Вступ

Спробуємо визначити місце теорії чисельних методів в системі інших галузей знань і розповісти про проблеми, що виникають у зв'язку з її застосуванням, перш ніж переходити до безпосереднього її викладу.

Математика як наука виникла у зв'язку з необхідністю рішення практичних завдань: вимірів на місцевості, навігації і так далі. Внаслідок цього математика була чисельною математикою, її метою було отримання рішення у вигляді числа. Чисельне рішення прикладних завдань завжди цікавило математиків.

Видатні представники минулого поєднували у своїх дослідженнях вивчення явищ природи, отримання їх математичного опису або, як то кажуть, математичної моделі явища, і його дослідження. Основна вимога, що пред'являється до математичної моделі, – *адекватність даному явищу*, т. е. вона повинна досить точно (у рамках допустимих похибок) відбивати характерні риси явища. В той же час вона повинна бути порівняно *простою* і *доступною* для дослідження.

Наведемо приклади деяких математичних моделей, що зробили величезний вплив на розвиток різних галузей науки і техніки. При побудові математичних моделей отримують деякі математичні співвідношення (як правило, рівняння).

Приклад

Нехай в початковий момент часу $t = 0$ тіло, що знаходиться на висоті h_0, починає рухатися вертикально вниз з початковою швидкістю v_0. Вимагається знайти закон руху тіла, т. е. побудувати математичну модель, яка дозволила б математично описати це завдання і визначити параметри руху у будь-який момент часу.

Перш ніж будувати вказану модель, треба прийняти деякі допущення, якщо вони не задані. Зокрема, припустимо, що це тіло має середню щільність, що значно перевищує щільність повітря, а його форма близька до кулі. В цьому випадку можна нехтувати опором повітря і розглядати вільне падіння тіла з урахуванням прискорення g. Відповідні співвідношення для висоти h і швидкості v у будь-який момент часу t добре відомі з шкільного курсу фізики. Вони мають вигляд:

$$h = h_0 - v_0 t - \frac{gt^2}{2}, \quad v = v_0 + gt. \tag{1.1}$$

Ці формули є шуканою математичною моделлю вільного падіння тіла. Сфера застосування цієї моделі обмежена випадками, в яких можна нехтувати опором повітря.

У багатьох завданнях про рух тіл в атмосфері планети модель (1.1) не може бути використана, оскільки при її застосуванні ми отримали б невірний результат. До таких завдань відносяться рух краплі, вхід в атмосферу тіл малої щільності, спуск на парашуті та ін. Тут необхідно побудувати точнішу математичну модель, що враховує опір повітря. Якщо позначити через $F(t)$ силу опору, діючу на тіло масою m, то його рух можна описати за допомогою рівнянь

$$m\frac{dv}{dt} = mg - F, \quad \frac{dh}{dt} = -v. \tag{1.2}$$

До цієї системи рівнянь необхідно додати початкові умови при $t = 0$:

$$v = v_0, \quad h = h_0. \tag{1.3}$$

Співвідношення (1.2) і (1.3) є математичною моделлю для завдання руху тіла в атмосфері. Існують і інші, складніші моделі подібних завдань (наприклад, про рух планера і т. п.). Помітимо також, що модель (1.1) легко виходить з (1.2) при $F = 0$.

Відоме велике число математичних моделей різних процесів або явищ. Вкажемо деякі з них, широко використовувані в механіці. Модель абсолютно твердого тіла дозволила отримати рівняння руху тіл в динаміці польоту. Модель ідеального газу привела до системи рівнянь Ейлера, що описує нев'язкі потоки газів. У гідродинаміці широко відома модель на основі рівнянь Навьє- Стоксу, в кінетичній теорії газів - рівняння Больцмана і т. д. У механіці твердого тіла, що деформується, відомі математичні моделі, що описують різні середовища (пружну, пружно-пластичну та ін.).

Є математичні моделі і для опису задач економіки, соціології, медицини, лінгвістики та ін.

Адекватність і порівняльна простота моделі не вичерпують вимог, що пред'являються до неї. Звернемо ще увагу на необхідність правильної *оцінки області застосовності* математичної моделі. Наприклад, модель тіла, що вільно падає, в якій нехтують опором повітря, дуже ефективна для твердих тіл з великою середньою щільністю і формою поверхні, близькою до сферичної. В той же час у ряді інших випадків (руху крапельки рідини, парашутного пристрою та ін.) для вирішення завдання вже недостатньо відомих з курсу фізики простих формул. Тут потрібні складніші математичні моделі: що враховують опір повітря і інші чинники.

Відмітимо, що успіх розв'язання задачі значною мірою визначається *вибором математичної моделі*; тут в першу чергу потрібні глибокі *знання в тій області*, до якої належить поставлене завдання. Крім того, потрібні знання відповідних *розділів математики* і *можливостей ЕОМ*.

Аналіз ускладнених моделей вимагав створення спеціальних, як правило, чисельних або асимптотичних методів розв'язання задач. Назви деяких з таких методів - методи Ньютона, Ейлера, Лобачевского, Гауса, Чебышева, Эрмита, Крылова - свідчать про те, що їх розробкою займалися видатні учені свого часу.

За допомогою математичного моделювання розв'язання науково-технічної задачі зводиться до розв'язання математичної задачі, що являється її моделлю. Для вирозв'язання математичних задач використовуються наступні основні групи методів обчислювальної математики: графічні, аналітичні і чисельні.

При використанні *аналітичних методів* розв'язання задачі вдається виразити за допомогою формул. Зокрема, якщо математичне завдання полягає в розв'язанні простих рівнянь алгебри або трансцендентних, диференціальних рівнянь і тому подібне, то використання відомих з курсу математики прийомів відразу призводить до мети. На жаль, на практиці це занадто окремі випадки.

Графічні методи дозволяють у ряді випадків оцінити порядок шуканої величини. Основна ідея цих методів полягає в тому, що розв'язання знаходиться шляхом геометричних побудов. Наприклад, для знаходження коренів рівняння $f(x) = 0$ будується графік функції $y = f(x)$, точки, перетини якого з віссю абсцис і будуть шуканими коренями. Графічні методи можуть застосовуватися для отримання початкового наближення до розв'язання, яке потім уточнюється за допомогою чисельних методів.

Основним інструментом для вирозв'язання складних математичних задач нині є *чисельні методи*, що дозволяють звести розв'язання задачі до виконання кінцевого числа арифметичних дій над числами; при цьому результати виходять у вигляді числових значень.

Підкреслимо важливі відмінності чисельних методів від аналітичних. По-перше, чисельні методи дозволяють отримати лише *наближене розв'язання* задачі. По-друге, вони зазвичай дозволяють отримати лише розв'язання задачі з *конкретними значеннями* параметрів і початкових даних.

Пояснимо другу відмінність на прикладі. За формулами (по аналітичному рішенню) можна проаналізувати, як зміниться закон руху при зміні параметра g і початкових значень v_0, h_0. Якщо в моделі (1.2), (1.3) вираз для $F(t)$ має простий вид (наприклад, F = const), то можна отримати аналітичне розв'язання, аналогічне (1.1). Це розв'язання теж легко досліджувати на предмет залежності від зміни параметрів і початкових умов. Якщо ж вираз для $F(t)$ досить складний, то завдання (1.2), (1.3) можна вирішити тільки чисельно. При цьому замість загальної формули розв'язання в результаті розрахунку будуть набуті значень v і h для деякого набору моментів часу t при конкретних значеннях g, m, v_0, h_0 . Для отримання розв'язання при інших значеннях параметрів і (чи) інших початкових умовах необхідно провести новий розрахунок. Для аналізу залежності розв'язання від параметрів і початкових умов потрібна велика серія розрахунків.

Незважаючи на ці недоліки, чисельні методи незамінні в складних завданнях, які не допускають аналітичного розв'язання.

Багато чисельних методів розроблено давно. У фізиці або механіці, наприклад, побудова математичних моделей для опису різних явищ і вивчення цих моделей з метою пояснення старих або передбачення нових ефектів є традиційними. Проте в цілому робота в цьому напрямі частенько просувалася відносно повільно, оскільки зазвичай при обчисленні вручну не вдавалося отримати розв'язання виникаючих трудомістких математичних задач і доводилося обмежуватися розглядом простих моделей. З появою ЕОМ почався період бурхливого розвитку чисельних методів і їх впровадження в практику. В результаті появи ЕОМ (електронно-обчислювальних машин або, як часто говорять, комп'ютерів) з програмним управлінням менш ніж за п'ятдесят років швидкість виконання арифметичних операцій зросла від 0,1 операції в секунду при ручному рахунку до 10^{12} операцій на сучасних серійних ЕОМ, т. е. приблизно у 10^{13} рази.

Приклад

> *1 TFLOPS = 1012 операції з 64 розрядними числами з плаваючою точкою в секунду.*
> *Phenom 9500 (2,2 ГГц x 4 ядра) = 0,0352 TFLOPS*
> *Core 2 Quad Q 6600 (4 ядра) = 0,0384 TFLOPS*

Все частіше результати розрахунків дозволяють виявляти і передбачати раніше явища, що ніколи не спостерігалися; це дає підстави говорити про математичний експеримент. У деяких дослідженнях довіра до результатів чисельних розрахунків така велика, що при розбіжності між результатами розрахунків і експериментів в першу чергу шукають похибка в результатах експериментів.

Поширена думка про всемогутність сучасних ЕОМ часто породжує враження, що математики позбавилися майже від усього клопоту, пов'язаного з чисельним розв'язанням задач, і розробка нових методів для їх вирозв'язання вже не така істотна. Насправді справа йде інакше, оскільки потреби еволюції, як правило, ставлять перед наукою завдання, що знаходяться на межі її можливостей. Вимога чисельного розв'язання нових задач привела до появи великої кількості нових методів. Разом з цим останні півстоліття відбувалося інтенсивне теоретичне переосмислення і старих методів, а також систематизація усіх методів. Ці теоретичні дослідження надають велику допомогу при розв'язанні конкретних задач і грають істотну роль в спостережуваному зараз широкому поширенні сфери застосувань ЕОМ і математики взагалі. Розширення можливостей застосування математики зумовило математизацію хімії, економіки, біології, геології, географії, психології, екології, метеорології, медицини, конкретних розділів техніки та ін. Суть математизації полягає в побудові математичних моделей процесів і явищ і в розробці методів їх дослідження.

Як вже відзначалося, за допомогою сучасних ЕОМ вдалося успішно вирішити ряд важливих науково-технічних задач. Досягнення в області використання ЕОМ обумовлені поєднанням ряду істотних чинників, без пропорційного розвитку яких вони були б багато скромніше:

1) *вдосконалення ЕОМ*, що включає збільшення швидкодії ЕОМ, розширення пам'яті, вдосконалення структури ЕОМ, неухильне зниження вартості арифметичної операції і одиниці пам'яті;

2) *розробка програмних засобів* спілкування з ЕОМ, що включає створення операційних систем, мов програмування, бібліотек і пакетів стандартних програм, зниження вимог (у разі персональних ЕОМ) до математичної і програмістської культури;

3) *зростання розуміння процесів і явищ* науки, техніки, природи і суспільства і *створення* їх *математичних моделей*;

4) *вдосконалення методів розв'язання* традиційних математичних і прикладних *задач* і створення методів розв'язання нових задач;

5) *зростання розуміння* можливостей *застосування ЕОМ* серед широких шарів суспільства; поширення так званої комп'ютерної письменності; координація зусиль фахівців різного профілю по використанню обчислювальної техніки.

Перегляд методів розв'язання складних прикладних задач показує, що, як правило, ефект, що досягається за рахунок вдосконалення чисельних методів, по порядку порівнянний з ефектом, що досягається за рахунок підвищення продуктивності ЕОМ. Важко сформулювати критерій, по якому можна було б оцінювати ефект застосування нових чисельних методів, і ще важче дати його достовірну кількісну оцінку. Все ж, якщо сказати, що ефект від застосування нових чисельних методів (при вимірі ефекту в логарифмічній шкалі) при розв'язанні прикладних природничонаукових задач *дає 40% загального ефекту*, що досягається за рахунок застосування нової обчислювальної техніки і нових чисельних методів, то ця оцінка не буде завищеною.

Розглянемо приклад, що ілюструє це твердження.

Приклад

> Розв'язання диференціальних рівнянь в приватних похідних зводиться до розв'язання систем лінійних рівнянь алгебри з матрицею, в кожному рядку якої є 5-10 ненульових елементів. Напередодні появи ЕОМ такі системи рівнянь вирішували у разі числа невідомих близько $10\text{-}10^2$; зараз нерідкі випадки, коли вирішуються системи з числом невідомих близько $10^5\text{-}10^6$. У гіпотетичному випадку розв'язання цих задач на сучасних ЕОМ методами, відомими тридцять років тому, довелося б обмежитися системами рівнянь з числом невідомих близько $10^3\text{-}10^4$ (при тих же витратах часу ЕОМ).

Скінченність швидкості поширення сигналу - 300 000 км/с - ставить вже зараз істотне обмеження на можливе зростання швидкодії однопроцесорних ЕОМ, тому значення подальшого розвитку теорії чисельних методів важко переоцінити. Зокрема, стає усе більш актуальною проблема розробки чисельних методів і програмних засобів для багатопроцесорних ЕОМ.

Швидке проникнення математики в багато областей знання, зокрема, пояснюється тим, що *математичні моделі* і методи їх дослідження застосовані відразу до *багатьох явищ, схожих* по своїй формальній структурі. Часто математична модель, що описує яке-небудь явище, з'являється при вивченні інших явищ або при абстрактних математичних побудовах задовго до конкретного розгляду цього явища. Зокрема, і в теорії чисельних методів, так само як в "чистій" математиці, корисна розробка загальних побудов. Проте є різниця в підході "*чистого*" і "*прикладного*" математика до вирозв'язання якої-небудь проблеми. На мові першого поняття "*Вирішити завдання*" означає *довести існування розв'язання* і запропонувати *процес, що сходиться* до розв'язання. Самі по собі ці результати корисні для прикладника, але, окрім цього, йому треба, щоб *процес* отримання наближення *не вимагав великих витрат*, наприклад часу або пам'яті ЕОМ. Йому важливо не лише те, що процес сходиться, але і те, як *швидко він сходиться*. Чисельний метод разом з можливістю отримання результату за прийнятний час повинен володіти і ще однією важливою якістю – *не вносити* в обчислювальний процес *значних похибок*. При чисельному розв'язанні задач виникають також питання, пов'язані із *стійкістю результату відносно збурень* початкових даних і округлень при обчисленнях.

Етапи розв'язання задачі на комп'ютері. При розв'язанні задачі на комп'ютері основна роль все-таки належить людині. Машина лише виконує його завдання за розробленою програмою. Роль людини і машини легко утямити, якщо процес розв'язання задачі розбити на наступні етапи.

1) *Постановка завдання.* Цей етап полягає в змістовній (фізичній) постановці завдання і визначенні кінцевої мети розв'язання.

2) *Побудова математичної моделі (математичне формулювання завдання).* Модель повинна правильно (адекватно) описувати основні закони фізичного процесу. Побудова або вибір математичної моделі з існуючих вимагає глибокого розуміння проблеми і знання відповідних розділів математики.

3) *Розробка чисельного методу.* Оскільки комп'ютер може виконувати лише прості операції, він "не розуміє" постановки завдання навіть в математичному формулюванні. Для її вирозв'язання має бути знайдений чисельний метод, що дозволяє звести завдання до деякого обчислювального алгоритму. Розробкою чисельних методів займаються фахівці в області обчислювальної математики. Фахівцеві-прикладникові для вирозв'язання завдання, як правило, необхідно з наявного арсеналу методів вибрати той, який найбільш придатний в даному конкретному випадку.

4) *Розробка алгоритму.* Процес розв'язання задачі (обчислювальний процес) записується у вигляді послідовності елементарних арифметичних і логічних операцій, що призводить до кінцевого результату і називається алгоритмом розв'язання задачі. Алгоритм можна наочно зображувати у вигляді блок-схеми, структурограммы і т. п. Досвідчений обчислювач частенько може і не удаватися до такого наочного представлення алгоритму, безпосередньо переходячи до наступного етапу.

5) *Програмування.* Алгоритм розв'язання задачі записується на зрозумілій машині мові у вигляді точно визначеної послідовності операцій - програми для комп'ютера. Складання програми (програмування) зазвичай проводиться за допомогою деякої проміжної (алгоритмічного) мови, а її трансляція (переклад машинною мовою) здійснюється самою обчислювальною системою.

6) *Відладка програми.* Складена програма містить різного роду помилки, неточності, описки. Відладка програми на машині включає контроль програми, діагностику (пошук і визначення змісту) помилок, їх виправлення. Програма випробовується на розв'язанні контрольних (тестових) задач для отримання упевненості в достовірності результатів.

7) *Проведення розрахунків.* На цьому етапі готуються початкові дані для розрахунків і проводиться рахунок за відлагодженою програмою. При цьому для зменшення ручної праці по обробці результатів бажано використовувати зручні форми видачі результатів, особливе їх графічне представлення (візуалізацію).

8) *Аналіз результатів.* Результати розрахунків аналізуються, оформляється науково-технічна документація.

Якщо при розв'язанні конкретної задачі можливе використання вже наявних прикладних програмних засобів, то деякі з перерахованих етапів можуть бути опущені. Так, для вирозв'язання багатьох (хоча і досить вузьких) класів задач створені програмні продукти, що істотно полегшують працю обчислювача. Може йтися, наприклад, про те, що обчислювач повідомляє програмі тільки математичну модель (або навіть постановку завдання) і початкові дані, а вибір методу, проведення розрахунків, видачу результатів програма бере на себе. Але навіть в цьому випадку не можна забувати про те, що отримане розв'язання зазвичай є лише наближеним, що кожна модель і кожен метод мають свої області застосовності. Отже, фахівцеві, що використовує комп'ютер для вирозв'язання прикладних задач, необхідно мати уявлення про *основи математичного моделювання*, *чисельних методів*, про *можливості комп'ютерів*, *уміти аналізувати* отримані *результати* з точки зору їх достовірності.

Слід зазначити ще один важливий момент в процесі розв'язання задачі за допомогою комп'ютера. Це - економічність вибраного способу розв'язання задачі, чисельного методу, моделі комп'ютера. Зокрема, якщо завдання допускає просте аналітичне розв'язання або вимір, то навряд чи доцільно робити обчислення на машині. Іноді розв'язання задачі проводять за допомогою великого обчислювального комплексу, хоча це можна було здійснити з використанням персонального комп'ютера.

Не зменшуючи значення фізичного експерименту, треба все-таки відмітити неухильно зростаючу долю обчислень на комп'ютері в загальному об'ємі розв'язання науково-технічних задач. У зв'язку з цим разом із збільшенням парку обчислювальних машин і підвищенням їх "інтелектуальних" можливостей зростає інтерес до математичного моделювання і розробки чисельних методів.

Джерела і класифікація похибок

Похибки розв'язання задачі обумовлюється наступними причинами:

1) математичний опис завдання є неточним, зокрема неточно задані початкові дані опису;

2) вживаний для вирозв'язання метод часто не є точним: отримання точного розв'язання виникаючої математичної задачі вимагає необмеженого або неприйнятний великого числа арифметичних операцій; тому замість точного розв'язання задачі доводиться удаватися до наближеного;

3) при введенні даних в машину, при виконанні арифметичних операцій і при виведенні даних проводяться округлення.

Похибки, що відповідають цим причинам, називають:

1) *неусувною похибкою*

2) *похибкою методу*

3) *обчислювальною похибкою.*

Часто неусувну похибка підрозділяють на дві частини:

а) неусувною похибкою називають лише похибка, що є наслідком неточності завдання числових даних, що входять в математичний опис завдання;

б) похибка, що є наслідком невідповідності математичного опису завдання реальності, називають, відповідно, похибкою математичної моделі.

Дамо ілюстрацію цих визначень.

Приклад

Нехай у нас є маятник (рис. 1.1), початкуючий рух у момент $t = t_0$ вимагається передбачити кут відхилення φ від вертикалі у момент t_1.

Рис. 1.1

Диференціальне рівняння, що описує коливання цього маятника, береться у виді

$$l\frac{d^2\varphi}{dt^2} + g\sin\varphi + \mu\frac{d\varphi}{dt} = 0, \qquad (1.4)$$

де l - довжина маятника, g - прискорення сили тяжіння, μ - коефіцієнт тертя.

Як тільки приймається такий опис завдання, розв'язання вже придбаває неусувну похибка, зокрема, тому, що реальне тертя залежить від швидкості не зовсім лінійно; інше джерело неусувної похибки полягає в погрішностях визначення l, g, μ, $\varphi(t_0)$, $\varphi'(t_0)$. Назва цієї похибки - "неусувна" - відповідає її сутності: вона неконтрольована в процесі чисельного розв'язання задачі і може зменшитися тільки за рахунок точнішого опису фізичного завдання і точнішого визначення параметрів. Диференціальне рівняння (1.4) не вирішується в явному виді; для його вирозв'язання вимагається застосувати який-небудь чисельний метод. Внаслідок цієї причини і виникає похибка методу. Обчислювальна похибка може виникнути, наприклад, із-за кінцевої кількості розрядів чисел, що беруть участь в обчисленнях.

Введемо формальні визначення.

Нехай I - точне значення відшукуваного параметра (в даному випадку - реальний кут відхилення маятника φ у момент часу t_1), $\tilde{I}$ – значення цього параметра, що відповідає прийнятому математичному опису (у даному випадку значення ф(t_1) розв'язання рівняння (1.4)), $\tilde{I}_h$ – розв'язання задачі, що отримується при реалізації чисельного методу в припущенні відсутності округлень, $\tilde{I}_h^*$ – наближення до розв'язання задачі, отримуване при реальних обчисленнях. Тоді

$$\begin{aligned} \rho_1 &= \tilde{I} - I \quad \textit{— неусувна похибка} \\ \rho_2 &= \tilde{I}_h - \tilde{I} \quad \textit{— похибка методу} \\ \rho_3 &= \tilde{I}_h^* - \tilde{I}_h \quad \textit{— обчислювальна похибка.} \end{aligned} \qquad (1.5)$$

Повна похибка $\rho_0 = \tilde{I}_h^* - I$, рівна різниці між реально отримуваним і точним розв'язаннями задачі, задовольняє рівності

$$\rho_0 = \rho_1 + \rho_2 + \rho_3. \qquad (1.6)$$

Може виникнути таке питання з приводу проблеми дослідження неусувної похибки: навіщо вивчати неусувну похибка розв'язання задачі, якщо вона "неусувна"? Принаймні, така точка зору здається виправданою, якщо математик отримує для чисельного вирозв'язання завдання вже готові рівняння, не беручи участь в обговоренні фізичної постановки завдання.

Це заперечення не можна визнати розумним. Часто математик сам займається дослідженням постановки завдання, аналізом і спрощенням даних рівнянь. Оскільки усі явища в природі взаємозв'язані, в принципі неможливо математично точно описати ніякий реальний процес, що відбувається в природі. Проте аналіз впливу різних чинників на похибка розв'язання може дозволити отримати простий опис процесу з допустимою похибкою. Зазвичай математик має уявлення про необхідну остаточну точність результату, і, виходячи з цього, він може проводити необхідні спрощення початкової задачі.

При розв'язанні більшості задач немає особливого сенсу застосовувати метод розв'язання задачі з похибкою, істотно меншою, ніж величина неусувної похибки. Тому, маючи уявлення про величину неусувної похибки, можна розумно сформулювати вимоги до точності результату чисельного розв'язання задачі. При цьому необхідно брати до уваги наступне:

1) часто для практичного використання висока точність не потрібна;

2) математична модель явища настільки груба, що вимагати високу точність розв'язання задачі безглуздо;

3) параметри моделі не можуть бути визначені з високою точністю;

4) часто потрібний взагалі не кількісний, а якісний результат, наприклад, такого типу: чи працюватиме цей пристрій в заданому режимі або ні.

Розглянемо, наприклад, одну реальну задачу.

Приклад

Перед математиками було поставлено завдання створення алгоритму і програми швидкого (менш ніж за 1 з

машинного часу) обчислення інтегралів спеціального виду з відносною похибкою 10^{-4}. Це завдання було ними успішно вирішене, тобто був розроблений метод обчислення таких інтегралів і на його основі створена стандартна програма. У свою чергу дослідники, що поставили завдання, не скупившись на витрати машинного часу, для перевірки якості запропонованого математиками методу і надійності програми самі вичислили приблизно один з таких інтегралів з відносною похибкою, на їх думку, 10^{-6}. Але виявилось, що усі спроби вирішити цю, так зване тестове завдання з похибкою, кращою, ніж 10^{-2}, за допомогою створеної математиками програми закінчувалися невдачею. Виникло припущення про похибка в самому тестовому завданні. Виявилось, що число π було узяте рівним 3.14, що вносило в тестовий приклад неусувну похибка, яка, природно, не могла бути усунена ніякими зусиллями математиків, що створювали алгоритм і програму.

Наближені числа

Числа з плаваючою точкою. Числа можуть бути представлені в пам'яті комп'ютерів різними способами. Сучасні комп'ютери (процесори), як правило, дозволяють обробляти цілі числа, а також дробові числа у формі з плаваючою точкою.

Як відомо, множина цілих чисел нескінченна. Проте процесор через обмеженість його розрядної сітки може оперувати лише з деякою кінцевою підмножиною цієї великої кількості. У сучасних комп'ютерах для зберігання цілого числа зазвичай відводиться 4 байти пам'яті, що дозволяє представляти цілі числа, що знаходяться приблизно в діапазоні від $-2\cdot10^9$ до $2\cdot10^9$.

При розв'язанні науково-технічних задач в основному використовуються дійсні числа. У комп'ютерах вони представляються у формі з плаваючою точкою. Десяткове число D в цій формі запису має вигляд $D = \pm m\cdot10^n$, де m и n - відповідно *мантиса* числа і його *порядок*. Наприклад, число -273.9 можна записати у виді: $(2739\cdot10^{-1}$, $-2.739\cdot10^2$, $-0.2739\cdot10^3$. Останній запис – *нормалізована форма* числа з плаваючою точкою. Таким чином, якщо представити мантису числа у вигляді $m = 0.d_1d_2..d_k$, то при $d_1\neq0$ отримаємо нормалізовану форму числа з плаваючою точкою. Надалі, кажучи про числа з плаваючою точкою, матимемо на увазі саме цю форму. Звичайний же запис числа у виді -273.9 називається формою запису з фіксованою точкою. Нині таке представлення використовується в комп'ютерах, як правило, тільки на етапі введення і виведення чисел.

Усе сказане вище про числа з плаваючою точкою поширюється і на числа, записані в інших системах числення. Число А в системі числення з основою α можна представити у виді $A = \pm0.a_1a_2..\ a_k\cdot\alpha^n$, де $a_1\ a_2..$ю a_k - цілі числа з діапазону 0,.., α – 1. З цього запису виходить, що підмножина дійсних чисел, з якою оперує конкретний комп'ютер, не є нескінченною: вона кінцева і визначається розрядністю k, а також межами порядку n_1, n_2 ($n_1 \le n \le n_2$). Можна показати, що ця підмножина містить

$$N = 2(\alpha - 1)(n_1 - n_2 + 1)\alpha^{k-1} + 1 \qquad (1.7)$$

чисел, найменшим і найбільшим по модулю являються відповідно числа

$$M_0 = (\alpha - 1)\alpha^{n_1-1} \text{ і } M_\infty = (1 - \alpha^{-k})\alpha^{n_2} \qquad (1.8)$$

звані машинним нулем і машинною нескінченністю.

Межі порядку n_1, n_2 визначають обмеженість дійсних чисел за величиною, а розрядність k - дискретність розподілу їх на відрізку числової осі. Наприклад, у разі десяткових чисел при чотирирозрядному представленні усі значення, що знаходяться в проміжку (0.28505, 0.28515), представляються числом 0.2851 (при виконанні округлення). Якщо до цього числа 0.2851 додати число, менше по модулю половини одиниці останнього розряду (т. е. менше по модулю 0.00005), в результаті вийде те ж саме число 0.2851.

Нині більшість виробників процесорів в основному дотримуються стандарту IEEE 754 для арифметичних операцій над двійковими числами з плаваючою точкою. Цей стандарт передбачає наявність, зокрема, двох двійкових (а = 2) форматів: з одинарною точністю і з подвійною точністю. Приведемо для цих форматів розмір пам'яті, що відводиться, значення k, n_1, n_2 і наближені значення M_0 і M_∞. Помітимо, що стандарт IEEE 754 передбачає обробку чисел, менших по модулю M_0, але не менших M_0*, правда, з меншою розрядністю k.

Точність	Байти	k	n_1	n_2	M_0	M_0*	M_∞
Одинарна	4	24	–125	128	$1.2\cdot10^{-38}$	$1.4\cdot10^{-45}$	$3.4\cdot10^{38}$
Подвійна	8	53	–1021	1024	$2.2\cdot10^{-308}$	$4.9\cdot10^{-324}$	$1.8\cdot10^{308}$

Оскільки для людини зручнішою є десяткова система числення, виникає питання про те, скільком десятковим розрядам відповідає вказана двійкова розрядність k. Можна вважати, що k відповідає 6-9 десятковим розрядам при одинарній і 15-17 розрядам при подвійній точності.

У сучасних мовах програмування передбачені типи даних для представлення дійсних чисел з одинарною і подвійною точністю. Наприклад, в мові Сі це типи float і double, в мові Паскаль - single і double. Зазвичай ці представлення відповідають стандарту IEEE 754.

Таким чином, комп'ютер оперує з наближеними значеннями дійсних чисел. Мірою точності наближених чисел є похибка.

Поняття похибки

Розрізняють два види похибок - абсолютну і відносну. *Абсолютна похибка* деякого числа дорівнює різниці між його істинним значенням і наближеним значенням, отриманим в результаті обчислення або виміру. *Відносна похибка* – це відношення абсолютної похибки до наближеного значення числа.

Таким чином, якщо a – наближене значення числа x, то вирази для абсолютної і відносної похибок запишуться відповідно у виді

$$\Delta x = x - a, \qquad \delta x = \Delta x / a. \tag{1.9}$$

На жаль, істинне значення величини x *зазвичай невідоме*. Тому приведені вирази для похибок практично не можуть бути використані. Є лише наближене значення a і треба знайти його *граничну похибку* Δa, що є верхньою оцінкою модуля абсолютної похибки, т. е. $|\Delta x| \le \Delta a$. Надалі значення Δa приймається як абсолютна похибка наближеного числа a. В цьому випадку істинне значення x знаходиться в інтервалі $(a - \Delta a, a + \Delta a)$.

Для наближеного числа, отриманого в результаті *округлення*, абсолютна похибка Δa приймається *рівній половині одиниці останнього розряду числа*. Наприклад, значення a = 0.734 могло бути отримано округленням чисел 0.73441, 0.73353 та ін. При цьому $|\Delta x| \le 0.0005$, і вважаємо Δa = 0.0005. Якщо при обчисленнях на комп'ютері округлення не проводиться, а цифри, що виходять за розрядну сітку машини, відкидаються, то максимально можлива похибка результату виконання операції в два рази більше в порівнянні з випадком округлення.

Наведемо приклади оцінки абсолютної похибки при деяких значеннях наближеної величини a:

a	51.7	–0.0031	16	16.00
Δa	0.05	0.00005	0.5	0.005

Граничне значення відносної похибки - відношення граничної абсолютної похибки до абсолютної величини наближеного числа :

$$\delta a = \Delta a / |a|. \tag{1.10}$$

Наприклад, $\delta(-2.3) = 0.05/2.3 \approx 0.022$ (2.2%). Помітимо, що похибка округляється завжди у бік збільшення. В даному випадку $\delta(-2.3) \approx 0.03$.

Приведені оцінки похибок наближених чисел справедливі, якщо в записі цих чисел усі значущі цифри вірні. Нагадаємо, що *значущими* цифрами вважаються усі цифри цього числа, починаючи з першої ненульової цифри. Наприклад, в числі 0.037 дві значущі цифри: 3 і 7, а в числі 14.80 усі чотири цифри значущі. Значущу цифру називають *вірною*, якщо абсолютна похибка числа не перевершує одиниці розряду, що відповідає цій цифрі. Наприклад, a = 0,03045 (при Δa = 0.000003), a = 0,03045000 (при Δa = 0.0000007); підкреслені цифри є вірними. Іноді вживається термін число вірних цифр після коми: підраховується число цифр після коми від першої цифри до останньої вірної цифри.

При зміні форми запису числа (наприклад, при записі у формі з плаваючою точкою) число значущих цифр не повинне мінятися, тобто треба дотримувати равносильность перетворень. Наприклад, записи $7500 = 0.7500 \cdot 10^4$ і $0.110 \cdot 10^2 = 11.0$ рівносильні, а записи $7500 = 0.75 \cdot 10^4$ і $0.110 \cdot 10^2 = 11$ нерівносильні.

Разом з приведеними вище оцінками похибок наближених чисел можна записати аналогічні оцінки і для обчислення функцій, аргументами яких є наближені числа. Проте повнішим виявляється загальне правило, засноване на обчисленні приросту (похибки) функції при заданих приростах (погрішностях) аргументів.

Розглянемо функцію однієї змінної $y = f(x)$. Нехай a - наближене значення аргументу x, Δa - його абсолютна похибка. Абсолютну похибка функції можна рахувати її приростом, який вона отримує при зміні аргументу на Δa. Цей приріст можна замінити диференціалом: $\Delta y \approx dy$. Тоді для оцінки абсолютної похибки отримаємо вираз $\Delta y = |f'(a)|\Delta a$.

Аналогічний вираз можна записати для функції декількох аргументів. Наприклад, оцінка абсолютної похибки функції $u = f(x, y, z)$, наближені значення аргументів якої відповідно a, b, з, має вигляд

$$\Delta u = |f'_x(x,y,z)|\Delta a + |f'_y(x,y,z)|\Delta b + |f'_z(x,y,z)|\Delta c. \tag{1.11}$$

Тут Δa, Δb, Δc - абсолютні похибки аргументів. Відносна похибка знаходиться по формулі

$$\delta u = \frac{\Delta u}{|f(a,b,c)|}. \tag{1.12}$$

Дії над наближеними числами

Сформулюємо правила оцінки граничних похибок при виконанні операцій над наближеними числами.

При складанні або відніманні чисел їх абсолютні похибки складаються. При множенні або діленні чисел один на одного їх відносні похибки складаються. При піднесенні до степеня наближеного числа його відносна похибка множиться на показник міри.

Для випадку двох наближених чисел a і b ці правила можна записати у вигляді формул

$$\Delta(a \pm b) = \Delta a + \Delta b, \quad \delta(a \cdot b) = \delta a + \delta b,$$
$$\delta(a/b) = \delta a + \delta b, \quad \delta(a^k) = k\delta a. \tag{1.13}$$

Отримані співвідношення можна вивести використовуючи формули оцінки похибки довільної функції (1.11). Наприклад, у такий спосіб легко отримати $\Delta(a - b)$: при $c = a - b$ по формулі (1.11) отримуємо $\Delta c = |c'_a|\Delta a + |c'_b|\Delta b = \Delta a + \Delta b$.

Відносна похибка суми позитивних доданків ув'язнена між найбільшим і найменшим значеннями відносних похибок цих доданків. Дійсно, нехай a > 0, b > 0, m = min(δa, δb), M = max(δa, δb). Тоді

$$\delta(a+b) = \frac{\Delta(a+b)}{a+b} = \frac{\Delta a + \Delta b}{a+b} = \frac{a\delta a + b\delta b}{a+b} \leqslant \frac{aM + bM}{a+b} = M. \qquad (1.14)$$

Аналогічно, δ(*a* + *b*) ≥ *m*. На практиці для оцінки похибки набуває найбільшого значення *M*.

Приклад

Знайти відносну похибка функції

$$y = \sqrt{\frac{a+b}{x^3(1-x)}}. \qquad (1.14)$$

Використовуючи формули (1.13), отримуємо

$$\delta y = \frac{1}{2}\left[\delta(a+b) + 3\delta x + \delta(1-x)\right] = \frac{1}{2}\left[\frac{\Delta a + \Delta b}{|a+b|} + 3\,\frac{\Delta x}{|x|} + \frac{\Delta x}{|1-x|}\right] \qquad (1.16)$$

Отримана оцінка відносної похибки містить в знаменнику вираз |1 (x|. Ясно, що при x ≈ 1 можна отримати дуже велику похибка. У зв'язку з цим розглянемо детальніше випадок віднімання близьких чисел. Запишемо вираз для відносної похибки різниці двох чисел у виді

$$\delta(a-b) = \frac{\Delta(a-b)}{|a-b|} = \frac{\Delta a + \Delta b}{|a-b|}. \qquad (1.17)$$

При a ≈ b ця похибка може бути скільки завгодно великою.

Приклад

Нехай *a* = 2520, *b* = 2518. В цьому випадку маємо абсолютні похибки початкових даних Δa = Δb = 0.5 і відносні похибки δ*a* ≈ δ*b* = 0.5/2518 ≈ 0.0002 (0.02%). Відносна похибка різниці рівна

$$\delta(a-b) = \frac{0.5+0.5}{2} = 0.5\ (50\,\%). \qquad (1.18)$$

Отже, при малих погрішностях в початкових даних ми отримали дуже неточний результат. Неважко підрахувати, що навіть при випадкових змінах *a* і *b* на одиницю в останніх розрядах їх різниця може набувати значень 0, 1, 2, 3, 4. Тому при організації обчислювальних алгоритмів *слід уникати віднімання близьких чисел*; при нагоді алгоритм треба видозмінити щоб уникнути втрати точності на деякому етапі обчислень.

З розглянутих правил виходить, що *при складанні або відніманні* наближених чисел бажано, щоб ці числа мали *однакові абсолютні похибки*, тобто з однаковим числом розрядів після десяткової точки. Наприклад, 38.723 + 4.9 = 43.6; 425.4 − 0.047 = 425.4. Облік відкинутих розрядів не підвищить точність результатів. При *множенні і діленні* наближених чисел *кількість* значущих *цифр* вирівнюється по *найменшому* з них.

Оцінимо тепер *похибки округлень*, які неминучі при обчисленнях за допомогою комп'ютера і пов'язані з *обмеженістю розрядної сітки* машини. При звичайному округленні (яке, як правило, і реалізується в комп'ютерах) максимальна відносна похибка є

$$\delta_{max} = 0.5\alpha^{1-k} \qquad (1.19)$$

де α - основа системи числення, *k* - кількість розрядів мантиси числа. При простому відкиданні зайвих розрядів ця похибка збільшується удвічі.

Вичислимо по формулі (1.19) максимальну похибка округлення δ_{max} для чисел, представлених у форматах з одинарною і подвійною точністю стандарту IEEE 754. Маємо: α = 2 в обох випадках, для одинарної точності *k* = 24 і $\delta_{max} \approx 6\cdot10^{-8}$, для подвійної точності *k* = 53 і $\delta_{max} \approx 10^{-16}$. Зверніть увагу, що відносна похибка має фіксовану величину для чисел з плаваючою точкою.

Не дивлячись на те, що при розв'язанні великих задач виконуються мільярди і трильйони операцій, це зовсім не означає механічного множення похибки при одному округленні на число операцій, оскільки при окремих діях похибки можуть компенсувати один одного (наприклад, при складанні чисел різних знаків). В той же час іноді похибки округлень у поєднанні з погано організованим алгоритмом можуть сильно спотворити результати. Надалі ми розглянемо такі випадки.

Переведення чисел з однієї *системи числення* в іншу також може бути *джерелом похибки* через те, що основа однієї системи числення не є мірою основи інший (наприклад, 10 і 2). Це може привести до того, що в новій системі числення число стає ірраціональним.

Приклад

Число 0.1 при перекладі в двійкову систему числення набере вигляду 0.00011001100 .. Може виявитися, що з кроком 0.1 треба при обчисленнях пройти відрізок [0,1] від x = 1 до x = 0; десять кроків не дадуть точного значення x = 0.

Зменшення похибок. При розгляді похибок результатів арифметичних операцій відзначалося, що *віднімання близьких чисел* приводить до *збільшення відносної похибки*; тому в алгоритмах слід уникати подібних ситуацій. Розглянемо також деякі інші випадки, коли можна уникнути втрати точності правильною організацією обчислень.

Приклад

Нехай вимагається знайти суму п'яти чотирирозрядних чисел: S = 0.2764 + 0.3944 + 1.475 + 26.46 + 1364. Складаючи усі ці числа, а потім округлюючи отриманий результат до чотирьох значущих цифр, отримуємо S = 1393. Проте при обчисленні на комп'ютері округлення відбувається після кожного складання. Припускаючи умовно сітку чотирирозрядною, простежимо за обчисленням на комп'ютері суми чисел від найменшого до найбільшого, т. е. в порядку їх запису: 0.2764 + 0.3944 = 0.6708, 0.6708 + 1.475 = 2.156, 2.156 + 26.46 = 28.62, 28.62 + 1364 = 1393; отримали S_1 = 1393, т. е. вірний результат. Змінимо тепер порядок обчислень і почнемо складати числа послідовно від останнього до першого: 1364+26.46 = 1390, 1390 + 1.475 = 1391, 1391 + 0.3944 = 1391, 1391 + 0.2764 = 1391; тут остаточний результат S_2 = 1391, він менш точний.

Аналіз процесу обчислень показує, що втрата точності тут відбувається через те, що додавання до великого числа малих чисел не відбувається, оскільки вони виходять за рамки розрядної сітки ($a + b = a$ при $a >> b$). Цих малих чисел може бути дуже багато, але на результат вони все одно не вплинуть, оскільки додаються по одному, що і мало місце при обчисленні S_2. Тут необхідно дотримуватися правила, відповідно до якого *складання чисел* треба проводити у *міру їх зростання*. У машинній арифметиці із-за похибки округлення суттєвим є порядок виконання операцій, і відомі з алгебри закони комутативності (і дистрибутивності) тут не завжди виконуються.

При розв'язанні задачі на комп'ютері треба використовувати подібного роду маленькі хитрощі для поліпшення алгоритму і зниження похибок результатів.

Приклад

Наприклад, при обчисленні на комп'ютері значення $(a + x)^2$ величина x може виявитися такою, що результатом складання $a + x$ вийде a (при $x << a$); в цьому випадку може допомогти заміна $(a + x)^2 = a^2 + 2ax + x^2 = a(a + 2x) + x^2$. Дійсно, тепер до a додається не x, а $2x$. Якщо ж при $x << a$ обчислюється величина $(a + x)^2 - a^2$, то доцільно привести її до виду $2ax + x^2$, уникнувши тим самим віднімання близьких величин.

Розглянемо ще один важливий приклад – використання степеневих рядів для обчислення значень функцій.

Приклад

Запишемо, наприклад, розкладання функції sin x по мірах аргументу:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \qquad (1.20)$$

За ознакою Лейбніца залишок знакозмінного ряду, що сходиться, тобто похибка суми кінцевого числа членів, не перевищує значення першого з відкинутих членів (за абсолютною величиною).

Вичислимо значення функції sinx при x = 0.5236 (30°). Члени ряду, менші 10^{-4}, не враховуватимемо. Обчислення проведемо з чотирма вірними знаками. Отримаємо

$$\sin 0.5236 = 0.5236 - 0.2392 \cdot 10^{-1} + 0.3279 \cdot 10^{-3} = 0.500. \qquad (1.21)$$

Це відмінний результат у рамках прийнятої точності. Знаючи з курсу вищої математики, що це розкладання синуса справедливе при будь-якому значенні аргументу ($-\infty < x < +\infty$), використовуємо його для обчислення функції при x = 6.807 (390°). Опускаючи обчислення, отримуємо sin 6.807 ≈ 0.5167. Відносна похибка складає тут близько 3% (замість очікуваного значення 0.01% за ознакою Лейбніца). Це пояснюється погрішностями округлень і способом підсумовування ряду (зліва направо, без урахування величини членів).

Не завжди допомагає і підвищена точність обчислень. Зокрема, для цього ряду при x = 25.6563 .. (1470° = 4·360°+30°) навіть при обліку членів ряду до 10^{-8} і обчисленнях з вісьмома значущими цифрами в результаті аналогічних обчислень (підсумовування зліва направо) виходить результат, що не має сенсу: sin x ≈ 129.

У програмах, що використовують степеневі ряди для обчислення значень функцій, можуть бути вжиті різні заходи щоб запобігти подібній втраті точності. Так, вплив похибок округлення істотно зменшується, якщо |x| < 1. Дійсно, при обчисленні x^k допускається абсолютна похибка

$$\Delta(x^k) = x^k \delta(x^k) = x^k k \delta x \qquad (1.22)$$

((див. (1.13)), яка при невиконанні нерівності |x|<1 може стати неприйнятно великою. Для тригонометричних функцій можна використовувати формули приведення, завдяки чому аргумент знаходитиметься на відрізку [0, 1]. При обчисленні експоненти аргумент x можна розбити на суму цілої і дробової частин ($e^x = e^{n+a} = e^n \cdot e^a$, 0 < a < 1) і використовувати розкладання в ряд тільки для e^a, а e^n обчислювати множенням. Таким чином, при організації обчислень можна своєчасно розпізнати "підводні камені", що дають втрату точності, і спробувати виправити положення.

Про розв'язання квадратного рівняння. Ми переконалися в тому, що при чисельному розв'язанні задач на комп'ютері обчислювача чекають всякі "пастки", які можуть привести до помітної втрати точності ре-зультатов або навіть до припинення рахунку. Хорошою ілюстрацією до цього є аналіз алгоритму розв'язання такої простої задачі, як розв'язання квадратного рівняння $ax^2 + bx + c = 0$. Його корені визначаються співвідношеннями

$$x_1 = \frac{-b - \sqrt{D}}{2a}, \quad x_2 = \frac{-b + \sqrt{D}}{2a}, \quad D = b^2 - 4ac. \qquad (1.23)$$

З аналізу цих формул видно, що тут є ряд особливостей обчислювального характеру, які необхідно мати на увазі при складанні алгоритму.

Розглянемо простий випадок $a = 0$. Тут рівняння стає лінійним, і його єдиний корінь є $x = -c/b$, якщо $b \neq 0$. При $a = b = 0$ і $c \neq 0$ рівняння не має розв'язання, a у випадку $a = b = c = 0$ його розв'язанням буде будь-яке число. Помітимо, що в машинній арифметиці рідко виходять точно нульові значення. Тому коэффіцієнти можна порівнювати не з нулем, а з деякою малою величиною ε.

Це у свою чергу породжує ряд ситуацій, залежних від співвідношення між коефіцієнтами.

Далі необхідно передбачити розгалуження алгоритму в залежності від знаку дискримінанта D : $D > 0$ - корені дійсні (см (1.23)); $D = 0$ - корені рівні : $x_1 = x_2 = -b/(2a)$; $D < 0$ - корені комплексні : $x_{1;2} = R \pm iI$, де $R = (b/(2a)$, $I = \frac{\sqrt{-D}}{2a}$.

Менш очевидним питанням є можливість появи похибок залежно від співвідношення між коефіцієнтами рівняння. Розглянемо один з найважливіших випадків, коли коефіцієнт b значно перевищує за абсолютною величиною інші. При цьому $b^2 >> 4ac$ і виникає небезпека віднімання близьких чисел в чисельнику одного з виразів (1.23) через те, що $D \approx |b|$.

Положення можна виправити різними способами. Наприклад, при b > 0 формулу для x_2 можна перетворити таким чином:

$$x_2 = \frac{\sqrt{D} - b}{2a} \frac{\sqrt{D} + b}{\sqrt{D} + b} = -\frac{2c}{\sqrt{D} + b}. \tag{1.24}$$

При $b < 0$ аналогічним чином можна записати формулу для x_1.

Більше універсальним способом являється використання значення sign b ("знак величини b") :

$$\operatorname{sign} b = \begin{cases} 1, & b \geqslant 0, \\ -1, & b < 0. \end{cases} \tag{1.25}$$

Тоді один з коренів може бути вичислений по формулі

$$x_1 = -\frac{b + \operatorname{sign} b \cdot \sqrt{D}}{2a}. \tag{1.26}$$

Вираз для обчислення значення другого кореня можна отримати за допомогою теореми Вієта. Із співвідношення $x_1 x_2 = c/a$ витікає, що

$$x_2 = \frac{c}{a x_1}. \tag{1.27}$$

На рис. 1.2 представлена блок-схема одного з варіантів алгоритму розв'язання квад-ратного рівняння з урахуванням розглянутих тут особливостей. При $D > 0$ значення коренів обчислюються по формулах (1.26) (1.27).

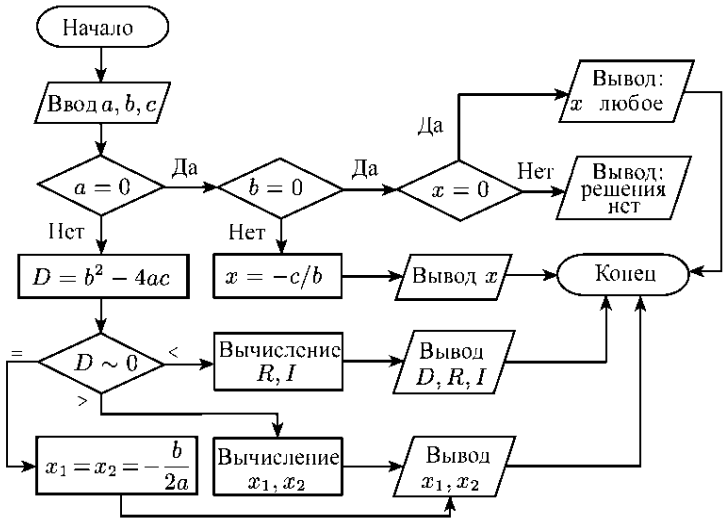

Рис. 1.2

Можна навести деякі приклади, коли реалізація цього алгоритму на комп'ютері неможлива. Припускатимемо, що обчислення проводят-ся з подвійною точністю.

Приклад

1.) $a = 10^{-200}$, $b = -3 \cdot 10^{-200}$, $c = 2 \cdot 10^{-200}$. При обчисленні добутків b^2 і $4ac$ виходить машинний нуль, т. е. $D = 0$; розв'язання піде по гілці рівних коренів : $x_1 = x_2 = 1.5$. Точні значення коренів, як неважко бачити, рівні $x_1 = 1$, $x_2 = 2$.

2.) $a = 10^{200}$, $b = -3 \cdot 10^{200}$, $c = 2 \cdot 10^{-200}$. Цей варіант аналогічний попередньому випадку з тією лише різницею, що замість отримання машинного нуля станеться переповнювання і переривання рахунку.

3.) $a = 10^{-200}$, $b = 10^{200}$, $c = -10^{200}$. Це важкий для реалізації на комп'ютері випадок. У практичних розрахунках зустрічаються рівняння з малим коефіцієнтом при x^2. В цьому випадку $b^2 >> 4ac$, але при обчисленні b^2 станеться переповнювання. Простим виходом з цього положення може бути зведення до випадку $a = 0$ з обов'язковою перевіркою інших коефіцієнтів.

Таким чином, аналіз навіть такого завдання, як розв'язання квадратного рівняння, показує, що використання чисельного алгоритму може бути зв'язане з деякими труднощами.

Стійкість. Коректність. Збіжність

Стійкість. Розглянемо похибки початкових даних. Оскільки це так звані неусувні похибки і обчислювач не може з ними боротися, то треба хоч би мати уявлення про їх вплив на точність остаточних результатів. Звичайно, ми маємо право сподіватися на те, що похибка результатів має порядок похибки початкових даних. Чи завжди це так? На жаль, ні. Деякі завдання дуже чутливі до неточностей в початкових даних. Ця чутливість характеризується так званою стійкістю.

Нехай в результаті розв'язання задачі по початковому значенню величини x знаходиться значення шуканої величини y. Якщо початкова величина має абсолютну похибка Δx, то розв'язання має похибка Δy. Задача називається *стійкою по початковому параметру x*, якщо розв'язок *y* безперервно від нього залежить, тобто малий приріст початкової величини Δx призводить до малого приросту шуканої величини Δy. Іншими словами, малі похибки в початковій величині призводять до малих похибок в розв'язанні.

Відсутність стійкості означає, що навіть незначні похибки в початкових даних призводять до великих похибок в розв'язанні або навіть до невірного результату. Про нестійкі завдання також говорять, що вони чутливі до похибок початкових даних.

Наведемо приклад нестійкого завдання. Розглянемо квадратне рівняння з параметром *a*:

$$x^2 - 2x + sign\ a = 0.$$

Функція sign визначена в (1.25). Розв'язання цього рівняння залежно від значення a таке: $x_1 = x_2 = 1$ при $a \geq 0$; $x_{1,2} = 1 \pm \sqrt{2}$ при $a < 0$. Очевидно, що при $a = 0$ скільки завгодно мала негативна похибка в завданні *a* приведе до кінцевої, а не скільки завгодно маліої похибки в розв'язанні рівняння.

Іноді буває, що теоретично завдання стійке, але проте чутливе до похибок початковим даних. Прирости початкової величини, що гарантують малість приросту шуканої величини, виявляються в цьому випадку настільки малими, що реальні малі прирости, з якими має справу обчислювач, приводять до великих похибок в розв'язанні.

Цікавою ілюстрацією такого завдання є так званий приклад Уилкинсона. Розглядається многочлен

$$P(x) = (x - 1)(x - 2)... (x - 20) = x^{20} - 210x^{19} + ...$$

Очевидно, що коренями цього многочлена є $x_1 = 1$, $x_2 = 2$, ... , $x_{20} = 20$. Припустимо, що один з коефіцієнтів многочлена вичислений з деякою малою похибкою. Наприклад, коефіцієнт -210 при x^{19} збільшимо на 10^{-7}. В результаті обчислень з подвійною точністю набудемо істотно інших значень коренів. Приведемо для наглядності ці значення, закруглені до трьох значущих цифр, :

$x_1 = 1.00$, $x_9 = 8.93$,
$x_2 = 2.00$, $x_{10,\ 11} = 10.1 \pm 0.601i$
$x_3 = 3.00$, $x_{12,\ 13} = 11.8 \pm 1.60i$,
$x_4 = 4.00$, $x_{14,\ 15} = 14.0 \pm 2.45i$,
$x_5 = 5.00$, $x_{16,\ 17} = 16.7 \pm 2.73i$,
$x_6 = 6.00$, $x_{18,\ 19} = 19.5 \pm 1.87i$,
$x_7 = 7.00$, $x_{20} = 20.8$.
$x_8 = 8.01$,

Таким чином, зміна коефіцієнта при x^{19} з -210 до $-210 + 10^{-7}$ (а це, поза сумнівом, мала зміна в звичайній обчислювальній практиці) привело до того, що половина коренів стали комплексними. Причина такого явища - чутливість завдання до похибок початкових даних; обчислення виконувалися досить точно, і похибки округлення не могли привести до таких наслідків. Помітимо, що якщо коефіцієнт -210 змінити на значно менше число, чим 10^{-7}, то зміна значень коренів стане малою. Наприклад, при збільшенні коефіцієнта -210 на 10^{-11} значень коренів, закруглених до трьох знаків, співпадуть зі значеннями коренів початкового многочлена. Приблизно такого результату і слід було чекати, оскільки завдання, що розглядається нами, стійке.

Коректність. Завдання називається *поставленим коректно*, якщо для будь-яких значень початкових даних з деякого класу, її розв'язання існує, є єдиним і стійким за початковими даними.

Розглянуте вище нестійке завдання являється некоректно поставленної. Застосовувати для вирозв'язання таких задач чисельні методи, як правило, недоцільно, оскільки похибки округлення, що виникають в розрахунках, сильно зростатимуть в ході обчислень, що приведе до значного спотворення результатів.

В той же час відмітимо, що нині розвинені методи розв'язання деяких некоректних задач. Це в основному так звані *методи регуляризації*. Вони грунтуються на заміні початкової задачі коректно поставленою задачею. Остання містить деякий параметр, при прагненні якого до нуля розв'язання цієї задачі переходить в розв'язання початкової задачі.

Нестійкість методів. Іноді при розв'язанні коректно поставленной завдання може виявитися нестійким метод її розв'язання. Такі випадки мали місце вище. Зокрема, з цієї причини при обчисленні синуса великого аргументу був отриманий результат, що не має сенсу. Розглянемо ще один приклад нестійкого алгоритму. Побудуємо численный метод обчислення інтеграла

$$I_n = \int_0^1 x^n e^{x-1} dx, \quad n = 1, 2, \ldots \tag{1.28}$$

Інтегруючи по частинах, знаходимо

$$\begin{aligned} I_1 &= \int_0^1 x e^{x-1} dx = x e^{x-1}\big|_0^1 - \int_0^1 e^{x-1} dx = \frac{1}{e}, \\ I_2 &= \int_0^1 x^2 e^{x-1} dx = x^2 e^{x-1}\big|_0^1 - 2\int_0^1 x e^{x-1} dx = 1 - 2I_1, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ I_n &= \int_0^1 x^n e^{x-1} dx = x^n e^{x-1}\big|_0^1 - n\int_0^1 x^{n-1} e^{x-1} dx = 1 - nI_{n-1}. \end{aligned} \tag{1.29}$$

Користуючись отриманим рекурентним співвідношенням, обчислюємо з подвійною точністю (приводиться результат, закруглений до трьох означаючих цифр)

I_1= 0.368, I_5= 0.146,
I_2= 0.368, ...
I_3= 0.368, I_{17}= 0.0558,
I_4= 0.368, I_{18}= 0.00369.

Значення інтеграла I_{18} не може бути негативним, оскільки підінтегральна функція $x^{18}e^{x-1}$ на усьому відрізку інтеграції [0, 1] невід'ємна. Досліджуємо джерело похибки. Максимальна абсолютна похибка при обчисленні I_1 дорівнює $0.5 \cdot 2^{-53} \approx 5 \cdot 10^{-17}$. Проте на кожному етапі ця похибка множиться на число, модуль якого більше одиниці (- 2, - 3, .., - 18), що у результаті дає $18! \approx 6.4 \cdot 10^{15}$. Це і призводить до результату, що не має сенсу. Тут знову причиною накопичення похибок є алгоритм розв'язання задачі, який виявився нестійким.

Чисельний алгоритм (метод) називається *коректним* у разі існування і єдиності чисельного розв'язання при будь-яких значеннях початкових даних, а також у разі стійкості цього розв'язання відносно похибок початкових даних.

Поняття збіжності. При аналізі точності обчислювального процесу одним з найважливіших критеріїв є *збіжність* чисельного методу. Вона означає близькість отримуваного чисельного розв'язання задачі до істинного розв'язання. Строгі визначення різних оцінок близкості можуть бути дані лише із залученням апарату функціонального аналізу. Тут ми обмежимося деякими поняттями збіжності, необходимыми для розуміння наступного матеріалу.

Розглянемо поняття *збіжності ітераційного процесу*. Цей процес полягає в тому, що для вирозв'язання деякого завдання і знаходження шуканого значення визначуваного параметра (наприклад, кореня нелінійного рівняння) будується метод послідовних наближень. В результаті багатократного повторення цього процесу (чи ітерацій) отримуємо последовательность значень x_1, x_2,.., x_n,.. Говорять, що ця последовательність сходиться до точного розв'язання x = a, якщо при необмеженому зростанні числа ітерацій межа цієї послідовності існує і рівна *a*: $\lim_{n\to\infty} x_n = a$. В цьому випадку маємо чисельний метод, що сходиться.

Інший підхід до поняття збіжності використовується в методах дискретизації. Ці методи полягають в заміні завдання з безперервними параметрами на завдання, в якому значення функцій обчислюються у фіксованих точках. Це відноситься, зокрема, до чисельного інтегруванню, рішенню диференціальних рівнянь і т. п. Тут під *збіжністю методу* розуміється прагнення значень розв'язання дискретної моделі завдання до відповідних значень розв'язання вихідної задачі при прагненні до нуля параметра дискретизації (наприклад, кроку інтегрування).

При розгляді збіжності важливими поняттями являються її вид, порядок і інші характеристики. Із загальної точки зору ці поняття розглядати тут недоцільно; до них звертатимемося при вивченні конкретних чисельних методів.

Таким чином, для отримання розв'язання задачі з необхідною точністю її постановка має бути коректною, а використовуваний чисельний метод повинен мати стійкість (коректність) і збіжність.

Підсумки

Чи́сельні ме́тоди — методи наближеного або точного розв'язування задач чистої або прикладної математики, які грунтуються на побудові послідовності дій над скінченною множиною чисел. Основні вимоги до чисельних методів, щоб вони були стійкими та збіжними.

Література

1. Мусіяка В.Г. Основи чисельних методів механіки.
2. Е.А. Волков. Численные методы: Учеб. пособие для вузов. – М.: Наука, 1987.- 248с.
3. Н.С. Бахвалов, Н.П. Жидков, Кобельков Г.М. Численные методы: Учеб. пособие. - М.: Наука, 1987 – 600с.
4. Н.С. Бахвалов, А.В. Лапин, Е.В. Чижонков. Численные методы в задачах и упражнениях. М., Высшая школа, 2000. – 190 с.
5. Н.Н. Калиткин Численные методы. М.: Наука, 1978
6. И.А. Гулин, А.А. Самарский. Численные методы. М.: Наука, 1989.
7. Березин И.С., Жидков Н.П. Методы вычислений. В 2–х т. М., 1959, т.1.– 464 с. т.2 – 602 с.
8. Турчак Л.И., Плотников П.В. Основы численных методовб 2003.

(1.1)

Приклад