, y)=۲y^۳$ لازم‌است. این مقادیر را از جواب دقیق $y(x)=۱/x$ به‌دست‌آورده سپس نتایج خود را با جواب دقیق مقایسه‌کنید.

۱۲.۸-۳ بررسی کنیدکه آیا مرکزمحاسباتی‌شما برنامه‌های IMSL را آبونه شده‌است یا نه. اگر چنین است، معادلهٔ تمرین ۱۲.۸-۱ با استفاده از DVERK و به‌ازای مقادیر XEND، K/۱۰ (K = ۱, ... , ۱۰)، حل‌نمایید.

۱۳.۸* معادلات دیفرانسیل سخت[1]

در عده‌ای از زمینه‌هـای مهم، از جمله واکنشهای شیمیایی، سیستمهای‌کنترل و شبکه‌های الکترونیکی،کاربردها به‌دستگاههایی‌ازمعادلات دیفرانسیل‌منجرمی‌شوندکه حل‌آنها دشوار است، زیرا فرایندهای مختلف دردستگاه، با مقیاسهای زمانی[2]کاملا متفاوتی رفتارمی‌کنند. مثلا اگرجواب یك معادلهٔ دیفرانسیل به‌صورت $y(x)=C_۱e^{-x}+C_۲e^{-۱۰۰۰x}$ داده شده باشد، مؤلفهٔ دوم این جواب درمقایسه با مؤلفهٔ اول آن با افزایش x بسیارسریعتر تحلیل می‌رود. بیشتر روشهایی‌کـه برای حل معادلات دیفرانسیل بیان‌کرده‌ایـم، وقتی برای حل مسائلی بـا جوابهایی از این نـوع به‌کار روند، ناپایداری زیادی ازخود نشان می‌دهند. مسائلی را که مؤلفه‌های جواب آنها مقیاسهای زمانی سخت متفاوتی دارند مسائل سخت نامند.

برای مثال، معادلهٔ مرتبهٔ دوم

$$\frac{d^۲y}{dx^۲}+۱۰۰۱\frac{dy}{dx}+۱۰۰۰y=۰ \qquad (۸۷.۸)$$

را درنظر می‌گیریم. جواب عمومی معادلهٔ (۸۷.۸) برابراست با

$$y(x)=Ae^{-x}+Be^{-۱۰۰۰x}$$

اگرشرایط اولیه را $y(۰)=۱$ و $y'(۰)=-۱$ بگیریم، جواب دقیق خواهد شد:

$$y(x)=e^{-x}$$

اکنون با استفاده از روش مرتبهٔ چهارم رونگه‌ـکوتـا، سعی می‌کنیم معادلـهٔ (۸۷.۸) را به‌ازای این‌شرایط اولیه حل‌کنیم. وقتی دستگاه فوق را به‌صورت یك دستگاه مرتبـهٔ اول (بخش۱۲.۸ را ببینید) بنویسیم، چنین می‌شود

$$\frac{dy_۱}{dx}=y_۲ \qquad y_۱(۰)=۱$$

$$\frac{dy_۲}{dx}=-۱۰۰۱y_۲-۱۰۰۰y_۱ \qquad y_۲(۰)=-۱ \qquad (۸۸.۸)$$

---

1. stiff 2. time scales

به‌ازای نموهای $h < ۰٫۰۰۲$، روش رونگه‌ـکوتا جوابهایی به‌دست می‌دهد که جواب e^{-x} را به‌خوبی تقریب می‌زند. ولی معنی $h = ۰٫۰۰۲$ این‌است که در هر بازه، ۵۰۰ مرحلهٔ انتگرالگیری می‌باید وجود داشته باشد. از آنجایی که جواب مطلوب برابر است با $y(x) = e^{-x}$ به‌نظرمی‌رسد که انتخاب خیلی بزرگتر نمو h مانعی نداشته باشد. ولی اگر $h = ۰٫۰۰۳$، این‌مقدار باز خیلی کوچك‌است، جواب عددی به‌سمت ∞ جهش می‌یابد[1].* توضیح این رفتار به‌عواملی که برای شرایط پایداری روش اختیارشده بستگی دارد. برای روش رونگه‌ـکوتا، ناحیهٔ پایداری چنان‌است که می‌باید داشته باشیم

$$۱۰۰۰h < ۲٫۸$$

یا $h < ۰٫۰۰۲۸$. یعنی برای مسئلهٔ بالا، نمو h به‌علت سرعت تغییرمؤلفهٔ جواب، یعنی $e^{-۱۰۰۰y}$، به‌دلایل پایداری محدود می‌شود. روش‌ادمز-مولتن وسایر روشهای چندمرحله‌ای استاندارد نیز به‌همین صورت نمو h را محدود می‌کنند.

بررسی گسترده برای یافتن روشهای مناسب درحل معادلات دیفرانسیل سخت هنوز ادامه دارد. موفقترین‌روشها ظاهراً روشهای ضمنی‌هستند. برای مثال، کاربرد روش‌ذوزنقه‌ای (۵۲.۸) تاحدی با موفقیت همراه بوده است. برای این روش تمام نیمصفحهٔ منفی ناحیهٔ پایداری‌است، لذا h به‌شرایط پایداری محدود نمی‌شود (کتاب گیر[۳۰] را ببینید). با اعمال این روش برای یك دستگاه دومعادلهٔ دومجهولی به‌شکل

$$y_۱' = f_۱(x, y_۱, y_۲)$$
$$y_۲' = f_۲(x, y_۱, y_۲)$$

روش ذوزنقه‌ای چنین خواهد شد

$$y_{۱,n+۱} = y_{۱n} + \frac{h}{۲}\{f_۱(x_n, y_{۱n}, y_{۲n}) + f_۱(x_{n+۱}, y_{۱,n+۱}, y_{۲,n+۱})\}$$

$$y_{۲,n+۱} = y_{۲n} + \frac{h}{۲}\{f_۲(x_n, y_{۱n}, y_{۲n}) + f_۲(x_{n+۱}, y_{۱,n+۱}, y_{۲,n+۱})\}$$

(۸۹.۸)

با تخصیص این معادلات به‌دستگاه خطی (۸۸.۸)، خواهیم داشت

$$y_{۱,n+۱} = y_{۱n} + \frac{h}{۲}\{y_{۲n} + y_{۲,n+۱}\}$$

---

1. explodes

* منظور آن است‌که با انتخاب $h = ۰٫۰۰۳$، ناپایداری عددی ایجاد می‌شود یعنی مقادیر متغیرهای غیرمستقل مرتباً افزایش می‌یابند.ـم.

$$y_{۲,n+۱}=y_{۲n}+\frac{h}{۲}\{-۱۰۰۱y_{۲n}-۱۰۰۰y_{۱n}$$

$$+(-۱۰۰۱y_{۲,n+۱}-۱۰۰۰y_{۱,n+۱})\}$$

درحالت معمولی، این‌معادلات با روش بارستی حل‌می‌شوند، اما به‌علت خطی‌بودن‌معادلات، یك دستگاه ضمنی برحسب مجهولات $y_{۱,n+۱}$ و $y_{۲,n+۱}$ به‌صورت زیرمی‌توانیم به‌دست آوریم:

$$y_{۱,n+۱}-\frac{h}{۲}y_{۲,n+۱}=y_{۱n}+\frac{h}{۲}y_{۲n}$$

$$\frac{h}{۲}(۱۰۰۰)y_{۱,n+۱}+\left(۱+\frac{۱۰۰۱h}{۲}\right)y_{۲,n+۱}=y_{۲n}-\frac{h}{۲}\{۱۰۰۱y_{۲n}$$

$$+۱۰۰۰y_{۱n}\}\qquad(۹۰.۸)$$

اکنون h را برابر با ۰ر۱ انتخاب می‌کنیم، لذا دستگاه (۹۰.۸) چنین خواهد شد

$$y_{۱,n+۱}-۰ر۰۵y_{۲,n+۱}=y_{۱n}+۰ر۰۵y_{۲n}$$

$$۵۰y_{۱,n+۱}+۵۱ر۰۵y_{۲,n+۱}=-۴۹ر۰۵y_{۲n}-۵۰y_{۱n}\qquad(۹۱.۸)$$

به‌ازای $n=۰$ داریم، $y_{۱۰}=۱$، $y_{۲۰}=-۱$ و با توجه به‌دستگاه (۹۱.۸) خواهیم داشت

$$y_{۱۱}=۰ر۹۰۴۷۶۲\qquad y_{۲۱}=-۰ر۹۰۴۷۶۲$$

که با رعایت نمو بزرگی‌که به‌کار رفته، تقریب مناسبی برای جواب دقیق

$$y(۰ر۱)=e^{-۰ر۱}=۰ر۹۰۴۸۳۷$$

است. بعد از ۱۰ مرحله، به‌ازای $h=۰ر۱$ خواهیم داشت $y_۱(۱ر۰)\approx ۰ر۳۶۷۵۷۳$ که درمقایسه با جواب‌دقیق $y(۱ر۰)=e^{-۱ر۰}=۰ر۳۶۷۸۷۹$، جواب مطلوبی‌است.

ولی در استفاده ازروش ذوزنقه‌ای برای مسائل غیرخطی سخت، یك اصلاح‌اساسی وجود داردکه می‌باید به‌کار رود. برای یك معادلهٔ‌تنهای $y'=f(x,y)$، روش‌ذوزنقه‌ای ضمنی‌است و به‌وسیلهٔ رابطهٔ زیرتعریف می‌شود

$$y_{n+۱}=y_n+\frac{h}{۲}\{f(x_n,y_n)+f(x_{n+۱},y_{n+۱})\}\qquad(۹۲.۸)$$

این معادله، معادله‌ای‌است ضمنی با n ثابت‌که باید با یك روش بارستی نسبت به $y_{n+۱}$

حل‌شود. درحالت معمولی، بارست نقطهٔ ثابت که به‌صورت زیرتعریف می‌شود، مورداستفاده قرار می‌گیرد

$$y_{n+1}^{(m+1)} = y_n + \frac{h}{2}\{f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(m)})\} \qquad m = 0, 1, \ldots$$

که در آن $y_{n+1}^{(0)}$ تقریبی‌است برای y_{n+1} که با روشی غیر از روشی نظیر اویلر به‌دست می‌آید. همان‌گونه‌که درقسمت ۸.۸ نشان داده شد، اگر $\left|\frac{h}{2}\frac{\partial f}{\partial y}\right| < 1$، این بارست نقطهٔ ثابت همگرا خواهد شد و چون در مسائل سخت، $|\partial f/\partial y|$ بسیار بزرگ است، این روش بارستی برای همگرایی به‌نموهای خیلی کوچك نیاز پیدا می‌کند. ولی می‌توانیم معادلهٔ (۹۲.۸) را با روش بارستی نیوتن نسبت به y_{n+1} به‌صورت زیرحل کنیم: قرار می‌دهیم $\bar{y} = y_{n+1}$، و معادلهٔ (۹۲.۸) را دوباره به‌شکل

$$F(\bar{y}) = \bar{y} - y_n - \frac{h}{2} f(x_n, y_n) - \frac{h}{2} f(x_{n+1}, \bar{y}) = 0 \qquad (۹۳.۸)$$

می‌نویسیم. اگر $\bar{y}^{(0)} = y_{n+1}^{(0)}$ تقریب اولیه‌ای برای $\bar{y}$ باشد، آنگاه تقریبهایی متوالی طبق روش نیوتن و از راه بارستی

$$\bar{y}^{(m+1)} = \bar{y}^{(m)} + \frac{F(\bar{y}^{(m)})}{F'(\bar{y}^{(m)})} \qquad m = 0, 1, \ldots$$

به‌دست می‌آید، که با توجه به‌فرمول (۹۳.۸) نتیجه می‌شود:

$$F'(\bar{y}) = 1 - \frac{h}{2}\frac{\partial f}{\partial \bar{y}}(x_{n+1}, \bar{y})$$

دراین حالت، وقتی $\partial f/\partial y$ بزرگ و منفی باشد، که ویژگی عادی مسائل سخت‌است، در رابطه با همگرایی مشکلی وجود نخواهد داشت. ولی روش نیوتن برای یك معادلهٔ تنها به‌محاسبهٔ $\partial f/\partial y$ نیازدارد، و برای یك‌دستگاه معادلات به‌محاسبهٔ عناصرماتریس ژاکوبی. بنابراین می‌توان انتظارداشت که زیربرنامه‌های لازم برای حل معادلات دیفرانسیل سخت تاحدی پیچیده باشند.

برای یك دستگاه معادلات خطی به‌صورت

$$\mathbf{y}' = A\mathbf{y}$$

که درآن A ماتریس ثابتی‌است، سختی مسئله به‌توسط ویژه‌مقدارهای ماتریس A معین می‌شود. اگر ویژه‌مقدارهای ماتریس A منفی و خیلی ازهم فاصله داشته باشند، آنگاه دستگاه سخت بوده و می‌توان انتظارداشت که حل آن باروشهای معمولی مشکل باشد. برای مثال دستگاه (۸۸.۸)، ماتریس A به‌صورت

$$A=\begin{bmatrix} ۰ & ۱ \\ -۱۰۰۰ & -۱۰۰۱ \end{bmatrix}$$

است و ویژه‌مقدارهای آن برابر با -۱۰۰۰ و -۱.

برای دستگاههای غیرخطی کلیتر به‌شکل

$$\mathbf{y}'=\mathbf{f}(x,\mathbf{y})$$

سختی دستگاه با ویژه‌مقدارهای ماتریس ژاکوبی

$$\left(\frac{\partial \mathbf{f}}{\partial \mathbf{y}}\right)$$

معین می‌شود. برای بحث کاملتر درمورد مسائل سخت و حل آنها به‌روشهای دیگر، خوانند به‌کتاب گیر[۳۰] ارجاع داده می‌شود.

تمرین

۱-۱۳۰۸ سعی کنید با استفاده از روش مرتبهٔ ۴ رونگه-کوتا به‌ازای نموهای

$$h=۰٫۰۰۱, ۰٫۰۰۲, ۰٫۰۰۳, ۰٫۰۱$$

دستگاه(۸۸۰۸) را از $x=۰$ تا $x=۲$ حل کنید. تحقیق کنید که به‌ازای $h=۰٫۰۰۳$ و $h=۰٫۰۱$، جواب جهش می‌یابد درحالی که به‌ازای $h=۰٫۰۰۱$ و $h=۰٫۰۰۲$ برای جواب دقیق $y=e^{-x}$، تقریب مطلوبی به‌دست می‌آید.

۲-۱۳۰۸ برای دستگاه

$$y_۱'= \quad y_۲$$

$$y_۲'=-۲۰۰y_۱-۱۰۲y_۲$$

نشان دهید که ویژه‌مقدارهای ماتریس ضرایب -۲ و -۱۰۰ هستند و از این‌رو جواب عمومی به‌صورت زیراست

$$y_۱(x)=y(x)=Ae^{-۲x}+Be^{-۱۰۰x}$$

درشرایط $y(۰)=۱$ و $y'(۰)=-۲$ که متناظر با $y_۱(۰)=۱$ و $y_۲(۰)=-۲$ هستند، جواب دقیق برابراست با $y(x)=e^{-۲x}$. با استفاده‌از روش‌ذوزنقه‌ای و به‌ازای $h=۰٫۱$ این دستگاه را از $x=۰$ تا $x=۱$ حل و نتایج خود را با جواب دقیق مقایسه‌کنید.

۹

مسائل با مقدار مرزی[1] در معادلات دیفرانسیل معمولی

در فصل هشتم روشهای عددی برای حل مسائل با مقدار اولیه را بررسی کردیم. در این گونه مسائل کلیهٔ شرایط اولیه در یک نقطه داده شده‌اند. در این فصل مسائلی را بررسی می‌کنیم که در آنها، شرایط در بیش از یک نقطه مشخص شده‌اند. یک مسئلهٔ ساده از مسئلهٔ با مقدار مرزی مرتبهٔ دوم به‌صورت

$$y''(x)=y(x) \qquad y(0)=0 \qquad y(1)=1 \qquad (۱.۹)$$

و یک مثال از مسئلهٔ با مقدار مرزی مرتبهٔ چهارم به‌صورت

$$y^{iv}(x)+k\,y(x)=q \qquad (۲.۹\text{الف})$$

$$y(0)=y'(0)=0 \qquad (۲.۹\text{ب})$$

$$y(L)=y''(L)=0 \qquad (۲.۹\text{پ})$$

است. در اینجا y می‌تواند معرف انحنای میله‌ای به‌طول L باشد که بار یکنواخت q بر آن وارد شده‌است. شرط (۲.۹ ب) گویای آن است که انتهای میله در $x=0$ تثبیت شده‌است درحالی‌که (۲.۹ پ) حکایت از این دارد که میله در $x=L$ فقط تکیه داده شده‌است.

---

1. boundary-value problems

برای حل مسائلی از این قبیل سه روش را بررسی می‌کنیم: روش **تفاضلهای متناهی**[1] و روشهای «نشانه‌گیری[2]» که از روشهای فصل هشتم اقتباس شده‌اند و روش **پهلوی‌هم‌گذاری**[3].

۱۰۹ روشهای تفاضل متناهی

فرض می‌کنیم که یك معادلهٔ دیفرانسیل خطی از مرتبهٔ بالاتر از یك با شرایطی که در دو سر بازهٔ $[a, b]$ مشخص شده، داده شده است. بازهٔ $[a, b]$ را به N قسمت مساوی با پهنای h تقسیم می‌کنیم. قرار می‌دهیم $x_0 = a$ و $x_N = b$ و **نقاط داخلی شبکه**[4] را به‌صورت زیر تعریف می‌کنیم

$$x_n = x_0 + nh \qquad n = 1, 2, \dots N-1$$

مقادیر متناظر y در این نقاط شبکه، به‌صورت زیر نشان داده می‌شوند:

$$y_n = y(x_0 + nh) \qquad n = 0, 1, \dots, N$$

گاهی با نقاطی بیرون از بازهٔ $[a, b]$ سروکار خواهیم داشت. این نقاط، **نقاط خارجی شبکه**[5] نامیده می‌شوند. نقاطی که در سمت چپ x_0 قرار دارند با علامت $x_{-1} = x_0 - h$، $x_{-2} = x_0 - 2h$ و غیره نشان داده می‌شوند و آنهایی که در سمت راست x_N قرار دارند با $x_{N+1} = x_N + h$، $x_{N+2} = x_N + 2h$ و غیره. بدیهی است که مقادیر متناظر y در نقاط خارجی شبکه با y_{-1}، y_{-2}، y_{N+1}، y_{N+2} و غیره نشان داده می‌شوند.

برای حل یك مسئلهٔ با مقدار مرزی با روش تفاضلهای متناهی، به‌جای همهٔ مشتقهایی که در معادله و نیز در شرایط مرزی ظاهر می‌شوند، تقریبی مناسب از تفاضلها گذاشته می‌شود. معمولا تفاضلهای مرکزی، به‌علت دقت بیشتری که عاید می‌سازند، ترجیح داده می‌شوند. برخی نمونه‌های تقریبی تفاضل مرکزی در زیر داده شده‌اند (فصل ۷ را ببینید):

$$y'(x_n) \approx \frac{y_{n+1} - y_{n-1}}{2h}$$

$$y''(x_n) \approx \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2}$$

$$y^{iv}(x_n) \approx \frac{y_{n+2} - 4y_{n+1} + 6y_n - 4y_{n-1} + y_{n-2}}{h^4} \qquad (۳.۹)$$

در هر مورد نمایش تفاضلهای محدود، یك تقریب $\mathcal{O}(h^2)$ برای مشتق مربوطه است. برای

---

1. finite differences
2. shooting
3. collocation
4. interior mesh points
5. exterior mesh points

نشان دادن این روش، معادلهٔ دیفرانسیل خطی مرتبهٔ ۲

$$y''(x)+f(x)y'+g(x)y=q(x) \qquad (۴.۹)$$

را درشرایط مرزی

$$y(x_0)=\alpha \qquad (۵.۹)$$

$$y(x_N)=\beta \qquad (۶.۹)$$

درنظر می‌گیریم. تقریب تفاضل متناهی برای معادلهٔ (۴.۹) به‌صورت زیراست

$$\frac{y_{n-1}-2y_n+y_{n+1}}{h^2}+\frac{f(x_n)(y_{n+1}-y_{n-1})}{2h}+g(x_n)y_n=q(x_n)$$

$$n=1,2,\dots,N-1$$

از ضرب طرفین معادله در h^2 و قراردادن $f(x_n)=f_n$ به‌ازای $n=1,\dots,N-1$ و گروه‌بندی جملات، خواهیم داشت:

$$\left(1-\frac{h}{2}f_n\right)y_{n-1}+(-2+h^2g_n)y_n+\left(1+\frac{h}{2}f_n\right)y_{n+1}=h^2q_n$$

$$n=1,2,\dots,N-1 \qquad (۷.۹)$$

از آنجایی‌که y_0 و y_N با شرایط (۵.۹) و(۶.۹) مشخص شده‌اند، بنابراین معادلهٔ(۷.۹) یك دستگاه خطی‌است با $N-1$ معادله و $N-1$ مجهول y_n، $(n=1,2,\dots,N-1)$. ازقراردادن α به‌جای y_0 و β به‌جای y_N درمعادلهٔ(۷.۹)، دستگاه معادلات فوق به‌صورت زیرین درمی‌آید

$$(-2+h^2g_1)y_1+\left(1+\frac{h}{2}f_1\right)y_2=h^2q_1-\left(1-\frac{h}{2}f_1\right)\alpha$$

$$\left(1-\frac{h}{2}f_2\right)y_1+(-2+h^2g_2)y_2+\left(1+\frac{h}{2}f_2\right)y_3=h^2q_2$$

$$\left(1-\frac{h}{2}f_3\right)y_2+(-2+h^2g_3)y_3+\left(1+\frac{h}{2}f_3\right)y_4=h^2q_3$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$\left(1-\frac{h}{2}f_{N-2}\right)y_{N-3}+(-2+h^2g_{N-2})y_{N-2}+\left(1+\frac{h}{2}f_{N-2}\right)y_{N-1}$$

$$=h^2q_{N-2}$$

$$\left(۱-\frac{h}{۲}f_{N-۱}\right)y_{N-۲}+(-۲+h^۲g_{N-۱})y_{N-۱}=h^۲q_{N-۱}-\left(۱+\frac{h}{۲}f_{N-۱}\right)\beta$$

(۸.۹)

البته می‌توان ضرایب دستگاه (۸.۹)را محاسبه کرد. زیرا $f(x)$، $g(x)$، و $q(x)$ توابع معلومی از x هستند. اکنون می‌توان این دستگاه خطی را با هریك از روشهایی که درفصل ۴ مورد بحث قرار گرفت، حل کرد. برای نمایش دستگاه فوق به‌شکل ماتریسی داریم $A\mathbf{y}=\mathbf{b}$، که در آن $\mathbf{y}=[y_۱, y_۲, \ldots, y_{N-۱}]^T$ معرف بردار مقادیرمجهول و $\mathbf{b}$ معرف بردار مقادیر معلوم سمت راست معادلهٔ (۸.۹) است و A ماتریس ضرایب. دراین حالت ماتریسA سه قطری و از مرتبهٔ $N-۱$ است. با شکل‌خاص زیر

$$A=\begin{bmatrix} d_۱ & c_۱ & & & & & \\ a_۲ & d_۲ & c_۲ & & & & \\ & a_۳ & d_۳ & c_۳ & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & a_{N-۲} & d_{N-۲} & c_{N-۲} & & \\ & & & a_{N-۱} & d_{N-۱} & & \end{bmatrix}$$

دستگاه $A\mathbf{y}=\mathbf{b}$ را می‌توان مستقیماً با استفاده از الگوریتم ۳.۴ی بخش ۲.۴ حل کرد. تنها کافی است‌که به‌جای n، $N-۱$ بگذاریم و $\mathbf{x}$ و $\mathbf{y}$ را معین‌کنیم و فرمولهای بازگشتی الگوریتم ۳.۴ را به‌کار بندیم.

حال به‌شرایط مرزی بر گردیم و ببینیم که اگر به‌جای شرط (۵.۹)، شرط

$$y'(x_۰)+\gamma y(x_۰)=۰ \qquad (۹.۹)$$

را در $x=x_۰$ قراردهیم، چه تأثیری بردستگاه (۸.۹) خواهدگذاشت.

اگر به‌جای $y'(x_۰)$ یك تفاضل پیشرو بگذاریم خواهیم داشت

$$\frac{y(x_۰+h)-y(x_۰)}{h}+\gamma y(x_۰)=۰$$

یا به‌صورت مرتب شده

$$y_۱+(-۱+\gamma h)y_۰=۰ \qquad (۹.۹\text{الف})$$

اکنون اگر معادلهٔ (۷.۹) را به‌ازای $n=1$ بنویسیم، وسپس به‌جای y_0، $y_1/(1-\gamma h)$ را قراردهیم خواهیم داشت

$$\left[(-2+h^2g_1)+\frac{1-(h/2)f_1}{1-\gamma h}\right]y_1+\left(1+\frac{h}{2}f_1\right)y_2=h^2q_1 \qquad (10.9)$$

حال دردستگاه(۸.۹) به‌جای‌اولین‌معادله می‌توان معادلهٔ(۱۰.۹) راگذارد. تمامی‌معادلات دیگر از دستگاه (۸.۹) بدون تغییر باقی می‌مانند، و با استفاده از الگوریتم ۳.۴ می‌توان دستگاه حاصل را دوباره حل کرد. ولی باید توجه داشت‌که رابطهٔ (۹.۹الف) تنها یك تقریب $\mathcal{O}(h)$ برای شرط مرزی (۹.۹) است (به‌قسمت ۱.۷، مراجعه‌کنید).

دراین‌صورت دقت جواب نیز از مرتبهٔ h خواهد بود. برای به‌دست‌آوردن جوابی که درهمه‌جا از مرتبهٔ h^2 باشد، به‌جای رابطهٔ (۹.۹) تقریب

$$\frac{y(x_0+h)-y(x_0-h)}{2h}+\gamma y(x_0)=0$$

یا مرتب شدهٔ آن

$$y_1-y_{-1}+2h\gamma y_0=0 \qquad (11.9)$$

را قرار می‌دهیم. از آنجاکه یك نقطهٔ خارجی y_{-1} در رابطه واردکرده‌ایم، اکنون باید y_0 و نیز $y_1, y_2, \ldots, y_{N-1}$ را به‌عنوان مجهول درنظر بگیریم. امـا چون N مجهول داریم می‌باید N معادله نیز داشته باشیم. با در نظرگرفتن $n=0$ در (۷.۹) می‌توان یك معادلهٔ اضافی به‌دست آورد. سپس اگر با استفاده از رابطهٔ (۱۱.۹)، y_{-1} راحذف کنیم، برای دومعادلهٔ اول خواهیم داشت

$$\left[2h\gamma\left(1-\frac{h}{2}f_0\right)+(-2+h^2g_0)\right]y_0+2y_1=h^2q_0 \qquad n=0$$

$$\left(1-\frac{h}{2}f_1\right)y_0+(-2+h^2g_1)y_1+\left(1+\frac{h}{2}f_1\right)y_2=h^2q_1 \qquad n=1$$

بقیهٔ معادلات همان معادلاتی هستندکـه در دستگاه (۸.۹) ظاهرشده‌اند. این دستگاه بـاز سه‌قطری ولی حالا از مرتبهٔ N است ومی‌توانند صریحاً بااستفاده ازالگوریتم۳.۴، دوباره حل‌شود.

به‌روشنی دیده می‌شودکـه دقت حاصله از روش تفاضلهای متناهی به‌کوچك بودن شبکه و مرتبهٔ تقریب تفاضـل محدود، بستگی خواهد داشت. هرقـدر شبکه‌کوچکتر شود، به‌همان نسبت تعداد معادلاتی که باید حل شوند افزایش می‌یابند. درنتیجه مدت لازم برای کامپیوتر ممکن‌است زیاد شود و مشکل بتوان دقت خوبی به‌دست آورد. برای شبکه‌هایی

بـا همان اندازه، استفاده از تقریبهای مرتبـهٔ بالاتـر، موجب دقت بیشتری می‌شود، امـا به‌پیچیدگـی قابل‌ملاحظه‌ای، بخصوص نزدیك به‌نقاط انتهایی بازهٔ مورد نظر کـه مقادیر خارجی در آن نقاط نامعلوم هستند، منجر می‌شود.

درعمل بهتر است که دستگاه معادلات به‌ازای چندین مقدار مختلف h حل‌شوند. در این صورت مقایسهٔ جوابها درنقاط واحدی از شبکه، دقت حاصل را نشان می‌دهند. بعلاوه از روند برونیابی‌کـه در بخش ۵.۷ شرح داده شد، می‌توان معمولا برای کسب جوابهای بهبودیافته استفاده کرد. به‌موجب توافقی کـه برای حل دستگاه معادلات تفاضلهای متناهی شده‌است، در برونیابی به‌سمت حد به‌صورت زیرعمل می‌شود. گیریم $y_h(x)$ معرف جواب تقریبی مسئله بامقدار مرزی برروی تعداد $N=(b-a)/h$ زیر تقسیم در بازهٔ $[a,b]$ در یکی از نقاط x ازشبـکه باشد. گیریم $y_{h/2}(x)$ جواب تقریبی همـان مسئله روی تعـداد $2N=(b-a)/(h/2)$ زیر تقسیـم از بـازهٔ $[a,b]$ باشد. اکنـون در $N-1$ نقطـهٔ $x_1=a+h$، $x_2=a+2h$، $\ldots$، $x_{N-1}=a+(N-1)h$، دوجواب تقریبی $y_h(x_n)$ و $y_{h/2}(x_n)$ را در دست داریم. بـا استفاده از روش برونیابـی در این نقـاط خواهیم داشت

$$y(x_n)\approx y_n^{(1)}:=\frac{4y_{h/2}(x_n)-y_h(x_n)}{3}\qquad n=1,2,\ldots,N-1$$

معمولا برونیابی فوق، بهبودی قابل‌توجهی درتقریب ایجاد می‌کند.

□ **مثال ۱۰.۹:** بـا استفاده از روش تفاضلهای متناهی، مسئلهٔ بامقدار مرزی (۱.۹) را حل کنید.

هر گـاه بگیریم $f(x)=0$، $g(x)=-1$، $q(x)=0$، از قراردادن $y_0=0$، $y_N=1$ در دستگاه (۸.۹)، دستگاه زیر را به‌دست می‌آوریم

$$\begin{aligned}
(-2-h^2)y_1+y_2 &= 0\\
y_{n-1}+(-2-h^2)y_n+y_{n+1} &= 0 \qquad n=2,3,\ldots,N-2\\
y_{N-2}+(-2-h^2)y_{N-1} &= -1
\end{aligned}$$

این‌دستگاهی‌است از $N-1$ معادله با $N-1$ مجهول: y_1، y_2، $\ldots$، y_{N-1}. این دستگاه، بااستفاده از زیر برنامه‌ای براساس‌الگوریتم ۳.۴ و به‌ازای مقادیر $h=0.1$ و $h=0.05$ روی‌کامپیوتر IBM 7090 حل‌شده بود. نتایج حاصله در زیر آمده‌است.

نتایج کامپیوتری برای مثال ۱۰۹

XN	YN(H = 0.05)	YN(H = 0.10)	YN(1)	Y(XN)
0	0	0	0	0
0.05	.04256502			.04256363
0.10	.08523646	.08524469	.08523372	.08523369
0.15	.12812098			.12811689
0.20	.17132582	.17134184	.17132048	.17132045
0.25	.21495896			.21495239
0.30	.25912950	.25915240	.25912187	.25912183
0.35	.30394787			.30393920
0.40	.34952610	.34955449	.34951663	.34951659
0.45	.39597815			.39596794
0.50	.44342014	.44345213	.44340946	.44340942
0.55	.49197068			.49195965
0.60	.54175115	.54178427	.54174010	.54174004
0.65	.59288599			.59287506
0.70	.64550304	.64553425	.64549263	.64549258
0.75	.69973386			.69972418
0.80	.75571401	.75573958	.75570550	.75570543
0.85	.81358345			.81357635
0.90	.87348684	.87350228	.87348166	.87348163
0.95	.93557395			.93557107
1.00	1	1	1	1

ستون چهارم، مقادیر برونیابی‌شده‌ای را نشان می‌دهد که دربازه‌هایی به‌طول ۰٫۱ از فرمول

$$y_n^{(1)} = \frac{۴y_{h/۲}(x_n) - y_h(x_n)}{۳}$$

به‌دست آمده‌اند. مقادیر داده شده درستون آخر ازجواب دقیق مسئلهٔ

$$y(x) = \frac{\sinh x}{\sinh ۱}$$

به‌دست آمده‌اند. این نتایج نشان می‌دهندکه به‌ازای $h = ۰٫۱$، جواب از سه تا چهار رقم بامعنی و به ازای $h = ۰٫۰۵$ از چهار تا پنج رقم بامعنی صحیح است. در حالی که جواب به دست آمده از برونیابی در حدود هفت رقم بامعنی صحیح دارد. برای به‌دست‌آوردن دقتی برابر با هفت رقم بامعنی، بدون استفاده از برونیابی لازم است که بازهٔ $[۰, ۱]$ تقریباً به‌یك شبکهٔ ۱۰۰ نقطه‌ای $(h = ۰٫۰۱)$ تقسیم شود. □

تمرین

۱۰۹-۱ مسئلهٔ با مقدار مرزی

$$\frac{d^2y}{dx^2}+y=0 \qquad y(0)=0 \qquad y(1)=1$$

را از طریق روشهای تفاضلی حل کنید. h را برابر ۱/۴ بگیرید و دستگاه حاصل را با استفاده از ماشین حساب جیبی حل نمایید.

جواب: $y_1=0.2923$، $y_2=0.5702$، $y_3=0.8104$

این جواب را با جواب دقیق $y=(\sin x)/(\sin 1)$ مقایسه کنید.

۳-۱۰.۹ با استفاده از یك شبکهٔ $h=0.1$، مسئلهٔ با مقدارمرزی (۱۰.۹) را وقتی شرط $y'(0)+y(0)=0$ جایگزین شرط $y(0)=0$ شده باشد حل کنید.

۳-۱۰.۹ یك دستگاه $\Theta(h^2)$ تفاضل متناهی برای تقریب‌زدن جواب مسئلهٔ با مقدار مرزی زیر بنویسید

$$y''+xy'+y=2x \qquad y(0)=1 \qquad y(1)=0$$

h را برابر ۰.۱ بگیرید و دستگاه را به‌شکل ماتریسی بنویسید و سپس با استفاده از یك برنامهٔ کامپیوتری براساس الگوریتم ۳.۴، این دستگاه را حل کنید.

۴-۱۰.۹ نشان دهید که می‌توان روش بارستی گاوس-زایدل را نیز برای حل دستگاه مثال ۱۰.۹ به‌کاربرد و این جواب حاصل از راه روش بارستی را با دقتی برابر با چهار رقم بامعنی به‌دست آورد. برای این مسئله، آیا روش مستقیم کاراتر از روش بارستی است؟

۵-۱۰.۹ مسئلهٔ با مقدارمرزی

$$y''+2y'+y=x \qquad y(0)=0 \qquad y(1)=0$$

را به‌ازای $h=1/8$ و $h=1/16$ با استفاده از روشهای تفاضل حل کنید، و نتیجه را با استفاده از برونیابی بهبود بخشید.

۲.۹ روشهای نشانه‌گیری

درمسائل خطی با مقدار مرزی، از چندین روش می‌توان استفاده کرد. روش تفاضلها که در بخش قبل بیان شد، در این گونه موارد روش مناسبی است. درروشهای دیگر سعی بر این است که جوابهای مستقل خطی برای معادلات دیفرانسیل به‌دست آورند و آنها را طوری ترکیب کنند که درشرایط مرزی صدق نمایند. برای معادلات غیرخطی روش اخیر را نمی‌توان به‌کاربرد. روش تفاضلها را می‌توان برای مسائل غیرخطی تطبیق داد، اما این امر مستلزم تخمین‌زدن جوابهای آزمایشی وسپس بهترساختن آنها به‌وسیلهٔ روندهای بارستی است. در این کارعلاوه برآنکه برنامه‌نویسی پیچیده‌ای لازم است، تضمینی هم برای همگرایی بارستها

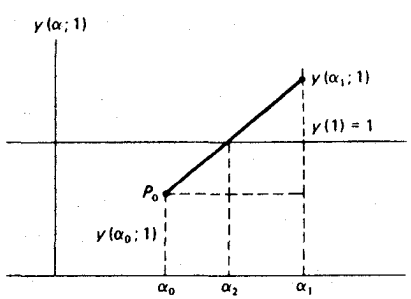
وجود ندارد. روش نشانه‌گیری که در این بخش توضیح داده می‌شود هم برای مسائل خطی و هم برای مسائل غیرخطی قابل استفاده است. در اینجا نیز هیچ تضمینی برای همگرایی وجود ندارد، اما استفاده از این روش بسیار ساده است و زمانی که این روش همگرا باشد، معمولا کاراتر از روشهای دیگر است.

باردیگر مسئله‌ای را که در (۱.۹) داده شده بود در نظر می‌گیریم. می‌خواهیم روشهای با مقدار اولیه را که در فصل ۸ مورد بحث قرار گرفته بود به‌کار بریم، اما برای انجام این کار می‌باید هم مقدار $y(0)$ در دست باشد و هم مقدار $y'(0)$. از آنجایی که $y'(0)$ قبلا معین شده است، آن را به‌صورت پارامتر مجهولی، مثلا α، که باید چنان تعیین شود که جواب حاصل، مقدار $y(1)$ را با دقت مطلوبی به‌دست دهد، در نظر می‌گیریم. بنابراین شیب اولیه را تخمین می‌زنیم و یک روند بارستی چنان بنا می‌کنیم که به‌سمت شیب دقیق همگرا شود. گیریم α_0 و α_1 دو تخمین برای شیب اولیهٔ $y'(0)$ باشند، و فرض می‌کنیم $y(\alpha_0;1)$ و $y(\alpha_1;1)$ مقادیری از y باشند که از انتگرالگیری معادلهٔ دیفرانسیل در $x=1$ به‌دست آمده‌اند. از لحاظ نموداری می‌توان این وضعیت را به‌طور ترسیمی با شکلهای ۱.۹ و ۲.۹ نمایش داد.

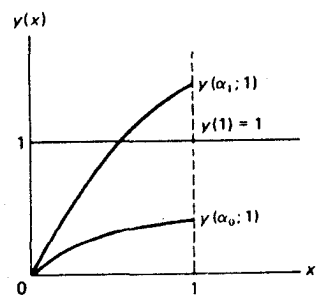
در شکل ۱.۹، جوابهای با مسائل مقدار اولیه ترسیم شده‌اند، درحالی که در شکل ۲.۹، $y(\alpha;1)$ به‌صورت تابعی از α ترسیم شده است. اما یک تقریب معمولا بهتر برای α را می‌توان از راه درونیابی خطی به‌دست آورد. محل تقاطع خط واصل بین نقاط P_0 و P_1 با خط $y(1)=1$ مختص[1] αای دارد که با رابطهٔ

$$\alpha_2=\alpha_0+(\alpha_1-\alpha_0)\frac{1-y(\alpha_0;1)}{y(\alpha_1;1)-y(\alpha_0;1)} \qquad (۱۲.۹)$$

معین می‌شود. حال با استفاده از مقادیر اولیهٔ $y(0)=0$ و $y'(0)=\alpha_2$، از معادلهٔ دیفرانسیل انتگرال می‌گیریم تا $y(\alpha_2;1)$ به‌دست آید. باردیگر با استفاده از درونیابی خطی بر پایهٔ



شکل ۲.۹



شکل ۱.۹

1. coordinate

α_1 و α_2، تقریب دیگر α_2 را می‌توانیم به‌دست آوریم. این روند تا زمانی که همگرایی به‌دست آید، یعنی تاجایی که $y(\alpha_i;1)=y(1)$ با ۱ تا تعداد ارقام مطلوب اعشاری تطابق نماید، ادامه می‌یابد. هیچ تضمینی برای همگرایی این روند بارستی وجود ندارد. واضح است که سرعت همگرایی به‌این‌امر که تا چه اندازه براوردهای اولیه خوب بوده بستگی دارد. این براوردها گاهی با بررسیهای فیزیکی مسئله و زمانی با نمایش ترسیمی سادهٔ جواب، به‌دست می‌آیند.

روند بالا برای یك مسئلهٔ کلی مرتبهٔ دوم با مقدار مرزی

$$y''=f(x,y,y') \qquad y(0)=y_0 \qquad y(b)=y_b \qquad (۱۳.۹)$$

در الگوریتم ۱.۹ خلاصه شده‌است.

الگوریتم ۱.۹: روش نشانه‌گیری برای مسائل با مقدارمرزی مرتبهٔ دوم

۱. گیریم α_k تقریبی برای شیب اولیهٔ مجهول $y'(0)=\alpha$ باشد. (دو مقدار اولیهٔ α_0 و α_1 را با استفاده از شهود فیزیکی انتخاب می‌کنیم.)

۲. با استفاده از یکی از روشهای فصل۸، مسئلهٔ با مقدار اولیهٔ

$$y''=f(x,y,y') \qquad y(0)=y_0 \qquad y'(0)=\alpha_k$$

را از $x=0$ تا $x=b$ حل می‌کنیم. جواب درنقطهٔ $x=b$ را $y(\alpha_k;b)$ می‌نامیم.

۳. تقریب بعدی را از درونیابی خطی زیر به‌دست می‌آوریم

$$\alpha_{k+1}=\alpha_{k-1}+(\alpha_k-\alpha_{k-1})\frac{y_b-y(\alpha_{k-1};b)}{y(\alpha_k;b)-y(\alpha_{k-1};b)}$$

$$k=1,2,\cdots$$

۴. مراحل۲ و ۳ را تا برقراری نامساوی $|y(\alpha_k;b)-y_b|<\varepsilon$ به‌ازای یك ε که قبلا معین‌شده، ادامه می‌دهیم.

روش بارستی که در الگوریتم ۱.۹ به‌کار رفته‌کاربردی‌است از روش خط قاطع که درفصل۳ توضیح داده شده بود.

این روش درمورد دستگاه معادلات مرتبهٔ بالاتر پیچیده‌تر و به‌دست آوردن همگرایی دشوارتر می‌شود. وضعیت کلی برای دستگاه غیرخطی را می‌توان به‌صورت زیر نمایش‌داد. یك دستگاه چهارمعادلهٔ چهارمجهولی:

$$\begin{aligned} x'&=f(x,y,z,w,t)\\ y'&=g(x,y,z,w,t)\\ z'&=h(x,y,z,w,t)\\ w'&=l(x,y,z,w,t) \end{aligned} \qquad (۱۴.۹)$$

را که t در آن معرف متغیر مستقلی است، در نظر می‌گیریم. در $t=0$ دو شرط، مثلا

$$x(0)=x_0$$

$$y(0)=y_0$$

و در $t=T$، دو شرط، مثلا

$$z(T)=z_T$$

$$w(T)=w_T$$

داده شده‌اند. گیریم $z(0)=\alpha$ و $w(0)=\beta$ مقادیر اولیه و دقیق برای $z(0)$ و $w(0)$ باشند، و فرض می‌کنیم α_0 و β_0 تخمینهایی برای مقادیر اولیه باشند. حال از دستگاه (۱۴.۹) انتگرال می‌گیریم و مقادیر z و w ای حاصل در $t=T$ را با $z(\alpha_0, \beta_0; T)$ و $w(\alpha_0, \beta_0; T)$ نشان می‌دهیم.

چون z و w در $t=T$ آشکارا توابعی از α و β هستند، می‌توانیم $z(\alpha, \beta; T)$ و $w(\alpha, \beta; T)$ را نسبت به‌این دو متغیر با جملات خطی، به‌سری تیلر بسط دهیم

$$z(\alpha, \beta; T)=z(\alpha_0, \beta_0; T)+(\alpha-\alpha_0)\frac{\partial z}{\partial \alpha}(\alpha_0, \beta_0; T)$$

$$+(\beta-\beta_0)\frac{\partial z}{\partial \beta}(\alpha_0, \beta_0; T)$$

$$w(\alpha, \beta; T)=w(\alpha_0, \beta_0; T)+(\alpha-\alpha_0)\frac{\partial w}{\partial \alpha}(\alpha_0, \beta_0; T)$$

$$+(\beta-\beta_0)\frac{\partial w}{\partial \beta}(\alpha_0, \beta_0; T) \qquad (۱۵.۹)$$

می‌توانیم $z(\alpha, \beta; T)$ و $w(\alpha, \beta; T)$ را برابر مقادیر مطلوبشان z_T و w_T قرار دهیم، اما پیش از اینکه بتوانیم (۱۵.۹) را نسبت به‌مقادیر تصحیحی $\alpha-\alpha_0$ و $\beta-\beta_0$ حل کنیم، باید مشتقهای جزیی در (۱۵.۹) را به‌دست آوریم. از آنجایی که جوابهای z و w معلوم نیستند، لذا نمی‌توانیم به‌طور تحلیلی، این مشتقهای جزیی را به‌دست آوریم. ولی، می‌توانیم مقادیر عددی تقریبی آنها را معین کنیم. برای این کار دستگاه (۱۴.۹) را یکبار با شرایط اولیهٔ x_0، y_0، α_0، β_0، و یکبار با شرایط x_0، y_0، $\alpha_0+\Delta\alpha_0$، β_0، و سپس با شرایط x_0، y_0، α_0، $\beta_0+\Delta\beta_0$ که در آن $\Delta\alpha_0$ و $\Delta\beta_0$ نمو‌های کوچکی هستند حل می‌کنیم. از حذف متغیرهای x_0 و y_0 که در طی عملیات ثابت می‌مانند، خارج‌قسمتهای تفاضلی زیر را تشکیل می‌دهیم

$$\frac{z(\alpha_0, \beta_0+\Delta\beta_0; T)-z(\alpha_0, \beta_0; T)}{\Delta\beta_0} \approx \frac{\partial z}{\partial \beta}(\alpha_0, \beta_0; T)$$

$$\frac{w(\alpha_0, \beta_0+\Delta\beta_0; T)-w(\alpha_0, \beta_0; T)}{\Delta\beta_0} \approx \frac{\partial w}{\partial \beta}(\alpha_0, \beta_0; T)$$

$$\frac{z(\alpha_0+\Delta\alpha_0, \beta_0; T)-z(\alpha_0, \beta_0; T)}{\Delta\alpha_0} \approx \frac{\partial z}{\partial \alpha}(\alpha_0, \beta_0; T)$$

$$\frac{w(\alpha_0+\Delta\alpha_0, \beta_0; T)-w(\alpha_0, \beta_0; T)}{\Delta\alpha_0} \approx \frac{\partial w}{\partial \alpha}(\alpha_0, \beta_0; T)$$

بعدازاینکه z_T را جایگزین $z(\alpha, \beta, T)$ و w_T را جایگزین $w(\alpha, \beta, T)$ کردیم، می‌توانیم (۱۵.۹) را نسبت به‌مقادیر تصحیحی $\delta\alpha_0 = \alpha - \alpha_0$ و $\delta\beta_0 = \beta - \beta_0$ حل کنیم و براوردهای جدید $\alpha_1 = \alpha_0 + \delta\alpha_0$ و $\beta_1 = \beta_0 + \delta\beta_0$ را برای پارامترهای α و β به‌دست آوریم. اکنون تمامی روند فوق با شروع ازمقادیر x_0، y_0، α_1، β_1 به‌عنوان شرایط اولیه تکرار شده‌است.

بنابراین هر بارست مشتمل برسه‌بار حل دستگاه (۱۴.۹) است. درحالت کلی اگر تعداد n پارامتر اولیهٔ مجهول وجود داشته باشد، هر بارست مستلزم $n+1$ جواب دستگاه اصلی‌است. روشی‌که در اینجا به‌کارگرفته شده، باروش‌اصلاحی‌نیوتن برای‌یافتن‌ریشه‌های معادلات چند متغیری هم‌ارز است (به‌بخش۲.۵ نگاه‌کنید.)

درحل باکامپیوتر، مسائل با مقدار مرزی یکی از دشوارترین رده‌های مسائل است. دراین مورد هیچ تضمینی برای همگرایی وجود ندارد، تخمینهای اولیهٔ خوبی باید در دسترس باشند، و معمولا آزمون و خطای فراوان و نیز مقدار زیادی وقت برای‌کامپیوتر لازم‌است.

□ **مثال۲.۹:** با استفاده از روش نشانه‌گیری، مسئلهٔ (۱.۹) را حل کنید. این‌کار را با تقریبهای اولیهٔ $\alpha_0 = 0.3$ و $\alpha_1 = 0.4$ برای $y'(0)$ و $h = 0.1$ شروع کنید.

با استفاده از زیربرنامهٔ استانداردِ RK۴* برای حل معادلات دیفرانسیل‌که در فصل۸ توضیح داده شد و همراه با درونیابی‌خطی براساس رابطهٔ (۱۲.۹)، جوابهای زیربه‌دست آمده‌اند. عمل بارستی درشرط $|\alpha_{k+1} - \alpha_k| < 1 \times 10^{-6}$ متوقف شده‌است.

---

* مرتبهٔ چهارم رونگه‌ـکوتا.ـم.

k	α_k	$y(\alpha_k;1)$
۰	۰٫۳۰۰۰۰۰۰۰	۰٫۳۵۲۵۶۰۷۷
۱	۰٫۴۰۰۰۰۰۰۰	۰٫۴۷۰۰۸۱۰۳
۲	۰٫۸۵۰۹۱۷۱۲	۰٫۹۹۹۹۹۹۹۹
۳	۰٫۸۵۰۹۱۷۱۲	۰٫۹۹۹۹۹۹۹۹

مقدار دقیق y' در $x=0$ برابراست با $\sinh^{-1}1=0.8509181303$. برای این‌مسئله همگرایی بسیارسریع‌است. باتوجه به‌نمو تخمینی‌که درنظرگرفته شده‌است، دقت به‌دست آمده به‌طور استثنایی خوب‌است. با استفاده از روشهای تفاضل متناهی بخش۱۰.۹، برای به‌دست‌آوردن دقت قابل‌مقایسه، نمو $h=0.01$ لازم خواهد بود. مع‌هذا روش تفاضل متناهی باز از لحاظ محاسبه‌کاراتر است. □

□ **مثال۳.۹:** با استفاده از روش نشانه‌گیری، مسئلهٔ با مقدار مرزی غیرخطی زیر را حل کنید

$$yy''+1+y'^2=0 \qquad y(0)=1 \qquad y(1)=2 \qquad (16.9)$$

حل: گیریم $\alpha_0=0.5$، $\alpha_1=1.0$ دوتقریب برای شیب مجهول $y'(0)$ باشند. باز با استفاده از زیر برنامهٔ RK۴ برای حل معادلات دیفرانسیل و درونیابی خطی به‌ازای نمو $h=1/64$ نتایج زیر به‌دست آمده‌اند:

α_i	$y(\alpha_i;1)$
۰٫۵۰۰۰۰۰۰۰	۰٫۹۹۹۹۹۹۹
۰٫۹۹۹۹۹۹۹	۱٫۴۱۴۲۱۳۳
۱٫۷۰۷۱۰۷۱	۱٫۸۴۷۷۵۸۲
۱٫۹۵۵۴۱۱۸	۱٫۹۷۷۵۷۸۶
۱٫۹۹۸۲۹۶۸	۱٫۹۹۹۱۴۶۳
۱٫۹۹۹۹۹۴۰	۱٫۹۹۹۹۹۵۲
۲٫۰۰۰۰۰۳۵	۲٫۰۰۰۰۰۰۰

شیب دقیق در $x=۰$ برابر است با $y'(۰)=۲$. بعد از هفت بارست مشاهده می‌شود که شیب اولیه تا شش رقم بامعنی صحیح است، در حالی که مقدار y در $x=۱$ حداقل تا هفت رقم بامعنی صحیح است. با استفاده از درونیابی درجهٔ دوم، بعد از سه بارست اول همگرایی می‌توانست تسریع شود.

آشکار است که تعداد بارست لازم، به‌انتخاب تقریبهای اولیهٔ $\alpha_۰$، $\alpha_۱$ بستگی دارد. این تقریبها را می‌توان گاهی از ملاحظات نموداری یا فیزیکی به‌دست آورد. □

تمرین

۱-۳۰۹ جواب عددی معادلهٔ زیر را به‌دست آورید

$$۲yy''-y'^۲+۴y^۲=۰ \qquad y\left(\frac{\pi}{۶}\right)=\frac{۱}{۴} \qquad y\left(\frac{\pi}{۲}\right)=۱$$

به‌عنوان تقریبهای اولیه برای $y'(\pi/۶)$، بگیرید $\alpha_۰=۰٫۵$ و $\alpha_۱=۰٫۸$، و بارستها را تا جایی ادامه دهید که شرط در $x=\pi/۲$ تا پنج رقم بامعنی صحیح باشد.

جواب: $y=(\sin x)^۲$؛ شیب اولیه برابر است با $y'(\pi/۶)=\sqrt{۳}/۲$.

۲-۳۰۹ درمثال ۳.۹ با استفاده از درونیابی درجهٔ دوم براساس $\alpha_۰$، $\alpha_۱$، $\alpha_۲$ تقریب بعدی را به‌دست آورید. چند بارست صرفه‌جویی شده‌است؟

۳-۳۰۹ با استفاده از روش نشانه‌گیری، مسائل زیر را حل کنید:

(الف) $$y''=۲y^۳,\ y(۱)=۱,\ y(۲)=\frac{۱}{۲}$$

$y'(۱)=۰$ را به‌عنوان اولین تخمین بگیرید. (جواب دقیق $y=۱/x$).

(ب) $$y''=e^y,\ y(۰)=y(۱)=۰$$

$y'(۰)=۰$ را به‌عنوان اولین تخمین بگیرید.

۳.۹ روشهای پهلوی هم‌گذاری

درسالهای اخیر برای حل مسائل با مقدارمرزی، هم درحالت یك‌بعدی وهم درمورد ابعاد بالاتر، توجه زیادی بر روشهای تقریبی‌متمرکزشده‌است. در این روشهای تقریبی، به‌جای آنکه جواب دریك مجموعه از نقاط گسسته جستجوشود، سعی بر آن است که یك ترکیب خطی ازتوابع مستقل خطی به‌دست آورندکه تقریبی را برای جواب ارائه دهد. درحقیقت

این اندیشهٔ اصلی ریشهٔ بسیار قدیمی دارد و به گیلرکین[1] و ریتس[2] [۳۱] برمی‌گردد، اما دراین اواخر این اندیشه‌ها شکل جدیدی، تحت عنوان «عنصر متناهی[3]» به‌خود گرفته‌اند (به‌کتاب استرنگ[4] و فیکس[5] [۳۱] مراجعه کنید)، و از نظر قالب و محتوی به‌آن حدی رسیده‌اند که اکنون با روشهای تفاضل متناهی کاملا رقابت می‌کنند.

ما به‌طور خیلی اجمالی به‌فکر موجود در پشت این روشهای تقریبی اشاره می‌کنیم وتوجه خود را بیشتر روش‌موسوم به‌**روش پهلوی هم‌گذاری** متمرکز می‌سازیم. (به‌کتاب استرنگ و فیکس مراجعه‌کنید[۳۱]).

برای ساده شدن‌کار، فرض می‌کنیم کـه یك مسئلهٔ بـا مقدار مرزی مرتبهٔ دوم خطی داریم که به‌فرم زیرنوشته‌ایم

$$Ly \equiv -y'' + p(x)y' + q(x)y = r(x) \qquad a \leqslant x \leqslant b \qquad (۱۷.۹\text{الف})$$

$$a_0 y(a) - a_1 y'(a) = \alpha \qquad (۱۷.۹\text{ب})$$

$$b_0 y(b) + b_1 y'(b) = \beta \qquad |a_0| + |b_0| \neq 0$$

گیریم $\{\psi_j(x)\}$، $j = 1, \ldots, N$، مجموعه‌ای از توابع مستقل خطی باشند، که باید به‌گونه‌ای که بعداً توضیح داده خواهد شد، انتخاب شوند. دراین‌صورت سعی می‌کنیم یك جواب تقریبی برای رابطهٔ (۱۷.۹) به‌شکل

$$U_N(x) = \sum_{j=1}^{N} c_j \psi_j(x) \qquad (۱۸.۹)$$

پیدا کنیم. ضرایب $\{c_j\}$ دربسط فوق چنان‌انتخاب می‌شوند که‌اندازهٔ خطای صادق درمسئلهٔ با مقدار مرزی را به‌حداقل برساند. برحسب اینکه اندازهٔ خطا چگونه تعریف شده باشد، روشهای مختلفی پدید می‌آیند.

در روش پهلوی هم‌گذاری ضرایب چنان انتخاب می‌شوند که $U_N(x)$ در شرایط (۱۷.۹ب) ودرمعادلهٔ (۱۷.۹الف) دقیقاً در نقاط درون بازهٔ $[a, b]$ صدق کنند. ازاین‌رو $\{c_j\}$ها درمعادلات زیر صدق کنند

$$a_0 U_N(a) - a_1 U_N'(a) = \alpha$$

$$b_0 U_N(b) + b_1 U_N'(b) = \beta$$

$$LU_N(x_i) - r(x_i) = 0 \qquad i = 1, \ldots, N-2 \qquad (۱۹.۹)$$

که دراینجا x_iها مجموعه‌ای از نقاط متمایز دربازهٔ $[a, b]$ هستند. زمانی کـه (۱۹.۹)

---

1. Galerkin 2. Ritz 3. finite element 4. Strang
5. Fix

به‌تفصیل نوشته شود، به‌صورت یك دستگاه N معادلهٔ N مجهولی خطی برحسب $\{c_j\}$ درخواهد آمد. وقتی (۱۹.۹)، مثلا با یکی از روشهای فصل ۴، حل شود، جواب آن یعنی $\{c_j\}$، در (۱۸.۹) گذاشته می‌شود تا جواب مطلوب به‌دست آید. تحلیل خطا برای این روش بسیار پیچیده و بیرون از ظرفیت این کتاب است. درعمل با افزایش تعداد توابع پایه یعنی N، می‌توان یك رشته از تقریبهایی را به‌دست آورد. سپس از مقایسهٔ این جوابهای تقریبی در مجموعه‌ای از نقاط ثابت در بازهٔ $[a, b]$، براوردی برای دقت می‌توان تعیین کرد.

اکنون به‌تأمل در انتخاب توابع پایهٔ $\{\psi_j(x)\}$ می‌پردازیم. این توابع معمولا طوری انتخاب می‌شوند که دارای یك یا چندتا از ویژگیهای زیر باشند:

(i) $\psi_j(x)$ها در بازهٔ $[a, b]$ به‌طور پیوسته مشتقپذیر باشند

(ii) $\psi_j(x)$ها در بازهٔ $[a, b]$ متعامد باشند، یعنی

$$\int_a^b \psi_j(x)\psi_k(x)\,dx = 0 \qquad j \neq k \text{ به‌ازای}$$

(iii) $\psi_j(x)$ها توابعی «ساده» مانند بسجمله‌ایها یا توابع مثلثاتی باشند.

(iv) $\psi_j(x)$ها در آن شرایط مرزی (درصورت وجود) که همگن باشند، صدق کنند.

یکی از توابع پایه، که معمولا به‌کار برده می‌شود، مجموعهٔ $\{\psi_j(x)\} = \{\sin j\pi x\}$، $j = 1, \ldots, N$، است که در بازهٔ $[0, 1]$ قائم است. باید توجه داشت که به‌ازای تمامی jها در $x = 0$ و در $x = 1$، تساوی $\sin j\pi x = 0$ برقرار است. یك مجموعه پایهٔ مهم دیگر، مجموعهٔ $\{\psi_j(x)\} = \{P_j(x)\}$، $j = 0, \ldots, N$، است که در آن $P_j(x)$ بسجمله‌ایهای لژاندر هستند که در فصل ۶ تشریح شدند. این بسجمله‌ایها در بازهٔ $[-1, 1]$ متعامدند، بالاخره $\psi_j(x)$ها را می‌توان چنان انتخاب کرد که بسجمله‌ایهای درجهٔ سوم-تکه‌ای[1] باشند (به فصل ۶ نگاه کنید).

به‌عنوان مثال، روش پهلوی هم‌گذاری را برای معادلهٔ (۱۰.۹) که به‌صورت زیر نوشته می‌شود به‌کار می‌بریم

$$U''(x) - U(x) = 0 \qquad \text{(۲۰.۹ الف)}$$

$$U(0) = 0 \qquad \text{(۲۰.۹ ب)}$$

$$U(1) = 1$$

برای توابع پایه بسجمله‌ایهایی انتخاب می‌کنیم و به‌دنبال جواب تقریبی $U_N(x)$ به‌شکل

$$U_N(x) = c_1 x + c_2 x^2 + c_3 x^3 \qquad (۲۱.۹)$$

می‌رویم. ملاحظه می‌کنیم که بدون توجه به انتخاب c_iها، تساوی $U_N(0) = 0$ برقرار است.

1. piecewise-cubic polynomials

چون سه ضریب وجود دارد، باید سه‌شرط روی $U_N(x)$ اعمال‌کنیم. یکی از شروط این است‌که $U_N(x)$ باید درشرط مرزی $x=1$ صدق‌کند، بنابراین یك معادله برای c_iها به‌صورت زیرداریم

$$U_N(1)=c_1+c_2+c_3=1 \qquad (۲۲.۹)$$

با تکیه بر اینکه $U_N(x)$ دقیقاً در دونقطه از بازهٔ $[0,1]$ درمعادلهٔ (۲۰.۹الف) صدق می‌کند، می‌توانیم دوشرط اضافی اعمال‌کنیم. نقاط $x_0=1/4$ و $x_1=3/4$ را به‌دلخواه انتخاب می‌کنیم، مستقیماً از محاسبه نتیجه می‌شودکه

$$U_N''(x)-U_N(x)=-c_1x+(2-x^2)c_2+(6x-x^3)c_3$$

و بنابراین داریم

$$U_N''\left(\frac{1}{4}\right)-U_N\left(\frac{1}{4}\right)=-\frac{1}{4}c_1+\frac{31}{16}c_2+\frac{95}{64}c_3=0$$

$$U_N''\left(\frac{3}{4}\right)-U_N\left(\frac{3}{4}\right)=-\frac{3}{4}c_1+\frac{23}{16}c_2+\frac{261}{64}c_3=0 \qquad (۲۳.۹)$$

دستگاه معادلات (۲۲.۹) و(۲۳.۹) را می‌توان مستقیماًحل‌کرد و جوابهای زیررا به‌دست آورد

$$c_1=0.8522370\ldots \qquad c_2=-0.1385270\ldots \qquad c_3=0.1616160\ldots$$

اگر این مقادیر را در (۲۱.۹) قراردهیم جواب تقریبی زیر به‌دست می‌دهد

$$U_N(x)=0.852237x-0.138527x^2+0.161616x^3 \qquad (۲۴.۹)$$

اکنون این جواب تقریبی را برای پیداکردن مقدار تقریبی $U(x)$ یا حتی $U'(x)$ در هر نقطه از بازهٔ $[0,1]$ می‌توان به‌کاربرد.

برای اینکه ببینیم تقریب به‌دست‌آمدهٔ $U_N(x)$ تا چه اندازه به مقداردقیق

$$U(x)=\sinh x/\sinh 1$$

نزدیك‌است، در زیرفهرست چند مقدار تطبیقی را می‌آوریم (به جدول ۱.۹ نگاه‌کنید). بنابراین به‌نظر می‌رسدکه در بدترین مقادیرنزدیك به‌نقطهٔ میانی بازه، دو تا ازمقادیر تا سه رقم با هم توافق دارند. با توجه به‌تعدادکم توابع پایه‌که برای تقریب $U_N(x)$ به‌کار برده شده، به‌نظر می‌رسدکه نتیجه‌کاملا مطلوب‌است. برای به‌دست آوردن نتایج دقیقتر، فقط باید تعداد توابع پایه را افزایش دهیم.

x	$U_N(x)$	$U(x)$
۰٫۱۰	۰٫۰۸۵۲۴۷	۰٫۰۸۵۳۳۷
۰٫۲۵	۰٫۲۱۴۷۱۹	۰٫۲۱۴۹۵۲
۰٫۵۰	۰٫۴۲۲۶۷۵	۰٫۴۲۴۳۰۹
۰٫۷۵	۰٫۶۹۹۵۶۷	۰٫۶۹۹۷۲۴
۰٫۹۰	۰٫۸۷۳۶۱۱	۰٫۸۷۳۴۸۱

تمرین

۱-۳۰۹ مسئلهٔ با مقدارمرزی

$$U''(x)-U(x)=x \qquad U(0)=0 \qquad U(1)=1$$

را با روش پهلوی هم‌گذاری حل کنید. برای تابع آزمون[1] از بسجمله‌ای پایهٔ

$$U_N(x)=c_1x+c_2x^2+c_3x^3+\cdots+c_Nx^N$$

استفاده کنید. ابتدا به‌ازای $N=3$ و سپس به‌ازای $N=4$ مسئله را حل و جوابها را در نقاط انتخاب شده از بازهٔ موردنظر مقایسه کنید. همچنین نتایج تقریبی را با جواب دقیق زیرمقایسه کنید

$$U(x)=\frac{\sinh x}{\sinh 1}-x$$

۲-۳۰۹ با استفاده از روش پهلوی هم‌گذاری سعی کنید که مسئلهٔ با مقدار مرزی

$$U''(x)+U(x)=x$$

$$U(0)=0 \qquad U(1)=1$$

راحل کنید. عملیات را با تابع آزمون

$$U_N(x)=x+\sum_{j=1}^{N} c_j \sin j\pi x$$

که خودبه‌خود به‌ازای جمیع مقادیر c_j درشرایط مرزی صدق می‌کند، شروع کنید. مسئله را به‌ازای $N=2$ و $N=4$ حل، و نتایج را مقایسه کنید.

1. trial function

ضمیمه

کتابهای زیربرنامه‌ای

فهرست زیر شرح مختصری‌است از برخی ازمهمترین مجموعه‌های نرم‌افزار[1] که شامل زیربرنامه‌های آزمون شده برای حل کلیهٔ مسائل مهم این کتاب‌است. اطلاعات بیشتر درباب امکان دستیابی به‌آنها را می‌توانید از منابعی که نام برده‌ایم به‌دست آورید.

۱. IMSL[2] (کتابهای بین‌المللی ریاضی و آمار)

احتمالاً این کاملترین مجموعه‌ای است که می‌توان از بازار تهیه کرد. این مجموعه شامل تعدادی درحدود ۲۳۵ زیربرنامه‌است که برای حل کلیهٔ مسائل این کتاب وزمینه‌های دیگری مانند محاسبات آماری و نیز بهینه‌سازی مفید کاربرد دارد. تمامی این برنامه‌ها به‌زبان ANSI FORTRAN نوشته وطوری تطبیق داده شده‌اند که روی کلیهٔ کامپیوترهای بزرگ قابل اجرا هستند.

Source :IMSL, Inc, GNB Building, 7500 Bellaire Blvd. Houston, Texas 77036.

۲. PORT

یك مجموعهٔ نسبتاً کامل و آزمایش‌شده‌ای‌است از زیربرنامه‌ها که برای تمامی مسائل متداول در زمینهٔ آنالیزعددی کاربرد دارد. این مجموعه، به‌زبان PFORT که زیرمجموعه‌ای قابل

---

1. software package
2. International Mathematical and Statistical Library

حمل‌از زبان ANSI FORTRANاست، نوشته وطوری‌طراحی‌شده‌است‌که به‌آسانی می‌توان آن را از یك‌کامپیوتر به‌کامپیوتر دیگرمنتقل کرد.

Source: Bell Telephone Laboratories, Murray Hill, New Jersey.

EISPACK .۳

مجموعه‌ای‌است برای حل مسائل‌استانده مربوط به‌ویژه‌مقدارها وویژه‌بردارها. این‌مجموعه به‌زبان ANSI FORTRAN کدگذاری شده و شکلی‌کاملا مستقل ازکامپیوتراجرا کننده دارد. این مجموعه یك مجموعهٔ نرم‌افزاری‌است باکیفیت بسیار بالا و بسیار قابل‌اطمینان و شامل تعداد زیادی امکانات‌کمکی برای‌استفاده‌کننده (به [۳۲] نگاه‌کنید).

Source: National Energy Software Center, Argonne National Laboratories 9700 S. Cass Ave, Argonne Illinois 60439.

LINPACK .۴

یك مجموعهٔ نرم‌افزاری برای حل‌دستگاه معادلات خطی و نیزمسائل مربوط به‌کوچکترین توانهای دوم. این بسته به‌زبان ANSI FORTRAN نوشته شده و مستقل از کامپیوتر، درزمینهٔ حساب حقیقی، حساب مختلط و حساب با دقت مضاعف در دسترس‌است. این بسته به‌نحوی‌گسترده درچندین مرکزمحاسباتی، مورد آزمایشهای زیادی قرار گرفته است.

Source: National Energy Software Center, Argonne National Laboratories, 9700 S. Cass Ave., Argonne, Illinois 60439.

منابع

1. Hamming, R. W.: *Numerical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1962.
2. Henrici, P. K.: *Elements of Numerical Analysis*, John Wiley, New York, 1964.
3. Traub, J. F.: *Iterative Methods for the Solution of Equations*, Prentice-Hall, New Jersey, 1963.
4. Scarborough, J. B.: *Numerical Mathematical Analysis*, Johns Hopkins, Baltimore, 1958.
5. Hildebrand, F. B.: *Introduction to Numerical Analysis*, McGraw-Hill, New York, 1956.
6. Müller, D. E.: "A method of solving algebraic equations using an automatic computer," *Mathematical Tables and Other Aids to Computation* (*MTAC*), vol. 10, 1956, pp. 208–215.
7. Hastings, C. Jr.: *Approximations for Digital Computers*, Princeton University Press, New Jersey, 1955.
8. Milne, W. E.: *Numerical calculus*, Princeton University Press, New Jersey, 1949.
9. Lanczos, C.: *Applied Analysis*, Prentice-Hall, New Jersey, 1956.
10. Householder, A. S.: *Principles of Numerical Analysis*, McGraw-Hill, New York, 1953.
11. Faddeev, D. K., and V. H. Faddeeva: *Computational Methods of Linear Algebra*, Freeman, San Francisco, 1963.
12. Carnahan, B., et al.: *Applied Numerical Methods*, John Wiley, New York, 1964.
13. *Modern Computing Methods*, Philosophical Library, New York, 1961.
14. McCracken, D., and W. S. Dorn: *Numerical Methods and Fortran Programming*, John Wiley, New York, 1964.
15. Henrici, P. K.: *Discrete Variable Methods for Ordinary Differential Equations*, John Wiley, New York, 1962.
16. Hamming, R. W.: "Stable Predictor-Corrector Methods for Ordinary Differential Equations," *Journal of the Association for Computing Machinery* (*JACM*), vol. 6, no. 1, 1959, pp. 37–47.
17. Rice, J. R.: *The Approximation of Functions*, vols. 1 and 2, Addison-Wesley, Reading, Mass., 1964.
18. Forsythe, G., and C. B. Moler; *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, New Jersey, 1967.
19. Isaacson, E., and H. Keller: *Analysis of Numerical Methods*, John Wiley, New York, 1966.
20. Stroud, A. H., and D. Secrest: *Gaussian Quadrature Formulas*, Prentice-Hall, New Jersey, 1966.
21. Johnson, L. W., and R. D. Riess: *Numerical Analysis*, Addison-Wesley, Reading, Mass, 1977.
22. Forsythe, G. E., M. A. Malcolm, and C. D. Moler: *Computer Methods for Mathematical Computations*, Prentice-Hall, New Jersey, 1977.
23. Stewart, G. W., *Introduction to Matrix Computation*, Academic Press, New York, 1973.
24. Wilkinson, J. H.: *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.
25. Ralston, A.: *A First Course in Numerical Analysis*, McGraw-Hill, New York, 1965.

26. Shampine, L. and R. Allen: *Numerical Computing*, Saunders, Philadelphia, 1973.
27. Gautschi, W.: "On the Construction of Gaussian Quadrature Rules from Modified Moments," *Math. Comp.*, vol. 24, 1970, pp. 245–260.
28. Fehlberg, E.: "Klassische Runge-Kutta-Formeln vierter und niedriger Ordnung mit Schrittweitenkontrolle und ihre Anwendung auf Wärmeleitungsprobleme," *Computing*, vol. 6, 1970, pp. 61–71.
29. Hull, T. E., W. H. Enright, and R. K. Jackson: *User's Guide for* DVERK—*A Subroutine for Solving Non-Stiff ODE's*, TR 100, Department of Computer Science, University of Toronto, October, 1976.
30. Gear, C. W.: *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, New Jersey, 1971.
31. Strang, G., and G. Fix: *An Analysis of the Finite Element Method*, Prentice-Hall, New Jersey, 1973.
32. Smith, B. T., J. M. Boyle, J. J. Dongerra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler: "Matrix Eigensystem routines—EISPACK Guide," *Lecture Notes in Computer Science*, vol. 6, Springer-Verlag, Heidelberg, 1976.
33. Ortega, J. M., and W. C. Rheinboldt: *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
34. Robinson, S. R.: "Quadratic Interpolation Is Risky," *SIAM J. Numer. Analysis*, vol. 16, 1979, pp. 377–379.
35. Rivlin, T. J.: *An Introduction to the Approximation of Functions*, Blaisdell, Waltham, Mass., 1969.
36. Winograd, S.: "On Computing the Discrete Fourier Transform," *Math. Comp.*, vol. 32, 1978, pp. 175–199.
37. Cooley, J. W., and J. W. Tukey: "An Algorithm for the Machine Calculation of Complex Fourier Series," *Math. Comp.*, vol. 19, 1965, pp. 297–301.
38. Ehlich, H., and K. Zeller: "Auswertung der Normen von Interpolationsoperatoren," *Math. Annalen*, vol. 164, 1966, pp. 105–112.
39. de Boor, C., and A. Pinkus: "Proof of the Conjectures of Bernstein and Erdös," *J. Approximation Theory*, vol. 24, 1978, pp. 289–303.
40. de Boor, C.: *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.
41. Wendroff, B.: *Theoretical Numerical Analysis*, Academic Press, New York, 1966.
42. Wilkinson, J. H.: *Rounding Errors in Algebraic Processes*, Prentice-Hall, New Jersey, 1963.

فهرست اسامی خاص

Adams-Bashforth	ادمز-بشفورت
Adams-Moulton	ادمز-مولتن
Cramer	کرامر
Descartes	دکارت
Euler	اویلر
Jacobi	ژاکوبی
Lagrange	لاگرانژ
Laguerre	لاگر
Lebesgue	لبگ
Milne	میلن
Muller	مولر
Murnaghan-Wrench	مورناهان-رنچ
Newton	نیوتن
Parseval	پارسوال
Rayleigh	ریلای
Remez	رمه
Rolle	رول
Romberg	رامبرگ
Runge-Kutta	رونگه-کوتا
Steffensen	استفنسن
Taylor	تیلر
Vandermonde	واندرموند

واژه‌نامه

subscription	آبونمان
free-end	آزاد-انتها
trial and error	آزمون و خطا
hyperplane	ابرصفحه
standard	استانده
adjoint	الحاقی
compact scheme	الگوی فشرده
accumulation	انباشتگی
option	انتخاب
standard deviation	انحراف معیار
step size	اندازهٔ نمو
matrix updating	باب روزکردن ماتریس
iteration	بارست
inner iteration	ـداخلی
inverse iteration	ـمعکوس
fixed point iteration	ـنقطهٔ ثابت
track	باریکه
interval	بازه
ill-conditional	بدشرط
ill-conditioning	بدشرطی
body program	بدنهٔ برنامه

data fitting	برازاندن داده‌ها
curve fitting	برازاندن منحنی
subroutine	برنامهٔ فرعی
extrapolation	برونیابی
superposition	برهمنهی
software package	بستهٔ نرم‌افزاری
characteristic polynomial	بسجمله‌ای مشخصه
well-conditioned	بهشرط
bit	بیت
overrelaxation	بیش‌واهلش
flag	پرچم
back-substitution	پسجایگذاری
collocation	پهلوی هم گذاری
underflow	پی‌ریز
forward-substitution	پیشجایگذاری
biasing	پیشقدری
iteration function	تابع بازست
real valued function	تابع حقیقی‌مقدار
integrand	تابع زیرعلامت انتگرال
univariate function	تابعی از یك متغیر تصادفی
discrete transform	تبدیل گسسته
factorization	تجزیه به‌عوامل
binary search	تجسس دودویی
backward error analysis	تحلیل خطای پسرو
pivoting strategy	تدبیر لولا گزینی
transposed	ترانهاده
transposition	ترانهش
stochastic	تصادفی
filtering	تصفیه
forward shift	تعویض پیشرو
difference	تفاضل
forward difference	ـ پیشرو

divided difference	ـمنقسم
finite difference	ـهای متناهی
synthetic division	تقسیم ترکیبی
monomial	تکجمله‌ای
multiplicity	تکرر
piecewise	تکه‌ای
singularity	تکینی
steepest descent	تندترین‌کاهش
sparse	تنک
register	ثبات
permutation	جایگشت
summand	جمعوند
jump	جهش
nontrivial solution	جواب نابدیهی
weighted sum	حاصلجمع وزین
steady_state	حالت ماندگار
interval arithmetic	حساب بازه‌ای
call_statement	حکم فراخوانی
loop	حلقه
error	خطا
residual error	ـی باقیمانده
round_off error	ـی گردکردن
line search	خط‌جویی
contour lines	خطوط مرزی
self_starting	خودآغازگر
well_ behaved	خوشرفتار
entry	درایه
interpolant	درونیاب
interpolation	درونیابی

iterative interpolation	ـ بارستی
osculatory interpolation	ـ بوسانی
linear interpolation	ـ خطی
spline interpolation	ـ قلمی
single precision	دقت ساده
double precision	ـ مضاعف
plotter	رسام
method	روش
secant method	ـ خط متقاطع
chopping method	ـ قطع کردن
damped method	ـ میرا
undamped method	روش غیرمیرا
truncating method	روش قطع کردن
rounding method	روش گردکردن
flowchart	روند نما
phase angle	زاویهٔ فاز
superdiagonal	زبرقطری
superscript	زبرنمایه
subdiagonal	زیرقطری
subscript	زیرنمایه
structured	ساختیافته
consistent	سازگار
bookkeeping	ساماندهی
overflow	سرریز
truncated series	سریهای ناتمام
pseudorandom	شبه تصادفی
perturbed argument	شناسهٔ آشفته
shifted power form	صورت توانی منتقله

nested form	صورت تودرتو
premultiplication	ضرب ازچپ
postmultiplication	ضرب از راست
pulse	ضربه
leading coefficient	ضریب جملهٔ پیشرو
codition number	عدد شرط
transcendental number	عدد غیرجبری
complex number	عدد همتافت
operand	عملوند
finite elements	عنصرهای متناهی
angular frequency	فراوانی زاویه‌ای
shifted power form	فرم توانی تغییرمکان‌یافته
corrector formula	فرمول تصحیحی
central difference_formula	فرمول تفاضل‌مرکزی
compact	فشرده
panel	قطعه
peak	قله
comment card	کارت توضیحی
minor	کهاد
small_oh	کهصفر
discretization	گسسته‌سازی
matrix	ماتریس
upper triangular matrix	ـبالامثلثی
lower triangular matrix	ـپایین‌مثلثی
dense matrix	ـچگال
tridiagonal matrix	ـسه‌قطری
off_diagonal matrix	ـغیر قطری
mantissa	مانتیس

orthogonal	متعامد
boundary-value problem	مسئلهٔ مقدارمرزی
circular	مستدیر
pivotal equation	معادلهٔ لولایی
intermediate-value	مقدار میانی
time scale	مقیاس زمانی
floating-point	ممیزشناور
plausible	موجه نما
big-oh	مهصفر
damping	میرایی
instability	ناپایداری
annual region	ناحیهٔ طوقی
improper	ناسره
imbedding	نشانیدن
breakpoint	نقطهٔ انفصال
exterior mesh-points	نقطه‌های خارج شبکه
increment	نمو
sample	نمونه
harmonic oscillation	نوسانهای همساز
noise	نوفه
noisy	نوفه‌ای
variance	واریانس
waltz	والس
relaxation	واهلش
existence and uniqueness	وجودویکتایی
compilation	همگردانی
univariate	یك متغیری
monotone	یکنوا