1. САНДЫҚ ӘДІСТЕР ТҮСІНІГІ

Қазіргі заманғы әуе кемелерін, олардың жеке құрамдас бөліктері мен блоктарын, сондай-ақ басқа да техникалық жүйелерді жобалау және әзірлеу конструктивті параметрлерді анықтаудың алдында болатын теориялық есептеулермен және зерттеулермен байланысты. Бұл есептеулер есептеу құралдарының (компьютер және олардың жүйелері) және есептеу әдістерінің көмегімен жүзеге асырылады. Бұл әдетте келесі қадамдарды қамтиды.

1. Есептің физикалық қойылымы. Бұл кезеңнің нәтижесі мәселені мағыналы түрде жалпы тұжырымдау болып табылады, яғни не берілген және нені анықтау қажет. Мысалы, берілген қозғалтқыш күшіне, зымыран массасына, оның аэродинамикалық сипаттамаларына, белгілі бір ауа-райы жағдайында және т.б. үшін зымыранның ұшу жолын есептеу үшін. Әдетте, бұл кезеңді белгілі бір пән саласындағы маман орындайды.
2. Есептің физикалық тұжырымына сәйкес қандай да бір математикалық модельді іздеу, таңдау немесе өзгерту. Бұл кезеңде келесі жұмыстар жүргізіледі:
   - есепті сипаттайтын негізгі математикалық теңдеулерді, қатынастарды, жуықтау формулаларын таңдау (жазып алу);
   - қосымша математикалық теңдеулерді, қатынастарды, шекаралық немесе шекаралық шарттарды таңдау (жазып алу);
   - математикалық модельді алдын ала (априори) негіздеу.

   Бұл кезең өте маңызды, өйткені физикалық үлгіге сәйкес келмейтін қате немесе сәтсіз модель өнімді жобалаудағы барлық күш-жігерді жоққа шығарады. Көптеген есептерді шешу кезінде жалпы қабылданған математикалық модельдер таңдалатынын ескеріңіз
3. Математикалық (аналитикалық, жуықтау-аналитикалық немесе сандық) әдісті әзірлеу, таңдау немесе өзгерту, ең қолайлы және үнемді. Бұл кезең зерттеушілерге қолжетімді білімдер негізінде (субъективті көзқарас), сонымен қатар компьютерлік ресурстар – операциялық және сыртқы жады, жылдамдық, ақпаратты ұсыну мүмкіндіктері (объективті тәсіл) негізінде жүзеге асырылады.
4. Алгоритмді құрастыру.
5. Бағдарламалық қамтамасыз етуді әзірлеу.
6. Мәселені шешу: нақты объектке қатысты әдістемелік және параметрлік компьютерлік зерттеулер арқылы модель мен әдісті апестериорлық негіздеу. Бұл кезең жобалық бөлімдерге объектінің ұсынымдары мен сипаттамаларын беруді қамтиды.

Математикалық модельдерді зерттеудің және олардың негізінде нақты объектілердің қасиеттерін зерттеудің классикалық құралдары математикалық формулалар түрінде дәл шешімдерді алуға мүмкіндік беретін аналитикалық әдістер болып табылады. Бұл әдістер мәселені шешу туралы ең толық ақпарат береді және олар осы уақытқа дейін өз маңызын жойған жоқ. Алайда, өкінішке орай, оларды қолдануға болатын мәселелер класы өте шектеулі. Сондықтан қазіргі заманғы техникалық жүйелерді әзірлеуде есептердің кең класын шешу, әдетте, сандық әдістермен жүзеге асырылады.

Сандық әдістер – математикалық модельдерге сәйкес алгоритмдерді жүзеге асыруға негізделген қолданбалы математика есептерін жуықтап шешу әдістері. Сандық әдістерді зерттейтін ғылымды сандық талдау немесе есептеу математикасы деп те атайды.

Сандық әдістер, аналитикалық әдістерге қарағанда, жалпы емес, жеке шешімдерді береді, олар үздіксіз емес, тәуелсіз айнымалылардың өзгеруінің дискретті аймақтарында анықталады. Бұл жағдайда сандық және логикалық массивтерде арифметикалық және логикалық операциялардың жеткілікті санын орындау талап етіледі. Есептеулердің жуықтау сипатына байланысты бұл процесс, өз кезегінде, белгілі бір есептер мен сандық әдістермен (сұлбалармен) байланысты кейбір негізгі талаптармен немесе түсініктермен байланысты - есептің жақсы жағдайына байланысты орнықтылық; жинақтылық, жоғары дәлдік, үнемділік және әдіс параметрлері - дискретизация қадамдары h немесе мәселе шешілетін бастапқы аймақты бөлу, қайталану саны (итерациялық әдістер үшін), біркелкі емес бөлу үшін қадамдардың қатынасы және т.б.

Мұнда келтірілген талаптардың кейбірі қарама-қайшы, сондықтан зерттеуді орындау кезінде бір нәрсені құрбан ету керек, мысалы, әдістің дәлдігі немесе үнемділігі. Бұл ұғымдардың кейбірі

(жинақтылық, орнықтылық, жақсы шарттылық) курстың негізгі бөлімдерінде шешілетін мәселелерді қарастыру кезінде қарастырылады және нақты мазмұнмен толтырылады, сондықтан біз олардың қысқаша анықтамаларын ғана береміз.

2. ЕСЕПТЕУ ҚАТЕЛІКТЕРІ

Қателердің бір түрі бастапқы физикалық модельдің таңдалған математикалық моделінің сәйкес келмеуімен байланысты. Бұл сәйкессіздік азды-көпті дәрежеде барлық шамамен шешілетін мәселелерге тән. Бұл қате жойылмайды және апестериорлы анықталады (мәселені шешудің алтыншы кезеңін қараңыз.) Қателердің қалған үш түрі таза есептеу болып табылады және келесі себептерге байланысты.

Бастапқы деректерді орнатудың дәлсіздігі (белгісіздігі) өлшеулердің немесе есептеулердің дәлдігімен немесе деректерді дөңгелектеумен байланысты қалпына келтірілмейтін қателерге әкеледі.

Егер бастапқы деректердегі белгісіздікті жойсақ, мысалы, оларды түзету арқылы және қандай да бір сандық әдісті қолданып шешім тапсақ, бастапқы деректерге дәл сәйкес келмейтін нәтиже аламыз. Бұл сандық немесе басқа жуық әдістің қатесі (мысалы, шамамен аналитикалық); Дәл осы қателер сандық әдістерді қарастырған кезде бағаланады. Бұл бағалауларды есептеулерді орындау алдында (априорлық бағалаулар) және олардан кейін (aposteriori бағалаулар) алуға болады.

Компьютерде барлық сандар түпкілікті түрде көрсетіледі, сондықтан есептеу алгоритмін пайдалану кезінде сандармен арифметикалық және басқа амалдардағы қателер, сонымен қатар дөңгелектеу қателері жүзеге асырылады.

Исторически сложилось, что исследования по второму пункту относятся, главным образом, к различным теориям вычислительной сложности и к теории алгоритмов, которая в 30-е годы XX века вычленилась из абстрактной математической логики. Но традиционная вычислительная математика, предметом которой считается построение и исследование конкретных численных методов, также немало способствует прогрессу в этой области.

Опять же, исторические и организационные причины привели к тому, что различные вычислительные методы для решения тех или иных конкретных задач относятся к к другим математическим дисциплинам. Например, численные методы для отыскания экстремумов рзличных функций являются предметом вычислительной оптимизации, теории принятия решений и исследования операций.

1.1 Погрешности вычислений

Общеизвестно, что в практических задачах числовые данные почти всегда не вполне точны и содержат ошибки. Если эти данные являются, к примеру, результатами измерений, то за редким исключением они не могут быть произведены абсолютно точно.

Ошибкой или *погрешностью* приближённого значения $\tilde{x}$ какой-либо величины называют разность между $\tilde{x}$ и истинным значением x этой величины, т. е. $\tilde{x} - x$. Часто более удобно оперировать *абсолютной погрешностью* Δ приближённой величины, которая определяется как абсолютная величина погрешности, т. е.

$$\Delta = |\tilde{x} - x|, \tag{1.1}$$

поскольку во многих случаях знак погрешности неизвестен.

Практически точное значение интересующей нас величины x неизвестно, так что вместо точного значения абсолютной погрешности также приходится довольствоваться её приближёнными значениями. Оценку сверху для абсолютной погрешности называют *предельной абсолютной погрешностью*. В самом этом термине содержится желание иметь эту величину как можно более точной, т. е. как можно меньшей.

Таким образом, если $\tilde{\Delta}$ — предельная абсолютная погрешность значения $\tilde{x}$ точной величины x, то

$$\Delta = |\tilde{x} - x| \leq \tilde{\Delta},$$

и потому

$$\tilde{x} - \tilde{\Delta} \le x \le \tilde{x} + \tilde{\Delta}.$$

Вместо этого двустороннего неравенства удобно пользоваться следующей краткой и выразительной записью:

$$x = \tilde{x} \pm \tilde{\Delta}.$$

Фактически, вместо точного числа мы имеем здесь целый диапазон значений — числовой интервал $[\tilde{x}-\tilde{\Delta}, \tilde{x}+\tilde{\Delta}]$ возможных представителей для точного значения x.

Как правило, указание одной только абсолютной погрешности недостаточно для характеристики качества рассматриваемого приближения. Более полное понятие о нём можно получить из *относительной погрешности* приближения, которая определяется как отношение абсолютной погрешности к самому значению этой величины:

$$\delta = \frac{\Delta}{|x|}. \tag{1.2}$$

Относительная погрешность — безразмерная величина.

Предельной относительной погрешностью приближённого значения x называют всякое число $\tilde{\delta}$, оценивающее сверху его относительную погрешность. Как правило, и для абсолютной, и для относительной погрешностей в речи опускают эпитеты «предельная», поскольку именно предельные погрешности являются реальными доступными нам (наблюдаемыми) величинами.

При записи приближённых чисел имеет смысл изображать их так, чтобы сама форма их написания давала характеристику об их точности. Ясно, что ненадёжные знаки представления чисел указывать смысла нет. Обычно принимают за правило писать числа так, чтобы все их значащие цифры кроме, может быть, последней были верны, а последняя цифра была бы сомнительной не более чем на единицу. Согласно этому правилу число 12340000, у которого цифра 4 уже сомнительна, нужно записывать в виде $1.23 \cdot 10^8$.

Значащей цифрой приближённого числа называется цифра в его представлении в заданной системе счисления, отличная от нуля, либо нуль, если он стоит между значащими цифрами или является представителем сохранённого разряда этого числа. Содержательное определение может состоять в том, что значащая цифра — это цифра из

представления числа, которая даёт существенную информацию о его относительной погрешности.

Значащие цифры могут быть верными или неверными.

Как изменяются абсолютные и относительные погрешности при выполнении арифметических операций с приближёнными числами? Приближённое число с заданной абсолютной погрешностью — это, фактически, целый интервал значений. По этой причине для абсолютных погрешностей поставленный вопрос решается формулами интервальной арифметики, рссматриваемой в §1.4. Здесь мы рассмотрим упрощённые версии этих операций.

Предложение 1.1.1 *Абсолютная погрешность суммы и разности приближённых чисел равна сумме абсолютных погрешностей операндов.*

Доказательство. Если x_1, x_2 — точные значения рассматриваемых чисел, $\tilde{x}_1$, $\tilde{x}_2$ — их приближённые значения, а $\tilde{\Delta}_1$, $\tilde{\Delta}_2$ — соответствующие предельные абсолютные погрешности, то

$$\tilde{x}_1 - \tilde{\Delta}_1 \ \le\ x_1 \ \le\ \tilde{x}_1 + \tilde{\Delta}_1, \tag{1.3}$$

$$\tilde{x}_2 - \tilde{\Delta}_2 \ \le\ x_2 \ \le\ \tilde{x}_2 + \tilde{\Delta}_2. \tag{1.4}$$

Складывая эти неравенства почленно, получим

$$(\tilde{x}_1 + \tilde{x}_2) - \big(\tilde{\Delta}_1 + \tilde{\Delta}_2\big) \le\ x_1 + x_2 \ \le\ (\tilde{x}_1 + \tilde{x}_2) + \big(\tilde{\Delta}_1 + \tilde{\Delta}_2\big).$$

Полученное соотношение означает, что величина $\tilde{\Delta}_1 + \tilde{\Delta}_2$ является предельной абсолютной погрешностью суммы $\tilde{x}_1 + \tilde{x}_2$.

Умножая обе части неравенства (1.4) на (-1), получим

$$-\tilde{x}_2 - \tilde{\Delta}_2 \ \le\ -x_2 \ \le\ -\tilde{x}_2 + \tilde{\Delta}_2.$$

Складывая почленно с неравенством (1.3), получим

$$(\tilde{x}_1 - \tilde{x}_2) - \big(\tilde{\Delta}_1 + \tilde{\Delta}_2\big) \le\ x_1 + x_2 \ \le\ (\tilde{x}_1 - \tilde{x}_2) + \big(\tilde{\Delta}_1 + \tilde{\Delta}_2\big).$$

Отсюда видно, что величина $\tilde{\Delta}_1 + \tilde{\Delta}_2$ является предельной абсолютной погрешностью разности $\tilde{x}_1 - \tilde{x}_2$. ■

Для умножения и деления формулы преобразования абсолютной погрешности более громоздки. Точные результаты для операций между приближёнными величинами даются интервальной арифметикой, рассматриваемой ниже в §1.4.

Рассмотрим теперь эволюцию относительной погрешности в вычислениях.

Предложение 1.1.2 *Если слагаемые в сумме имеют одинаковый знак, то относительная погрешность суммы не превосходит наибольшей из относительных погрешностей слагаемых.*

Доказательство. Пусть складываются приближённые величины x_1 и x_2 с относительными погрешностями $\tilde{\delta}_1$ и $\tilde{\delta}_2$. Тогда их абсолютные погрешности

$$\Delta_1 = \delta_1 |x_1|, \qquad \text{и} \qquad \Delta_2 = \delta_2 |x_2|.$$

Если $\delta = \max\{\delta_1, \delta_2\}$, то

$$\Delta_1 \le \delta |x_1|, \qquad \Delta_2 \le \delta |x_2|.$$

Складывая полученные неравенства почленно, получим

$$\Delta_1 + \Delta_2 \le \delta \big(|x_1| + |x_2|\big),$$

откуда

$$\frac{\Delta_1 + \Delta_2}{|x_1| + |x_2|} \le \delta.$$

В числителе полученной дроби стоит предельная абсолютная погрешность суммы, а в знаменателе — модуль точного значения суммы, если слагаемые имеют один и тот же знак. ■

Ситуация с относительной погрешностью принципиально меняется, когда в сумме слагаемые имеют разный знак, т. е. она является разностью. Если результат имеет меньшую абсолютную величину, чем абсолютные величины операндов, то значение дроби (1.2) возрастёт. А если вычитаемые числа очень близки друг к другу, то знаменатель в (1.2) сделается очень маленьким и относительная погрешность результата может катастрофически возрасти.

Пример 1.1.1 Рассмотрим вычитание чисел 1001 и 1000, каждое из которых является приближённым и известным с абсолютной точностью 0.1. Таким образом, относительные точности обоих чисел примерно равны 0.01%. Выполняя вычитание, получим результат 1, который имеет абсолютную погрешность $0.1 + 0.1 = 0.2$. Как следствие, относительная погрешность результата достигла 20%. ■

Предложение 1.1.3 *Если погрешности приближённых чисел малы, то относительная погрешность их произведения приближённо (с точностью до членов более высокого порядка малости) равна сумме относительных погрешностей сомножителей.*

Доказательство. Пусть x_1, x_2, …, x_n — точные значения рассматриваемых чисел, $\tilde{x}_1$, $\tilde{x}_2$, …, $\tilde{x}_n$ — их приближённые значения. Обозначим также $x := x_1 x_2 \dots x_n$, $\tilde{x} := \tilde{x}_1 \tilde{x}_2 \dots \tilde{x}_n$, и пусть $f(y_1, y_2, \dots, y_n) = y_1 y_2 \dots y_n$ — функция произведения n чисел. Разлагая её в точке $(x_1, x_2, \dots, x_n)$ по формуле Тейлора с точность до членов первого порядка, получим

$$\begin{aligned}
\tilde{x} - x &= f(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n) - f(x_1, x_2, \dots, x_n) \\
&\approx \sum_{i=1}^{n} \frac{\partial f}{\partial y_i}(x_1, x_2, \dots, x_n) \cdot (\tilde{x}_i - x_i) \\
&= \sum_{i=1}^{n} x_1 \dots x_{i-1} x_{i+1} \dots x_n \, (\tilde{x}_i - x_i) \\
&= \sum_{i=1}^{n} x_1 x_2 \dots x_n \, \frac{\tilde{x}_i - x_i}{x_i}.
\end{aligned}$$

Разделив на $x = x_1 x_2 \dots x_n$ обе части этого приближённого равенства и беря от них абсолютное значение, получим с точностью до членов второго порядка малости

$$\left| \frac{\tilde{x} - x}{x} \right| = \sum_{i=1}^{n} \left| \frac{\tilde{x}_i - x_i}{x_i} \right|,$$

что и требовалось. ■

Предложение 1.1.4 *Если погрешности приближённых чисел малы, то относительная погрешность их частного приближённо (с точностью до членов более высокого порядка малости) равна сумме относительных погрешностей сомножителей.*

Доказательство. Если $u = x/y$, то

$$\Delta u \approx \frac{\partial u}{\partial x} \Delta x + \frac{\partial u}{\partial y} \Delta y = \frac{\Delta x}{y} - \frac{x \, \Delta y}{y^2}.$$

Поэтому

$$\frac{\Delta u}{u} = \frac{\Delta x}{y\frac{x}{y}} - \frac{x\,\Delta y}{y^2\frac{x}{y}} = \frac{\Delta x}{x} - \frac{\Delta y}{y},$$

так что

$$\left|\frac{\Delta u}{u}\right| \le \left|\frac{\Delta x}{x}\right| + \left|\frac{\Delta y}{y}\right|.$$

Это и требовалось показать. ■

1.2 Компьютерная арифметика

Для правильного учёта погрешностей реализации вычислительных методов на различных устройствах и для правильной организации этих методов нужно знать детали конкретного способа вычислений. В современных электронных цифровых вычислительных машинах (ЭЦВМ), на которых выполняется подавляющая часть современных вычислений, эти детали реализации регламентируются специальным международным стандартом. Он был принят в 1985 году Институтом инженеров по электротехнике и электронике[1], профессиональной ассоциацией, объединяющей в своих рядах также специалистов по аппаратному обеспечению ЭВМ. Этот стандарт, коротко называемый IEEE 754, в 1995 году был дополнен и развит следующим стандартом, названным IEEE 854 [26, 34].

Согласно этим стандартам вещественные числа представляются в ЭВМ в виде «чисел с плавающей точкой», в которых число хранится в форме мантиссы и показателя степени. Зафиксируем натуральное число β, которое будет называться основанием системы счисления. *Числами с плавающей точкой* называются числа вида

$$\left(\alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \ldots + \alpha_p\beta^{-p}\right)\cdot\beta^e,$$

которые условно можно записать в виде

$$0\,.\,\alpha_1\alpha_2\ldots\alpha_p\cdot\beta^e,$$

где $0 \le \alpha_i < \beta$, $i = 1, 2, \ldots, p$. В выписанном представлении величина $0\,.\,\alpha_1\alpha_2\ldots\alpha_p$ называется *мантиссой* числа, а p — количество значащих

[1] Чаще всего его называют английской аббревиатурой IEEE от Institute of Electrical and Electronics Engineers.

цифр мантиссы — это *точность* рассматриваемой модели с плавающей точкой. На показатель степени e также обычно накладывается двустороннее ограничение $e_{\min} \le e \le e_{\max}$.

Стандарты IEEE 754/854 предписывают для цифровых ЭВМ значения $\beta = 2$ или $\beta = 10$, и в большинстве компьютеров используется $\beta = 2$, т. е. двоичная система счисления. С одной стороны, это вызвано особенностями физичекой реализации современных ЭВМ, где 0 соответствует отсутствию сигнала (заряда и т. п.), а 1 — его наличию. С другой стороны, двоичная система оказывается выгодной при выполнении с ней приближённых вычислений (см. [26]).

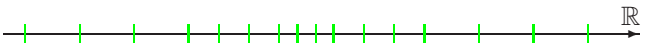

Рис. 1.1. Множество чисел, представимых в цифровой ЭВМ — дискретное конечное подмножество вещественной оси $\mathbb{R}$.

Как видим, числа с плавающей точкой обеспечивают практически фиксированную относительную погрешность представления вещественных чисел и изменяющуюся абсолютную погрешность.

Стандарты IEEE 754/854 предусматривают для чисел с плавающей точкой «одинарную точность» и «двойную точность», а также «расширенные» варианты этих представлений. При этом для хранения чисел одинарной точности отводится 4 байта памяти ЭВМ, для двойной точности — 8 байтов. Из этих 32 или 64 битов один бит зарезервирован для указания знака числа: 0 соответствует «−», а 1 соответствует «+». Таким образом, во внутреннем «машинном» представлении знак присутствует у любого числа, в том числе и у нуля.

Для двойной точности, наиболее широко распрстранённой в современных расчётах, диапазоп чисел, представимых в ЭВМ простирается от примерно $2.22 \cdot 10^{-308}$ до $1.79 \cdot 10^{308}$. Помимо обычных чисел стандарты IEEE 754/854 описывают несколько специальных объектов вычислений. Это, прежде всего, машинная бесконечность и специальный нечисловой объект под названием `NaN` (названный как сокращение английской фразы «Not a Number»). `NaN` полезен во многих ситуациях, в частности, он может использоваться для сигнализации о нетипичных и исключительных событиях, случившихся в процессе вычислений, которые, тем не менее, нельзя было прерывать.

Очень важной характеристикой множества машинных чисел явля-

ется так называемое «машинное ε» (машинное эпсилон), которое характеризует густоту множества машинно-представимых чисел. Это наименьшее положительное число $\varepsilon_{\text{маш}}$, такое что в компьютерной арифметике $1 + \varepsilon_{\text{маш}} \neq 1$ при округлении к ближайшему. Из конструкции чисел с плавающей точкой следует тогда, что компьютер, грубо говоря, не будет различать чисел a и b, удовлетворяющих условию $1 < a/b < 1 + \varepsilon_{\text{маш}}$. Для двойной точности представления в стандарте IEEE 754/854 машинное эпсилон примерно равно $1.11 \cdot 10^{-16}$.

Принципиальной особенностью компьютерной арифметики, вызванной дискретностью множества машинных чисел и наличием округлений, является невыполнение некоторых общеизвестных свойств вещественной арифметики. Например, сложение чисел с плавающей точкой неассоциативно, т. е. в общем случае неверно, что

$$(a + b) + c = a + (b + c).$$

Читатель может проверить на любом компьютере, что в арифметике IEEE 754/854 двойной точности при округлении «к ближайшему»

$$\left(1 + 1.1{\cdot}10^{-16}\right) + 1.1{\cdot}10^{-16} \;\neq\; 1 + \left(1.1{\cdot}10^{-16} + 1.1{\cdot}10^{-16}\right).$$

Левая часть этого отношения равна 1, тогда как правая — ближайшему к единице справа машинно-представимому числу. Эта ситуация имеет место в любых приближённых вычислениях, которые сопровождаются округлениями, а не только при расчётах на современных цифровых ЭВМ.

Из отсутствия ассоциативности следует, что результат суммирования длинных сумм вида $x_1 + x_2 + \ldots + x_n$ зависит от порядка, в котором выполняется попарное суммирование слагаемых, или, как говорят, от расстановки скобок в сумме. Каким образом следует организовывать такое суммирование в компьютерной арифметике, чтобы получать наиболее точные результаты? Ответ на этот вопрос существенно зависит от значений слагаемых, но в случае суммирования уменьшающихся по абсолютной величине величин суммировать нужно «с конца». Именно так, к примеру, лучше всего находить суммы большинства рядов.

1.3 Обусловленность математических задач

Вынесенный в заголовок этого параграфа термин — *обусловленность* — означает меру чувствительности решения задачи к изменениям (возмущениям) её входных данных. Ясно, что любая информация