

Численные методы (анализ, алгебра, обыкновенные дифференциальные уравнения), Н. С. Бахвалов. Главная редакция физико-математической литературы изд-ва «Наука», М., 1975 г.

В книге рассматриваются основные положения численных методов, относящиеся к приближению функций, интегрированию, задачам алгебры и оптимизации, решению обыкновенных дифференциальных уравнений.

Значительное внимание уделяется вопросам выбора методов и организации вычислений при решении большого числа однотипных задач.

Книга предназначена для студентов университетов и технических вузов с расширенной программой по математике, специализирующихся по прикладной и вычислительной математике, а также для лиц, интересующихся теорией и практикой численных методов.

Таблиц 1, рисунков 62, библиографических названий 126.

ОГЛАВЛЕНИЕ

Предисловие	7
Введение	9

ЧАСТЬ I. ЧИСЛЕННЫЕ МЕТОДЫ МАТЕМАТИЧЕСКОГО АНАЛИЗА

Глава I. Погрешность результата численного решения задачи	15
§ 1. Источники и классификация погрешности	15
§ 2. Запись чисел в ЭВМ	18
§ 3. Абсолютная и относительная погрешности. Формы записи данных	19
§ 4. О вычислительной погрешности	21
§ 5. Погрешность функции	23
Глава II. Интерполяция и смежные вопросы	30
§ 1. Постановка задачи приближения функций	31
§ 2. Интерполяционный многочлен Лагранжа	35
§ 3. Оценка остаточного члена интерполяционного многочлена Лагранжа	37
§ 4. Разделенные разности и их свойства	37
§ 5. Интерполяционная формула Ньютона с разделенными разностями	39
§ 6. Разделенные разности и интерполирование с кратными узлами	42
§ 7. Уравнения в конечных разностях	47
§ 8. Многочлены Чебышева	56
§ 9. Минимизация оценки остаточного члена интерполяционной формулы	59
§ 10. Конечные разности	62
§ 11. Интерполяционные формулы Ньютона для равных промежутков	65
§ 12. Интерполяционные формулы Бесселя и Эверетта. Составление таблиц	67
§ 13. О погрешности округления при интерполировании	75
§ 14. Применение аппарата интерполирования. Обратная интерполяция	77
§ 15. Ортогональные системы и их свойства	78
§ 16. Ортогональные многочлены	85
§ 17. Численное дифференцирование	88
§ 18. О вычислительной погрешности формул численного дифференцирования	92
Глава III. Численное интегрирование	95
§ 1. Квадратурные формулы Ньютона — Котеса	95
§ 2. Оценка погрешности квадратурной формулы на классе функций	103
§ 3. Квадратурные формулы Гаусса	107
§ 4. Практическая оценка погрешности элементарных квадратурных формул	119
§ 5. Интегрирование сильно осциллирующих функций	125

§ 6	Повышение точности интегрирования за счет разбиения отрезка на равные части	128
§ 7	О постановках задач оптимизации	134
§ 8	Оптимальные квадратуры на классах функций с одной производной	139
§ 9	Оптимизация распределения узлов квадратурной формулы	147
§ 10	Примеры оптимизации распределения узлов	153
§ 11	Главный член погрешности	159
§ 12	Формулы Эйлера и Грегори	163
§ 13	Правило Рунге практической оценки погрешности	167
§ 14	Формулы Ромберга	174
§ 15	Эксперименты и их обсуждение	178
§ 16	Вычисление интегралов в нерегулярном случае	185
§ 17	Принципы построения стандартных программ с автоматическим выбором шага	193
§ 18	Стандартные программы численного интегрирования	201
Глава IV	Приближение функций и смежные вопросы	210
§ 1	Наилучшие приближения в линейном нормированном пространстве	210
§ 2	Наилучшее приближение в гильбертовом пространстве и вопросы, возникающие при его практическом построении	212
§ 3	Дискретное преобразование Фурье	218
§ 4	Быстрое преобразование Фурье	222
§ 5	Наилучшее равномерное приближение	225
§ 6	Примеры наилучшего равномерного приближения	228
§ 7	Итерационный метод построения многочлена наилучшего равномерного приближения	235
§ 8	О форме записи многочлена	242
§ 9	О способах вычисления элементарных функций	249
§ 10	О скорости приближения функций различных классов	253
§ 11	Интерполяция и приближение сплайнами	256
§ 12	Энтропия и ϵ -энтропия	262
Глава V	Многомерные задачи	270
§ 1	Метод неопределенных коэффициентов	271
§ 2	Метод наименьших квадратов	272
§ 3	Метод регуляризации	274
§ 4	Пример регуляризации	275
§ 5	Сведение многомерных задач к одномерным	281
§ 6	Оценка погрешности численного интегрирования по равномерной сетке	289
§ 7	Оценка снизу погрешности численного интегрирования	292
§ 8	Об оптимизации оценки погрешности на более широких классах способов интегрирования	295
§ 9	Метод Монте-Карло	300
§ 10	Обсуждение правомерности использования недетерминированных методов решения задач	305
§ 11	Ускорение сходимости метода Монте-Карло	307
§ 12	Квадратурные формулы повышенной точности со случайными узлами	311
§ 13	О выборе метода решения задачи	316

ЧАСТЬ II ЗАДАЧИ АЛГЕБРЫ И ОПТИМИЗАЦИИ

Глава VI	Численные методы алгебры	323
§ 1	Методы последовательного исключения неизвестных	324
§ 2	Метод ортогонализации	333

§ 3	Метод простой итерации	335
§ 4	Исследование реального итерационного процесса	340
§ 5	Спектр семейства матриц	343
§ 6	δ^2 -процесс практической оценки погрешности и ускорения сходимости	349
§ 7	Оптимизация скорости сходимости итерационных процессов	352
§ 8	Метод Зейделя	363
§ 9	Метод наискорейшего градиентного спуска	369
§ 10	Метод сопряженных градиентов	372
§ 11	Метод Монте-Карло решения систем линейных уравнений	378
§ 12	Итерационные методы с использованием спектрально эквивалентных операторов	385
§ 13	Погрешность приближенного решения системы уравнений и обусловленность матриц. Регуляризация	388
§ 14	Проблема собственных значений	394
§ 15	Решение полной проблемы собственных значений для симметричной матрицы методом вращений	400
Глава VII. Решение систем нелинейных уравнений и задач оптимизации		405
§ 1	Метод простой итерации и смежные вопросы	437
§ 2	Метод Ньютона решения нелинейных уравнений	411
§ 3	Другие методы решения одного уравнения	416
§ 4	Методы спуска	420
§ 5	Другие методы сведения многомерных задач к задачам меньшей размерности	425
§ 6	Решение стационарных задач путем установления	429
§ 7	Что оптимизировать?	436
§ 8	Как оптимизировать?	440
Часть III. Численные методы решения обыкновенных дифференциальных уравнений		
Глава VIII. Численные методы решения задачи Коши		447
§ 1	Разложение решения в ряд Тейлора	448
§ 2	Методы Рунге — Кутты	450
§ 3	Методы с контролем погрешности на шаге	459
§ 4	Оценка погрешности одношаговых методов	461
§ 5	Конечно-разностные методы	466
§ 6	Метод неопределенных коэффициентов	471
§ 7	Исследование свойств конечно-разностных методов на модельных задачах	476
§ 8	Оценка погрешности конечно-разностных методов	483
§ 9	Главный член погрешности	488
§ 10	Изучение свойств конечно-разностных методов на более точных моделях	493
§ 11	Интегрирование систем уравнений	502
§ 12	Ряд общих вопросов	512
§ 13	Формулы численного интегрирования уравнений второго порядка	519
§ 14	Оценка погрешности численного решения задачи Коши для уравнения второго порядка	523
§ 15	Двусторонние методы	528
Глава IX. Численные методы решения краевых задач для обыкновенных дифференциальных уравнений		535
§ 1	Простейшие методы решения краевой задачи для уравнения второго порядка	535
§ 2	Функция Грина сеточной краевой задачи	542

§ 3. Решение простейшей краевой сеточной задачи	548
§ 4. Замыкания вычислительных алгоритмов	557
§ 5. Обсуждение постановок краевых задач для линейных систем первого порядка	565
§ 6. Алгоритмы решения краевых задач для систем уравнений пер- вого порядка	571
§ 7. Методы дифференциальной ортогональной прогонки	577
§ 8. Нелинейные краевые задачи	583
§ 9. Аппроксимации специального типа	593
§ 10. Конечно-разностные методы отыскания собственных значений	601
§ 11. Оптимизация распределения узлов интегрирования	606
§ 12. Влияние вычислительной погрешности в зависимости от формы записи конечно-разностного уравнения	612
§ 13. Оценка вычислительной погрешности при решении краевой за- дачи методом прогонки	618
Список литературы	622
Предметный указатель	628

ПРЕДИСЛОВИЕ

Настоящая книга написана на основе лекций, читавшихся автором последние десять лет на механико-математическом факультете и на факультете вычислительной математики и кибернетики МГУ и на основе опыта работы в ВЦ МГУ.

Как и всякое другое пособие по численным методам, эта книга содержит изложение основных положений теории, в данном случае относящихся к приближению функций, интегрированию, задачам алгебры и оптимизации, решению дифференциальных уравнений.

Текущий период характерен бурным развитием вычислительной техники; в частности, мощность парка ЭВМ возрастает на порядки в течение десятилетия. Это обстоятельство сопровождается усиленным применением ЭВМ и численных методов для решения резко расширяющегося круга задач. В результате происходит быстрое изменение взглядов на весь комплекс вопросов, связанных с применением ЭВМ, и требований к численным методам решения задач. В частности, в связи с вышесказанным нельзя предложить пособия по численным методам, содержащего рецепты решения всех возможных реальных задач. При выборе метода решения сложной конкретной задачи всякое пособие зачастую играет роль лишь общего руководства, отталкиваясь от которого исследователь рассматривает свои проблемы.

Работа, связанная с созданием и применением численных методов, складывается из чисто теоретических исследований методов решения типичных задач, анализа работы алгоритмов при решении модельных задач, вычислительного эксперимента и ряда других моментов. Большое значение имеют вопросы выбора направления исследования, построения математических моделей рассматриваемых явлений, организации контактов с представителями других наук.

В каждом конкретном случае эти проблемы приобретают свои оттенки и трудно дать какие-либо общие рекомендации, пригодные во всех случаях. Поэтому рассуждения автора по затрагиваемым вопросам следует рассматривать не как руководство к действию, а просто как одну из возможных точек зрения.

Эффект от применения ЭВМ и численных методов достигается как в результате решения отдельных трудных задач, так и за счет расширения круга пользователей ЭВМ путем предоставления им больших удобств. Для последней цели важно составление стандартных программ решения типичных математических задач. Поэтому в рассматриваемой книге определенное внимание уделено также вопросам теории численных методов, возникающим при разработке таких стандартных программ.

Автор многим обязан своему отцу, профессору С. В. Бахвалову, академикам А. Н. Колмогорову, С. Л. Соболеву, А. Н. Тихонову, члену-корреспонденту АН СССР Н. Г. Четаеву, профессорам И. С. Березину, С. К. Годунову, доцентам Н. П. Жидкову, В. Г. Карманову, А. Ф. Филиппову, оказавшим большое влияние на формирование взглядов автора по рассматриваемым вопросам.

Автор выражает благодарность рецензенту профессору Е. А. Волкову за весьма ценные замечания, которые существенно помогли автору.

Автор благодарен профессорам В. В. Воеводину, Н. Н. Кузнецову, доцентам Ф. П. Васильеву, Е. Г. Дьяконову, Я. М. Жилейкину, В. А. Морозову за полезные обсуждения, способствовавшие улучшению рукописи.

Автор признателен за большую помощь при работе над рукописью Б. И. Березину, Г. М. Кобелькову, А. Б. Кукаркину, К. Ю. Осипенко, Р. В. Разумейко.

Н. С. Бахвалов

ВВЕДЕНИЕ

Прежде чем переходить к непосредственному изложению теории численных методов, попытаемся определить ее место в системе других областей знания и рассказать о проблемах, возникающих в связи с ее применением.

Первоначально элементы математики появились в связи с необходимостью решения практических задач: измерений на местности, навигации и т. д. Вследствие этого математика была численной математикой — ее целью являлось получение решения в виде числа.

Численное решение прикладных задач всегда интересовало математиков. Крупнейшие математики прошлого сочетали в своих исследованиях изучение явлений природы, получение их математического описания (как иногда говорят, математической «модели» явления) и его исследование. Анализ усложненных моделей потребовал создания специальных, численных методов решения задач. Названия некоторых из таких методов — методы Ньютона, Эйлера, Лобачевского, Гаусса, Чебышева, Эрмита — свидетельствуют о том, что их разработкой занимались крупнейшие ученые своего времени.

Настоящее время характерно резким расширением приложений математики, во многом связанным с созданием и развитием средств вычислительной техники. В результате создания ЭВМ с программным управлением менее чем за тридцать лет скорость выполнения арифметических операций возросла от 0,1 операции в секунду при ручном счете до 3 000 000 операций на современных серийных ЭВМ, т. е. примерно в $(5000)^2$ раз.

Рост возможностей в связи с созданием вычислительной техники носит качественный характер и иногда сравнивается с промышленной революцией, вызванной открытием паровой машины. Уместно вспомнить, что в итоге промышленной революции и последующего развития науки и техники на протяжении двух веков скорость передвижения возросла от скорости пешехода 6 км/час до скорости космонавта 30 000 км/час, т. е. в 5000 раз.

Распространенное мнение о всемогуществе современных ЭВМ часто порождает впечатление, что математики избавились почти от всех хлопот, связанных с численным решением задач,

и разработка новых методов решения задач уже не столь существенна. В действительности это не так.

Один из общих законов развития науки состоит в том, что потребности развития общества часто ставят перед ней задачи, несколько превышающие ее возможности. Расширение возможностей приложения математики обусловило математизацию других разделов науки: химии, экономики, биологии, геологии, географии, психологии, медицины и конкретных разделов техники. Процесс математизации состоит в построении математических моделей процессов и явлений и разработке методов их исследования.

В других разделах науки, например в физике или механике, построение математических моделей для описания различных явлений природы и изучение этих моделей с целью объяснения старых или предсказания новых эффектов явлений являются традиционными. Примером успешного изучения математических моделей в сочетании с обработкой результатов наблюдений задолго до появления ЭВМ явилось открытие неизвестной ранее планеты — Нептуна.

Однако в целом работа в этом направлении зачастую продвигалась относительно медленно, поскольку обычно не удавалось получить решение возникающих математических задач и приходилось ограничиваться рассмотрением простейших моделей. Применение ЭВМ и расширение математического образования резко увеличило возможности в направлении построения и исследования математических моделей. Все чаще результаты расчетов позволяют обнаруживать и предсказывать ранее никогда не наблюдавшиеся явления; это дает основания говорить о математическом эксперименте. В некоторых направлениях исследования доверие к результатам численных расчетов так велико, что при расхождении между результатами расчетов и экспериментов в первую очередь ищут ошибку в результатах экспериментов.

Современные успехи в решении таких важных для общества задач, как атомные, космические, экономические, вряд ли были бы возможны без применения ЭВМ и численных методов.

Требование численного решения новых задач привело к появлению большого количества новых методов. Наряду с этим последние тридцать лет происходило интенсивное теоретическое осмысливание как старых, так и новых методов, их систематизация.

Теоретические исследования в области численных методов в основном группируются вокруг численных методов решения так называемых типичных математических задач. В настоящее время сюда принято относить задачи анализа (приближение, дифференцирование, интегрирование), задачи алгебры, решение дифференциальных и интегральных уравнений, задачи оптимизации, обратные задачи; отметим, что два последних направле-

ния теории численных методов начали формироваться в последние пятнадцать лет. Эти теоретические исследования оказывают большую помощь при решении конкретных задач и играют существенную роль в наблюдаемом сейчас широком распространении сферы приложений ЭВМ и математики вообще.

Рассмотрим поучительный пример. Решение дифференциальных уравнений в частных производных сводится к решению систем линейных уравнений с матрицей, в каждой строке которой имеется 5—10 ненулевых элементов. Накануне появления ЭВМ такие системы уравнений решали в случае числа неизвестных порядка 10 — 10^2 ; сейчас нередки случаи, когда решаются системы с числом неизвестных порядка 10^5 — 10^6 . В гипотетическом случае решения этих задач на современных ЭВМ методами, известными тридцать лет назад, пришлось бы ограничиться системами уравнений с числом неизвестных порядка 10^3 — 10^4 (при тех же затратах времени ЭВМ).

Быстрое проникновение математики во многие области знания, в частности, объясняется тем, что математические модели и методы их исследования применимы сразу ко многим явлениям, сходным по своей формальной структуре. Часто математическая модель, описывающая какое-либо явление, появляется при изучении других явлений или при абстрактных математических построениях задолго до конкретного рассмотрения данного явления. В частности, и в теории численных методов, так же как в чистой математике, полезна разработка общих построений. Однако есть разница в подходе «чистого» и «прикладного» математика к решению какой-либо проблемы. На языке первого понятие решить задачу означает доказать существование решения и предложить процесс, сходящийся к решению. Сами по себе эти результаты полезны для прикладника, но, кроме этого, ему нужно, чтобы процесс получения приближения не требовал больших затрат, например, времени или памяти ЭВМ. Ему важно не только то, что процесс сходится, но и то, как быстро он сходится. При численном решении задач возникают также новые вопросы, связанные с устойчивостью результата относительно возмущений исходных данных и округлений при вычислениях.

Наряду с теорией численных методов период бурного роста переживает и ряд других разделов математики, непосредственно обязанных ЭВМ своим возникновением. Применение численных методов и ЭВМ к решению естественнонаучных задач оказывает влияние и на традиционные разделы математики. Например, направление исследования квазилинейных гиперболических систем во многом развивается под влиянием контакта с численной математикой.

Математика возникла и развивается как часть естествознания, и долгое время ее развитие существенным образом

определялось потребностями физики и механики. Требование математизации новых разделов науки неизбежно приводит к обратному влиянию этих разделов науки на развитие математики и должно существенно изменить лицо самой математики.

Отметим еще один канал влияния прикладной стороны математики на развитие ее теоретических разделов. Прикладные исследования имеют непосредственную отдачу; это усиливает доверие общества к науке, расширяет понимание ее проблем и имеет следствием усиленное вложение средств с целью ее развития.

При реальной работе в области приложений математики возникает большое число осложнений самого различного, зачастую нематематического характера.

Хотя трудно надеяться, что какие-либо теоретические нравования могут заменить собственный опыт работы, попытаемся, иногда не совсем всерьез, обратить внимание на некоторые вопросы общего характера, важные для работы в области приложений математики. Проводимая ниже систематизация этих вопросов является довольно случайной, условной; по-видимому, можно предложить еще добрый десяток подобных классификаций, имеющих не меньшее право на существование.

1. Первостепенное значение имеет выбор направления исследования. Свобода выбора направления исследования обычно довольно невелика, так как основные контуры направления исследования обычно задаются «извне».

При выборе направления исследования в пределах имеющихся возможностей полезно иметь в виду следующее правило трех частей, по своему внешнему виду похожее на шутку: проблемы делятся на I — легкие, II — трудные, III — очень трудные. Проблемами I заниматься не стоит, они будут решены ходом событий и без вашего вмешательства, проблемы III вряд ли удастся решить в настоящее время; поэтому стоит обратиться к проблемам II.

2. Нужно уметь сформулировать на языке математики конкретные задачи физики, механики, экономики, инженерные задачи и т. д., т. е. построить математическую модель рассматриваемого явления.

Начинающий работу математик часто жалуется на трудности контактов с представителями других наук, которые «даже» не могут сформулировать стоящих перед ними задач. Правильное формулирование задачи — это научная проблема не менее сложная, чем само решение задачи, и не нужно надеяться, что кто-то другой целиком сделает это за вас. При постановке проблемы первостепенное внимание должно быть уделено выяснению цели исследования; приписываемая математическая модель явления не есть что-то однозначное, раз навсегда связанное с

этим явлением, а зависит от цели исследования. Прежде чем выписывать дифференциальные уравнения, выбирать метод решения и обращаться к ЭВМ, стоит подумать, а не будут ли бесполезны все результаты вычислений? В то же время надо воспринимать как должное, что большая часть результатов вычислений будет выброшена сразу же после их получения. Дело в том, что производимая работа зачастую носит исследовательский характер, и трудно заранее предсказать, что и в какой форме следует получить, на каком пути нужно искать численное решение задачи. Цель исследования и описание проблемы обычно уточняются в процессе контактов представителей конкретных наук или руководства организацией («заказчиков») и математиков («исследователей» или «исполнителей»).

3. Успех в прикладной науке требует широкой математической подготовки, поскольку только такая подготовка может обеспечить приспособляемость к непрерывно меняющимся типам задач, предъявляемых к решению. Одной из причин необходимости изучения «бесполезных» для практики на первый взгляд разделов математики является достижение более уверенного и более свободного владения «нужными» разделами математики.

4. Не следует думать, что совершенное знание математики, численных методов и навыки работы с ЭВМ позволяют сразу решить любую прикладную математическую задачу. Во многих случаях требуется доводка методов, приспособление их к решению конкретных задач. При этом типична обстановка, когда используются методы, применение которых теоретически не обосновано, или теоретические оценки погрешности численного метода неприемлемы для практического использования; при выборе метода решения задачи и анализа результатов приходится полагаться на опыт предшествующего решения задач, на интуицию и сравнение с экспериментом; и при этом приходится отвечать за достоверность результата. Поэтому для успеха в работе необходимы развитое неформальное мышление, умение рассуждать по аналогии, дающие основания ручаться за достоверность результата там, где с позиций логики и математики, вообще говоря, ручаться нельзя.

В рассматриваемом вопросе есть и другая сторона. При численном решении конкретных трудных задач, возникающих в других областях знания, математик действует зачастую как естествоиспытатель, полагаясь во многом лишь на опыт и «правдоподобные» рассуждения. Крайне желательно, чтобы такая эмпирическая работа подкреплялась теоретическими разработками методов, аккуратной проверкой качества методов на контрольных задачах с известным решением, объективным сравнением с экспериментом. При длительном продвижении в каком-то направлении без такого подкрепления может теряться перспектива

работы, полная уверенность в правильности получаемых результатов. Известное высказывание, что «хороший теоретик может истолковать в желаемом ему направлении любые результаты как расчетов, так и эксперимента», содержит большую долю истины.

5. После завершения расчетов наступает этап использования результатов вычислений в практической деятельности, или, как часто говорят, этап внедрения результатов. Правильнее будет сказать, что подготовка к использованию результатов начинается уже с анализа постановки задачи и в процессе ее решения, и по существу все моменты решения задачи и внедрения неразрывно связаны между собой; в процессе формулирования задачи и ее решения заказчик и исполнитель взаимно уточняют постановку задачи и тем самым подготавливают почву для ее внедрения. Математика в сочетании с ЭВМ применяется в самых разнообразных областях, и часто приходится иметь дело с заказчиками, не имеющими опыта применения ЭВМ. В процессе контакта с такими «начинающими» заказчиками особенно важно преодолеть их первоначальное недоверие к вторжению математики; результаты вычислений будут использоваться только тогда, когда заказчик осмыслит их со своих позиций и убедится в том, что их действительно можно и нужно использовать. При правильном подходе к взаимным контактам к концу процесса решения задачи начинающий заказчик приходит к пониманию, что ЭВМ и математика могут дать ему не все, но довольно много, а «начинающий» математик — к пониманию того, что он дает заказчику кое-что, но далеко не все нужное для реального решения задачи.

Поучительна история все с той же планетой Нептун. За год до Леверье Адамс путем расчетов открыл существование этой планеты; однако он не установил контакта с астрономами, и его результат не был своевременно проверен экспериментом.

6. Существенным моментом в прикладной работе является необходимость получения результатов в установленный срок.

Заказчик, для которого проводятся исследования (расчеты), часто имеет конкретный срок завершения исследований и принятия решения на их основе. Если исследования не будут завершены к сроку, то решение все равно будет принято, но на основании более грубого, эмпирического или просто «волевого» подхода. Потерянное в таком случае доверие со стороны заказчика часто бывает невозможно восстановить.

В такой ситуации лучше найти удовлетворительное решение задачи, но в срок, чем получить полное решение задачи к тому времени, когда оно станет бесполезным.

Дополнительная литература к введению

I [10], II [30], [31], [35].

Часть I

ЧИСЛЕННЫЕ МЕТОДЫ МАТЕМАТИЧЕСКОГО АНАЛИЗА

Глава I

ПОГРЕШНОСТЬ РЕЗУЛЬТАТА ЧИСЛЕННОГО РЕШЕНИЯ ЗАДАЧИ

§ 1. Источники и классификация погрешности

Погрешность решения задачи обуславливается следующими причинами:

1. Математическое описание задачи является неточным, в частности, неточно заданы исходные данные описания.

2. Применяемый для решения метод часто не является точным: получение точного решения возникающей математической задачи требует неограниченного или неприемлемо большого числа арифметических операций, и поэтому вместо получения точного решения задачи приходится прибегать к приближенному.

3. При вводе данных в машину, при выполнении арифметических операций и при выводе данных производятся округления.

Погрешности, соответствующие этим причинам, называют:

- 1) *неустранимой погрешностью*,
- 2) *погрешностью метода*,
- 3) *вычислительной погрешностью*.

Дадим иллюстрацию этих определений. Пусть имеется маятник (рис. 1.1.1)*), начинающий движение в момент $t = t_0$. Требуется предсказать его положение в момент t_1 .

Дифференциальное уравнение, описывающее колебание этого маятника, берется в виде

$$l \frac{d^2\varphi}{dt^2} + g \sin \varphi + \mu \frac{d\varphi}{dt} = 0, \quad (1)$$

l — длина маятника, g — ускорение силы тяжести, μ — коэффициент трения.

Как только принято такое описание, решение задачи уже приобрело неустранимую погрешность, в частности, потому, что

*) Тройная нумерация ссылок означает главу, параграф, номер формулы или рисунка; двойная — параграф и номер (в данной главе); одинарная — только номер (в данном параграфе).

реальное трение зависит от скорости не совсем линейно; другой источник неустранимой погрешности состоит в ошибках определения $l, g, \mu, t_0, \varphi(t_0), \varphi'(t_0)$.

Название этой погрешности — «неустраняемая» соответствует ее существу: она неконтролируема в процессе численного решения задачи и может уменьшиться только за счет более точного описания физической задачи и более точного определения параметров. Дифференциальное уравнение (1) не решается в явном виде; для его решения потребуется применить какой-либо численный метод. Вследствие этой причины и возникает погрешность метода.

Вычислительная погрешность может возникнуть, например, из-за конечности количества разрядов чисел, участвующих в вычислениях.

Введем формальные определения.

Пусть I — точное значение отыскиваемого параметра, например, в данном случае, реальный угол отклонения маятника φ в момент времени t_1 , $\tilde{I}$ — значение этого параметра, соответствующее принятому математическому описанию, в данном случае — значение $\varphi(t_1)$ решения уравнения (1), $\tilde{I}_h$ — решение задачи, получаемое при реализации численного метода в предположении отсутствия округлений, $\tilde{I}_h^*$ — приближение к решению задачи, получаемое при реальных вычислениях. Тогда

$$\begin{aligned} \rho_1 &= \tilde{I} - I - \text{неустраняемая погрешность,} \\ \rho_2 &= \tilde{I}_h - \tilde{I} - \text{погрешность метода,} \\ \rho_3 &= \tilde{I}_h^* - \tilde{I}_h - \text{вычислительная погрешность.} \end{aligned} \quad (2)$$

Полная погрешность, т. е. разность между реально получаемым и точным решением задачи, $\rho_0 = \tilde{I}_h^* - I$ удовлетворяет очевидному равенству

$$\rho_0 = \rho_1 + \rho_2 + \rho_3. \quad (3)$$

Во многих случаях под термином «погрешность» того или иного вида удобно понимать не рассмотренные выше разности между приближениями, а некоторые меры близости между ними. Например, в скалярном случае, полагают

$$\rho_0 = |\tilde{I}_h^* - I|, \quad \rho_1 = |\tilde{I} - I|, \quad \rho_2 = |\tilde{I}_h - \tilde{I}|, \quad \rho_3 = |\tilde{I}_h^* - \tilde{I}_h|;$$

при таких обозначениях вместо (3) получаем

$$\rho_0 \leq \rho_1 + \rho_2 + \rho_3. \quad (4)$$

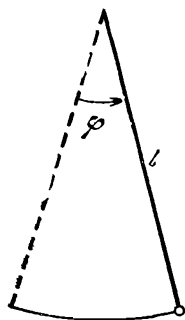


Рис. 1.1.1.

В других случаях решение I и приближения $\tilde{I}$, $\tilde{I}_h$, $\tilde{I}_h^*$ оказываются элементами некоторых функциональных пространств, часто различных. Например, I может быть элементом пространства F непрерывных на $[0, 1]$ функций, а I_h — элементом пространства F_h сеточных функций f_h , определенных в точках $x_n = nh$, $n = 0, \dots, h^{-1}$, h^{-1} — целое. Тогда в качестве меры погрешности вводят некоторую меру близости $\rho(z_1, z_2)$, где z_1 и z_2 могут быть элементами как одного, так и различных подпространств. Требуемая на эту меру близости — возможность принять ее за естественную меру погрешности и выполнение неравенства треугольника

$$\rho(z_1, z_3) \leq \rho(z_1, z_2) + \rho(z_2, z_3) \quad (5)$$

при любых $z_1, z_2, z_3 \in F, F_h$. При этом не накладывается условие: если $\rho(z_1, z_2) = 0$, то $z_1 \equiv z_2$; так что функция $\rho(z_1, z_2)$ не обязательно является расстоянием в некотором метрическом пространстве.

В рассматриваемом сейчас примере можно положить

$$\rho(f_1, f_2) = \max_{0 \leq n \leq h^{-1}} |f_1(nh) - f_2(nh)|$$

независимо от того, каким пространствам принадлежат f_1 и f_2 .

Меры погрешности определяются, соответственно, равенствами

$$\rho_0 = \rho(\tilde{I}_h^*, I), \quad \rho_1 = \rho(\tilde{I}, I), \quad \rho_2 = \rho(\tilde{I}_h, \tilde{I}), \quad \rho_3 = \rho(\tilde{I}_h^*, \tilde{I}_h)$$

и, вследствие (5), будут удовлетворять неравенству (4).

Может возникнуть такой вопрос по поводу проблемы исследования неустранимой погрешности: зачем изучать неустранимую погрешность решения задачи, если она «неустранима»? По крайней мере такая точка зрения кажется оправданной, если математик получает для численного решения уже готовые уравнения, не участвуя в обсуждении физической постановки задачи.

Это возражение нельзя признать разумным. Часто математик сам занимается исследованием постановки задачи, анализом и упрощением рассматриваемых уравнений. Поскольку все явления в природе взаимосвязаны, в принципе невозможно описать математически точно никакой реальный процесс, происходящий в природе. Однако анализ роли различных факторов на погрешность решения может позволить получить простейшее описание процесса с допустимой погрешностью. Обычно математик имеет представление о требуемой окончательной точности результата, и, исходя из этого, он может производить необходимые упрощения исходной задачи.

Если математик не участвует в обсуждении физической постановки задачи, то представление о величине неустранимой погрешности ему все равно необходимо по следующей причине. При решении большинства задач из соображений «здравого смысла» видно, что нет особого резона применять метод решения задачи с погрешностью, существенно меньшей, чем величина неустранимой погрешности. Поэтому, имея представление о величине неустранимой погрешности, математик-исследователь

может понизить требования к точности применяемого им алгоритма.

В одном случае для решения была предложена следующая задача: для большого числа областей G и правых частей f составить таблицы решений уравнения Пуассона:

$$\Delta u = u_{x_1 x_1} + u_{x_2 x_2} = f(x_1, x_2) \quad \text{в области } G, \quad u|_{\Gamma} = 0;$$

Γ — граница G ; требуемая точность табличных значений 10^{-8} . В исходной постановке задача была непосильна для решения в приемлемое время, и поэтому был рассмотрен вопрос о разумности исходной постановки.

Оказалось, что реальная физическая задача описывается дифференциальным уравнением

$$(k(x_1, x_2) u_{x_1})_{x_1} + (k(x_1, x_2) u_{x_2})_{x_2} = f(x_1, x_2);$$

про коэффициент $k(x_1, x_2)$ известно лишь, что при существующей технологии изготовления рассматриваемого материала он находится в пределах от 0,7 до 1,5. Ясно, что численное решение исходной задачи с погрешностью, меньшей 0,01 (а может быть, и меньшей 0,1), является бессмысленным. После уменьшения требуемой точности оказалось возможным уменьшить также число различных областей и правых частей, пользуясь тем обстоятельством, что решения в близких между собой областях при близких правых частях близки. После такого изменения постановки задача стала посильной для решения.

Таким образом, исследование неустранимой погрешности исходной постановки задачи является часто удобным путем для снижения предъявляемых требований к точности.

Иногда в литературе встречается несколько иная классификация погрешностей: *неустранимой погрешностью* называют лишь погрешность, являющуюся следствием неточности задания исходных данных, входящих в математическое описание задачи, а погрешность, являющуюся следствием несоответствия математического описания реальности, называют *погрешностью математической модели*.

§ 2. Запись чисел в ЭВМ

В современных ЭВМ наиболее приняты две формы записи чисел.

Первая форма записи — *с фиксированной запятой*: все числа в ЭВМ имеют модуль, меньший 1; число знаков после запятой фиксировано. Таким образом, машина оперирует с числами

$$x = \pm \sum_{k=1}^l \alpha_k q^{-k} = \pm (\alpha_1, \dots, \alpha_l); \quad (1)$$

здесь q — целое — основание системы счисления, $\alpha_1, \dots, \alpha_l$ — целые в пределах $0 \leq \alpha_k < q$.

При операциях над числами x с $|x| < 1$ могут появляться числа y с $|y| \geq 1$, и тогда произойдет остановка работы ЭВМ («машинный останов» или «АВОСТ»). Чтобы избежать этого,

производится масштабирование задачи — введение новых масштабов. Иногда заранее нельзя указать нужные масштабы; в других случаях введение очень больших масштабов с самого начала приведет к тому, что в исходных данных большое число первых из α_i обратится в нуль и произойдет существенная потеря информации. Поэтому часто предусматривают изменение масштабов уже в процессе решения задачи.

Вторая форма записи — *с плавающей запятой*: машина оперирует с числами

$$x = \pm q^p \sum_{k=1}^t \alpha_k q^{-k} = \pm q^p (\alpha_1, \dots, \alpha_t); \quad (2)$$

порядок числа p удовлетворяет неравенству $|p| \leq p_0$.

Наиболее распространен случай двоичной системы счисления: $q = 2$; случаи $q = 8$ и $q = 10$ применяются как вспомогательные на этапах подготовки и выдачи данных. На небольшой части машин принята троичная система счисления: $q = 3$. В наиболее развитых ЭВМ имеются представления чисел как с фиксированной запятой, так и с плавающей; в ряде случаев число разрядов t может задаваться пользователем по его усмотрению.

При работе в режиме с плавающей запятой пользователь получает дополнительные удобства, так как может не заботиться о масштабах; однако при этом происходит некоторое замедление работы ЭВМ.

Наиболее типичны значения $q = 2$, $t = 35$, $p_0 = 64$; тогда $q^{-t} = 2^{-35} \approx 3 \cdot 10^{-11}$, $2^{p_0} = 2^{64} \approx 7 \cdot 10^{19}$.

§ 3. Абсолютная и относительная погрешности.

Формы записи данных

Если a — точное значение некоторой величины, а a^* — известное приближение к нему, то *абсолютной погрешностью* приближения a^* называют обычно некоторую величину $\Delta(a^*)$, про которую известно, что

$$|a^* - a| \leq \Delta(a^*).$$

Относительной погрешностью называют некоторую величину $\delta(a^*)$, про которую известно, что

$$\left| \frac{a^* - a}{a^*} \right| \leq \delta(a^*).$$

Относительную погрешность часто выражают в процентах.

По ходу изложения материала будут употребляться выражения: *большое число*, *очень большое число*, *сильный рост функции*.

Чаще всего число x мы называем *большим*, если $|x| \gg 1$, но относительная погрешность результата решения задачи порядка $|x|2^{-t}$ является допустимой;

если относительная погрешность порядка $|x|2^{-t}$ является недопустимо большой, число x называем *очень большим*;

выражение *функция сильно растет* чаще всего означает, что она возрастает в очень большое число раз.

Иногда в литературе абсолютной погрешностью называют величину $a^* - a$, а относительной — величину $\frac{a^* - a}{a^*}$; мы будем придерживаться исходных определений, и поэтому у нас всегда $0 \leq \Delta(a^*), \delta(a^*)$.

Значащими цифрами числа называют все цифры в его записи, начиная с первой ненулевой слева.

Пример. У чисел $a^* = 0,03045$, $a^* = 0,03045000$ значащими цифрами являются подчеркнутые цифры. Число *значащих цифр* в первом случае равно 4, во втором 7.

Значащую цифру называют *верной*, если абсолютная погрешность числа не превосходит единицы разряда, соответствующего этой цифре.

Примеры. $a^* = 0,03045$, $\Delta(a^*) = 0,000003$; $a^* = 0,03045000$, $\Delta(a^*) = 0,0000007$; подчеркнутые цифры являются верными.

Если все значащие цифры верные, то говорят, что число записано *со всеми верными цифрами*.

Пример. При $a^* = 0,03045$, $\Delta(a^*) = 0,000003$ число a^* записано со всеми верными цифрами.

Иногда употребляется термин *число верных цифр после запятой*: подсчитывается число цифр после запятой от первой цифры до последней верной цифры. В последнем примере это число равно 5.

Довольно часто информация о некоторой величине задается пределами ее изменения:

$$a_1 \leq a \leq a_2, \quad (1)$$

например,

$$1,119 \leq a \leq 1,127.$$

Принято записывать эти пределы с одинаковым числом знаков после запятой; так как обычно достаточно грубого представления о погрешности, то при записи пределов a_1, a_2 часто берут столько значащих десятичных цифр, сколько нужно, чтобы разность $a_2 - a_1$ содержала одну — три значащие цифры.

Многочисленные оговорки «часто», «обычно», «принято» специально употребляются нами, чтобы не создавать впечатления об обязательности каких-то стандартных форм задания информации. Эти формы задания информации рассматриваются лишь потому, что они наиболее распространены, а следовательно, наиболее удобны при контактах.

Абсолютную или относительную погрешность обычно записывают в виде числа, содержащего одну или две значащие цифры.

Информацию о том, что a^* является приближенным значением числа a с абсолютной погрешностью $\Delta(a^*)$, иногда удобно записывать в виде

$$a = a^* \pm \Delta(a^*); \quad (2)$$

числа a^* и $\Delta(a^*)$ принято записывать с одинаковым числом знаков после запятой. Например, записи

$$a = 1,123 \pm 0,004,$$

$$a = 1,123 \pm 4 \cdot 10^{-3}$$

относятся к общепринятым и означают, что

$$1,123 - 0,004 \leq a \leq 1,123 + 0,004.$$

Соответственно, информацию, что a^* является приближенным значением числа a с относительной погрешностью $\delta(a^*)$, записывают в виде

$$a = a^* (1 \pm \delta(a^*)). \quad (3)$$

Например, записи

$$a = 1,123 (1 \pm 0,003),$$

$$a = 1,123 (1 \pm 3 \cdot 10^{-3}),$$

$$a = 1,123 (1 \pm 0,3\%)$$

означают, что

$$(1 - 0,003) 1,123 \leq a \leq (1 + 0,003) 1,123.$$

При переходе от одной из форм записи к другой надо следить, чтобы пределы изменения, указываемые новой формой записи, были шире старых. Иначе такой переход будет незаконным. Например, при переходе от (1) к (2) должны выполняться неравенства

$$a^* - \Delta(a^*) \leq a_1, \quad a_2 \leq a^* + \Delta(a^*),$$

при переходе от (2) к (3) неравенства

$$a^* (1 - \delta(a^*)) \leq a^* - \Delta(a^*), \quad a^* + \Delta(a^*) \leq a^* (1 + \delta(a^*)),$$

при переходе от (3) к (2) должны выполняться противоположные неравенства (пределы всегда расширяются!).

§ 4. О вычислительной погрешности

Ограничение на порядки чисел в ЭВМ $|p| \leq p_0$ иногда приводит к прекращению вычислений; в других случаях относительно небольшая разрядность чисел в ЭВМ приводит к недопустимому искажению результата вычислительной

погрешностью. Такие алгоритмы, где, вследствие ограниченности p или малости t , возникают подобные эффекты, называют «неустойчивыми».

Построение «устойчивых» алгоритмов, при использовании которых вычисления доводятся до конца без существенного искажения результата, составляет существенную часть теории численных методов.

Ниже показано, что повышение точности иногда может быть достигнуто за счет несложного алгебраического преобразования.

Отыскивается наименьший корень уравнения $y^2 - 140y + 1 = 0$. Будем считать, что вычисления производятся в десятичной системе счисления, причем в мантиссе числа после округлений удерживается 4 разряда. Имеем

$$y = 70 - \sqrt{4899}, \quad \sqrt{4899} = 69,992 \dots;$$

после округления получаем

$$\sqrt{4899} \approx 69,99, \quad y \approx 70 - 69,99 = 0,01.$$

То же самое значение y можно, «избавившись от иррациональности в числителе», представить в виде

$$y = 1/(70 + \sqrt{4899}), \quad \sqrt{4899} \approx 69,99, \quad 70 + 69,99 = 139,99$$

и после округления

$$70 + 69,99 \approx 140,0.$$

Наконец

$$1/140 = 0,00714285 \dots,$$

и после округления

$$y \approx 0,007143.$$

Производя вычисления с дополнительными разрядами, можно проверить, что в обоих случаях все подчеркнутые цифры результатов верные; однако во втором случае точность результата существенно выше. Дело в том, что в первом случае пришлось вычитать близкие большие числа; так как эти числа были большие, то они были округлены с большой абсолютной погрешностью; в результате и ответ получился с большой абсолютной погрешностью. Здесь нам впервые встретилось явление потери значащих цифр (или «пропадания значащих цифр»), имеющее место при вычитании близких величин; это явление, например, довольно часто приводит к существенному искажению результата при решении линейных алгебраических систем уравнений.

§ 5. Погрешность функции

Довольно часто возникает следующая задача. Искомая величина y является функцией от параметров $a_1, \dots, a_n$: $y = y(a_1, \dots, a_n)$; известна область G в пространстве переменных $(a_1, \dots, a_n)$, которой принадлежат эти параметры. Требуется получить приближение к y и оценить его погрешность.

Если y^* — приближенное значение величины y , то *предельной абсолютной погрешностью* $A(y^*)$ называют наилучшую при имеющейся информации оценку погрешности величины y^* ; согласно этому определению, в данном случае

$$A(y^*) = \sup_{(a_1, \dots, a_n) \in G} |y(a_1, \dots, a_n) - y^*|; \quad (1)$$

предельной относительной погрешностью называют величину $\frac{A(y^*)}{|y^*|}$.

Задача. Доказать, что предельная абсолютная погрешность $A(y^*)$ минимальна при

$$y^* = \frac{Y_1 + Y_2}{2},$$

где

$$Y_1 = \inf_G y(a_1, \dots, a_n), \quad Y_2 = \sup_G y(a_1, \dots, a_n).$$

Рассмотрим наиболее распространенный случай, когда область G — прямоугольник:

$$|a_j - a_j^*| \leq \Delta(a_j^*), \quad j = 1, \dots, n,$$

и за приближенное значение принимается

$$y^* = y(a_1^*, \dots, a_n^*).$$

Если y — непрерывно дифференцируемая функция своих аргументов, то, согласно формуле Лагранжа,

$$y(a_1, \dots, a_n) - y^* = \sum_{j=1}^n b_j(\theta)(a_j - a_j^*), \quad (2)$$

где

$$b_j(\theta) = y_{a_j}(a_1^* + \theta(a_1 - a_1^*), \dots, a_n^* + \theta(a_n - a_n^*)), \quad 0 \leq \theta \leq 1.$$

Отсюда следует оценка погрешности

$$|y(a_1, \dots, a_n) - y^*| \leq A_0(y^*) = \sum_{j=1}^n B_j \Delta(a_j^*), \quad (3)$$

где

$$B_j = \sup_G |y_{a_j}(a_1, \dots, a_n)|.$$

Покажем, что при малых значениях

$$\rho = \sqrt{(\Delta(a_1^*))^2 + \dots + (\Delta(a_n^*))^2}$$

эта оценка не может быть существенно улучшена. Если производные y_{a_j} непрерывны в G , то

$$B_j = |b_j(0)| + o(1) \quad \text{при} \quad \rho \rightarrow 0.$$

Поэтому величина $A_0(y^*)$ может быть представлена в виде

$$\begin{aligned} A_0(y^*) &= A^0(y^*) + \varepsilon_1(\rho), \\ A^0(y^*) &= \sum_{j=1}^n |b_j(0)| \Delta(a_j^*), \quad \varepsilon_1(\rho) = o(\rho). \end{aligned} \quad (4)$$

При

$$a_j - a_j^* = \text{sign}(b_j(0)) \Delta(a_j^*)$$

имеем

$$\begin{aligned} b_j(0)(a_j - a_j^*) &= (b_j(0) + o(1)) \text{sign}(b_j(0)) \Delta(a_j^*) = \\ &= (|b_j(0)| + o(1)) \Delta(a_j^*), \end{aligned}$$

и тогда

$$y(a_1, \dots, a_n) - y^* = A^0(y^*) + o(\rho). \quad (5)$$

Отсюда и из определения предельной абсолютной погрешности (1) следует, что

$$A^0(y^*) + \varepsilon_2(\rho) \leq A(y^*), \quad \varepsilon_2(\rho) = o(\rho), \quad (6)$$

в то же время из (3) имеем

$$A(y^*) \leq A_0(y^*).$$

В совокупности получаем соотношение

$$A^0(y^*) + \varepsilon_2(\rho) \leq A(y^*) \leq A_0(y^*) = A^0(y^*) + \varepsilon_1(\rho). \quad (7)$$

Таким образом, оценка (3) не является существенно завышенной по сравнению с предельно точной оценкой (1). Согласно (7), величины $A(y^*)$ и $A_0(y^*)$ с большой точностью приближаются более просто вычисляемой величиной $A^0(y^*)$.

Очень часто вместо оценки погрешности (3) пользуются, вообще говоря, неверной «оценкой»

$$|y(a_1, \dots, a_n) - y^*| \leq A^0(y^*), \quad (8)$$

называемой *линейной оценкой погрешности*.

Рассмотрим некоторые примеры определения величин

$$A(y^*), \quad A_0(y^*) \quad \text{и} \quad A^0(y^*)$$

и произведем сравнение этих величин.

1. $y = a^{10}$, $a^* = 1$, $\Delta(a^*) = 0,001$. Тогда

$$y^* = 1, \quad y_a = 10 \cdot a^9, \quad b(0) = 10,$$

$$B = \sup_{|a-1| \leq 0,001} |10 \cdot a^9| = 10,09 \dots,$$

$$A(y^*) = \sup_{|a-1| \leq 0,001} |a^{10} - 1| = 1,001^{10} - 1 = 0,010045 \dots,$$

$$A_0(y^*) = B\Delta(a^*) = 0,01009 \dots,$$

$$A^0(y^*) = |b(0)|\Delta(a^*) = 0,01.$$

Здесь оценка погрешности через величину $A_0(y^*)$, предельно точная оценка (1) и «линейная оценка» (8) отличаются несущественно.

2. $y = a^{10}$, $a^* = 1$, $\Delta(a^*) = 0,1$. Тогда

$$y^* = 1, \quad B = \sup_{|a-1| \leq 0,1} 10 \cdot a^9 = 10 \cdot (1,1)^9 = 23, \dots,$$

$$A(y^*) = \sup_{|a-1| \leq 0,1} |a^{10} - 1| = (1,1)^{10} - 1 = 1,5 \dots,$$

$$A_0(y^*) = B\Delta(a^*) = 2,3 \dots, \quad A^0(y^*) = |b(0)|\Delta(a^*) = 1.$$

Здесь различие между этими оценками более заметно.

3. Произведем конкретную оценку в случае простейших функций. Пусть

$$y = \gamma_1 a_1 + \dots + \gamma_n a_n,$$

где $\gamma_1, \dots, \gamma_n$ принимают значения $+1$ или -1 ; пусть также известны оценки $|a_j - a_j^*| \leq \Delta(a_j^*)$. В данном конкретном случае $b_j(0) \equiv \gamma_j$, $|b_j(0)| \equiv 1$, поэтому в оценках (4), (6) отсутствуют члены $\epsilon_j(\rho)$ высшего порядка и из (7) следует, что

$$A(y^*) = A_0(y^*) = A^0(y^*) = \Delta(a_1^*) + \dots + \Delta(a_n^*).$$

Поскольку, по определению, погрешностью называют любую оценку для $y - y^*$, то это соотношение можно также записать в виде

$$\Delta(\pm a_1^* \pm \dots \pm a_n^*) = \Delta(a_1^*) + \dots + \Delta(a_n^*). \quad (9)$$

В случае $y = a_1 \pm a_2$ это равенство иногда формулируется в виде правила:

предельная погрешность суммы или разности равна сумме предельных погрешностей.

Если ошибки в величинах a_j^* зависимы, то оценка (9), вообще говоря, может быть улучшена. Простейший пример: $a_1 = a$, $a_2 = 1 - a$, где известно, что в обоих случаях a одно и то же;

тогда независимо от погрешности в значении a сумма $a_1 + a_2$ равна 1.

Пусть теперь $y = a_1^{p_1} \dots a_n^{p_n}$; тогда $b_j(0) = p_j(a_j^*)^{-1} y^*$ и

$$A(y^*) \approx A^0(y^*) = \sum_{j=1}^n |p_j| |a_j^*|^{-1} |y^*| \Delta(a_j^*).$$

После деления на $|y^*|$ получаем

$$\frac{A(y^*)}{|y^*|} \approx \frac{A^0(y^*)}{|y^*|} = \sum_{j=1}^n |p_j| \frac{\Delta(a_j^*)}{|a_j^*|} = \sum_{j=1}^n |p_j| \delta(a_j^*). \quad (10)$$

По отношению к частным случаям $y = a_1 \cdot a_2$ или $y = a_1 \cdot a_2^{-1}$ это соотношение иногда формулируют в виде правила:

предельная относительная погрешность произведения или частного приближенно равна сумме предельных относительных погрешностей.

4. Довольно часто возникает задача оценки погрешности функции, заданной уравнением

$$F(y, a_1, \dots, a_n) = 0. \quad (11)$$

Дифференцируя по a_j , имеем

$$\frac{\partial F}{\partial y} \frac{\partial y}{\partial a_j} + \frac{\partial F}{\partial a_j} = 0,$$

откуда

$$\frac{\partial y}{\partial a_j} = - \left(\frac{\partial F}{\partial a_j} \right) \left(\frac{\partial F}{\partial y} \right)^{-1} \quad (12)$$

При заданных $a_1^*, \dots, a_n^*$ можно найти y^* как корень (11), а затем значения

$$b_j(0) = - \left(\frac{\partial F}{\partial a_j} \right) \left(\frac{\partial F}{\partial y} \right)^{-1} \Big|_{(y^*, a_1^*, \dots, a_n^*)}. \quad (13)$$

С помощью этих величин можно получить «линейную оценку» погрешности (8).

Вследствие зависимости производных $\frac{\partial y}{\partial a_j}$ от самого значения y получение строгих оценок (1), (3) здесь довольно трудоемко.

5. Обратимся к оценке погрешности корней квадратного уравнения

$$F(y, a_1, a_2) = y^2 + a_1 y + a_2 = 0 \quad (14)$$

при заданных приближенных значениях коэффициентов a_1^*, a_2^* и их погрешностях $\Delta(a_1^*), \Delta(a_2^*)$.

Пусть y^* — решение уравнения

$$y^{*2} + a_1^* y^* + a_2^* = 0.$$

Из формулы (13) имеем

$$b_1(0) = \frac{\partial y}{\partial a_1} \Big|_{(a_1^*, a_2^*)} = - \frac{y^*}{2y^* + a_1^*},$$

$$b_2(0) = \frac{\partial y}{\partial a_2} \Big|_{(a_1^*, a_2^*)} = - \frac{1}{2y^* + a_1^*}$$

и, следовательно,

$$A^0(y^*) = \frac{|y^*| \Delta(a_1^*) + \Delta(a_2^*)}{|2y^* + a_1^*|}. \quad (15)$$

Рассмотрим некоторую область $|a_1| \leq b_1$, $|a_2| \leq b_2$ изменения коэффициентов a_1 , a_2 . Из явного выражения корней

$$y = -\frac{a_1}{2} \pm \sqrt{\frac{a_1^2}{4} - a_2}$$

следует, что корни являются непрерывными функциями коэффициентов, и поэтому

$$|y(a_1, a_2) - y(a_1^*, a_2^*)| \leq \omega(|a_1 - a_1^*|, |a_2 - a_2^*|) \quad (16)$$

при (a_1, a_2) , (a_1^*, a_2^*) из этой области, $\omega(\lambda_1, \lambda_2) \rightarrow 0$ при $\lambda_1, \lambda_2 \rightarrow 0$. Правая часть (15) стремится к ∞ при $2y^* + a_1^* \rightarrow 0$; поэтому из сравнения (15) и (16) вытекает, что «линейная оценка» погрешности при помощи формулы (8) может оказаться в некоторых случаях сильно завышенной по сравнению с точной оценкой погрешности (3). Дело в том, что ранее рассуждения проводились в предположении непрерывной дифференцируемости $y(a_1, \dots, a_n)$ по аргументам $a_1, \dots, a_n$. Тогда погрешность y^* оказывалась величиной того же порядка, что и погрешности аргументов $\Delta(a_1^*)$. В случае, когда величина y^* определяется неявным образом, при некоторых значениях параметров она оказывается недифференцируемой функцией своих аргументов и характер оценки меняется.

Пусть y^* является двукратным корнем (11) при $a_1 = a_1^*$, $a_2 = a_2^*$. Разложим левую часть (11) в ряд Тейлора в окрестности точки y^* , a_1^* , a_2^* . Поскольку

$$F(y^*, a_1^*, a_2^*) = F_y(y^*, a_1^*, a_2^*) = 0$$

при y^* — двукратном корне (11), то уравнение (11) примет вид

$$d_{200}(y - y^*)^2 + d_{010}(a_1 - a_1^*) + d_{001}(a_2 - a_2^*) + \dots = 0,$$

где

$$d_{i,jk} = \frac{F_{y^i a^j a_2^k}(y^*, a_1^*, a_2^*)}{i!j!k!},$$

а отброшенные члены — порядка $o(\rho)$. В рассматриваемом примере (14) $d_{200} = 1$ и тогда

$$y - y^* = \pm \sqrt{d_{010}(a_1 - a_1^*) + d_{001}(a_2 - a_2^*) + o(\rho)}.$$

Таким образом, погрешность корня оказалась величиной порядка $O(\sqrt{\rho})$. Соответственно погрешность корня кратности l обычно имеет порядок $O(\sqrt[l]{\rho})$.

6. При оценке погрешностей корней алгебраических уравнений на первый взгляд могла бы показаться актуальной и естественной следующая проблема. Рассматривается уравнение

$$y^n + \sum_{j=1}^n a_j^* y^{n-j} = 0. \quad (17)$$

Задана некоторая область изменения значений a_j^* коэффициентов уравнения и оценка $\Delta(a_j^*) \leq \gamma$ их погрешностей. Оценить верхнюю границу ω погрешностей корней y этого уравнения.

З а д а ч а. Получить такую оценку в случае уравнения (17) при $|a_1^*|, |a_2^*| \leq 2$, $\Delta(a_1^*), \Delta(a_2^*) \leq \gamma$.

Если область изменения значений a_j^* такова, что при $\{a_j^*\}$ из этой области уравнение может иметь l -кратный корень, то нельзя рассчитывать на оценку погрешности, по порядку лучшую $O(\sqrt[l]{\gamma})$. Однако, за исключением множества значений коэффициентов меры нуля, корень уравнения является простым и можно рассчитывать на оценку погрешности $O(\gamma)$. Таким образом, оценка погрешности через величину γ для большинства значений $a_1^*, \dots, a_n^*$ оказывалась бы существенно завышенной и пользы от этой оценки было бы мало.

7. Рассмотрим еще одну типичную задачу, встречающуюся в приложениях и близкую по способу решения к рассматриваемым выше.

Имеется приближение y^* к корню уравнения

$$f(y) = 0;$$

требуется оценить его погрешность.

Вычислим величину

$$a^* = f(y^*).$$

При малых $y^* - y$ из равенства

$$f(y) - f(y^*) = a - a^*$$

следует, что

$$f'(y^*)(y - y^*) \approx a - a^*,$$

и таким образом,

$$y - y^* \approx \frac{a - a^*}{f'(y^*)} = \frac{a - f(y^*)}{f'(y^*)}.$$

В наиболее распространенном случае $a = 0$ име

$$y - y^* \approx - \frac{f(y^*)}{f'(y^*)}.$$

ИНТЕРПОЛЯЦИЯ И СМЕЖНЫЕ ВОПРОСЫ

Задача приближения функций возникает при решении многих задач, а иногда и как самостоятельная. Настоящая глава посвящена частному, но довольно распространенному способу приближения функций путем интерполяции их значений; интерполяция является также важным вспомогательным аппаратом при решении других задач численного анализа: численного интегрирования и дифференцирования, решения дифференциальных уравнений и др. Прежде чем переходить непосредственно к интерполяции, напомним некоторые определения и в § 1 обсудим различные постановки задачи приближения функций.

Множество M называется *линейным*, если в нем определены операции сложения и умножения на числа (действительные или комплексные), не выводящие за пределы M и удовлетворяющие условиям:

- 1) сложение ассоциативно: $(x + y) + z = x + (y + z)$,
- 2) коммутативно $x + y = y + x$,
- 3) существует нулевой элемент 0 , для которого $x + 0 = x$ при любом $x \in M$,
- 4) $0 \cdot x = 0$ при любом $x \in M$.
- 5) $(\alpha + \beta)x = \alpha x + \beta x$,
- 6) $\alpha(x + y) = \alpha x + \alpha y$,
- 7) $\alpha(\beta x) = (\alpha\beta)x$,
- 8) $1 \cdot x = x$.

В линейном множестве можно ввести понятие линейной зависимости и линейной независимости элементов. Система элементов $x_1, \dots, x_n$ линейного множества M называется *линейно зависимой*, если существуют $c_1, \dots, c_n$, не равные одновременно нулю, такие, что

$$c_1 x_1 + \dots + c_n x_n = 0.$$

В противном случае систему называют *линейно независимой*.

Линейным подпространством называется подмножество N линейного множества, для которого из условия $x, y \in N$ следует $\alpha x + \beta y \in N$ при любых α и β .

Пространство R называется *метрическим*, если для любых двух элементов определено расстояние $\rho(x, y)$, удовлетворяющее условиям:

- 1) $\rho(x, y) \geq 0$, причем $\rho(x, y) = 0$ тогда и только тогда, когда $x = y$,
- 2) $\rho(x, y) = \rho(y, x)$.
- 3) $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$ для любых $x, y, z \in R$.

Множество R называют *линейным нормированным пространством*, если а) оно линейно и б) каждому элементу $f \in R$ поставлено в соответствие действительное число $\|f\|$, называемое *нормой* f и удовлетворяющее условиям:

- 1) $\|f\| \geq 0$, причем $\|f\| = 0$ тогда и только тогда, когда $f = 0$,
- 2) $\|\alpha f\| = |\alpha| \|f\|$ для любого комплексного α ,
- 3) $\|f_1 + f_2\| \leq \|f_1\| + \|f_2\|$.

Очевидно, линейное нормированное пространство является одновременно метрическим с расстоянием

$$\rho(f_1, f_2) = \|f_1 - f_2\|.$$

Линейное нормированное пространство называется *строго нормированным*, если равенство

$$\|f_1 + f_2\| = \|f_1\| + \|f_2\|$$

возможно тогда и только тогда, когда $f_2 = \alpha f_1$, $\alpha \geq 0$.

Говорят, что в линейном множестве R определено скалярное произведение, если каждой упорядоченной паре элементов $f_1, f_2 \in R$ поставлено в соответствие некоторое комплексное число (f_1, f_2) и при этом выполняются соотношения:

1. $(f_1, f_2) = \overline{(f_2, f_1)}$.
2. Для любых $f_1, f_2, f_3 \in R$ и комплексных α_1, α_2 имеет место равенство

$$(\alpha_1 f_1 + \alpha_2 f_2, f_3) = \alpha_1 (f_1, f_3) + \alpha_2 (f_2, f_3).$$

3. $(f, f) \geq 0$ и $(f, f) = 0$ только при $f = 0$.

Из этих свойств скалярного произведения вытекает ряд других его свойств:

4. $(f_3, \alpha_1 f_1 + \alpha_2 f_2) = \bar{\alpha}_1 (f_3, f_1) + \bar{\alpha}_2 (f_3, f_2)$.
- Для любых $f_1, f_2 \in R$ имеют место неравенства:
5. $|(f_1, f_2)|^2 \leq (f_1, f_1)(f_2, f_2)$.
6. $\|f_1 + f_2\| \leq \|f_1\| + \|f_2\|$, где $\|f\| = \sqrt{(f, f)}$.
7. Знак равенства в п. 6 имеет место лишь при $f_2 = \alpha f_1$, $\alpha \geq 0$.

Свойство 6 означает, что линейное множество R со скалярным произведением является линейным нормированным и, следовательно, метрическим пространством с

$$\rho(f_1, f_2) = \|f_1 - f_2\| = \sqrt{(f_1 - f_2, f_1 - f_2)}.$$

Свойство 7 означает, что оно является строго нормированным

§ 1. Постановка задачи приближения функций

Задача приближения функции возникает в различных ситуациях, часть из которых будет рассмотрена далее. Многообразие методов, предлагаемых для ее решения, столь велико, что иногда возникает следующий вопрос. Может быть, наличие большого количества различных способов приближения объясняется просто отсутствием научного подхода к постановке и решению проблемы; если бы такой подход был, то, может быть, удалось бы предложить один оптимальный способ приближения, пригодный во всех случаях? Такой вопрос возникает и при рассмотрении других разделов численного анализа. Сколь бы ни было заманчиво разработать единый подход к решению всех задач,

следует все-таки признать, что многообразие методов вызывается существом дела — многообразием различных постановок проблемы. Различные разделы теории приближений, в частности интерполяции, можно рассматривать как изучение абстрактных моделей некоторых реальных классов проблем.

1. Простейшая задача, приводящая к приближению функций, заключается в следующем. В дискретные моменты времени $x_1, \dots, x_n$ наблюдаются значения функции $f(x)$; требуется восстановить ее значения при других x . Аналогичная задача возникает в следующей обстановке. По ходу вычислений на ЭВМ приходится многократно вычислять одну и ту же сложную функцию в различных точках. Вместо ее непосредственного вычисления иногда целесообразнее вычислить ее значения в отдельных точках, а в других точках вычислять, используя информацию об этих известных значениях.

Иногда из каких-то дополнительных соображений известно, что приближающую функцию целесообразно искать в виде

$$f(x) \approx g(x; a_1, \dots, a_n).$$

Если параметры $a_1, \dots, a_n$ определяются из условия совпадения $f(x)$ и приближающей функции в точках $x_1, \dots, x_n$:

$$g(x_j; a_1, \dots, a_n) = f(x_j), \quad j = 1, \dots, n,$$

то такой способ приближения называют *интерполяцией* или *интерполированием*.

2. Часто бывает известно, что функция хорошо приближается функциями определенного вида, например многочленами, но неясно, как лучше всего выбрать степень приближающего многочлена. Проблема особенно усложняется, когда заданные значения функции содержат большие ошибки.

3. В задачах планирования экспериментов возникает следующая проблема. Известен вид хорошего приближения функции, например, функция хорошо приближается многочленом второй степени. В то же время измеряемые значения функции содержат большие ошибки. Требуется получить наилучшее в определенной норме приближение при минимальном числе измеряемых значений функции. Такая задача возникает при планировании экспериментов в биологии, химии, физике, географии, военном деле.

Коснемся подробнее планирования экспериментов на химических предприятиях. Наиболее существенный выходной параметр продукции, например, сортность капрона, определяется целым рядом параметров, часть которых зависит от свойств поступающего сырья, а часть поддается управлению; требуется максимально повысить выпуск капрона первого сорта. Для этого нужно иметь в распоряжении функцию, отражающую зависимость выхода продукции от всех этих параметров. Рассмотрим случай, когда эта функция

является многочленом второй степени от 10 параметров. Такой многочлен содержит 66 коэффициентов, поэтому для их отыскания требуется по крайней мере 66 измерений. Каждое измерение содержит существенную случайную ошибку из-за наличия ряда неконтролируемых и неизмеряемых факторов; поэтому для понижения влияния случайности необходимо увеличение числа измерений; в то же время каждое измерение обходится относительно дорого, так как требует изменения режима работы реального предприятия; к тому же такое изменение часто требует длительного промежутка времени, поскольку предприятие должно выйти на стационарный режим работы, соответствующий задаваемым значениям параметров. В результате анализа всех этих противоречий и возникла та проблема приближения функций, о которой шла речь вначале.

4. Задача приближения появляется при составлении стандартных программ вычисления элементарных и специальных функций. Обычно такие функции обладают специальными свойствами, позволяющими резко уменьшить объем вычислений.

Возникающая здесь проблема может быть сформулирована следующим образом. Рассматриваются все функции $g(x)$, программа вычисления которых записывается в k ячейках ЭВМ, такие, что некоторая норма погрешности $\|f - g\|$ не превосходит ϵ . Среди всех таких функций нужно выбрать ту, вычисление которой требует минимальных затрат времени ЭВМ.

В зависимости от обстановки норма может выбираться по-разному. В большинстве случаев берется $\|f\| = \sup_{[a, b]} |f|$, $[a, b]$ — отрезок, на котором приближается функция.

Довольно часто требуется повышенная точность в отдельных точках. Например, одна из стандартных программ $\sin x$ обеспечивает малость погрешности в норме

$$\|f\| = \sup_{\left[0, \frac{\pi}{2}\right]} |\rho(x)f(x)|, \quad \rho(x) = \min(10^{10}, x^{-1}).$$

Введение множителя $\rho(x)$ вызывается требованием малости относительной погрешности $\sin x$ при малых x .

Точно так же может толковаться по-разному требование минимальности затрат времени ЭВМ. Затраты, вообще говоря, могут зависеть от точки, в которой вычисляется значение функции. Обозначим их через $t(x)$. Если не имеется информации о том, с какой частотой вычисляются значения функции в тех или иных частях отрезка, то целесообразно в качестве общей меры затрат принять

$$T = \sup_x t(x).$$

Если такая информация есть, то можно принять

$$T = Mt(x).$$

5. Если функция задается графиком или сложным аналитическим выражением, то возникают вариационные задачи других

типов. Например, пусть решено разбить отрезок $[a, b]$ на l частей:

$$[a_{i-1}, a_i], \quad i = 1, \dots, l, \quad a_0 = a, \quad a_l = b,$$

и на каждом отрезке $[a_{i-1}, a_i]$ приближать функцию многочленом

$$f(x) \approx g(x) = \sum_{k=0}^{n_i} a_k^i x^k$$

или дробно-рациональной функцией

$$f(x) \approx g(x) = \frac{\sum_{k=0}^{n_i} a_k^i x^k}{\sum_{k=0}^{m_i} b_k^i x^k}.$$

Среди таких способов приближения отыскивается оптимальный в том или ином смысле. Чаще всего заранее накладывается требование $n_1 = \dots = n_l$, $m_1 = \dots = m_l$, фиксируется число отрезков разбиения l и производится оптимизация метода по $a_1, \dots, a_{l-1}$, n и m .

В частном случае $l = 1$ возникает задача наилучших равномерных приближений многочленами или дробно-рациональными функциями на фиксированном отрезке. Об этой задаче пойдет речь в главе IV.

6. Вид приближающей функции существенно зависит от цели, с которой осуществляется приближение. Предположим, что с требуемой точностью функция может быть приближена многочленом десятой степени или выражением $a_1 \sin \omega_1 x + a_2 \sin \omega_2 x$. Если полученное приближение используется в теоретических исследованиях, для решения задачи на моделирующем устройстве или в технологическом процессе, то вторая форма записи может быть более удобной. Однако если значения функции вычисляются на ЭВМ, то вторая форма записи потребует при своей реализации большего числа арифметических операций.

Далее будет конкретно рассмотрена задача интерполирования. Ее выделение вызвано наличием непосредственных многочисленных приложений, а также и следующим обстоятельством. Аппарат интерполирования многочленами является важнейшим аппаратом численного анализа; на его основе строится большинство методов решения других задач; его роль в численном анализе аналогична роли разложения Тейлора в классическом анализе.

Попутно будут рассмотрены некоторые другие вопросы общего характера, имеющие значение для других разделов численного анализа.

§ 2. Интерполяционный многочлен Лагранжа

Среди способов интерполирования наиболее распространен случай линейного интерполирования, когда приближение ищется в виде

$$g(x; a_1, \dots, a_n) = \sum_{i=1}^n a_i \varphi_i(x),$$

$\varphi_i(x)$ — фиксированные функции, значения коэффициентов a_i определяются из условия совпадения с приближаемой функцией в узлах интерполяции x_j :

$$f(x_j) = \sum_{i=1}^n a_i \varphi_i(x_j), \quad j = 1, \dots, n. \quad (1)$$

Способ решения задачи, при котором коэффициенты a_i определяются непосредственным решением системы (1), называется *способом неопределенных коэффициентов*.

Наиболее изучен случай интерполирования многочленами

$$\sum_{i=1}^n a_i x^{i-1}. \quad (2)$$

Тогда

$$\varphi_i(x) = x^{i-1}, \quad i = 1, \dots, n,$$

и система уравнений (1) имеет вид

$$\sum_{i=1}^n a_i x_j^{i-1} = f(x_j), \quad j = 1, \dots, n. \quad (3)$$

Мы неявно предполагаем, что все x_j различные. Определитель этой системы $\det[x_j^{i-1}]$ отличен от 0 (определитель Вандермонда). Следовательно, система (3) всегда имеет решение и притом единственное. Таким образом, доказано существование и единственность интерполяционного многочлена вида (2).

Нам удастся получить ряд явных представлений интерполяционного многочлена (2), не прибегая к непосредственному решению системы (3). Однако следует сразу же отметить, что в других случаях получение интерполирующей функции в явном виде обычно крайне затруднительно, и приходится прибегать к непосредственному решению системы (1).

Пусть $\delta_{q_1, \dots, q_k}$ есть символ

$$\delta_{q_1, \dots, q_k} = \begin{cases} 1 & \text{при } q_1 = \dots = q_k = 0, \\ 0 & \text{для остальных совокупностей значений } q_1, \dots, q_k. \end{cases}$$

Задача интерполирования будет решена, если удастся построить многочлены $\Phi_i(x)$ степени $n-1$ такие, что $\Phi_i(x_j) = \delta_{j-i}$

при $i, j = 1, \dots, n$. Многочлен

$$g_n(x) = \sum_{j=1}^n f(x_j) \Phi_j(x)$$

будет искомым интерполяционным многочленом; в самом деле,

$$g_n(x_i) = \sum_{j=1}^n f(x_j) \Phi_j(x_i) = \sum_{j=1}^n f(x_j) \delta_{i-j} = f(x_i);$$

кроме того, $g_n(x)$ — многочлен степени $n-1$.

Поскольку $\Phi_i(x_j) = 0$ при $j \neq i$, то $\Phi_i(x)$ делится на $x - x_j$ при $j \neq i$; таким образом, нам известно $n-1$ делителей многочлена степени $n-1$; отсюда

$$\Phi_i(x) = \text{const} \cdot \prod_{j \neq i} (x - x_j).$$

Из условия $\Phi_i(x_i) = 1$ получаем

$$\Phi_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

Интерполяционный многочлен (2), записанный в форме

$$g_n(x) = L_n(x) = \sum_{i=1}^n f(x_i) \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}, \quad (4)$$

называют *интерполяционным многочленом Лагранжа*.

Введем в рассмотрение функцию

$$\omega_n(x) = \prod_{j=1}^n (x - x_j).$$

Очевидно,

$$\omega'_n(x) = \sum_{k=1}^n \left[\prod_{j \neq k} (x - x_j) \right];$$

при $x = x_i$, $k \neq i$ выражение в квадратных скобках обращается в нуль, поскольку содержит множитель $(x_i - x_i)$. Таким образом,

$$\omega'_n(x_i) = \prod_{j \neq i} (x_i - x_j).$$

Поэтому коэффициент $\prod_{j \neq i} \frac{x - x_j}{x_i - x_j}$ можно записать в виде

$$\frac{\omega_n(x)}{\omega'_n(x_i)(x - x_i)},$$

а многочлен Лагранжа в виде

$$L_n(x) = \sum_{i=1}^n \frac{f(x_i) \omega_n(x)}{\omega'_n(x_i)(x - x_i)}, \quad (5)$$

§ 3. Оценка остаточного члена интерполяционного многочлена Лагранжа

В предположении непрерывности $f^{(n)}(x)$ оценим разность между $f(x)$ и построенным интерполяционным многочленом $g_n(x)$. Положим

$$\varphi(z) = f(z) - g_n(z) - K\omega_n(z),$$

выберем K из условия $\varphi(x) = 0$, где x — точка, в которой оценивается погрешность. Очевидно,

$$K = \frac{f(x) - g_n(x)}{\omega_n(x)}.$$

При таком выборе K функция $\varphi(z)$ обращается в нуль в $(n+1)$ -й точке $x_1, \dots, x_n, x$. На основании теоремы Ролля ее производная $\varphi'(z)$ обращается в нуль по крайней мере в n точках; применяя теорему Ролля к $\varphi'(z)$, получаем, что ее производная $\varphi''(z)$ обращается в нуль по крайней мере в $(n-1)$ -й точке; продолжая эти рассуждения дальше, имеем: $\varphi^{(n)}(z)$ обращается в нуль по крайней мере в одной точке ξ , причем эта точка принадлежит отрезку $[y_1, y_2]$:

$$y_1 = \min(x_1, \dots, x_n, x), \quad y_2 = \max(x_1, \dots, x_n, x).$$

Поскольку

$$\varphi^{(n)}(z) = f^{(n)}(z) - Kn!,$$

из условия $\varphi^{(n)}(\xi) = 0$ следует

$$K = \frac{f^{(n)}(\xi)}{n!}.$$

Соотношение $\varphi(x) = 0$ можно переписать в виде

$$f(x) - g_n(x) = \frac{f^{(n)}(\xi)\omega_n(x)}{n!}, \quad \xi \in [y_1, y_2], \quad (1)$$

который и дает представление остаточного члена.

§ 4. Разделенные разности и их свойства

Интерполяционный многочлен можно рассматривать как обобщение отрезка ряда Тейлора.

Обобщением понятия производной является понятие *разделенной разности*. Разделенные разности нулевого порядка $f(x_i)$ совпадают со значениями функции $f(x_i)$; разности первого порядка определяются равенством

$$f(x_i; x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i}, \quad (1)$$

разности второго порядка — равенством

$$f(x_i; x_j; x_k) = \frac{f(x_j; x_k) - f(x_i; x_j)}{x_k - x_i}$$

и, вообще, разности k -го порядка $f(x_1; \dots; x_{k+1})$ определяются через разности $(k-1)$ -го порядка по формуле

$$f(x_1; \dots; x_{k+1}) = \frac{f(x_2; \dots; x_{k+1}) - f(x_1; \dots; x_k)}{x_{k+1} - x_1}. \quad (2)$$

Иногда вместо $f(x_1; \dots; x_k)$ используют обозначения $(f)(x_1; \dots; x_k)$ или $[x_1; \dots; x_k]$.

Лемма. *Справедливо равенство*

$$f(x_1; \dots; x_k) = \sum_{j=1}^k \frac{f(x_j)}{\prod_{i \neq j} (x_j - x_i)}. \quad (3)$$

Доказательство будем проводить по индукции; при $k=1$ это равенство превращается в равенство $f(x_1) = f(x_1)$, при $k=2$ совпадает с (1). Пусть (3) доказано при $k \leq l$. Тогда

$$\begin{aligned} f(x_1; \dots; x_{l+1}) &= \frac{f(x_2; \dots; x_{l+1}) - f(x_1; \dots; x_l)}{x_{l+1} - x_1} = \\ &= \frac{1}{x_{l+1} - x_1} \left(\sum_{j=2}^{l+1} \frac{f(x_j)}{\prod_{\substack{i \neq j \\ 2 \leq i \leq l+1}} (x_j - x_i)} - \sum_{j=1}^l \frac{f(x_j)}{\prod_{\substack{i \neq j \\ 1 \leq i \leq l}} (x_j - x_i)} \right) \end{aligned}$$

Для $j \neq 1, l+1$ коэффициент при $f(x_j)$ в правой части есть

$$\begin{aligned} &\frac{1}{\prod_{\substack{i \neq j \\ 2 \leq i \leq l+1}} (x_j - x_i)} - \frac{1}{\prod_{\substack{i \neq j \\ 1 \leq i \leq l}} (x_j - x_i)} = \\ &= \frac{x_j - x_1}{x_{l+1} - x_1} - \frac{x_j - x_{l+1}}{x_{l+1} - x_1} = \frac{(x_j - x_1) - (x_j - x_{l+1})}{x_{l+1} - x_1} = \frac{1}{\prod_{\substack{i \neq j \\ 1 \leq i \leq l+1}} (x_j - x_i)}. \end{aligned}$$

т. е. имеет требуемый вид; для $j=1$ или $j=l+1$ $f(x_j)$ входит только в одно слагаемое в правой части и коэффициент при нем также имеет требуемый вид. Доказательство закончено.

Непосредственно из (3) вытекает ряд следствий:

1. Разделенная разность является линейным оператором от функции f :

$$(\alpha_1 f_1 + \alpha_2 f_2)(x_1; \dots; x_k) = \alpha_1 f_1(x_1; \dots; x_k) + \alpha_2 f_2(x_1; \dots; x_k).$$

2. Разделенная разность есть симметрическая функция своих аргументов $x_1, \dots, x_k$ (т. е. не меняется при любой их перестановке).

Если функция задана в точках $x_1, \dots, x_n$, то таблицу

$$\begin{array}{ccccccc}
 f(x_1) & & & & & & \\
 f(x_2) & f(x_1, x_2) & & f(x_1, x_2, x_3) & \dots & & \\
 f(x_3) & f(x_2, x_3) & & \vdots & & & f(x_1, \dots, x_n) \\
 \vdots & \vdots & & \vdots & & & \\
 f(x_n) & f(x_{n-1}, x_n) & \dots & & & &
 \end{array}$$

называют *таблицей ее разделенных разностей*.

§ 5. Интерполяционная формула Ньютона с разделенными разностями

При помощи разделенных разностей можно получить другую форму записи интерполяционного многочлена (2.2).

Справедливо равенство

$$\begin{aligned}
 f(x) - L_n(x) &= f(x) - \sum_{i=1}^n f(x_i) \prod_{j \neq i} \frac{x - x_j}{x_i - x_j} = \\
 &= \prod_{i=1}^n (x - x_i) \left(\frac{f(x)}{\prod_{i=1}^n (x - x_i)} + \sum_{i=1}^n \frac{f(x_i)}{(x_i - x) \prod_{j \neq i} (x_i - x_j)} \right).
 \end{aligned}$$

Сравнивая с (4.3), убеждаемся, что выражение в скобках есть $f(x; x_1; \dots; x_n)$. Таким образом, можно написать

$$f(x) - L_n(x) = f(x; x_1; \dots; x_n) \omega_n(x). \quad (1)$$

Пусть $L_m(x)$ — интерполяционный многочлен Лагранжа с узлами интерполяции $x_1, \dots, x_m$. Интерполяционный многочлен Лагранжа $L_n(x)$ можно представить в виде

$$L_n(x) = L_1(x) + (L_2(x) - L_1(x)) + \dots + (L_n(x) - L_{n-1}(x)). \quad (2)$$

Разность $L_m(x) - L_{m-1}(x)$ есть многочлен степени $m-1$, обращающийся в нуль в точках $x_1, \dots, x_{m-1}$, поскольку $L_{m-1}(x_j) = L_m(x_j) = f(x_j)$ при $1 \leq j \leq m-1$; следовательно,

$$L_m(x) - L_{m-1}(x) = A_{m-1} \omega_{m-1}(x), \quad \omega_{m-1}(x) = (x - x_1) \dots (x - x_{m-1}).$$

Пологая $x = x_m$, получим

$$f(x_m) - L_{m-1}(x_m) = A_{m-1} \omega_{m-1}(x_m).$$

Эта схема часто применяется при решении следующей задачи:

Дана таблица значений некоторой функции $f(x)$; требуется при каждом значении x наилучшим образом определить значение $f(x)$ путем интерполяции значений этой таблицы.

Пусть x фиксировано; перенумеруем узлы интерполяции в порядке возрастания $|x_i - x|$. Интерполяционные многочлены $L_{(1, \dots, m)}(x)$ будем обозначать, как обычно, $L_m(x)$.

Выше получено представление погрешности (1)

$$f(x) - L_m(x) = f(x; x_1; \dots; x_m) \omega_m(x),$$

а также равенство

$$L_{m+1}(x) - L_m(x) = f(x_1; \dots; x_{m+1}) \omega_m(x). \quad (6)$$

Так как при малых $|x - x_k|$

$$f(x; x_1; \dots; x_m) \approx \frac{f^m(x)}{m!} \approx f(x_1; \dots; x_{m+1}),$$

то отсюда следует, что

$$L_{m+1}(x) - L_m(x) \approx f(x) - L_m(x). \quad (7)$$

Поэтому величину $\varepsilon_m = |L_{m+1}(x) - L_m(x)|$ можно рассматривать как приближенную оценку погрешности интерполяционной формулы $f(x) \approx L_m(x)$. Последовательно вычисляют значения $L_0(x)$, $L_1(x)$, ε_1 , $L_2(x)$, ε_2 , ...; если при некотором m $\varepsilon_m \leq E$, где E — удовлетворяющая нас точность, то вычисления прекращают и полагают

$$f(x) \approx L_m(x).$$

Если это неравенство не выполняется ни при каком m , то находят $\varepsilon_{m_0} = \min_m \varepsilon_m$ и полагают $f(x) \approx L_{m_0}(x)$;

если этот минимум достигается при нескольких m , то среди них выбирают наименьшее. Часто величины ε_m , начиная с некоторого m , имеют устойчивую тенденцию к увеличению, и поэтому с этого момента вычисление значений $L_m(x)$, ε_m прекращают.

Заметим, что такой выбор m да и сам способ упорядочивания точек x_i по их близости к x не являются оптимальными.

Рассмотрим функцию $f(x) = \frac{1}{x-h}$; h — некоторое комплексное число.

Задача. Доказать, что для такой функции

$$f(x) - L_n(x) = \frac{1}{x-h} \left(\frac{x-x_1}{h-x_1} \right) \dots \left(\frac{x-x_n}{h-x_n} \right).$$

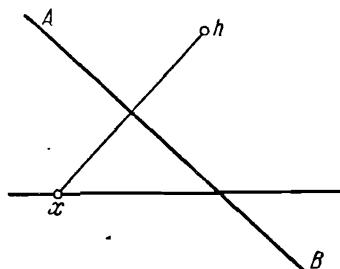


Рис. 2.5.1.

$\varepsilon \rightarrow 0$. В частности, можно положить

$$x_{lj}^\varepsilon = x_l + (j-1)\varepsilon.$$

Построим интерполяционный многочлен $g_s^\varepsilon(x)$ степени $s-1$, совпадающий с $f(x)$ в точках x_{lj}^ε . Таблица разделенных разностей, соответствующая этому набору узлов, имеет вид

$$\begin{array}{rcl} f(x_{11}^\varepsilon) & & \\ f(x_{12}^\varepsilon) & f(x_{11}^\varepsilon; x_{12}^\varepsilon) & \\ f(x_{13}^\varepsilon) & f(x_{12}^\varepsilon; x_{13}^\varepsilon) & f(x_{11}^\varepsilon, x_{12}^\varepsilon, x_{13}^\varepsilon) \\ \vdots & & \\ f(x_{1m_1}^\varepsilon) & & \\ f(x_{21}^\varepsilon) & f(x_{1m_1}^\varepsilon; x_{21}^\varepsilon) & \\ \vdots & & \\ f(x_{nm_n}^\varepsilon) & & f(x_{11}^\varepsilon, x_{12}^\varepsilon, \dots, x_{nm_n}^\varepsilon) \end{array} \quad (2)$$

Напишем интерполяционную формулу Ньютона с разделенными разностями:

$$g_s^\varepsilon(x) = A_0^\varepsilon + A_1^\varepsilon(x - x_{11}^\varepsilon) + \\ + A_2^\varepsilon(x - x_{11}^\varepsilon)(x - x_{12}^\varepsilon) + \dots + A_{s-1}^\varepsilon(x - x_{11}^\varepsilon) \dots (x - x_{nm_n-1}^\varepsilon)$$

где

$$A_0^\varepsilon = f(x_{11}^\varepsilon), \quad A_1^\varepsilon = f(x_{11}^\varepsilon; x_{12}^\varepsilon), \\ A_2^\varepsilon = f(x_{11}^\varepsilon; x_{12}^\varepsilon; x_{13}^\varepsilon), \dots, \quad A_{s-1}^\varepsilon = f(x_{11}^\varepsilon; x_{12}^\varepsilon; \dots; x_{nm_n}^\varepsilon).$$

Выражая разделенные разности через производные, имеем

$$f(x_{11}^\varepsilon; \dots; x_{im}^\varepsilon) = \frac{f^{(m-l)}(\bar{x}_{11m}^\varepsilon)}{(m-l)!}.$$

Переходя к пределу при $\varepsilon \rightarrow 0$, получаем

$$\lim_{\varepsilon \rightarrow 0} f(x_{11}^\varepsilon; \dots; x_{im}^\varepsilon) = \frac{f^{(m-l)}(x_l)}{(m-l)!}. \quad (3)$$

Таким образом, все разности таблицы (2) вида $f(x_{11}^\varepsilon; \dots; x_{im}^\varepsilon)$ при $\varepsilon \rightarrow 0$ имеют пределы, которые естественно обозначать $\underbrace{f(x_l; \dots; x_l)}_{m-l+1 \text{ раз}}$; (из (3) следует, что величина предела зависит только

$m-l$ и x_l). Имеем

$$\underbrace{f(x_l; \dots; x_l)}_{m-l+1 \text{ раз}} = \frac{f^{(p)}(x_l)}{p!}. \quad (4)$$

Индукцией по порядку разности покажем, что все разности, входящие в таблицу (2), имеют пределы. Для разностей нулевого порядка, совпадающих со значениями функции, это очевидно. Пусть утверждение доказано для разностей порядка $q-1$.

Выразим каждую из разностей порядка q через разности порядка $q-1$:

$$f(x_{il}^e; \dots; x_{jp}^e) = \frac{f(x_{i, l+1}^e; \dots; x_{jp}^e) - f(x_{il}^e; \dots; x_{j, p-1}^e)}{x_{jp}^e - x_{il}^e}; \quad (5)$$

для определенности мы рассматриваем здесь случай $l < m_i$ и $1 < p$. При $j = i$ существование предела разностей порядка q следует из (3). При $j \neq i$ предел знаменателя в правой части (5) равен $x_j - x_i \neq 0$, а предел числителя существует по предположению индукции. Тогда предел левой части также существует. Таким образом, показано, что существуют пределы у всех разностей порядка q и, следовательно, у всех разностей таблицы. Предел левой части (5) обозначаем

$$\underbrace{f(x_i; \dots; x_i)}_{m_i - l + 1 \text{ раз}}; \underbrace{x_{l+1}; \dots; x_{i+1}}_{m_{l+1} \text{ раз}}; \dots; \underbrace{x_j; \dots; x_j}_{p \text{ раз}}.$$

Таблица пределов разностей

может быть заполнена при помощи формулы (4) и рекуррентного соотношения, следующего из (5):

$$\begin{aligned} f(\underbrace{x_i; \dots; x_i}_{m_i - l + 1 \text{ раз}}; \dots; \underbrace{x_j; \dots; x_j}_{p \text{ раз}}) = \\ = \frac{f(\underbrace{x_i; \dots; x_i}_{m_i - l \text{ раз}}; \dots; \underbrace{x_j; \dots; x_j}_{p \text{ раз}}) - f(\underbrace{x_i; \dots; x_i}_{m_i - l + 1 \text{ раз}}; \dots; \underbrace{x_j; \dots; x_j}_{p - 1 \text{ раз}})}{x_j - x_i} \end{aligned} \quad (6)$$

при $j \neq i$.

Отсюда следует, что коэффициенты A_k^ε имеют при $\varepsilon \rightarrow 0$ пределы A_k , где

$$A_0 = f(x_1), \quad A_1 = f'(x_1), \\ A_2 = \frac{f''(x_1)}{2}, \quad \dots, \quad A_{s-1} = f(x_1; \dots; x_1; \dots; x_n; \dots; x_n).$$

Таким образом, многочлены $g_s^\varepsilon(x)$ стремятся при $\varepsilon \rightarrow 0$ к некоторому многочлену

$$g_s(x) = A_0 + A_1(x - x_1) + A_2(x - x_1)^2 + \dots \\ \dots + A_{s-1}(x - x_1)^{m_1} \dots (x - x_{n-1})^{m_{n-1}}(x - x_n)^{m_n-1} = \\ = f(x_1) + f(x_1; x_1)(x - x_1) + f(x_1; x_1; x_1)(x - x_1)^2 + \dots \quad (7)$$

Многочлен $g_s(x)$ записывается в виде

$$g_s(x) = \sum_{i=1}^{m_1} \frac{f^{(i-1)}(x_1)}{(i-1)!} (x - x_1)^{i-1} + O((x - x_1)^{m_1}).$$

Отсюда вытекает, что он удовлетворяет условиям, заданным в точке x_1 . Вследствие единственности интерполяционного многочлена многочлен $g_s^\varepsilon(x)$ не изменится при переобозначении $x_1 = x_j$, $x_j = x_1$. Поэтому предельный многочлен будет удовлетворять заданным условиям в любой точке x_j . Следовательно, этот многочлен является искомым.

Для упрощения вывода остаточного члена выберем последовательности x_{ij}^ε такими, чтобы при $0 < \varepsilon < \varepsilon_0$ выполнялось соотношение

$$x_{ij}^\varepsilon \in [y_1, y_2],$$

где, как и в § 3,

$$y_1 = \min(x, x_1, \dots, x_n), \quad y_2 = \max(x, x_1, \dots, x_n).$$

Согласно (3.1), имеем

$$f(x) - g_s^\varepsilon(x) = \left(\frac{f^{(s)}(\xi_\varepsilon)}{s!} \right) \omega_s^\varepsilon(x); \quad (8)$$

здесь

$$\omega_s^\varepsilon(x) = \prod_{i=1}^n \prod_{j=1}^{m_i} (x - x_{ij}^\varepsilon).$$

При $x \neq x_1, \dots, x_n$ перейдем в (8) к пределу при $\varepsilon \rightarrow 0$. Поскольку левая часть в (8) имеет предел $f(x) - g_s(x)$, а

$$\lim_{\varepsilon \rightarrow 0} \omega_s^\varepsilon(x) = \omega_s(x) = \prod_{i=1}^n (x - x_i)^{m_i},$$

то существует предел $\lim_{\varepsilon \rightarrow 0} f^{(s)}(\xi_\varepsilon) = z(x)$; таким образом,

$$f(x) - g_s(x) = \frac{z(x)}{s!} \omega_s(x). \quad (9)$$

Переходя к пределу при $\varepsilon \rightarrow 0$ в соотношениях

$$M_1 = \min_{[y_1, y_2]} f^{(s)}(x) \leq f^{(s)}(\xi_\varepsilon) \leq \max_{[y_1, y_2]} f^{(s)}(x) = M_2,$$

получим $M_1 \leq z(x) \leq M_2$. По теореме Ролля существует точка $\xi \in [y_1, y_2]$ такая, что $z(x) = f^{(s)}(\xi)$.

Таким образом, равенство (9) приобретает вид

$$f(x) - g_s(x) = \frac{f^{(s)}(\xi)}{s!} \omega_s(x). \quad (10)$$

Согласно (5.1), справедливо равенство

$$f(x) - g_s^e(x) = f(x; x_{11}^e; \dots; x_{nm_n}^e) \omega_s^e(x).$$

Переходя к пределу при $\varepsilon \rightarrow 0$, получим

$$f(x) - g_s(x) = f(x; x_1; \dots; x_n) \omega_s(x). \quad (11)$$

Сравнивая это равенство с (10), имеем

$$f(x; x_1; \dots; x_n) = \frac{f^{(s)}(\xi)}{s!}.$$

Это соотношение остается в силе при предельном переходе $x \rightarrow x_j$, j — любое. Из этих соотношений следует, что формула (5.4)

$$f(x_1; \dots; x_{N+1}) = \frac{f^{(N)}(\xi)}{N!}$$

(переписанная в других обозначениях) справедлива и в случае, когда не все $x_1, \dots, x_{N+1}$ — различные.

Мы доказали существование интерполяционного многочлена, удовлетворяющего условиям (1). Задачу интерполяции можно было бы поставить и таким образом:

Задана таблица чисел a_{ij} , $i = 1, \dots, n$, $j = 1, \dots, m_i$. Требуется построить многочлен $g_s(x)$ степени $s-1$, удовлетворяющий условиям

$$g_s^{(j-1)}(x_i) = a_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, m_i.$$

Эта задача равносильна исходной задаче, поскольку всегда можно указать гладкую функцию $f(x)$ такую, что

$$f^{(j-1)}(x_i) = a_{ij}$$

§ 7. Уравнения в конечных разностях

Уравнениями в конечных разностях принято называть уравнения, связывающие значения функций целочисленных аргументов. Роль конечно-разностных уравнений в дискретном анализе аналогична роли дифференциальных и интегральных уравнений в классическом анализе.

Рассмотрим простейший случай одного линейного уравнения относительно неизвестной функции одного целочисленного аргумента

$$ly = \sum_{i=0}^k a_i(n) y(n+i) = f(n). \quad (1)$$

Это уравнение является разностным аналогом линейного дифференциального уравнения k -го порядка

$$\bar{ly} = \sum_{i=0}^k b_i(x) y^{(i)}(x) = f(x) \quad (2)$$

и называется *линейным разностным уравнением k -го порядка*.

Изучение этого уравнения, так же как и построения предыдущего параграфа, показывает, что рассмотрение всяких исключительных случаев: кратные корни, кратные узлы интерполяции, часто можно провести с помощью предельного перехода от регулярного случая.

Исследование уравнений (1) и (2) будем проводить параллельно.

Каждое из этих уравнений имеет вид $Ly = h$, где L — линейный оператор; уравнение $Ly = 0$ называют *однородным*; формулы

$$y = y(C_1, \dots, C_l, n) \quad \text{и} \quad y = y(C_1, \dots, C_l, x),$$

соответственно, называют *общим решением*, если при подстановке значений параметров C_i можно получить любое решение рассматриваемого уравнения. Если v — частное решение неоднородного уравнения $Lv = h$, то разность $y - v$ является решением однородного уравнения: $L(y - v) = h - h = 0$. Таким образом, общее решение неоднородного уравнения есть сумма частного решения неоднородного уравнения и общего решения однородного. Решения $y_1, \dots, y_m$ однородного уравнения $Ly = 0$ называют *линейно зависимыми* в рассматриваемой области изменения независимого аргумента, если существуют $C_1, \dots, C_m$, не все равные нулю, при которых $C_1 y_1 + \dots + C_m y_m \equiv 0$. В противном случае эти решения называют *линейно независимыми*. Если функции u_i являются решениями однородного уравнения $Lu = 0$, то любая функция $\sum_i C_i u_i$ также является решением этого

уравнения, поскольку

$$L\left(\sum_i C_i u_i\right) = \sum_i C_i (Lu_i) = 0.$$

Пусть для определенности уравнение (2) рассматривается в области $x \geq 0$, а уравнение (1) в области $n \geq 0$.

Теорема

Пусть $b_k(x) \neq 0$ при $x \geq 0$ и все $b_j(x)$ непрерывны при $x \geq 0$ | Пусть $a_k(n) \neq 0$ при $n \geq 0$ и все $a_i(n)$ ограничены при каждом $n \geq 0$.

Тогда общее решение однородного уравнения $Ly = 0$ записывается в виде

$$y = \sum_{i=1}^k C_i u_i,$$

где $u_1, \dots, u_k$ линейно независимые решения уравнения $Ly = 0$.

Доказательство

Согласно теореме существования уравнение $ly = 0$ имеет решение при любых начальных условиях $y(0), \dots, y^{(k-1)}(0)$.

Согласно теореме единственности, это решение единственно.

Обозначим через $y_i(x)$ решения уравнения $ly = 0$ при начальных условиях

$$y_i^{(j-1)} = \delta_{i-j}, \quad i, j = 1, \dots, k.$$

Однородное уравнение $ly = 0$ можно представить в виде $y(n+k) =$

$$= - \sum_{i=0}^{k-1} \frac{a_i(n)}{a_k(n)} y(n+i). \quad (3)$$

Если мы зададимся $y(0), \dots, y(k-1)$, то из (3) сможем вычислить последовательно $y(k), y(k+1), \dots$. Таким образом, при любых $y(0), \dots, y(k-1)$ уравнение $l(y) = 0$ имеет решение.

Это решение единственно, поскольку значения любого решения удовлетворяют уравнению (3), а из этого уравнения значения $y(k), y(k+1), \dots$ определяются однозначно.

Обозначим через $y_i(n)$ решения уравнения $ly = 0$ при начальных условиях

$$y_i(j-1) = \delta_{i-j}, \quad i, j = 1, \dots, k.$$

Эти решения образуют линейно независимую систему. В самом деле, если

$$\begin{array}{l|l} \sum_{i=1}^k C_i y_i(x) \equiv 0, & \sum_{i=1}^k C_i y_i(n) \equiv 0, \\ \text{то при } j=1, \dots, k \text{ имеем} & \text{то при } j=1, \dots, k \text{ имеем} \\ 0 = \sum_{i=1}^k C_i y_i^{(j-1)}(0) = & 0 = \sum_{i=1}^k C_i y_i(j-1) = \\ = \sum_{i=1}^k C_i \delta_{i-j} = C_j. & = \sum_{i=1}^k C_i \delta_{i-j} = C_j. \end{array}$$

Следовательно, в случае

$$\sum_{i=1}^k C_i y_i \equiv 0$$

все C_i обязательно равны нулю, и поэтому функции $y_1, \dots, y_k$ линейно независимы.

Пусть $y(x)$ — какое-либо решение уравнения $\bar{L}y = 0$. Пусть $y(n)$ — какое-либо решение уравнения $ly = 0$.
Функция Функция

$$z(x) = \sum_{j=1}^k y_j^{(j-1)}(0) y_j(x) \quad \Bigg| \quad z(n) = \sum_{j=1}^k y_j(j-1) y_j(n)$$

является решением этого уравнения при начальных условиях $y(0), \dots, y^{(k-1)}(0)$. $\Bigg|$ $y(0), \dots, y(k-1)$.

Вследствие единственности решения уравнения

$$\bar{L}y = 0 \text{ имеем} \quad \Bigg| \quad y = 0 \text{ имеем}$$

$$y(x) = \sum_{j=1}^k y_j^{(j-1)}(0) y_j(x) \quad \Bigg| \quad y(n) = \sum_{j=1}^k y_j(j-1) y_j(n).$$

Теорема доказана.

Далее в курсе дифференциальных уравнений доказывается такая теорема. Если нам известно k линейно независимых решений однородного уравнения $\bar{L}y = 0$, то задача нахождения решения неоднородного уравнения (2) сводится к решению уравнений

$$dC_j/dx = g_j(x),$$

где $g_j(x)$ — известные функции, т. е. к отысканию квадратур. Точно так же в случае, когда известно k линейно независимых решений однородного уравнения $ly = 0$, задача нахождения решения неоднородного уравнения

сводится к решению простейших разностных уравнений

$$C_j(n+1) - C_j(n) = g_j(n),$$

где $g_j(n)$ — известные функции. Мы не будем останавливаться на этих теоремах в силу их малой практической значимости. ▲

Перейдем к рассмотрению уравнений с постоянными коэффициентами:

$$\begin{aligned} \bar{ly} = \sum_{i=0}^k b_i y^{(i)}(x) = f(x), \quad b_k \neq 0 \quad (4) \quad & \left| \quad ly = \sum_{i=0}^k a_i y(n+i) = f(n), \quad a_k \neq 0 \quad (5) \right. \end{aligned}$$

и соответствующих однородных уравнений

$$\bar{ly} = \sum_{i=0}^k b_i y^{(i)}(x) = 0, \quad (6) \quad \left| \quad ly = \sum_{i=0}^k a_i y(n+i) = 0. \quad (7) \right.$$

Займемся отысканием частных решений однородного уравнения.

Подставляя в (6) предполагаемый вид частного решения $\exp(\lambda x)$, получаем уравнение

$$\left(\sum_{i=0}^k b_i \lambda^i \right) \exp(\lambda x) = 0.$$

В случае уравнения (7) функцию $\exp(\lambda n)$ удобнее записывать в виде μ^n , $\mu = \exp \lambda$. Подставляя в (7), получаем уравнение

$$\left(\sum_{i=0}^k a_i \mu^i \right) \mu^n = 0.$$

Таким образом, каждому корню уравнения

$$\sum_{i=0}^k b_i \lambda^i = 0, \quad (8) \quad \left| \quad \sum_{i=0}^k a_i \mu^i = 0, \quad (9) \right.$$

называемому *характеристическим*, соответствует частное решение

$$\exp(\lambda x). \quad \left| \quad \mu^n.\right.$$

Если все корни характеристического уравнения простые, мы получаем k различных решений. Покажем, что каждому l -кратному корню характеристического уравнения соответствует l различных решений однородного уравнения

$$\exp(\lambda x), \quad x \exp(\lambda x), \quad \dots \quad \left| \quad \mu^n, \quad C_1^n \mu^{n-1}, \quad \dots, \quad C_{l-1}^{l-1} \mu^{n-l+1}.\right.$$

$$\dots, \quad x^{l-1} \exp(\lambda x).$$

Пусть для определенности $\lambda_1 = \dots = \lambda_l$, $\mu_1 = \dots = \mu_l$.

Разложим на множители характеристический многочлен

$$\sum_{i=0}^k b_i \lambda^i = b_k \prod_{j=1}^k (\lambda - \lambda_j). \quad \left| \quad \sum_{i=0}^k a_i \mu^i = a_k \prod_{j=1}^k (\mu - \mu_j). \right.$$

Зададимся действительным параметром $\varepsilon > 0$, $\varepsilon \rightarrow 0$.

Возьмем $\lambda_{j\varepsilon}$ такие, что

а) $\lambda_{j\varepsilon}$ различны при $j = 1, \dots, l$,

б) стремятся к λ_j при $\varepsilon \rightarrow 0$ для всех $j \leq k$.

Возьмем $\mu_{j\varepsilon}$ такие, что

а) $\mu_{j\varepsilon}$ различны при $j = 1, \dots, l$,

б) стремятся к μ_j при $\varepsilon \rightarrow 0$ для всех $j \leq k$.

Образуем характеристические уравнения, соответствующие этим корням:

$$0 = b_k \prod_{j=1}^k (\lambda - \lambda_{j\varepsilon}) = \sum_{i=0}^k b_{i\varepsilon} \lambda^i. \quad \left| \quad 0 = a_k \prod_{j=1}^k (\mu - \mu_{j\varepsilon}) = \sum_{i=0}^k a_{i\varepsilon} \mu^i. \right.$$

Очевидно,

$$b_{i\varepsilon} \rightarrow b_i \quad \left| \quad a_{i\varepsilon} \rightarrow a_i \right. \quad (10)$$

при $\varepsilon \rightarrow 0$.

Этим характеристическим уравнениям соответствуют уравнения

$$\sum_{i=0}^k b_{i\varepsilon} y_\varepsilon^{(i)}(x) = 0. \quad (11) \quad \left| \quad \sum_{i=0}^k a_{i\varepsilon} y_\varepsilon(n+i) = 0. \quad (12) \right.$$

Пусть при $\varepsilon > 0$ мы можем указать решения

$y_\varepsilon(x)$ уравнения (11) такие, что при любом $x \geq 0$ существует предел

$$\lim_{\varepsilon \rightarrow 0} y_\varepsilon(x) = Y(x),$$

причем $y_\varepsilon(x)$ сходится к $Y(x)$ равномерно вместе со всеми производными до порядка k включительно, на любом конечном отрезке $[x_1, x_2]$.

Переходя к пределу в (11) с учетом (10), получаем, что предельная функция $Y(x)$ удовлетворяет уравнению (6).

$y_\varepsilon(n)$ уравнения (12) такие, что при любом $n \geq 0$ существует предел

$$\lim_{\varepsilon \rightarrow 0} y_\varepsilon(n) = Y(n).$$

Переходя к пределу в (12) с учетом (10), получаем, что предельная функция $Y(n)$ удовлетворяет уравнению (7).

Построим такие последовательности $y_\varepsilon(x)$ и $y_\varepsilon(n)$, которые будут сходиться к частным решениям (6), (7), соответствующим кратным корням. При проведении этих построений удобно

использовать разделенные разности. Рассмотрим сначала случай двукратного корня. Положим

$$y_{2\varepsilon}(x) = (\exp \lambda x) (\lambda_{1\varepsilon}; \lambda_{2\varepsilon}) = \left| \begin{array}{l} y_{2\varepsilon}(n) = (\mu^n) (\mu_{1\varepsilon}; \mu_{2\varepsilon}) = \\ = \frac{\exp(\lambda_{2\varepsilon} x) - \exp(\lambda_{1\varepsilon} x)}{\lambda_{2\varepsilon} - \lambda_{1\varepsilon}} \cdot \end{array} \right| = \frac{\mu_{2\varepsilon}^n - \mu_{1\varepsilon}^n}{\mu_{2\varepsilon} - \mu_{1\varepsilon}}.$$

Очевидно, эти функции являются решениями уравнений (11), (12), соответственно. Запишем их в виде

$$y_{2\varepsilon}(x) = x \exp(\lambda_{1\varepsilon} x) \times \left| \begin{array}{l} y_{2\varepsilon}(n) = \\ \times \frac{\exp((\lambda_{2\varepsilon} - \lambda_{1\varepsilon})x) - 1}{(\lambda_{2\varepsilon} - \lambda_{1\varepsilon})x} \cdot \end{array} \right| = \mu_{2\varepsilon}^{n-1} + \mu_{2\varepsilon}^{n-2} \mu_{1\varepsilon} + \dots + \mu_{1\varepsilon}^{n-1}.$$

Переходя к пределу при $\varepsilon \rightarrow 0$, получим

$$y_{2\varepsilon}(x) \rightarrow x \exp(\lambda_1 x) \quad \left| \quad y_{2\varepsilon}(n) \rightarrow \mu_1^{n-1} + \dots + \mu_1^{n-1} = \right. \\ \left. = n \mu_1^{n-1} \right|$$

Мы получили второе линейно независимое решение, соответствующее двукратному корню. Случай корня кратности, большей двух, рассматривается аналогично. При $1 \leq q \leq l$ положим

$$y_{q\varepsilon}(x) = (\exp(\lambda x)) (\lambda_{1\varepsilon}; \dots; \lambda_{q\varepsilon}). \quad \left| \quad y_{q\varepsilon}(n) = (\mu^n) (\mu_{1\varepsilon}; \dots; \mu_{q\varepsilon}). \right|$$

Вследствие представления разделенной разности (4.3)

$$h_{q\varepsilon}(x) \text{ записывается в виде} \quad \left| \quad y_{q\varepsilon}(n) \text{ записывается в виде} \right|$$

$$\sum_{j=1}^q c_{j\varepsilon} \exp(\lambda_{j\varepsilon} x)$$

$$\sum_{j=1}^q c_{j\varepsilon} \mu_{j\varepsilon}^n$$

и, таким образом, является решением (11).

и, таким образом, является решением (12).

Рассмотрим сначала случай, когда все

$$\lambda_{j\varepsilon}, \quad j = 1, \dots, q, \quad \left| \quad \mu_{j\varepsilon}, \quad j = 1, \dots, q, \right|$$

действительные.

Воспользовавшись формулой (5.4)

$$f(x_1, \dots, x_q) = \frac{f^{(q-1)}(\zeta)}{(q-1)!}, \quad \text{где } y_1 \leq \zeta \leq y_2, \quad (13)$$

получим

$$y_{q\varepsilon}(x) = \frac{x^{q-1}}{(q-1)!} \exp(\lambda_\varepsilon x),$$

$$y_{q\varepsilon}(n) = C_n^{q-1} \mu_\varepsilon^{n-q+1},$$

где

$$\lambda_\varepsilon \in [\min(\lambda_{1\varepsilon}, \dots, \lambda_{q\varepsilon}), \\ \max(\lambda_{1\varepsilon}, \dots, \lambda_{q\varepsilon})].$$

где

$$\mu_\varepsilon \in [\min(\mu_{1\varepsilon}, \dots, \mu_{q\varepsilon}), \\ \max(\mu_{1\varepsilon}, \dots, \mu_{q\varepsilon})].$$

При $\varepsilon \rightarrow 0$ имеем

$$y_{q\varepsilon}(x) \rightarrow Y_q(x) = \frac{x^{q-1}}{(q-1)!} \exp(\lambda_1 x). \quad (14) \quad \left| \quad y_{q\varepsilon}(n) \rightarrow Y_q(n) = C_n^{q-1} \mu_1^{n-q+1} \quad (15) \right.$$

Если не все $\lambda_{j\varepsilon}$ или, соответственно, не все $\mu_{j\varepsilon}$ действительны, соотношения (14), (15) остаются в силе, однако проведенное доказательство непригодно вследствие того, что при комплексных $x_1, \dots, x_q$ формула (13) неверна.

Мы ограничимся доказательством соотношений (14), (15) при специальном выборе $\lambda_{j\varepsilon}, \mu_{j\varepsilon}$. Доказательство проводится в предположении $\mu_1 \neq 0$; при $\mu_1 = 0$ можно взять $\mu_{j\varepsilon} = (j-1)\varepsilon$, и проведенное выше доказательство остается в силе. При $\varepsilon > 0$ положим

$$\lambda_{j\varepsilon} = \lambda_1 + (j-1)\varepsilon \quad (16) \quad \left| \quad \mu_{j\varepsilon} = \mu_1 (1 + (j-1)\varepsilon) \quad (17) \right.$$

при $j = 1, \dots, l$ при $j = 1, \dots, l$

Тогда, воспользовавшись (4.3), получаем равенства

$$\begin{aligned} (\exp(\lambda x))(\lambda_{1\varepsilon}; \dots; \lambda_{q\varepsilon}) &= \sum_{j=1}^q \frac{\exp[(\lambda_1 + (j-1)\varepsilon)x]}{\prod_{i \neq j} ((\lambda_1 + (j-1)\varepsilon) - (\lambda_1 + (i-1)\varepsilon))} = \\ &= \left[\sum_{j=1}^q \frac{\exp(j-1)\varepsilon x}{\prod_{i \neq j} ((j-1)\varepsilon - (i-1)\varepsilon)} \right] \exp(\lambda_1 x) = \\ &= ((\exp x)(0; \varepsilon; \dots; (q-1)\varepsilon)) \exp(\lambda_1 x); \end{aligned}$$

соответственно,

$$\begin{aligned} (\mu^n)(\mu_{1\varepsilon}; \dots; \mu_{q\varepsilon}) &= \sum_{j=1}^q \frac{\mu_1^n (1 + (j-1)\varepsilon)^n}{\prod_{i \neq j} (\mu_1 (1 + (j-1)\varepsilon) - \mu_1 (1 + (i-1)\varepsilon))} = \\ &= \left[\sum_{j=1}^q \frac{(1 + (j-1)\varepsilon)^n}{\prod_{i \neq j} ((1 + (j-1)\varepsilon) - (1 + (i-1)\varepsilon))} \right] \mu_1^{n-q+1} = \\ &= (\mu^n)(1; 1 + \varepsilon; \dots; 1 + (q-1)\varepsilon) \mu_1^{n-q+1}. \end{aligned}$$

Отсюда на основании (13) имеем

$$\begin{aligned} (\exp(\lambda x))(\lambda_{1\varepsilon}; \dots; \lambda_{q\varepsilon}) &= \frac{x^{q-1}}{(q-1)!} \exp((\lambda_1 + \bar{\lambda}_\varepsilon)x), \quad \left| \quad (\mu^n)(\mu_{1\varepsilon}; \dots; \mu_{q\varepsilon}) = C_n^{q-1} \mu_1^{n-q+1} \bar{\mu}_\varepsilon^{n-q+1}, \right. \\ \text{где} \quad 0 \leq \bar{\lambda}_\varepsilon \leq (q-1)\varepsilon. & \quad \left| \quad \text{где} \quad 1 \leq \bar{\mu}_\varepsilon \leq 1 + (q-1)\varepsilon. \right. \end{aligned}$$

Отсюда следует, при $\varepsilon \rightarrow 0$ что выполнены соотношения

$$(\exp(\lambda x))(\lambda_{1\varepsilon}; \dots; \lambda_{q\varepsilon}) \rightarrow Y_q(x). \quad | \quad (\mu^n)(\mu_{1\varepsilon}; \dots; \mu_{q\varepsilon}) \rightarrow Y_q(n).$$

Здесь q может принимать значения $q = 1, \dots, l$

Итак, как в случае уравнения (7), так и в случае уравнения (6), мы смогли поставить в соответствие каждому l -кратному корню характеристического уравнения l различных решений.

Теорема. Совокупность всех полученных решений, соответствующих корням характеристического уравнения (9), образует фундаментальную систему (т. е. они линейно независимы и решение (7) может быть получено как линейная комбинация таких решений).

Наметим доказательство этого утверждения. Пусть $y_1(n), \dots, y_k(n)$ — частные решения (7).

Лемма. Система решений $y_1(n), \dots, y_k(n)$ является фундаментальной, если

$$\det \|y_i(j-1)\| \neq 0.$$

В самом деле, предположим, что эти функции линейно зависимы. Тогда существуют C_i , не все равные нулю, такие, что

$$\sum_{i=1}^k C_i y_i(n) = 0$$

при всех n , в частности при $n=0, \dots, k-1$. Но из условия $\det \|y_i(j-1)\| \neq 0$ следует, что система линейных уравнений

$$\sum_{i=1}^k C_i y_i(j-1) = 0, \quad j = 1, \dots, k,$$

имеет только нулевое решение $C_1 = \dots = C_k = 0$. Мы пришли к противоречию. Покажем, что всякое решение (7) является линейной комбинацией этих решений. В самом деле, система уравнений

$$\sum_{i=1}^k C_i y_i(j-1) = y(j-1), \quad j = 1, \dots, k,$$

имеет решение при любых $y(0), \dots, y(k-1)$. Таким образом, для любого решения $y(n)$ мы можем указать такие $C_1, \dots, C_k$, что решение однородного уравнения

$$z(n) = \sum_{i=1}^k C_i y_i(n)$$

совпадает с $y(n)$ при $n = 0, \dots, k-1$. Вследствие единственности решения разностного уравнения при начальных условиях $y(0), \dots, y(k-1)$ получаем, что

$$y(n) = z(n) = \sum_{i=1}^k C_i y_i(n)$$

при всех n . Лемма доказана.

Если все корни характеристического уравнения (9) различны, то при $y_l(n) = \mu_l^n$ имеем

$$\det \| y_l(j-1) \| = \prod_{i>j} (\mu_i - \mu_j) \neq 0.$$

Рассмотрим общий случай, когда среди корней (9) могут быть кратные: $\mu_1 = \dots = \mu_{l_1}$, $\mu_{l_1+1} = \dots = \mu_{l_1+l_2}$, $\dots$; μ_1 , μ_{l_1+1} , $\mu_{l_1+l_2+1}$, $\dots$ — различны. Образует частные решения возмущенного разностного уравнения (12):

$$y_{l_1 e}(n) = (\mu^n) (\mu_{l_1 e}), \quad y_{l_1 e}(n) = (\mu^n) (\mu_{l_1 e}; \dots; \mu_{l_1 e}),$$

$$y_{l_1+1, e}(n) = (\mu^n) (\mu_{l_1+1, e}), \dots, y_{l_1+l_2, e}(n) = (\mu^n) (\mu_{l_1+1, e}; \dots; \mu_{l_1+l_2, e}), \dots$$

Как показано выше, при специальном выборе корней возмущенного уравнения в виде (17) имеем $y_{j e}(n) \rightarrow Y^{(j)}(n)$ при $e \rightarrow 0$, где

$$\begin{aligned} Y^{(1)}(n) &= \mu_1^n, & Y^{(2)}(n) &= C_n^1 \mu_1^{n-1}, \dots, \\ Y^{(l_1)}(n) &= C_n^{l_1-1} \mu_1^{n-l_1+1}, & Y^{(l_1+1)}(n) &= \mu_{l_1+1}^n, \dots \end{aligned} \quad (18)$$

Очевидно, что $W_e = \det \| y_{j e}(j-1) \| \rightarrow \det \| Y^{(j)}(j-1) \| = W$. Производя линейные преобразования над строками матрицы $\| y_{j e}(j-1) \|$, получаем, что

$$|W_e| = \left| \prod_{i>j}' (\mu_{i e} - \mu_{j e}) \right|;$$

здесь знак ' означает, что в этом произведении не встречается множителей $(\mu_{i e} - \mu_{j e})$ таких, что при $e \rightarrow 0$ $\mu_{i e}$ и $\mu_{j e}$ имеют один и тот же предел. Отсюда вытекает, что

$$W = \lim_{e \rightarrow 0} W_e \neq 0,$$

и, таким образом, система функций $Y^{(j)}(n)$ является фундаментальной. Теорема доказана.

Теорема. Множество линейных комбинаций

$$\sum_{q=1}^l c_q C_n^{q-1} \mu_1^{n-q+1} \quad (19)$$

частных решений (7), соответствующих корню μ_1 кратности l , совпадает с множеством функций

$$P_{l-1}(n) \mu_1^n, \quad (20)$$

где $P_{l-1}(n)$ — произвольные многочлены степени $l-1$.

Поскольку всякая функция $C_n^{q-1} \mu_1^{n-q}$ есть многочлен степени $q-1 < l$ по переменной n , то всякая функция вида (19) представима в виде (20). С другой стороны, пусть $P_{l-1}(n)$ — некоторый многочлен степени $l-1$ от переменной n . Всякий многочлен степени $l-1$ является интерполяционным многочленом степени $l-1$ для самого себя при любых l узлах интерполяции. Поэтому в (5.3) мы можем положить $n = l$, $L_n = P_{l-1}$, $f = P_{l-1}$. Кроме того, положим $x_j = j-1$ и $x = n$. Получим

$$P_{l-1}(n) = A_0 + A_1 n + A_2 n(n-1) + \dots + A_{l-1} n(n-1) \dots (n-(l-2)),$$

где $A_j = P_{l-1}(0; \dots; j)$; иначе это равенство можно переписать в виде

$$P_{l-1}(n) = \sum_{q=0}^{l-1} c_q C_n^{q-1} \mu_1^{l-q}, \quad c_q = A_q (q-1)! \mu_1^{q-1};$$

отсюда следует, что каждое решение вида (20) может быть представлено в виде (19). Теорема доказана.

Таким образом, вместо системы фундаментальных решений (18) можно взять систему фундаментальных решений

$$Y_1(n) = \mu_1^n, Y_2(n) = n\mu_1^n, \dots, Y_{l_1}(n) = n^{l_1-1}\mu_1^n, Y_{l_1+1}(n) = \mu_{l_1+1}^n, \dots$$

Часто такой выбор решения фундаментальной системы более удобен.

Рассмотрим теперь разностное уравнение

$$\sum_{i=0}^k a_i y(n+i) = \left(\sum_{j=0}^{m-1} C_j n^j \right) \sigma^n. \quad (21)$$

Пусть σ является корнем характеристического уравнения (9) кратности l ; в частности, если σ не является корнем этого уравнения, то $l=0$. Можно показать, что уравнение (21) имеет частное решение вида

$$\left(\sum_{j=l}^{l+m-1} d_j n^j \right) \sigma^n; \quad (22)$$

коэффициенты d_j могут быть найдены методом неопределенных коэффициентов.

§ 8. Многочлены Чебышева

Многочлены Чебышева $T_n(x)$, где $n \geq 0$, определяются соотношениями

$$\begin{aligned} T_0(x) &= 1, & T_1(x) &= x, \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x) & \text{при } n > 0. \end{aligned} \quad (1)$$

Пользуясь рекуррентной формулой (1), получаем, например,

$$\begin{aligned} T_2(x) &= 2x^2 - 1, & T_3(x) &= 4x^3 - 3x, \\ T_4(x) &= 8x^4 - 8x^2 + 1, & T_5(x) &= 16x^5 - 20x^3 + 5x, \dots \end{aligned}$$

Старший член $T_{n+1}(x)$ получается из старшего члена $T_n(x)$ умножением на $2x$ и, следовательно, старший член $T_n(x)$ при $n > 0$ есть $2^{n-1}x^n$.

Все многочлены $T_{2n}(x)$ являются четными функциями, а $T_{2n+1}(x)$ — нечетными.

При $n = 0$ это очевидно. Предположив справедливость этого утверждения при некотором n , мы получим: $2xT_{2n+1}(x)$ — четная функция и, вследствие (1), $T_{2n+2}(x)$ — четная. Тогда $2xT_{2n+2}(x)$ — нечетная и $T_{2n+3}(x)$ — нечетная вследствие (1).

При любом θ имеем

$$\cos((n+1)\theta) = 2 \cos \theta \cos n\theta - \cos((n-1)\theta).$$

Полагая $\theta = \arccos x$, получим

$$\cos((n+1)\arccos x) = 2x \cos(n\arccos x) - \cos((n-1)\arccos x).$$

Функция $\cos(n\arccos x)$ удовлетворяет тому же разностному уравнению (1) по переменной n , что и $T_n(x)$; начальные условия при $n = 0$ и $n = 1$ одни и те же:

$$\cos(0 \cdot \arccos x) = 1 = T_0(x),$$

$$\cos(1 \cdot \arccos x) = x = T_1(x);$$

поэтому при всех n

$$T_n(x) = \cos(n\arccos x). \quad (2)$$

Следовательно,

$$|T_n(x)| \leq 1 \quad \text{при} \quad |x| \leq 1. \quad (3)$$

Не нужно думать, что $|T_n(x)| \leq 1$ при всех x . Если $|x| > 1$, то $\arccos x$ не является действительным числом, а косинус такого числа может оказаться больше 1.

Рекуррентное соотношение (1) является разностным уравнением с характеристическим уравнением

$$\mu^2 - 2x\mu + 1 = 0,$$

отсюда

$$\mu_{1,2} = x \pm \sqrt{x^2 - 1}.$$

При $x \neq \pm 1$ корни простые, и поэтому $T_n(x) = C_1 \mu_1^n + C_2 \mu_2^n$. Из начальных условий $T_0(x) = 1$, $T_1(x) = x$ получаем $C_1 = C_2 = 1/2$; таким образом,

$$T_n(x) = \frac{(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n}{2}. \quad (4)$$

При комплексных значениях $x = z$ функция $\sqrt{z^2 - 1}$ является неоднозначной. Под $\sqrt{z^2 - 1}$ будем понимать ветвь, которая при действительных $z > 1$ принимает положительные значения и определена в плоскости с разрезом по $[-1, 1]$.

Из уравнения

$$T_n(x) = \cos(n\arccos x) = 0$$

получаем что

$$x_m = \cos\left(\frac{\pi(2m+1)}{2n}\right), \quad m = 0, \dots, n-1.$$

— корни $T_n(x)$. Вследствие (2), (3) точками экстремума $T_n(x)$ на $[-1, 1]$ будут точки, где $|T_n(x)| = 1$. Решая это уравнение, получим

$$x_{(m)} = \cos\left(\frac{\pi m}{n}\right), \quad m = 0, \dots, n$$

причем

$$T_n(x_{(m)}) = \cos \pi m = (-1)^m.$$

Многочлены

$$\bar{T}_n(x) = 2^{1-n} T_n(x) = x^n + \dots$$

называют многочленами, *наименее уклоняющимися от нуля*. Это определение объясняется следующим свойством.

Лемма. Если $P_n(x)$ — многочлен степени n со старшим коэффициентом 1, то

$$\max_{[-1, 1]} |P_n(x)| \geq \max_{[-1, 1]} |\bar{T}_n(x)| = 2^{1-n}. \quad (5)$$

Доказательство. Предположим противное. Многочлен $T_n(x) - P_n(x)$ имеет степень $n - 1$; в то же время

$$\text{sign}(\bar{T}_n(x_{(m)}) - P_n(x_{(m)})) = \text{sign}((-1)^m 2^{1-n} - P_n(x_{(m)})) = (-1)^m,$$

так как, согласно предположению, $|P_n(x_{(m)})| < 2^{1-n}$ при всех m . Таким образом, между каждыми двумя точками $x_{(m)}, x_{(m+1)}$ многочлен $T_n(x) - P_n(x)$ меняет знак. Многочлен $T_n(x) - P_n(x)$ степени $n - 1$, отличный от нуля (поскольку он отличен от нуля в точках $x_{(m)}$), имеет n различных корней. Мы пришли к противоречию.

Задача. Доказать более сильное утверждение: если

$$P_n(x) = x^n + \dots \neq \bar{T}_n(x),$$

то

$$\max_{[-1, 1]} |P_n(x)| > 2^{1-n}.$$

Линейной заменой переменных $x' = \frac{b+a}{2} + \frac{b-a}{2} x$ отрезок $[-1, 1]$ можно перевести в заданный отрезок $[a, b]$. Многочлен $T_n(x)$ при этом преобразуется в многочлен $\bar{T}_n\left(\frac{2x - (b+a)}{b-a}\right)$ со старшим коэффициентом $(2/(b-a))^n$. В соответствии с леммой можно утверждать, что многочлен

$$\bar{T}_n^{[a, b]}(x) = (b-a)^n 2^{1-2n} T_n\left(\frac{2x - (b+a)}{b-a}\right)$$

со старшим коэффициентом 1 является многочленом, *наименее уклоняющимся от нуля* на отрезке $[a, b]$. Это означает, что для любого многочлена $P_n(x)$ степени n со старшим коэффициентом

1 справедливо неравенство

$$\max_{[a, b]} |P_n(x)| \geq \max_{[a, b]} |\bar{T}_n^{[a, b]}(x)| = (b-a)^n 2^{1-2n}. \quad (6)$$

Нетрудно проверить, что нулями $\bar{T}_n^{[a, b]}(x)$ являются точки

$$x_m = \frac{b+a}{2} + \frac{b-a}{2} \cos\left(\frac{\pi(2m+1)}{2n}\right), \quad m=0, \dots, n-1.$$

Отметим также, что многочлены $\tilde{T}_0(x) = 1/\sqrt{2}$ и $\tilde{T}_n(x) = T_n(x)$ при $n \geq 1$ образуют на $[-1, 1]$ ортонормированную систему с весом $\frac{2}{\pi\sqrt{1-x^2}}$. Проверим свойство ортонормированности этих функций при $m, n \neq 0$. После замены $x = \cos \theta$ имеем

$$\begin{aligned} \int_{-1}^1 \frac{2T_n(x)T_m(x)}{\pi\sqrt{1-x^2}} dx &= \frac{2}{\pi} \int_0^\pi \cos n\theta \cos m\theta d\theta = \\ &= \frac{1}{\pi} \int_0^\pi (\cos((n-m)\theta) + \cos((n+m)\theta)) d\theta = \delta_{n-m} + \delta_{n+m}. \end{aligned}$$

Второе слагаемое обращается в нуль при $n, m \geq 0$, если $n^2 + m^2 \neq 0$. Отсюда вытекает требуемое равенство

$$\int_{-1}^1 \frac{2\tilde{T}_n(x)\tilde{T}_m(x)}{\pi\sqrt{1-x^2}} dx = \delta_{n-m}.$$

§ 9. Минимизация оценки остаточного члена интерполяционной формулы

Пусть функция $f(x)$ приближается на $[a, b]$ с помощью интерполяционного многочлена степени $n-1$ с узлами интерполяции $x_1, \dots, x_n \in [a, b]$; пусть погрешность оценивается в норме $\|f\| = \sup_{[a, b]} |f(x)|$. Согласно (3.1), имеем

$$f(x) - L_n(x) = \frac{f^{(n)}(\xi) \omega_n(x)}{n!},$$

где $\xi \in [a, b]$, если $x \in [a, b]$. Отсюда следует оценка погрешности интерполяции

$$\|f - L_n\| \leq \frac{\|f^{(n)}\| \|\omega_n\|}{n!}. \quad (1)$$

Займемся минимизацией правой части этой оценки за счет выбора узлов $x_1, \dots, x_n$; для этой цели и были введены в рассмотрение многочлены Чебышева, наименее уклоняющиеся от нуля. Многочлен $\omega_n(x) = (x-x_1)\dots(x-x_n)$ имеет старший

коэффициент 1, и поэтому $\|\omega_n\| \geq (b-a)^n 2^{1-2n}$, согласно (8.6). Если взять в качестве узлов интерполяции

$$x_m = \frac{b+a}{2} + \frac{b-a}{2} \cos\left(\frac{\pi(2m+1)}{2n}\right), \quad m=0, \dots, n-1. \quad (2)$$

то

$$\omega_n(x) = (b-a)^n 2^{1-2n} T_n\left(\frac{2x-(b+a)}{b-a}\right)$$

и

$$\|\omega_n\| = (b-a)^n 2^{1-2n}.$$

Следовательно, при таком расположении узлов справедлива наилучшая из оценок, которая может быть получена как следствие оценки (1):

$$\|f - L_n\| \leq \frac{\|f^{(n)}\| (b-a)^n 2^{1-2n}}{n!}. \quad (3)$$

При получении (1) максимум произведения заменен на произведение максимумов сомножителей. Поэтому может возникнуть надежда получить оценку погрешности, лучшую чем (3). Однако это не так; если $f(x) = a_n x^n + \dots$ — многочлен степени n , то

$$f^{(n)}(\xi) = a_n n! = \text{const},$$

и неравенство (1) превращается в равенство; тогда при любых узлах интерполяции имеем

$$\|f - L_n\| \geq \frac{\|f^{(n)}\| (b-a)^n 2^{1-2n}}{n!} = a_n (b-a)^n 2^{1-2n}.$$

Как отмечалось, важной проблемой вычислительной математики является проблема оптимизации методов решения задач некоторого класса. Общая постановка ее такова.

Задается некоторый класс P решаемых задач p . Задается некоторое множество M методов решения. Пусть $e(p, m)$ — погрешность метода m при решении задачи p . Величину

$$e(P, m) = \sup_{p \in P} e(p, m)$$

называют *погрешностью метода на классе задач P* . Величину

$$E(P, M) = \inf_{m \in M} e(P, m)$$

называют *оптимальной оценкой погрешности методов из множества M на классе задач P* . Если существует метод $m \in M$, на котором эта оценка достигается, т. е. $E(P, M) = e(P, m)$, то такой метод m называют *оптимальным*.

Полученное нами решение задачи об оптимизации узлов распределения интерполяционной формулы можно сформулировать в описанных выше терминах.

Пусть P — множество задач приближения функций, определенных на $[a, b]$ и удовлетворяющих условию $|f^{(n)}(x)| \leq A_n$, M — множество методов приближения, состоящих в том, что функция заменяется ее интерполяционным многочленом $L_n(x)$ по совокупности узлов $x_1, \dots, x_n$; таким образом, метод решения m определяется заданием узлов интерполяции $x_1, \dots, x_n$. Наконец, пусть мера погрешности $e(p, m) = \|f - L_n\|$. Согласно (1), имеем

$$e(P, m) \leq \frac{A_n \|\omega_n\|}{n!}.$$

С другой стороны, для задачи приближения многочлена

$$P_{n+1}(x) = \frac{A_n}{n!} x^n + \dots,$$

относящейся к рассматриваемому классу, имеем

$$e(p, m) = \frac{A_n \|\omega_n\|}{n!};$$

следовательно,

$$e(P, m) = \frac{A_n \|\omega_n\|}{n!}.$$

Как мы видели выше,

$$E(P, M) = \inf_m e(P, m) = \inf_{x_1, \dots, x_n} \left(\frac{A_n \|\omega_n\|}{n!} \right) = \frac{A_n (b-a)^n 2^{1-2n}}{n!}.$$

Таким образом, способ интерполяции по узлам многочлена Чебышева (2) является оптимальным в рассматриваемом смысле.

Строгая постановка и решение подобных вариационных задач очень важны для развития вычислительной математики. По существу основную часть нашего курса можно было бы изложить как решение различных задач оптимизации. Однако мы не будем проводить последовательно такую точку зрения, чтобы не создалось преувеличенное впечатление о роли изучения задачи оптимизации методов.

В заключение произведем сравнение оценки (3) с оценкой погрешности разложения функции в ряд Тейлора. Согласно рассмотренным § 6, отрезок ряда Тейлора

$$t_n(x) = \sum_{j=0}^{n-1} \frac{f^{(j)}\left(\frac{a+b}{2}\right)}{j!} \left(x - \frac{a+b}{2}\right)^j$$

совпадает с интерполяционным многочленом Лагранжа при единственном, n -кратном узле интерполяции $\frac{a+b}{2}$. Поэтому, естественно, при оптимальном распределении узлов интерполяции

(2) мы должны иметь лучшую оценку. В самом деле, оценка

$$\|f - t_n\| \leq \frac{\|f^{(n)}\| (b-a)^n 2^{-n}}{n!}$$

погрешности отрезка ряда Тейлора уступает оценке (3) в 2^{n-1} раз. При любом другом выборе точки, где функция раскладывается в ряд Тейлора, оценка погрешности отрезка ряда Тейлора будет еще хуже.

§ 10. Конечные разности

Пусть узлы таблицы x_i расположены на равных расстояниях: $x_i = x_0 + ih$, f_i — соответствующие значения функции; величину h называют *шагом таблицы*.

Разности $f_{i+1} - f_i$ называют *разностями первого порядка*; в зависимости от обстановки эту величину обозначают как *разность вперед*: Δf_i , *разность назад*: ∇f_{i+1} или *центральную разность*: $\delta f_{i+1/2} = f_{i+1/2}^1$. Таким образом,

$$f_{i+1} - f_i = \Delta f_i = \nabla f_{i+1} = \delta f_{i+1/2} = f_{i+1/2}^1. \quad (1)$$

Разности высшего порядка образуют при помощи рекуррентных соотношений:

$$\Delta^m f_i = \Delta(\Delta^{m-1} f_i) = \Delta^{m-1} f_{i+1} - \Delta^{m-1} f_i,$$

$$\nabla^m f_i = \nabla(\nabla^{m-1} f_i) = \nabla^{m-1} f_i - \nabla^{m-1} f_{i-1},$$

$$\delta^m f_i = \delta(\delta^{m-1} f_i) = \delta^{m-1} f_{i+1/2} - \delta^{m-1} f_{i-1/2},$$

$$f_i^m = f_{i+1/2}^{m-1} - f_{i-1/2}^{m-1}.$$

Таблицу разностей обычно располагают в виде

x	f	f^1	f^2	f^3	
x_0	f_0				.
		$f_{1/2}^1$			.
x_1	f_1		f_1^2		.
		$f_{3/2}^1$		$f_{3/2}^3$	.
x_2	f_2		f_2^2		.
		$f_{5/2}^1$		$f_{5/2}^3$	.
x_3	f_3		f_3^2		.
		$f_{7/2}^1$			.
x_4	f_4				.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

В некоторых интерполяционных формулах, наряду с упомянутыми выше величинами, используются средние арифметические двух последовательных величин одного и того же столбца:

$$\begin{aligned}\mu f_i^m &= (f_{i+1/2}^m + f_{i-1/2}^m)/2 && \text{при } m \text{ нечетном,} \\ \mu f_{i+1/2}^m &= (f_{i+1}^m + f_i^m)/2 && \text{при } m \text{ четном.}\end{aligned}$$

Лемма. Разности m -го порядка выражаются через значения функции по формуле

$$f_i^m = \sum_{j=0}^m (-1)^j C_m^j f_{i+m/2-j}. \quad (2)$$

(При четном m i — целое, при нечетном m i — полуцелое.)

Доказательство проводим методом индукции. При $m=1$ соотношение (2) выполняется, согласно (1). Пусть оно доказано при $m=l$. Имеем

$$f_i^{l+1} = f_{i+1/2}^l - f_{i-1/2}^l = \sum_{j=0}^l (-1)^j C_l^j f_{i+1/2+l/2-j} - \sum_{j=0}^l (-1)^j C_l^j f_{i-1/2+l/2-j}.$$

Собирая коэффициенты при одинаковых f_k и пользуясь равенством

$$C_l^j + C_l^{j+1} = C_{l+1}^{j+1},$$

получим требуемое выражение для величины f_i^{l+1} . Лемма доказана.

Из (2) следует, что оператор конечной разности является линейным. В частности, из (2) имеем

$$\begin{aligned}f_i^2 &= f_{i+1} - 2f_i + f_{i-1}, \\ f_i^3 &= f_{i+3/2} - 3f_{i+1/2} + 3f_{i-1/2} - f_{i-3/2}, \\ f_i^4 &= f_{i+2} - 4f_{i+1} + 6f_i - 4f_{i-1} + f_{i-2},\end{aligned}$$

Лемма. При $x_i \equiv x_0 + ih$ справедливо равенство

$$f(x_i; \dots; x_{i+m}) = \frac{f_{i+m/2}^m}{h^m \cdot m!}. \quad (3)$$

Доказательство проводим по индукции. При $m=1$ имеем

$$f(x_i; x_{i+1}) = \frac{f_{i+1} - f_i}{x_{i+1} - x_i} = \frac{f_{i+1/2}^1}{h}.$$

Предположив, что соотношение (3) верно при всех $m \leq l$ имеем

$$\begin{aligned}f(x_i; \dots; x_{i+l+1}) &= \frac{f(x_{i+1}; \dots; x_{i+l+1}) - f(x_i; \dots; x_{i+l})}{x_{i+l+1} - x_i} = \\ &= \frac{f_{i+1+1/2}^l - f_{i+1/2}^l}{(h^l l!) (h(l+1))} = \frac{f_{i+(l+1)/2}^{l+1}}{h^{l+1} (l+1)!};\end{aligned}$$

таким образом, (3) справедливо и при $m = l + 1$. Лемма доказана.

Согласно (5.4), имеем $f(x_i; \dots; x_{i+m}) = \frac{f^{(m)}(\xi)}{m!}$, где $x_i \leq \xi \leq x_{i+m}$. Сопоставляя это равенство с (3), получаем

$$\Delta^m f_i = \nabla^m f_{i+m} = \delta_{i+m/2}^m f_{i+m/2} = h^m f^{(m)}(\xi). \quad (4)$$

Следствие. Конечные разности n -го порядка от многочлена степени n постоянны, а разности любого более высокого порядка равны нулю.

При ручном счете последнее обстоятельство часто используется для составления таблиц многочленов. Непосредственно вычисляем значения многочлена для $n + 1$ последовательных значений аргумента, по этим значениям составляем таблицу разностей до n -го порядка. Далее заполняем столбец разностей n -го порядка, учитывая, что они постоянны, затем столбец разностей $(n - 1)$ -го порядка, затем $(n - 2)$ -го порядка и т. д.; при заполнении столбцов разностей пользуемся формулой

$$f_i^t = f_{i-1}^t + f_{i-1/2}^{t+1}.$$

Оказывается, что при составлении таким образом таблицы $N \gg n$ значений многочлена степени n вычислительная погрешность в самом неблагоприятном случае может быть порядка N^ϵ , где ϵ — единица последнего разряда, с которым производятся вычисления. Это обстоятельство требует аккуратности при применении этого способа; в частности, при составлении длинной таблицы ее разбивают на части и применяют этот прием к отдельным частям.

Рассмотрим влияние ошибки какого-либо значения f_i на конечные разности различных порядков; пусть вместо f_i стоит $f_i + \epsilon$. Тогда имеем таблицу разностей

$$\begin{array}{rcccl} & & f_{i-2} & & \\ & & f_{i-3/2}^1 & & \\ f_{i-1} & & & f_{i-1}^2 + \epsilon & \\ & & f_{i-1/2}^1 + \epsilon & & \\ f_i + \epsilon & & & f_i^2 - 2\epsilon & \\ & & f_{i+1/2}^1 - \epsilon & & \\ f_{i+1} & & & f_{i+1}^2 + \epsilon & \\ & & f_{i+3/2}^1 & & \\ f_{i+2} & & & & \end{array}$$

Мы видим, что в соответствии с (2) на разности порядка m ошибка распространяется с коэффициентами $(-1)^j C_m^j$. Если функция достаточно гладкая, то ее разности не очень высокого порядка могут оказаться малыми. В то же время на их фоне величины $C_m^j \epsilon$ будут выглядеть достаточно большими. Из наблюдений над таблицей разностей можно указать значение функции, содержащее ошибку, и исправить его,

Точно так же можно обнаруживать ошибки, имевшие место при составлении таблицы разностей. Пусть, например,

$$\begin{aligned} f_{1/2}^3 &= 1 \cdot 10^{-5}, & f_{3/2}^3 &= 2 \cdot 10^{-5}, & f_{5/2}^3 &= 12 \cdot 10^{-5}, \\ f_{7/2}^3 &= -23 \cdot 10^{-5}, & f_{9/2}^3 &= 11 \cdot 10^{-5}, & f_{11/2}^3 &= 1 \cdot 10^{-5}. \end{aligned}$$

Если бы какое-либо значение f_i содержало относительно большую ошибку ϵ , то в третьих разностях это обстоятельство проявилось бы в наличии величин вида ϵ , -3ϵ , 3ϵ , $-\epsilon$. В рассматриваемом случае третьи разности практически равны нулю, за исключением $f_{5/2}^3$, $f_{7/2}^3$, $f_{9/2}^3$, которые примерно имеют вид ϵ , -2ϵ , ϵ ($\epsilon = 11 \cdot 10^{-5}$). Это наводит на мысль, что была допущена ошибка при вычислении значения $f_{7/2}^1$, которая и имела следствием эти возмущения в третьих разностях.

Отметим следующее свойство таблиц разностей функций. Поскольку значения функции $f(x_i)$ вычисляются не абсолютно точно, а округляются в процессе и после вычислений, то на самом деле вместо значений f_i стоят некоторые значения $y_i = f_i + \eta_i$. Тогда значения разностей записываются в виде

$$y_i^m = f_i^m + \eta_i^m.$$

Для достаточно гладкой $f(x)$ (мы не будем определять точно, что это такое) значения f_i^m будут убывать с ростом m . Например, при $f(x) = \sin x$ имеем

$$f_i^m = h^m f^{(m)}(\xi), \quad |f^{(m)}(\xi)| \leq 1;$$

поэтому $|f_i^m| \leq h^m \rightarrow 0$ при $h < 1$, $m \rightarrow \infty$. Величины η_i^m записываются как комбинации слагаемых вида $(-1)^j C_m^j \eta_k$ и поэтому могут расти с ростом m . Поэтому при наличии округлений функция $\varphi(m) = \max_i |y_i^m|$ убывает при малых m и растет при больших m . В той области значений m , где разности высшего порядка начинают расти, основная информация, которая содержится в таблице разностей k -го порядка очень гладкой функции, — это информация о разностях ошибок η_i^k , и использование этих разностей, как правило, не прибавляет ничего нового к нашей информации о функции.

§ 11. Интерполяционные формулы Ньютона для равных промежутков

Высокая точность приближения функции может быть достигнута двумя путями:

1. Добавляются новые узлы интерполяции, и функция интерполируется по всем узлам многочленом максимально высокой степени.

2. После добавления новых узлов интерполяции значения функции вычисляются путем интерполяции по ближайшим узлам; степень интерполяционного многочлена не увеличивается.

Рассмотрим второй из этих путей. Если нет особых причин для какого-то специального выбора узлов, узлы большой таблицы значений функций обычно берут равномерно распределенными: $x_j = x_0 + jh$. Если узлы интерполирования выбираются

вблизи точки x , где вычисляется значение функции, то промежуточная точка ξ в оценке остаточного члена (3.1) также находится вблизи точки x ; таким образом, величина $f^{(n)}(\xi)$ изменяется не очень сильно при выборе узлов в окрестности точки x .

Следовательно, решающее влияние на значение погрешности оказывает величина

$$|\omega_n(x)| = \prod_{j=1}^n |x - x_j|,$$

т. е. произведение расстояний от точки x до узлов интерполирования. Следовательно, величина $|\omega_n(x)|$ будет минимальной, если в качестве узлов интерполирования для нахождения $f(x)$ мы возьмем n узлов, ближайших к x . Очевидно, что при четном $n = 2l$ следует взять по l узлов справа и слева от точки x . При нечетном $n = 2l + 1$ следует взять узел, ближайший к x , и по l узлов слева и справа от него. Если точка находится в начале таблицы или конце, то выбрать такую совокупность узлов невозможно.

При интерполировании в начале или конце таблицы принято записывать интерполяционный многочлен в виде так называемых *формул Ньютона* для интерполирования вперед или назад. Пусть $L_n(x)$ — интерполяционный многочлен Лагранжа по узлам $x_0, \dots, x_{n-1}$. Согласно (5.3), имеем

$$L_n(x) = f(x_0) + f(x_0; x_1)(x - x_0) + \dots \\ \dots + f(x_0; \dots; x_{n-1})(x - x_0) \dots (x - x_{n-2}).$$

Произведем замену переменных $x = x_0 + ht$ и перейдем от разделенных разностей к конечным, согласно (10.3). Получим

$$L_n(x_0 + ht) = f_0 + f_{1/2}^1 t + \dots + f_{(n-1)/2}^{n-1} \frac{t(t-1) \dots (t-(n-2))}{(n-1)!}. \quad (1)$$

Остаточный член (3.1) представится в виде

$$f^{(n)}(\xi) \frac{t(t-1) \dots (t-(n-1)) h^n}{n!}.$$

Формулу (1) называют *интерполяционной формулой Ньютона для интерполирования вперед*. Если мы произведем такую же замену переменных в интерполяционном многочлене $L_n(x)$ по узлам $x_0, x_{-1}, \dots, x_{-(n-1)}$:

$$L_n(x) = f(x_0) + f(x_0; x_{-1})(x - x_0) + \dots \\ \dots + f(x_0; \dots; x_{-(n-1)})(x - x_0) \dots (x - x_{-(n-2)}),$$

то получим *интерполяционную формулу Ньютона для интерполирования назад*

$$L_n(x_0 + ht) = f_0 + f_{-1/2}^1 t + \dots + f_{-(n-1)/2}^{n-1} \frac{t(t+1) \dots (t+(n-2))}{(n-1)!} \quad (2)$$

с остаточным членом

$$f^{(n)}(\xi) \frac{t(t+1) \dots (t+(n-1)) h^n}{n!}.$$

Обратимся к вопросу о практической оценке погрешности. Если функция $f(x)$ задается аналитически, то можно оценить ее производную и отсюда получить оценку остаточного члена интерполяции. Однако, если функция достаточно сложная, получение такой оценки может потребовать много времени и достаточной квалификации; кроме того, такая оценка часто оказывается существенно завышенной. Поэтому обычно поступаются строгостью оценки в угоду ее простоте: верхнюю грань производной заменяют каким-то приближением к ней (верхней грани).

Если оценивается погрешность $f(x) - L_n(x)$ в конкретной точке x , то, согласно (5.6), (5.7), имеем

$$f(x) - L_n(x) \approx f(x_1; \dots; x_{n+1}) \omega_n(x).$$

В других случаях важна малость погрешности интерполирования во всех точках отрезка $[a, b]$, и необходимо иметь оценку величины

$$M_n = \max_{[a, b]} |f^{(n)}(x)|.$$

Для определенности речь пойдет об интерполировании функции, заданной на сетке с постоянным шагом: $x_q = x_0 + qh$. Если $f^{(n)}(x)$ непрерывна, то справедливо равенство

$$\lim_{h \rightarrow 0} \sigma_n^h = M_n; \quad \sigma_n^h = \max_{a \leq x_m \leq b-nh} \left| \frac{\Delta^n f_m}{h^n} \right|.$$

Поэтому при малых h можно принять $M_n \approx \sigma_n^h$.

§ 12. Интерполяционные формулы Бесселя и Эверетта. Составление таблиц

Во многих случаях приобретает особо важное значение малость степени полиномов, приближающих функцию. Уменьшения степени таких полиномов без потери точности иногда можно достигнуть, образуя линейные комбинации интерполяционных полиномов. Рассмотрим простейшие из таких способов приближения функций.

Если точка интерполирования x лежит на $(x_0, x_0 + h/2]$, то узлы интерполирования удобно брать в последовательности

$$x_0, x_0 + h, x_0 - h, \dots, x_0 + lh, x_0 - lh.$$

Тогда при любом q группа из первых q узлов этой последовательности состоит из узлов, ближайших к x , и оценка остаточного члена вида

$$|f(x) - L_n(x)| \leq \max_x \frac{|f^{(n)}(x)|}{n!} |\omega_n(x)|.$$

для интерполяционного многочлена по этой совокупности узлов будет лучше, чем для любой другой совокупности.

Интерполяционная формула, написанная по первым узлам этой последовательности, называется *формулой Гаусса для интерполирования вперед*. Последовательность узлов $x_0, x_0-h, x_0+h, \dots$ в том же самом смысле будет наилучшей для интерполирования в точке из $[x_0 - h/2, x_0]$. Такую интерполяционную формулу называют *формулой Гаусса для интерполирования назад*. Формулы Гаусса имеют самостоятельное значение и широко применяются при ручном счете. Выпишем формулу Гаусса для интерполирования вперед по $2n+2$ узлам

$$L_{2n+2}(x) = f(x_0) + f(x_0; x_1)(x - x_0) + \dots \\ \dots + f(x_0; x_1; x_{-1}; \dots; x_{n+1})(x - x_0) \dots (x - x_n)(x - x_{-n}). \quad (1)$$

После замены $x = x_0 + ht$ получим

$$G_{2n+2}(x_0 + ht) = L_{2n+2}(x_0 + ht) = \\ = f_0 + f_{1/2}^1 t + f_0^2 \frac{t(t-1)}{2!} + \dots + f_{1/2}^{2n+1} \frac{t(t^2-1) \dots (t^2-n^2)}{(2n+1)!}. \quad (2)$$

Выпишем теперь формулу Гаусса того же порядка для интерполирования назад по отношению к узлу x_1 , т. е. по совокупности узлов $x_1, x_0, x_2, x_{-1}, \dots, x_{n+1}, x_{-n}$:

$$L_{2n+2}(x) = f(x_1) + f(x_1; x_0)(x - x_1) + \dots \\ \dots + f(x_1; x_0; \dots; x_{n+1}; x_{-n})(x - x_1)(x - x_0) \dots (x - x_{n+1}). \quad (3)$$

После замены переменных $x = x_0 + ht$ получим

$$G_{2n+2}(x_0 + ht) = f_1 + f_{1/2}^1(t-1) + f_1^2 \frac{t(t-1)}{2!} + \dots \\ \dots + f_{1/2}^{2n+1} \frac{t(t^2-1) \dots (t^2-(n-1)^2)(t-n)(t-(n+1))}{(2n+1)!}. \quad (4)$$

Формулы (1), (2), (3), (4) являются различными формами записи одного и того же интерполяционного многочлена Лагранжа по узлам $x_{-n}, \dots, x_{n+1}$. Поэтому можно написать, что этот многочлен Лагранжа равен полусумме правых частей (2) и (4). При образовании этой полусуммы заметим следующее: разности нечетного порядка в обеих формулах заметны и те же, и поэтому можно привести подобные члены, соответствующие одинаковым разностям; в то же время коэффициенты при разностях четного порядка одинаковы и их также можно объединить. В результате получим *интерполяционную формулу Бесселя*

$$B_{2n+2}(x_0 + ht) = G_{2n+2}(x_0 + ht) = L_{2n+2}(x_0 + ht) = \\ = \mu f_{1/2} + f_{1/2}^1(t-1/2) + \mu f_{1/2}^2 \frac{t(t-1)}{2!} + \dots \\ \dots + \mu f_{1/2}^{2n} \frac{t(t^2-1) \dots (t^2-(n-1)^2)(t-n)}{(2n)!} + \\ + f_{1/2}^{2n+1} \frac{t(t^2-1) \dots (t^2-(n-1)^2)(t-n)(t-1/2)}{(2n+1)!} \quad (5)$$

(здесь и далее опускаются промежуточные выкладки). Естественно, что остаточный член этой формулы совпадает с остаточным членом интерполяционного многочлена Лагранжа

$$f(x) - B_{2n+2}(x) = f^{(2n+2)}(\xi) \frac{(x - x_{-n}) \dots (x - x_n)(x - x_{n+1})}{(2n+2)!},$$

или

$$\begin{aligned} f(x_0 + ht) - B_{2n+2}(x_0 + ht) &= \\ &= f^{(2n+2)}(\xi) h^{2n+2} \frac{t(t^2 - 1) \dots (t^2 - n^2)(t - (n+1))}{(2n+2)!}. \end{aligned} \quad (6)$$

По сравнению с формулами Лагранжа и Гаусса (1), (2) эта формула дает лишь определенные удобства пользования; в частности, при интерполировании на середину отрезка, т. е. при $t = 1/2$, коэффициенты при разностях нечетного порядка обращаются в нуль. Однако, если отбросить последнее слагаемое в правой части (5), то получится формула иного типа:

$$\begin{aligned} B_{2n+1}(x_0 + ht) &= \mu f_{1/2} + f_{1/2}'(t - 1/2) + \\ &+ \mu f_{1/2}^2 \frac{t(t-1)}{2!} + \dots + \mu f_{1/2}^{2n} \frac{t(t^2-1) \dots (t^2-(n-1)^2)(t-n)}{(2n)!}. \end{aligned} \quad (7)$$

Многочлен $B_{2n+1}(x_0 + ht)$ степени $2n$ уже не является собственно интерполяционным многочленом. Он образуется как полусумма интерполяционных многочленов с узлами

$$x_{-n}, \dots, x_n \quad \text{и} \quad x_{-(n-1)}, \dots, x_{n+1},$$

соответственно, т. е. совпадает с $f(x)$ лишь в $2n$ узлах $x_{-(n-1)}, \dots, x_n$, в то же время при его написании используются также значения функции в узлах x_{-n} и x_{n+1} (рис. 2.12.1). Его остаточный член, естественно, складывается из отброшенного члена в правой части (5) и правой части (6):

$$\begin{aligned} f(x_0 + ht) - B_{2n+1}(x_0 + ht) &= \\ &= f_{1/2}^{2n+1} \frac{t(t^2-1) \dots (t^2-(n-1)^2)(t-n)(t-1/2)}{(2n+1)!} + \\ &+ f^{(2n+2)}(\xi) h^{2n+2} \frac{t(t^2-1) \dots (t^2-n^2)(t-(n+1))}{(2n+2)!}. \end{aligned} \quad (8)$$

Можно убедиться, что первый член правой части (8), являющийся главным, меньше, чем главный член погрешности каждой из формул, полусуммой которых является $B_{2n+1}(x_0 + ht)$. Мы проверим далее последнее утверждение при $n = 1$.

Рассмотрим следующую типичную задачу. Требуется построить таблицу значений некоторой функции так, чтобы погрешность при интерполяции значений функции многочленом заданной степени m не превосходила ϵ . В этом случае говорят, что

таблица допускает интерполяцию степени m (с погрешностью ϵ). Таблицы, выпускаемые для широкого круга пользователей, обычно состояются так, чтобы они допускали интерполяцию первой степени, иначе — *линейную интерполяцию*. Примером таких таблиц могут служить таблицы В. М. Брадиса, известные из школьного курса. В дальнейшем рассматриваем случай таблицы с постоянным шагом.

Для вычисления значения $f(x)$ при помощи такой таблицы берутся узлы x_q и x_{q+1} справа и слева от точки x : $x_q < x < x_{q+1}$ (они будут ближайшими к x); затем $f(x)$ заменяется интерполя-

ционным многочленом первой степени по этим узлам:

$$f(x_q + th) \approx L_2(x_q + th) = f_q + f'_{q+1/2} t.$$

Погрешность этой формулы

$$f''(\xi) h^2 \frac{t(t-1)}{2},$$

где

$$x_q \leq \xi \leq x_{q+1}.$$

Эта величина не превосходит ϵ , если

$$(\max |f''(\xi)|) \times \times h^2 \max_{[0, 1]} \left| \frac{t(t-1)}{2} \right| \leq \epsilon;$$

$\max_{[0, 1]} \left| \frac{t(t-1)}{2} \right|$ достигается при $t = 1/2$ и равен $1/8$. Таким образом, достаточно выполнения условия

$$\max |f''(\xi)| h^2/8 \leq \epsilon. \quad (9)$$

Пусть мы хотим составить или ввести в машину таблицу $\sin x$ на $[0, \pi/2]$ так, чтобы погрешность линейной интерполяции не превосходила $0,5 \cdot 10^{-6}$. Поскольку $\max |(\sin x)''| = 1$, то из (9) вытекает требование на шаг таблицы

$$h^2/8 \leq 0,5 \cdot 10^{-6} \quad \text{или} \quad h \leq 0,002.$$

Часто требование допустимости линейной интерполяции является слишком жестким и вместо него требуют допустимости *квадратичной интерполяции* (т. е. интерполяции многочленом второй степени). Простейшим случаем квадратичной интерполяции будет интерполяция многочленом Лагранжа по ближайшим

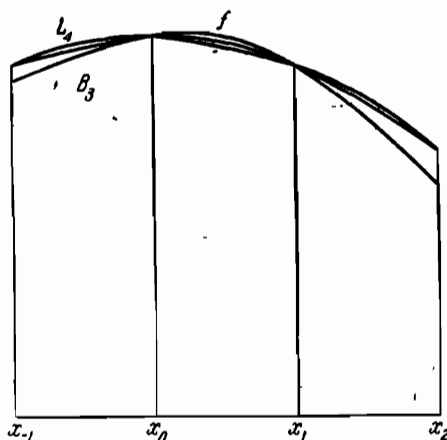


Рис. 2.12.1.

трем узлам. Пусть x_q — узел, ближайший к x . Положим $x = x_q + ht$ и

$$f(x_q + ht) \approx L_3(x_q + ht) = f_q + f_{q+1/2}^1 t + f_q^2 \frac{t(t-1)}{2}. \quad (10)$$

Остаточный член этой формулы

$$f^{(3)}(\xi) h^3 \frac{t(t^2-1)}{3!}.$$

Чтобы таблица допускала квадратичную интерполяцию (10), достаточно выполнения условия

$$\max |f^{(3)}(\xi)| h^3 \max_{|t| \leq 1/2} \left| \frac{t(t^2-1)}{6} \right| \leq \varepsilon.$$

Так как $\max_{|t| \leq 1/2} \left| \frac{t(t^2-1)}{6} \right| = \frac{1}{16}$, то это требование на шаг переписывается в виде

$$\max |f^{(3)}(\xi)| h^3/16 \leq \varepsilon. \quad (11)$$

В конкретном случае $f(x) = \sin x$, $\varepsilon = 0,5 \cdot 10^{-6}$ получаем $h \leq 0,02$.

Рассмотрим другой способ замены функции многочленом второй степени. Пусть $x \in (x_q, x_{q+1})$, $x = x_q + ht$; положим

$$\begin{aligned} f(x_q + ht) &\approx B_3(x_q + ht) = \\ &= \mu f_{q+1/2} + f_{q+1/2}^1 (t-1/2) + \mu f_{q+1/2}^2 \frac{t(t-1)}{2}, \end{aligned} \quad (12)$$

т. е. заменим $f(x)$ многочленом Бесселя второй степени (7), написанным по узлам x_{q-1} , x_q , x_{q+1} , x_{q+2} . Согласно (8), остаточный член (12) есть

$$f_{q+1/2}^3 \frac{t(t-1)(t-1/2)}{6} + f^{(4)}(\xi) h^4 \frac{t(t^2-1)(t-2)}{24}.$$

Так как $f_{q+1/2}^3 = h^3 f^{(3)}(\xi)$, то для допустимости интерполяции по формуле (12) достаточно выполнения соотношения

$$\begin{aligned} \max |f^{(3)}(\xi)| h^3 \max_{0 \leq t \leq 1} \left| \frac{t(t-1)(t-1/2)}{6} \right| + \\ + \max |f^{(4)}(\xi)| h^4 \max_{0 \leq t \leq 1} \left| \frac{t(t^2-1)(t-2)}{24} \right| \leq \varepsilon. \end{aligned}$$

Поскольку

$$\max_{0 \leq t \leq 1} \left| \frac{t(t-1)(t-1/2)}{6} \right| = \frac{1}{72\sqrt{3}}, \quad \max_{0 \leq t \leq 1} \left| \frac{t(t^2-1)(t-2)}{24} \right| = \frac{3}{128},$$

то интерполяция по формуле (12) допустима, если

$$\frac{\max |f^{(3)}(\xi)| h^3}{72\sqrt{3}} + \frac{3 \max |f^{(4)}(\xi)|}{128} h^4 \leq \varepsilon. \quad (13)$$

При малых h главной частью является первое слагаемое; оно меньше, чем левая часть (11) в $9\sqrt{3}/2 = 7,794 \dots$ раз. Следовательно, при малых h для выполнимости (13) можно взять шаг в $\sqrt{9\sqrt{3}/2} \approx 1,98$ раз больше, чем для выполнимости (11). В рассматриваемом примере условие (13) имеет вид

$$h^3/72\sqrt{3} + 3h^4/128 \leq 0,5 \cdot 10^{-6}.$$

Решая это неравенство, получим $h \leq h_0 = 0,038\dots$ Заметим, что при ручном счете шаг $h = 0,038$ неудобен вследствие «некруглости» этого числа. Поэтому при составлении таблиц его заменили бы заведомо на меньшее, но «более круглое» число 0,03.

Если и такой шаг таблицы слишком велик, можно пойти по пути увеличения порядка интерполяционного многочлена.

Сейчас мы рассмотрим еще один способ записи интерполяционного многочлена, носящий название *способа Эверетта* и являющийся весьма удобным при решении следующей задачи уплотнения таблиц: дана таблица значений функции с узлами $x_0 + nh$, требуется составить таблицу значений функции с узлами $x_0 + nh'$ при $h' = h/l$, где l — целое. Последующие построения, связанные с формулами Эверетта, непосредственно вряд ли понадобятся читателю. Однако они любопытны в качестве примера частных приемов и частных вариационных задач, помогающих усилить эффективность вычислительных алгоритмов в конкретной обстановке.

Обратимся к интерполяционной формуле Гаусса (2), исключив в этой формуле разности нечетного порядка при помощи соотношения

$$f_{1/2}^{2l+1} = f_1^{2l} - f_0^{2l}.$$

После приведения подобных членов получится формула Эверетта

$$E_{2n+2}(x_0 + ht) = G_{2n+2}(x_0 + ht) = S_0(u) + S_1(t); \quad (14)$$

здесь $u = 1 - t$

$$S_q(t) = f_q t + f_q^2 \frac{t(t^2 - 1)}{3!} + \dots + f_q^{2n} \frac{t(t^2 - 1) \dots (t^2 - n^2)}{(2n+1)!}. \quad (15)$$

Еще раз подчеркнем, что все эти формулы дают представление одного и того же интерполяционного многочлена. Остаточный член этой формулы есть остаточный член формулы Лагранжа, т. е.

$$f^{(2n+2)}(\xi) h^{2n+2} \frac{t(t^2 - 1) \dots (t^2 - n^2)(t - (n+1))}{(2n+2)!}. \quad (16)$$

Запись интерполяционного многочлена в виде (14) примерно вдвое сокращает работу по сравнению с другими записями при решении задачи уплотнения таблиц: при использовании формулы Эверетта значения $f(x_0 - ht)$ при $0 < t < 1$ вычисляются по формуле

$$f(x_0 - ht) = S_0(u) + S_{-1}(t);$$

таким образом, значение $S_0(u)$ используется при нахождении двух значений $f(x_0 \pm ht)$. Мы можем вычислить все значения $S_q(u)$ при $u = 1/l, \dots, (l-1)/l$ и составить значения $f(x)$ в узлах сгущенной сетки $x_0 + nh'$ как их линейные комбинации. Непосредственное вычисление значений интер-

полюционного многочлена из других форм записи требует суммирования $2n+2$ слагаемых по сравнению с $n+1$ слагаемым в (15).

Проведем дополнительные построения в случае $n=2$, когда

$$S_q(t) = f_q t + f_q^2 \frac{t(t^2-1)}{6} + f_q^4 \frac{t(t^2-1)(t^2-4)}{120}.$$

При ручном счете целесообразно приблизить коэффициент при f_q^4 выражением $-k \frac{t(t^2-1)}{120}$ и вместо $S_q(t)$ вычислять

$$\bar{S}_q(t) = f_q t + \left(f_q^2 - \frac{k}{20} f_q^4 \right) \frac{t(t^2-1)}{6}.$$

Можно представить пользователю таблицы, где, наряду со столбцом значений f_q , содержался бы столбец значений $f_q^2 - \frac{k}{20} f_q^4$. Вычисление значений функции при помощи таких таблиц по формуле

$$f(x_q + ht) \approx \bar{S}_{q+1}(t) + \bar{S}_q(u) = \bar{E}_6(x_q + ht)$$

было бы весьма простым. Поставим перед собой задачу наиболее разумного выбора параметра k . Мы будем формулировать эту задачу как задачу выбора k , при котором главная часть

$$\sup_{t \in [0, 1]} |E_6(x_q + ht) - \bar{E}_6(x_q + ht)|$$

минимальна. Имеем равенство

$$\begin{aligned} E_6(x_q + ht) - \bar{E}_6(x_q + ht) &= \\ &= f_{q+1}^4 \frac{t(t^2-1)(t^2-4+k)}{120} + f_q^4 \frac{u(u^2-1)(u^2-4+k)}{120}. \end{aligned} \quad (17)$$

На первый взгляд может показаться разумным выбрать k из условия

$$\min_k \max_{t \in [0, 1]} |t(t^2-1)(t^2-4+k)|.$$

Однако, вследствие близости f_{q+1}^4 и f_q^4 , слагаемые в правой части (17) могут компенсировать друг друга, и такой выбор k не будет наилучшим. Поэтому предварительно выделим главный член в правой части (17). Подставим в правую часть (17)

$$f_{q+1}^4 = \mu f_{q+1/2}^4 + 1/2 f_{q+1/2}^5, \quad f_q^4 = \mu f_{q+1/2}^4 - 1/2 f_{q+1/2}^5, \quad u = 1-t.$$

После приведения подобных членов получим

$$E_6(x_q + ht) - \bar{E}_6(x_q + ht) = s_1 + s_2,$$

где

$$s_1 = \mu f_{q+1/2}^4 \frac{5t(t^2-1)(t-2) + 3kt(t-1)}{120}, \quad (18)$$

$$s_2 = f_{q+1/2}^5 \frac{t(t^2-1)(t-2)(t-1/2) + kt(t-1)(t-1/2)}{120}. \quad (19)$$

При $f^{(4)}(x) \neq 0$ s_1 имеет порядок h^4 , и поэтому является главной частью разности $E_6(x_q + ht) - \bar{E}_6(x_q + ht)$. Наиболее естественно выбрать k

из условия минимума коэффициента в s_1

$$\min_k \max_{[0, 1]} |5t(t^2 - 1)(t - 2) + 3kt(t - 1)|.$$

Справедливо равенство

$$t(t^2 - 1)(t - 2) = t(t - 1)(t(t - 1) - 2).$$

Поэтому после введения новой переменной $y = 4t(1 - t)$ получим

$$5t(t^2 - 1)(t - 2) + 3kt(t - 1) = \frac{5}{16}y^2 + \frac{5}{2}y - \frac{3}{4}ky.$$

Поскольку при изменении t на отрезке $[0, 1]$ y пробегает (дважды) отрезок $[0, 1]$, то

$$\begin{aligned} \min_k \max_{[0, 1]} |5t(t^2 - 1)(t - 2) + 3kt(t - 1)| &= \\ &= \min_k \max_{[0, 1]} \left| \frac{5}{16}y^2 + \frac{5}{2}y - \frac{3}{4}ky \right| = \frac{5}{16} \min_r \max_{[0, 1]} |y^2 - ry|, \end{aligned} \quad (20)$$

где $r = \frac{12}{5}k - 8$.

Обратимся к рис. 2.12.2. При $0 < r < 1$ функция $y^2 - ry$ на $[0, 1]$ имеет локальные экстремумы в точке 1 и в точке, где $(y^2 - ry)' = 0$, т. е. при $y = r/2$. Величина первого экстремума равна $1 - r$, второго $-r^2/4$. Определим $r_0 \in (0, 1)$ из условия, что значения функции $y^2 - ry$ в точках этих локальных экстремумов по модулю равны между собой, т. е. из уравнения $1 - r_0 = r_0^2/4$. Отсюда $r_0 = \sqrt{8} - 2$ и

$$\max_{[0, 1]} |y^2 - r_0 y| = 1 - r_0 = 3 - \sqrt{8} = \frac{1}{3 + \sqrt{8}} = b.$$

Если $r = r_1 < r_0$, то $\max_{[0, 1]} |y^2 - ry| \geq 1 - r > 1 - r_0 = b$; если $r = r_2 > r_0$, то

$$\max_{[0, 1]} |y^2 - ry| \geq \left| \left(\frac{r}{2} \right)^2 - \frac{r^2}{2} \right| = \frac{r^2}{4} > \frac{r_0^2}{4} = 1 - r_0 = b; \text{ справедливость этих}$$

соотношений также видна непосредственно из рисунка. Таким образом, мы нашли r , при котором достигается

$$\min_r \max_{[0, 1]} |y^2 - ry| = \frac{1}{3 + \sqrt{8}}.$$

Обращаясь к (20), получаем значение k :

$$k = 5(3 + \sqrt{2})/6 = 3,6785 \dots$$

и при этом k

$$\max_{[0, 1]} |5t(t^2 - 1)(t - 2) + 3kt(t - 1)| = \frac{5}{16(3 + \sqrt{8})}.$$

Из (18) получаем оценку

$$|s_1| \leq \frac{\max_{[0, 1]} |f^{(4)}(x)|}{384(3 + \sqrt{8})} h^4 = \frac{\max_{[0, 1]} |f^{(4)}(x)|}{2238,0 \dots} h^4.$$

Оценим величину $|s_2|$ при таком значении k :

$$|s_2| \leq \frac{\max_{[0, 1]} |f^{(5)}(x)|}{1646, \dots} h^5.$$

В то же время исходный остаточный член можно оценить через

$$\frac{\max |f^{(6)}(x)|}{204,8} h^6.$$

Суммируя все эти оценки, получаем

$$|f(x_q + ht) - \bar{E}_6(x_q + ht)| \leq \frac{\max |f^{(4)}(x)|}{2238,0 \dots} h^4 + \frac{\max |f^{(5)}(x)|}{1646, \dots} h^5 + \frac{\max |f^{(6)}(x)|}{204,8} h^6.$$

Отсюда можно получить, что для рассматриваемого нами примера $f(x) = \sin x$ погрешность интерполяции по этой формуле не будет превосходить $0,5 \cdot 10^{-6}$ при $h \leq h_0 = 0,16 \dots$. Таким образом, по сравнению со случаем линейной интерполяции, мы добились существенного увеличения шага, с которым допустимо составление таблиц при точности интерполяции $0,5 \cdot 10^{-6}$.

Может случиться, что не важна форма, в которой ищется приближение для суммы слагаемых, содержащих четвертые разности, например, если в дальнейшем мы будем исследовать свойства функции, находить ее экстремумы и т. д. Для этих целей можно поставить перед собой задачу приблизить эти слагаемые выражением $f_{q+1}^4 Q_3(t) + f_q^4 Q_3(u)$ так, чтобы максимум главного члена погрешности был минимальным.

Задача. Показать, что решением упомянутой выше задачи будет

$$Q_3(t) = \frac{960t(1-t^2) - 5}{30720};$$

показать, что при этом главный член погрешности не превосходит

$$\frac{\max |f^{(4)}(x)|}{3072} h^4.$$

§ 13. О погрешности округления при интерполировании

Предположим, что выбран некоторый способ интерполяции. Выше мы получили достаточное представление о погрешности, являющейся следствием замены функции многочленом. Однако существует еще одна причина погрешности: задаваемые значения функции содержат погрешности, в частности, вследствие округления этих значений. Пусть требуется вычислить значение

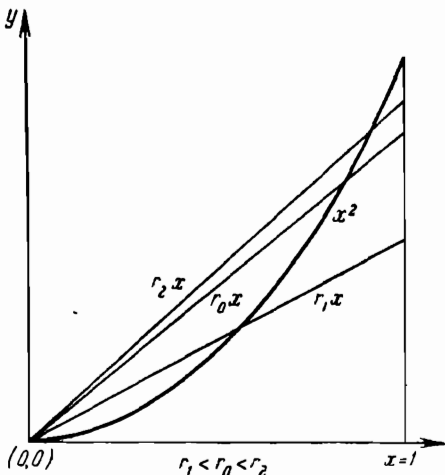


Рис. 2.12.2.

$f(x)$ по формуле

$$f(x) \approx \sum_{j=1}^n f_j P_j(x),$$

являющейся общим видом рассматриваемых нами интерполяционных формул. Поскольку реально заданы не f_j , а приближенные значения $f_j^* = f_j + \eta_j$, то в результате будет получено значение

$$f^* = \sum_{j=1}^n f_j P_j(x) + \sum_{j=1}^n \eta_j P_j(x).$$

Если известны границы изменения значений η_j , то можно оценить верхнюю грань погрешности

$$\varepsilon = \sum_{j=1}^n \eta_j P_j(x).$$

Например, при условии $|\eta_j| \leq \eta$ имеем оценку

$$|\varepsilon| \leq \eta \sum_{j=1}^n |P_j(x)|.$$

Для широкого употребления иногда выпускаются p -значные (все значения функции содержат p десятичных знаков после запятой) таблицы, удовлетворяющие следующим условиям: таблица допускает линейную интерполяцию с погрешностью $0,5 \cdot 10^{-p}$, все значения функции заданы с погрешностью не более $0,5 \cdot 10^{-p}$. Если мы вычисляем значение $f(x)$ интерполяцией по узлам x_0, x_1 , где $x_0 \leq x < x_1$, то

$$P_0(x) = \frac{x_1 - x}{x_1 - x_0}, \quad P_1(x) = \frac{x - x_0}{x_1 - x_0}$$

и

$$\eta(|P_0(x)| + |P_1(x)|) = \eta.$$

Таким образом, при линейной интерполяции погрешность, вследствие округления значений функций, не превосходит погрешности этих значений. С учетом того, что погрешность интерполяции не превосходит $0,5 \cdot 10^{-p}$, получаем, что погрешность получаемых приближенных значений не превосходит 10^{-p} , если только при интерполяции не вносится дополнительной погрешности.

Составление p -значных таблиц с погрешностями значений функции, не превосходящими $0,5 \cdot 10^{-p}$, является очень трудоемким делом. Пусть, например, точное значение $f(x_1)$ есть

$$0, y_1 \dots y_p \underbrace{499 \dots 9}_{k \text{ раз}}; \quad (1)$$

здесь $y_1, \dots, y_p$ — десятичные цифры. Мы должны взять в качестве f_1 значение $0, y_1 \dots y_p$. Однако если мы при вычислении $f(x_1)$ совершим ошибку $2 \cdot 10^{-k-p-1}$, то мы возьмем в качестве f_1 значение $0, y_1 \dots y_p + 10^{-p}$, т. е. ошибка будет больше $0,5 \cdot 10^{-p}$. Поскольку k может оказаться сколь угодно

большим, то может потребоваться проводить вычисления значений $f(x_i)$ с очень большим числом знаков. В большинстве случаев p -значные таблицы или содержат значения функций с погрешностью, не большей 10^{-p} , или в описании таблиц говорится, что погрешность не превосходит $0,5 \cdot 10^{-p}$, но на самом деле допускается погрешность $0,5 \cdot 10^{-p}(1 + \gamma)$, где γ — достаточно малое. Источником погрешности значений функции является также погрешность от округления при вычислении по интерполяционным формулам. Естественно, что эта погрешность зависит от числа значащих цифр, с которыми производятся вычисления. Как и выше, никакое фиксированное количество знаков при вычислениях не позволяет гарантировать, что ошибка, являющаяся следствием округлений при вычислениях, меньше $0,5 \cdot 10^{-p}$. В самом деле, если значение интерполяционного многочлена при отсутствии округлений есть (1), то при ошибке от округления $2 \cdot 10^{-k-p-1}$ округленное значение может иметь ошибку, несколько большую $0,5 \cdot 10^{-p}$. Таким образом, даже при интерполяции p -значной таблицы, удовлетворяющей наложенным выше самым жестким условиям, погрешность округленных с p знаками после запятой значений функций может превосходить $1,5 \cdot 10^{-p}$.

Применение вычислительных машин и автоматизация способов печати таблиц делают не столь существенным рассмотрение затронутых выше вопросов. Если мы возьмем старые руководства по методам вычислений, то найдем там существенно большее, чем у нас, количество форм записей интерполяционных многочленов. Дело заключается в следующем: при вычислениях на большинстве современных ЭВМ каждая операция осуществляется с большим числом разрядов в фиксированное время, не зависящее от числа значащих цифр чисел; вероятность сбоя машины также не зависит от этого числа. Поэтому при той же погрешности метода и округления качество интерполяционной формулы определяется лишь числом различных арифметических и логических операций при ее реализации. При ручном счете время осуществления операций и вероятность сбоя (ошибка человека, осуществляющего вычисления) растут с числом значащих цифр, с которыми производятся вычисления. Появление в свое время большого числа записей одной и той же интерполяционной формулы Лагранжа в существенной степени связано со стремлением получать формулы, в которых участвуют числа возможно меньшей разрядности.

§ 14. Применения аппарата интерполирования.

Обратная интерполяция

Употребление интерполяционных многочленов оказывается полезным при решении ряда задач.

Пусть, например, требуется найти значение экстремума функции одной переменной и точку экстремума. Составим таблицу значений функции с крупным шагом. Из рассмотрения этой таблицы можно увидеть место расположения экстремума. Составим таблицу с более мелким шагом в этом месте. Из рассмотрения этой таблицы можно уточнить расположение экстремума и т. д. Такое последовательное уплотнение таблицы не следует производить неограниченно. Обычно нас интересует значение экстремума с некоторой точностью γ . При некотором шаге таблица допускает кубическую интерполяцию, например по формулам из § 12, с точностью γ . Заменяем функцию интерполяционным многочленом, продифференцируем его и найдем корень. Значение функции в этом корне примем за экстремальное

значение. Поскольку $|f(x) - P_3(x)| \leq \gamma$, то $|\max f(x) - \max P_3(x)| \leq \gamma$. В зависимости от обстановки иногда бывает удобнее мельчить шаг таблицы до тех пор, пока не станет допустимой квадратичная интерполяция.

Другой типичной задачей, где может быть применен аппарат интерполирования, является нахождение корня X уравнения $f(x) = d$.

Путь решения этой задачи тот же самый. Составляем таблицу значений функции; определяем по ней грубо, где находится корень уравнения, затем составляем таблицу с более мелким шагом и т. д.

Если вычисление функции относительно нетрудоемко, неразумно применять в процессе вычислений интерполяцию степени выше второй; в противном случае возникает задача нахождения корней многочленов, сама требующая достаточно большого числа арифметических операций. Если вычисление функции трудоемко, может оказаться более выгодным пойти по пути увеличения степени интерполяционного многочлена.

В случае, когда в окрестности $y = d$ функция $g(y)$, обратная к $f(x)$, является достаточно гладкой, более эффективным явится применение обратной интерполяции. *Обратной интерполяцией* называется следующий алгоритм. Пусть известны значения функции $y_i = f(x_i)$ при $i = 1, \dots, n$. Эта информация эквивалентна тому, что известны значения $x_i = g(y_i)$ обратной функции. При условии допустимости интерполяции по переменной y можно заменить обратную функцию $g(y)$ интерполяционным многочленом $L_n(y)$, удовлетворяющим условиям

$$L_n(y_i) = x_i, \quad i = 1, \dots, n,$$

и положить $X = g(d) \approx L_n(d)$. Такой способ особенно удобен, если нас интересуют значения решения уравнения при достаточно большом числе значений d или если желательно получение явного выражения корня уравнения $f(x) = d$ в зависимости от параметра d . Если интерполяция по узлам $y_1, \dots, y_n$ не обеспечивает нужной точности, полагаем $x_{n+1} = L_n(d)$, $y_{n+1} = f(x_{n+1})$; далее, в зависимости от обстановки, целесообразно заменить $g(d)$ значением интерполяционного многочлена по всем узлам $y_1, \dots, y_{n+1}$ или по узлам, ближайшим к d .

§ 15. Ортогональные системы и их свойства

Пусть H — гильбертово пространство со скалярным произведением (f, g) и, соответственно, нормой $\|f\| = \sqrt{(f, f)}$.

Наиболее важные следствия рассмотрений настоящего параграфа относятся к случаям:

А) H — пространство комплекснозначных функций, определенных на $[a, b]$ с ограниченным интегралом

$$\int_a^b p(x) |f(x)|^2 dx,$$

скалярное произведение задается равенством

$$(f, g) = \int_a^b p(x) f(x) \bar{g}(x) dx; \quad (1)$$

$p > 0$ почти всюду на $[a, b]$; функции, отличающиеся друг от друга на множестве меры 0, считаются равными.

Б) H — евклидово пространство n -мерных комплекснозначных векторов $(x_1, \dots, x_n)$ со скалярным произведением

$$(x, y) = \sum_{j=1}^n p_{jj} x_j \bar{y}_j, \quad (2)$$

$P = [p_{ij}]$ — положительно определенная матрица.

Пусть задана система элементов из H : $\{\varphi_j\}_n = \varphi_1, \dots, \varphi_n$. Система $\{\varphi_j\}_n$ называется ортогональной, если $(\varphi_i, \varphi_j) = 0$ при $i \neq j$. Система $\{\varphi_j\}_n$ называется линейно независимой, если

$$\sum_{j=1}^n C_j \varphi_j = 0$$

только тогда, когда все $C_j = 0$.

Важным аппаратом многих исследований является ортогонализация заданной системы элементов гильбертова пространства.

Лемма 1. Пусть задана линейно независимая система элементов $\varphi_1, \dots, \varphi_n$. Можно построить ортогональную линейно независимую систему элементов

$$\psi_j = \sum_{i=1}^j b_{ji} \varphi_i, \quad j = 1, \dots, n, \quad \text{где все } b_{jj} = 1. \quad (3)$$

Мы будем проводить построение такой системы методом индукции. При $n = 1$ имеем тривиальную систему $\psi_1 = \varphi_1$. Пусть требуемая система $\{\psi_j\}_n$ построена при некотором $n = k$, тогда ψ_{k+1} отыскиваем в виде

$$\psi_{k+1} = \varphi_{k+1} - \sum_{i=1}^k a_{ki} \psi_i. \quad (4)$$

Коэффициенты a_{ki} выбираем из условия ортогональности $(\psi_{k+1}, \psi_i) = 0$ при $i \leq k$. Вследствие ортогональности системы

элементов $\{\psi_j\}_k$ последнее соотношение представится в виде

$$(\varphi_{k+1}, \psi_i) - a_{ki}(\psi_i, \psi_i) = 0,$$

откуда

$$a_{ki} = \frac{(\varphi_{k+1}, \psi_i)}{(\psi_i, \psi_i)},$$

следовательно, элемент

$$\psi_{k+1} = \varphi_{k+1} - \sum_{i=1}^k \frac{(\varphi_{k+1}, \psi_i)}{(\psi_i, \psi_i)} \psi_i \quad (5)$$

будет ортогонален всем предшествующим. Подставляя сюда

$$\psi_i = \sum_{p=1}^i b_{ip} \varphi_p$$

при $i \leq k$, получим требуемое соотношение. Совокупность соотношений (3) при $j \leq n$ можно представить в виде

$$\psi_{(n)} = B_n \varphi_{(n)},$$

где

$$\psi_{(n)} = \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_n \end{pmatrix}, \quad \varphi_{(n)} = \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_n \end{pmatrix}, \quad B_n = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ b_{21} & 1 & 0 & 0 & \dots \\ b_{31} & b_{32} & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}.$$

В то же время, перенося все ψ_i из правой части (5) в левую, получим

$$\varphi_{(n)} = A_n \psi_{(n)}, \quad \text{где} \quad A_n = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ a_{21} & 1 & 0 & 0 & \dots \\ a_{31} & a_{32} & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}; \quad (6)$$

матрицу B_n иногда называют *матрицей ортогонализации*. Так как $\det B_n = 1$, преобразование, задаваемое матрицей B_n , является невырожденным и поэтому переводит линейно независимую систему элементов $\{\varphi_j\}_n$ в линейно независимую систему $\{\psi_j\}_n$.

В силу линейной независимости системы $\{\varphi_j\}_n$ отсюда следует, что $B_n = A_n^{-1}$.

Важным аппаратом математического анализа являются системы ортогональных многочленов. В качестве системы функций $\{\varphi_j\}$ берется система $1, \dots, x^{j+1}, \dots$, производится ортогонализация в пространстве со скалярным произведением (1) по

описанной выше процедуре. Получаемые многочлены

$$\psi_l(x) = x^{l-1} + \sum_{i=1}^{l-1} b_{li} x^{i-1} \quad (7)$$

называют *ортogonalными многочленами*, соответствующими весу $\rho(x)$ и отрезку $[a, b]$. Иногда ортогональными многочленами, соответствующими весу $\rho(x)$, называют многочлены $\Psi_j(x) = \alpha_j \psi_j(x)$, где величины α_j подбирают из каких-либо дополнительных соотношений, например, из условия $\|\Psi_j\| = 1$. Систему ортогональных элементов, удовлетворяющих такому условию, называют *ортонормированной*.

Мы уже имели дело с системой многочленов

$$T_n(x) = 2^{n-1} x^n + \dots,$$

ортogonalных на отрезке $[-1, 1]$ с весом $2/(\pi \sqrt{1-x^2})$: Как отмечалось, значения этих многочленов можно вычислять по рекуррентной формуле

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \quad (8)$$

Вычисление значений ортогональных многочленов Чебышева при помощи формулы (8) более предпочтительно по сравнению с непосредственным вычислением их по явной формуле (7) по следующим причинам:

1. Вычисление по формуле (8) не требует хранения в памяти или вычисления коэффициентов b_{ji} .

2. Обычно требуется вычислять одновременно значения всех многочленов $\psi_1(x), \dots, \psi_n(x)$ в одной и той же точке. При независимом вычислении значения каждого многочлена по формуле (7) вычисление значений всех многочленов потребует $\asymp n^2$ арифметических операций.

Здесь и далее $a(n) \asymp b(n)$ означает, что $a(n)$ и $b(n)$ одного порядка, т. е.

$$a(n) = O(b(n)), \quad b(n) = O(a(n)).$$

При одновременном вычислении всех значений при помощи рекуррентного соотношения (8) потребуется $O(n)$ арифметических операций.

3. Значения $T_n(x)$, получаемые при непосредственном вычислении по формуле (7), могут содержать большую вычислительную погрешность. Дело заключается в следующем: число $T_n(x)$ образуется как сумма слагаемых:

$$T_n(x) = \sum_{l=0}^n d_{nl} x^l. \quad (9)$$

При записи в машине эти слагаемые $d_{nj}x^j$ смогут приобрести абсолютную погрешность порядка $|d_{nj}x^j|2^{-t}$; следствием этого может быть погрешность значения $T_n(x)$ порядка $D_n(x)2^{-t}$, где

$$D_n(x) = \sum_{j=0}^n |d_{nj}x^j|.$$

Оценим снизу $D_n(x)$; из равенства

$$T_n(|x|i) = \sum_{j=0}^n d_{nj}(|x|i)^j$$

следует оценка

$$|T_n(|x|i)| \leq \sum_{j=0}^n |d_{nj}x^j| = D_n(x).$$

В то же время, согласно (8.4), при действительном x имеем

$$T_n(|x|i) = \frac{(|x| + \sqrt{1+|x|^2})^n + (|x| - \sqrt{1+|x|^2})^n}{2} i^n;$$

так как

$$(|x| + \sqrt{1+|x|^2})(|x| - \sqrt{1+|x|^2}) = -1,$$

то отсюда получаем

$$D_n(x) \geq |T_n(|x|i)| \geq \frac{(|x| + \sqrt{1+|x|^2})^n - (|x| - \sqrt{1+|x|^2})^n}{2}.$$

Таким образом, при больших n и $x \neq 0$ при непосредственном использовании формулы (7) вычислительная погрешность может достигать значений порядка

$$(|x| + \sqrt{1+|x|^2})^n 2^{-t}.$$

Задача. Доказать равенство

$$D_n(x) = |T_n(|x|i)|.$$

Мы столкнулись здесь опять с явлением пропадания значащих цифр: $|T_n(x)| \leq 1$, но при вычислении значения $T_n(x)$ из (9) оно получается как сумма больших по модулю слагаемых переменного знака и поэтому приобретает большую погрешность.

В то же время можно показать, что при вычислении по рекуррентной формуле (8) погрешность $T_n(x)$ имеет порядок $O(n^2 2^{-t})$. Из изложенного видна важность получения рекуррентных соотношений типа (8), связывающих значения ортогональных многочленов, соответствующих другим весовым функциям $\rho(x)$.

Мы рассмотрим некоторые более общие образования. Пусть L — самосопряженный оператор над гильбертовым пространством H : $\varphi_i \in H$, $\varphi_j = L^{j-1}\varphi_1$; предположим, что элементы $\varphi_1, \dots, \varphi_n$ линейно независимы. Согласно лемме, можно построить систему ортогональных элементов

$$\psi_j = \sum_{i=1}^j b_{ji} \varphi_i,$$

где все $b_{jj} = 1$.

Теорема. Элементы ψ_j удовлетворяют системе соотношений

$$\psi_{j+1} + \alpha_{j+1, j} \psi_j + \alpha_{j+1, j-1} \psi_{j-1} - L\psi_j = 0, \quad (10)$$

$$\alpha_{j+1, j-1} > 0.$$

Доказательство. Справедливо равенство

$$L\psi_j = \sum_{i=1}^j b_{ji} L\varphi_i = \sum_{i=1}^j b_{ji} \varphi_{i+1}.$$

Вследствие соотношения (6): $\varphi_{(j+1)} = A_{j+1} \psi_{(j+1)}$ можно написать

$$L\psi_j = \sum_{i=1}^{j+1} \alpha_{j+1, i} \psi_i. \quad (11)$$

Заметим, что $\alpha_{j+1, j+1} = 1$ вследствие (3), (6). Умножая скалярно обе части (11) на ψ_k при $k \leq j+1$, получим

$$(\psi_j, L\psi_k) = (L\psi_j, \psi_k) = \sum_{i=1}^{j+1} \alpha_{j+1, i} (\psi_i, \psi_k) = \alpha_{j+1, k} (\psi_k, \psi_k). \quad (12)$$

Выразим $L\psi_k$ из (11) и подставим в левую часть последнего равенства

$$\left(\psi_j, \sum_{i=1}^{k+1} \alpha_{k+1, i} \psi_i \right) = \alpha_{j+1, k} (\psi_k, \psi_k).$$

При $k+1 < j$ левая часть обращается в нуль, и поэтому $\alpha_{j+1, k} = 0$ для $k < j-1$. Таким образом, соотношение (11) может быть представлено в виде

$$\alpha_{j+1, j+1} \psi_{j+1} + \alpha_{j+1, j} \psi_j + \alpha_{j+1, j-1} \psi_{j-1} - L\psi_j = 0, \quad (13)$$

где, вследствие (12),

$$\alpha_{j+1, k} = (\psi_j, L\psi_k) / (\psi_k, \psi_k).$$

Мы имеем

$$(\psi_j, L\psi_{j+1}) / (\psi_{j+1}, \psi_{j+1}) = \alpha_{j+1, j+1} = 1,$$

т. е. $(L\psi_j, \psi_{j+1}) = (\psi_{j+1}, \psi_{j+1})$; поэтому

$$\alpha_{j+1, j-1} = (\psi_j, L\psi_{j-1}) / (\psi_{j-1}, \psi_{j-1}) = (\psi_j, \psi_j) / (\psi_{j-1}, \psi_{j-1}) > 0.$$

В то же время

$$\alpha_{j+1, j} = (L\psi_j, \psi_j) / (\psi_j, \psi_j).$$

Таким образом, рекуррентное соотношение (10) имеет вид

$$\psi_{j+1} + \frac{(L\psi_j, \psi_j)}{(\psi_j, \psi_j)} \psi_j + \frac{(\psi_j, \psi_j)}{(\psi_{j-1}, \psi_{j-1})} \psi_{j-1} - L\psi_j = 0. \quad (14)$$

Теорема доказана.

Далее наибольший интерес представит случай A при $\varphi_1 \equiv 1$ и L — операторе умножения на x . Тогда ψ_j — многочлен $P_{j-1}(x)$ степени $j-1$, и после переобозначения $\alpha_{jk} = d_{j-1, k-1}$, $j = n+1$ соотношение (14) приобретет вид

$$P_{n+1}(x) + d_{n+1, n} P_n(x) + d_{n+1, n-1} P_{n-1}(x) - x P_n(x) = 0, \quad (15)$$

$$d_{n+1, n} = \frac{\int_a^b x p(x) P_n^2(x) dx}{\int_a^b p(x) P_n^2(x) dx}, \quad d_{n+1, n-1} = -\frac{\int_a^b p(x) P_n^2(x) dx}{\int_a^b p(x) P_{n-1}^2(x) dx} > 0. \quad (16)$$

Лемма 2. Пусть $\{\psi_j\}_n$ — система элементов

$$\psi_j = \sum_{i=1}^l b_{ji} \varphi_i \quad \text{при} \quad b_{jj} = 1,$$

построенная методом последовательной ортогонализации системы $\{\varphi_j\}_n$, а $\{z_j\}_n$ — некоторая ортогональная система элементов

$$z_j = \sum_{i=1}^l d_{ji} \varphi_i, \quad \text{причем все} \quad d_{jj} = 1. \quad (17)$$

Тогда $z_j = \psi_j$ при всех $j \leq n$.

Доказательство. При $j=1$ утверждение справедливо; пусть оно выполняется при всех $j \leq k$. Из (17) следует, что

$$z_{k+1} = \varphi_{k+1} - \sum_{i=1}^k \tilde{a}_{k+1, i} z_i.$$

Умножая на z_i , получим

$$0 = (\varphi_{k+1}, z_i) - \tilde{a}_{k+1, i} (z_i, z_i);$$

отсюда

$$\tilde{a}_{k+1, i} = \frac{(\varphi_{k+1}, \psi_i)}{(\psi_i, \psi_i)}.$$

Возвращаясь к доказательству леммы 1, мы видим, что z_{k+1} также определяется равенством (5). Таким образом, $z_{k+1} = \psi_{k+1}$, и единственность ортогональной системы рассматриваемого вида доказана.

§ 16. Ортогональные многочлены

Приведем наиболее употребительные системы ортогональных многочленов, соответствующие различным весовым функциям.

1. *Многочлены Якоби*. При отрезке $[-1, 1]$ и весовой функции

$$\rho(x) = (1-x)^\alpha (1+x)^\beta, \quad \alpha, \beta > -1,$$

ортогональную систему образуют *многочлены Якоби*

$$P_n^{(\alpha, \beta)}(x) = \frac{(-1)^n}{2^n n!} (1-x)^{-\alpha} (1+x)^{-\beta} \frac{d^n}{dx^n} ((1-x)^{\alpha+n} (1+x)^{\beta+n}).$$

Имеют место соотношения

$$\|P_n^{(\alpha, \beta)}(x)\| = \left(\frac{2^{\alpha+\beta+1} \Gamma(\alpha+n+1) \Gamma(\beta+n+1)}{n! (\alpha+\beta+2n+1) \Gamma(\alpha+\beta+n+1)} \right)^{1/2}$$

$$\begin{aligned} (\alpha+\beta+2n)(\alpha+\beta+2n+1)(\alpha+\beta+2n+2) x P_n^{(\alpha, \beta)}(x) = \\ = 2(n+1)(\alpha+\beta+n+1)(\alpha+\beta+2n) P_{n+1}^{(\alpha, \beta)}(x) + \\ + (\beta^2 - \alpha^2)(\alpha+\beta+2n+1) P_n^{(\alpha, \beta)}(x) + \\ + 2(\alpha+n)(\beta+n)(\alpha+\beta+2n+2) P_{n-1}^{(\alpha, \beta)}(x); \quad (1) \end{aligned}$$

здесь Γ — гамма-функция Эйлера.

Многочлены Якоби удовлетворяют дифференциальному уравнению

$$\begin{aligned} L_{\alpha, \beta}(P_n^{(\alpha, \beta)}(x)) = (1-x^2)(P_n^{(\alpha, \beta)}(x))_{xx} + \\ + ((\beta-\alpha) - (\alpha+\beta+2)x)(P_n^{(\alpha, \beta)}(x))_x = -n(\alpha+\beta+n+1)P_n^{(\alpha, \beta)}(x). \end{aligned}$$

иначе говоря, они являются собственными функциями дифференциального оператора $L_{\alpha, \beta}$.

2. *Многочлены Лежандра*. Частным случаем многочленов Якоби при $\alpha = \beta = 0$, т. е. при весовой функции $\rho(x) \equiv 1$, являются *многочлены Лежандра*

$$L_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n$$

с нормой

$$\|L_n\| = \sqrt{2/(2n+1)}$$

и рекуррентным соотношением

$$(n+1)L_{n+1}(x) - (2n+1)xL_n(x) + nL_{n-1}(x) = 0.$$

Заметим, что при больших n это соотношение, так же как и (1) при любых α и β , близко к рекуррентному соотношению для многочленов Чебышева.

3. *Многочлены Чебышева*. При $\alpha = \beta = -1/2$, $\rho(x) = 1/\sqrt{1-x^2}$ многочлены Якоби превращаются в *многочлены Чебышева первого рода* $T_n(x)$.

4. *Многочлены Чебышева второго рода*. При $\alpha = \beta = 1/2$, $\rho(x) = \sqrt{1-x^2}$, многочлены Якоби превращаются в *многочлены Чебышева второго рода*

$$U_n(x) = (\sin((n+1)\arccos x))/\sqrt{1-x^2} = T'_{n+1}(x)/(n+1), \quad n = 0, 1, \dots,$$

с нормой $\|U_n\| = \pi/2$. Рекуррентное соотношение для многочленов Чебышева второго рода такое же, как для многочленов Чебышева первого рода.

5. *Многочлены Эрмита*. При $(a, b) = (-\infty, \infty)$ и $\rho(x) = \exp(-x^2)$ ортогональную систему образуют *многочлены Эрмита*

$$H_n(x) = (-1)^n \exp(x^2) \frac{d^n}{dx^n} (\exp(-x^2))$$

с нормой

$$\|H_n\| = \sqrt{2^n \cdot n!} \sqrt{\pi}$$

и рекуррентным соотношением

$$H_{n+1}(x) - 2xH_n(x) + 2nH_{n-1}(x) = 0.$$

Многочлены Эрмита удовлетворяют дифференциальному уравнению

$$H_n'' - 2xH_n' = -2nH_n.$$

6. *Многочлены Лагерра*. При $[a, b] = [0, \infty)$ и $\rho(x) = x^\alpha \exp(-x)$ ортогональную систему образуют *многочлены Лагерра*

$$L_n^{(\alpha)}(x) = (-1)^n x^{-\alpha} \exp x \frac{d^n}{dx^n} (x^{\alpha+n} \exp(-x))$$

с нормой

$$\|L_n^{(\alpha)}\| = \sqrt{n! \Gamma(\alpha + n + 1)}$$

и рекуррентным соотношением

$$L_{n+1}^{(\alpha)}(x) - (x + \alpha + 2n + 1)L_n^{(\alpha)}(x) + n(\alpha + n)L_{n-1}^{(\alpha)}(x) = 0.$$

Многочлены Лагерра удовлетворяют дифференциальному уравнению

$$x(L_n^{(\alpha)}(x))_{xx} + (\alpha + 1 - x)(L_n^{(\alpha)}(x))_x = -nL_n^{(\alpha)}(x).$$

Установим ряд свойств ортогональных многочленов. Пусть $P_0(x), \dots, P_n(x), \dots$ — система ортогональных многочленов на отрезке $[a, b]$:

$$P_n(x) = x^n + \sum_{j=0}^{n-1} p_{nj}x^j.$$

Лемма 1. *Каждый многочлен $P_n(x)$ имеет ровно n различных нулей на открытом интервале (a, b) .*

Доказательство. Предположим, что $P_n(x)$ имеет на (a, b) только $l < n$ нулей $x_1, \dots, x_l$ нечетной кратности. Тогда многочлен

$$P_n(x) \prod_{j=1}^l (x - x_j),$$

не меняет знак на $[a, b]$; поэтому

$$\int_a^b \rho(x) P_n(x) \prod_{j=1}^l (x - x_j) dx \neq 0.$$

С другой стороны, этот интеграл равен нулю, поскольку $P_n(x)$ ортогонален ко всем многочленам меньшей степени. Получили противоречие.

Лемма 2. Пусть $x_1^{(n)} < \dots < x_n^{(n)}$ — нули $P_n(x)$. Тогда нули многочленов $P_{n-1}(x)$ и $P_n(x)$ перемежаются, т. е.:

$$a < x_1^{(n)} < x_1^{(n-1)} < \dots < x_{n-1}^{(n-1)} < x_n^{(n)} < b. \quad (2)$$

Доказательство. Пусть это утверждение доказано для некоторого n . Подставим в (15.15) $x = x_i^{(n)}$; получим

$$P_{n+1}(x_i^{(n)}) + d_{n+1, n-1} P_{n-1}(x_i^{(n)}) = 0.$$

Поскольку $d_{n+1, n-1} > 0$, то

$$\text{sign } P_{n+1}(x_i^{(n)}) = -\text{sign } P_{n-1}(x_i^{(n)}). \quad (3)$$

Так как все нули многочлена $P_{n-1}(x) = x^{n-1} + \dots$ лежат на (a, b) , то

$$\text{sign } P_{n-1}(b) = 1, \quad \text{sign } P_{n-1}(a) = (-1)^{n-1}.$$

Отсюда и из (2) следует, что

$$\text{sign } P_{n-1}(x_i^{(n)}) = (-1)^{n-i}.$$

Точно так же

$$\text{sign } P_{n+1}(b) = 1, \quad \text{sign } P_{n+1}(a) = (-1)^{n+1}.$$

Отсюда и из (3) получаем, что в последовательно расположенных точках $a, x_1^{(n)}, \dots, x_n^{(n)}, b$ многочлен $P_{n+1}(x)$ имеет чередующиеся знаки. Между каждыми двумя последовательными из этих точек многочлен $P_{n+1}(x)$ должен иметь корень; в итоге доказана справедливость (2) для следующего значения n . Индукцию можно начинать с $n = 1$, поскольку при $n = 1$ все выписанные соотношения, включая (3) выполняются. Лемма 2 доказана.

Рассмотрим практически важный случай, когда вес является четной функцией относительно середины отрезка $[a, b]$; пусть для определенности, $[a, b] = [-1, 1]$, соответственно, $\rho(x) = \rho(-x)$, тогда справедлива

Лемма 3. Все многочлены $P_{2l}(x)$ — четные, многочлены $P_{2l-1}(x)$ — нечетные, т. е.

$$P_j(-x) = (-1)^j P_j(x) \quad \text{при всех } j, \quad (4)$$

и рекуррентное соотношение (15.15) имеет вид

$$P_{j+1}(x) - xP_j(x) + d_{j+1, j-1} P_{j-1}(x) = 0. \quad (5)$$

Доказательство. Многочлены $P_0(x) = 1$, $P_1(x) = x$ ортогональны, вследствие четности $\rho(x)$

$$(P_0, P_1) = \int_{-1}^1 x \rho(x) dx = 0,$$

и поэтому являются многочленами ортогональной системы. Пусть мы уже доказали (4) при $j \leq n$ и (5) при $j < n$. Обращаясь к формуле (15.15), имеем

$$P_{n+1}(x) + (d_{n+1, n} - x) P_n(x) + d_{n+1, n-1} P_{n-1}(x) = 0.$$

Вследствие (4) и четности $\rho(x)$ функция $x P_n(x)$ — нечетная и поэтому, согласно (15.16), $d_{n+1, n} = 0$. Отсюда следует справедливость (5) при $j = n$. Функция $x P_n(x)$ той же четности, что и $P_{n-1}(x)$; поэтому и третье слагаемое $P_{n+1}(x)$, входящее в левую часть (5) при $j = n$, той же четности, что и $P_{n-1}(x)$. Таким образом, утверждение (4) доказано для следующего значения $j = n + 1$. Лемма 3 доказана.

Отсюда, в частности, следует, что в рассматриваемом случае нули ортогональных многочленов $P_n(x)$ расположены симметрично относительно начала координат. В частности, при n нечетном одним из нулей будет точка $x = 0$ вследствие нечетности $P_n(x)$.

Рекуррентные соотношения для конкретных ортогональных многочленов, выписанные в пп. 2—4, имеют несколько иной вид, чем (5), поскольку (5) выписано для ортогональных многочленов, нормированных так, что их старший коэффициент равен 1.

Отметим одно важное свойство распределения нулей ортогональных многочленов. Пусть $[a, b] = [-1, 1]$, вес $\rho(x)$ почти всюду положителен на $[-1, 1]$; обозначим через $\omega_n(x_1, x_2)$ число нулей многочлена $P_n(x)$, принадлежащих отрезку $[x_1, x_2]$. Тогда справедливо соотношение

$$\lim_{n \rightarrow \infty} \frac{\omega_n(x_1, x_2)}{n} = \int_{x_1}^{x_2} \frac{2}{\pi \sqrt{1-x^2}} dx. \quad (6)$$

Таким образом, нули ортогональных многочленов независимо от весовой функции $\rho(x)$ асимптотически распределены одинаково, с плотностью $\frac{2}{\pi \sqrt{1-x^2}}$.

§ 17. Численное дифференцирование

Простейшие формулы численного дифференцирования получаются в результате дифференцирования интерполяционных формул.

Пусть известны значения функции в точках $x_1, \dots, x_n$ и требуется вычислить производную $f^{(k)}(x)$. Построим интерполяционный многочлен $L_n(x)$ и положим $f^{(k)}(x) \approx L_n^{(k)}(x)$. Точно так же мы можем заменять значения производных функции значениями производных других многочленов интерполяционного типа: Бесселя, Эверетта и т. д.

Произведем оценку погрешности формул численного дифференцирования. Имеем равенство

$$f(x) - L_n(x) = f(x; x_1; \dots; x_n) \omega_n(x);$$

отсюда по формуле Лейбница

$$f^{(k)}(x) - L_n^{(k)}(x) = \sum_{j=0}^k C_k^j (f(x; x_1; \dots; x_n))^{(j)} \omega_n^{(k-j)}(x). \quad (1)$$

Пусть $g(x)$ непрерывно дифференцируема q раз. Согласно выражению (5.4) производной через разделенную разность, имеем

$$g(x; x + e; \dots; x + qe) = g^{(q)}(x_e)/q!,$$

где $x \leq x_e \leq x + qe$; отсюда

$$g^{(q)}(x) = q! \lim_{e \rightarrow 0} g(x; x + e; \dots; x + qe).$$

Следовательно,

$$(f(x; x_1; \dots; x_m))^{(q)} = q! [\lim_{e \rightarrow 0} f(x; x + e; \dots; x + qe; x_1; \dots; x_m)]. \quad (2)$$

Выражение в квадратных скобках по определению разделенной разности с кратными узлами есть

$$\underbrace{f(x; \dots; x; x_1; \dots; x_m)}_{q+1 \text{ раз}}.$$

Таким образом, соотношение (1) можно переписать в виде

$$f^{(k)}(x) - L_n^{(k)}(x) = \sum_{j=0}^k \frac{k!}{(k-j)!} \underbrace{f(x; \dots; x; x_1; \dots; x_n)}_{j+1 \text{ раз}} \omega_n^{(k-j)}(x). \quad (3)$$

Мы получили представление погрешности формул численного дифференцирования через разделенные разности функций. После оценки разделенных разностей через производные, согласно (5.4), получится оценка погрешности

$$|f^{(k)}(x) - L_n^{(k)}(x)| \leq \sum_{j=0}^k \frac{k!}{(k-j)!(n+j)!} \max_{[y_1, y_2]} |f^{(n+j)}(x)| |\omega_n^{(k-j)}(x)|, \quad (4)$$

где, как и ранее,

$$y_1 = \min(x, x_1, \dots, x_n), \quad y_2 = \max(x, x_1, \dots, x_n).$$

Рассмотрим некоторые частные случаи. Пусть производная k -го порядка определяется из дифференцирования интерполяционного многочлена k -й степени при $k+1$ узлах интерполяции

$x_1, \dots, x_{k+1}$. Поскольку

$$L_{k+1}(x) = f(x_1) + \dots + f(x_1; \dots; x_{k+1})(x - x_1) \dots (x - x_k),$$

то соотношение $f^{(k)}(x) \approx L_{k+1}^{(k)}(x)$ приобретает вид

$$f^{(k)}(x) \approx k! f(x_1; \dots; x_{k+1}),$$

т. е. является следствием (5.4).

При $k=1$ из (3) имеем

$$\begin{aligned} f'(x) - L'_n(x) &= f(x; x_1; \dots; x_n) \omega'_n(x) + f(x; x; x_1; \dots; x_n) \omega_n(x) = \\ &= \frac{f^{(n)}(\xi_1)}{n!} \omega'_n(x) + \frac{f^{(n+1)}(\xi_2)}{(n+1)!} \omega_n(x), \quad \xi_1, \xi_2 \in [y_1, y_2]. \end{aligned} \quad (5)$$

Если x является одним из узлов интерполирования $x_1, \dots, x_n$, то $\omega_n(x) = 0$, и из соотношения (5) получаем

$$|f'(x) - L'_n(x)| \leq (\max_{[y_1, y_2]} |f^{(n)}(x)|) \frac{|\omega'_n(x)|}{n!}.$$

Рассмотрим случай, когда узлы расположены на равных расстояниях $x_j - x_{j-1} \equiv h$. Выражение $\omega_n^{(j)}(x)$ есть сумма произведений $n-j$ сомножителей, каждое порядка h , и поэтому $\omega_n^{(j)}(x)$ имеет порядок $O(h^{n-j})$. Значит, правая часть (1) есть

$$f(x; x_1; \dots; x_n) \omega_n^{(k)}(x) + O(h^{n-k+1}) = O(h^{n-k}).$$

Если $\omega_n^{(k)}(x) = 0$ в точке x , то, согласно этому соотношению, порядок погрешности повышается. Приведем примеры употребительных формул, обладающих таким свойством.

Пусть вычисляется значение производной $f''(x)$ в точке x_0 при узлах интерполяции $x_{-l}, \dots, x_0, \dots, x_l$, $x_j - x_{j-1} \equiv h$. Тогда $n = 2l + 1$, $\omega_n(x) = (x - x_{-l}) \dots (x - x_l)$ — функция нечетная относительно точки x_0 . Поэтому $\omega_n(x_0) = \omega'_n(x_0) = 0$, и соотношение (3) при $k = 2$ будет иметь вид

$$f''(x_0) - L''_n(x_0) = 2f(x; x; x_{-l}; \dots; x_l) \omega'_n(x_0).$$

Непосредственным дифференцированием получаем значение

$$\omega'_n(x_0) = (-1)^l (l!)^2 h^{2l}.$$

Таким образом, с учетом (5.4) имеем

$$f''(x_0) - L''_n(x_0) = \frac{2(-1)^l (l!)^2}{(2l+2)!} f^{(2l+2)}(\xi) h^{2l}. \quad (6)$$

Величину $L''_n(x_0)$ можно выразить через разности f_0^{2l} . В результате

получится формула

$$f''(x_0) \approx L_n''(x_0) = h^{-2} \sum_{j=1}^l \frac{2(-1)^{j-1}((j-1)!)^2}{(2j)!} f_0^{2j} \quad (7)$$

с остаточным членом (6).

Наиболее употребительны случаи формулы (7), соответствующие $l=1$, $l=2$:

$$f''(x_0) \approx f_0^2/h^2, \quad f''(x_0) \approx (f_0^2 - \frac{1}{12} f_0^4)/h^2. \quad (8)$$

Точно такое же повышение порядка погрешности на 1 по сравнению с h^{n-2k} имеет место при вычислении любой производной $f^{(2k)}(x_0)$ по рассмотренной выше совокупности узлов.

Пусть вычисляется значение производной $y'(x)$ в точке $\bar{x} = x_0 + h/2$ при узлах интерполяции

$$x_{-(l-1)}, \dots, x_0, x_1, \dots, x_l, \quad \dot{x}_{j+1} - x_j \equiv h.$$

В этом случае $\omega_n(x)$ — функция четная относительно точки $\bar{x}$, и поэтому соотношение (5) приобретает вид

$$f'(\bar{x}) - L_n'(\bar{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega_n(\bar{x}).$$

Как и в случае второй производной, выразим $L_n'(\bar{x})$ через $f_{1/2}^{2l+1}$ и вычислим $\omega_n(\bar{x})$. Получим формулу численного дифференцирования

$$\begin{aligned} f'(x_0 + h/2) &\approx L_{2l}'(x_0 + h/2) = \\ &= h^{-1} \sum_{j=0}^{l-1} \frac{(-1)^j ((1/2)(3/2) \dots (j-1/2))^2}{(2j+1)!} f_{1/2}^{2j+1} \end{aligned} \quad (9)$$

с остаточным членом

$$\frac{(-1)^l ((1/2)(3/2) \dots (l-1/2))^2}{(2l+1)!} f^{(2l+1)}(\xi) h^{2l}.$$

Выпишем наиболее простые формулы (соответствующие $l=1$, $l=2$)

$$\begin{aligned} f'(x_0 + h/2) &\approx f_{1/2}^1/h, \\ f'(x_0 + h/2) &\approx (f_{1/2}^1 - \frac{1}{24} f_{1/2}^3)/h. \end{aligned} \quad (10)$$

Точно такое же повышение порядка точности на 1 будет иметь место при этой совокупности узлов для любой производной нечетного порядка.

Упомянем также представляющие практический интерес так называемые *односторонние формулы численного дифференцирования*. Пусть требуется вычислить $f'(x_0)$ при узлах интерполяции $x_0, \dots, x_n$ с равным шагом. Согласно (5), имеем

$$f'(x_0) - L'_{n+1}(x_0) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega'_n(x_0). \quad (11)$$

Подставляя сюда выражение $L_{n+1}(x)$ в виде интерполяционного многочлена Ньютона для интерполирования вперед, получим формулу

$$f'(x_0) \approx L'_{n+1}(x_0) = h^{-1} \sum_{j=1}^n \frac{(-1)^{j-1}}{j} f_{j/2} \quad (12)$$

с остаточным членом

$$\frac{(-1)^n}{(n+1)} f^{(n+1)}(\xi) h^n.$$

Простейшие формулы (соответствующие $l=1, l=2$)

$$f'(x_0) \approx f_{1/2}^1/h, \quad (13)$$

$$f'(x_0) \approx (f_{1/2}^1 - \frac{1}{2} f_1^2)/h.$$

Аналогично строятся формулы для вычисления $f'(x_0)$ при узлах интерполяции $x_0, x_{-1}, \dots, x_{-n}$:

$$f'(x_0) \approx L'_{n+1}(x_0) = h^{-1} \sum_{j=1}^n \frac{1}{j} f_{-j/2}^l \quad (14)$$

с остаточным членом

$$\frac{1}{n+1} f^{(n+1)}(\xi) h^n.$$

§ 18. О вычислительной погрешности формул численного дифференцирования

В случае формул численного дифференцирования мы сталкиваемся с тем обстоятельством, что уменьшение погрешности метода решения может приводить к росту влияния погрешности исходных данных и вычислительной погрешности.

Пусть для определенности $f'(x_0)$ определяется из соотношения

$$f'(x_0) \approx (f(x_1) - f(x_0))/h. \quad (1)$$

Согласно (17.11), остаточный член этой формулы

$$r_1 = -f''(\xi) h/2.$$

Пусть $|f''(\xi)| \leq M_2$ и, следовательно, $|r_1| \leq M_2 h/2$.

Если значения функции $f(x_i)$ вычисляются с некоторыми погрешностями e_i , $|e_i| \leq E$, то погрешность $f'(x_0)$ будет содержать дополнительное слагаемое

$$r_2 = (e_1 - e_0)/h, \quad |r_2| \leq 2E/h.$$

Для простоты пренебрежем округлениями при реальном вычислении правой части (1). Тогда имеем оценку погрешности

$$|r| \leq |r_1| + |r_2| \leq g(h) = M_2 h/2 + 2E/h. \quad (2)$$

Для малости погрешности необходима малость h , но при уменьшении h растет второе слагаемое (рис. 2.18.1). Из уравнения $g'(h) = 0$ получаем точку экстремума $g(h)$:

$$h_0 = 2\sqrt{E/M_2}$$

и затем значение

$$g(h_0) = 2\sqrt{M_2 E}.$$

Таким образом, ни при каком h нельзя гарантировать, что погрешность результата будет величиной порядка $o(\sqrt{E})$.

Ошибки e_i возникают вследствие ошибок в задаваемых значениях функций, например, если функция определяется из измерений, или вычисляется по некоторой приближенной формуле; поскольку эти значения округляются дополнительно при вводе в машину, то следует считать, что $E \geq \text{const } 2^{-t}$, где t — число разрядов. Таким образом, мы можем получить $f'(x_0)$ в лучшем случае с половиной верных разрядов.

В случае применения формул более высокого порядка точности положение несколько улучшается. Пусть производная $y^{(k)}(x)$ вычисляется по формуле

$$y^{(k)}(x) \approx \left(\sum_0^n c_j y_j \right) / h^k, \quad c_j = O(1), \quad (3)$$

с остаточным членом $O(h^l)$; в случае $k=1,2$ мы видели, что формулы для вычисления производных порядка k действительно содержат в знаменателе h^k . Погрешность, являющаяся следствием ошибок в правой части (3), оценивается величиной $\text{const} \cdot E/h^k$. Таким образом, вместо (2) мы имеем

$$|r| \leq g_1(h) = A_1 h^l + A_2 E/h^k.$$

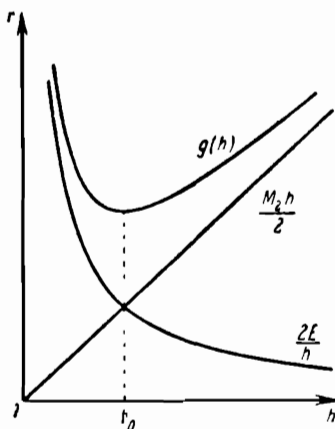


Рис. 2.18.1.

Минимум правой части достигается при h порядка $E^{1/(l+k)}$, и при этом сама правая часть имеет порядок $E^{l/(l+k)}$. Таким образом, с ростом l порядок погрешности по отношению к E повышается; при этом значение шага, соответствующее минимуму оценки погрешности, становится все больше. Конечно, следует иметь в виду, что величины A_1 и A_2 могут расти с ростом l , и поэтому увеличение l разумно лишь в определенных пределах.

В практической работе получили распространение различные способы «сглаживания» исходных значений функции. Эти приемы позволяют, с одной стороны, уменьшить величину E , а с другой — сделать функцию более гладкой, т. е. сделать допустимым применение интерполяционных формул дифференцирования более высокого порядка точности. Об этом пойдет речь в главе V.

Дополнительная литература к главе II

§ 1: I [10], II [32]; §§ 4—7, 10: II [9, 11]; § 18: III [25, 34].

ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

§ 1. Квадратурные формулы Ньютона — Котеса

Простейшие квадратурные формулы можно получить из наглядных соображений. Пусть вычисляется интеграл

$$I = \int_a^b f(x) dx.$$

Если $f(x) \approx \text{const}$ на рассматриваемом отрезке, то можно положить $I \approx (b-a)f(\xi)$, ξ — произвольная точка отрезка $[a, b]$. Естественно взять в качестве ξ среднюю точку отрезка; тогда получим формулу прямоугольников:

$$I \approx (b-a)f\left(\frac{b+a}{2}\right).$$

Предположим, что функция $f(x)$ близка к линейной; тогда естественно заменить интеграл площадью трапеции с высотой $(b-a)$ и основаниями $f(a)$ и $f(b)$ (рис. 3.1.1). Получим формулу трапеций

$$I \approx (b-a) \frac{f(a) + f(b)}{2}.$$

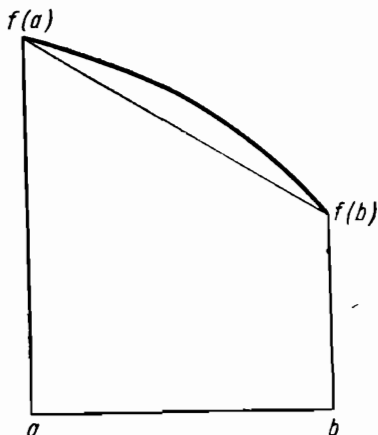


Рис. 3.1.1.

Если функция $f(x)$ близка к линейной, то из наглядных соображений видно, что формула прямоугольников также должна давать неплохой результат: дело в том, что $(b-a)f\left(\frac{b+a}{2}\right)$ есть площадь любой трапеции с высотой $b-a$ и средней линией $f\left(\frac{b+a}{2}\right)$ в частности, она равна площади трапеции, у которой

одна из сторон лежит на касательной к графику кривой в точке $\left(\frac{b+a}{2}, f\left(\frac{b+a}{2}\right)\right)$ (рис. 3.1.2).

Более сложные квадратурные формулы, так же как и формулы численного дифференцирования, строятся при помощи аппарата интерполирования. На этом же пути можно получить оценку погрешности квадратурных формул.

Пусть вычисляется интеграл

$$\int_a^b f(x) p(x) dx;$$

смысл введения функции $p(x)$, называемой *весовой*, будет объяснен в § 3. Зададимся некоторыми $d_1, \dots, d_n \in [-1, 1]$ и построим интерполяционный многочлен $L_n(x)$ степени $n-1$, совпадающий с $f(x)$ в точках

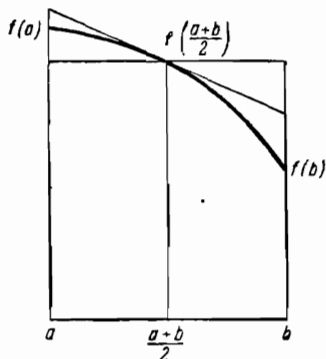


Рис. 3.1.2.

$$x_1 = \frac{b+a}{2} + \frac{b-a}{2} d_1.$$

Положим

$$\int_a^b f(x) p(x) dx \approx \int_a^b L_n(x) p(x) dx.$$

Имеем

$$\begin{aligned} R_n(f) &= \int_a^b f(x) p(x) dx - \int_a^b L_n(x) p(x) dx = \\ &= \int_a^b p(x) (f(x) - L_n(x)) dx. \end{aligned}$$

Разность $f(x) - L_n(x)$ можно оценить, воспользовавшись оценкой погрешности интерполяционного многочлена Лагранжа

$$|f(x) - L_n(x)| \leq \left(\max_{[a, b]} |f^{(n)}(x)| \right) \frac{|\omega_n(x)|}{n!},$$

где

$$\omega_n(x) = (x - x_1) \dots (x - x_n).$$

Отсюда

$$|R_n(f)| \leq \left(\max_{[a, b]} |f^{(n)}(x)| \right) \int_a^b \frac{|\omega_n(x)| |p(x)|}{n!} dx.$$

Произведем в последнем интеграле замену переменных

$$x = X(t) = \frac{b+a}{2} + \frac{b-a}{2} t;$$

получим

$$\frac{1}{n!} \int_a^b |\omega_n(x) p(x)| dx = D(d_1, \dots, d_n) \left(\frac{b-a}{2} \right)^{n+1},$$

где

$$D(d_1, \dots, d_n) = \int_{-1}^1 \frac{|\omega_n^0(t) p^0(t)|}{n!} dt,$$

$$\omega_n^0(t) = (t-d_1) \dots (t-d_n), \quad p^0(t) = p\left(\frac{b+a}{2} + \frac{b-a}{2} t\right).$$

Таким образом, справедлива оценка

$$|R_n(f)| \leq D(d_1, \dots, d_n) \left(\max_{[a, b]} |f^{(n)}(x)| \right) \left(\frac{b-a}{2} \right)^{n+1}. \quad (1)$$

Пусть все d_j различны. Тогда

$$L_n(x) = \sum_{j=1}^n f(x_j) \prod_{i \neq j} \frac{x - x_i}{x_j - x_i}.$$

После замены переменных $x = X(t)$ получим

$$\int_a^b p(x) L_n(x) dx = \left(\frac{b-a}{2} \right) \sum_{j=1}^n D_j f(x_j), \quad (2)$$

где

$$D_j = \int_{-1}^1 p^0(t) \prod_{i \neq j} \frac{t - d_i}{d_j - d_i} dt.$$

Таким образом, построенная квадратурная формула имеет вид

$$\int_a^b f(x) p(x) dx \approx \frac{b-a}{2} \sum_{j=1}^n D_j f\left(\frac{b+a}{2} + \frac{b-a}{2} d_j\right). \quad (3)$$

Как и при численном дифференцировании можно обнаружить следующее обстоятельство: если задача имеет определенную симметрию, то метод с симметрией того же типа часто обладает дополнительными преимуществами.

Л е м м а. Пусть весовая функция $p(x)$ — четная относительно середины отрезка $[a, b]$, и узлы x_j расположены симметрично относительно середины отрезка: $d_j = -d_{n+1-j}$. Тогда коэффициенты квадратуры, соответствующие симметричным узлам, равны между собой

$$D_j = D_{n+1-j}. \quad (4)$$

Доказательство. Запишем коэффициенты квадратуры в виде

$$D_j = \int_{-1}^1 p^0(t) \frac{\omega(t)}{\omega'(d_j)(t-d_j)} dt,$$

где $\omega(t) = (t-d_1) \dots (t-d_n)$. Если n четно, то $\omega(t) = \omega(-t)$ и $\omega'(d_j) = -\omega'(d_{n+1-j})$. Если n нечетно, то $\omega(t) = -\omega(-t)$ и $\omega'(d_j) = \omega'(d_{n+1-j})$. В обоих случаях после замены $t = -t'$ получим

$$\begin{aligned} D_j &= \int_{-1}^1 p^0(t') \frac{\omega(t')}{\omega'(d_{n+1-j})(t' + d_j)} dt' = \\ &= \int_{-1}^1 p^0(t') \frac{\omega(t')}{\omega'(d_{n+1-j})(t' - d_{n+1-j})} dt' = D_{n+1-j}. \end{aligned} \quad (5)$$

Такие «симметричные» квадратуры обладают следующим дополнительным свойством, которое, формально говоря, не предусматривалось при их построении. Они точны для любой функции, нечетной относительно середины отрезка $[a, b]$. В самом деле, для таких функций $\int_a^b p(x) f(x) dx = 0$ вследствие четности

$p(x)$, а $\sum_{j=1}^n D_j f(x_j) = 0$ вследствие (4); поэтому и $R_n(f) = 0$.

В частности, квадратуры будут точны для любого одночлена $\text{const} \cdot \left(x - \frac{b+a}{2}\right)^{2l+1}$.

Рассмотрим такую квадратуру, соответствующую нечетному n . Она точна для $f(x) = \text{const} \cdot \left(x - \frac{b+a}{2}\right)^n$ и, согласно построению, также и для любого многочлена степени $n-1$. Следовательно, такая квадратура будет точна и для любого многочлена степени n . Таким образом, построенные квадратуры с $2q-1$ и $2q$ узлами и с симметричным расположением узлов относительно середины отрезка оказываются точными для многочленов степени $2q-1$.

Чтобы получить уточненную оценку погрешности квадратур с нечетным числом узлов не через $f^{(n)}(x)$, а через $f^{(n+1)}(x)$, следует заменить подынтегральную функцию интерполяционным многочленом Лагранжа, имеющим точку $(b+a)/2$ двукратным узлом интерполирования. Ниже для случая $p(x) \equiv 1$ строится ряд элементарных квадратурных формул и дается оценка их погрешностей; при $n=1$ и $n=3$ для «симметричных» формул производится оценка погрешности через $f^{(n+1)}(x)$.

1. *Формула прямоугольников.* Пусть $n = 1$, $d_1 = 0$. Тогда

$$D = \int_{-1}^1 |t| dt = 1, \quad D_1 = \int_{-1}^1 1 \cdot dt = 2$$

и имеем квадратурную формулу

$$\int_a^b f(x) dx \approx (b-a) f\left(\frac{a+b}{2}\right) \quad (6)$$

с оценкой остаточного члена

$$\max_{[a, b]} |f'(x)| \frac{(b-a)^2}{4}.$$

2. *Формула прямоугольников с кратным узлом.* Пусть $n = 2$, $d_1 = d_2 = 0$. Тогда

$$D = \int_{-1}^1 \frac{t^2}{2} dt = \frac{1}{3},$$

$$L_2(x) = f\left(\frac{a+b}{2}\right) + f'\left(\frac{a+b}{2}\right)\left(x - \frac{a+b}{2}\right),$$

$$\int_a^b L_2(x) dx = (b-a) f\left(\frac{a+b}{2}\right).$$

Таким образом, имеем квадратурную формулу

$$\int_a^b f(x) dx \approx (b-a) f\left(\frac{a+b}{2}\right)$$

с оценкой остаточного члена

$$\max_{[a, b]} |f''(x)| \frac{(b-a)^3}{24}.$$

В обоих случаях получилась одна и та же квадратурная формула, но с различной оценкой остаточного члена.

3. *Формула трапеций.* Пусть $n = 2$, $d_1 = -1$, $d_2 = 1$. Тогда

$$D = \int_{-1}^1 \frac{|t^2 - 1|}{2} dt = \frac{2}{3},$$

$$D_1 = \int_{-1}^1 \frac{1-t}{2} dt = 1,$$

$$D_2 = \int_{-1}^1 \frac{1+t}{2} dt = 1.$$

Получена формула трапеций

$$\int_a^b f(x) dx \approx \frac{(b-a)}{2} (f(a) + f(b)) \quad (7)$$

с оценкой остаточного члена

$$\max_{[a, b]} |f''(x)| \frac{(b-a)^3}{12}.$$

4. Формула Симпсона. Пусть $n=4$, $d_1=-1$, $d_2=d_4=0$, $d_3=1$. Тогда

$$D = \int_{-1}^1 \frac{t^2(1-t^2)}{4!} dt = \frac{1}{90}.$$

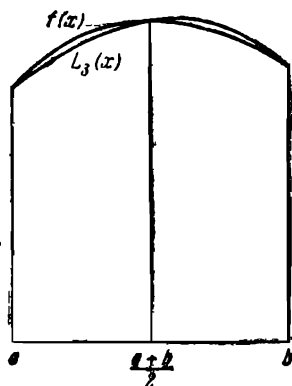
Согласно формуле интерполирования с кратными узлами, мы можем написать

$$L_4(x) = L_3(x) + f\left(a; b; \frac{a+b}{2}; \frac{a+b}{2}\right) (x-a)(x-b)\left(x - \frac{a+b}{2}\right),$$

где

$$L_3(x) = f(a) + f(a; b)(x-a) + f\left(a; b; \frac{a+b}{2}\right)(x-a)(x-b).$$

Второе слагаемое в выражении $L_4(x)$ является функцией, нечетной относительно середины отрезка $[a, b]$. Поэтому



$$\int_a^b L_4(x) dx = \int_a^b L_3(x) dx.$$

Многочлен $L_3(x)$ является интерполяционным многочленом второй степени, соответствующим

$$d_1=-1, \quad d_2=0, \quad d_3=1$$

(рис. 3.1.3). Этим значениям d_1, d_2, d_3 соответствуют

Рис. 3.1.3.

$$D_1 = \frac{1}{3}, \quad D_2 = \frac{4}{3}, \quad D_3 = \frac{1}{3}.$$

Полученная квадратурная формула

$$\int_a^b f(x) dx \approx \frac{b-a}{6} \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right) \quad (8)$$

с оценкой остаточного члена

$$\max_{[a, b]} |f^{(4)}(x)| \frac{(b-a)^4}{2880}$$

называется *формулой Симпсона*.

5. Квадратурную формулу, точную для многочленов третьей степени, можно получить также, например, при $n = 4$, взяв

$$d_1 = -1, \quad d_2 = -\frac{1}{3}, \quad d_3 = \frac{1}{3}, \quad d_4 = 1.$$

Тогда по описанной процедуре получается квадратура

$$\begin{aligned} \int_a^b f(x) dx &\approx \\ &\approx \frac{b-a}{8} \left(f(a) + 3f\left(a + \frac{b-a}{3}\right) + 3f\left(a + \frac{2(b-a)}{3}\right) + f(b) \right). \end{aligned} \quad (9)$$

При той же степени многочленов, для которых эта формула точна, она требует на 1 узел больше по сравнению с формулой Симпсона, и поэтому ее употребление менее оправдано.

В связи с рассмотрением этой квадратуры полезно обратить внимание на следующее. Так как $f(x) - L_n(x) = f^{(n)}(\xi(x)) \frac{\omega_n(x)}{n!}$, то остаточный член квадратуры записывается в виде:

$$\int_a^b f^{(n)}(\xi(x)) \frac{\omega_n(x)}{n!} dx.$$

В рассматривавшихся выше случаях 2—4 $\omega_n(x)$ не меняла знака на $[a, b]$. Согласно теореме о среднем:

$$\int_a^b h(x) g'(x) dx = h(\bar{x}) \int_a^b g(x) dx, \quad \bar{x} \in [a, b], \quad \text{если } g(x) \geq 0;$$

это выражение может быть представлено в виде

$$f^{(n)}(\bar{x}) \int_a^b \frac{\omega_n(x)}{n!} dx. \quad (10)$$

Подставляя в (10) конкретное выражение $\omega_n(x)$, получаем представление погрешности формул прямоугольников, трапеций и Симпсона, соответственно:

$$R_2(f) = f''(\bar{x}_1) \frac{(b-a)^3}{24},$$

$$R_3(f) = -f''(\bar{x}_2) \frac{(b-a)^3}{12}, \quad R_4(f) = -f^{(4)}(\bar{x}_3) \frac{(b-a)^5}{2880}.$$

Если $f^{(n)}(x)$ на $[a, b]$ не очень сильно меняется, то для этих формул оценка (1) не будет слишком завышенной. В случае формулы (9)

$$\omega_n(x) = (x-a) \left(x - \frac{2a+b}{3}\right) \left(x - \frac{a+2b}{3}\right) (x-b)$$

меняет знак на $[a, b]$, и поэтому нельзя непосредственно представить остаточный член в виде (10).

Поэтому возможно, что оценка (1) для этой квадратуры окажется завышенной. В § 2 будут проведены построения общего характера, с помощью которых удастся получить точную оценку погрешности этой квадратуры через максимум $|f^{(4)}(x)|$.

Основной целью настоящей главы будет рассмотрение способов вычисления интегралов от функций, заданных аналитическим выражением, и выработка принципов построения стандартных программ интегрирования таких функций. Естественно, что кроме этих задач в теории квадратурных формул имеются и другие задачи, например, связанные с обработкой экспериментального материала.

Для примера обратим внимание на квадратурные формулы Чебышева, широко применявшиеся при подсчете водоизмещения судов. Постановка задачи, приведшей к построению этих квадратур, довольно близка к постановке задачи, возникающей при планировании экспериментов (см. гла-

ва 2, § 1). Вычисляется интеграл $\int_{-1}^1 f(x) dx$, причем известно, что функция

$f(x)$ с приемлемой точностью может быть приближена многочленом степени q . Получение каждого значения $f(x)$, например, путем измерений, обходится довольно дорого, и получаемые значения содержат довольно большие случайные ошибки. Предположим, что ошибки измерения независимы, имеют одинаковую дисперсию d и математическое ожидание, равное нулю. Тогда дисперсия приближенного значения $S_n(f)$, вычисляемого по квадратурной формуле

$$I(f) \approx S_n(f) = \sum_{j=1}^n c_j f(x_j),$$

равна

$$d \sum_{j=1}^n c_j^2.$$

Условие $I(f) = S_n(f)$ при $f = \text{const}$ имеет вид:

$$2 = \sum_{j=1}^n c_j \quad (11)$$

Как нетрудно проверить, минимум величины $d \sum_{j=1}^n c_j^2$ при условии (11) достигается при $c_1 = \dots = c_n = 2/n$. Эти рассуждения привели к следующей

постановке задачи: среди всех квадратур

$$I(f) \approx \frac{2}{n} \sum_{j=1}^n f(x_j),$$

точных для многочленов степени q , найти квадратуру, соответствующую наименьшему n . При $q = 0$ и $q = 1$ искомой будет квадратура прямоугольников $I(f) = 2f(0)$; при $q = 2$ и $q = 3$ — квадратура Гаусса (см. § 3)

$$I(f) \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

§ 2. Оценка погрешности квадратурной формулы на классе функций

Рассмотрим квадратурную формулу

$$I(f) = \int_a^b f(x) dx \approx S_N(f) = (b-a) \sum_{j=1}^N D_j f(x_j), \quad *)$$

про которую известно, что она точна для всех многочленов степени $n-1$. Пусть $C_n(A_n, [a, b])$ — класс функций $n-1$ раз непрерывно дифференцируемых с кусочно-непрерывной производной $f^{(n)}(x)$, удовлетворяющей условию $|f^{(n)}(x)| \leq A_n$ при $x \in [a, b]$. Поставим перед собой задачу получить явное выражение величины

$$\sup_{f \in C_n(A_n, [a, b])} |I(f) - S_N(f)|,$$

т. е. получить оценку погрешности, предельно точную на рассматриваемом классе.

Напомним формулу Тейлора с остаточным членом в интегральной форме:

$$f(x) = T_{cf}^n + \int_c^x f^{(n)}(y) \frac{(x-y)^{n-1}}{(n-1)!} dy, \quad (1)$$

где

$$T_{cf}^n = \sum_{j=0}^{n-1} f^{(j)}(c) \frac{(x-c)^j}{j!},$$

Подставим это представление f в выражение погрешности

$$R_N(f) = I(f) - S_N(f). \quad (2)$$

*) Иногда в записи квадратуры перед знаком суммы стоит множитель $b-a$, а иногда $(b-a)/2$; в зависимости от этого коэффициенты D_j имеют различный смысл.

Величину $R_N(f)$ можно записать в виде

$$R_N(f) = R_N(T_c^n f) + R_N \left(\int_c^x f^{(n)}(y) \frac{(x-y)^{n-1}}{(n-1)!} dy \right).$$

Первое слагаемое равно нулю вследствие предположения, что квадратура точна для многочленов степени $n-1$. Таким образом,

$$\begin{aligned} R_N(f) &= \int_a^b \left(\int_c^x f^{(n)}(y) \frac{(x-y)^{n-1}}{(n-1)!} dy \right) dx = \\ &= (b-a) \sum_{l=1}^N D_l \int_c^{x_l} f^{(n)}(y) \frac{(x_l-y)^{n-1}}{(n-1)!} dy. \end{aligned}$$

Возьмем для определенности $c = a$; производя перестановку переменных интегрирования x и y в первом интеграле, получим

$$\begin{aligned} \int_a^b \int_a^x f^{(n)}(y) \frac{(x-y)^{n-1}}{(n-1)!} dy dx &= \\ &= \int_a^b f^{(n)}(y) \left(\int_y^b \frac{(x-y)^{n-1}}{(n-1)!} dx \right) dy = \int_a^b f^{(n)}(y) \frac{(b-y)^n}{n!} dy. \end{aligned}$$

Каждое слагаемое $\int_a^{x_l} f^{(n)}(y) \frac{(x_l-y)^{n-1}}{(n-1)!} dy$ можно записать в виде

$$\begin{aligned} \int_a^b f^{(n)}(y) \frac{(\overline{x_l-y})^{n-1}}{(n-1)!} dy, \\ \bar{t} = \begin{cases} t & \text{при } t > 0, \\ 0 & \text{при } t \leq 0. \end{cases} \end{aligned}$$

Таким образом,

$$R_N(f) = \int_a^b K_N(y) f^{(n)}(y) dy, \quad (3)$$

где

$$K_N(y) = \frac{(b-y)^n}{n!} - (b-a) \sum_{l=1}^N D_l \frac{(\overline{x_l-y})^{n-1}}{(n-1)!}, \quad (4)$$

Отсюда следует оценка погрешности

$$|R_N(f)| \leq A_n \int_a^b |K_N(y)| dy. \quad (5)$$

Покажем, что эта оценка не может быть улучшена. На любом из отрезков $[a, x_1], [x_1, x_2], \dots, [x_N, b]$ функция $K_N(y)$ является многочленом степени n и поэтому имеет не более n нулей. Поэтому на всем отрезке $[a, b]$ функция $K_N(y)$ имеет не более $n(N+1)$ нулей. Положим $z(x) = A_n \operatorname{sign} K_N(x)$. Функция

$$Z(x) = \int_a^x z(y) \frac{(x-y)^{n-1}}{(n-1)!} dy$$

имеет своей n -й производной кусочно-непрерывную функцию $z(x)$, $|z(x)| = A_n$, т. е. принадлежит рассматриваемому классу.

$$R_N(Z) = \int_a^b K_N(y) A_n \operatorname{sign} K_N(y) dy = A_n \int_a^b |K_N(y)| dy,$$

таким образом,

$$\sup_{f \in C_n(A_n, [a, b])} |R_N(f)| = A_n \int_a^b |K_N(y)| dy. \quad (6)$$

Задача. Рассмотрим класс n раз непрерывно дифференцируемых функций, $|f^{(n)}(x)| \leq A_n$. Доказать, что на этом классе также

$$\sup |R_N(f)| = A_n \int_a^b |K_N(y)| dy. \quad (7)$$

Обратимся к оценке погрешности квадратурной формулы

$$\int_a^b f(x) dx \approx \frac{b-a}{8} \left(f(a) + 3f\left(\frac{2a+b}{3}\right) + 3f\left(\frac{a+2b}{3}\right) + f(b) \right)$$

через $\max_{[a, b]} |f^{(4)}(x)|$. Предварительно произведем в интеграле

$$\int_a^b |K_N(y)| dy \text{ замену } y = a + (b-a)t. \text{ Получим}$$

$$\int_a^b |K_N(y)| dy = (b-a)^{n+1} \int_0^1 |K_N^0(t)| dt,$$

где

$$K_N^0(t) = \frac{(1-t)^n}{n!} - \sum_{j=1}^N D_j \frac{(\overline{d_j^0 - t})^{n-1}}{(n-1)!}, \quad d_j^0 = \frac{x_j - a}{b - a}. \quad (8)$$

В рассматриваемом случае

$$K_4^0(t) = \frac{(1-t)^4}{24} - \frac{1}{48} \left((\overline{0-t})^3 + 3 \left(\overline{\frac{1}{3}-t} \right)^3 + \right. \\ \left. + 3 \left(\overline{\frac{2}{3}-t} \right)^3 + (\overline{1-t})^3 \right);$$

заметим, что $(\overline{0-t})^3 \equiv 0$ при $0 \leq t$. Можно проверить, что $K_4^0(t) \leq 0$ при $0 \leq t \leq 1$. Далее, для оценки погрешности следует вычислить интеграл

$$\int_0^1 |K_4^0(t)| dt = - \int_0^1 K_4^0(t) dt.$$

В случае, когда $K_N^0(t)$ сохраняет знак, оценка погрешности может быть произведена более удобным способом.

Лемма. Пусть $K_N^0(t)$ сохраняет знак на $[a, b]$. Тогда справедливо равенство

$$R_N(f) = f^{(n)}(\xi) \frac{(b-a)^{n+1}}{n!} \left(\int_0^1 t^n dt - \sum_{j=1}^N D_j (d_j^0)^n \right). \quad (9)$$

Доказательство. Применяя обобщенную формулу Лагранжа к правой части (3), получим

$$R_N(f) = f^{(n)}(\xi) \int_a^b K_N(y) dy = f^{(n)}(\xi) (b-a)^{n+1} \int_0^1 K_N^0(t) dt.$$

Подставляя сюда $f(y) = f_0(y) = \frac{(y-a)^n}{n!}$, имеем

$$R_N(f_0) = (b-a)^{n+1} \int_0^1 K_N^0(t) dt.$$

Подставив ту же функцию в равенство

$$R_N(f_0) = I(f_0) - S_N(f_0),$$

получим

$$R_N(f_0) = \frac{(b-a)^{n+1}}{n!} \left(\int_0^1 t^n dt - \sum_{j=1}^N D_j (d_j^0)^n \right).$$

Из этих соотношений следует

$$\int_0^1 K_N^0(t) dt = \frac{1}{n!} \left(\int_0^1 t^n dt - \sum_{j=1}^N D_j (d_j^0)^n \right) = \sigma.$$

Теперь справедливость (9) очевидна. Заметим, что величина σ есть погрешность квадратуры $\int_0^1 f(x) dx \approx \sum_{j=1}^N D_j f(d_j^0)$ при вычислении $\int_0^1 \frac{x^n}{n!} dx$.

Таким образом, в рассматриваемом случае

$$\begin{aligned} R_N(f) = f^{(4)}(\xi) \frac{(b-a)^6}{24} \left(\frac{1}{5} - \frac{3}{8} \left(\frac{1}{3} \right)^4 - \frac{3}{8} \left(\frac{2}{3} \right)^4 - \frac{1}{8} \right) = \\ = -f^{(4)}(\xi) \frac{(b-a)^6}{6480}. \end{aligned}$$

При пяти узлах аналогично формуле Симпсона строится формула

$$\begin{aligned} \int_a^b f(x) dx \approx \frac{b-a}{90} \left(12f\left(\frac{a+b}{2}\right) + 32\left(f\left(\frac{3a+b}{4}\right) + \right. \right. \\ \left. \left. + f\left(\frac{a+3b}{4}\right)\right) + 7(f(a) + f(b)) \right) \quad (10) \end{aligned}$$

с остаточным членом $-\frac{1}{1935360} (b-a)^7 f^{(6)}(\xi)$.

При выводе остаточного члена этой формулы узел $\frac{a+b}{2}$ рассматривается как кратный, функция $K_5^0(y)$ сохраняет постоянный знак и оценка погрешности проводится по описанной выше схеме.

Формулы с большим числом равноотстоящих узлов применяются относительно редко.

§ 3. Квадратурные формулы Гаусса

На примере формул Симпсона и прямоугольников видно, что при удачном выборе узлов степень многочленов, для которых точна квадратурная формула, повышается.

Поэтому можно поставить следующую оптимизационную задачу. При заданном числе n узлов построить квадратуру

$$I(f) = \int_a^b f(x) p(x) dx \approx S_n(f) = \frac{b-a}{2} \sum_{j=1}^n D_j f(x_j), \quad (1)$$

точную для многочленов наиболее высокой степени. Такие квадратуры называют *формулами Гаусса*.

Рассмотрение подобной постановки вызывается следующими соображениями. Пусть имеется квадратурная формула (1), точная для любого многочлена P_m степени m :

$$I(P_m) = S_n(P_m) \quad (2)$$

Вследствие (2) при любом P_m имеем цепочку равенств

$$R_n(f) = R_n(P_m) + R_n(f - P_m) = R_n(f - P_m).$$

Справедлива очевидная оценка

$$|R_n(g)| \leq \int_a^b |g(x)| |p(x)| dx + \frac{b-a}{2} \sum_{j=1}^n |D_j| |g(x_j)| \leq V_n \sup_{[a, b]} |g(x)|,$$

где

$$V_n = \int_a^b |p(x)| dx + \left(\frac{b-a}{2}\right) \sum_{j=1}^n |D_j|. \quad (3)$$

Отсюда

$$|R_n(f)| \leq V_n \sup_{[a, b]} |f(x) - P_m(x)|.$$

Взяв в правой части нижнюю грань по всем многочленам степени m , получим оценку

$$|R_n(f)| \leq V_n E_m(f), \quad (4)$$

где

$$E_m(f) = \inf_{P_m} \sup_{[a, b]} |f(x) - P_m(x)|.$$

Функция приближается многочленами более высокой степени точнее, чем многочленами низкой. Поэтому можно ожидать, что погрешность квадратуры будет меньше, если она точна для многочленов более высокой степени.

Это высказывание, вообще говоря, неверно: во-первых — оценка (4) не является точной, во-вторых — коэффициент V_n зависит от конкретной квадратуры и, строго говоря, может расти с ростом степени многочленов, для которых точна квадратура. Однако все-таки в нем содержится некоторое оправдание разумности рассматриваемой постановки.

Может случиться, что подынтегральная функция исходного интеграла $\int_a^b F(x) dx$ плохо приближается многочленами, а функция $F(x)/p(x)$, где $p(x)$ — некоторая функция простого вида, приближается многочленами хорошо. Тогда для данной задачи квадратура вида (1), точная для многочленов некоторой

степени, будет иметь преимущество по сравнению с квадратурой

$$\int_a^b F(x) dx \approx \frac{b-a}{2} \sum_{j=1}^n c_j F(x_j), \quad (5)$$

точной для многочленов той же степени.

Если при $p(x) \neq 0$ положим в (1) $p(x)f(x) = F(x)$, то получим квадратуру

$$\int_a^b F(x) dx \approx \frac{b-a}{2} \sum_{j=1}^n C_j F(x_j), \quad C_j = \frac{D_j}{p(x_j)}.$$

Очевидно, эта квадратура будет точна для всех функций $p(x)P_m(x)$, где $P_m(x)$ — многочлен степени m . Таким образом, исходная постановка равносильна постановке отыскания квадратуры вида (5), точной для функций $F(x) = p(x)P_m(x)$ при возможно более высокой степени многочлена m .

Остановимся на первой из постановок. При

$$P_m(x) = \sum_{q=0}^m a_q x^q$$

имеем

$$R_n(P_m) = \sum_{q=0}^m a_q R_n(x^q).$$

Таким образом, чтобы квадратура (1) была точна для многочленов степени m , необходимо и достаточно, чтобы она была точна для всех функций x^q , $q = 0, \dots, m$. Следовательно, должны выполняться соотношения

$$R_n(x^q) = \int_a^b x^q p(x) dx - \frac{b-a}{2} \sum_{j=1}^N D_j x_j^q = 0, \quad q = 0, \dots, m. \quad (6)$$

Мы имеем систему из $m+1$ -го уравнения относительно неизвестных $x_1, \dots, x_n, D_1, \dots, D_n$.

При $m \leq 2n-1$ число уравнений не более числа неизвестных, и поэтому можно ожидать, что алгебраическая система (6) имеет решение. Можно попытаться построить квадратурные формулы, соответствующие значению $m = 2n-1$, решая эту систему, однако, неясно, будут ли узлы квадратур, получаемых из (6), принадлежать отрезку $[a, b]$. В противном случае может случиться, что функция $f(x)$ не определена в узлах интегрирования, и поэтому употребление квадратуры невозможно.

Заметим, что в главе VIII при построении конечно-разностных методов решения обыкновенных дифференциальных уравнений возникнут квадратуры с узлами вне отрезка $[a, b]$.

Займемся построением квадратур, соответствующих максимальному значению $m = 2n - 1$.

Лемма 1. Если $x_1, \dots, x_n$ — узлы квадратуры (1), точной для всех многочленов степени $2n - 1$, то

$$\int_a^b \omega_n(x) P_{n-1}(x) p(x) dx = 0$$

при $\omega_n(x) = (x - x_1) \dots (x - x_n)$ и $P_{n-1}(x)$ — произвольном многочлене степени не выше $n - 1$.

Доказательство. Пусть $P_{n-1}(x)$ — некоторый многочлен степени не выше $n - 1$. Вследствие условия леммы квадратура (1) точна для

$$Q_{2n-1}(x) = \omega_n(x) P_{n-1}(x).$$

Поэтому

$$\begin{aligned} \int_a^b \omega_n(x) P_{n-1}(x) p(x) dx &= \int_a^b Q_{2n-1}(x) p(x) dx = \\ &= \frac{b-a}{2} \sum_{j=1}^n D_j Q_{2n-1}(x_j) = 0; \end{aligned}$$

последнее соотношение вытекает из равенства $Q_{2n-1}(x_j) = 0$ при всех j . Лемма 1 доказана.

Далее предполагается, что $p(x) > 0$ почти всюду на $[a, b]$.

Из рассмотрений § 15 гл. II вытекает единственность многочлена $\psi_n(x) = x^n + \dots$, ортогонального ко всем многочленам низшей степени при скалярном произведении, заданном соотношением

$$(f, g) = \int_a^b f(x) \bar{g}(x) p(x) dx.$$

Поэтому $\psi_n(x) = \omega_n(x)$ и узлы отыскиваемой квадратуры должны быть нулями $\psi_n(x)$; согласно рассмотрениям § 15 гл. II, многочлен $\psi_n(x)$ имеет на $[a, b]$ n различных нулей.

Лемма 2. Пусть $x_1, \dots, x_n$ — нули $\psi_n(x)$ и (1) — квадратура, точная для многочленов степени $n - 1$. Тогда квадратура (1) точна для многочленов степени $2n - 1$.

Доказательство. Произвольный многочлен $Q_{2n-1}(x)$ степени $2n - 1$ представим в виде

$$Q_{2n-1}(x) = \psi_n(x) g_{n-1}(x) + r_{n-1}(x),$$

где g_{n-1} и r_{n-1} — многочлены степени $n-1$. Имеем

$$R_n(Q_{2n-1}) = R_n(\psi_n g_{n-1}) + R_n(r_{n-1}) = R_n(\psi_n g_{n-1}),$$

$$R_n(\psi_n g_{n-1}) =$$

$$= \int_a^b \psi_n(x) g_{n-1}(x) p(x) dx - \frac{b-a}{2} \sum_{j=1}^n D_j \psi_n(x_j) g_{n-1}(x_j) = 0,$$

поскольку

$$\int_a^b \psi_n(x) g_{n-1}(x) p(x) dx = 0$$

вследствие свойства ортогональности $\psi_n(x)$ к многочленам низшей степени, а все $\psi_n(x_j) = 0$ по предположению леммы; тогда $R_n(Q_{2n-1}) = 0$. Лемма 2 доказана.

Теперь можно построить требуемую квадратурную формулу. Для этого зададимся узлами интерполяции $x_1, \dots, x_n$, $\psi_n(x_j) = 0$, и построим квадратуру Ньютона — Котеса, согласно описанной в § 1 процедуре; получим требуемую квадратуру

$$\int_a^b f(x) p(x) dx \approx \frac{b-a}{2} \sum_{j=1}^n G_j f(x_j), \quad (7)$$

точную для многочленов степени $2n-1$.

Если почти всюду $p(x) > 0$, то не существует квадратуры, точной для всех многочленов степени $2n$. В самом деле, возьмем $Q_{2n}(x) = (x-x_1)^2 \dots (x-x_n)^2$, тогда левая часть (1)

$$\int_a^b ((x-x_1) \dots (x-x_n))^2 p(x) dx > 0,$$

а правая равна 0.

Исследуем некоторые свойства квадратуры (7), в частности, докажем положительность коэффициентов G_j . Функция $\left(\frac{\psi_n(x)}{x-x_l}\right)^2$ является многочленом степени $2n-2$, обращающимся в нуль во всех точках $x_j \neq x_l$. Квадратура (7) будет точна для этой функции, и поэтому

$$\int_a^b \left(\frac{\psi_n(x)}{x-x_l}\right)^2 p(x) dx = \frac{b-a}{2} G_l \left(\frac{\psi_n(x)}{x-x_l}\right)^2 \Big|_{x_l}.$$

Раскрывая выражение $\frac{\psi_n(x)}{x-x_l}$, получим

$$G_l = \frac{1}{b-a} \int_a^b \left(\prod_{i \neq l} \frac{x-x_i}{x_l-x_i} \right)^2 p(x) dx > 0.$$

Квадратура (7) точна для $f(x) = 1$, и поэтому

$$\int_a^b p(x) dx = \frac{b-a}{2} \sum_{j=1}^n G_j;$$

вследствие положительности G_j мы можем переписать это равенство в виде

$$\int_a^b p(x) dx = \frac{b-a}{2} \sum_{j=1}^n |G_j|.$$

Возвращаясь к формулам (3), (4), получим

$$V_n = \int_a^b p(x) dx + \frac{b-a}{2} \sum_{j=1}^n |G_j| = 2 \int_a^b p(x) dx.$$

Тогда из формулы (4) следует оценка

$$|R_n(f)| \leq 2 \left(\int_a^b p(x) dx \right) E_{2n-1}(f). \quad (8)$$

Приведем еще одну оценку погрешности формул Гаусса, на этот раз через

$$\max_{[a, b]} |f^{(2n)}(x)|.$$

Пусть $L_{2n}(x)$ — интерполяционный многочлен Лагранжа с двукратными узлами интерполяции $x_1, \dots, x_n$, т. е.

$$L_{2n}(x_l) = f(x_l), \quad L'_{2n}(x_l) = f'(x_l).$$

Согласно (2.6.11), имеем

$$f(x) - L_{2n}(x) = f(x; x_1; x_1; \dots; x_n; x_n) \psi_n^2(x),$$

Поскольку для формул Гаусса $R_n(L_{2n}(x)) = 0$, то

$$\begin{aligned} R_n(f) &= R_n(f(x; x_1; x_2; \dots; x_n; x_n) \psi_n^2(x)) = \\ &= \int_a^b f(x; x_1; x_1; \dots; x_n; x_n) \psi_n^2(x) p(x) dx - \\ &\quad - \frac{b-a}{2} \sum_{j=1}^n D_j f(x_j; x_1; x_1; \dots; x_n; x_n) \psi_n^2(x_j). \end{aligned}$$

Вычитаемое равно нулю, так как $\psi_n(x_j) = 0$ при всех j , поэтому

$$R_n(f) = \int_a^b f(x; x_1; x_1, \dots, x_n, x_n) \psi_n^2(x) p(x) dx.$$

Воспользовавшись неотрицательностью $\psi_n^2(x) p(x)$, применим обобщенную формулу Лагранжа; получим

$$R_n(f) = f(\xi; x_1; x_1; \dots; x_n; x_n) \int_a^b \psi_n^2(x) p(x) dx;$$

выразив $f(\xi; x_1; x_1; \dots; x_n; x_n)$ через производную, имеем

$$R_n(f) = f^{(2n)}(\bar{\xi}) \int_a^b \frac{\psi_n^2(x)}{(2n)!} p(x) dx, \quad (9)$$

Для практического применения формул Гаусса необходимо иметь в распоряжении узлы и коэффициенты этих квадратур. В случае $p(x)$ — четной относительно точки $\frac{a+b}{2}$, согласно рассмотрению § 16 гл. II, нули ортогональных многочленов, т. е. узлы квадратур Гаусса, расположены симметрично относительно середины отрезка $[a, b]$. Вследствие леммы из § 1 коэффициенты интерполяционной формулы Гаусса (7) будут удовлетворять условию четности $G_j = G_{n+1-j}$. Это обстоятельство наполовину уменьшает объем таблиц формул Гаусса.

Если $p(x) \equiv 1$, то коэффициенты G_j и числа $d_j = \frac{2x_j - (b+a)}{b-a}$ не зависят от отрезка $[a, b]$. В самом деле, если многочлен $\psi_n(x) = (x-x_1) \dots (x-x_n)$ принадлежит системе многочленов, ортогональных с весом 1 на $[a, b]$, то многочлен $(t-d_1) \dots (t-d_n)$ принадлежит системе многочленов, ортогональных с весом 1 на $[-1, 1]$. Поэтому он сам, его нули, а согласно (1.2), и коэффициенты D_j определяются однозначно, вне зависимости от исходного отрезка $[a, b]$.

При составлении таблиц квадратур Гаусса для различных весов удобно пользоваться свойством перемежаемости нулей ортогональных многочленов $\psi_n(x)$, доказанным в § 16 гл. II. В соответствии с этим свойством после определения корней $\psi_{n-1}(x)$ становятся известными интервалы, которым принадлежат корни $\psi_n(x)$.

Приведем для сведения набор квадратур Гаусса для отрезка $[-1, 1]$ при $\rho(x) = 1$. В этом случае остаточный член (9) есть

$$f^{(2n)}(\xi) \frac{2^{2n+1} (n!)^4}{(2n!)^3 (2n+1)}.$$

Вследствие свойства симметрии мы указываем лишь неотрицательные d_j и коэффициенты при них (табл. 1).

Т а б л и ц а 1

	d_1 D_1	d_2 D_2	d_3 D_3
1	0,0000000000 2,0000000000		
2	0,5773502692 1,0000000000		
3	0,0000000000 0,8888888888	0,7745966692 0,5555555556	
4	0,8611363115 0,3478548451	0,3399810436 0,6521451549	
5	0,0000000000 0,5688888888	0,9061798459 0,4786286705	0,5384693101 0,2369268851
6	0,9324695142 0,1713244924	0,6612093864 0,3607615730	0,2386191861 0,4679139346

В настоящее время составлены таблицы узлов и весов квадратур Гаусса при $n \leq 512$ с 20 десятичными знаками.

Рассмотрим примеры применения проведенных построений для получения несколько иных квадратур. Для простоты возьмем случай $[a, b] = [-1, 1]$.

Формулы Лобатто. Пусть при отыскании квадратуры

$$I(f) = \int_{-1}^1 f(x) dx \approx \sum_{j=1}^n D_j f(d_j)$$

наложено дополнительное требование, что $d_1 = -1$ и $d_n = 1$. Тогда число свободных параметров, от которых зависит квадратура, равно $2n - 2$. Поэтому есть основания ожидать, что удастся

найти решение системы (6) при $m = 2n - 3$, т. е. построить квадратуру, точную для многочленов степени $2n - 3$. Интерполяционным многочленом первой степени, совпадающим с $f(x)$, в точках ± 1 , является

$$Q_1(x) = f(-1) \frac{1-x}{2} + f(1) \frac{1+x}{2}.$$

Представим исходный интеграл в виде

$$I(f) = \int_{-1}^1 Q_1(x) dx + \int_{-1}^1 \frac{f(x) - Q_1(x)}{1-x^2} p(x) dx, \quad (10)$$

где $p(x) = 1 - x^2$.

Построим формулу Гаусса с $n - 2$ узлами, соответствующую весовой функции $p(x)$:

$$\int_{-1}^1 f(x) p(x) dx \approx \sum_{j=2}^{n-1} \bar{D}_j f(d_j); \quad (11)$$

для удобства мы обозначили узлы через $d_2, \dots, d_{n-1}$. Положим

$$\int_{-1}^1 f(x) dx \approx \int_{-1}^1 Q_1(x) dx + \sum_{j=2}^{n-1} \bar{D}_j \frac{f(d_j) - Q_1(d_j)}{1 - d_j^2}. \quad (12)$$

После подстановки выражения $Q_1(x)$ и осуществления интегрирования в первом интеграле получим требуемую квадратуру

$$\int_{-1}^1 f(x) dx \approx D_1 f(-1) + \sum_{j=2}^{n-1} D_j f(d_j) + D_n f(1),$$

где, в частности,

$$D_j = \bar{D}_j (1 - d_j^2)^{-1}, \quad j = 2, \dots, n-1.$$

Действительно, пусть $f(x)$ — многочлен степени $2n - 3$. Первый интеграл в (10) вычисляется без ошибки. Разность $f(x) - Q_1(x)$ есть многочлен степени $2n - 3$, обращающийся в 0 при $x = \pm 1$; поэтому $(f(x) - Q_1(x))(1 - x^2)^{-1}$ есть многочлен степени $2n - 5$. Для такого многочлена квадратура (11) точна, и поэтому при замене второго интеграла в (10) вторым слагаемым в правой части (12) ошибки не произойдет. При $n = 2$ формула Лобатто совпадает с формулой трапеций, при $n = 3$ — с формулой Симпсона.

Пусть теперь $p(x)$ — нечетная функция на $[-1, 1]$, $p(x) > 0$ почти всюду на $[0, 1]$, число узлов $n = 2l - \text{четное}$.

При $p(x)$ четной у хороших квадратур оказывалось $D_j = D_{n+1-j}$, $d_j = -d_{n+1-j}$.

При $p(x)$ нечетной естественно искать такие квадратуры в виде

$$I(f) \approx \sum_{j=1}^l D_j (f(d_j) - f(-d_j)), \quad 0 < d_1 < \dots < d_l \leq 1. \quad (13)$$

Такая квадратура автоматически точна при $f(x)$ четной. Квадратура (13) содержит $n = 2l$ свободных параметров, поэтому есть основания надеяться получить квадратуру, точную для одночленов $x, \dots, x^{2n-1}$. Так как она автоматически точна для одночленов $1, \dots, x^{2n}$, то она будет точна для многочленов степени $2n$, т. е. на 1 выше, чем в случае формул Гаусса при положительном весе.

Чтобы построить такую квадратуру, представим исходный интеграл в виде

$$I(f) = \int_0^1 (f(x) - f(-x)) p(x) dx = \int_0^1 \frac{f(x) - f(-x)}{x} x p(x) dx.$$

После замены переменных $x = \sqrt{y}$ получим

$$I(f) = \int_0^1 F(y) \bar{p}(y) dy,$$

где

$$F(y) = \frac{f(\sqrt{y}) - f(-\sqrt{y})}{2\sqrt{y}}, \quad \bar{p}(y) = p(\sqrt{y}).$$

Построим квадратуру Гаусса

$$\int_0^1 F(y) \bar{p}(y) dy \approx \sum_{j=1}^l \bar{D}_j F(\bar{d}_j), \quad (14)$$

точную для многочленов $F(y)$ степени $2l-1$. Возвращаясь к старой переменной x , имеем

$$I(f) \approx \sum_{j=1}^l \bar{D}_j \frac{f(\sqrt{\bar{d}_j}) - f(-\sqrt{\bar{d}_j})}{2\sqrt{\bar{d}_j}} = \sum_{j=1}^l D_j (f(d_j) - f(-d_j)).$$

Если $f(x) = \sum_{j=0}^{2n} a_j x^j$, то

$$F(y) = \sum_{j=0}^{n-1} a_{2j+1} y^j.$$

Поэтому соотношение (14) и последующее обращаются в точные равенства.

Такие квадратуры могут быть применены, например, при вычислении интегралов $\int_0^{\infty} f(x) \cos x \, dx$; интеграл разбивается на части:

$$\int_{\pi n}^{(n+1)\pi} f(x) \cos x \, dx = (-1)^{n-1} \frac{\pi}{2} \int_{-1}^1 f\left(\left(x + 2n + 1\right) \frac{\pi}{2}\right) \sin \frac{\pi x}{2} \, dx,$$

и для вычисления последних интегралов применяются квадратуры (13), соответствующие $p(x) = \sin \frac{\pi x}{2}$.

Практически важный случай, в частности для задачи приближения функций, представляют формулы Гаусса на отрезке $[-1, 1]$ при весовой функции $\frac{1}{\sqrt{1-x^2}}$, называемые *формулами Эрмита*:

$$I(f) = \int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} \, dx \approx S_n(f) = \frac{\pi}{n} \sum_{j=1}^n f(x_j), \quad (15)$$

где $x_j = \cos \frac{(2j-1)\pi}{2n}$ — нули многочлена Чебышева $T_n(x)$. Проверим, что квадратура (15) точна для многочленов $T_m(x)$, если $m < 2n$. При $m=0$ справедливо

$$I(T_0) = \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} \, dx = \pi, \quad S_n(T_0) = \pi$$

и $I(T_0) = S_n(T_0)$. При $m > 0$, вследствие ортогональности функций $T_m(x)$ при весе $\frac{1}{\sqrt{1-x^2}}$, имеем

$$I(T_m) = I(T_m T_0) = (T_m, T_0) = 0.$$

В то же время

$$\begin{aligned} S_n(T_m) &= \frac{\pi}{n} \sum_{j=1}^n \cos(m \arccos x_j) = \frac{\pi}{n} \sum_{j=1}^n \cos \frac{m(2j-1)\pi}{2n} = \\ &= \frac{\pi}{2n} \sum_{j=1-n}^n \exp\left(\frac{m(2j-1)\pi i}{2n}\right). \end{aligned}$$

Суммируемые числа образуют геометрическую прогрессию со знаменателем $\exp\left(\frac{m\pi i}{n}\right)$; при $0 < m < 2n$ этот знаменатель отличен от 1. Воспользовавшись формулой для суммы членов геометрической прогрессии:

$$\sum_{j=1}^n aq^{j-1} = \frac{a(q^n - 1)}{q - 1},$$

получим

$$S_n(T_m) = \frac{\pi}{2n} \frac{\exp\left(\frac{m(2n+1)\pi i}{2n}\right) - \exp\left(\frac{m(1-2n)\pi i}{2n}\right)}{\exp\frac{m\pi i}{n} - 1}.$$

Так как

$$\exp\left(\frac{m(2n+1)\pi i}{2n}\right) = \exp\left(\frac{m(1-2n)\pi i}{2n}\right),$$

то при $0 < m < 2n$ имеем $S_n(T_m) = 0 = I(T_m)$. Поскольку квадратура точна для всех многочленов Чебышева $T_m(x)$ при $m < 2n$, то она точна для всех многочленов степени $2n-1$.

В настоящее время рассчитано много таблиц формул Гаусса и формул типа Лобатто, в частности, при

$$\begin{aligned} [a, b] &= [-1, 1], & p(x) &= 1, \\ [a, b] &= [0, 1], & p(x) &= x^a, \\ [a, b] &= [0, 1], & p(x) &= (x(1-x))^a, \\ [a, b] &= [0, \infty), & p(x) &= x^s \exp(-x). \end{aligned}$$

Если подынтегральная функция интеграла

$$I(f) = \int_0^{\omega} f(x) dx$$

хорошо приближается тригонометрическими многочленами с периодом ω , то целесообразно применить квадратуру

$$I(f) \approx S_N(f) = \frac{\omega}{N} \sum_{j=0}^{N-1} f\left(\frac{j\omega}{N}\right). \quad (16)$$

Имеем равенство

$$\begin{aligned} S_N\left(\exp\left(2\pi m i \frac{x}{\omega}\right)\right) &= \frac{\omega}{N} \sum_{j=0}^{N-1} \exp\left(2\pi m i \frac{j\omega}{N}\right) = \\ &= \begin{cases} \frac{\omega}{N} \sum_{j=0}^{N-1} 1 = \omega & \text{при } \frac{m}{N} \text{ целое,} \\ \frac{\exp(2\pi m i) - 1}{\exp\left(\frac{2\pi m i}{N}\right) - 1} = 0 & \text{при } \frac{m}{N} \text{ не целое.} \end{cases} \end{aligned}$$

В то же время

$$I\left(\exp\left(2\pi m i \frac{x}{\omega}\right)\right) = \begin{cases} \omega & \text{при } m=0, \\ 0 & \text{при } m \neq 0. \end{cases}$$

Из этих равенств следует, что квадратура (16) точна для функций $\cos\left(2\pi m \frac{x}{\omega}\right)$ при $m=0$ или при $\frac{m}{N}$ не целое и всех функций $\sin\left(2\pi m \frac{x}{\omega}\right)$; в результате этого квадратура точна для любого тригонометрического многочлена

$$t_N(x) = a_0 + \sum_{m=1}^{N-1} \left(a_m \cos\left(2\pi m \frac{x}{\omega}\right) + b_m \sin\left(2\pi m \frac{x}{\omega}\right) \right) + \\ + b_N \sin\left(2\pi N \frac{x}{\omega}\right),$$

следовательно,

$$R_N(f) = R_N(f - t_N) + R_N(t_N) = R_N(f - t_N).$$

Аналогично (8) получаем оценку

$$|R_N(f)| \leq 2\omega \inf_{t_N} \max_{[0, \omega]} |f(x) - t_N(x)|.$$

Квадратуры (16) особенно употребительны при вычислении интегралов по угловым переменным при интегрировании в полярных и сферических координатах.

Можно показать, что не существует квадратур с N узлами, точных для всех тригонометрических многочленов степени N .

§ 4. Практическая оценка погрешности элементарных квадратурных формул

Выше получен ряд квадратурных формул и строгие оценки погрешности для них. Однако это не решает всех проблем по отношению к задаче численного интегрирования, стоящих перед нами. Важнейшей задачей вычислительной математики является создание алгоритмов и их систем, обеспечивающих получение решения задач с заданной точностью при минимальном объеме затрат человеческого труда и работы машины. Практическое применение полученных выше оценок требует аналитических выкладок по оценке производных и поэтому достаточно большого объема работы исследователя; кроме того, эти оценки часто оказываются слишком завышенными. Поэтому при создании таких систем обычно отказываются от использования подобных оценок, зачастую жертвуя строгой гарантией малости погрешности приближенного решения.

Можно говорить, что задача от ее возникновения до получения результата проходит через некоторую систему, состоящую из людей, решающих задачу, и ЭВМ. На первоначальном этапе применения ЭВМ узким местом, т. е. местом, наиболее тормозившим работу этой системы, являлся недостаток ЭВМ. Поэтому применение аналитических методов решения или аналитическое проведение оценок погрешности было оправданным.

Однако теперь, с повсеместным распространением вычислительной техники и внедрением ее в различные сферы деятельности общества, обстановка меняется. Узким местом этой системы становятся длительность выбора математической модели, выбора метода решения задачи, программирования и других этапов, предшествующих непосредственному решению задачи на ЭВМ. Эти этапы особенно замедляются в случае, когда решением задач на ЭВМ занимаются представители конкретных наук, например филологи, медики, экономисты, географы, мало знакомые с численными методами или программированием. Их обучение тонкостям теории численных методов может превратиться в самоцель, отвлекающую от решения основных задач их науки, и в конечном счете обойтись обществу довольно дорого. Поэтому в настоящее время важнейшей проблемой является создание систем решения задач с максимально простым обращением, предполагающих малую квалификацию пользователя в отношении численных методов и программирования.

Например, естественно потребовать, чтобы к программе вычисления интеграла с заданной точностью мог обратиться исследователь, знающий, что такое интеграл, но не умеющий ни интегрировать, ни дифференцировать.

Конечно, в развитии многих областей знания и техники решающая роль математики состоит в создании математической модели явления, а потом уже в применении ЭВМ для ее обсчета. При разработке модели от специалиста этой отрасли знания требуется определенная математическая культура, и наше высказывание не следует понимать как предложение полностью избавить его от математики.

Подоплекой проводимых здесь рассуждений является следующая известная мысль. Когда мы занимаемся решением каких-то задач, то нужно учитывать эффективность нашей работы не только по совокупным затратам на решение этих задач, но и принимать во внимание убыток, понесенный обществом в результате того, что нами не решены некоторые другие, возможно более важные, задачи.

При практическом анализе погрешности численного интегрирования часто пользуются различными полуэмпирическими приемами. Наиболее распространенным из этих приемов является следующий. Производятся вычисления по двум квадратурным

формулам

$$I(f) = \int_a^b f(x) dx \approx S^k(f) = \frac{b-a}{2} \sum_{j=1}^{N_k} D_{jf}^k(x_j^k), \quad k=1, 2;$$

некоторая линейная комбинация

$$S(f) = S^1(f) + \theta(S^2(f) - S^1(f))$$

этих квадратур принимается за приближенное значение интеграла, а величина $\rho = |S^2(f) - S^1(f)|$ — за меру погрешности приближенной формулы $I(f) \approx S(f)$. Довольно типичным является случай $\theta = 0$.

Описанный выше подход нельзя считать полностью оправданным вследствие его неоднозначности.

Пусть $S^1(f)$ — формула Симпсона:

$$\int_{-1}^1 f(x) dx \approx S^1(f) = \frac{f(-1) + 4f(0) + f(1)}{3},$$

$S^2(f)$ — формула трапеций: $S^2(f) = f(-1) + f(1)$ и $\theta = 0$. Тогда в качестве меры погрешности выступает величина

$$\rho = \frac{2}{3} |f(1) - 2f(0) + f(-1)|.$$

Если $S^2(f) = 2f(0)$ — формула прямоугольников, то соответствующее значение $\rho = \frac{1}{3} |f(1) - 2f(0) + f(-1)|$. Таким образом, мы получили две различные эмпирические оценки погрешности одной и той же формулы Симпсона.

Попытаемся прояснить ситуацию. Выражение $S(f)$ является некоторой квадратурной суммой

$$S(f) = \frac{b-a}{2} \sum_{j=1}^N D_{jf}(x_j) \quad (1)$$

по совокупности узлов x_j , принадлежащих объединению узлов, соответствующих квадратурам $S^1(f)$ и $S^2(f)$. В то же время

$$\rho = |l(f)|,$$

где

$$l(f) = \frac{b-a}{2} \sum_{j=1}^N B_{jf}(x_j). \quad (2)$$

Возьмем произвольную линейную комбинацию вида (2) и положим

$$S^1(f) = S(f), \quad S^2(f) = S(f) - \theta l(f).$$

Тогда мы получим приближенное значение интеграла $I(f) \approx S(f)$ с оценкой погрешности $\rho = \theta |l(f)|$. Мы видим, что на таком пути можно получить неограниченное множество оценок погрешности одной и той же квадратуры (1).

Рассматриваемую задачу можно формулировать следующим образом. Приближенное значение интеграла вычисляется по формуле

$$I(f) \approx S(f) = \frac{b-a}{2} \sum_{j=1}^N D_j f(x_j). \quad (3)$$

Требуется построить выражение вида (2), дающее представление о погрешности квадратуры (3).

Предположим, что погрешность квадратуры (3) представляется в виде

$$R(f) = \bar{D}(b-a)^{m+1} f^{(m)}(\xi). \quad (4)$$

Рассмотрим случай $m \leq N$. Тогда в качестве $l(f)$ можно взять величину

$$l(f) = \bar{D}(b-a)^{m+1} m! f(x_{i_1}; \dots, x_{i_{m+1}}),$$

где $x_{i_1}, \dots, x_{i_{m+1}}$ — различные узлы квадратуры (3). Согласно выражению разделенной разности через производную, имеем:

$$l(f) = \bar{D}(b-a)^{m+1} f^{(m)}(\xi), \quad a < \xi < b;$$

следовательно, при $a = \text{const}$, $f^{(m)}(a) \neq 0$, $b-a \rightarrow 0$ справедливо соотношение: $R(f) \sim l(f)$, и величину $\rho = |l(f)|$ можно принять за меру погрешности. Пусть, например, оценивается погрешность формулы трапеций

$$\int_a^b f(x) dx \approx S(f) = \frac{b-a}{2} \left(f(a) + f\left(\frac{a+b}{2}\right) + f(b) \right);$$

согласно оценкам из §§ 1, 2, имеем

$$R(f) = -\frac{(b-a)^3}{12} f''(\xi);$$

таким образом, мы можем принять за меру погрешности

$$\frac{b-a}{3} \left| f(b) - 2f\left(\frac{a+b}{2}\right) + f(a) \right|.$$

Иначе обстоит дело, когда $m \geq N$. Тогда нельзя получить никакого приближения к $f^{(m)}(\xi)$ через величины $f(x_1), \dots, f(x_N)$.

и проблема получения эмпирической оценки погрешности в рассматриваемой выше постановке не может быть решена. Например, мы не можем получить удовлетворительного представления об оценке погрешности формулы Симпсона через значения $f(a)$, $f\left(\frac{a+b}{2}\right)$, $f(b)$. Однако можно получить некоторую завышенную оценку погрешности.

Рассмотрим один подход к разрешению возникшей проблемы. Обратимся к формуле (2.5) при $n = N - 1$. Мы получим оценку погрешности вида:

$$|R(f)| \leq (b-a)^N \bar{D} \max_{[a, b]} |f^{(N-1)}(x)| = \sigma.$$

Положим

$$\rho = (b-a)^N \bar{D} (N-1)! |f(x_1; \dots; x_N)|. \quad (5)$$

При $a = \text{const}$, $f^{(N-1)}(a) \neq 0$, $b-a \rightarrow 0$ имеем $\sigma \sim \rho$, и таким образом, величину ρ можно принять за приближенную оценку погрешности формулы (3). Эта оценка будет сильно завышенной, поскольку при предположении $m \geq N$ имеем $R(f) = O((b-a)^{N+1})$; однако мы не можем предложить лучшей оценки погрешности формулы (3) по сравнению с оценкой через σ . В случае формулы Симпсона из (2.5) получаем оценку (5) с $\bar{D} = \frac{1}{8!}$, и, таким образом, за меру погрешности принимаем

$$\frac{4(b-a)}{8!} \left| f(b) - 2f\left(\frac{a+b}{2}\right) + f(a) \right|. \quad (6)$$

Если вычисляется интеграл от аналитической функции и число узлов задано, то представляется наиболее разумным применение формул Гаусса, поскольку они точны для многочленов максимальной степени; оценку погрешности разумно проводить по описанной выше процедуре.

Наше предложение оценивать погрешность формулы Симпсона с помощью выражения (6) основывается на предположении достаточной гладкости $f(x)$. Однако подобные эмпирические заключения зачастую применяются в условиях, когда не выполнены требования, предполагавшиеся при их получении, например при интегрировании негладких функций.

Пусть, например,

$$[a, b] = [0, 1], \quad f_{\alpha, \beta}(x) = (x + \alpha)^\beta, \quad \alpha > 0, \quad \beta > 0.$$

Тогда погрешность формулы Симпсона есть

$$E_{\alpha, \beta} = \frac{(1 + \alpha)^{\beta+1} - \alpha^{\beta+1}}{\beta + 1} - \left(\frac{1}{6} \alpha^\beta + \frac{2}{3} \left(\frac{1}{2} + \alpha \right)^\beta + \frac{1}{6} (1 + \alpha)^\beta \right).$$

В то же время

$$\left| f_{\alpha, \beta}(1) - 2f_{\alpha, \beta}\left(\frac{1}{2}\right) + f_{\alpha, \beta}(0) \right| = S_{\alpha, \beta} = \left| (1 + \alpha)^\beta - 2\left(\frac{1}{2} + \alpha\right)^\beta + \alpha^\beta \right|.$$

Имеем

$$\lim_{\beta \rightarrow 0} \lim_{\alpha \rightarrow 0} \frac{E_{\alpha, \beta}}{S_{\alpha, \beta}} = \frac{1}{6}.$$

Таким образом, применение (6) для оценки погрешности формулы Симпсона при малых α и β является неоправданным.

В случае малых α и β более естественно оценивать погрешность формулы Симпсона величиной

$$\frac{1}{6} (b-a) \left| f(b) - 2f\left(\frac{a+b}{2}\right) + f(a) \right|. \quad (7)$$

В одной из стандартных программ интегрирования, где стремились к «максимальной» надежности результата, в качестве меры погрешности была принята величина

$$D^0 (b-a)^N |f(x_1; \dots; x_N)|,$$

где D^0 было определено следующим способом: был взят отрезок $[0, 1]$ и положено

$$D^0 = \sup_{\substack{0 < \alpha \leq 1 \\ 0 < \beta \leq N-1}} \frac{|R(f_{\alpha, \beta})|}{|f_{\alpha, \beta}(x_1; \dots; x_N)|};$$

в случае формулы Симпсона такой выбор D^0 также приводит к оценке погрешности выражением (7).

В случае многомерных интегралов все практические способы оценки погрешности опираются на исходную, раскритикованную нами процедуру. Дело в том, что в многомерном случае погрешность оценивается через значения нескольких производных подынтегральной функции. Получение «обоснованных» оценок, подобных (6), для таких формул крайне затруднительно. Поэтому обращаются к исходной процедуре с последующей экспериментальной проверкой результатов ее применения. Исходная постановка вычисления интеграла по формуле

$$I(f) \approx \sum_{j=1}^N D_j f(P_j)$$

с оценкой погрешности по $|I(f)|$, $I(f) = \sum_{j=1}^N B_j f(P_j) \neq 0$ содержит в качестве

свободных параметров величины D_j , B_j и координаты узлов P_j квадратуры. При выборе этих параметров обычно руководствуются следующим правилом. Задаются степенью многочленов m , для которых квадратура должна быть точна, и требуют, чтобы выражение $I(f)$ обращалось в 0 на многочленах некоторой заранее заданной степени m_1 ($m_1 < m$), но не обращалось в 0 на всех многочленах степени m ; обычно берут $m_1 = m - 2$. За счет выбора B_j , D_j и узлов квадратуры подбирают минимальное N , при котором выполняются заданные условия. При удовлетворении этих условий нет однозначности в способе оценки погрешности, поскольку $I(f)$ определяется этими условиями с точностью до некоторого множителя α : $I(f) = \alpha I^0(f)$, где

$I^0(f) = \sum_{j=1}^N B_j^0 f(P_j)$. Обычно α берут равным минимальному по модулю из чисел D_j/B_j^0 .

§ 5. Интегрирование сильно осциллирующих функций

Пусть требуется вычислить интеграл

$$\int_a^b f(x) \exp(i\omega x) dx,$$

где $\omega(b-a) \gg 1$, $f(x)$ — гладкая функция. Функции

$$\operatorname{Re}(f(x) \exp(i\omega x)), \quad \operatorname{Im}(f(x) \exp(i\omega x))$$

имеют на рассматриваемом отрезке примерно $\omega(b-a)/\pi$ нулей. Поскольку многочлен степени n имеет не более n нулей, то такие функции могут быть хорошо приближены многочленами степени n лишь при $n \gg \omega(b-a)/\pi$. Поэтому для непосредственного вычисления интегралов от таких функций потребуется применение квадратур, точных для многочленов высокой степени.

Более выгодным может оказаться путь рассмотрения функции $\exp(i\omega x)$ как весовой.

Как и в § 1, зададимся узлами интерполирования

$$x_j = \frac{b+a}{2} + \frac{b-a}{2} d_j, \quad j = 1, \dots, n,$$

и заменим исходный интеграл на $\int_a^b L_n(x) \exp(i\omega x) dx$; последний интеграл может быть вычислен в явном виде

$$\begin{aligned} \int_a^b L_n(x) \exp(i\omega x) dx &= S_n^\omega(f) = \\ &= \frac{b-a}{2} \exp\left(i\omega \frac{a+b}{2}\right) \sum_{j=1}^n D_j\left(\omega \frac{b-a}{2}\right) f(x_j), \end{aligned}$$

где

$$D_j(\rho) = \int_{-1}^1 \left(\prod_{k \neq j} \frac{t - d_k}{d_j - d_k} \right) \exp(i\rho t) dt. \quad (1)$$

Получилась квадратурная формула

$$\int_a^b f(x) \exp(i\omega x) dx \approx S_n^\omega(f) \quad (2)$$

с остаточным членом

$$R_n(f) = \int_a^b (f(x) - L_n(x)) \exp(i\omega x) dx.$$

В соответствии с (1.1)

$$|R_n(f)| \leq \int_a^b |f(x) - L_n(x)| dx \leq D(d_1, \dots, d_n) (\max_{[a, b]} |f^{(n)}(x)|) \left(\frac{b-a}{2}\right)^{n+1}.$$

Вычисление интегралов подобного рода является типичной задачей, встречающейся при разложении функций в ряды Фурье, при построении диаграмм направленности антенн и т. д. В стандартных программах вычисления таких интегралов используются формулы (1), (2), соответствующие случаям: $n = 3$, $d_1 = -1$, $d_2 = 0$, $d_3 = 1$ (эту формулу называют *формулой Филлона*) или $n = 5$, $d_1 = -1$, $d_2 = -0,5$, $d_3 = 0$, $d_4 = 0,5$, $d_5 = 1$.

При вычислении интегралов по формулам (1), (2) может возникнуть следующая неприятность, которую мы проиллюстрируем на случае $n = 2$; $d_1 = -1$, $d_2 = 1$; тогда:

$$D_1(p) = \int_{-1}^1 \frac{1-t}{2} \exp(ipt) dt = \frac{\sin p}{p} + \frac{p \cos p - \sin p}{p^2} i,$$

$$D_2(p) = \int_{-1}^1 \frac{1+t}{2} \exp(ipt) dt = \frac{\sin p}{p} - \frac{p \cos p - \sin p}{p^2} i.$$

При $p \rightarrow 0$ имеем

$$\frac{p \cos p - \sin p}{p^2} = -\frac{p}{3} + O(p^3) \rightarrow 0.$$

Таким образом, $D_1(p), D_2(p) \rightarrow 1$ при $p \rightarrow 0$.

Пусть p — малое число. Функции $\cos p$ и $\sin p$ вычисляются в машине с погрешностью $O(2^{-t})$ и $O(p2^{-t})$, соответственно. Вследствие этого коэффициенты $D_1(p), D_2(p)$ приобретают погрешность $O(2^{-t}/p)$. При $n > 2$ оказывается, что погрешность коэффициентов $D_j(p)$, вычисляемых по формулам (1), может оказаться величиной порядка $2^{-t}/p^{n-1}$. Например, при $t = 30$, $n = 5$, $p \approx 0,01$ такая погрешность уже недопустима. Поэтому

стандартные программы вычисления интегралов от осциллирующих функций должны иметь специальный блок, предусматривающий изменение расчетных формул при малых p с тем, чтобы избежать существенного влияния вычислительной погрешности. Если n не очень велико, например $n = 2$, то можно пойти по следующему пути: при $|p|$, большем некоторого p_n , подбираемого экспериментально, вычисления производим по формулам (1), (2); если $|p| \leq p_n$, то вычисляем исходный интеграл по формуле трапеций (1.7), рассматривая всю функцию $f(x) \exp(i\omega x)$ как подынтегральную. В рассматриваемом случае ($n = 2, d_1 = -1, d_2 = 1$) формула (1.7) приобретает вид

$$\int_a^b f(x) \exp(i\omega x) dx \approx \frac{b-a}{2} (\exp(i\omega a) f(a) + \exp(i\omega b) f(b)). \quad (3)$$

Формулы (1), (3) можно объединить в одну формулу

$$\begin{aligned} \int_a^b f(x) \exp(i\omega x) dx = \\ = \frac{b-a}{2} \exp\left(i\omega \frac{a+b}{2}\right) (D_1(p) f(a) + D_2(p) f(b)), \end{aligned} \quad (4)$$

где

$$\begin{aligned} p &= \omega \frac{b-a}{2}, \\ D_{1,2}(p) &= \begin{cases} \frac{\sin p}{p} \pm \frac{p \cos p - \sin p}{p^2} i & \text{при } |p| > p_2, \\ \exp(\mp pi) & \text{при } |p| \leq p_2. \end{cases} \end{aligned} \quad (5)$$

В стандартных программах вычисления интегралов рассматриваемого типа применяются квадратуры типа (4), (5). Возникает вопрос: зачем усложнять стандартную программу? Может быть, проще написать стандартную программу вычисления по формулам (2), стандартную программу вычисления по формуле трапеций и дать инструкцию: при больших $|p|$ обращаться к первой из стандартных программ, при малых $|p|$ — ко второй. Обсудим целесообразность такого подхода. Нашей целью является создание максимальных удобств пользователю ЭВМ. Если в описании правил использования стандартной программы написано слишком много, то пользователь может не понять того, что написано, и

1) обратиться к другой программе, худшей по качеству, но имеющей лучшее описание или более простое обращение;

2) воспользоваться программой неправильно и не получить результатов; например, в рассматриваемом случае, применив

метод (1), (2) при $p = 0$, он получит аварийную остановку машины;

3) воспользоваться программой неправильно и получить неверный результат, предполагая, что он верный; например, так случится, если он воспользуется формулами (1), (2) при $2^{-1}/p^n$ порядка 1.

Очевидно, последний случай влечет за собой наиболее неприятные последствия для пользователя.

При выборе метода для стандартной программы надо стремиться к тому, чтобы описание программы, составляемой на базе этого метода, было по возможности простым и кратким; надо иметь в виду, что всякое дополнительное высказывание может быть истолковано неправильно в ущерб для расширения области применения программы. Предложение об использовании двух программ, одной — при $|p| > p_2$, другой — при $|p| \leq p_2$, приводило к путанице, особенно в случае вычисления серий интегралов, включающих интегралы как с малыми, так и с большими p . Поэтому в конце концов эта пара программ была заменена программой, где вычисления проводятся по формулам (4), (5) и в описании ничего не говорится о величине $|p|$.

Приведем один поучительный реальный пример.

В начале описания одной из стандартных программ вычисления кратных интегралов было написано «применение настоящей программы нецелесообразно, если число узлов берется большим 100 000». Спустя короткое время программа была практически изъята из употребления. Оказалось, что среди первых пользователей программы подавляющее большинство сразу задавало число узлов $n = 100\,000$; при этом на вычисление простого интеграла уходило слишком много машинного времени. Весть об этом распространялась во все стороны с возрастающей скоростью и вскоре к программе перестали обращаться. На самом деле, большинство реальных интегралов вычислялось при ее использовании с приемлемой точностью при числе узлов порядка 1000; число 100 000 было указано лишь как ориентировочная верхняя граница значений n , при которых вычислительная погрешность еще не оказывает катастрофического влияния на результат.

§ 6. Повышение точности интегрирования за счет разбиения отрезка на равные части

Из оценки (3.4) вытекает, что погрешность квадратуры оценивается через точность, с которой функция $f(x)$ может быть приближена многочленами степени n . Поэтому может показаться естественным добиваться увеличения точности за счет повышения степени многочленов, для которых эта квадратура точна. Однако такой путь содержит свои «подводные камни». При неудачном выборе узлов может оказаться, что величина

$\sum_{i=1}^n |D_i|$ растет вместе с n . Тогда в оценке (3.4) величина V_n

также растет с n и может оказаться, что уменьшение $E_m(f)$ с ростом n не компенсирует увеличения V_n .

Например, для простейшего равномерного распределения узлов

$$d_j^{(n)} = -1 + \frac{2(j-1)}{n-1}, \quad j=1, \dots, n, \quad [a, b] = [-1, 1],$$

оказывается, что при $m = n-1$ V_n растет примерно как 2^n (точнее, $\log_2 V_n \sim n$). В результате этого для аналитической функции $f(x) = (1 + 25x^2)^{-1}$ при таких узлах

$$R_n(f) = \int_{-1}^1 f(x) dx - S_n(f) \rightarrow \infty$$

при $n \rightarrow \infty$.

Рассмотрим случай отрезка интегрирования $[-1, 1]$ и сформулируем общую теорему, указывающую на необходимость осторожного обращения с формулами, точными для многочленов очень высокой степени. Пусть при каждом n мы имеем квадратуру

$$\int_{-1}^1 f(x) dx \approx \sum_{j=1}^n D_j^{(n)} f(d_j^{(n)}), \quad (1)$$

точную для многочленов степени $n-1$. Обозначим через $w_n(x_1, x_2)$ число узлов квадратуры (1), принадлежащих отрезку $[x_1, x_2]$.

Теорема (без доказательства). Пусть существует отрезок $[x_1, x_2] \in [-1, 1]$ такой, что

$$\frac{w_n(x_1, x_2)}{n} \not\sim \int_{x_1}^{x_2} \frac{1}{\pi \sqrt{1-x^2}} dx \quad \text{при } n \rightarrow \infty.$$

Тогда можно указать $b \neq 0$ и a такие, что для одной из аналитических функций

$$\operatorname{Re} \left(\frac{1}{x - (a + bi)} \right), \quad \operatorname{Im} \left(\frac{1}{x - (a + bi)} \right),$$

$$\lim_{n \rightarrow \infty} |R_n(f)| = \infty.$$

Таким образом, узлы квадратур (1), точных для многочленов степени $n-1$, должны располагаться обязательно с такой же плотностью, как нули ортогональных многочленов, т. е. как и нули квадратур Гаусса. Иначе такие квадратуры нельзя рассматривать как универсальные.

Перепишем оценку (3.8) погрешности формул Гаусса

$$|R_n(f)| \leq 2 \left(\int_a^b p(x) dx \right) E_{2n-1}(f). \quad (2)$$

Пусть подынтегральная функция $f(x)$ непрерывна; тогда, согласно теореме Вейерштрасса, при любом $\varepsilon > 0$ найдется

многочлен $P_q(x)$, для которого $\max_{[a, b]} |f(x) - P_q(x)| \leq \varepsilon$, отсюда следует, что $E_m(f) \rightarrow 0$ при $m \rightarrow \infty$. Таким образом, для любой непрерывной функции $f(x)$ погрешность формул Гаусса

$$R_n(f) \rightarrow 0 \quad \text{при} \quad n \rightarrow \infty.$$

Задача. Пусть $f(x)$ интегрируема по Риману. Доказать, что для формул Гаусса

$$R_n(f) \rightarrow 0 \quad \text{при} \quad n \rightarrow \infty.$$

Итак, формулы Гаусса могли бы быть положены в основу универсальных программ вычисления интегралов от функций. Некоторое практическое неудобство представляет следующее обстоятельство. Мы не можем обычно заранее сказать, с каким числом узлов потребуются квадратуры Гаусса при вычислении интегралов с заданной точностью. Поэтому потребуется вводить в машину или переписывать с запоминающих устройств в оперативную память узлы и веса этих квадратур, соответствующие все большим значениям n , что несколько затормозит вычисление интегралов, особенно на малых машинах. Следует также отметить, что во многих случаях возникает задача вычисления интегралов, где подынтегральная функция или ее производные невысокого порядка имеют участки резкого изменения, например обращаются в бесконечность. Такие функции плохо приближаются многочленами сразу на всем отрезке интегрирования. Здесь часто оказывается более выгодным разбить исходный отрезок на части и там применять формулы Гаусса или какие-либо иные формулы.

Пусть вычисляется интеграл $I = \int_A^B f(x) dx$. Разобьем отрезок $[A, B]$ на M частей $[a_{q-1}, a_q]$, где $a_0 = A, a_M = B$. Для вычисления интеграла по каждой из частей применим какую-либо из квадратурных формул из §§ 1,3:

$$\begin{aligned} I_q &= \int_{a_{q-1}}^{a_q} f(x) dx \approx s_q = \\ &= \frac{a_q - a_{q-1}}{2} \sum_{j=1}^m D_j f\left(\frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \end{aligned} \quad (3)$$

с оценкой остаточного члена

$$D\left(\max_{[a_{q-1}, a_q]} |f^{(n)}(x)|\right) \left(\frac{a_q - a_{q-1}}{2}\right)^{n+1}$$

В результате интеграл по всему отрезку будет аппроксимирован суммой

$$S_M^m(f) = \sum_{q=1}^M s_q = \sum_{q=1}^M \frac{a_q - a_{q-1}}{2} \sum_{j=1}^m D_j f\left(\frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \quad (4)$$

с оценкой остаточного члена

$$\begin{aligned} |R_M^m(f)| &= |I - S_M^m(f)| \leq \\ &\leq D \sum_{q=1}^M \left(\max_{[a_{q-1}, a_q]} |f^{(n)}(x)| \right) \left(\frac{a_q - a_{q-1}}{2} \right)^{n+1}. \end{aligned} \quad (5)$$

Формулу (4) часто называют *составной* или *обобщенной*.

Рассмотрим наиболее простой для исследования случай, когда отрезки разбиения имеют одинаковую длину $a_q - a_{q-1} \equiv H$, тогда оценка погрешности (5) после замены $\max_{[a_{q-1}, a_q]} |f^{(n)}(x)|$ на величину $A_n = \max_{[a, b]} |f^{(n)}(x)|$ приобретает вид

$$|R_M^m(f)| \leq \bar{D} A_n (B - A) H^n, \quad \text{где } \bar{D} = D 2^{-(n+1)}, \quad (6)$$

или

$$|R_M^m(f)| \leq \bar{D} A_n \frac{(B - A)^{n+1}}{M^n}. \quad (7)$$

Не следует думать, что в случае функций с малым числом ограниченных производных составные формулы численного интегрирования имеют лучший порядок сходимости по сравнению с формулами Гаусса. Предположим, что подынтегральная функция имеет p ограниченных производных. Тогда, применяя составную формулу (4), соответствующую $n = p$, получим приближенное значение интеграла с погрешностью $O(N^{-p})$. С другой стороны, известно, что для такой функции $E_{2N-1}(f) = O(N^{-p})$. Поэтому из (3.8) следует оценка погрешности формулы Гаусса с N узлами

$$R_N(f) = O(N^{-p}).$$

Таким образом, порядок оценки погрешности в обоих случаях одинаков.

Обратим внимание еще на одно удобство использования формул Гаусса сразу по всему отрезку интегрирования. Не нужно оценивать число p_0 ограниченных производных подынтегральной функции и в соответствии с этим выбирать наиболее подходящую формулу численного интегрирования по отрезкам разбиения — при применении формул Гаусса порядок погрешности $O(N^{-p_0})$ обеспечивается автоматически. Если производные более высокого порядка, чем p_0 , обращаются в бесконечность только на концах отрезка интегрирования, то обычно порядок сходимости формул Гаусса еще лучше. Конечно, не нужно думать, что формула, имеющая более высокий порядок скорости сходимости, всегда точнее формулы более низкого порядка скорости сходимости (по этому поводу см. объяснения к рис. 3.15.2—3.15.5).

Приведем конкретные квадратурные формулы и оценки погрешности для частных случаев формул (3).

1. *Составная формула трапеций с постоянным шагом интегрирования.* В этом случае при постоянном шаге $a_q - a_{q-1} \equiv H$ формула (3) приобретает вид

$$\int_{a_0}^{a_M} f(x) dx \approx H \left(\frac{f(a_0)}{2} + f(a_1) + \dots + f(a_{M-1}) + \frac{f(a_M)}{2} \right),$$

оценка остаточного члена

$$A_2 \frac{(a_M - a_0) H^2}{12} = A_2 \frac{(a_M - a_0)^3}{12M^2}. \quad (8)$$

2. *Составная формула Симпсона с постоянным шагом интегрирования.* При постоянном шаге $a_q - a_{q-1} \equiv H = 2h$ формула (3) приобретает вид

$$\int_{a_0}^{a_M} f(x) dx \approx H \left(\frac{1}{6} f(a_0) + \frac{4}{6} f\left(\frac{a_1}{2}\right) + \frac{1}{3} f(a_1) + \right. \\ \left. + \frac{4}{6} f\left(\frac{a_3}{2}\right) + \frac{1}{3} f(a_2) + \dots + \frac{1}{3} f(a_{M-1}) + \frac{4}{6} f\left(\frac{a_{M-1}}{2}\right) + \frac{1}{6} f(a_M) \right),$$

где $a_j = a_0 + jH$. Оценка остаточного члена

$$A_4 \frac{(a_M - a_0) H^4}{2880} = \frac{A_4 (a_M - a_0)^5}{2880M^4} = \frac{A_4 (a_M - a_0) h^4}{180}.$$

Последняя запись оценки наиболее употребительна.

Мы получили формулы с порядком погрешности $O(N^{-n})$ по отношению к общему числу узлов интегрирования $N = Mt$ в предположении ограниченности $|f^{(n)}(x)|$. Заметим, что в случае формул трапеций и Симпсона общее число узлов N оказалось меньше чем Mt , поскольку концы элементарных отрезков $[a_{q-1}, a_q]$ были узлами интерполяции и значения функции в этих концах использовались для вычисления интегралов по двум элементарным отрезкам.

Оценку погрешности того же порядка можно получить и при более слабом предположении $\int_A^B |f^{(n)}(x)| dx < \infty$.

В самом деле, обратимся к представлению (2.3) погрешности на элементарном отрезке

$$R_N(f) = \int_a^b K_N(y) f^{(n)}(y) dy,$$

где

$$K_N(y) = \frac{(b-y)^n}{n!} - (b-a) \sum_{j=1}^N D_j \frac{(\overline{x_j - y})^{n-1}}{(n-1)!}. \quad (9)$$

Отсюда получаем оценку

$$|R_N(f)| \leq \max_{[a, b]} |K_N(y)| \int_a^b |f^{(n)}(y)| dy. \quad (10)$$

В то же время

$$\max_{[a, b]} |K_N(y)| = \tilde{D} (b-a)^n,$$

где $\tilde{D} = \max_{[-1, 1]} \left| \frac{(1-t)^n}{n!} - \sum_{j=1}^N D_j \frac{(\overline{d_j^0} - t)^{n-1}}{(n-1)!} \right|$, $d_j^0 = \frac{x_j - a}{b-a}$, это равенство получено с помощью замены переменных $y = a + (b-a)t$ в правой части (10). В итоге имеем

$$|R_N(f)| \leq \tilde{D} (b-a)^n \int_a^b |f^{(n)}(x)| dx. \quad (11)$$

Таким образом, для квадратуры (3) справедлива так же оценка остаточного члена

$$\tilde{D} (a_q - a_{q-1})^n \int_{a_{q-1}}^{a_q} |f^{(n)}(x)| dx.$$

Пусть $a_q - a_{q-1} \equiv H$. Тогда суммарная оценка остаточного члена есть

$$|R_M^m(f)| \leq \tilde{D} \sum_{q=1}^M H^n \int_{a_{q-1}}^{a_q} |f^{(n)}(x)| dx = \tilde{D} H^n \int_A^B |f^{(n)}(x)| dx. \quad (12)$$

Пусть, например, вычисляется

$$\int_0^1 (\sin x)^\alpha dx, \quad 0 < \alpha < 1.$$

Так как $((\sin x)^\alpha)'_x \rightarrow \infty$ при $x \rightarrow 0$, то мы не можем получить никаких оценок погрешности через $\max |f'(x)|$. В то же время функция $(\sin x)^\alpha$ монотонная на $[0, 1]$, и поэтому

$$\int_0^1 |((\sin x)^\alpha)'| dx = (\sin x)^\alpha \Big|_0^1 = 1.$$

Следовательно, при использовании квадратуры (4), соответствующей $n=1$, с постоянным шагом $a_q - a_{q-1} \equiv H$, согласно (12) имеем оценку погрешности $O(H)$.

В данном примере из оценки (6) малость погрешности не следовала; в то же время на основании (12) мы заключаем, что эта погрешность порядка $O(1/N)$. Таким образом, большое

значение какой-либо оценки погрешности не всегда свидетельствует о том, что в действительности величина погрешности велика.

В частности, переход к оценкам через нормы производных в метриках L_p при $p \neq \infty$ иногда приводит к существенному уточнению оценки погрешности в многомерном случае.

Возникает вопрос: можно ли получить лучшие по порядку оценки погрешности при тех же предположениях о подынтегральной функции или хотя бы улучшить константы в этих оценках. Изучение этого вопроса приводит к задаче построения квадратур с оптимальной оценкой погрешности, или, как говорят, *оптимальных квадратур*.

§ 7. О постановках задач оптимизации

Мы получили ряд формул численного интегрирования и могли бы получить еще большее количество таких формул. В связи с этим возникает следующая проблема — чем больше способов решения задачи, тем труднее решиться выбрать какой-то из них; каждый может иметь свои положительные качества: простота программирования, малое время работы ЭВМ, малая загрузка памяти, простота оценки погрешности, применимость к широкому кругу задач. Таким образом, излишняя информация о способах решения задач при чрезмерном ее росте может также и затормозить их реальное решение. Поэтому необходима какая-то систематизация методов решения, их отбор. Естественно пытаться решать задачи наилучшим, оптимальным способом. Мы уже рассматривали некоторые модельные формулировки задач об оптимальных методах решения на примере вычисления значений функций; при этом рассмотрении возникают определенные математические постановки задач, требующих решения.

При рассмотрении каждой новой проблемы представляется желательным получить ее наиболее подробное, наилучшее описание и затем решить возникшую задачу наилучшим образом. Однако обычно это не удается и приходится довольствоваться меньшим: описать проблему наилучшим образом и решить ее удовлетворительно или описать проблему удовлетворительно и затем полностью решить возникшую задачу. При рассмотрении проблемы оптимизации методов можно говорить о выборе между удовлетворительным решением проблемы оптимизации методов для классов задач, близких к реальным, или полным решением проблемы для эталонных математических классов, подобных рассматриваемому в следующем параграфе. Может вызвать недоумение высказывание об «удовлетворительном решении проблемы оптимизации методов» для каких-то классов;

ведь задача оптимизации методов на классе сводится к вполне конкретной математической задаче, которую, по-видимому, можно решить окончательно, а не «удовлетворительно».

В принципе это высказывание верно, однако обычно полностью решить задачу в приемлемое для практики время не удастся: время, необходимое для построения оптимального метода, обычно существенно превосходит время, в течение которого возникает новое, уточненное, описание классов рассматриваемых задач. Также надо иметь в виду, что не всегда удается формализовать это математическое описание класса реальных задач: имеется какое-то качественное представление о классе, имеются неплохие численные методы и интуитивно ясно, что в практическом плане оптимальность методов достигнута; в то же время нет даже четко формализованного описания класса решаемых задач. Примером такой обстановки является положение с вопросом интегрирования кусочно-гладких функций, которое в неформализованном виде будет рассмотрено в § 17. Точно так же, например при анализе физических моделей, можно описать задачу с помощью сложной системы дифференциальных уравнений и решить эту систему с малой точностью или описать с помощью простой системы, и решить эту систему с большой точностью.

В разных случаях, в зависимости от конкретной обстановки, бывает целесообразен тот или иной подход к решению задачи. Представляется, что для прикладной науки часто более существенно не окончательное решение вариационных задач, а правильная их постановка. Обычно больший эффект дает удовлетворительное решение правильно поставленной в практическом смысле вариационной задачи, чем полное решение вариационной задачи, не охватывающей все существенные характерные черты задачи.

Что такое оптимальный метод решения задачи?

Под оптимальным решением задачи можно понимать решение, требующее минимальной затраты машинного времени. Это будет неправильная постановка, поскольку в принципе всякую задачу можно решить без применения машин, затратив на это очень большое количество времени.

Под оптимальным можно понимать решение, требующее минимальных затрат времени (или, соответственно, материальных затрат). При этом следует помнить, что время и затраты на поиск оптимального алгоритма уже входят в решение задачи.

Когда мы ставим вопрос о решении задачи оптимальным образом, мы не учитываем разных возможностей исследователей, решающих задачу. А ведь, в зависимости от индивидуальных возможностей исследователя (наличие ЭВМ, библиотеки, лаборанта, сотрудников, обладающих опытом решения подобных

задач, и, наконец, его способностей), оптимальное решение, вообще говоря, будет различным.

Мы говорили о вопросе оптимального решения каждой конкретной задачи. Такая постановка не совсем правильна также по следующей причине. На самом деле коллектив исследователей сталкивается с целой совокупностью задач. Мы можем поставить перед собой цель решить первую, вторую и т. д. задачи за кратчайшее время. Однако постановка вопроса совсем изменится, если мы зададимся целью решить с минимальными затратами целую совокупность задач. Мы можем решать первую, вторую и т. д. задачи оптимальными методами. Но в это время наука движется вперед, и если мы не будем изучать новых методов, создавать новые алгоритмы и стандартные программы, рассчитанные на решение будущих задач, то в целом мы проиграем. При наличии большой серии задач, не требующей сверхсрочного решения, будет выгоднее заняться теоретическими исследованиями, создать новые алгоритмы, а затем уже решать эти задачи.

Не имеет однозначного ответа также вопрос о том, кто должен заниматься оптимальным решением задач. Если мы имеем дело с единственной конкретной задачей, заведомо не требующей нового алгоритма и больших затрат машинного времени, ее лучше выполнить работнику низкой квалификации. Увеличение числа задач и их сложности требует привлечения работников высокой квалификации, поскольку здесь их отдача будет наиболее полной. Конечно, важно учесть и роль изучения опыта решения задач как сложных, так и простых. Для того чтобы задачу оптимизации методов можно было рассматривать как чистоматематическую задачу, необходимо определить целевую функцию исследования, класс рассматриваемых задач и возможности исследователя.

Качество какого-то алгоритма на классе задач мы характеризуем его качеством на самой плохой для него задаче этого класса. Поэтому чем уже класс, тем лучше качество алгоритмов на этом классе.

Казалось бы, что следует построить оптимальные алгоритмы для возможно большего числа классов задач. Однако при излишней детализации классов задач слишком много усилий уйдет на отыскание оптимальных на этих классах способов решения задач. В этом случае может не хватить сил и времени на решение самих задач, разработку и создание стандартных программ. (Конечно, эти замечания против детализации задач в меньшей мере относятся к случаю решения задач на специализированных машинах и устройствах, где следует максимально сузить класс рассматриваемых задач и рассмотреть проблему всесторонне.)

При первом взгляде на проблему кажется, что в практике вычислений всегда имеют дело с конкретными задачами и никогда не рассматривают класс задач. По этому поводу можно сказать следующее: действительно, при выборе метода решения задачи исследователь не относит задачу формально к какому-либо классу; однако метод решения всегда выбирается в зависимости от некоторых типичных свойств задачи: тип дифференциального уравнения, наличие особенности у решения, число конечных производных и т. п. При выборе алгоритма исследователь учитывает эти свойства и тем самым вольно или невольно относит рассматриваемую задачу к классу задач, обладающих этими свойствами.

При анализе задач и разбиении их на классы часто возникает вопрос: какому классу задач уделить первоочередное внимание, в частности, на какой класс задач следует рассчитывать при составлении стандартной программы численного интегрирования? Пусть, например, при помощи этой программы будут вычисляться интегралы как от гладких, так и от не очень гладких функций (это, конечно, условное деление) примерно с одинаковой частотой. На какие функции нужно ориентироваться при выборе алгоритма? Ответ на этот вопрос можно получить из следующих соображений. На гладких функциях алгоритм будет работать более эффективно; поэтому время, затрачиваемое на их решение, будет меньше времени, затрачиваемого на решение задач с негладкими функциями. 50% экономии времени при вычислении интегралов от гладких функций принесет меньшую выгоду, чем 50% экономии времени при вычислении интегралов от негладких функций, и следовательно, целесообразно уделить большее внимание эффективности алгоритма в случае негладких функций. Конечно, вывод будет другим, если доля негладких функций мала.

При отыскании оптимального метода решения класса задач в случае, когда предполагается непосредственное внедрение алгоритма в вычислительную практику, обычно возникает также следующая проблема: нам не удастся сразу найти оптимальный метод решения задач рассматриваемого класса. После некоторого времени работы мы находим какой-то метод, иногда близкий к оптимальному, иногда просто лучший, чем ранее известные, а затем постепенно усовершенствуем его. В какой момент надо остановиться в усовершенствовании метода и перейти к составлению стандартных программ решения задач этого класса? При ответе на этот вопрос надо учитывать следующие простые соображения. Если мы поспешим и быстро начнем внедрять только что полученный алгоритм, то, может быть, нам придется вскоре создавать новые алгоритмы из-за плохого качества этого алгоритма. Затягивание времени при составлении стандартных

программ также нежелательно: при этом мы допустим большой перерасход времени при решении задач по имеющимся стандартным программам.

Можно возразить, что создание совершенных стандартных программ приведет к большой экономии машинного времени в будущем. Однако при анализе этого возражения в свою очередь следует учесть, что машинное время становится все менее дефицитным и более дешевым. Кроме того, чем раньше мы решим ту или иную задачу, тем больший реальный выход получает общество в целом.

Нужно учитывать также важность быстрого внедрения вновь созданных алгоритмов для самой задачи отыскания оптимальных методов решения. Дело в том, что анализ результатов расчетов часто может указать на необходимость изменения класса рассматриваемых задач и открыть путь для новых теоретических исследований. Мы видим, что подход к проблеме оптимальности должен носить динамический характер: должны строиться оптимальные методы или близкие к ним для все новых классов задач, предъявляемых наукой и техникой при изменении возможностей, предоставляемых ЭВМ.

При этом необходима как текущая работа, так и работа по доведению предложенных ранее постановок до окончательного решения. Для численной математики, как и для всякой прикладной науки, характерна следующая обстановка. Обычно задачи новых типов предъявляются сначала в незначительном количестве, требуется срочное их решение любой ценой; затраты идут не в счет. На первом этапе решения применяется первый попавшийся приемлемый метод. Далее эти задачи поступают в большом количестве, производится более или менее удовлетворительная постановка задачи оптимизации методов и находится некоторое ее решение. Затем задача переводится на поток, т. е. решение задач этого типа производится при помощи систем соответствующих стандартных программ. Не следует думать, что на этом этапе полностью кончается исследование данного типа задач — чтобы создавать эффективные методы решения новых задач, нужно осмысливать те задачи, которые остались несколько позади, и производить их теоретические обследования.

Соотношение между текущей и перспективной работой также является важнейшим фактором жизнедеятельности любой научной организации. В каждый момент времени организации предъявляются некоторые требования, выполнение которых необходимо для ее существования: самоокупаемость в случае хозрасчетных организаций, своевременная отчетность по годовому плану и т. п. Однако существуют и критические моменты времени, когда могут предъявляться повышенные требования в виде требований своевременного решения новых классов задач. По-

этому при планировании работы должны предусматриваться какие-то теоретические разработки впрок, резервы людей, машинного времени.

Вернемся к вопросу об оптимизации методов.

Часто математик, создав оптимальный или близкий к нему метод решения задачи, сетует на то, что этот метод плохо внедряется в вычислительную практику. Ответ на этот вопрос может быть самым разным. Очень часто это происходит вследствие консерватизма практических работников, их желания работать старыми, привычными методами. Иногда это объясняется недостатками самого метода. Например, случается, что кроме (и даже вместо) оптимальности метода желательны и существенны простота метода и возможность надежного контроля точности получаемых результатов. Может случиться, что сам класс рассмотренных задач не совпадает с классом, к которому относится большинство реальных задач, поступающих для решения.

Конечно, нужно учитывать также вопрос о том, насколько широк возможный круг решаемых задач рассматриваемого класса. Если сейчас и в перспективе ожидается решение небольшого числа задач рассматриваемого класса, разработка оптимальных алгоритмов и создание стандартных программ решения задач этого класса могут себя не оправдать. В то же время изучение задачи оптимизации методов на различных классах часто полезно тем, что при ее решении возникают новые методы, которые затем находят широкое применение и при решении задач из других классов. Примером, подтверждающим это высказывание, может служить разработка теоретико-числовых методов интегрирования.

§ 8. Оптимальные квадратуры на классах функций с одной производной

Рассуждения § 7 показывают важность изучения различных постановок проблемы оптимизации методов на классах задач.

Рассмотрим задачу вычисления интеграла

$$I(f) = \int_{\Omega} f(P) p(P) dP.$$

Область интегрирования Ω и весовая функция $p(P)$ считаются фиксированными. Класс рассматриваемых задач определим заданием класса F подынтегральных функций. Погрешностью квадратуры

$$I(f) \approx S_N(f) = \sum_{j=1}^N D_j f(P_j) \quad .$$

на классе F называют величину

$$R_N(F) = \sup_{f \in F} |R_N(f)|,$$

где, как обычно,

$$R_N(f) = I(f) - S_N(f).$$

Нижняя грань

$$W_N(F) = \inf_{D_I, P_I} R_N(F)$$

называется *оптимальной оценкой погрешности квадратур на рассматриваемом классе*. Если существует квадратура, для которой $R_N(F) = W_N(F)$, то такую квадратуру называют *оптимальной, или наилучшей, на рассматриваемом классе*.

Обратимся к случаю

$$I(f) = \int_0^1 f(x) dx, \quad f \in F = C_1(A, [0, 1]);$$

$C_1(A, [0, 1])$ — класс непрерывных функций с кусочно-непрерывной первой производной, удовлетворяющей условию $|f'(x)| \leq A$.

Квадратурную сумму записываем в виде

$$S_N(f) = \sum_{j=1}^N D_j f(x_j), \quad x_1 < \dots < x_N.$$

Потребуем выполнения условия

$$\sum_{j=1}^N D_j = 1. \quad (1)$$

Если (1) не выполнено, то при $f(x) = c = \text{const}$ имеем

$$R_N(f) = \left(1 - \sum_{j=1}^N D_j\right) c \neq 0.$$

Все функции $f(x) = \text{const}$ принадлежат рассматриваемому классу, и следовательно,

$$R_N(F) \geq \sup_c \left(\left| 1 - \sum_{j=1}^N D_j \right| |c| \right) = \infty.$$

Ясно, что об оптимальности такой квадратуры говорить не приходится. Квадратура, удовлетворяющая условию (1), точна для постоянных, т. е. всех многочленов нулевой степени. Для таких квадратур, согласно (2.6), справедливо соотношение

$$R_N(C_1(A, [0, 1])) = A \int_0^1 |K_N(y)| dy,$$

где

$$K_N(y) = 1 - y - \sum_{j=1}^N D_j (\overline{x_j - y})^0,$$

$$(\bar{t})^0 = \begin{cases} 1 & \text{при } t > 0, \\ 0 & \text{при } t \leq 0. \end{cases}$$

Для $x_{m-1} \leq y < x_m$ имеем

$$(\overline{x_j - y})^0 = \begin{cases} 0 & \text{при } j < m, \\ 1 & \text{при } j \geq m, \end{cases}$$

(здесь мы положили $0^0 = 0$); поэтому

$$\begin{aligned} \sum_{j=1}^N D_j (\overline{x_j - y})^0 &= \\ &= \begin{cases} 1 & \text{при } y < x_1, \\ Q_m = \sum_{j=m}^N D_j & \text{при } x_{m-1} \leq y < x_m, \quad m = 2, \dots, N, \\ 0 & \text{при } x_N \leq y. \end{cases} \quad (2) \end{aligned}$$

Мы видим, что исходная задача минимизации погрешности на классе свелась к решению следующей задачи: приблизить наилучшим образом в метрике L_1 $\|f\| = \int_0^1 |f(x)| dx$ функцию $1 - y$ функциями вида (2). Обозначим

$$\int_0^1 |K_N(y)| dy \quad \text{через} \quad V(Q_2, \dots, Q_N; x_1, \dots, x_N).$$

Имеем равенство

$$\begin{aligned} V(Q_2, \dots, Q_N; x_1, \dots, x_N) &= \\ &= \int_0^{x_1} |y| dy + \sum_{m=2}^N \int_{x_{m-1}}^{x_m} |(1-y) - Q_m| dy + \int_{x_N}^1 |1-y| dy. \quad (3) \end{aligned}$$

От фиксированного Q_m зависит только одно слагаемое

$$V(Q_m) = \int_{x_{m-1}}^{x_m} |(1-y) - Q_m| dy.$$

Непосредственным интегрированием получим

$$V(Q_m) = \begin{cases} \frac{(x_m - x_{m-1})^2}{2} + (Q_m - (1 - x_{m-1}))(x_m - x_{m-1}) & \text{при } 1 - x_{m-1} \leq Q_m, \\ \frac{(1 - Q_m - x_m)^2 + (1 - Q_m - x_{m-1})^2}{2} & \text{при } 1 - x_m \leq Q_m \leq 1 - x_{m-1}, \\ \frac{(x_m - x_{m-1})^2}{2} - (Q_m - (1 - x_m))(x_m - x_{m-1}) & \text{при } Q_m \leq 1 - x_m. \end{cases}$$

Отсюда

$$V'(Q_m) > 0 \quad \text{при} \quad Q_m > 1 - \frac{x_{m-1} + x_m}{2},$$

$$V'(Q_m) < 0 \quad \text{при} \quad Q_m < 1 - \frac{x_{m-1} + x_m}{2},$$

следовательно,

$$\min_{Q_m} V(Q_m) = V\left(1 - \frac{x_{m-1} + x_m}{2}\right) = \frac{(x_m - x_{m-1})^2}{4}. \quad (4)$$

Кроме величин $V(Q_m)$ правая часть (3) содержит также слабое

$$\int_0^{x_1} |y| dy + \int_{x_N}^1 |1 - y| dy = \frac{x_1^2 + (1 - x_N)^2}{2}.$$

Отсюда, из (3) и (4) следует, что

$$\begin{aligned} \inf_{Q_2, \dots, Q_N} V(Q_2, \dots, Q_N; x_1, \dots, x_N) = \\ = \bar{V}(x_1, \dots, x_N) = \frac{x_1^2 + (1 - x_N)^2}{2} + \sum_{m=2}^N \frac{(x_m - x_{m-1})^2}{4}. \end{aligned}$$

Таким образом,

$$\begin{aligned} \inf_{Q_2, \dots, Q_N; x_1, \dots, x_N} V(Q_2, \dots, Q_N; x_1, \dots, x_N) = \\ = \inf_{x_1, \dots, x_N} \bar{V}(x_1, \dots, x_N). \end{aligned}$$

Приравняв нулю производные $\frac{\partial \bar{V}}{\partial x_m}$, получим систему уравнений

$$\begin{aligned} \frac{3x_1 - x_2}{2} = 0, \quad \frac{3x_N - 2 - x_{N-1}}{2} = 0, \quad \frac{x_{m-1} - 2x_m + x_{m+1}}{2} = 0, \\ m = 2, \dots, N-1. \end{aligned}$$

Решением этой системы будет

$$x_m = \frac{2m-1}{2N}, \quad m=1, \dots, N,$$

и при этих значениях

$$\bar{V}(x_1, \dots, x_N) = \frac{1}{4N}.$$

Мы получили точку экстремума функции $\bar{V}$, но это еще не означает, что мы нашли ее нижнюю грань; функция $\bar{V}(x_1, \dots, x_N)$ рассматривается в области $0 \leq x_1 < x_2 < \dots < x_N \leq 1$, и если ее нижняя грань достигается в некоторой точке границы, то выполнение условий $\frac{\partial \bar{V}}{\partial x_m} = 0$ в этой точке не является обязательным.

Проверим непосредственно, что полученное значение $\bar{V}$ является минимальным. В неравенство Коши — Буняковского

$$\left| \sum_{j=1}^{2N} a_j b_j \right|^2 \leq \left(\sum_{j=1}^{2N} a_j^2 \right) \left(\sum_{j=1}^{2N} b_j^2 \right)$$

подставим

$$a_1 = 1, \quad b_1 = x_1, \quad b_2 = b_3 = \frac{x_2 - x_1}{2},$$

$$b_4 = b_5 = \frac{x_3 - x_2}{2}, \dots, \quad b_{2N-2} = b_{2N-1} = \frac{x_N - x_{N-1}}{2}, \quad b_{2N} = 1 - x_N,$$

очевидно, $\sum_{j=1}^{2N} a_j b_j = 1$. В результате получим

$$1 \leq 2N \sum_{j=1}^{2N} b_j^2$$

или

$$\frac{1}{4N} \leq \frac{1}{2} \left(\sum_{j=1}^{2N} b_j^2 \right) = \bar{V}(x_1, \dots, x_N).$$

Итак, минимальное значение $\int_0^1 |K_N(y)| dy$, равное $\frac{1}{4N}$, достигается при

$$x_j = \frac{2j-1}{2N}, \quad Q_j = 1 - \frac{x_{j-1} + x_j}{2} = 1 - \frac{j-1}{N}.$$

Для соответствующей квадратурной формулы

$$D_1 = 1 - Q_1 = \frac{1}{N}, \quad D_N = Q_N = \frac{1}{N},$$

$$D_j = Q_j - Q_{j+1} = \frac{1}{N}, \quad j=2, \dots, N-1.$$

Таким образом, оптимальной на рассматриваемом классе является квадратура

$$\int_0^1 f(x) dx \approx \frac{1}{N} \sum_{j=1}^N f\left(\frac{2j-1}{2N}\right)$$

с оценкой остаточного члена $\frac{A}{4N}$.

Рассмотрим в качестве примера еще один класс функций $C_1^{(1)}(A, [0, 1])$ — класс непрерывных кусочно-дифференцируемых функций, удовлетворяющих условию $\int_0^1 |f'(x)| dx \leq A$.

Согласно (6.10), имеем оценку погрешности

$$R_N(C_1^{(1)}(A, [0, 1])) \leq \sup_{[0, 1]} |K_N(y)| \int_0^1 |f'(x)| dx. \quad (5)$$

Покажем, что на самом деле здесь имеет место равенство. Функция $K_N(y) = (1-y) - Q_m$ линейная на $[x_{m-1}, x_m]$, поэтому $\sup_{[x_{m-1}, x_m]} |K_N(y)| = \max(|1-x_{m-1}-Q_m|, |1-x_m-Q_m|)$, кроме того, $\sup_{[0, x_1]} |K_N(y)| = \sup_{[0, x_1]} |y| = x_1$, $\sup_{[x_N, 1]} |K_N(y)| = \sup_{[x_N, 1]} |1-y| = 1-x_N$.

Таким образом,

$$\sigma = \sup_{[0, 1]} |K_N(y)| = \max(x_1, |1-x_1-Q_2|, |1-x_2-Q_2|, \dots, 1-x_N). \quad (6)$$

Вследствие линейности $K_N(y)$ на каждом из промежутков $[0, x_1], [x_1, x_2], \dots, [x_{N-1}, x_N], [x_N, 1]$ справедливо следующее утверждение: для любого ε можно указать некоторый отрезок $[y_1, y_2]$, $y_2 > y_1$, принадлежащий одному из указанных выше сегментов, такой, что на этом отрезке

$$|K_N(y)| \geq \sup |K_N(y)| - \varepsilon, \quad \text{sign } K_N(y) = \text{const.}$$

Возьмем

$$f_\varepsilon(x) = \begin{cases} 0 & \text{при } x \leq y_1, \\ A \frac{x-y_1}{y_2-y_1} & \text{при } y_1 \leq x \leq y_2, \\ 1 & \text{при } x \geq y_2, \end{cases}$$

тогда $f'_\varepsilon(x) = \frac{A}{y_2-y_1}$ при $x \in (y_1, y_2)$, $f'_\varepsilon(x) = 0$ при $x \notin [y_1, y_2]$. Для такой функции имеем

$$R_N(f_\varepsilon) = \int_{y_1}^{y_2} \frac{K_N(y) A}{y_2-y_1} dy,$$

и следовательно,

$$|R_N(f_\varepsilon)| \geq A \left(\sup_{[0, 1]} |K_N(y)| - \varepsilon \right).$$

Вследствие произвольности ε справедливо неравенство

$$R_N(C_1^{(1)}(A, [0, 1])) \geq A \sup_{[0, 1]} |K_N(y)|.$$

Отсюда и из (5) получаем

$$R_N(C_1^{(1)}(A, [0, 1])) = A \sup_{[0, 1]} |K_N(y)|.$$

Имеем очевидное равенство

$$\begin{aligned} & \{x_1\} + \{(1 - x_1 - Q_2) + \{-(1 - x_2 - Q_2)\}\} + \dots \\ & \dots + \{(1 - x_{N-1} - Q_N) + \{-(1 - x_N - Q_N)\}\} + \{1 - x_N\} = \\ & = x_1 + (x_2 - x_1) + \dots + (x_N - x_{N-1}) + 1 - x_N = 1. \end{aligned}$$

Максимум модулей слагаемых, стоящих в фигурных скобках, согласно (6), равен $\sup |K_N(y)|$. Если сумма $2N$ слагаемых равна 1, то максимум модуля этих слагаемых не менее $\frac{1}{2N}$. Поэтому $\sup_{[0, 1]} |K_N(y)| \geq \frac{1}{2N}$. Максимум модуля этих слагаемых будет равен $\frac{1}{2N}$ в том и только том случае, если все они равны между собой и равны $\frac{1}{2N}$. Следовательно, если существует квадратурная формула, для которой

$$\sup_{[0, 1]} |K_N(y)| = \frac{1}{2N},$$

то обязательно

$$x_1 = \frac{1}{2N}, \quad 1 - x_1 - Q_2 = \frac{1}{2N}, \quad Q_2 + x_2 - 1 = \frac{1}{2N}, \quad \dots, \quad 1 - x_N = \frac{1}{2N}.$$

Подставляя $x_1 = \frac{1}{2N}$ во второе уравнение, получим $Q_2 = 1 - \frac{1}{N}$, затем подставляем это значение Q_2 в третье уравнение, получим $x_2 = \frac{3}{2N}$ и т. д. Итак,

$$x_j = \frac{2j-1}{2N}, \quad Q_j = 1 - \frac{j-1}{N}.$$

Этим значениям Q_j , как и ранее, соответствуют $D_j = \frac{1}{N}$. Таким образом, мы получили оптимальную на рассматриваемом классе квадратуру

$$\int_0^1 f(x) dx \approx \frac{1}{N} \sum_{j=1}^N f\left(\frac{2j-1}{2N}\right)$$

с оценкой остаточного члена $\frac{A}{2N}$.

В настоящее время оптимальные квадратуры получены для небольшого набора классов функций, в основном, одной переменной. Непосредственное значение этих квадратур для приложений ничтожно; однако это не значит, что не нужно заниматься построением таких квадратур.

Представители других, особенно гуманитарных, наук привыкли говорить, что математика количественная наука, «математики считают», в то время как их науки занимаются качественными выводами.

Это в корне неправильное мнение. Конечно, в настоящее время математика оказывает влияние на многие стороны жизни общества, поставляя многообразные результаты расчетов: прочность конструкций, оптимальные планы и т. д. Однако, кроме этого, математика является существенным аппаратом качественного исследования явлений природы и общества. Зачастую имеется много противоречащих друг другу моделей этих явлений, выбор между которыми не может быть сделан путем теоретических построений. Обсчет моделей на ЭВМ может дать ответ о соответствии их реальности, и здесь выводы математики носят, в основном, качественный характер. Ценность рассмотрения некоторых математических моделей экономики часто выражается не в выработке оптимального плана, а в выработке, в основном, качественной рекомендации, например, о направлении изменения структуры топливного баланса или об изменении системы оплаты с целью большего стимулирования развития соответствующей отрасли.

Точно так же проводившиеся выше построения оптимальных квадратур и дальнейшее их развитие на случай функций большей гладкости и большего числа переменных ценны не получением конкретных квадратурных формул, а выяснением качественной стороны вопроса: где какие методы лучше, на какую точность можно рассчитывать при использовании определенной информации о подынтегральной функции, какова плотность распределения узлов у хороших квадратур.

Например, мы убедились, что, используя информацию только об ограниченности первой производной, нельзя получить оценок погрешности, по порядку лучших $O\left(\frac{1}{N}\right)$. Таким образом, оценки погрешности, полученные в § 6 при $n = 1$, по порядку числа узлов не могут быть улучшены. Мы убедились также, что предположение об ограниченности первой производной не может привести к существенному улучшению оценки погрешности по сравнению с предположением об ограниченности

$$\int_0^1 |f'(x)| dx.$$

§ 9. Оптимизация распределения узлов квadrатурной формулы

Развитие методов численного интегрирования могло бы пойти по пути создания оптимальных методов на различных классах $C_r^{(p)}(A, [a, b])$ и создания программ на основе этих методов. Очень скоро после начала рассмотрения задач оптимизации методов стало ясно, что на этих классах уже известные методы с оценками погрешности (6.7), (6.12) недалеко от оптимальных; как увидим в § 7 гл. V, за счет оптимизации квадратурной формулы оценки погрешности (6.7), (6.12) на рассматриваемых классах не могут быть улучшены по порядку. При этом также стало ясно, что некоторые методы, практически совпадающие с классическими методами Эйлера и Грегори, являются *асимптотически оптимальными* по оценке главного члена погрешности. Мы вкладываем в это понятие следующий смысл. Для этих методов на соответствующих классах функций были получены оценки погрешности

$$\frac{c_r}{N^r} + \frac{c_{r+1}}{N^{r+1}},$$

в то время как для оптимальных на этих классах методов оценка погрешности имеет вид

$$\frac{c_r}{N^r} + O\left(\frac{1}{N^{r+1}}\right).$$

Казалось бы, что так как есть почти оптимальные методы, то следующим этапом должен стать перевод всех систем интегрирования на использование этих оптимальных методов.

Однако подобное утверждение нельзя рассматривать как бесспорное. Вследствие большого многообразия задач, требующих решения, при переходе к этапу внедрения методов в практику всегда следует проявлять известную осторожность. Нельзя полностью ручаться, что принятое нами описание классов этих задач наилучшим образом соответствует классу реальных задач; например, следует признать, что классы C_1 плохо описывают реальные задачи. Конечно, вследствие привычки к традиционным методам внедрение новых методов обычно требует энергичных действий со стороны их приверженцев. В то же время надо иметь в виду, что старые методы прошли испытание временем и могут оказаться более пригодными для решения задач некоторых классов; поэтому отбрасывание старых методов должно производиться лишь после достаточного теоретического и практического анализа.

В случае рассматриваемой проблемы оптимизации методов интегрирования практические работники еще до полного

выяснения вопроса об оптимальности методов пришли к следующим заключениям.

При разбиении исходного отрезка интегрирования на одинаковые $a_q - a_{q-1} \equiv H$ мы получаем информацию о подынтегральной функции равномерно по всему отрезку интегрирования. В то же время подынтегральная функция может быть более гладкой на части отрезка интегрирования и ясно, что там следует поместить относительно меньшее количество узлов. Поэтому для практически оптимальных методов разбиение отрезка интегрирования на части должно быть приспособлено к специфике поведения подынтегральной функции.

Рассмотрим возможные постановки задачи распределения узлов в зависимости от особенностей поведения производной подынтегральной функции. Эти постановки имеют много общего, однако для конкретных задач тот или иной подход иногда оказывается более удобным. Для простоты изложения мы будем проводить рассмотрение на примере формулы трапеций.

Пусть вычисляется интеграл

$$I(f) = \int_0^1 f(x) dx$$

и подынтегральная функция удовлетворяет условиям $|f''(x)| \leq A_l$ на отрезках (B_{l-1}, B_l) , $l = 1, \dots, q$, где $0 = B_0 < B_1 < \dots < B_q = 1$. Для вычисления интегралов

$$I_l(f) = \int_{B_{l-1}}^{B_l} f(x) dx$$

применим составную формулу трапеций с равными отрезками разбиения длины $H_l = \frac{b_l}{N_l}$, $b_l = B_l - B_{l-1}$:

$$I_l(f) \approx S_l(f) = H_l \left(\frac{f(B_{l-1})}{2} + f(B_{l-1} + H_l) + \dots + \frac{f(B_l)}{2} \right).$$

Согласно (6.8), оценка остаточного члена есть $\frac{A_l b_l^3}{12N_l^2}$. Суммарная

погрешность при замене $I(f)$ суммой $\sum_{l=1}^q S_l(f)$ не превзойдет

$$\Phi = \sum_{l=1}^q \frac{A_l b_l^3}{12N_l^2}.$$

Поставим задачу: при заданном числе $N = N_1 + \dots + N_q$ отрезков разбиения распорядиться выбором величин N_l так, чтобы суммарная оценка погрешности Φ была минимальной.

Найдем минимум Φ при условии $\Psi = N_1 + \dots + N_q - N = 0$, не предполагая пока, что величины N_l — целые. Приравнявая нулю производные функции Лагранжа $\Phi + \lambda\Psi$, получаем систему уравнений

$$\frac{\partial(\Phi + \lambda\Psi)}{\partial N_l} = -\frac{A_l b_l^3}{6N_l^3} + \lambda = 0.$$

Отсюда

$$N_l = b_l \sqrt[3]{\frac{A_l}{6\lambda}}. \quad (1)$$

Из условия $\Psi = 0$ получаем уравнение относительно λ :

$$\left(\sum_{l=1}^q b_l \sqrt[3]{\frac{A_l}{6\lambda}} \right) = N.$$

Определяем λ и затем из (1) находим N_l . Поскольку N_l должны быть целыми, то возьмем

$$N_l = N_l^0 = \left[b_l \sqrt[3]{\frac{A_l}{6\lambda}} \right]. \quad (2)$$

Очевидно,

$$N - q \leq N_1^0 + \dots + N_q^0 \leq N.$$

Мы не нашли настоящего минимума Φ по множеству всевозможных целых $N_1, \dots, N_q$, удовлетворяющих условию $N_1 + \dots + N_q = N$, однако дальнейшее уточнение вряд ли разумно.

Обычно практический интерес представляет другая вариационная задача. Найти минимальное значение $N = N_1 + \dots + N_q$, при котором оценка погрешности Φ не превосходит ε . Поскольку Φ монотонно убывает с ростом величин N_l , то достаточно ограничиться случаем $\Phi = \varepsilon$. Возьмем функцию Лагранжа в виде

$$N_1 + \dots + N_q + \lambda^{-1}(\Phi - \varepsilon).$$

Приравнявая нулю ее производные по N_l , получим те же уравнения (1); подставляя значение $N_l = b_l \sqrt[3]{\frac{A_l}{6\lambda}}$ в уравнение $\Phi = \varepsilon$, получим

$$\lambda^{2/3} \sum_{l=1}^q \left(\frac{A_l}{48} \right)^{1/3} b_l = \varepsilon.$$

Определяем λ , а затем соответствующие N_l . Соотношения относительно величин N_l , получившиеся при рассмотрении этих двух задач, имеют одинаковый вид. Поэтому для получения

качественных выводов об оптимальном распределении узлов достаточно ограничиться рассмотрением одной из этих задач.

Нашей целью является разработка оптимальных методов решения и разработка на их основе систем программ решения типовых математических задач. Можно представить себе программу вычисления интеграла с заданной точностью, работающую по следующей схеме. Производится вычисление таблицы значений функции на некоторой сетке $x_0, \dots, x_n$. По этой таблице составляется таблица разделенных разностей

$$f(x_0; x_1; x_2), \dots, f(x_{n-2}; x_{n-1}; x_n).$$

Из рассмотрения этой таблицы делается вывод о наиболее целесообразном разбиении отрезка на части (B_{l-1}, B_l) и значениях A_l , соответствующих этим частям. Затем, в соответствии с (1), выбираются значения N_l и производится интегрирование. Имеются работающие программы, реализующие алгоритм вычисления интеграла, лишь несущественно отличающийся от описанного. По-видимому, при решении задач некоторых классов такие программы могут оказаться наиболее эффективными.

Однако большинство алгоритмов реально работающих стандартных программ базируется не на таком непосредственном использовании полученных соотношений, а на одном качественном выводе, являющемся следствием (1). Перепишем это равенство в виде

$$\frac{A_l}{12} \left(\frac{b_l}{N_l} \right)^3 = \frac{\lambda}{2}.$$

Левая часть этого выражения равна оценке погрешности по элементарному отрезку интегрирования длины $\frac{b_l}{N_l}$, на которые был разбит отрезок $[B_{l-1}, B_l]$. Таким образом, это соотношение означает, что при оптимальном распределении узлов интегрирования оценки погрешностей, приходящихся на элементарные отрезки интегрирования, должны быть одинаковыми.

Для получения этого вывода достаточно было ограничиться рассмотрением случая $q = 2$. Это обстоятельство подтверждает общее свойство качественных характеристик методов решения задач (не обязательно математических): для их получения достаточно ограничиться рассмотрением простейших моделей, учитывающих основные стороны явления.

Рассмотрим еще одну, близкую постановку задачи оптимизации распределения узлов интегрирования. Чтобы не утомлять читателя второстепенными деталями, мы не будем проводить подробных оценок членов высшего порядка в оценке погрешности.

Пусть отрезок интегрирования $[0, 1]$ разбит на части $[a_{q-1}, a_q]$, $q = 1, \dots, N$, $a_0 = 0$, $a_N = 1$ и интеграл по каждой части вычисляется по формуле трапеций

$$I_q(f) = \int_{a_{q-1}}^{a_q} f(x) dx \approx s_q(f) = \frac{a_q - a_{q-1}}{2} (f(a_{q-1}) + f(a_q)).$$

Тогда интеграл по всему отрезку $[0, 1]$ вычисляется по формуле

$$I(f) \approx \sum_{q=1}^N s_q(f)$$

с оценкой остаточного члена

$$r = \sum_{q=1}^N \left(\max_{[a_{q-1}, a_q]} |f''(x)| \right) \frac{(a_q - a_{q-1})^3}{12}. \quad (3)$$

Пусть известно, что $|f''(x)| \leq F(x)$ на $[0, 1]$, где $F(x)$ непрерывна; пусть в качестве a_q взяты значения $\varphi\left(\frac{q}{N}\right)$ непрерывно дифференцируемой функции φ , удовлетворяющей условиям $\varphi(0) = 0$, $\varphi(1) = 1$. Очевидно,

$$a_q - a_{q-1} = \varphi\left(\frac{q}{N}\right) - \varphi\left(\frac{q-1}{N}\right) = \varphi'\left(\frac{q}{N}\right) \frac{1}{N} + o\left(\frac{1}{N}\right) \text{ при } N \rightarrow \infty,$$

поэтому

$$\max_{[a_{q-1}, a_q]} |f''(x)| \leq \max_{[a_{q-1}, a_q]} F(x) = F(a_q) + o(1) = F\left(\varphi\left(\frac{q}{N}\right)\right) + o(1).$$

Из этих соотношений получаем

$$\frac{(a_q - a_{q-1})^3}{12} \left(\max_{[a_{q-1}, a_q]} |f''(x)| \right) \leq \varepsilon_q = \left(\varphi'\left(\frac{q}{N}\right) \right)^3 \frac{F\left(\varphi\left(\frac{q}{N}\right)\right)}{12N^3} + o\left(\frac{1}{N^3}\right).$$

Подставляя последние соотношения в (3), имеем

$$r \leq \bar{r} = \frac{1}{N^2} \left\{ \sum_{q=1}^N \frac{1}{N} \left(\varphi'\left(\frac{q}{N}\right) \right)^3 \frac{F\left(\varphi\left(\frac{q}{N}\right)\right)}{12} \right\} + o\left(\frac{1}{N^2}\right).$$

Выражение в фигурных скобках является квадратурной суммой Римана для интеграла

$$\int_0^1 (\varphi'(t))^3 \frac{F(\varphi(t))}{12} dt$$

от непрерывной функции. Следовательно,

$$\bar{r} = \frac{1}{N^2} \left(\int_0^1 (\varphi'(t))^3 \frac{F(\varphi(t))}{12} dt \right) + o\left(\frac{1}{N^2}\right). \quad (4)$$

Рассмотрим задачу минимизации первого, главного члена выражения (4). Для удобства решения уравнения Эйлера примем за независимую переменную функцию φ . Тогда коэффициент при $\frac{1}{N^2}$ в главном члене погрешности запишется в виде

$$\int_0^1 (t'(\varphi))^{-2} \frac{F(\varphi)}{12} d\varphi.$$

Уравнение Эйлера для функции, минимизирующей функционал

$$\int_0^1 G(\varphi, t, t') d\varphi,$$

имеет вид

$$\frac{d}{d\varphi} \left(\frac{\partial G}{\partial t'} \right) - \frac{\partial G}{\partial t} = 0. \quad (5)$$

В рассматриваемом случае $\frac{\partial G}{\partial t} = 0$, и поэтому из (5) следует $\frac{\partial G}{\partial t'} = \text{const}$. Подставляя конкретное значение функции G , получим

$$(t'(\varphi))^{-3} \frac{F(\varphi)}{6} = \text{const}$$

или

$$F(\varphi) (\varphi'(t))^3 = C_1. \quad (6)$$

Общее решение этого уравнения зависит от C_1 и еще некоторой постоянной C_2 . Значения этих постоянных можно определить из граничных условий $\varphi(0) = 0$, $\varphi(1) = 1$. Решение рассмотренной вариационной постановки может практически использоваться различными способами. Например, существуют программы, осуществляющие численное интегрирование (6) на сетке с шагом, существенно большим $\frac{1}{N}$, и затем распределяющие узлы в соответствии с полученным решением.

Из соотношения (6) можно сделать тот же вывод о равенстве оценок погрешностей на элементарных отрезках интегрирования при оптимальном распределении узлов. В самом деле, умножим (6) на $\frac{1}{12N^3}$, положим $t = \frac{q}{N}$ и заменим $F\left(\varphi\left(\frac{q}{N}\right)\right)$ и

$\frac{1}{N} \varphi' \left(\frac{q}{N} \right)$, соответственно, эквивалентными величинами $\max_{[a_{q-1}, a_q]} F(x)$, $a_q - a_{q-1}$. В результате получим

$$\left(\max_{[a_{q-1}, a_q]} F(x) \right) \frac{(a_q - a_{q-1})^3}{12} \approx \frac{C_1}{12N^3}. \quad (7)$$

Еще один путь практического использования уравнения (6) состоит в следующем. Пусть требуется вычислить большую серию интегралов с одинаковым характерным поведением подинтегральных функций. Выделим простейшую, модельную функцию, для которой задача оптимизации узлов может быть решена в явном виде, и далее будем производить интегрирование с распределением узлов, соответствующим этой функции. Если характер изменения функций рассматриваемой серии зависит от некоторого параметра, то этот параметр следует учесть при выборе модельной функции; естественно, что модельная функция не обязательно относится к рассматриваемому классу. Чем большее число задач предъявляется для решения, тем более могут быть оправданы затраты, связанные с удачным выбором и рассмотрением модельной задачи.

§ 10. Примеры оптимизации распределения узлов

Рассмотрим примеры решения уравнения (9.6) для конкретных задач.

Пример 1. Пусть вычисляется серия интегралов

$$I(b) = \int_0^1 f(b, x) dx,$$

b — параметр серии, $f(b, x) = x^b g(b, x)$, $-1 < b < 2$, $g(b, x)$ — гладкая, $g(b, 0) \neq 0$. Если $b \neq 0, 1$, то вторая производная $f_{xx}(b, x)$ не ограничена в окрестности точки 0; при выборе модельной задачи следует учесть эту специфику поведения подинтегральной функции. В окрестности $x = 0$ имеем

$$f_{xx}(b, x) = b(b-1)x^{b-2}g(b, 0) + O(x^{b-1}).$$

Таким образом, в окрестности точки $x = 0$ вторая производная f_{xx} пропорциональна второй производной функции x^b , поэтому функцию x^b естественно рассматривать в качестве модельной. Примем за $F(x)$ величину $|b(b-1)|x^{b-2}$, тогда уравнение (9.6) запишется в виде

$$|b(b-1)|\varphi^{b-2} \left(\frac{d\varphi}{dt} \right)^3 = C_1.$$

Отсюда

$$\left(\frac{3 \sqrt[3]{|b(b-1)|}}{b+1} \right) \varphi^{\frac{b+1}{3}} = C_1 t + C_2;$$

из условия $\varphi(0) = 0$ получаем, что $C_2 = 0$, из условия $\varphi(1) = 1$, что $\varphi^{\frac{b+1}{3}} = t$. Таким образом,

$$\varphi(t) = t^{\frac{3}{1+b}} \quad (1)$$

и для модельной задачи вычисления интеграла $\int_0^1 x^b dx$ оптимальным в рассматриваемом нами смысле является распределение узлов $a_q = \left(\frac{q}{N} \right)^{\frac{3}{1+b}}$.

Проведенные выше рассуждения не являются строгими, так как при получении оценки (9.4) предполагалась ограниченность оценки второй производной $F(x)$, а для рассматриваемой модельной задачи это условие не выполнено.

Приведем правдоподобные рассуждения в пользу того, что и в рассматриваемом примере разумно ориентироваться на уравнение (9.6) при выборе оптимального распределения узлов $\varphi(t)$. Пусть $F(x)$ непрерывна во всех точках $(0, 1]$, как в примере 1. Предположим, что распределение узлов, соответствующее некоторой функции $\varphi_0(t)$, таково, что главный член оценки погрешности интегрирования существует, не больше, чем у любого другого распределения, и имеет порядок $O(1/N^2)$. Покажем, что тогда $\varphi_0(t)$ должно быть решением (9.6). В самом деле, возьмем некоторое $\eta > 0$. Главная часть погрешности интегрирования по отрезку $[\varphi_0(\eta), 1]$ имеет вид $\frac{I_\eta(\varphi_0)}{12N^2}$, где

$$I_\eta(\varphi) = \int_{\eta}^1 F(\varphi(t)) (\varphi'(t))^3 dt.$$

Возьмем произвольную функцию $\varphi(t)$ рассматриваемого класса, совпадающую с $\varphi_0(t)$ при $t \leq \eta$. Очевидно, что погрешности интегрирования на отрезке $[0, \varphi(\eta)]$ для всех таких функций совпадают. Вследствие минимальности главного члена погрешности, соответствующего функции $\varphi_0(t)$, должно выполняться неравенство $I_\eta(\varphi) \geq I_\eta(\varphi_0)$. Отсюда следует, что функция $\varphi_0(t)$ должна удовлетворять уравнению (9.6) при $\eta < t \leq 1$. Вследствие произвольности η она должна удовлетворять уравнению при $0 < t \leq 1$.

На примере вычисления интеграла $\int_0^1 x^b dx$ при $-1 < b < 2$ покажем, что для $\varphi_0(t)$ — решения (9.6) при граничных условиях $\varphi_0(0) = 0$, $\varphi_0(1) = 1$, главный член оценки погрешности существует и действительно имеет порядок $\frac{1}{N^2}$. Соответствующая функция $\varphi_0(t)$, согласно (1), есть $t^{\frac{3}{1+b}}$; ясно,

что $d = \frac{3}{1+b} > 1$. По формуле Тейлора

$$a_q - a_{q-1} = \Phi'_0 \left(\frac{q}{N} \right) \frac{1}{N} - \Phi''_0(\theta_q) \frac{1}{2N^2}, \quad (2)$$

где $\frac{q-1}{N} \leq \theta_q \leq \frac{q}{N}$. Так как $\Phi''_0(t) = d(d-1)t^{d-2} > 0$, то

$$a_q - a_{q-1} \leq \Phi'_0 \left(\frac{q}{N} \right) \frac{1}{N} = d \left(\frac{q}{N} \right)^{d-1} \cdot \frac{1}{N}. \quad (3)$$

При $2 \leq q$ имеем

$$|\Phi''_0(\theta_q)| \leq \max \left(\left| \Phi''_0 \left(\frac{q-1}{N} \right) \right|, \left| \Phi''_0 \left(\frac{q}{N} \right) \right| \right) = O \left(\left(\frac{q}{N} \right)^{d-2} \right). \quad (4)$$

Из (2) и (4) следует, что

$$\begin{aligned} a_q - a_{q-1} &= d \left(\frac{q}{N} \right)^{d-1} \frac{1}{N} + O \left(\left(\frac{q}{N} \right)^{d-2} \frac{1}{N^2} \right) = \\ &= \left(d \left(\frac{q}{N} \right)^{d-1} \frac{1}{N} \right) \left(1 + O \left(\frac{1}{q} \right) \right) \quad \text{при } 2 \leq q. \end{aligned}$$

В рассматриваемом случае

$$F(x) = |b(b-1)|x^{b-2}, \quad |F'(x)| = |b(b-1)(b-2)|x^{b-3}$$

возрастает. Поэтому

$$\max_{[a_{q-1}, a_q]} |F(x) - |b(b-1)| \left(\frac{q}{N} \right)^{d(b-2)}| \leq |F'(a_{q-1})| (a_q - a_{q-1}).$$

При $2 \leq q$ с учетом (3) имеем

$$|F'(a_{q-1})| (a_q - a_{q-1}) \leq \text{const} \left(\frac{q}{N} \right)^{d-1+d(b-3)} \frac{1}{N}.$$

Отсюда получаем

$$\begin{aligned} \max_{[a_{q-1}, a_q]} F(x) &= |b(b-1)| \left(\frac{q}{N} \right)^{d(b-2)} + O \left(\left(\frac{q}{N} \right)^{d(b-2)-1} \frac{1}{N} \right) = \\ &= |b(b-1)| \left(\frac{q}{N} \right)^{d(b-2)} \left(1 + O \left(\frac{1}{q} \right) \right). \end{aligned}$$

Следовательно,

$$\begin{aligned} \max_{[a_{q-1}, a_q]} F(x) \frac{(a_q - a_{q-1})^3}{12} &= \\ &= \frac{|b(b-1)|}{12} \left(\frac{q}{N} \right)^{d(b-2)} \left(d \left(\frac{q}{N} \right)^{d-1} \frac{1}{N} \right)^3 \left(1 + O \left(\frac{1}{q} \right) \right), \quad (5) \end{aligned}$$

поскольку $b = \frac{3-d}{d}$, то показатель степени, в которую возводится $\frac{q}{N}$, равен

$$3(d-1) + d(b-2) = 3(d-1) + d \left(\frac{3(1-d)}{d} \right) = 0.$$

Таким образом,

$$\max_{[a_{q-1}, a_q]} F(x) \frac{(a_q - a_{q-1})^3}{12} = \frac{d^3 |b(b-1)|}{12N^3} \left(1 + O \left(\frac{1}{q} \right) \right) \quad \text{при } 2 \leq q.$$

При больших q правая часть мало зависит от q ; это естественно, так как для функции $\varphi(t)$, удовлетворяющей (9.6), выполняется (9.7).

Оценка погрешности интегрирования на отрезке $[0, a_1]$ вида

$$\left(\max_{[0, a_1]} F(x) \right) \frac{(a_1 - 0)^3}{12}$$

непригодна вследствие неограниченности $F(x)$. При $b < 0$ подынтегральная функция x^b обращается в ∞ при $x \rightarrow 0$. Определим ее равной 0 при $x = 0$.

Тогда при всех b разность между интегралом $\int_0^{a_1} x^b dx$ и квадратурой $s_1(f)$

есть

$$\frac{a_1^{b+1}}{b+1} - \frac{a_1 \cdot a_1^b}{2} = \left(\frac{1}{b+1} - \frac{1}{2} \right) a_1^{b+1} = \left(\frac{1}{b+1} - \frac{1}{2} \right) \frac{1}{N^3}.$$

Сумма этого выражения и оценок погрешности (5) по остальным отрезкам равна величине

$$r = \frac{d^3 b |b-1|}{12N^2} + O\left(\sum_{q=1}^N \frac{1}{qN^3}\right) = \frac{d^3 b |b-1|}{12N^2} + O\left(\frac{\ln N}{N^3}\right). \quad (6)$$

Мы получили нужное соотношение.

Задача. Пусть

$$f(x) = \begin{cases} 0 & \text{при } x = 0, \\ x^b & \text{при } x \in (0, 1], \end{cases}$$

$-1 < b < 1$, интеграл $\int_0^1 f(x) dx$ вычисляется по формуле трапеций с посто-

янным шагом $a_q - a_{q-1} = \frac{1}{N}$. Показать, что суммарная погрешность удовлетворяет соотношению

$$R_N(f) \sim \frac{D(b)}{N^{1+b}},$$

где $D(b) \neq 0$.

Из сравнения последней оценки с (6) следует, что в случае функций с особенностями в производных оптимизация распределения узлов интегрирования может приводить к увеличению порядка скорости сходимости.

Пример 2. Вычисляется серия интегралов

$$\int_0^1 x^b \ln x g(b, x) dx,$$

$g(b, x)$ — гладкая, $g(b, 0) \neq 0$, b — параметр серии. Поскольку $\ln x$ имеет особенность в точке 0, то кажется естественным взять в качестве модельной функции $x^b \ln x$. Ее вторая производная имеет вид

$$x^{b-2} (b(b-1) \ln x + (2b-1)).$$

При $F(x) = |(x^b \ln x)''|$ уравнение (9.6) не решается в квадратурах: поэтому постарайся его упростить. При $x \rightarrow 0$ функция $\ln x$ растет медленнее, чем любая степень $x^{-\varepsilon}$, $\varepsilon > 0$. Исходя из этого, в уравнении (9.6) для упрощения решения возьмем $F(x) = \text{const} \cdot x^{b-2}$.

Пример 3. Вычисляется серия интегралов

$$\int_0^1 \exp\left(-\frac{x}{b}\right) g(b, x) dx, \quad (7)$$

$g(b, x)$ — гладкая, $g(b, 0) \neq 0$, b — параметр серии — может принимать очень малые значения. При малых b подынтегральная функция и ее производные резко меняются в окрестности точки $x = 0$ за счет множителя $\exp\left(-\frac{x}{b}\right)$.

Поэтому имеет смысл произвести оптимизацию распределения узлов интегрирования на модельной задаче вычисления интеграла $\int_0^1 \exp\left(-\frac{x}{b}\right) dx$.

Положим $F(x) = \left| \left(\exp\left(-\frac{x}{b}\right) \right)'' \right|$. Уравнение (9.6) приобретает вид

$$\frac{1}{b^2} \exp\left(-\frac{\varphi}{b}\right) \left(\frac{d\varphi}{dt} \right)^3 = C_1.$$

Отсюда

$$1 - \exp\left(-\frac{\varphi}{3b}\right) = C_3 t + C_4.$$

Из условия $\varphi(0) = 0$ следует $C_4 = 0$, а из условия $\varphi(1) = 1$ получаем

$$1 - \exp\left(-\frac{\varphi}{3b}\right) = \left(1 - \exp\left(-\frac{1}{3b}\right) \right) t.$$

Отсюда

$$\varphi(t) = 3b \ln \frac{1}{1 - \left(1 - \exp\left(-\frac{1}{3b}\right) \right) t}.$$

Пример 4. Вычисляется та же серия, но $g(b, x) = x h(b, x)$, $h(b, 0) \neq 0$. Можно рассмотреть в качестве модельного интеграл

$\int_0^1 x \exp\left(-\frac{x}{b}\right) dx$. При

$$F(x) = \left| \left(x \exp\left(-\frac{x}{b}\right) \right)'' \right| = \left| \frac{x}{b^2} - \frac{2}{b} \right| \exp\left(-\frac{x}{b}\right)$$

соответствующее уравнение (9.6) не решается в элементарных квадратурах. В случае необходимости срочного решения задачи целесообразно ограничиться оптимальным распределением узлов, построенным для примера 3.

Пример 5. Вычисляется серия интегралов

$$\int_0^1 \exp\left(-\frac{x^2}{b^2}\right) g(b, x) dx.$$

$g(b, x)$ — гладкая, $g(b, 0) \neq 0$, b — параметр серии — может принимать малые значения. При малых b подынтегральная функция и ее производные резко меняются в окрестности точки $x = 0$ за счет множителя $\exp\left(-\frac{x^2}{b^2}\right)$.

Поэтому в качестве модельной возьмем задачу вычисления интеграла

$$\int_0^1 \exp\left(-\frac{x^2}{b^2}\right) dx. \text{ Положим } F(x) = \left| \left(\exp\left(-\frac{x^2}{b^2}\right) \right)'' \right|, \text{ тогда в качестве (9.6)}$$

получим уравнение

$$\left| 1 - 2 \frac{\varphi^2}{b^2} \right| \exp\left(-\frac{\varphi^2}{b^2}\right) \left(\frac{d\varphi}{dt} \right)^3 = \frac{C_1 b^2}{2}.$$

Отсюда

$$\int \sqrt[3]{\left| 1 - 2 \frac{\varphi^2}{b^2} \right|} \exp\left(-\frac{\varphi^2}{3b^2}\right) d\varphi = \int \frac{\sqrt[3]{C_1 b^2}}{\sqrt[3]{2}} dt.$$

Этот интеграл не вычисляется в явном виде, и поэтому попытаемся произвести упрощения; например, можно заменить $\sqrt[3]{\left| 1 - 2 \frac{\varphi^2}{b^2} \right|}$ на 1. При

больших значениях $\frac{\varphi}{b}$, когда погрешность такой замены большая, ее влияние не столь значительно вследствие наличия малого множителя $\exp\left(-\frac{\varphi^2}{3b^2}\right)$.

После такого упрощения функция $\varphi(t)$ будет выражаться через функцию, обратную к функции

$$\int_0^u \exp(-v^2) dv.$$

Заметим, что задачи, подобные рассмотренным в примерах 3, 4, 5, возникают довольно часто. Например, при расчетах диаграмм направленности антенн вычисляются серии интегралов

$$\int_0^1 \exp(ibg(b, x)) h(b, x) dx$$

в широком диапазоне изменения b ; функции $g(b, x)$, $h(b, x)$ являются довольно гладкими. При b не очень больших эти интегралы могут вычисляться при помощи известных нам квадратурных формул. С ростом b производные подынтегральной функции растут, и поэтому требуется все большее число узлов интегрирования. При очень больших b можно воспользоваться методом перевала или иными асимптотическими методами. Однако при «промежуточных» значениях b оба эти метода будут плохи: первый — вследствие трудоемкости, второй — вследствие малой точности. Поэтому иногда оказывается целесообразным следующий метод: контур интегрирования преобразуется

так, чтобы он проходил по линиям наискорейшего спуска $\exp(ibg(b, x))$, как это делается при использовании метода перевала. Мы получим интегралы от резко меняющихся функций, аналогичные рассмотренным в примерах 2, 3, 4. Далее производим оптимизацию распределения узлов интегрирования и применяем соответствующие квадратурные формулы.

На примерах 3, 4 видно, что оптимизация распределения узлов интегрирования на основании (9.6) требует достаточно высокой квалификации исследователя. Поэтому далее в § 17 будет рассмотрен вопрос о передаче этих функций ЭВМ.

§ 11. Главный член погрешности

В §§ 1, 2, 6, 8 был рассмотрен ряд квадратурных формул и получены оценки погрешности через максимумы производных или интегралы от их модулей. При выводе этих оценок не предполагалось наличия у подынтегральной функции производных высших порядков. Однако если такие производные имеются, то может произойти компенсация некоторых составляющих погрешности. Для обсуждения вопроса о качестве оценок и о реальной величине погрешности целесообразно выделить в погрешности главный член.

Обратимся к формуле (1.3)

$$I(f) = \int_a^b f(x) p(x) dx \approx S_n(f) = \frac{b-a}{2} \sum_{j=1}^n D_j f \left(\frac{b+a}{2} + \frac{b-a}{2} d_j \right) \quad (1)$$

с остаточным членом (1.1):

$$R_n(f) = I(f) - S_n(f) = \int_a^b p(x) (f(x) - L_n(x)) dx.$$

Для определенности рассматриваем случай $p(x) = 1$. После замены переменных $x = \frac{b+a}{2} + \frac{b-a}{2} t$ остаточный член запишется в виде

$$R_n(f) = \frac{b-a}{2} \int_{-1}^1 \left(f \left(\frac{b+a}{2} + \frac{b-a}{2} t \right) - Q_n \left(f \left(\frac{b+a}{2} + \frac{b-a}{2} t \right) \right) \right) dt. \quad (2)$$

Здесь для удобства введено обозначение:

$$Q_n(g(t)) = \sum_{j=1}^n g(d_j) \prod_{t \neq d_j} \frac{t - d_l}{d_j - d_l}.$$

Пусть m таково, что квадратурная формула (1) точна для всех многочленов степени $m-1$, но не точна для многочленов

степени m . Из способа построения интерполяционной квадратурной формулы (1.3) вытекает, что $m \geq n$. Подставим в правую часть (2) разложение Тейлора

$$f\left(\frac{b+a}{2} + \frac{b-a}{2}t\right) = \sum_{j=0}^{l-1} f^{(j)}\left(\frac{a+b}{2}\right) \frac{\left(\frac{b-a}{2}t\right)^j}{j!} + f^{(l)}(\xi(t)) \frac{\left(\frac{b-a}{2}t\right)^l}{l!} \quad (3)$$

при некотором $l > m$, здесь $a \leq \xi(t) \leq b$ для $-1 \leq t \leq 1$. Группируя слагаемые с одинаковыми множителями $f^{(j)}$, получим

$$R_n(f) = \sum_{j=0}^{l-1} f^{(j)}\left(\frac{b+a}{2}\right) \left(\frac{b-a}{2}\right)^{j+1} \gamma_j + \rho_l,$$

где

$$\gamma_j = \int_{-1}^1 \frac{(t^j - Q_n(t^j))}{j!} dt,$$

$$\rho_l = \left(\frac{b-a}{2}\right)^{l+1} \int_{-1}^1 \frac{f^{(l)}(\xi(t)) t^l - Q_n(f^{(l)}(\xi(t)) t^l)}{l!} dt.$$

С учетом явного выражения интерполяционного оператора Q_n получаем грубую оценку последнего слагаемого:

$$|\rho_l| \leq \gamma_l^0 \max_{[a, b]} |f^{(l)}(x)| \left(\frac{b-a}{2}\right)^{l+1},$$

где

$$\gamma_l^0 = \int_{-1}^1 \left(|t|^l + \sum_{j=1}^n |d_j|^l \prod_{i \neq j} \left| \frac{t-d_i}{d_j-d_i} \right| \right) / l! dt.$$

Коэффициент при $f^{(l)}\left(\frac{b+a}{2}\right)$ равен погрешности при вычислении интеграла от функции $\frac{\left(x - \frac{b+a}{2}\right)^j}{j!}$ по рассматриваемой квадратурной формуле. Поэтому $\gamma_0 = \dots = \gamma_{m-1} = 0$. В то же время $\gamma_m \neq 0$, так как иначе квадратура была бы точна для любого многочлена степени m . В итоге этих рассуждений мы имеем право написать равенство

$$R_n(f) = \sum_{j=m}^{l-1} f^{(j)}\left(\frac{b+a}{2}\right) \left(\frac{b-a}{2}\right)^{j+1} \gamma_j + \max_{[a, b]} |f^{(l)}(x)| \left(\frac{b-a}{2}\right)^{l+1} \theta_l \gamma_l^0, \quad (4)$$

где $|\theta_l| \leq 1$, $\gamma_m \neq 0$, $m \leq l-1$.

Заметим, что в § 1 для формул прямоугольников, трапеций и Симпсона были получены соотношения вида

$$R_n(f) = \text{const} \cdot f^{(m)}(\xi) \left(\frac{b-a}{2} \right)^{m+1}; \quad (5)$$

для формулы Гаусса при $p(x) \equiv 1$ такое соотношение следует из формулы (3.9). Из соотношения (5) вытекает справедливость соотношения, аналогичного (4) при $l = m + 1$.

Пусть для вычисления интеграла

$$I(f) = \int_A^B f(x) dx$$

применяется квадратурная формула (6.4)

$$I(f) \approx S_M^n(f) = \sum_{q=1}^M s_q(f),$$

где

$$s_q = \left(\frac{a_q - a_{q-1}}{2} \right) \sum_{j=1}^n D_j f \left(\frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j \right),$$

причем $a_q = \varphi\left(\frac{q}{M}\right)$, где $\varphi(t)$ — гладкая, $\varphi(0) = A$, $\varphi(1) = B$. Тогда $a_q - a_{q-1} = O\left(\frac{1}{M}\right)$. Воспользовавшись (4), можно написать равенство

$$\int_{a_{q-1}}^{a_q} f(x) dx - s_q = \gamma_m f^{(m)}\left(\frac{a_{q-1} + a_q}{2}\right) \left(\frac{a_q - a_{q-1}}{2}\right)^{m+1} + O\left(\frac{1}{M^{m+2}}\right).$$

Следовательно, сумма погрешностей по всем отрезкам разбиения записывается в виде

$$R_M^n(f) = \gamma_m \sum_{q=1}^M f^{(m)}\left(\frac{a_{q-1} + a_q}{2}\right) \left(\frac{a_q - a_{q-1}}{2}\right)^{m+1} + O\left(\frac{1}{M^{m+1}}\right).$$

Далее проводим преобразования такие же, как в § 9:

$$a_q - a_{q-1} = \frac{1}{M} \varphi'\left(\frac{q}{M}\right) + O\left(\frac{1}{M^2}\right),$$

$$f^{(m)}\left(\frac{a_{q-1} + a_q}{2}\right) = f^{(m)}\left(\varphi\left(\frac{q}{M}\right)\right) + O\left(\frac{1}{M}\right).$$

В результате имеем

$$R_M^n(f) = \frac{(\gamma_m 2^{-(m+1)})}{M^m} \left\{ \sum_{q=1}^M \frac{1}{M} \Phi\left(\frac{q}{M}\right) \right\} + O\left(\frac{1}{M^{m+1}}\right),$$

где $\Phi(t) = f^{(m)}(\varphi(t))(\varphi'(t))^{m+1}$. Выражение в фигурных скобках является квадратурной суммой интеграла

$$\int_0^1 \Phi(t) dt,$$

к которому оно стремится. Если $f^{(m+1)}(x)$ и $\varphi''(t)$ ограничены, то можно показать, что

$$R_M^n(f) = \frac{\gamma_m 2^{-(m+1)}}{M^m} \int_0^1 \Phi(t) dt + O\left(\frac{1}{M^{m+1}}\right). \quad (6)$$

В частности, если точки $d_1, \dots, d_n$ расположены симметрично относительно 0 и производные $f^{(m+2)}(x)$ и $\varphi^{(3)}(t)$ ограничены, то второе слагаемое в правой части (6) есть $O\left(\frac{1}{M^{m+2}}\right)$.

Докажем последнее утверждение для случая формулы трапеций при постоянном шаге интегрирования, т. е. при $\varphi'_t = \text{const}$, $a_q - a_{q-1} \equiv H$. Возьмем $l = 4$;

$$\gamma_2 = \int_{-1}^1 \frac{(t^2 - 1)}{2!} dt = -\frac{2}{3},$$

$$\gamma_3 = \int_{-1}^1 \frac{(t^3 - t)}{3!} dt = 0,$$

$$\gamma_4^0 = \int_{-1}^1 \frac{\left(t^4 + \left|\frac{t-1}{2}\right| + \left|\frac{t+1}{2}\right|\right)}{4!} dt = \frac{1}{10}.$$

Итак,

$$\begin{aligned} \int_{a_{q-1}}^{a_q} f(x) dx - \frac{H}{2} (f(a_{q-1}) + f(a_q)) = \\ = -\frac{H^3}{12} f''\left(\frac{a_{q-1} + a_q}{2}\right) + \theta_q \frac{H^5}{320} \max_{[a_{q-1}, a_q]} |f^{(4)}(x)|. \end{aligned} \quad (7)$$

Выражение $R_M^n(f)$ суммарной погрешности, получающееся после суммирования (7), можно записать в виде

$$R_M(f) = -\frac{H^2}{12} E_2(f) + O((B-A)H^4),$$

где

$$E_2(f) = \sum_{q=1}^M H f'' \left(\frac{a_{q-1} + a_q}{2} \right).$$

Величина $E_2(f)$ является квадратурной суммой для интеграла $\int_A^B f''(x) dx$, если при вычислении интегралов по элементарным отрезкам применяется формула прямоугольников (1.6):

$$\int_a^b g(x) dx \approx (b-a) g\left(\frac{b+a}{2}\right)$$

с оценкой остаточного члена

$$(\max_{[a, b]} |g''(x)|) \frac{(b-a)^3}{24}.$$

Следовательно, согласно (6.6), имеем

$$\left| E_2(f) - \int_A^B f''(x) dx \right| \leq (\max_{[A, B]} |f^{(4)}(x)|) \frac{(B-A)H^2}{24}.$$

В итоге получаем равенство

$$R_M(f) = -\frac{H^2}{12} \left(\int_A^B f''(x) dx \right) + O((B-A)H^4). \quad (8)$$

§ 12. Формулы Эйлера и Грегори

Полученное соотношение (11.8) может использоваться в различных направлениях. Например, его можно представить в виде

$$R_M(f) = -\frac{H^2}{12} (f'(B) - f'(A)) + O((B-A)H^4). \quad (1)$$

После вычисления $f'(B) - f'(A)$ получаем значение главного члена погрешности. Предположим, что достигнутая точность не является удовлетворительной. Запишем (1) в виде

$$I(f) = S_M^4(f) + O((B-A)H^4),$$

где

$$S_M^4(f) = S_M^2(f) - \frac{H^2}{12} (f'(B) - f'(A)).$$

Выражение $S_M^4(f)$ является квадратурной формулой с погрешностью $O((B-A)H^4)$, т. е. дает лучшее приближение к значению $I(f)$, чем исходная формула трапеций.

Можно попытаться выделить главный член погрешности полученной формулы. Имеем равенства

$$S_M^4(f) = \sum_{q=1}^M s_q^2(f),$$

$$s_q^2(f) = \frac{H}{2} (f(a_{q-1}) + f(a_q)) - \frac{H^2}{12} (f'(a_q) - f'(a_{q-1})).$$

Величину $s_q^2(f)$ будем рассматривать как приближенное значение интеграла

$$I_q(f) = \int_{a_{q-1}}^{a_q} f(x) dx.$$

Подставим в разность $I_q(f) - s_q^2(f)$ представление $f(x)$ в виде отрезка ряда Тейлора (11.3) при $l=6$. Как и в § 11, соберем коэффициенты при значениях $f^{(l)}\left(\frac{a_{q-1} + a_q}{2}\right)$:

$$I_q(f) - s_q^2(f) = \frac{H^5}{720} f^{(4)}\left(\frac{a_{q-1} + a_q}{2}\right) + O(H^7).$$

После суммирования по q получим

$$I(f) - S_M^4(f) = \frac{H^4}{720} \left\{ \sum_{q=1}^N H f^{(4)}\left(\frac{a_{q-1} + a_q}{2}\right) \right\} + O((B-A)H^6).$$

Выражение в фигурных скобках является квадратурной формулой прямоугольников для интеграла

$$\int_A^B f^{(4)}(x) dx = f^{(3)}(B) - f^{(3)}(A).$$

Таким образом, имеем

$$I(f) - S_M^4(f) = \frac{H^4}{720} (f^{(3)}(B) - f^{(3)}(A)) + O((B-A)H^6).$$

Мы получили выражение остаточного члена квадратуры $S_M^4(f)$, а также новую квадратурную формулу

$$I(f) \approx S_M^6(f) = S_M^2(f) - \frac{H^2}{12} (f'(B) - f'(A)) + \frac{H^4}{720} (f^{(3)}(B) - f^{(3)}(A))$$

с остаточным членом $O((B-A)H^6)$ и т. д. Продолжая этот процесс, для каждого целого p получим *квадратурную формулу Эйлера*

$$I(f) \approx S_M^{2p}(f),$$

где

$$S_M^{2p}(f) = S_M^2(f) - \sum_{k=1}^{p-1} \gamma_{2k} H^{2k} (f^{(2k-1)}(B) - f^{(2k-1)}(A)). \quad (2)$$

Существует следующее соотношение, которому удовлетворяют числа γ_j :

$$\frac{x}{\exp x - 1} = \sum_{j=0}^{\infty} \gamma_j x^j.$$

Обычно принято записывать γ_j в виде $B_j/j!$; числа B_j называют *числами Бернулли*.

Оценка остаточного члена квадратуры (2) может быть произведена следующим образом. Рассмотрим квадратуру

$$I_q(f) \approx s_q^p(f) = \frac{H}{2} (f(a_{q-1}) + f(a_q)) - \sum_{k=1}^{p-1} \gamma_{2k} H^{2k} (f^{(2k-1)}(a_q) - f^{(2k-1)}(a_{q-1})).$$

Подставим в остаточный член

$$r_q^p(f) = I_q(f) - s_q^p(f)$$

представление $f(x)$ в виде отрезка ряда Тейлора

$$\sum_{j=0}^{2p-1} f^{(j)}\left(\frac{a_{q-1} + a_q}{2}\right) \frac{1}{j!} \left(x - \frac{a_{q-1} + a_q}{2}\right)^j + \int_{(a_{q-1} + a_q)/2}^x f^{(2p)}(y) \frac{(x-y)^{2p-1}}{(2p-1)!} dy.$$

Как и в § 2, получим выражение остаточного члена вида

$$r_q^p(f) = \int_{a_{q-1}}^{a_q} K_p\left(x - \frac{a_{q-1} + a_q}{2}, H\right) f^{(2p)}(x) dx;$$

коэффициенты при младших производных $f^{(j)}\left(\frac{a_{q-1} + a_q}{2}\right)$ при $j < 2p$ обратились в нуль. Путем довольно громоздких вычислений можно установить, что $K_p\left(x - \frac{a_{q-1} + a_q}{2}, H\right)$ не меняет знака при $a_{q-1} \leq x \leq a_q$. Применяя теорему о среднем, имеем:

$$r_q^p(f) = f^{(2p)}(\xi_q^p) \int_{a_{q-1}}^{a_q} K_p\left(x - \frac{a_{q-1} + a_q}{2}, H\right) dx, \quad (3)$$

где $a_{q-1} \leq \xi_q^p \leq a_q$. Оказывается, что

$$\int_{a_{q-1}}^{a_q} K_p \left(x - \frac{a_{q-1} + a_q}{2}, H \right) dx = -\gamma_{2p} H^{2p+1}.$$

После суммирования (3) получим

$$I(f) - S_M^{2p}(f) = -\gamma_{2p} H^{2p} E_{2p};$$

величина

$$E_{2p} = H \sum_{q=1}^M f^{(2p)}(\xi_q^p)$$

заключена между

$$(B-A) \min_{[A, B]} f^{(2p)}(x) \quad \text{и} \quad (B-A) \max_{[A, B]} f^{(2p)}(x)$$

и, следовательно, равна $(B-A) f^{(2p)}(\xi_p)$, где $A \leq \xi_p \leq B$. Окончательное представление погрешности:

$$I(f) - S_M^{2p}(f) = -\gamma_{2p} f^{(2p)}(\xi_p) H^{2p} (B-A).$$

Для сведения приведем значения

$$\gamma_2 = \frac{1}{12}, \quad \gamma_4 = -\frac{1}{720}, \quad \gamma_6 = \frac{1}{30\,240}, \quad \gamma_8 = -\frac{1}{1\,209\,600},$$

$$\gamma_{10} = \frac{1}{47\,900\,160}.$$

При реальных вычислениях использование формул Эйлера несколько неудобно в связи с необходимостью вычисления не только значений функции, но и значений ее производных.

В выражении $S_M^{2p}(f)$ заменим производные $f^{(2k-1)}(A)$ и $f^{(2k-1)}(B)$ производными интерполяционных многочленов степени l с узлами $a_0, \dots, a_l$ и $a_N, \dots, a_{N-l}$, соответственно. При $l=2p-3$ и $l=2p-2$ после проведения промежуточных преобразований мы получим формулы численного интегрирования Грегори

$$I(f) \approx G_M^l(f),$$

$$G_M^l(f) = S_M^2(f) + H \sum_{k=1}^l \beta_k (\nabla^k f(a_M) - (-1)^k \Delta^k f(a_0)),$$

где

$$\beta_k = (-1)^{k+1} \int_0^1 \frac{x(x-1)\dots(x-k)}{(k+1)!} dx.$$

В частности,

$$\beta_1 = -\frac{1}{12}, \quad \beta_2 = -\frac{1}{24}, \quad \beta_3 = -\frac{19}{720}, \quad \beta_4 = -\frac{3}{160},$$

$$\beta_5 = -\frac{863}{60\,480}, \quad \beta_6 = -\frac{275}{24\,192}.$$

Главный член погрешности квадратуры Грегори $I(f) \approx G_M^I(f)$ есть

$$\beta_{l+1} H^{l+2} (f^{(l+1)}(B) - (-1)^{l+1} f^{(l+1)}(A)).$$

Задача В предположении, что сказанное ранее верно, доказать справедливость последнего соотношения

Поскольку реальные подынтегральные функции имеют различный характер поведения производных в различных частях отрезка интегрирования, то в современных стандартных программах численного интегрирования обычно применяются различные способы интегрирования с автоматическим выбором шага (см. §§ 17, 18). Поэтому область применения формул Эйлера — Грегори сузилась по сравнению с периодом, предшествующим применению ЭВМ. Однако они находят себе приложение при интегрировании функций, заданных таблично, вычислении неопределенных интегралов, при решении интегральных уравнений Вольтерра и в других задачах, где существенно, чтобы значения подынтегральной функции вычислялись именно на равномерной сетке.

§ 13. Правило Рунге практической оценки погрешности

Мы получили соотношение (11.6) и ряд его следствий, на основании которых можно судить о величине главного члена погрешности. Однако непосредственное вычисление интеграла

$$K_m = \frac{\gamma_m}{2^{m+1}} \int_0^1 f^{(m)}(\varphi(t)) (\varphi'(t))^{m+1} dt, \quad (1)$$

характеризующего главный член погрешности, довольно обременительно.

В практической работе обычно идут по несколько иному пути. Зададимся некоторыми M_1 и M_2 , $M_1 < M_2$ и произведем дважды численное интегрирование с разбиением отрезка интегрирования на части $\left[\varphi\left(\frac{q-1}{M_1}\right), \varphi\left(\frac{q}{M_1}\right)\right]$ и $\left[\varphi\left(\frac{q-1}{M_2}\right), \varphi\left(\frac{q}{M_2}\right)\right]$, соответственно. В результате получим приближенные значения интеграла $S_{M_1}(f)$ и $S_{M_2}(f)$. Согласно (11.6), имеем

$$\begin{aligned} I(f) - S_{M_1}(f) &= K_m M_1^{-m} + \rho_{M_1}, \\ I(f) - S_{M_2}(f) &= K_m M_2^{-m} + \rho_{M_2}, \end{aligned} \quad (2)$$

где $\rho_M = o(M^{-m})$. Предположим $K_m \neq 0$. Если бы в этих соотношениях отсутствовали величины ρ_{M_1} и ρ_{M_2} , то получилось бы два уравнения с двумя неизвестными $I(f)$ и K_m , которые можно

было бы определить; в противном случае можно получить лишь некоторые приближения к ним. Исключая $I(f)$, получаем соотношение для определения K_m :

$$S_{M_2}(f) - S_{M_1}(f) = K_m (M_1^{-m} - M_2^{-m}) + \rho_{M_1} - \rho_{M_2}.$$

Отсюда

$$K_m = \frac{S_{M_2}(f) - S_{M_1}(f) + \rho_{M_2} - \rho_{M_1}}{M_1^{-m} - M_2^{-m}};$$

таким образом, погрешность квадратуры $S_{M_2}(f)$ может быть представлена в виде

$$\begin{aligned} I(f) - S_{M_2}(f) &= K_m M_2^{-m} + \rho_{M_2} = \\ &= \frac{S_{M_2}(f) - S_{M_1}(f)}{(M_2/M_1)^m - 1} + \frac{\rho_{M_2} - \rho_{M_1}}{(M_2/M_1)^m - 1} + \rho_{M_2}. \end{aligned}$$

Первое слагаемое

$$z_1 = \frac{S_{M_2}(f) - S_{M_1}(f)}{(M_2/M_1)^m - 1}$$

можно вычислить. Если оно является главным членом погрешности, то можно считать, что $I(f) - S_{M_2}(f) \approx z_1$, и использовать величину z_1 как оценку погрешности. Оценка погрешности через величину z_1 носит название *правила Рунге*. Чтобы z_1 было главным членом погрешности $I(f) - S_{M_2}(f)$, имеющей порядок M_2^{-m} , необходимо выполнение условия

$$z_2 = \frac{\rho_{M_2} - \rho_{M_1}}{(M_2/M_1)^m - 1} + \rho_{M_2} = o(M_2^{-m}). \quad (3)$$

Если, например, $M_2 \sim M_1 \lambda$, где $\lambda > 1$, то

$$z_2 = \frac{o(M_2^{-m})}{\lambda^m - 1} + o(M_2^{-m}) = o(M_2^{-m}),$$

и требуемое условие выполняется.

Если имеется более сильная оценка для величин ρ_M , то соотношение $I(f) - S_{M_2}(f) \approx z_1$ будет выполняться при более слабых предположениях о соотношении между M_1 и M_2 . При вычислении однократных интегралов типичен случай, когда

$$\rho_M = O(M^{-m-1}). \quad (4)$$

Тогда

$$z_2 = \frac{O(M_2^{-m-1}) + O(M_1^{-m-1})}{(M_2/M_1)^m - 1} + o(M_2^{-m-1}).$$

Предположим сначала $M_2/M_1 \rightarrow 1$; согласно соотношению $(1 + \varepsilon)^m - 1 \sim m\varepsilon$ при $\varepsilon \rightarrow 0$ имеем

$$\left(\frac{M_2}{M_1}\right)^m - 1 \sim m\left(\frac{M_2}{M_1} - 1\right).$$

Тогда

$$z_2 = O\left(\frac{1}{M_2^{m+1}\left(\frac{M_2}{M_1} - 1\right)}\right) + O\left(\frac{1}{M_2^{m+1}}\right) = O\left(\frac{1}{(M_2 - M_1)M_2^m}\right). \quad (5)$$

Таким образом, при условии $M_2/M_1 \rightarrow 1$ для выполнения (3) достаточно, чтобы $M_2 - M_1 \rightarrow \infty$ при $M_2 \rightarrow \infty$.

Пусть $M_2/M_1 \rightarrow \infty$, тогда

$$z_2 = O\left(\frac{1}{M_1^{m+1}\left(\left(\frac{M_2}{M_1}\right)^m - 1\right)}\right) + O\left(\frac{1}{M_2^{m+1}}\right) = O\left(\frac{1}{M_1 M_2^m}\right), \quad (6)$$

т. е. для выполнения (3) достаточно, чтобы $M_1 \rightarrow \infty$. Объединяя оценки (5) и (6), получаем

$$z_2 = O\left(\max\left(\frac{1}{M_1}, \frac{1}{M_2 - M_1}\right) \frac{1}{M_2^m}\right).$$

Таким образом, в случае (4) для применимости правила Рунге достаточно выполнения условий

$$M_1, M_2 - M_1 \gg 1. \quad (7)$$

Задача. Показать, что при $\rho_M = O(M^{-m-2})$ для применимости правила Рунге достаточно выполнения условия $M_1 \gg 1$.

Может случиться, что величина K_m , определяемая равенством (1), равна нулю или очень мала. Тогда при реально используемых M слагаемое $K_m M^{-m}$ не будет существенно больше по модулю, чем ρ_M , и проводимые нами построения будут выглядеть как довольно сомнительные. В этом случае при оценке погрешности через величину z_1 зачастую будет создаваться превратное представление о величине погрешности.

Если z_1 является главным членом величины $K_m M_2^{-m}$ и $z_2 = O(M_2^{-m})$, мы имеем право написать равенство

$$I(f) - (S_{M_2}(f) + z_1) = o(M_2^{-m}), \quad (8)$$

т. е. величина $S_{M_2}(f) + z_1$ дает лучшее по порядку приближение к исходному значению интеграла.

Существуют различные процедуры вычисления интегралов, опирающиеся на использование правила Рунге. В одних случаях задаются некоторыми M_0 и $\lambda > 1$ и последовательно вычисляют $S_{M_k}(f)$ при $M_k = [M_0 \lambda^k]$, $k = 0, 1, \dots$. После каждого вычисления величины $S_{M_k}(f)$, $k > 0$, определяют соответствующую величину

$$z_1^k = \frac{S_{M_k}(f) - S_{M_{k-1}}(f)}{(M_k/M_{k-1})^m - 1}$$

и проверяют, будет ли выполняться условие $|z_1^k| < \varepsilon$, где ε — заданная точность. Если при некотором $k = l$ это условие выполняется, то величину $S_{M_l}(f)$ или $S_{M_l}(f) + z_1^l$ принимают за приближенное значение исходного интеграла.

Против этой методики часто выдвигается такое возражение: если значения $S_{M_k}(f)$ при различных k вычисляются независимо, то происходит большой перерасход работы вследствие вычисления большого числа «ненужных» значений $S_{M_k}(f)$.

На самом деле, этот перерасход не является столь катастрофическим, как это кажется на первый взгляд. Объем работы при вычислении каждого значения $S_{M_k}(f)$ пропорционален $M_k \sim \sim M_0 \lambda^k$, общий объем работы пропорционален

$$\sum_{k=0}^l M_0 \lambda^k = M_0 \frac{\lambda^{l+1} - 1}{\lambda - 1} \sim M_0 \frac{\lambda^{l+1}}{\lambda - 1}. \quad (9)$$

Если относить к «полезной» работе только работу по вычислению $S_{M_l}(f)$, то доля «полезной» работы будет равна отношению $M_0 \lambda^l$ к $M_0 \frac{\lambda^{l+1}}{\lambda - 1}$, т. е. $1 - \lambda^{-1}$. Если относить к «полезной» работе также работу по вычислению значения $S_{M_{l-1}}(f)$, необходимого для оценки погрешности, то доля полезной работы равна отношению $M_0 \lambda^l + M_0 \lambda^{l-1}$ к $M_0 \frac{\lambda^{l+1}}{\lambda - 1}$, т. е. $1 - \lambda^{-2}$.

Чтобы нейтрализовать высказывания возможных оппонентов о большом объеме ненужной работы, можно привести следующий довод: возьмем $\lambda = 10$; тогда доля полезной работы составит 99%. Такой прием обосновывать эффективность метода, подсчитывая высокий процент «полезной» работы, обычно выглядит убедительным, хотя и не является правомерным. Качество метода должно оцениваться по общим затратам, а не долей полезной работы.

По-видимому, описанная выше процедура вычисления интеграла не является наилучшей; однако будет поучительно рассмотреть вопрос о правильном выборе параметра λ для этой процедуры.

Чтобы упростить анализ вопроса, сделав его возможным, будем считать что погрешность $R_M(f)$ удовлетворяет условию $R_M(f) = K_m M^{-m}$; тогда z_1^k в точности равно $K_m M_k^{-m}$. Пусть k — наименьшее целое, для которого $|K_m| (M_0 \lambda^k)^{-m} < \varepsilon$. Последнее соотношение эквивалентно неравенству

$$\frac{1}{m} \log_{\lambda} \left(\frac{|K_m|}{M_0^m \varepsilon} \right) < k.$$

Отсюда

$$k = \left\lceil \frac{1}{m} \log_{\lambda} \left(\frac{|K_m|}{M_0^m \varepsilon} \right) \right\rceil + 1.$$

Это соотношение удобнее записать в виде

$$k = 1 + \frac{1}{m} \log_{\lambda} \left(\frac{|K_m|}{M_0^m \varepsilon} \right) - \left\{ \log_{\lambda} \sqrt[m]{\frac{|K_m|}{M_0^m \varepsilon}} \right\},$$

где, как обычно, $\{y\} = y - [y]$. Общий объем вычислений, согласно (9), пропорционален

$$\begin{aligned} \frac{M_0 \lambda^k}{1 - \frac{1}{\lambda}} &= \frac{M_0 \lambda^{\frac{1}{m} \log_{\lambda} \left(\frac{|K_m|}{M_0^m \varepsilon} \right) + 1 - \left\{ \log_{\lambda} \sqrt[m]{\frac{|K_m|}{M_0^m \varepsilon}} \right\}}}{1 - \frac{1}{\lambda}} = \\ &= \frac{\left(\frac{|K_m|}{\varepsilon} \right)^{\frac{1}{m}} \lambda^{1 - \left\{ \log_{\lambda} \sqrt[m]{\frac{|K_m|}{M_0^m \varepsilon}} \right\}}}{1 - \frac{1}{\lambda}}. \end{aligned}$$

От величины λ зависит второй множитель

$$g(\lambda) = \frac{\lambda^{1 - \left\{ \log_{\lambda} \sqrt[m]{\frac{|K_m|}{M_0^m \varepsilon}} \right\}}}{1 - \frac{1}{\lambda}};$$

поскольку ничего не известно о величине $|K_m|$, то нужно сделать уточнение постановки.

Первый возможный путь — потребовать минимальности по λ величины

$$\max_{|K_m|} g(\lambda) = g_1(\lambda) = \frac{\lambda}{1 - \frac{1}{\lambda}}.$$

Непосредственное исследование показывает, что величина $g_1(\lambda)$ принимает наименьшее значение 4 при $\lambda = 2$.

Другой путь — это исследование статистики поведения величины $g(\lambda)$.

Статистика отмечает следующий факт: если мы зададимся произвольным $\lambda > 1$ и из различных независимых источников получим некоторые числа y , то величины $\{\log_{\lambda} y\}$ будут асимптотически равномерно распределены на $[0, 1]$.

Величина

$$y = \left\{ \log_{\lambda} \sqrt[m]{\frac{|K_m|}{M_0^m \varepsilon}} \right\}$$

записывается в виде

$$\left\{ \log_{\lambda} \sqrt[m]{|K_m|} - \log_{\lambda} M_0 - \log_{\lambda} \sqrt[m]{\varepsilon} \right\}.$$

Задаваемые нами значения точности ε нельзя считать случайными, поскольку мы привыкли задавать их в определенном виде, например $\varepsilon = 2^{-s}$, $\varepsilon = 10^{-s}$ или $\varepsilon = 0,5 \cdot 10^{-1}$. Значение M_0 также нельзя считать случайным вследствие наличия подобных привычек. Однако величина $\sqrt[m]{|K_m|}$ может

рассматриваться как таковая, поскольку K_m характеризует главный член погрешности различных задач. Поэтому будет естественно предположить, что величина y равномерно распределена на $[0, 1]$. Тогда

$$g_2(\lambda) = M(g(\lambda)) = \int_0^1 \frac{\lambda^{1-y}}{1 - \frac{1}{\lambda}} dy = \frac{\lambda}{\ln \lambda}.$$

Непосредственно проверяем, что наименьшее значение $g_2(\lambda)$, равное e , достигается при $\lambda = e$.

Возникает вопрос, какому подходу отдать предпочтение. Если мы имеем дело с большим числом задач, то при прочих равных условиях будет более разумным второй, статистический подход. В данном случае имеем таблицу значений

λ	1,5	2	2,718	3	4	8	16
$g_1(\lambda)$	4,5	4	4,300	4,5	5,333	9,143	17,667
$g_2(\lambda)$	3,699	2,886	2,718	2,730	2,886	3,848	5,772

Отсюда

$$\frac{g_1(e)}{g_1(2)} = 1,075, \quad \frac{g_2(2)}{g_2(e)} = 1,062.$$

Мы видим, что при замене точки экстремума одной функции $g_1(\lambda)$ на точку экстремума другой значения функций изменяются несущественно. Наиболее предпочтительными значениями λ являются 2 или e , однако в данном случае не столь существенно, какое из этих двух значений следует выбрать. Хотя $\lambda = e$ статистически более предпочтительно, $\lambda = 2$ является традиционным и имеет некоторое преимущество в отношении накопления вычислительной погрешности. Традиционно при вычислении s -кратных интегралов при переходе к большему числу узлов это число увеличивают в 2^s раз, что соответствует $\lambda = 2^s$. Мы видим, что при $s = 1, 2$ этот выбор приемлем по отношению к статистической оценке, а при $s > 2$ становится менее удачным.

Другой путь использования правила Рунге заключается в следующем. Производим интегрирование с относительно небольшим числом узлов M_1 и M_2 и затем определяем величину K_m . В одномерном случае обычно берут $M_2/M_1 = 2$. Затем из уравнения $|K_m| M^{-m} = \varepsilon$ определяем требуемое число $M = M_3$ и осуществляем интегрирование с M_3 отрезками разбиения. Затем контролируем точность результата, применяя правило Рунге (если оно является законным) по результатам расчетов с M_2 и M_3 узлами. Если при данных M_2 и M_3 применение этого правила является сомнительным, то с целью получения представления о погрешности производим расчет еще при каком-либо числе отрезков разбиения M_4 . В случае если точность оказывается неудовлетворительной, по результатам S_{M_1} и S_{M_4} пересчитываем новое значение K_m и затем новое требуемое число отрезков разбиения $M = M_5$. При выборе оптимальной стратегии этого

алгоритма надо иметь в виду следующее: может быть, стоит завышать значение M по сравнению с решением уравнения $|K_m| M^{-m} = \varepsilon$, чтобы избежать лишних пересчетов.

Сформулируем правило выбора критерия практической оценки погрешности. Пусть u — решение задачи, u_n — получаемые приближения, $\lim_{n \rightarrow \infty} u_n = u$. За погрешность приближенного решения принимается некоторое выражение v_n , удовлетворяющее условию

$$\frac{|v_n - (u_n - u)|}{|u_n - u|} \rightarrow 0$$

при $n \rightarrow \infty$. Очевидно, $|v_n| \sim |u_n - u|$.

В рассматриваемом случае задачи численного интегрирования за меру погрешности, удовлетворяющую этому условию, мы можем принять величину z_1 , если только $z_2 = o(z_1)$. Однако соотношение $I(f) - S_M(f) \sim z_1$ является лишь асимптотическим, а применяется при конкретных M_1 и M_2 ; поэтому в особо ответственных случаях прибегают к способам проверки его реальной выполнимости. Например, в описанном выше случае, когда последовательно вычисляются величины $S_{M_i}(f)$ при $M_i = M_0 \lambda^i$, имеем

$$\begin{aligned} S_{M_i}(f) - S_{M_{i-1}}(f) &= \\ &= (I(f) - S_{M_{i-1}}(f)) - (I(f) - S_{M_i}(f)) = \\ &= K_m(\lambda^m - 1) M_i^{-m} + o(M_i^{-m}). \end{aligned}$$

Если число узлов настолько велико, что член $o(M_i^{-m})$ существенно меньше, чем $K_m(\lambda^m - 1) M_i^{-m}$, то

$$\omega_i = \frac{S_{M_{i+1}}(f) - S_{M_i}(f)}{S_{M_i}(f) - S_{M_{i-1}}(f)} \approx \frac{K_m(\lambda^m - 1) M_{i+1}^{-m}}{K_m(\lambda^m - 1) M_i^{-m}} = \lambda^{-m}.$$

По близости величины ω_i к λ^{-m} можно судить о возможности ограничиться при оценке погрешности величиной z_1 .

При решении одной задачи предъявлялось требование обеспечить высокую точность и дать гарантию, что эта точность действительно достигается. Вследствие некоторой специфики задачи применение стандартных программ было невозможно и применялась интерполяционная квадратурная формула (см. § 1), соответствующая $m = 6$, с дроблением отрезка интегрирования на части. Последовательно производились вычисления интеграла при числе отрезков дробления $M_k = 2^k$ и оценка погрешности по величине z_1^k , однако требуемая точность не достигалась даже при очень большом числе узлов. Была сделана попытка проверить применимость асимптотической оценки по поведению величин ω_i . Оказалось, что величины ω_i быстро приближались к значению $1/4$. Возникло предположение, что в программе вместо квадратуры, соответствующей порядку сходимости $m = 6$, по ошибке стоит

квадратура, соответствующая порядку сходимости $m = 2$; так оно и оказалось.

В другом подобном случае величины ω_i группировались вблизи значения $1/3$ вместо того, чтобы стремиться к значению 2^{-4} . Попытки отыскать ошибки в программе оказались безуспешными. Оказалось, что подынтегральная

функция интеграла $\int_0^1 f(x) dx$ имела вид $g(x) \sqrt{x + \delta}$, где $\delta > 0$ очень

малое, а $g(x)$ — гладкая. При разбиении отрезка интегрирования на равные части для такой функции погрешность действительно ведет себя как M^{-6} , но при очень больших M . При реальных M она ведет себя как $M^{-5/2}$ и, следовательно, величины ω должны быть близки к $2^{-5/2}$. В этом случае проверка применимости правила Рунге по признаку $\omega_i \rightarrow 2^{-m}$ была крайне полезной: оценка погрешности по правилу Рунге оказывалась примерно в 20 раз меньше действительного значения погрешности, и если бы не была предпринята такая проверка, результат вычислений имел бы недопустимо большую ошибку.

Приведем пример удачного применения правила Рунге. Вычислялась серия из большого числа однотипных интегралов, причем в каждом случае требовалась определенная гарантия малости погрешности. Было решено вычислять эти интегралы по первой из описанных выше процедур, с оценкой погрешности по правилу Рунге. Чтобы избежать недоразумений, подобных описанным выше, было выбрано несколько десятков типичных интегралов, проведены вычисления и выведены на печать таблицы значений ω_i . Величины ω_i довольно хорошо сходились к пределам, из чего был сделан вывод о применимости правила Рунге при решении задач данной серии.

§ 14. Формулы Ромберга

Если подынтегральная функция достаточно гладкая, то погрешность квадратурной формулы может быть представлена в виде

$$R_M(f) = \sum_{k=1}^l D_k(f) M^{-i_k} + r(M), \quad (1)$$

где

$$i_1 < \dots < i_l, \quad r(M) = o(M^{-i_l}).$$

Например, в предположении ограниченности $f^{(2m+2)}(x)$ погрешность формулы трапеций, согласно (12.2), представляется в виде

$$R_M(f) = - \sum_{k=1}^m \gamma_{2k} \left(\frac{B-A}{M} \right)^{2k} (f^{(2k-1)}(B) - f^{(2k-1)}(A)) + \\ + O \left((B-A) \left(\frac{B-A}{M} \right)^{2m+2} \right).$$

Обычно при гладкой подынтегральной функции

$$i_2 - i_1 = \dots = i_l - i_{l-1} = s,$$

где s или 1, или 2. Предположим, что произведено вычисление $S_M(f)$ при значениях $M = M_0, \dots, M_l$. Мы имеем равенства

$$I(f) = S_{M_j}(f) + \sum_{k=1}^l D_k(f) M_j^{-ik} + r(M_j), \quad j=0, \dots, l.$$

Образует линейную комбинацию этих соотношений с некоторыми коэффициентами c_j , потребовав, чтобы

$$\sum_{j=0}^l c_j = 1. \quad (2)$$

Получим соотношение

$$I(f) = \sum_{j=0}^l c_j S_{M_j}(f) + \sum_{k=1}^l D_k(f) \left(\sum_{j=0}^l c_j M_j^{-ik} \right) + \sum_{j=0}^l c_j r(M_j).$$

Предположим, что выполняются равенства

$$\sum_{j=0}^l c_j M_j^{-ik} = 0 \quad \text{при} \quad k=1, \dots, l; \quad (3)$$

тогда

$$I(f) = \sum_{j=0}^l c_j S_{M_j}(f) + \sum_{j=0}^l c_j r(M_j). \quad (4)$$

Если величина $\sum_{j=0}^l c_j r(M_j)$ пренебрежимо мала, то можно положить

$$I(f) \approx \sum_{j=0}^l c_j S_{M_j}(f). \quad (5)$$

Такая квадратура называется *формулой Ромберга*. Система соотношений (2), (3) образует систему из $l+1$ линейных уравнений с $l+1$ неизвестными, и поэтому есть основания ожидать, что она имеет решение.

Обратим внимание на связь между задачей решения этой системы и задачей интерполяции. Пусть, например, как для формулы трапеций,

$$I(f) = S_M(f) + \sum_{k=1}^l D_k M^{-2k} + r(M).$$

Перепишем это соотношение в виде

$$Q_l(M^{-2}) = S_M(f) + r(M), \quad \text{где} \quad Q_l(y) = I(f) - \sum_{k=1}^l D_k y^k. \quad (6)$$

Из рассмотрения (6) видно, что задача нахождения $I(f)$ может формулироваться следующим образом. Заданы приближенные

значения многочлена $Q_l(y)$ при $y = M_0^{-2}, \dots, M_l^{-2}$ требуется определить значение $Q_l(0) = I(f)$. Согласно интерполяционной формуле Лагранжа (§ 2 гл. II), имеем

$$Q_l(y) = \sum_{j=0}^l Q_l(y_j) \prod_{i \neq j} \frac{y - y_i}{y_j - y_i}.$$

Отсюда

$$Q_l(0) = \sum_{j=0}^l c_j Q_l(y_j), \quad \text{где} \quad c_j = \prod_{i \neq j} \frac{y_i}{y_i - y_j}, \quad y_j = M_j^{-2}, \quad (7)$$

и следовательно,

$$I(f) = Q_l(0) = \sum_{j=0}^l c_j Q_l(y_j) = \sum_{j=0}^l c_j S_{M_j}(f) + \sum_{j=0}^l c_j r(M_j).$$

Мы получили то же соотношение (4), причем одновременно получили в явном виде значения коэффициентов c_j .

Приближения (5), которые мы получили, представляют некоторые квадратурные формулы.

Пусть для определенности S_M — формула трапеций, $l = 1$, $M_1 = 2M_0$. Согласно (7), имеем

$$c_0 = \frac{1}{1 - 2^2} = -\frac{1}{3}, \quad c_1 = \frac{1}{1 - \left(\frac{1}{2}\right)^2} = \frac{4}{3}.$$

Тогда (5) имеет вид

$$I(f) \approx \frac{4}{3} S_{2M_0}(f) - \frac{1}{3} S_{M_0}(f) = H \left(\frac{1}{6} f(A) + \frac{4}{6} f\left(A + \frac{H}{2}\right) + \frac{2}{6} f(A + H) + \dots + \frac{1}{6} f(B) \right), \quad (8)$$

где $H = (B - A)/M_0$; таким образом, в правой части мы получили формулу Симпсона. Пусть $l = 2$, $M_1 = 2M_0$, $M_2 = 4M_0$. Тогда $c_0 = \frac{1}{45}$, $c_1 = -\frac{4}{9}$,

$$c_2 = \frac{64}{45},$$

$$I(f) = \frac{64}{45} S_{4M_0}(f) - \frac{20}{45} S_{2M_0}(f) + \frac{1}{45} S_{M_0}(f) = H \left(\frac{7}{90} f(A) + \frac{32}{90} f\left(A + \frac{1}{4}H\right) + \frac{12}{90} f\left(A + \frac{2}{4}H\right) + \right. \\ \left. + \frac{32}{90} f\left(A + \frac{3}{4}H\right) + \frac{14}{90} f(A + H) + \frac{32}{90} f\left(A + \frac{5}{4}H\right) + \dots + \frac{7}{90} f(B) \right). \quad (9)$$

Эта квадратура шестого порядка точности совпадает с квадратурой (6.4) при $a_q - a_{q-1} \equiv H$, если для вычисления интегралов по элементарным отрезкам интегрирования применяется формула (2.10). Точно так же квадратуру (9) можно получить, если прибавить к квадратуре Симпсона с $M_1 = 2M_0$ отрезками разбиения ее уточнение по правилу Рунге при $M_2 = 2M_1$.

Применение правила Ромберга дает некоторые преимущества по сравнению с непосредственным применением составных квадратур. Например, пусть мы задались какой-то квадратурой, вычислили и выдали на печать значения $S_{M_0 2^i}(f)$, $i = 0, \dots, l$, но нужной точности еще не достигли. Тогда можно попытаться получить приближение к интегралу, применив правило Ромберга по некоторой совокупности значений $S_{M_0 2^s}(f)$, $\dots$, $S_{M_0 2^p}(f)$, и для контроля точности применить это правило по совокупности значений

$$S_{M_0 2^{s+1}}(f), \dots, S_{M_0 2^{p+1}}(f).$$

Наиболее распространенная стандартная программа метода Ромберга устроена следующим образом: задаются некоторым M_0 и последовательно вычисляются приближенные значения интеграла по формуле трапеций $I(f) \approx S_{M_k}^{(1)}(f)$ при M_k отрезках разбиения: $M_k = M_0 \cdot 2^k$. Целесообразно вести вычисления по формуле

$$S_{M_k}^{(1)}(f) = \frac{1}{2} S_{M_{k-1}}^{(1)}(f) + \frac{B-A}{M_k} \sum_{j=1}^{M_{k-1}} f\left(A + \frac{2j-1}{M_k}(B-A)\right).$$

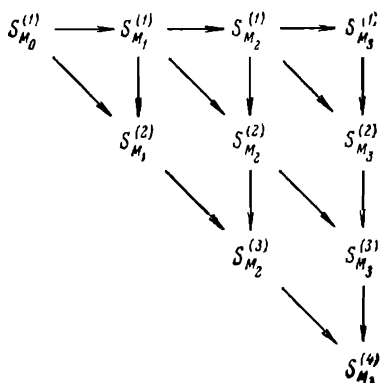
При каждом k после вычисления $S_{M_k}^{(1)}(f)$ последовательно вычисляют

$$S_{M_k}^{(2)}(f), \dots, S_{M_k}^{(k+1)}(f)$$

по рекуррентной формуле

$$S_{M_k}^{(l)}(f) = S_{M_k}^{(l-1)}(f) + \frac{1}{2^l - 1} (S_{M_k}^{(l-1)}(f) - S_{M_{k-1}}^{(l-1)}(f)),$$

Таким образом, последовательность вычислений определяется схемой



Вычисления значений $S_{M_k}^{(l)}(f)$ обычно продолжаются до тех пор, пока при некотором k не окажется

$$\min_l |S_{M_k}^{(l)}(f) - S_{M_{k-1}}^{(l)}(f)| < \varepsilon.$$

Задача. Показать, что $S_{M_k}^{(l)}(f)$ есть результат применения правила Ромберга к значениям

$$S_{M_k}^{(l)}(f), \dots, S_{M_{k-l+1}}^{(l)}(f),$$

в частности, величина $S_{M_k}^{(2)}(f)$ совпадает с результатом вычислений по формуле Симпсона, а $S_{M_k}^{(3)}(f)$ — по формуле (9).

Метод Ромберга широко распространен в западных странах, в частности, он является одним из наиболее рекомендуемых в системах стандартных программ фирмы IBM.

§ 15. Эксперименты и их обсуждение

Рассмотрим один пример применения результатов § 14. При решении задачи последовательно вычислялись некоторые приближения S_{M_k} с числом узлов $M_k = 2^k$. Погрешность решения оценивалась по близости последующих приближений $|S_{M_k} - S_{M_{k-1}}|$. Оказалось, эта величина стремится к нулю очень медленно и реальный порядок сходимости, по-видимому, $O(M^{-1})$. В то же время конкретные условия решения задачи требовали применения метода с порядком погрешности $O(M^{-2})$; в данном случае под конкретными условиями мы имеем в виду следующее: требуется решение большой серии задач при заданном ограничении на время решения; это определяет максимальное число операций, допустимое для вычисления отдельной задачи серии. Естественно было произвести уточнение по правилу Рунге, совпадающему с уточнением по Ромбергу при $l = 1$, предположив, что

$$R_M(f) \sim D_1 M^{-1} + O(M^{-2}).$$

Тогда соотношение (14.5) принимает вид

$$I(f) \approx S_{M_k}^{(1)} = 2S_{M_k} - S_{M_{k-1}}. \quad (1)$$

Погрешность решения оценивалась по разности последующих приближений $|S_{M_k}^{(1)} - S_{M_{k-1}}^{(1)}|$. Однако она также стремилась к 0, как M^{-1} , и желаемого ускорения сходимости не получилось.

Поэтому требовался более тщательный теоретический анализ погрешности. Он показал, что погрешность приближенного реше-

ния качественно ведет себя, как погрешность при вычислении интеграла $\int_0^1 f(x) dx$ по квадратуре

$$S_M(f) = \frac{1}{M} \sum_{q=1}^M f\left(\frac{q}{M}\right)$$

для $f(x) = g(x) \ln x$; $g(x)$ — гладкая, $g(0) \neq 0$. Иначе говоря, задача вычисления этого интеграла была выбрана как модельная задача.

Можно показать, что в этом случае

$$I(f) - S_M(f) = \frac{D_1 \log_2 M}{M} + \frac{D_2}{M} + O\left(\frac{\log_2 M}{M^2}\right).$$

Подставив $M_k = 2^k$, получим

$$\begin{aligned} S_{M_k}(f) - S_{M_{k-1}}(f) &= \frac{D_1 \log_2 M_{k-1}}{M_{k-1}} + \frac{D_2}{M_{k-1}} - \left(\frac{D_1 \log_2 M_k}{M_k} + \frac{D_2}{M_k} \right) + \\ &+ O\left(\frac{k}{2^{2k}}\right) = \frac{\left(D_1 \log_2 \frac{M_k}{4} + D_2\right)}{M_k} + O\left(\frac{\log_2 M_k}{M_k^2}\right). \end{aligned}$$

Это соотношение соответствует экспериментальному убыванию величины $S_{M_k} - S_{M_{k-1}}$, как $1/M_k$: множитель $\log_2 M_k$ изменяется много медленнее, чем $1/M_k$, и его наличие трудно обнаружить. Образует величины

$$S_{M_k}^{(1)}(f) = 2S_{M_k}(f) - S_{M_{k-1}}(f);$$

имеем

$$\begin{aligned} S_{M_k}^{(1)}(f) &= I(f) - \frac{2D_1 \log_2 M_k}{M_k} - \frac{2D_2}{M_k} + \frac{D_1 \log_2 M_{k-1}}{M_{k-1}} + \frac{D_2}{M_{k-1}} + \\ &+ O\left(\frac{\log_2 M_k}{M_k^2}\right) = I(f) - \frac{2D_1}{M_k} + O\left(\frac{\log_2 M_k}{M_k^2}\right). \end{aligned}$$

Таким образом,

$$I(f) - S_{M_k}^{(1)}(f) = \frac{2D_1}{M_k} + O\left(\frac{\log_2 M_k}{M_k^2}\right).$$

По отношению к приближениям $S_M^{(l)}(f)$ выполняется соотношение (14.1) при $l=1$. Мы имеем основания применить правило Рунге — Ромберга. Образует приближения

$$S_{M_k}^{(2)}(f) = 2S_{M_k}^{(1)}(f) - S_{M_{k-1}}^{(1)}(f) = 4S_{M_k}(f) - 4S_{M_{k-1}}(f) + S_{M_{k-2}}(f).$$

Для этих приближений имеем

$$I(f) - S_{M_k}^{(2)}(f) = O\left(\frac{\log_2 M_k}{M_k^2}\right).$$

Точно ту же квадратуру $S_{M_k}^{(2)}(f)$ можно было бы получить следующим образом: напишем равенства

$$I(f) = S_{M_i}(f) + \frac{D_1 \log_2 M_i}{M_i} + \frac{D_2}{M_i} + r(M_i), \quad i = k-2, k-1, k.$$

Подберем c_j и составим линейную комбинацию этих равенств, чтобы получить соотношение

$$I(f) = \sum_{j=0}^2 c_j S_{M_{k-j}}(f) + \sum_{j=0}^2 c_j r(M_{k-j}).$$

Теоретический анализ показывает, что в случае вычисления рассматриваемого интеграла выполняется соотношение

$$R_M(f) = \frac{D_1 \log_2 M}{M} + \frac{D_2}{M} + \frac{D_3 \log_2 M}{M^2} + \frac{D_4}{M^2} + O\left(\frac{\log_2 M}{M^3}\right). \quad (2)$$

Возникает мысль использовать это соотношение для дальнейшего повышения порядка точности квадратур. Однако прежде, чем пытаться повышать порядок точности, используя такое асимптотическое соотношение, полезно проверить, выполняется ли оно с приемлемой точностью при реально допустимых значениях M .

Для подхода к этому вопросу есть два пути.

Можно постараться подобрать модельную задачу с известным ответом, близкую к рассматриваемой. Тогда после проведения расчета мы будем иметь в распоряжении приближенное значение S_M и ошибку

$$R_M = I - S_M.$$

Может случиться, что имеются какие-то предположения о характере поведения этой величины, например, что

$$R_M \sim \text{const} \cdot M^{-m}. \quad (3)$$

В таком случае можно подсчитать для некоторой последовательности M_k значения $M_k^m R_{M_k}$ и посмотреть, стабилизируются ли эти величины с ростом M . Если не имеется предположения о характере поведения погрешности в данной задаче, то можно применить следующую методику.

Возьмем координатную плоскость $\ln M, \ln \left(\frac{1}{|R|} \right)$ (рис. 3.15.1); нанесем на нее точки $\left(\ln M_k, \ln \frac{1}{|R_{M_k}|} \right)$. Если эти точки расположены хаотически, то, значит, числа M_k не настолько велики, чтобы в погрешности выделился главный член. Предположим, что асимптотическое равенство (3) в данной области изменения параметра M выполняется с большой точностью. Из (3) следует, что

$$\ln \left| \frac{1}{R_M} \right| \sim m \ln M;$$

после дифференцирования имеем:

$$\frac{d \ln \left| \frac{1}{R_M} \right|}{d \ln M} \sim m. \quad (4)$$

Заметим, что операция дифференцирования асимптотических равенств, вообще говоря, незаконна.

Согласно (4), в случае, когда (3) выполняется достаточно точно, точки $\left(\ln M_k, \ln \frac{1}{|R_{M_k}|} \right)$, получае-

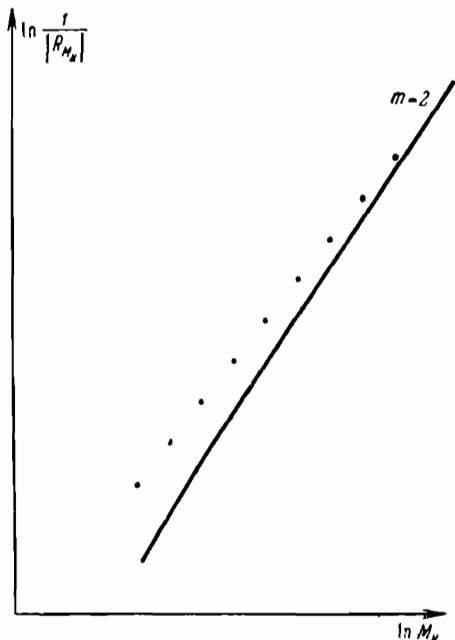


Рис. 3.15.1.

мые в результате эксперимента на ЭВМ, должны лежать на плавной кривой, наклон которой стремится к m . Близость наклона этой кривой к постоянному характеризует точность выполнения асимптотического равенства $R_M \sim \text{const} \cdot M^{-m}$. Если наклон кривой резко меняется, то, значит, еще нет оснований применять правило Рунге. В частности, случаи

$$R_M \sim \text{const} \cdot \frac{\ln M}{M} \text{ и } R_M \sim \text{const} \cdot \frac{1}{M}$$

практически неразличимы при таком рассмотрении, потому что

$$\frac{d \ln \left| \frac{1}{R_M} \right|}{d \ln M} \rightarrow 1 \text{ при } M \rightarrow \infty.$$

Проверку справедливости предположения о характере поведения погрешности можно осуществлять и таким путем. Если

справедливо равенство

$$R_M \sim \frac{c}{M^m}, \quad (5)$$

то

$$S_{M\lambda} - S_M \sim \frac{c(1 - \lambda^{-m})}{M^m}. \quad (6)$$

С другой стороны, если (6) выполняется при $M \geq M_0$ и $S_M \rightarrow I$, то будет выполняться и (5). Поэтому вместо проверки практической выполнимости (5) можно производить проверку практической выполнимости (6), в частности, при помощи изучения графиков функций $g(M) = (S_{M\lambda} - S_M)M^m$ или расположения точек

$$\left(\ln M_k, \ln \frac{1}{|S_{M_k\lambda} - S_{M_k}|} \right). \quad (7)$$

В рассматриваемой задаче точки (7) при $\lambda = 2$ лежали на плавной кривой с наклоном примерно 1. Отсюда был сделан вывод, что рассматриваемый метод имеет первый порядок сходимости. Для уточненных значений $S_M^{(1)}$ расположение точек

$$\left(\ln M, \ln \frac{1}{|S_{M\lambda}^{(1)} - S_M^{(1)}|} \right) \quad (8)$$

имело тот же характер.

Такое непонятное поведение погрешности побудило к рассмотрению модельной задачи и построению нового метода $I(f) \approx S_M^{(2)}(f)$. Так как точки (7) и (8) лежали на плавных кривых, то были основания надеяться, что выделение и устранение главного члена погрешности указанным путем приведет к увеличению точности. Чтобы надеяться на увеличение точности при выделении и устранении членов $\frac{D_3 \log_2 M}{M^2}$ и $\frac{D_4}{M^2}$ в главном члене погрешности формулы $I(f) \approx S_M^{(2)}(f)$, нужно быть уверенным, что они действительно являются главными при употребляемых значениях M . Для этой цели было построено множество точек

$$\left(\ln M, \ln \frac{1}{|S_{M\lambda}^{(2)} - S_M^{(2)}|} \right).$$

Для практически достижимых в условиях этой задачи значений M в расположении этих точек не наблюдалось никакой «плавности» и закономерности. Это означало, что надежды на дальнейшее повышение точности для этого метода являются необоснованными.

На рис. 3.15.2—3.15.5 изображены графики зависимости погрешности от числа узлов при вычислении различными методами интегралов от функций с типичными особенностями. N — общее

число узлов интегрирования, R_N — погрешность. Численное интегрирование производилось с помощью стандартных программ интегрирования с автоматическим выбором шага (см. § 18). На графиках N — общее число узлов, R_N — погрешность интегрирования; графики, отмеченные буквами L_k , соответствуют расчетам, где для вычисления интегралов по элементарным отрезкам

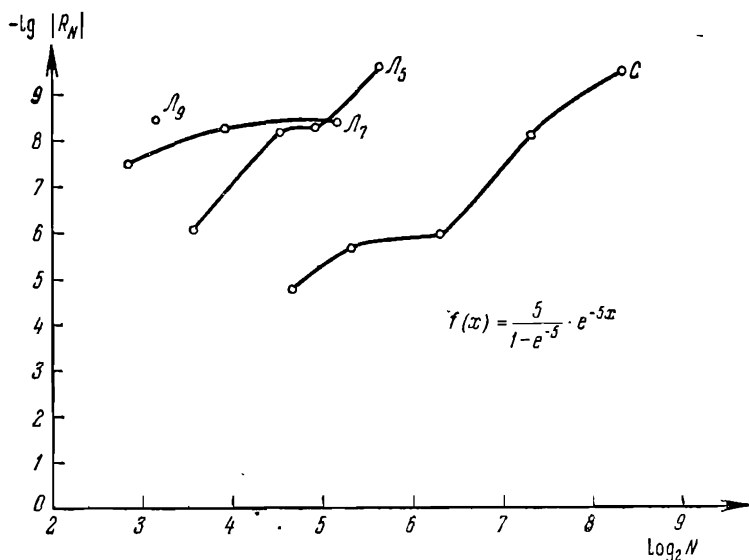


Рис. 3.15.2.

применялись формулы Лобатто с k узлами; график, соответствующий формуле Лобатто с тремя узлами, совпадающей с формулой Симпсона, помечен C ; отрезок интегрирования $[0, 1]$.

Из рассмотрения рис. 3.15.4 особенно четко видно, что формулы более высокого порядка точности могут оказываться менее эффективными по сравнению с формулами низкого порядка точности. Из рассмотрения этих рисунков также видно, что при малых N графики погрешности могут иметь довольно сложный вид и не являться графиками гладких функций; в последнем случае применение правила Рунге для практической оценки погрешности является неоправданным.

Полезно указать на следующую практически важную деталь. Если решение задачи содержит какие-то неисследованные особенности, ухудшающие сходимость методов, то лучше сразу выделить простейшую модельную задачу, содержащую эти особенности, и провести выбор метода и проверку применимости

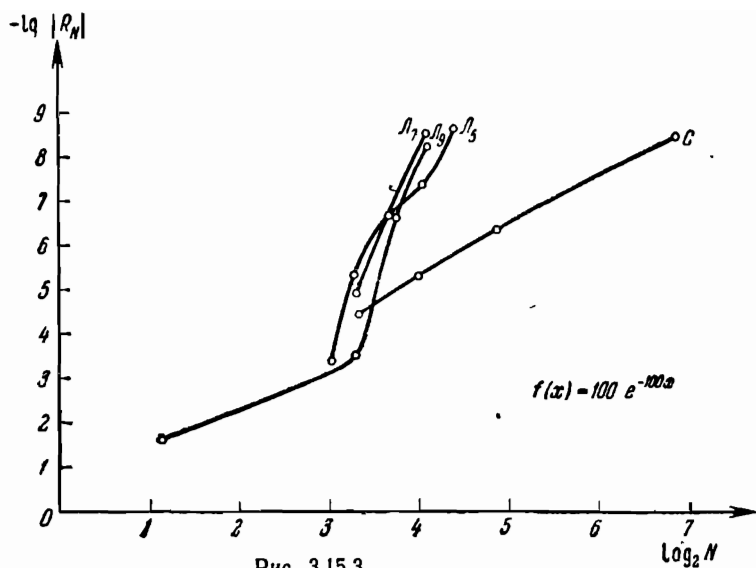


Рис. 3.15.3.

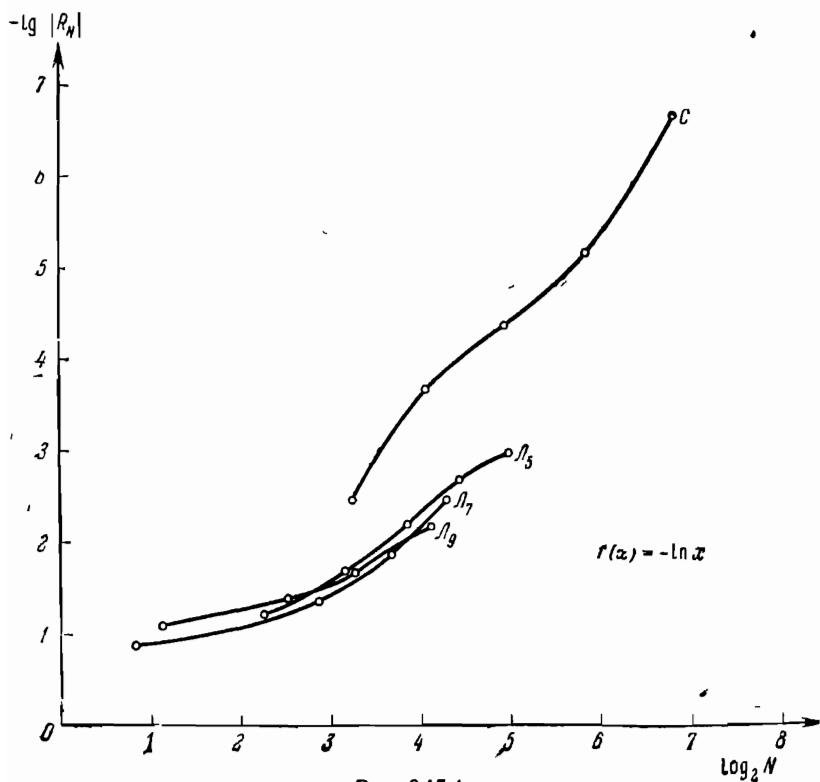


Рис. 3.15.4.

различных асимптотических критериев на этой задаче. Этот путь обычно приводит к более быстрому пониманию существа вопроса и избавляет от необходимости проведения многочисленных экспериментов на самой исходной задаче. В частности, достигается

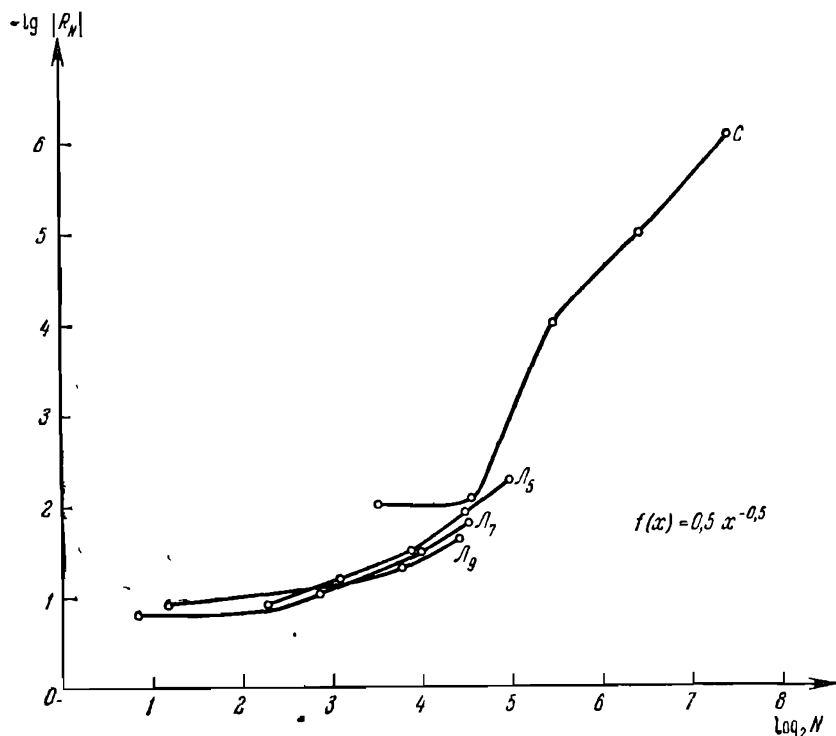


Рис. 3.15.5.

экономия при программировании задачи, в машинном времени и упрощается исследование за счет возможности построения более представительных графиков поведения погрешности — мы имеем возможность получить больше точек (7), (8), поскольку для простой задачи их получение менее трудоемко.

§ 16. Вычисление интегралов в нерегулярном случае

Существенную часть реально встречающихся подынтегральных функций составляют функции с особенностями у них или у их производных, или функции, производные которых очень велики. Если такой нерегулярный характер поведения подынтегральной функции не вызывается колебательным характером

поведения, неплохой результат дают стандартные программы интегрирования с автоматическим выбором шага, которые будут обсуждаться в § 17. В случае малой серии интегралов с особенностями обращения к этим стандартным программам может оказаться способом решения задачи с наименьшими затратами. Вычисление большой серии интегралов с особенностями может потребовать привлечения внимания исследователей более высокой квалификации. Укажем ряд приемов, которые могут оказаться полезными при рассмотрении этого вопроса.

1. *Выделение весовой функции.* Пусть вычисляется интеграл $\int_a^b f(x) dx$, где a и b могут быть и бесконечными. Представим подынтегральную функцию в виде $f(x) = g(x)p(x)$, где $p(x)$ — достаточно простая, а $g(x)$ — гладкая. Далее применяем какой-либо из рассмотренных ранее способов вычисления интегралов с весом. Рассмотрим некоторые примеры.

Пусть вычисляется $\int_{-1}^1 \frac{dx}{\sqrt{1-x^4}}$. Представим $f(x)$ в виде $\frac{1}{\sqrt{1+x^2}} \cdot \frac{1}{\sqrt{1-x^2}}$; функция $\frac{1}{\sqrt{1+x^2}}$ является гладкой. Функцию $\frac{1}{\sqrt{1-x^2}}$ можно рассматривать как весовую. Этой весовой функции соответствует квадратура Эрмита.

Пусть вычисляется $\int_0^1 f(x) dx$, где $f(x)$ может быть представлена в виде $g(x)x^\alpha$, где $-1 < \alpha < 1$, $g(x)$ — гладкая, $g(0) \neq 0$. При вычислении интеграла по формуле трапеций с постоянным шагом $H = M^{-1}$ погрешность стремится к нулю медленнее, чем M^{-2} . Один из возможных способов — обращение к квадратурам Гаусса, соответствующим данной весовой функции.

2. Можно пойти по пути разбиения интеграла на части и вычисления интеграла по каждой части при помощи построений из § 1. Представим интеграл в виде

$$I = \sum_{q=1}^M I_q, \quad I_q = \int_{(q-1)H}^{qH} g(x) x^\alpha dx, \quad H = \frac{1}{M}.$$

Заменим $g(x)$ на интерполяционный многочлен

$$P_{(q)}(x) = g((q-1)H) + (x - (q-1)H) \frac{(g(qH) - g((q-1)H))}{H},$$

получим

$$\begin{aligned}
 I_q &\approx \int_{(q-1)H}^{qH} P_{(q)}(x) x^\alpha dx = \\
 &= H^{\alpha+1} q^{\alpha+2} \left(\left(\frac{1 - \left(1 - \frac{1}{q}\right)^{\alpha+1}}{\alpha+1} - \frac{1 - \left(1 - \frac{1}{q}\right)^{\alpha+2}}{\alpha+2} \right) g(qH) + \right. \\
 &\quad \left. + \left(\frac{1 - \left(1 - \frac{1}{q}\right)^{\alpha+2}}{\alpha+2} - \frac{1 - \frac{1}{q} - \left(1 - \frac{1}{q}\right)^{\alpha+2}}{\alpha+1} \right) g((q-1)H) \right). \quad (1)
 \end{aligned}$$

Суммируя по q правые части (1), получим квадратуру для вычисления исходного интеграла. В ряде случаев будет удобнее положить

$$g(qH) = (qH)^{-\alpha} f(qH)$$

и, таким образом, получить квадратуру, имеющую вид

$$I \approx \sum_{q=0}^N D(q, H) f(qH). \quad (2)$$

З а д а ч а. Для квадратуры (2) получить оценку погрешности

$$\text{const} \cdot \max_{[0, 1]} |g''(x)| M^{-2}.$$

Далее будут рассмотрены более простые по виду способы вычисления интегралов с особенностями. Описанный выше способ аппроксимации интеграла по значениям функции на фиксированной, в частности, равномерной сетке обладает определенными преимуществами, например, при решении интегральных уравнений путем сведения к решению системы алгебраических уравнений.

В некоторых случаях, важных для приложений, необходимая точность уже достигается, когда $g(x)$ заменяется на отрезках разбиения на постоянную. В этом случае полагаем

$$\int_{a_{q-1}}^{a_q} g(x) p(x) dx \approx g(\xi_q) \int_{a_{q-1}}^{a_q} p(x) dx,$$

и квадратура для вычисления исходного интеграла приобретает вид

$$\int_0^1 g(x) p(x) dx \approx \sum_{q=1}^M g(\xi_q) \int_{a_{q-1}}^{a_q} p(x) dx. \quad (3)$$

Пусть вычисляется интеграл

$$I = \int_0^1 \frac{g(x)b}{b^2 + x^2} dx, \quad b - \text{малое.}$$

Задача. Показать, что при вычислении интеграла по формуле трапеций с постоянным шагом $a_q - a_{q-1} \equiv H = M^{-1}$ погрешность оценивается через

$$\text{const} \cdot \min \left(\frac{1}{Mb}, \frac{1}{(Mb)^2} \right). \quad (4)$$

При $a_q - a_{q-1} \equiv H$ квадратура (3) приобретает вид

$$I \approx \sum_{q=1}^M g(\xi_q) \left(\text{arctg} \left(\frac{qH}{b} \right) - \text{arctg} \left(\frac{(q-1)H}{b} \right) \right),$$

где $(q-1)H \leq \xi_q \leq qH$.

Задача. Получить оценку погрешности

$$\text{const} \cdot \max_{[0, 1]} |g'(x)| M^{-1}.$$

В рассматривавшихся выше случаях коэффициенты квадратур записываются в виде интегралов

$$\int_{a_{q-1}}^{a_q} P_l(x) p(x) dx,$$

где $P_l(x)$ — некоторые многочлены; эти интегралы вычисляются в явном виде. Для ряда классов задач, где эти интегралы не вычисляются в явном виде, может оказаться разумным найти эти интегралы при помощи численного интегрирования. Эта дополнительная работа оправдывается, если получившиеся формулы используются многократно, например, при вычислении большой серии интегралов, при вычислении кратных интегралов как повторных (см. гл. V), при решении интегральных уравнений.

3. Пусть теперь вычисляется

$$I_\omega(f) = \int_0^1 f(x) \exp(i\omega x) dx,$$

где ω — большое, $f(x)$ — достаточно гладкая; будем рассматривать функцию $\exp(i\omega x)$ как весовую. Представим интеграл в виде

$$I = \sum_{q=1}^M I_q, \quad I_q = \int_{(q-1)H}^{qH} f(x) \exp(i\omega x) dx,$$

для вычисления интегралов I_q применим квадратуру типа (5.2).

Задача. Для этого способа интегрирования получить оценку погрешности

$$D_n \cdot \max_{[0, 1]} |f^{(n)}(x)| M^{-n},$$

где D_n не зависит от ω .

4. В некоторых случаях подынтегральную функцию можно представить в виде $f(x) = G(x) + g(x)$, причем $\int_A^B G(x) dx$ берется в явном виде, а $g(x)$ — гладкая. Пусть вычисляется интеграл

$$I = \int_0^1 f(x) dx, \quad f(x) = \frac{\ln x}{1+x^2}. \quad (5)$$

Возьмем $G(x) = \ln x$. Тогда

$$g(x) = -\frac{x^2 \ln x}{1+x^2};$$

величина $\int_0^1 |g''(x)| dx$ будет конечна, и поэтому погрешность

вычисления интеграла $\int_0^1 g(x) dx$ по формуле трапеций с постоянным шагом $a_q - a_{q-1} = M^{-1}$ имеет порядок $O(M^{-2})$. Чтобы погрешность формулы Симпсона имела порядок $O(M^{-4})$, следует взять $G(x) = (1-x^2) \ln x$.

В случае $f(x) = (x^2 + b^2)^{-1} \exp x$, b — малое, целесообразно взять

$$G(x) = \left(\text{Выч. } f(z) \right)_{bt} (x - bi)^{-1} + \left(\text{Выч. } f(z) \right)_{-bt} (x + bi)^{-1};$$

достигаемое здесь расширение области аналитичности подынтегральной функции особенно эффективно ((3.8), (4.10.1)) в случае использования формул Гаусса.

5. Другим способом устранения особенности подынтегральной функции является замена переменной интегрирования. При замене переменных $x = \varphi(t)$, $\varphi(0) = 0$, $\varphi(1) = 1$ исходный интеграл $I = \int_0^1 f(x) dx$ преобразуется к виду

$$\int_0^1 g(t) dt, \quad (6)$$

где

$$g(t) = f(\varphi(t)) \varphi'(t).$$

За счет множителя $\varphi'(t)$ происходит устранение особенностей подынтегральной функции в отдельных точках. Произведем в интеграле (5) замену переменной $x = t^h$, тогда получим

$$I = \int_0^1 \frac{k^2 t^{k-1} \ln t}{1 + t^{2k}} dt;$$

при $k > 2$ интеграл $\int_0^1 |g''(t)| dt$ конечен, и следовательно, погрешность формулы трапеций имеет порядок $O(M^{-2})$. С увеличением k растет порядок производных $g^{(n)}(t)$, для которых интеграл $\int_0^1 |g^{(n)}(t)| dt$ ограничен, и поэтому можно применять квадратуры все более высокого порядка точности.

Однако если k — очень большое, то производные функции (t) хотя и конечны, но очень большие; поэтому должна соблюдаться определенная пропорция между величиной k и числом узлов N . Необходимость соблюдения осторожности при употреблении очень больших k видна хотя бы из следующего. Постоянный шаг $t_q - t_{q-1} = \frac{1}{M}$ в интеграле (6) соответствует узлам интегрирования $a_q = \varphi\left(\frac{q}{M}\right) = \left(\frac{q}{M}\right)^k$ в исходном интеграле; по-

этому при больших k используется мало значений подынтегральной функции в правой части отрезка (рис. 3.16.1).

6. Как мы уже видели в § 10, к повышению скорости сходимости при вычислении интегралов от

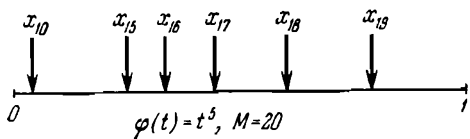


Рис. 3.16.1.

функций с особенностями приводит также оптимизация распределения узлов интегрирования.

7. В некоторых случаях приходится идти по пути сочетания некоторых из описанных способов. Пусть вычисляется интеграл

$$\int_0^1 g(x) x^\alpha \exp(i\omega x) dx,$$

где ω — большое, $g(x)$ — гладкая, $g(0) \neq 0$, $|\alpha| < 1$. Наличие мно-

жителя $\exp(i\omega x)$ требует выделения его как весового. Наличие множителя x^α требует принятия специальных мер для интегрирования в окрестности точки $x = 0$. Замена переменных $x = \varphi(t)$ в данном случае является неприемлемой, поскольку для соответствующей весовой функции $\exp(i\omega \varphi(t))$ невозможно вычисление в явном виде коэффициентов интерполяционных квадратурных формул. Приемлемым здесь является путь разбиения отрезка на неравные части, соответствующие оптимальному распределению узлов при вычислении интеграла от функции x^α , и применение на каждой части интерполяционных квадратурных формул (5.2), соответствующих весовой функции $\exp(i\omega x)$.

В случае интегралов типа $\int_0^1 g(x) x^{-\alpha} \sin \omega x dx$ при $\alpha > 1$, $g(x)$ гладкой, $g(0) \neq 0$, такой способ будет неприемлем, поскольку в окрестности $x = 0$ неинтегрируемая функция $x^{-\alpha}$ не аппроксимируется многочленами. Здесь целесообразно разбить исходный интеграл на части $\int_0^\varepsilon$ и $\int_\varepsilon^1$, где ε порядка $\frac{1}{\omega}$; для вычисления

второго интеграла разумно применить процедуру, только что описанную выше. В первом интеграле функция $\sin \omega x$ не играет роли сильно осциллирующего множителя, поскольку при таком выборе ε она имеет там конечное число колебаний; поэтому этот интеграл можно вычислять, например, распределив узлы интегрирования соответственно оптимальному распределению для функции $g(0)\omega x^{1-\alpha}$, аппроксимирующей подынтегральную при малых ωx .

8. Для полноты изложения упомянем метод Ромберга. Мы уже рассматривали возможность применения этого метода при вычислении интегралов от функций с особенностями на примере

вычисления интеграла от функции $\int_0^1 g(x) \ln x dx$. Погрешность формулы трапеций с постоянным шагом при вычислении инте-

грала $\int_0^1 g(x) x^\alpha dx$ для $g(x)$ гладкой, $g(0) \neq 0$, $-1 < \alpha \leq 1$,

представляется в виде $D_1 N^{-1-\alpha} + D_2 N^{-2-\alpha} + \dots$; таким образом, имеются основания для применения процедуры Ромберга.

9. Решение ряда задач сводится к вычислению сингулярных интегралов типа

$$I(a) = \int_a^B \frac{g(x)}{x-a} dx,$$

где $a \in (A, B)$, $g(a) \neq 0$. Интеграл понимается в смысле главного значения, т. е. как предел

$$\lim_{\varepsilon \rightarrow 0} \left(\int_A^{a-\varepsilon} \frac{g(x)}{x-a} dx + \int_{a+\varepsilon}^B \frac{g(x)}{x-a} dx \right).$$

Интеграл может быть записан как сумма интеграла по отрезку, симметричному относительно точки a , и интеграла от гладкой функции по оставшейся части. Для простоты предполагаем, что

первый интеграл преобразован к виду $\int_{-b}^b \frac{g(x)}{x} dx$. Если $g(x)$

удовлетворяет условию Гельдера в точке $x=0$, т. е. $|g(x) - g(0)| \leq A|x|^\alpha$, $\alpha > 0$, то последний интеграл равен несингу-

лярному интегралу $\int_0^b \frac{g(x) - g(-x)}{x} dx$. В частности, если

$g(x)$ — гладкая, то новая подынтегральная функция $\frac{g(x) - g(-x)}{x}$ является гладкой.

В ряде случаев, например при решении интегральных уравнений с сингулярными ядрами, возникает следующая ситуация. Значения функции $g(x)$ задаются на некоторой фиксированной сетке. Исходя из информации об этих значениях, требуется вычислять значения интеграла $I(a)$ при различных a . Если $g(x)$ достаточно гладкая, то здесь можно поступить следующим образом.

Разбиваем отрезок $[A, B]$ на части:

$$[A_0, A_1], \dots, [A_{M-1}, A_M], \quad A_0 = A, \quad A_M = B.$$

На каждой из частей $[A_{q-1}, A_q]$ приближаем $g(x)$ интерполяционным многочленом $L^q(x)$. Исходный интеграл заменяем суммой интегралов

$$i_q = \int_{A_{q-1}}^{A_q} \frac{L^q(x)}{x-a} dx.$$

Интегралы i_q вычисляем в явном виде. Если $a \in (A_{q-1}, A_q)$, то соответствующий интеграл i_q следует рассматривать как сингулярный. Если $a = A_q$, то следует объединить интегралы i_q и i_{q+1} (расходящиеся) в один сингулярный

$$i_q + i_{q+1} = \int_{A_{q-1}}^{A_{q+1}} \frac{\tilde{L}^q(x)}{x-a_q} dx,$$

$$\tilde{L}^q(x) = \begin{cases} L^q(x) & \text{при } x < A_q, \\ L^{q+1}(x) & \text{при } x > A_q. \end{cases}$$

Заметим, что для существования этого интеграла как сингулярного обязательно выполнение условия

$$L^q(A_q) = L^q(A_{q+1}).$$

Получившиеся интегралы вычислим в явном виде.

Задача. Пусть отрезок интегрирования разбит на равные части длины H и на каждой части функция $g(x)$ аппроксимируется при помощи линейной интерполяции. Таким образом, исходный интеграл аппроксимируется суммой интегралов

$$\sum_{q=1}^M \int_{A+(q-1)H}^{A+qH} \frac{g((q-1)H) + (x - (q-1)H) \frac{g(qH) - g((q-1)H)}{H}}{x - a} dx,$$

где $H = (B - A)M^{-1}$. В предположении ограниченности $|g''(x)|$ получить оценку погрешности $O(M^{-2} \ln M)$.

§ 17. Принципы построения стандартных программ с автоматическим выбором шага

Из рассмотрения примера 1 в § 10 видно, что интегралы от функций с особенностями типа x^b хорошо вычисляются методами интегрирования с переменным шагом, если узлы интегрирования распределены оптимальным образом. По-видимому, столь же хорошо будут интегрироваться функции с особенностями других типов. Поэтому представляется заманчивым строить стандартные программы численного интегрирования так, что для любой функции распределение узлов является оптимальным или близким к нему.

В § 9 даны наметки того, как можно было бы построить распределение узлов, близкое к оптимальному, после исследования поведения подынтегральной функции на редкой сетке. Однако в случае резко меняющихся функций, например функций типа x^b , реализация этих наметок приводит к недостаточно удовлетворительным методам. Если первоначальная сетка, на которой исследуется подынтегральная функция, частая, то сильно загружается память ЭВМ; если она редкая, то не удастся хорошо аппроксимировать оптимальное распределение узлов на участках резкого изменения подынтегральной функции.

Поэтому при построении стандартных программ интегрирования приняты несколько иные процедуры распределения узлов интегрирования, обеспечивающие лучшее приближение к оптимальному распределению узлов для функций с особенностями. Рассмотрим некоторые из них.

Для вычисления интегралов по элементарным отрезкам разбиения $[a_{q-1}, a_q]$ выбираются: квадратурная формула

$$I_q(f) \approx \frac{a_q - a_{q-1}}{2} \sum_{j=1}^m D_j f\left(\frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \quad (1)$$

и мера (§ 4) погрешности

$$\rho_q(j) = \left| \frac{a_q - a_{q-1}}{2} \sum_{j=1}^m B_j f \left(\frac{a_{q-1} + a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j \right) \right|. \quad (2)$$

Пусть вычисляется

$$\int_A^B f(x) dx, \quad A = \alpha_0.$$

Первая процедура, которую естественно назвать *горизонтальной*, определяется заданием параметров β , $\sigma < 1$, h_0 и ϵ_0 . Полагая $\epsilon_1 = \epsilon_0 \beta$. Предположим, что каким-то образом уже вычис-

лено приближенное значение интеграла $\int_{\alpha_0}^{\alpha_q} f(x) dx$. Программа

располагает в каждый момент времени некоторым значением h_q , с которым надо начинать считать оставшуюся часть интеграла. Вычисляем величину $\rho_q(f)$, соответствующую отрезку $[a_q, a_q + h_q]$. Если оказалось $\rho_q(f) \leq \epsilon_0$, то вычисляем приближенное значение $\int_{a_q}^{a_q+h_q} f(x) dx$ по формуле (1) и полагаем $a_{q+1} = a_q + h_q$. Мы полу-

чили приближенное значение величины $\int_{\alpha_0}^{a_{q+1}} f(x) dx$. В случае

$\epsilon_1 < \rho_q(f)$ полагаем $h_{q+1} = h_q$, в противном случае полагаем $h_{q+1} = \frac{h_q}{\sigma}$. Мы готовы к следующему шагу. Если оказалось $\rho_q(f) > \epsilon_0$, то принимаем σh_q за новое значение величины h_q и возвращаемся к исходной позиции: вычислено значение интеграла $\int_{\alpha_0}^{a_q} f(x) dx$ и задан шаг h_q . Начальные условия для применения процедуры:

$$q = 0, \quad \int_{\alpha_0}^{\alpha_q} f(x) dx = 0, \quad h_0.$$

Процедура должна также иметь блок окончания работы: если оказалось, что $a_q + h_q > B$, то следует положить $h_q = B - a_q$. Установилась практика брать $\sigma = \frac{1}{2}$.

Другая процедура, которую можно назвать *вертикальной*, определяется заданием числа ϵ_0 и заключается в следующем. Пусть на каком-то шаге возникает необходимость вычисления

интеграла по отрезку разбиения $[c, d]$: $\int_c^d f(x) dx$; вычисляется

величина $\rho(f)$, соответствующая этому отрезку. Если она оказалась меньше ϵ_0 , то этот интеграл вычисляется по соответствующей формуле (1) и программа переходит к следующему справа отрезку разбиения. В противном случае отрезки $[c, (c+d)/2]$ и $[(c+d)/2, d]$ объявляются отрезками разбиения и программа обращается к вычислению интеграла по левому из этих отрезков. В начале работы программа обращается к вычислению

исходного интеграла $\int_A^B f(x) dx$. Некоторым недостатком этой

процедуры является необходимость запоминания отрезков разбиения, интегрирование по которым на данный момент времени не произведено. Распределение узлов, осуществляемое этими процедурами, не является асимптотически оптимальным в смысле, определяемом в § 9, по следующим причинам: величина $\rho_q(f)$ не есть главный член оценки погрешности квадратуры (1), а, как правило, лишь некоторая грубая, завышенная по порядку оценка для него; отрезки разбиения могут принимать лишь довольно редкий ряд значений; для первой процедуры они имеют вид $h_0 \sigma^k$, а для второй — $(B-A)2^{-k}$.

Рассмотрим некоторые моменты, связанные с практическим использованием описанных выше процедур.

Чтобы в отдельных случаях сделать распределение узлов более близким к оптимальному, иногда величины $\rho_q(f)$, соответствующие отрезкам разбиения, сравнивают не с ϵ_0 , а с $\epsilon_0 h_q^\gamma$, где γ специально подбирается. На первоначальном этапе применения ЭВМ сложилась традиция производить сравнение величины $\rho_q(f)$ с величиной $\epsilon_0 h_q$, т. е. брать $\gamma = 1$. Оказалось, что для резко меняющихся функций получающееся распределение узлов было далеко от оптимального, а в случае функций с разрывами программа не могла сосчитать интеграл. В самом деле, пусть, например,

$$f(x) = \begin{cases} 0 & \text{при } x < c, \\ 1 & \text{при } x > c. \end{cases}$$

Если при каком-то q $\rho_q(f)$ содержит одновременно значения f в узлах, где $f = 0$ и $f = 1$, то величина $\rho_q(f)$ имеет порядок h_q , и поэтому нет оснований надеяться, что неравенство

$$|\rho_q(f)| \leq \epsilon_0 h_q$$

будет выполняться при малых h_q ; программа будет осуществлять дробление шага до машинного нуля. Если функция меняется на малом участке очень резко, шаг будет мельчиться неоправданно сильно. Это обстоятельство было отмечено пользователями, и после теоретического анализа было принято полагать $\gamma = 0$, если к противному нет особых причин.

В случае резко меняющихся функций следует также иметь в виду, что программа может не заметить участка резкого изменения функции. Пусть $\rho(f) = 0$, если $f(x) = P_s(x)$ — многочлен степени s , и пусть реальная подынтегральная функция есть

$$f(x) = P_s(x) + g(x),$$

где $g(x)$ практически отлична от 0 лишь на малом отрезке, например,

$$g(x) = -\frac{1}{i \cdot 2\pi\epsilon} \exp\left(\frac{(x-C)^2}{2\epsilon}\right)$$

(рис. 3.17.1). Для определенности обратимся к первой из описанных выше процедур. Если отрезок разбиения $[a_{q-1}, a_q]$ удален от точки C , то $\rho(f) \approx \rho(P_s) = 0$, и поэтому программа будет

увеличивать шаг. При подходе к точке C шаг может оказаться настолько большим, что окрестность, где $g(x)$ существенно больше 0, окажется заключенной между узлами. Тогда $\rho(f)$ будет близко к 0 и на отрезке, содержащем точку C . В результате в качестве значения $I(f)$ мы получим значение $I(P_s)$; ошибка этого приближенного значения близка к 1. Для контроля точности можно было бы попытаться произвести интегрирование с другим значением ϵ_0 ; однако с большой вероятностью получился бы тот же результат.

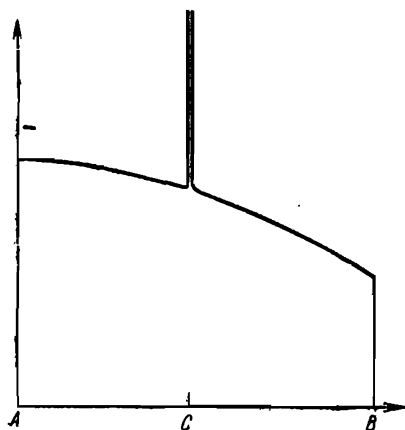


Рис. 3.17.1.

Не следует думать, что этот недостаток свойствен только методам интегрирования с автоматическим выбором шага. Если производить вычисление этого интеграла по формулам с постоянным шагом H , то при $H \gg \sqrt{\epsilon}$ все равно может оказаться, что $g(x) \approx 0$ во всех узлах интегрирования и мы получим приближенное значение $I^* \approx I(P_s)$. Производя численное интегрирование при нескольких шагах и оценивая погрешность по правилу Рунге, можно прийти к неправильному выводу, что прибли-

женное значение интеграла $I^* \approx I(P_s)$. В целом можно сказать, что в случае алгоритмов интегрирования с автоматическим выбором шага возможность получения подобных неправильных выводов несколько больше.

При составлении стандартных программ всегда приходится балансировать между двумя крайностями: гарантии требуемой точности для любой подынтегральной функции или быстрого вычисления интеграла с нужной точностью для большинства предъявляемых к решению задач.

По-видимому, можно говорить о каком-то принципе неопределенности по отношению к методам высокой эффективности: дальнейшее повышение скорости работы должно сопровождаться уменьшением надежности. Чтобы избежать «проскакивания» областей резкого изменения функции, можно предусмотреть в программе наличие отрезков, где шаг интегрирования дробится принудительно. Например, в описание одной из стандартных программ, рассматриваемых в § 18, входят еще концы некоторого отрезка $[\alpha, \beta]$. Если отрезок разбиения $[a_{q-1}, a_q]$ пересекается с $[\alpha, \beta]$, то следующий отрезок разбиения выбирается по более сложному правилу. В качестве $[\alpha, \beta]$ следует задавать некоторый отрезок, где подынтегральная функция резко меняется.

Обратим внимание на выбор значения σ . Пусть для гладкой функции $f(x)$

$$\rho(f) = c_s \left| f^{(s)} \left(\frac{a_{q-1} + a_q}{2} \right) \right| (a_q - a_{q-1})^{s+1} + o((a_q - a_{q-1})^{s+1}). \quad (3)$$

Для понимания механизма изменения шага предположим, что $f^{(s)}(x)$ — кусочно-постоянная. В области постоянства $f^{(s)}(x)$ имеем

$$\rho(f) = c_s |f^{(s)}(x)| (a_q - a_{q-1})^{s+1}.$$

Рассмотрим случай $\beta < \sigma^{s+1}$. Пусть для шага h^0 , с которым мы приступаем к интегрированию в данной области постоянства производной, выполняется соотношение $c_s |f^{(s)}(x)| (h^0)^{s+1} \leq \epsilon_1$. Тогда шаг будет укрупняться каждый раз в σ^{-1} раз, пока величина $\rho(f)$ не станет больше ϵ_1 . Поскольку при шаге, в σ^{-1} раз меньшем, было $\rho(f) \leq \epsilon_1$, то интегрирование будет осуществляться с наименьшим шагом $h = h^0 \sigma^k$, для которого $c_s |f^{(s)}(x)| h^{s+1} > \epsilon_1$. Точно так же, если для начального в этой области шага выполняется соотношение $c_s |f^{(s)}(x)| (h^0)^{s+1} > \epsilon_0$, то шаг станет наибольшим из шагов $h^0 \sigma^k$, для которых $c_s |f^{(s)}(x)| h^{s+1} \leq \epsilon_0$. Если, как мы предположили, $\frac{\epsilon_1}{\epsilon_0} = \beta < \sigma^{s+1}$, то

может оказаться, что, в зависимости от поведения функции при меньших x , в рассматриваемой области выбираются различные шаги. В то же время мы видели, что шаг интегрирования желательно распределять так, чтобы оценка погрешности на всех отрезках разбиения интервала интегрирования была примерно одинаковой. Исходя из сказанного, мы делаем вывод о желательности употребления $\beta \geq \sigma^{s+1}$.

Рассмотрим этот случай. Можно указать шаг $h = h^0 \sigma^k$ такой, что

$$\epsilon_0 \sigma^{s+1} < c_s |f^{(s)}(x)| h^{s+1} \leq \epsilon_0. \quad (4)$$

Если оказалось, что

$$e_1 < c_s |f^{(s)}(x)| h^{s+1} \leq e_0, \quad (5)$$

то при интегрировании по рассматриваемому отрезку будет выбран именно такой шаг. Однако при

$$e_0 \sigma^{s+1} < c_s |f^{(s)}(x)| h^{s+1} \leq e_1 \quad (6)$$

программа выработает следующий шаг, равный h/σ . Для этого шага, вследствие (6), окажется $e_0 < \rho(f)$, и поэтому шаг h/σ будет признан непригодным. Программа возьмет шаг, равный h , для него выполняется (6); поэтому интегрирование по текущему отрезку будет закончено и как исходный шаг для дальнейшего будет принят шаг h/σ . Таким образом, фактическое интегрирование происходит с шагом h и на каждом шаге делается попытка произвести интегрирование с шагом h/σ . Если вычисления с шагами h и h/σ независимы, то на каждом шаге будет удвоенный объем работы. Неравенство (5) равносильно неравенству

$$\log_{\sigma^{s+1}} e_1 \geq \log_{\sigma^{s+1}} |f^{(s)}(x)| + \log_{\sigma^{s+1}} (c_s h_0^{s+1}) + k \geq \log_{\sigma^{s+1}} e_0. \quad (7)$$

Величина $\{\log_{\sigma^{s+1}} |f^{(s)}(x)|\}$ равномерно распределена на $[0, 1]$, согласно сообщавшемуся нами на стр. 171 факту о равномерности при любом a распределения $\{\log_a g\}$ значений величин g , возникающих из независимых источников.

Можно показать, что вероятность события: неравенство (7) выполняется при некотором целом k , равна $\log_{\sigma^{s+1}} \beta$.

Если неравенство (7) не выполняется, то при $\sigma^{s+1} \leq \beta$ выполняется (6); следовательно, (6) выполняется с вероятностью $1 - \log_{\sigma^{s+1}} \beta$. Таким образом, для используемой процедуры интегрирование производится с шагом h , определяемым условием (4), причём математическое ожидание затраты работ на шаге пропорционально

$$\log_{\sigma^{s+1}} \beta + 2(1 - \log_{\sigma^{s+1}} \beta) = 2 - \log_{\sigma^{s+1}} \beta.$$

Мы получили количественное выражение среднего возрастания затрат при $\beta > \sigma^{s+1}$. Это выражение показывает желательность выбора $\beta = \sigma^{s+1}$. Равенство $\rho(f) = c_s |f^{(s)}(x)| (a_q - a_{q-1})^{s+1}$ при $f^{(s)}(x) \neq \text{const}$ является асимптотическим. Поэтому для некоторых классов задач проводилось исследование поведения $\rho(f)$ при реальных не очень малых $a_q - a_{q-1}$. Там, где реальный порядок убывания $\rho(f)$ оказывался иным, величина β , соответственно, изменялась.

Как и в случае интегрирования с постоянным шагом, возникает вопрос практической оценки погрешности. Обратимся сначала ко второй процедуре, имеющей более простое описание. Как и ранее, проводим рассмотрение на примере кусочно-постоянной $f^{(s)}(x)$. Шаг интегрирования в области постоянства $f^{(s)}(x)$ для этой процедуры определяется условием: это наибольший шаг h вида $(B - A)2^{-k}$, для которого

$$c_s |f^{(s)}(x)| h^{s+1} \leq e_0. \quad (8)$$

Очевидно, что для такого максимального шага выполняется соотношение

$$e_0 2^{-(s+1)} < c_s |f^{(s)}(x)| h^{s+1}.$$

Произведем интегрирование с некоторым $e'_0 < e_0$. В области, где

$$e'_0 < c_s | f^{(s)}(x) | h^{s+1} \leq e_0,$$

старый шаг интегрирования будет уже непригоден и поэтому произойдет дробление шага не менее чем вдвое. В области, где

$$e_0 2^{-(s+1)} < c_s | f^{(s)}(x) | h^{s+1} \leq e'_0,$$

дробления шага интегрирования не произойдет. Таким образом, при $e_0 2^{-(s+1)} < e'_0$ может случиться, что дробление шага интегрирования произойдет лишь на части отрезка интегрирования, или вообще не произойдет. Из сравнения результатов расчетов с e_0 и e'_0 мы не получим информации о величине погрешности интегрирования в той области, где дробление шага не произошло. При

$$e'_0 \leq e_0 2^{-(s+1)} \quad (9)$$

шаг дробится всюду; однако отсюда не следует, что сравнение результатов расчетов с e_0 и любым e'_0 , удовлетворяющим (9), дает надежную гарантию точности результата. Пусть

$$e_0 2^{-(l+1)(s+1)} < e'_0 < e_0 2^{-l(s+1)},$$

в области, где ранее было

$$e'_0 2^{l(s+1)} < c_s | f^{(s)}(x) | h^{s+1} \leq e_0, \quad (10)$$

произойдет дробление шага в l раз; в то же время в области, где ранее

$$e_0 2^{-(s+1)} < c_s | f^{(s)}(x) | h^{s+1} \leq e'_0 2^{l(s+1)}. \quad (11)$$

произойдет дробление шага в $l+1$ раз. Пусть погрешность интегрирования с параметром e_0 складывается из двух слагаемых

$$R_{e_0}(f) = R_{e_0}^1(f) + R_{e_0}^2(f);$$

первое слагаемое равно сумме погрешностей по области, где выполняется (10), а второе — по области, где выполняется (11). Погрешности интегрирования с e'_0 запишем, соответственно, в виде

$$R_{e'_0}(f) = R_{e'_0}^1(f) + R_{e'_0}^2(f).$$

Предположим, что для гладкой $f(x)$ погрешность квадратуры (1) имеет вид:

$$\gamma n f^{(n)} \left(\frac{a_{q-1} + a_q}{2} \right) (a_q - a_{q-1})^{n+1} + o((a_q - a_{q-1})^{n+1}).$$

Если выполнены все условия, обеспечивающие в § 11 существование главного члена погрешности, то при уменьшении шага интегрирования по некоторой области в 2^k раз погрешность интегрирования уменьшится в 2^{kn} раз. Соответственно, будем предполагать, что выполняются соотношения

$$R_{e'_0}^1(f) \sim R_{e_0}^1(f) 2^{-nl}, \quad R_{e'_0}^2(f) \sim R_{e_0}^2(f) 2^{-n(l+1)}. \quad (12)$$

В предположении справедливости (12) имеем

$$I(f) - S_{e_0}(f) = R_{e_0}^1(f) + R_{e_0}^2(f), \quad (13)$$

$$I(f) - S_{e'_0}(f) = R_{e'_0}^1(f) 2^{-nl} + R_{e'_0}^2(f) 2^{-n(l+1)}. \quad (14)$$

Исключая неизвестную величину $I(f)$, получим

$$S_{e'_0}(f) - S_{e_0}(f) = R_{e_0}^1(f) (1 - 2^{-nl}) + R_{e_0}^2(f) (1 - 2^{-n(l+1)}). \quad (15)$$

Правая часть (15) не пропорциональна ни правой части (13), ни правой части (14), и мы не имеем возможности получить главного члена погрешностей $R_{e_0}^1(f)$ и $R_{e'_0}^1(f)$. Если nl очень велико, то правая часть (15) близка к погрешности $R_{e_0}^2(f)$. Однако употребление этого выражения для оценки погрешности выражения $S_{e_0}(f)$ вряд ли имеет смысл, поскольку интересно оценить погрешность более точного приближения $S_{e'_0}(f)$. В случае $e'_0 = 2^{-nl}e_0$ область, где выполняется (11), вырождается, и поэтому $R_{e'_0}^2(f) = 0$. Тогда

$$I(f) - S_{e_0}(f) = R_{e_0}^1(f), \quad I(f) - S_{e'_0}(f) = R_{e'_0}^1(f) 2^{-nl}$$

и имеется возможность применять правило Рунге для оценки погрешности.

Довольно сложный дополнительный анализ показывает, что в предположении непрерывности $f^{(n)}(x)$, при введении в алгоритм требования уменьшения максимально допустимого шага пропорционально $e_0^{1/(s+1)}$, проводимые выше рассуждения можно обосновать.

Обратимся к первой процедуре выбора шага. Для определенности предполагаем $\beta = \sigma^{s+1}$, хотя проводимые далее рассуждения остаются в силе при любом $\beta < 1$. Если мы умножим (4) на λ^{s+1} , то получим

$$e_0 \lambda^{s+1} \sigma^{s+1} < c_s |f^{(s)}(x)| (\lambda h)^{s+1} \leq e_0 \lambda^{s+1}.$$

Отсюда следует, что при одновременном умножении e_0 на λ^{s+1} и h_0 на λ шаг интегрирования всюду изменится пропорционально: умножится на λ . Поэтому, если выполняются все предположения из § 11, суммарная погрешность интегрирования изменится примерно в λ^n раз. Таким образом, имеем соотношение

$$I(f) - S_{\lambda^{s+1}e_0}^{\lambda h_0}(f) \approx (I(f) - S_{e_0}^{h_0}(f)) \lambda^n, \quad (16)$$

где $S_e^H(f)$ — результат интегрирования с начальным шагом H при значении e параметра точности на шаге. Мы можем использовать соотношение (16) для оценки погрешности по принципу Рунге при численном интегрировании с автоматическим выбором шага.

Итак, мы получили соотношения между параметрами схемы, необходимые для возможности применения правила Рунге для оценки погрешности.

Заметим, что изменение начального шага h_0 в σ^k раз, где k — целое, при фиксированных остальных параметрах метода не приводит к существенному изменению погрешности результата: за счет уменьшения шага в σ раз или увеличения в $1/\sigma$ раз программа очень быстро выйдет на шаг, удовлетворяющий условию

$$e_0 2^{-(s+1)} < c_s |f^{(s)}(x)| h^{s+1} \leq e_0.$$

§ 18. Стандартные программы численного интегрирования

Описания стандартных программ обычно состоят из следующих частей: целевое назначение, обращение к процедуре, особенности применения, описание метода, тело процедуры, тестовые примеры.

Приведем для примера отдельные части описаний реально работающих процедур численного интегрирования на языке Алгол.

Так как основные принципы построения программ вычисления одно- и многократных интегралов примерно одинаковые, будет приведена одна процедура вычисления кратных интегралов.

1. Процедура *INT 120*. Обращение к процедуре имеет вид $Z := INT\ 120\ (A, B, F, P, I, X)$. Формальные параметры процедуры *REAL A* — нижний предел интегрирования; *REAL B* — верхний предел интегрирования; *REAL PPROCEDURE F* — процедура-функция без параметров, задающая подынтегральную функцию $f(x)$. *INTEGER P* — условное число, принимающее значение 1 или 2 в зависимости от применяемой квадратурной формулы; если $P = 1$ — применяется формула Гаусса, если $P = 2$ — формула Симпсона; *INTEGER I* — число, задающее набор узлов, по которому будет сосчитан интеграл; количество узлов N определяется процедурой по правилу: $N = 2I$, если счет ведется по формуле Гаусса; $N = 2I + 1$, если счет ведется по формуле Симпсона. В случае $P = 2$ число I может быть любым, большим 0, в случае $P = 1$ может принимать только следующие значения:

$$I = 1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48;$$

REAL X — аргумент функции $f(x)$; *REAL Z* — вычисленное значение интеграла.

Процедуры *INT 110* и *INT 111* предназначены для вычисления интегралов типов (а) и (б):

$$(a) \int_A^B f(x) dx, \quad (б) \int_A^B f(x) \cos(\rho x + \varphi) dx.$$

2. Процедура *INT 110*. Обращение к процедуре имеет вид

$$Z := INT\ 110(X, A, B, RO, FI, E, F, L).$$

Формальные параметры процедуры: *REAL X* — аргумент функции $f(x)$, *REAL A* — нижний предел интегрирования; если он равен $-\infty$, то нужно поставить произвольное отрицательное

$A \geq -10^{19}$ такое, что $\int_{-\infty}^A f(x) dx$ мал; *REAL B* — верхний предел

интегрирования; если он равен ∞ , то нужно поставить произвольное положительное $B \leq 10^{19}$ такое, что $\int_B^{\infty} f(x) dx$ мал.

Верхний предел интегрирования B всегда должен быть больше нижнего A ; *REAL RO* и *FI* — параметры ρ и ϕ из осциллирующего множителя $\cos(\rho x + \phi)$; при вычислении интегралов типа (а) формальные параметры *RO* и *FI* следует положить нулями; *REAL E* — задаваемая точность вычисления интеграла; приближенное значение интеграла отыскивается с погрешностью

$E \left(1 + \left| \int_A^B f(x) dx \right| \right)$; *REAL PROCEDURE F* — процедура-функция без параметров, задающая подынтегральную функцию $f(x)$;

LABEL L — метка, на которую произойдет уход из процедуры *INT 110* в случае, когда нельзя сосчитать интеграл с заданной точностью, или если пользователь в обращении к процедуре *INT 110* в фактических параметрах задаст нижний предел A больше верхнего B ; *REAL Z* — вычисленное значение интеграла.

Особенности применения. 1) Если подынтегральная функция имеет на отрезке интегрирования пикообразную особенность (см. рис. 3.17.1), то разумнее пользоваться описываемой далее процедурой *INT 111*.

2) Величина E характеризует абсолютную погрешность приближенного значения интеграла, если оно меньше единицы, и относительную в противном случае. Возможны случаи, когда пользователь задаст слишком большую точность, которая не может быть достигнута для данной подынтегральной функции. Тогда на печать выдается фраза «заданная точность не может быть достигнута», и произойдет уход из процедуры по метке L .

3) Пользователь должен иметь свой блок, помеченный меткой L , в котором предусмотрена работа программы в случае отказа. Например, простейший блок $Y := 1/0$ вызывает останов машины.

3. Процедура *INT 111*. Обращение к процедуре имеет вид $Z := \text{INT } 111 (X, SI, A, B, RO, FI, ALFA, BETA, TAU, EPS, E, E1, F, L)$.

Формальные параметры процедуры, не определенные ранее в описании процедуры *INT 110*: *INTEGER SI* — условное число, позволяющее выдавать промежуточные значения счета интеграла; *SI* может принимать значения только 0 или 1; если $SI = 0$, то выдач не будет; *REAL ALFA* — начало участка отрезка интегрирования, который нужно проходить с принудительным шагом (начало особенности); *REAL BETA* — конец участка интегрирования, который нужно проходить с принудительным шагом (конец особенности); *REAL*

TAU — принудительный шаг, с которым нужно идти на отрезке $[\alpha, \beta]$, где функция имеет особенность; $REAL\ EPS$ — величина допустимой абсолютной погрешности вычисления интеграла на шаге; $REAL\ E1$ — оценка абсолютной погрешности, с которой сосчитан интеграл, возможно несколько завышенная.

Особенности применения. 1) Если подынтегральная функция $f(x)$ имеет пикообразную особенность (см. рис. 3.17.1) в некоторой точке на отрезке интегрирования, то целесообразно окрестность этой точки (обозначим ее $[\alpha, \beta]$) пройти с принудительным шагом $\tau(TAU)$. Необходимо, чтобы $\alpha(ALFA)$ было меньше $\beta(BETA)$; если нет участка интегрирования, который нужно пройти с принудительным шагом, то формальным параметрам α, β, τ можно присвоить любые значения, но обязательно положить $\beta < \alpha$.

2) При вычислении интеграла пользователь должен задать значение абсолютной погрешности вычисления интеграла на шаге $\epsilon(EPS)$ и значение погрешности вычисления интеграла на всем интервале E ; если нет никакой информации о целесообразном выборе значения ϵ , то разумно задать $\epsilon = E$.

3) Если пользователь задаст параметр $SI = 1$, то в процессе счета будут выдаваться на печать само число SI и приближенное значение интеграла при значениях параметра EPS , равных $\epsilon, \frac{\epsilon}{128}, \frac{\epsilon}{128^2}, \dots$; выдач столько, сколько просчетов по данному отрезку, пока не будет достигнута заданная точность E .

Описание методов интегрирования в случае интегралов типа (а). Бесконечная прямая, по которой нужно считать интеграл, делится точкой M на 2 луча; имеем

$$\int_{-\infty}^{\infty} f(x) dx = \int_M^{\infty} f(x) dx - \int_M^{-\infty} f(x) dx.$$

Если есть особенность, то $M = \alpha$, в противном случае $M = 0$. Счет по бесконечному лучу происходит следующим образом. Луч делится на отрезки Δ_i длины $|\Delta_1| = 2, \dots, |\Delta_n| = 2^n$. Начало отрезка Δ_1 совпадает с точкой M , конец — либо с точкой $M + 2$, если считается $\int_M^{\infty} f(x) dx$, либо с точкой $M - 2$

для интеграла $\int_M^{-\infty} f(x) dx$. Таким образом,

$$\int_M^{\pm\infty} f(x) dx = \sum_{i=1}^{\infty} \int_{\Delta_i} f(x) dx.$$

На каждом отрезке Δ_i интеграл считается с автоматическим выбором шага по формулам Лобатто. Обозначим значение интеграла на отрезке Δ_i через I_i , а абсолютную погрешность, с которой сосчитан интеграл, через E_i . Сосчитаем последовательно интегралы по отрезкам $\Delta_1, \Delta_{1+1}, \Delta_{1+2}$. Счет интеграла

по бесконечному лучу прекращается, если выполнены следующие условия:

$$0 \leq \frac{|I_{i+1}|^2}{|I_i| - |I_{i+1}|} < E \quad \text{и} \quad 0 \leq \frac{|I_{i+2}|^2}{|I_{i+1}| - |I_{i+2}|} < E. \quad (1)$$

Пусть

$$\sum_{k=1}^{i+2} E_k < E \left(1 + \left| \sum_{k=1}^{i+2} I_k \right| \right); \quad (2)$$

тогда делается вывод, что интеграл по бесконечному лучу сосчитан с заданной точностью E и равен $I^1 = \sum_{k=1}^{i+2} I_k$ (если интеграл надо считать по всей

бесконечной прямой, то по лучу интеграл считается с точностью $E/2$). Если условие (2) не выполнено, то величина ϵ (точность на шаге) делится на 128 и счет по бесконечному лучу начинается сначала, при этом интеграл на каждом отрезке Δ_i считается с более высокой точностью. Вновь полученное значение интеграла обозначим через I^2 . Если

$$\min \left(\sum_{k=1}^{i+2} E_k, |I^2 - I^1| \right) < E (1 + |I^2|)$$

(E_k и i соответствуют второму просчету), то интеграл сосчитан и равен I^2 , в противном случае ϵ снова делится на 128, и весь процесс начинается сначала. Если на луче нет особенности и два последовательных значения I_i и I_{i+1} равны нулю, то счет обрывается без проверки условий (1). Теперь перейдем к описанию метода численного интегрирования на конечном отрезке $\Delta = [a, b]$. Счет по конечному отрезку Δ ведется по квадратурным формулам Лобатто с семью узлами. Заменой переменной $x = a + (b-a)t$ отрезок Δ переводится в единичный отрезок $[0, 1]$. Рассмотрим отрезок $[p, p+h] \in [0, 1]$. На этом отрезке интеграл заменяется соответствующей квадратурной формулой и подсчитывается соответствующий квадратурный член $\rho = |I(t)|$.

Если

$$\rho < \frac{\epsilon}{b-a}, \quad (3)$$

то интеграл по отрезку $[p, p+h]$ считается вычисленным. Пусть $p+h < 1$, тогда перейдем к следующему частичному отрезку

$$[p_1, p_1 + h_1], \quad p_1 = p + h, \quad h_1 = \min(h, 1 - p_1),$$

если $\frac{\epsilon}{128(b-a)} \leq \rho$, и $h_1 = \min(2h, 1 - p_1)$, если

$$\rho < \frac{\epsilon}{128(b-a)}.$$

Если $\frac{\epsilon}{b-a} \leq \rho$, то переходим к вычислению интеграла по отрезку $\left[p, p + \frac{h}{2} \right]$. Может случиться, что условие (3) не выполняется ни при каком $h < H$, где $H = 2^{-36}(p + 2^{-24})$ (величина, близкая к машинному нулю; данный выбор H произведен с учетом разрядности в машине БЭСМ 6), в этом случае предусмотрен уход из процедуры по метке L с предварительной

выдачей информации p и h . Процесс вычислений начинается при $h_{нач} = 1$, $p_{нач} = 0$ и продолжается до тех пор, пока интеграл не будет вычислен на всем отрезке $[0, 1]$. В случае задания особенности $[\alpha, \beta]$ для каждого частичного отрезка $[p, p + h]$ происходит проверка пересечения отрезков $[\alpha, \beta]$ и $[p, p + h]$. Если это пересечение не пусто и очередной шаг $h > \tau$, то h полагается равным τ .

Процедура *INT 120* при $P = 1$ ориентирована на вычисление интегралов от аналитических функций. Процедуры *INT 110* и *INT 111* несколько уступают ей в случае очень гладких функций и существенно превосходят ее для негладких функций.

Отметим некоторые особенности последних двух процедур. Тела процедур *INT 110* и *INT 111* отличаются несущественно. Если положить $\varepsilon = E$ и $\alpha > \beta$, то эти процедуры работают одинаково. Наличие упрощенного варианта *INT 110* имеет целью облегчить обращение в случае малых серий задач. Наличие усложненного варианта *INT 111* дает возможность существенно уменьшить объем вычислений в случае больших серий интегралов и вычислять интегралы от функций с пикообразными особенностями.

Возможность уменьшения объема вычислений в случае больших серий интегралов реализуется следующим образом. Выбирается типичный (обычно самый неблагоприятный) интеграл серии. В процедуре *INT 111* задаются $SI = 1$, E — равным или меньшим требуемой точности, $\varepsilon = \varepsilon_0$ — очень большим. В результате вычислений будут получены приближенные значения интеграла, соответствующие просчетам $\varepsilon = \varepsilon_0, \varepsilon_0 2^{-k}, \dots, \varepsilon_0 2^{-kn}$; $k = 7$ в случае интегралов типа (а), $k = 5$ в случае интегралов типа (б). Из рассмотрения этих результатов делается вывод, при каком $\varepsilon = \varepsilon_0 2^{-kl}$ достигается требуемая точность.

Если при этом значении ε получается точность, существенно большая требуемой, то иногда стоит произвести также отладочные просчеты со значениями $\varepsilon = \varepsilon_0 2^{-1}, \dots, \varepsilon_0 2^{-(k-1)}$. Тогда в совокупности будут получены результаты просчетов при $\varepsilon = \varepsilon_0, \varepsilon_0 2^{-1}, \dots$, среди них выбираем максимальное ε , соответствующее требуемой точности.

Далее вычисляем всю серию интегралов при полученном значении ε и очень большом E , например, $E = 10^{10}$; большое значение E гарантирует, что просчет по всему отрезку будет производиться только один раз.

Многократные интегралы можно вычислять с помощью указанных выше процедур вычисления однократных интегралов, если транслятор допускает рекурсивные процедуры. Например, интеграл

$$I = \int_0^1 \left(\int_0^{\sqrt{1-x^2}} \exp(10(x+y)) dy \right) dx$$

записываем в виде

$$I = \int_0^1 I_1(x) dx,$$

где

$$I_1(x) = \int_0^{\sqrt{1-x^2}} \exp(10(x+y)) dy,$$

и оба интегрирования производим численно. Подынтегральная функция интеграла $I_1(x)$ — гладкая, и этот интеграл удобно вычислять с помощью любой из указанных процедур. Функция $I_1(x)$ имеет неограниченную производную в точке $x = 1$, и поэтому при вычислении интеграла $\int_0^1 I_1(x) dx$ нецелесообразно применять процедуру *INT 120*, ориентированную на вычисление интегралов от гладких функций.

Отметим одно обстоятельство. На первый взгляд казалось бы естественным вычислять интегралы по элементарным отрезкам по формулам Гаусса, более точным, чем формулы Лобатто при том же числе узлов, формулы Гаусса и были заложены в основу первых стандартных программ интегрирования с автоматическим выбором шага. Однако практика интегрирования показала следующий их недостаток в случае резко меняющихся функций. Формулы Гаусса не содержат среди узлов концов отрезков $[p, p+h]$; если

вычислять интеграл $\int_0^1 \exp(-Ax) dx$ при большом A , то все узлы формулы

Гаусса при применении ее ко всему отрезку $[0, 1]$ могут оказаться в области, где подынтегральная функция ничтожно мала; программа прекратит интегрирование, выдав приближенное значение, равное 0. Для формул, содержащих конечные точки, также возможна аналогичная ситуация, например в случае

интеграла $\int_0^1 x \exp(-Ax) dx$ при большом A , однако она встречается значительно реже.

Если формула интегрирования по элементарному отрезку содержит конечную точку, то значение функции в этой точке может использоваться для вычисления интегралов по обоим отрезкам, которым она принадлежит. Поэтому, в случае формулы Лобатто с $n+1$ узлом и формулы Гаусса с n узлами на элементарный отрезок приходится одинаковый объем вычислений; в то же время формулы точны для многочленов одинаковой степени $2n-1$. Второе преимущество формулы Лобатто заключается в следующем. По $(n+1)$ -му узлу этой формулы мы можем получить выражение разложенной разности n -го порядка, т. е. получить оценку для производной $f^{(n)}$; в случае формулы Гаусса можно получить оценку лишь для производной $f^{(n-1)}$. В результате этого разное распределение узлов формулы Лобатто оказывается более близким к оптимальному.

4. Процедура *INT 21*. Предназначена для вычисления интегралов по S -мерному прямоугольному параллелепипеду

$$\int_{\Omega} f(X) dX, \quad \Omega = [a_1 \leq x_1 \leq b_1, \dots, a_S \leq x_S \leq b_S],$$

$$X = (x_1, \dots, x_S).$$

Обращение к процедуре имеет вид

$$Z := \text{INT } 21(S, A, F, X, EPS, P).$$

Формальные параметры процедуры: *INTEGER S* — кратность интеграла; *ARRAY A[1:2S]* — массив, задающий область Ω : $A[1] = a_1$, $A[2] = b_1$, ..., $A[2S-1] = a_S$, $A[2S] = b_S$; *REAL PROCEDURE F* — процедура-функция без параметров, задающая подынтегральную функцию f ; *ARRAY X[1:S]* — аргумент функции f ; *REAL EPS* — величина допустимой абсолютной погрешности вычисления интеграла на шаге; *REAL Z* — вычисленное значение интеграла; *LABEL P* — метка, на которую произойдет переход из процедуры *INT 21* в случае, когда заданная точность *EPS* не может быть достигнута.

Описание метода. Заменой переменных

$$x_i = a_i + (b_i - a_i)t_i, \quad i = 1, \dots, S,$$

данный S -мерный прямоугольный параллелепипед переводится в единичный куб

$$G = [0 \leq t_1 \leq 1, \dots, 0 \leq t_S \leq 1], \quad t = (t_1, \dots, t_S),$$

и интеграл вычисляется по этому кубу. Обозначим через P_S якобиан при такой замене: $P_S = (b_1 - a_1) \dots (b_S - a_S)$. В основу метода положена пара кубатурных формул

$$\int_{c_1-h}^{c_1+h} \dots \int_{c_S-h}^{c_S+h} g(t) dt \approx h^S \Sigma_q^S = Q_q^S, \quad q = 1, 2.$$

Для вычисления интеграла $\int_0^1 \dots \int_0^1 g(t) dt$ вычисляются Q_1^S и Q_2^S , соответствующие кубу G , и проверяется условие

$$P_S |Q_1^S - Q_2^S| \leq \varepsilon. \quad (4)$$

Если это условие выполнено, то за приближенное значение интеграла по G принимается значение, сосчитанное по кубатурной

формуле Q_3^S , являющейся линейной комбинацией формул Q_1^S и Q_2^S . Если условие (4) не выполняется, то куб G разбивается на 2^s равных кубов и описанный алгоритм применяется к каждому из этих кубов. Процесс дробления продолжается до тех пор, пока условие (4) не будет выполнено. Если при делении шага h пополам наступит такой момент, когда h^S станет машинным нулем, то счет прекращается. Кубатурные формулы Σ_q^S для различных S имеют следующий вид:

1) $S=2$:

$$\begin{aligned}\Sigma_1^2 &= A_1 \sum_{|i|+|j|=1} g(c_1 + iah, c_2 + jah) + \\ &+ A_2 \sum_{|i|+|j|=1} g(c_1 + i\beta h, c_2 + j\beta h) + A_3 \sum_{|i|+|j|=1} g(c_1 + i\gamma h, c_2 + j\gamma h), \\ \Sigma_2^2 &= B_1 \sum_{|i|+|j|=1} g(c_1 + iah, c_2 + jah) + \\ &+ B_2 \sum_{|i|+|j|=1} g(c_1 + i\beta h, c_2 + j\beta h) + B_3 \sum_{|i|+|j|=1} g(c_1 + i\gamma h, c_2 + j\gamma h) + \\ &+ B_4 \sum_{|i|+|j|=1} g(c_1 + i\nu h, c_2 + j\nu h),\end{aligned}$$

где

$$\begin{aligned}\alpha &= 0,658149897623035910, & A_3 &= 0,17361111111111111, \\ \beta &= 0,549119831921783496, & B_1 &= 3,99942795838189963, \\ \gamma &= 0,894427190999915878, & B_2 &= -6,29803906949301074, \\ \nu &= 0,316227766016837933, & B_3 &= 0,124007936507936507, \\ A_1 &= 1,06136206790541224, & B_4 &= 3,17460317460317460, \\ A_2 &= -0,234973179016523356,\end{aligned}$$

$$Q_3^2 = Q_2^2.$$

Формула Q_1^2 точна для всех многочленов степени ≤ 5 . Формула Q_2^2 точна для всех многочленов степени ≤ 7 .

2) $S=3$:

$$\begin{aligned}\Sigma_1^3 &= \frac{8}{225} \left(44g(c_1, c_2, c_3) + \right. \\ &+ \frac{121}{8} \sum_{|i|+|j|+|k|=1} g\left(c_1 + i\sqrt{\frac{5}{11}}h, c_2 + j\sqrt{\frac{5}{11}}h, c_3 + k\sqrt{\frac{5}{11}}h\right) + \\ &\left. + 10 \sum_{|i|+|j|+|k|=1} g(c_1 + ih, c_2 + jh, c_3 + kh) \right),\end{aligned}$$

$$\begin{aligned}
\Sigma_2^3 = & \frac{8}{1125} \left(-\frac{1552}{5} g(c_1, c_2, c_3) + \right. \\
& + \frac{1573}{40} \sum_{|i|, |j|, |k|=1} g\left(c_1 + i\sqrt{\frac{5}{11}}h, c_2 + j\sqrt{\frac{5}{11}}h, c_3 + k\sqrt{\frac{5}{11}}h\right) + \\
& + \frac{784}{5} \sum_{|i|+|j|+|k|=1} g\left(c_1 + i\sqrt{\frac{5}{14}}h, c_2 + j\sqrt{\frac{5}{14}}h, c_3 + k\sqrt{\frac{5}{14}}h\right) + 15T_h \Big), \\
T_h = & \sum_{|i|, |j|=1} g(c_1 + ih, c_2 + jh, c_3) + \\
& + \sum_{|i|, |k|=1} g(c_1 + ih, c_2, c_3 + kh) + \sum_{|j|, |k|=1} g(c_1, c_2 + jh, c_3 + kh), \\
Q_3^3 = & \frac{4}{9} Q_1^3 + \frac{5}{9} Q_2^3.
\end{aligned}$$

Формулы Q_1^3 и Q_2^3 точны для всех многочленов степени ≤ 5 .
 Формула Q_3^3 точна для всех многочленов степени ≤ 7 .

Дополнительная литература к главе III

I (8), II [21]; § 2: I [8, 11, 12], II [21, 33]; § 3: I [8], II [21, 47], III [5];
 § 5: I [17], III [27]; § 8: II [33]; § 9, 10: II [3], III [9, 11, 49]; § 12: I [1, 11];
 § 14: III [53]; §§ 17, 18: II [3, 36], III [9, 11].

ПРИБЛИЖЕНИЕ ФУНКЦИЙ И СМЕЖНЫЕ ВОПРОСЫ

§ 1. Наилучшие приближения в линейном нормированном пространстве

Рассмотрим следующую задачу. Имеется элемент f линейного нормированного пространства $\mathcal{R}$. Требуется найти его наилучшее приближение линейной комбинацией $\sum_{i=1}^n c_i g_i$ данных линейно независимых элементов $g_1, \dots, g_n \in \mathcal{R}$. Это означает: найти элемент $\sum_{i=1}^n c_i^0 g_i$ такой, что

$$\left\| f - \sum_{i=1}^n c_i^0 g_i \right\| = \Delta = \inf_{c_1, \dots, c_n} \left\| f - \sum_{i=1}^n c_i g_i \right\|.$$

Если такой элемент существует, то он называется *элементом наилучшего приближения*.

Теорема 1. *Элемент наилучшего приближения существует.*

Доказательство. Вследствие соотношений

$$\left\| f - \sum_{i=1}^n c_i^1 g_i \right\| - \left\| f - \sum_{i=1}^n c_i^2 g_i \right\| \leq \left\| \sum_{i=1}^n (c_i^1 - c_i^2) g_i \right\| \leq \sum_{i=1}^n |c_i^1 - c_i^2| \|g_i\|$$

функция

$$F_f(c_1, \dots, c_n) = \left\| f - \sum_{i=1}^n c_i g_i \right\|$$

является непрерывной функцией аргументов c_i при любом $f \in \mathcal{R}$. Пусть $\|c\|$ — евклидова норма вектора $c = (c_1, \dots, c_n)$. Функция

$$F_0(c_1, \dots, c_n) = \|c_1 g_1 + \dots + c_n g_n\|$$

непрерывна на единичной сфере $\|c\| = 1$ и, следовательно, в некоторой ее точке $(\tilde{c}_1, \dots, \tilde{c}_n)$ достигает своей нижней грани $\tilde{F}$ по сфере. $\tilde{F} \neq 0$, так как равенство $\tilde{F} = \|\tilde{c}_1 g_1 + \dots + \tilde{c}_n g_n\| = 0$

противоречит линейной независимости элементов $g_1, \dots, g_n$. Для любого $c = (c_1, \dots, c_n) \neq (0, \dots, 0)$ справедлива оценка

$$\|c_1 g_1 + \dots + c_n g_n\| = F_0(c_1, \dots, c_n) = \|c\| F_0\left(\frac{c_1}{\|c\|}, \dots, \frac{c_n}{\|c\|}\right) \geq \|c\| \tilde{F}.$$

Пусть $\gamma > \frac{2\|f\|}{\tilde{F}}$. Функция $F_f(c_1, \dots, c_n)$ непрерывна в шаре $\|c\| \leq \gamma$ и, следовательно, в некоторой точке шара $(c_1^0, \dots, c_n^0)$ достигает своей нижней грани по шару F^0 . Имеем $F^0 \leq F_f(0, \dots, 0) = \|f\|$. Вне этого шара

$$F_f(c_1, \dots, c_n) \geq \|c_1 g_1 + \dots + c_n g_n\| - \|f\| > \tilde{F} \frac{2\|f\|}{\tilde{F}} - \|f\| = \|f\| \geq F^0.$$

Таким образом,

$$F_f(c_1, \dots, c_n) \geq F^0 = F_f(c_1^0, \dots, c_n^0)$$

при всевозможных $c_1, \dots, c_n$. Теорема доказана.

Элементов наилучшего приближения, вообще говоря, может быть несколько.

Теорема 2. Если пространство $\mathcal{H}$ строго нормировано, то элемент наилучшего приближения единствен.

Доказательство. Предположим противное, т. е. что существуют два элемента

$$f_1 \neq f_2, \quad f_1 = \sum_{i=1}^n a_{i1} g_i$$

такие, что $\|f - f_1\| = \|f - f_2\| = \Delta$.

Очевидно, $\Delta \neq 0$, поскольку иначе $f_1 = f = f_2$. Кроме того,

$$\left\|f - \frac{f_1 + f_2}{2}\right\| = \left\|\frac{f - f_1}{2} + \frac{f - f_2}{2}\right\| \leq \frac{\|f - f_1\|}{2} + \frac{\|f - f_2\|}{2} = \Delta.$$

Поскольку $\frac{f_1 + f_2}{2}$ является линейной комбинацией элементов $g_1, \dots, g_n$, то $\left\|f - \frac{f_1 + f_2}{2}\right\| \geq \Delta$ вследствие определения Δ . Вместе с предшествующими соотношениями это означает, что

$$\left\|\frac{f - f_1}{2}\right\| + \left\|\frac{f - f_2}{2}\right\| = \left\|\frac{f - f_1}{2} + \frac{f - f_2}{2}\right\|.$$

Вследствие предположения строгой нормированности пространства $\frac{f - f_1}{2} = \alpha \frac{f - f_2}{2}$. Если $\alpha \neq 1$, то $f = \frac{f_1 - \alpha f_2}{1 - \alpha}$, f является линейной комбинацией $g_1, \dots, g_n$ и поэтому $\Delta = 0$. При $\alpha = 1$ имеем $f_1 = f_2$ и приходим к противоречию с допущением $f_1 \neq f_2$.

- Задачи. 1. Доказать, что гильбертово пространство строго нормировано.
2. Доказать, что пространство $L_p(0, 1)$ с нормой

$$\|f\|_{L_p} = \sqrt[p]{\int_0^1 |f(x)|^p dx}$$

строго нормировано при $1 < p < \infty$.

3. Доказать, что пространство L_1 и пространство непрерывных функций $C([0, 1])$ с нормой

$$\|f\|_C = \max_{[0, 1]} |f(x)|$$

не являются строго нормированными.

§ 2. Наилучшее приближение в гильбертовом пространстве и вопросы, возникающие при его практическом построении

Для гильбертова пространства задача нахождения элемента наилучшего приближения формально сводится к решению системы линейных уравнений.

Наиболее простой способ получения этой системы следующий. Коэффициенты a_i элемента наилучшего приближения реализуют минимум выражения

$$\delta^2 = \left\| f - \sum_{i=1}^n a_i g_i \right\|^2 = \left(f - \sum_{i=1}^n a_i g_i, f - \sum_{i=1}^n a_i g_i \right).$$

Приравнявая нулю производные по $\operatorname{Re} a_i$ и $\operatorname{Im} a_i$, получаем искомую систему уравнений для определения a_i . Вследствие существования элемента наилучшего приближения эта система имеет решение.

Построим эту систему и исследуем вопрос об единственности ее решения несколько иным способом.

Пусть H — линейное подпространство гильбертова пространства $\mathcal{H}$, $f \in \mathcal{H}$. Поставим задачу найти элемент $h_0 \in H$, наилучшим образом приближающий f , т. е. элемент h_0 , для которого

$$\|f - h_0\| = \inf_{h \in H} \|f - h\|.$$

Лемма 1. Пусть в H существует элемент h_0 наилучшего приближения к f . Тогда разность $f - h_0$ ортогональна ко всем элементам подпространства H .

Доказательство. Предположим, существует элемент $h_1 \in H$, для которого $(f - h_0, h_1) = \alpha \neq 0$. Поскольку вместо h_1 можно взять $h_1/\|h_1\|$, то предполагаем $\|h_1\| = 1$. Рассмотрим элемент $h_2 = h_0 + \alpha h_1 \in H$. Имеем

$$\begin{aligned} \|f - h_2\|^2 &= (f - h_0 - \alpha h_1, f - h_0 - \alpha h_1) = \\ &= (f - h_0, f - h_0) - \alpha (h_1, f - h_0) - \bar{\alpha} (f - h_0, h_1) + \alpha \bar{\alpha} (h_1, h_1), \end{aligned}$$

но $(h_1, f - h_0) = \overline{(f - h_0, h_1)} = \bar{a}$. Поэтому отсюда получаем

$$\|f - h_2\|^2 = \|f - h_0\|^2 - \alpha\bar{a} - \bar{a}\alpha + \alpha\bar{a} = \|f - h_0\|^2 - \alpha\bar{a} < \|f - h_0\|^2.$$

Мы пришли к противоречию с предположением, что h_0 — элемент наилучшего приближения.

Лемма 2. Если $(f - h_0, h) = 0$ для любого $h \in H$, то h_0 — элемент наилучшего приближения.

Доказательство. Для произвольного $h \in H$ имеем

$$\begin{aligned} \|f - h\|^2 &= (f - h_0 + h_0 - h, f - h_0 + h_0 - h) = \\ &= (f - h_0, f - h_0) + (h_0 - h, f - h_0) + \\ &\quad + (f - h_0, h_0 - h) + (h_0 - h, h_0 - h), \end{aligned}$$

так как $h_0 - h \in H$, то второе и третье слагаемые обращаются в нуль, согласно предположению леммы. В итоге имеем

$$\|f - h\|^2 = \|f - h_0\|^2 + \|h - h_0\|^2, \quad (1)$$

следовательно, $\|f - h\|^2 > \|f - h_0\|^2$ при $h \neq h_0$.

Так как гильбертово пространство строго нормировано, единственность элемента наилучшего приближения вида $a_1 g_1 + \dots + a_n g_n$ следует из теоремы 2 § 1. Другое доказательство следует из лемм 1, 2: если h_0 — элемент наилучшего приближения, то $(f - h_0, h) = 0$ при $h \in H$, тогда, согласно (1), любой элемент $h \neq h_0$, $h \in H$ не является элементом наилучшего приближения.

Вернемся к задаче наилучшего приближения f комбинациями $\sum_{i=1}^n a_i g_i$ линейно независимых элементов $g_1, \dots, g_n$. Эти комбинации образуют линейное подпространство H . Если элемент $\sum_{i=1}^n a_i g_i$ осуществляет наилучшее приближение, то, согласно лемме 1, выполняются соотношения

$$\left(f - \sum_{i=1}^n a_i g_i, g_j\right) = 0 \quad \text{при } j = 1, \dots, n. \quad (2)$$

Поскольку элемент наилучшего приближения существует, то отсюда следует существование решения системы (2). С другой стороны, если соотношения (2) выполнены, то

$$\left(f - \sum_{i=1}^n a_i g_i, \sum_{j=1}^n c_j g_j\right) = 0$$

при любых c_j , иначе $\left(f - \sum_{i=1}^n a_i g_i, h\right) = 0$ при любом $h \in H$; согласно

лемме 2, такой элемент $\sum_{i=1}^n a_i g_i$ является элементом наилучшего приближения. Предположим, что система (2) имеет решение $(b_1, \dots, b_n)$, отличное от $(a_1, \dots, a_n)$. Тогда элемент $g = \sum_{i=1}^n b_i g_i$ также является элементом наилучшего приближения и, вследствие его единственности,

$$\sum_{i=1}^n a_i g_i = \sum_{i=1}^n b_i g_i;$$

предположение, что $a_i \neq b_i$ при некотором i , противоречит линейной независимости элементов g_i .

Таким образом, мы показали, что система (2) всегда имеет и притом единственное решение. Ее можно переписать в виде

$$\sum_{i=1}^n a_i (g_i, g_j) = (f, g_j), \quad j = 1, \dots, n; \quad (3)$$

Итак, коэффициенты, соответствующие элементу наилучшего приближения, можно найти, решая невырожденную систему (3).

Матрица

$$G(g_1, \dots, g_n) = [(g_i, g_j)]$$

называется *матрицей Грамма* системы элементов.

При исследовании свойств матриц Грамма полезно использовать формальное равенство

$$G(g_1, \dots, g_n) = \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix} (g_1, \dots, g_n).$$

Задача. Доказать, что $G(g_1, \dots, g_n) \geq 0$, причем $G(g_1, \dots, g_n) = 0$ тогда и только тогда, когда $g_1, \dots, g_n$ линейно зависимы. Доказать неравенство

$$G(g_1, \dots, g_{n+1}) \leq G(g_1, \dots, g_n) (g_{n+1}, g_{n+1}).$$

Пусть g — наилучшее приближение к f . Тогда, согласно лемме 1, имеем $(f - g, h) = 0$ при $h \in H$; подставляя в (1) $h_0 = g$ и $h = 0$, получаем

$$\|f\|^2 = \|f - g\|^2 + \|g\|^2.$$

Если элементы g_i образуют ортонормированную систему $(g_i, g_j) = \delta_{i-j}$, то система (3) приобретает вид

$$a_j = (f, g_j). \quad (4)$$

Тогда наилучшее приближение записывается в форме

$$g = \sum_{i=1}^n (f, g_i) g_i,$$

и имеется следующее удобное представление величины $\|f - g\|^2$:

$$\begin{aligned} \|f - g\|^2 &= \|f\|^2 - \|g\|^2 = (f, f) - \left(\sum_{i=1}^n a_i g_i, \sum_{i=1}^n a_i g_i \right) = \\ &= (f, f) - \sum_{i,j=1}^n a_i \bar{a}_j \delta_{i-j} = (f, f) - \sum_{i=1}^n |a_i|^2 = (f, f) - \sum_{i=1}^n |(f, g_i)|^2. \end{aligned}$$

Поскольку $\|f - g\|^2 \geq 0$, то из равенства

$$\|f - g\|^2 = (f, f) - \sum_{i=1}^n |(f, g_i)|^2$$

следует, в частности, известное неравенство Бесселя

$$(f, f) \geq \sum_{i=1}^n |(f, g_i)|^2.$$

Если исходные элементы не образуют ортонормированной системы, то их можно ортогонализировать при помощи рассматривавшегося в главе II алгоритма ортогонализации.

Задача вычисления совокупности коэффициентов $a_j = (f, g_j)$, $j = 1, \dots, M$, при относительно большом M часто возникает как вспомогательная при решении более сложных задач.

В качестве модели изучения способов вычисления a_j рассмотрим задачу вычисления интегралов

$$a_j = (f, Q_j) = \int_a^b f(x) Q_j(x) p(x) dx, \quad p(x) > 0,$$

где действительные многочлены $Q_j(x)$ образуют ортонормированную систему в пространстве со скалярным произведением

$$(f, g) = \int_a^b f(x) g(x) p(x) dx.$$

Интегралы a_j можно вычислить приближенно, по формуле Гаусса (§ 3, гл. II), соответствующей данной весовой функции

$$a_j = \int_a^b f(x) Q_j(x) p(x) dx \approx \sum_{q=1}^N D_q^N f(x_q^N) Q_j(x_q^N). \quad (5)$$

Если f — многочлен степени l , то $f(x) Q_j(x)$ — многочлен степени $l + j$. Поскольку формула Гаусса точна для многочленов степени $2N - 1$, то формула (5) точна, если $f(x)$ — многочлен степени $2N - j - 1$.

При решении одной задачи требовалось вычислить несколько тысяч интегралов типа

$$a_j = \int_a^b f(x) Q_j(x) p(x) dx, \quad j = 0, \dots, M.$$

Попытки применения программ интегрирования с автоматическим выбором шага или формул (5) не привели к успеху. Анализ задачи показал, что ее решение при независимом вычислении интегралов a_j требует неприемлемых затрат времени ЭВМ.

Узким местом задачи в данном случае была высокая трудоемкость вычисления значений функции $f(x)$. Поэтому был принят следующий путь. Интегралы a_j разбивались на группы

$$(a_0, \dots, a_{n_1}), (a_{n_1+1}, \dots, a_{n_2}), \dots, (a_{n_{k-1}+1}, \dots, a_{n_k}), \quad n_k = M.$$

Для каждой группы определялось число N_i , достаточное, чтобы интегралы этой группы вычислялись по формуле (5) с требуемой точностью. Затем для интегралов каждой группы одновременно вычислялись приближения

$$a_j^{(N_i)} = \sum_{q=1}^{N_i} D_q^{N_i} f(x_q^{N_i}) Q_j(x_q^{N_i}).$$

Каждое значение $f(x_q^{N_i})$ вычислялось только один раз для элементов всей группы. Оказалось также, что при этом появляется дополнительный выигрыш машинного времени за счет одновременного вычисления значений всех многочленов $Q_j(x_q^{N_i})$ в одной точке при помощи рекуррентных соотношений. Выбор разбиения коэффициентов a_j на группы производился с учетом трудоемкости вычисления значений функции $f(x)$.

В другом подобном случае оказалось наиболее выгодным с позиций экономии машинного времени такое разбиение не производить, а взять N , соответствующее a_M , и с этим значением N произвести одновременное вычисление всех интегралов. Объем вычислений сократился примерно в 100 раз по сравнению с исходным методом.

Вследствие уже упоминавшейся привычки оценивать эффективность метода по доле полезной работы заказчика некоторое время возражали против применения описанной выше процедуры, ссылаясь на то, что при малых j интегралы вычисляются со «слишком высокой» точностью, и поэтому производится много ненужной работы. Исполнители убедили их в высокой эффективности метода при помощи упоминавшегося выше довода о доле полезной работы: при вычислении всей серии интегралов предлагаемым методом объем вычислений лишь на 20% больше, чем в случае вычисления интеграла a_M по формуле (5), следовательно, доля полезной работы около 80%. Довод был признан убедительным, и метод приняли к реализации. Если функция $f(x)$ плохо приближается многочленами, то можно прибегнуть к следующей процедуре: вычислять все интегралы одновременно при помощи алгоритма интегрирования с автоматическим выбором шага, причем узлы интегрирования брать во всех интегралах одними и теми же, такими же, как для «самого плохого» интеграла a_M . Конечно, для реализации этого алгоритма стандартная процедура интегрирования с автоматическим выбором шага уже непригодна, и требуется написание новой программы. Другой резерв — это не вычислять непосредственно всех значений функции $f(x)$, а определять некоторые из них путем интерполирования по уже вычисленным значениям.

Рассмотрим одну задачу другого рода, при решении которой в определенном смысле используются проводимые выше в этом параграфе построения. Требуется построить интерполяционный многочлен степени $N-1$ с узлами интерполяции $x_1^N, \dots, x_N^N$; эти узлы являются нулями многочлена $Q_N(x)$ степени N из ортонормированной системы многочленов $\{Q_n(x)\}$, соответствующей весу $\rho(x)$. Например, в главе II шла речь об интерполяции по нулям многочленов Чебышева. Будем отыскивать интерполяционный многочлен в виде линейной комбинации

$$P_{N-1}(x) = \sum_{j=0}^{N-1} a_j Q_j(x).$$

При $m, n < N$ многочлен $Q_m(x)Q_n(x)$ имеет степень не выше $2N-2$. Поэтому квадратура Гаусса с N узлами точна для этого многочлена

$$\sum_{q=1}^N D_q^N Q_m(x_q^N) Q_n(x_q^N) = \int_a^b Q_m(x) Q_n(x) \rho(x) dx = \delta_{n-m}. \quad (6)$$

Таким образом, векторы

$$Q_m = \{Q_m(x_1^N), \dots, Q_m(x_N^N)\}$$

при $m < N$ образуют ортонормированную систему относительно скалярного произведения

$$(y, z) = \sum_{q=1}^N D_q^N y_q z_q; \quad (7)$$

поэтому вектор $f = (f(x_1^N), \dots, f(x_N^N))$ может быть разложен по этой системе векторов:

$$f = \sum_{j=0}^{N-1} d_j Q_j,$$

где $d_j = (f, Q_j)$.

Многочлен

$$P_{N-1}(x) = \sum_{j=0}^{N-1} d_j Q_j(x)$$

будет искомым.

В практических вычислениях используются чаще всего весовые функции $\rho(x) = \frac{1}{\sqrt{1-x^2}}$ на $[-1, 1]$ и $\rho(x) = 1$. В первом случае ортогональные многочлены и квадратура Гаусса известны в явном виде (многочлены Чебышева, формула Эрмита), во втором известны ортогональные многочлены и имеются достаточно обширные таблицы узлов и весов квадратур Гаусса.

§ 3. Дискретное преобразование Фурье

Пусть $f(x)$ — периодическая функция с периодом 1 — может быть разложена в ряд Фурье

$$f(x) = \sum_{q=-\infty}^{\infty} a_q \exp(2\pi i q x), \quad (1)$$

причем

$$\sum_{q=-\infty}^{\infty} |a_q| < \infty. \quad (2)$$

Рассмотрим значения этой функции на сетке из точек $x_l = l/N$, где l, N — целые, N — фиксировано, и обозначим $f(x_l) = f_l$. Пусть $q_2 - q_1 = kN$, k — целое; тогда $q_2 x_l - q_1 x_l = kN x_l = kl$; kl — целое и, следовательно,

$$\exp(2\pi i q_1 x) = \exp(2\pi i q_2 x) \quad (3)$$

в узлах сетки. Поэтому, если функция $f(x)$ рассматривается лишь на сетке узлов x_l , в соотношении (1) можно привести подобные члены

$$f_l = \sum_{q=0}^{N-1} A_q \exp(2\pi i q x_l), \quad (4)$$

где

$$A_q = \sum_{s=-\infty}^{\infty} a_{q+sN}.$$

Если с самого начала была задана функция, определенная только на сетке, то на этой сетке ее можно также представить в форме (1). Действительно, такую функцию можно продолжить на всю прямую, доопределив ее между узлами сетки путем линейной интерполяции. Для непрерывной кусочно-дифференцируемой функции выполняется (2), и поэтому после приведения на сетке подобных членов получим (4). Определим скалярное произведение для функций на сетке:

$$(f, g) = \frac{1}{N} \sum_{l=0}^{N-1} f_l \bar{g}_l.$$

Множитель $1/N$ введен для согласованности получаемых соотношений с непрерывным случаем: если $f(x)$ и $g(x)$ — непрерывные функции на $[0, 1]$, то

$$(f, g) \rightarrow \int_0^1 f(x) \bar{g}(x) dx$$

при $N \rightarrow \infty$.

Функции $g_q(x_l) = \exp(2\pi i q x_l)$ при $0 \leq q < N$ образуют ортонормированную систему относительно так введенного скалярного произведения. Действительно,

$$(g_q, g_l) = \frac{1}{N} \sum_{l=0}^{N-1} \exp\left(2\pi i \frac{q-l}{N} l\right).$$

При $q \neq l$, суммируя геометрическую прогрессию, имеем

$$(g_q, g_l) = \frac{1}{N} \frac{\exp(2\pi i (q-l)) - 1}{\exp\left(2\pi i \frac{q-l}{N}\right) - 1} = 0.$$

Поскольку $(g_q, g_q) = 1$, то в итоге имеем

$$(g_q, g_l) = \delta_{q-l} \quad \text{при } 0 \leq q, l < N. \quad (5)$$

Умножая (4) скалярно на g_l , получим равенство

$$A_l = (f, g_l) = \frac{1}{N} \sum_{l=0}^{N-1} f_l \exp(-2\pi i j x_l), \quad (6)$$

очевидно,

$$A_l \rightarrow a_l = \int_0^1 f(x) \exp(-2\pi i j x) dx$$

при $N \rightarrow \infty$ и j фиксированном.

Однако это не означает, что

$$f(x) \approx \sum_{j=0}^{N-1} A_j \exp(2\pi i j x).$$

Посмотрим, почему это соотношение неверно. Пусть для простоты

$$f(x) = \sum_{-k}^k a_j \exp(2\pi i j x),$$

где $2k < N$. Из (4) получаем представление для коэффициентов A_j :

$$A_j = \begin{cases} a_j & \text{при } 0 \leq j \leq k, \\ 0 & \text{при } k < j < N - k, \\ a_{j-N} & \text{при } N - k \leq j < N. \end{cases}$$

Таким образом, правая часть этого неверного приближенного

равенства есть

$$\sum_{j=0}^k a_j \exp(2\pi i j x) + \exp(2\pi i N x) \sum_{j=-k}^{-1} a_j \exp(2\pi i j x).$$

Она совпадает с $f(x)$ в точках x_l , но далека от нее вне этих точек.

Заменой переменной суммирования $s' = s + k$ в определении A_q убеждаемся, что $A_{q_1} = A_{q_2}$, если $q_1 - q_2 = kN$, где k — целое. Воспользовавшись равенством (3), перепишем (4) в виде

$$f_l = \sum_{-N/2 < q \leq N/2} A_q \exp(2\pi i q x_l). \quad (7)$$

Если $f(x)$ — достаточно гладкая, то величины $|a_j|$ с ростом j убывают быстро, и поэтому $A_q \approx a_q$ при малых q ; кроме того, тогда величины A_q и a_q малы при больших q . В итоге есть основание надеяться на справедливость приближенного равенства

$$f(x) \approx \sum_{-N/2 < q \leq N/2} A_q \exp(2\pi i q x) \quad (8)$$

во всех точках x .

Напомним, что это приближенное равенство обращается в равенство в точках сетки. Способ аппроксимации функции правой частью (8) носит название *тригонометрической интерполяции*. Соотношение (7) называют *конечным* или *дискретным рядом Фурье*, а коэффициенты A_q — *дискретными коэффициентами Фурье*.

Игнорирование установленного нами факта о равенстве функций

$$\exp(2\pi i q_1 x) \quad \text{и} \quad \exp(2\pi i q_2 x)$$

в узлах сетки при $q_1 - q_2 = kN$ часто является источником получения неверных соотношений.

При решении одной инженерной задачи потребовалось определение первой собственной частоты колебаний конструкции. Было принято решение написать нестационарное уравнение, описывающее процесс колебаний, вывести на печать график и из рассмотрения графика определить частоту. Соответствующее уравнение, которое мы будем условно обозначать $x'' = F(x)$, решалось методом конечных разностей. Для контроля надежности результата производился повторный расчет с вдвое более мелким шагом. Графики кривых, полученных в результате расчетов, совпадали с точностью 10%. Однако из сравнения с экспериментом оказалось, что полученная частота отличается от истинной в десятки раз. Причина недоразумения заключалась в том, что график решения строился с шагом $1/N$, существенно большим периода колебаний решения задачи. Решение было близко к функции $\cos t \cdot \exp(2\pi i q t)$, где q/N близко к четному числу $2k$. Поэтому как на сетке с шагом $1/N$, так

и на вдвое более мелкой с шагом $1/(2N)$ получался график одной и той же функции $\text{const} \cdot \exp(2\pi i(q - 2kN)t)$.

В другом случае несоответствие со здравым смыслом возникло при расчете диаграммы направленности антенны. Предпринимавшиеся попытки найти ошибку в программе, методе решения или физическом описании задачи не приводили к положительному результату. Объяснение оказалось тем же: график сильно колеблющейся функции выдавался на очень редкой сетке. На рис. 4.3.1 сплошной кривой изображен реальный график сечения диаграммы, пунктиром — график, который строился путем интерполяции полученных расчетных значений x и противоречил эксперименту.

Существует соответствие между задачей приближения функций линейными комбинациями многочленов Чебышева и тригонометрическими многочленами. Пусть на отрезке $[-1, 1]$ функция $f(x)$ приближается

линейными комбинациями $\sum_{j=0}^{m-1} a_j T_j(x)$. Замена пе-

ременных $x = \cos t$ сводит исходную задачу к задаче приближения функции $f(\cos t)$ линейной комбинацией

$$\sum_{j=0}^{m-1} a_j T_j(\cos t) = \sum_{j=0}^{m-1} a_j \cos(jt).$$

Справедливо равенство

$$(f, g)_1 = \int_{-1}^1 \frac{f(x) \bar{g}(x)}{\sqrt{1-x^2}} dx = \int_0^\pi f(\cos \theta) \bar{g}(\cos \theta) d\theta = (f, g)_2.$$

Следовательно, задача наилучшего приближения $f(x)$ в норме, соответствующей скалярному произведению $(f, g)_1$, эквивалентна задаче приближения $f(\cos \theta)$ в норме, соответствующей скалярному произведению $(f, g)_2$. Точно так же существует соответствие в случае задач интерполяции и наилучшего приближения в равномерной метрике. Задача интерполирования функции многочленом по узлам $x_j = \cos\left(\pi \frac{2j-1}{2m}\right)$ — нулям многочлена Чебышева $T_m(x)$ — после такой замены сводится к задаче интерполирования функции $f(\cos \theta)$ при помощи тригонометрического многочлена $\sum_{j=0}^{m-1} a_j \cos(jt)$ по узлам $t_j = \pi \frac{2j-1}{2N}$, образующим равномерную сетку.

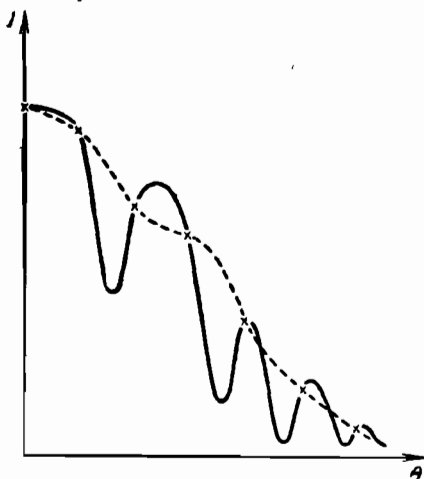


Рис. 4.3.1.

§ 4. Быстрое преобразование Фурье

Осуществление прямого и обратного дискретных преобразований Фурье

$$(f_0, \dots, f_{N-1}) \Leftrightarrow (A_0, \dots, A_{N-1})$$

является составной частью решения многих задач. Непосредственное осуществление этих преобразований по формулам (3.4), (3.6) требует $O(N^2)$ арифметических операций. Рассмотрим вопрос о возможности сокращения этого числа операций; для определенности речь пойдет о вычислении коэффициентов A_q по заданным значениям функции. Идея построения алгоритмов *быстрого преобразования Фурье* опирается на то, что при N составном в слагаемых правой части (3.6) можно выделить группы, которые входят в выражения различных коэффициентов A_q .

Рассмотрим сначала случай $N = p_1 p_2$, $p_1, p_2 \neq 1$. Представим q, j , лежащие в пределах $0 \leq q, j < N$ в виде $q = q_1 + p_1 q_2$, $j = j_2 + p_2 j_1$, где $0 \leq q_1, j_1 < p_1$, $0 \leq q_2, j_2 < p_2$. Имеем цепочку соотношений

$$\begin{aligned} A_q &= A(q_1, q_2) = \frac{1}{N} \sum_{j=0}^{N-1} f_j \exp\left(-2\pi i \frac{qj}{N}\right) = \\ &= \frac{1}{N} \sum_{j_1=0}^{p_1-1} \sum_{j_2=0}^{p_2-1} f_{j_2+p_2 j_1} \exp\left(-2\pi i \frac{(q_1 + p_1 q_2)(j_2 + p_2 j_1)}{p_1 p_2}\right). \end{aligned}$$

Из равенства

$$\frac{(q_1 + p_1 q_2)(j_2 + p_2 j_1)}{p_1 p_2} = q_2 j_1 + \frac{q j_2}{N} + \frac{j_1 q_1}{p_1}$$

и предыдущего соотношения получим

$$A(q_1, q_2) = \frac{1}{p_2} \sum_{j_2=0}^{p_2-1} A^{(1)}(q_1, j_2) \exp\left(-2\pi i \frac{q j_2}{N}\right),$$

где

$$A^{(1)}(q_1, j_2) = \frac{1}{p_1} \sum_{j_1=0}^{p_1-1} f_{j_2+p_2 j_1} \exp\left(-2\pi i \frac{j_1 q_1}{p_1}\right).$$

Непосредственное вычисление всех $A^{(1)}(q_1, j_2)$ требует $O(p_1^2 p_2)$ арифметических операций, а последующее вычисление $A(q_1, q_2)$ еще $O(p_1 p_2^2)$ операций. Следовательно, при $p_1, p_2 = O(\sqrt{N})$ общее число операций будет $O(N^{3/2})$. Точно так же при $N = p_1 \dots p_r$ строится алгоритм вычисления совокупности значений A_q , для которого общее число операций не превосходит $CN(p_1 + \dots + p_r)$, здесь C — абсолютная постоянная. Выпишем соответствующие расчетные формулы для наиболее употреби-

тельного случая $p_1 = \dots = p_r = 2$. Представим числа q, j в виде

$$q = \sum_{k=1}^r q_k 2^{k-1}, \quad j = \sum_{m=1}^r j_{r+1-m} 2^{m-1},$$

где $q_k, j_m = 0, 1$. Величину $qj2^{-r}$ можно представить в виде

$$\begin{aligned} qj2^{-r} &= \sum_{m=1}^r qj_{r+1-m} 2^{m-1-r} = \sum_{m=1}^r \left(\sum_{k=1}^r q_k 2^{k+m-r-2} \right) j_{r+1-m} = \\ &= \sum_{m=1}^r \left(\sum_{k=1}^{r-m+1} q_k 2^{k+m-r-2} \right) j_{r+1-m} + s, \end{aligned}$$

где s — целое, равное сумме всех слагаемых

$$q_k j_{r+1-m} 2^{k+m-r-2},$$

у которых $k + m - r - 2 \geq 0$. Очевидно,

$$\exp\left(-2\pi i \frac{qj}{N}\right) = \exp\left(-2\pi i \left(\frac{qj}{N} - s\right)\right).$$

Поэтому

$$\begin{aligned} A(q_1, \dots, q_r) &= A_q = \frac{1}{N} \sum_{j=0}^{N-1} f_j \exp\left(-2\pi i \frac{jq}{N}\right) = \\ &= \frac{1}{2} \sum_{j_r=0}^1 \dots \frac{1}{2} \sum_{j_1=0}^1 f_{j_r + 2j_{r-1} + \dots + 2^{r-1}j_1} \times \\ &\quad \times \exp\left(-2\pi i \sum_{m=1}^r \left(\sum_{k=1}^{r-m+1} q_k 2^{k+m-r-2} \right) j_{r+1-m}\right). \end{aligned}$$

После перегруппировки слагаемых имеем

$$\begin{aligned} A(q_1, \dots, q_r) &= \frac{1}{2} \sum_{j_r=0}^1 \exp\left(-2\pi i j_r 2^{-r} \sum_{k=1}^r q_k 2^{k-1}\right) \times \\ &\quad \times \left(\frac{1}{2} \sum_{j_{r-1}=0}^1 \exp\left(-2\pi i j_{r-1} 2^{1-r} \sum_{k=1}^{r-1} q_k 2^{k-1}\right) \times \dots \right. \\ &\quad \left. \dots \times \left(\frac{1}{2} \sum_{j_1=0}^1 \exp\left(-2\pi i j_1 2^{-1} q_1\right) f_{j_r + 2j_{r-1} + \dots + 2^{r-1}j_1} \right) \dots \right). \end{aligned}$$

Это соотношение можно записать в виде последовательности рекуррентных соотношений

$$\begin{aligned} A^{(m)}(q_1, \dots, q_m; j_{m+1}, \dots, j_r) &= \\ &= \frac{1}{2} \sum_{j_m=0}^1 \exp\left(-2\pi i j_m 2^{-m} \sum_{k=1}^m q_k 2^{k-1}\right) \times \\ &\quad \times A^{(m-1)}(q_1, \dots, q_{m-1}; j_m, \dots, j_r), \quad m = 1, \dots, r, \end{aligned}$$

где

$$A^{(0)}(j_1, \dots, j_r) = f_{l_r + j_{r-1}2 + \dots + j_1 2^{r-1}},$$

$$A^{(r)}(q_1, \dots, q_r) = A(q_1, \dots, q_r).$$

Переход от каждой совокупности $A^{(m-1)}$ к совокупности $A^{(m)}$ требует $O(N)$ арифметических и логических операций; всего таких шагов r , поэтому общее число операций

$$O(Nr) = O(N \log_2 N).$$

Вычисление при помощи этой совокупности соотношений дает выигрыш в отношении накопления вычислительной погрешности по сравнению с формулами (3.6). Определенные удобства имеются также при вычислении значений экспонент, входящих в расчетные формулы. При вычислении величин $A^{(m)}$ используются значения $\exp(-2\pi i j 2^{-m})$, $j = 0, 1, \dots, 2^m - 1$; в частности, при $m = 1$ это 1 и -1 . Для вычисления значений $A^{(m+1)}$ потребуются еще значения $\exp(-2\pi i j 2^{-m-1})$ при нечетных j , удовлетворяющих неравенству $0 \leq j < 2^{m+1}$; их можно вычислять через уже известные величины, в частности, при помощи соотношений ($m \geq 2$):

$$\exp(-2\pi i j 2^{-m-1}) = \frac{\exp\left(-2\pi i \frac{j+1}{2} 2^{-m}\right) + \exp\left(-2\pi i \frac{j-1}{2} 2^{-m}\right)}{2 \cos(\pi 2^{-m})},$$

где в свою очередь

$$\cos(\pi 2^{-m}) = \sqrt{\frac{1 + \cos(\pi 2^{1-m})}{2}}$$

при $m \geq 1$.

В ряде случаев удается еще уменьшить число операций в некоторое число раз. Один из таких случаев уже упоминался нами выше: была дана вещественная функция $f(t) = g(\cos t)$, известная в точках $t_l = \frac{\pi(2l-1)}{2N}$; требовалось найти коэффициенты интерполяционного многочлена

$$\sum_{j=0}^{N-1} A_j \cos jt.$$

Другой случай — N четно, заданы значения функции

$$\sum_{j=1}^{N/2-1} A_j \sin 2\pi jx$$

в точках $x_l = l/N$, $0 < l < N/2$; ищутся коэффициенты A_j .

Задача. Требуется найти коэффициенты c_j произведения двух многочленов

$$\left(\sum_{j=0}^{N-1} a_j x^j\right) \left(\sum_{j=0}^{N-1} b_j x^j\right) = \sum_{j=0}^{2N-2} c_j x^j.$$

Показать, что для этого достаточно $O(N \log_2 N)$ операций.

§ 5. Наилучшее равномерное приближение

Если норма в линейном нормированном пространстве определяется не через скалярное произведение, нахождение элемента наилучшего приближения существенно усложняется. Рассмотрим типичную задачу, встречающуюся, в частности, при составлении стандартных программ вычисления функций.

Пусть $\mathcal{R}$ — пространство ограниченных вещественных функций, определенных на отрезке $[a, b]$ вещественной оси, с нормой

$$\|f\| = \sup_{[a, b]} |f(x)|.$$

Ищется наилучшее приближение вида

$$Q_n(x) = \sum_{j=0}^n a_j x^j.$$

Согласно теореме 1 из § 1 существует элемент наилучшего приближения, т. е. многочлен $Q_n^0(x)$ такой, что

$$E_n(f) = \|f - Q_n^0\| \leq \|f - Q_n\|$$

при любом многочлене $Q_n(x)$ степени n . Такой многочлен $Q_n^0(x)$ называют *многочленом наилучшего равномерного приближения*. Далее будут установлены необходимые и достаточные условия того, чтобы многочлен являлся многочленом наилучшего приближения для непрерывной функции.

Теорема (Валле Пуссен). Пусть существуют $n+2$ точки $x_0 < \dots < x_{n+1}$ отрезка $[a, b]$ такие, что

$$\text{sign}((f(x_i) - Q_n(x_i))(-1)^i) = \text{const},$$

т. е. при переходе от точки x_i к следующей точке x_{i+1} величина $f(x) - Q_n(x)$ меняет знак. Тогда

$$E_n(f) \geq \mu = \min_{i=0, \dots, n+1} |f(x_i) - Q_n(x_i)|. \quad (1)$$

Доказательство. В случае $\mu = 0$ утверждение теоремы очевидно. Пусть $\mu > 0$. Предположим противное, т. е. что для многочлена наилучшего приближения $Q_n^0(x)$

$$\|Q_n^0 - f\| = E_n(f) < \mu.$$

Имеем

$$\text{sign}(Q_n(x) - Q_n^0(x)) = \text{sign}((Q_n(x) - f(x)) - (Q_n^0(x) - f(x))).$$

В точках x_i первое слагаемое превосходит по модулю второе; поэтому $\text{sign}(Q_n(x_i) - Q_n^0(x_i)) = \text{sign}(Q_n(x_i) - f(x_i))$. Следовательно,

многочлен $Q_n(x) - Q_n^0(x)$ степени n меняет знак $n+1$ раз. Получили противоречие.

Теорема Чебышева. Чтобы многочлен $Q_n(x)$ был многочленом наилучшего приближения непрерывной функции $f(x)$, необходимо и достаточно существование на $[a, b]$ по крайней мере $m = n+2$ точек $x_0 < \dots < x_{n+1}$ таких, что

$$f(x_i) - Q_n(x_i) = \alpha (-1)^i \|f - Q_n\|,$$

$i = 0, \dots, n+1$, $\alpha = 1$ (или $\alpha = -1$) одновременно для всех i .

Точки $x_0, \dots, x_{n+1}$, удовлетворяющие условиям теоремы, принято называть *точками чебышевского альтернанса*.

Доказательство. Достаточность. Обозначим через L величину $\|f - Q_n\|$. Применяя (1), имеем

$$L = \mu \leq E_n(f), \quad \text{но} \quad E_n(f) \leq \|f - Q_n\| = L$$

вследствие определения величины $E_n(f)$. Следовательно, $E_n(f) = L$, и данный многочлен является многочленом наилучшего равномерного приближения.

Необходимость. Пусть данный многочлен $Q_n(x)$ является многочленом наилучшего равномерного приближения. Обозначим через y_1 нижнюю грань точек $x \in [a, b]$, в которых $|f(x) - Q_n(x)| = L$; из определения L следует существование такой точки. Вследствие непрерывности $f(x) - Q_n(x)$ имеем $|f(y_1) - Q_n(y_1)| = L$. Для определенности далее рассматриваем случай $f(y_1) - Q_n(y_1) = +L$. Обозначим через y_2 нижнюю грань всех точек $x \in [y_1, b]$, в которых $f(x) - Q_n(x) = -L$, последовательно через y_{k+1} обозначим нижнюю грань точек $x \in [y_k, b]$, в которых $f(x) - Q_n(x) = (-1)^k L$, ... Вследствие непрерывности $f(x) - Q_n(x)$ при всех k имеем

$$f(y_{k+1}) - Q_n(y_{k+1}) = (-1)^k L.$$

Продолжаем этот процесс до значения $y_m = b$ или y_m такого, что $|f(x) - Q_n(x)| < L$ при $y_m < x \leq b$. Если $m \geq n+2$, то утверждение теоремы выполнено. Предположим противное.

Вследствие непрерывности $f(x) - Q_n(x)$ при любом k ($1 \leq k \leq m$) можно указать точку z_{k-1} такую, что $|f(x) - Q_n(x)| < L$ при $z_{k-1} \leq x < y_k$; положим $z_0 = a$, $z_m = b$. Согласно построению на отрезках $[z_{i-1}, z_i]$, $i = 1, \dots, m$, имеются точки, в частности, точки y_i , где $f(x) - Q_n(x) = (-1)^{i-1} L$, и нет точек, где $f(x) - Q_n(x) = (-1)^i L$. Положим

$$v(x) = \prod_{i=1}^{m-1} (z_i - x), \quad Q_n^d(x) = Q_n(x) + dv(x), \quad d > 0,$$

и рассмотрим поведение разности

$$f(x) - Q_n^d(x) = f(x) - Q_n(x) - dv(x)$$

на отрезках $[z_{j-1}, z_j]$. Для примера обратимся к отрезку $[z_0, z_1]$. На $[z_0, z_1]$ имеем $v(x) > 0$, и поэтому

$$f(x) - Q_n^d(x) \leq L - dv(x) < L.$$

В то же время $f(x) - Q_n(x) > -L$ на $[z_0, z_1]$, поэтому при d достаточно малых, например, при

$$d < d_1 = \frac{\min_{[z_0, z_1]} |f(x) - Q_n(x) + L|}{\max_{[z_0, z_1]} |v(x)|},$$

на $[z_0, z_1]$ имеем

$$f(x) - Q_n^d(x) > -L.$$

Кроме того,

$$|f(z_1) - Q_n^d(z_1)| = |f(z_1) - Q_n(z_1)| < L.$$

Таким образом,

$$|f(x) - Q_n^d(x)| < L$$

на этом отрезке при достаточно малом d . После проведения аналогичных рассуждений относительно остальных отрезков мы сможем указать малое d_0 такое, что на всех отрезках выполняется неравенство

$$|f(x) - Q_n^{d_0}(x)| < L;$$

$v(x)$ и, следовательно, $Q_n^{d_0}(x)$ являются многочленами степени n . Получилось противоречие с предположением, что $Q_n(x)$ — многочлен наилучшего приближения, а $m < n + 2$. Теорема доказана.

Это доказательство теоремы Чебышева и рассматриваемый в § 7 итерационный метод получения многочлена наилучшего равномерного приближения имеют характер тупика в общей структуре нашего курса: методы исследования и сами результаты никак не используются в дальнейших построениях. Однако они интересны как типичный пример рассмотрения задач математического программирования — нового, быстро развивающегося раздела прикладной математики.

Теорема единственности. *Многочлен наилучшего равномерного приближения непрерывной функции единствен.*

Доказательство. Предположим, что существуют два многочлена степени n наилучшего равномерного приближения:

$$Q_n^1(x) \neq Q_n^2(x), \quad \|f - Q_n^1\| = \|f - Q_n^2\| = E_n(f).$$

Отсюда

$$\left\| f - \frac{Q_n^1 + Q_n^2}{2} \right\| \leq \left\| \frac{f - Q_n^1}{2} \right\| + \left\| \frac{f - Q_n^2}{2} \right\| = E_n(f).$$

Поэтому многочлен $\frac{Q_n^1(x) + Q_n^2(x)}{2}$ также является многочленом наилучшего равномерного приближения. Пусть $x_0, \dots, x_{n+1}$ — соответствующие этому многочлену точки чебышевского альтернанса. Тогда

$$\left| \frac{Q_n^1(x_i) + Q_n^2(x_i)}{2} - f(x_i) \right| = E_n(f), \quad i = 0, \dots, n+1,$$

или

$$|(Q_n^1(x_i) - f(x_i)) + (Q_n^2(x_i) - f(x_i))| = 2E_n(f).$$

Так как $|Q_n^k(x_i) - f(x_i)| \leq E_n(f)$, $k = 1, 2$, то последнее соотношение возможно лишь при

$$Q_n^1(x_i) - f(x_i) = Q_n^2(x_i) - f(x_i).$$

Два различных многочлена степени n $Q_n^1(x)$ и $Q_n^2(x)$ совпадают в $n+2$ точках. Мы пришли к противоречию.

Задача. Функция $f(x) = \sin 100x$ приближается на $[0, \pi]$. Найти $Q_{90}(x)$.

§ 6. Примеры наилучшего равномерного приближения

1. Непрерывная на $[a, b]$ функция $f(x)$ приближается многочленом нулевой степени. Пусть

$$\sup_{[a, b]} f(x) = f(x_1) = M, \quad \inf_{[a, b]} f(x) = f(x_2) = m.$$

Многочлен $Q_0(x) = (M + m)/2$ является многочленом наилучшего приближения, а x_1, x_2 — точками чебышевского альтернанса.

Задачи. а) Доказать, что наилучшее приближение нулевой степени имеет тот же вид, если $f(x)$ не обязательно непрерывна.

б) Построить пример функции и соответствующего многочлена первой степени наилучшего равномерного приближения так, чтобы среди точек чебышевского альтернанса не было a и b .

в) Построить пример функции (естественно не непрерывной), для которой многочлен наилучшего приближения не удовлетворяет условиям теоремы Чебышева.

2. Выпуклая на $[a, b]$ функция $f(x)$ приближается многочленом первой степени $Q_1(x) = a_0 + a_1x$. Вследствие выпуклости $f(x)$ разность $f(x) - (a_0 + a_1x)$ может иметь только одну внутреннюю точку экстремума. Поэтому точки a, b являются точ-

ками чебышевского альтернанса. Пусть d — третья точка альтернанса. Согласно теореме Чебышева, имеем равенства

$$\begin{aligned} f(a) - (a_0 + a_1 a) &= \alpha L, \\ f(d) - (a_0 + a_1 d) &= -\alpha L, \\ f(b) - (a_0 + a_1 b) &= \alpha L. \end{aligned}$$

Вычитая первое уравнение из третьего, получим $f(b) - f(a) = a_1(b - a)$. Отсюда находим $a_1 = \frac{f(b) - f(a)}{b - a}$. Для определения неизвестных d , L , a_0 , a_1 и $\alpha = \pm 1$ получено всего три уравнения. Однако следует вспомнить, что точка d является точкой экстремума разности $f(x) - (a_0 + a_1 x)$. Находим значение d ; если $f(x)$ дифференцируема, то уравнение для определения d будет $f'(d) - a_1 = 0$. Теперь определяем a_0 , например, из уравнения, получающегося сложением первого и второго уравнений. Геометрически эта процедура выглядит следующим образом (рис. 4.6.1). Проводим секущую через точки $(a, f(a))$, $(b, f(b))$. Ее наклон равен a_1 . Проводим параллельную ей касательную к кривой $y = f(x)$, проводим прямую, проходящую посередине между этими прямыми.

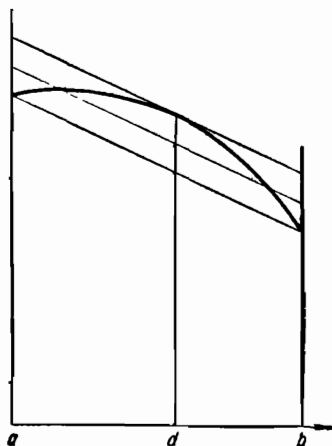


Рис. 4.6.1.

Задача. $f(x) = |x|$, $[a, b] = [-1, 5]$. Построить многочлен наилучшего равномерного приближения первой степени.

3. Функция $f(x)$, удовлетворяющая условию $f^{(n+1)}(x) \geq 0$, приближается на $[a, b]$ многочленом наилучшего приближения степени n ; требуется оценить величину $E_n(f)$. В § 9 гл. II мы имели оценку погрешности интерполяции по узлам многочлена Чебышева

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos \left(\frac{\pi(2k-1)}{2(n+1)} \right),$$

а именно,

$$|f(x) - P_n(x)| \leq \left(\max_{[a, b]} |f^{(n+1)}(x)| \right) \frac{(b-a)^{n+1}}{2^{2n+1} (n+1)!}. \quad (1)$$

Отсюда следует неравенство

$$E_n(f) \leq \left(\max_{[a, b]} |f^{(n+1)}(x)| \right) \frac{(b-a)^{n+1}}{2^{2n+1} (n+1)!}.$$

Пусть $Q_n(x)$ — многочлен наилучшего равномерного приближения. Вследствие теоремы Чебышева разность $f(x) - Q_n(x)$

обращается в нуль в $(n+1)$ -й точке $y_1, \dots, y_{n+1}$. Поэтому многочлен $Q_n(x)$ можно рассматривать как интерполяционный с узлами интерполяции $y_1, \dots, y_{n+1}$. Согласно (2.3.1), имеем представление погрешности интерполирования

$$f(x) - Q_n(x) = f^{(n+1)}(\xi) \frac{\omega_{n+1}(x)}{(n+1)!},$$

где

$$\omega_{n+1}(x) = (x - y_1) \dots (x - y_{n+1}), \quad \xi \in [a, b].$$

Пусть

$$\max_{[a, b]} |\omega_{n+1}(x)| = |\omega_{n+1}(x_0)|.$$

Имеем

$$\begin{aligned} E_n(f) &= \|f(x) - Q_n(x)\| \geq \\ &\geq |f(x_0) - Q_n(x_0)| = |f^{(n+1)}(\xi(x_0))| \frac{|\omega_{n+1}(x_0)|}{(n+1)!} \geq \\ &\geq \left(\min_{[a, b]} |f^{(n+1)}(x)| \right) \left(\max_{[a, b]} \frac{|\omega_{n+1}(x)|}{(n+1)!} \right). \end{aligned}$$

Согласно (2.8.6),

$$\max_{[a, b]} |\omega_{n+1}(x)| \geq \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Отсюда следует оценка

$$E_n(f) \geq \left(\min_{[a, b]} |f^{(n+1)}(x)| \right) \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}. \quad (2)$$

Таким образом, если $f^{(n+1)}(x)$ сохраняет знак и меняется не очень сильно, разность между погрешностью многочлена наилучшего приближения и интерполяционного многочлена по нулям многочленов Чебышева незначительна.

Задача. Доказать, что в рассматриваемом случае чебышевский альтернанс содержит точки a и b .

4. Рассмотрим задачу нахождения многочлена наилучшего приближения степени n в случае, когда

$$f(x) = P_{n+1}(x) = a_0 + \dots + a_{n+1}x^{n+1}, \quad a_{n+1} \neq 0.$$

Тогда $f^{(n+1)}(x) = a_{n+1}(n+1)!$ и оценки сверху (1) и снизу (2) для $E_n(f)$ совпадают:

$$E_n(f) = |a_{n+1}|(b-a)^{n+1} 2^{-2n-1}.$$

Таким образом, многочленом наилучшего приближения оказывается интерполяционный многочлен $Q_n(x)$ при узлах интерполяции

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos \left(\frac{\pi(2k-1)}{2(n+1)} \right), \quad k = 1, \dots, n+1.$$

Можно получить другое представление этого многочлена наилучшего равномерного приближения

$$Q_n(x) = P_{n+1}(x) - a_{n+1} T_{n+1} \left(\frac{2x - (a+b)}{b-a} \right) \frac{(b-a)^{n+1}}{2^{2n+1}}. \quad (3)$$

Действительно, выражение в правой части является многочленом степени n , поскольку коэффициент при x^{n+1} равен нулю; точки $x_i = \frac{a+b}{2} + \frac{b-a}{2} \cos \frac{\pi i}{n+1}$, $i = 0, \dots, n$, максимума модуля разности $P_{n+1} - Q_n(x)$ образуют чебышевский альтернанс на $[a, b]$.

5. Пусть $f(x)$ нечетна относительно $(a+b)/2$ и непрерывна. Для определенности будем говорить об отрезке $[-1, 1]$. Условие нечетности означает $f(x) = -f(-x)$. Покажем, что здесь многочлен наилучшего приближения любой степени нечетен, т. е. записывается в виде суммы нечетных степеней x . Действительно, пусть $Q_n(x)$ — многочлен наилучшего приближения для $f(x)$. Имеем $|f(x) - Q_n(x)| \leq E_n(f)$. После замены x на $-x$ и умножения выражения под знаком модуля на -1 получим

$$| -f(-x) - (-Q_n(-x)) | \leq E_n(f),$$

иначе

$$| f(x) - (-Q_n(-x)) | \leq E_n(f).$$

Следовательно, многочлен $-Q_n(-x)$ также является многочленом наилучшего равномерного приближения. Вследствие теоремы единственности имеем $Q_n(x) = -Q_n(-x)$, что и требовалось доказать.

6. Пусть требуется приблизить x^3 на $[-1, 1]$ многочленом наилучшего приближения первой степени. Предшествующим результатом можно воспользоваться двояко. Один путь: поскольку искомый многочлен наилучшего приближения будет нечетным, то его достаточно отыскивать среди многочленов $Q_1(x) = a_1 x$. Второй путь: поскольку для данной задачи многочлен наилучшего приближения второй степени оказывается многочленом первой степени, то исходная задача эквивалентна задаче построения многочлена наилучшего приближения второй степени. Последняя задача наилучшего приближения многочлена многочленом степени, на единицу меньшей, уже нами рассматривалась.

Задача. $f(x)$ четна относительно середины отрезка приближения $[-1, 1]$; $f(x) = f(-x)$, и непрерывна.

Доказать, что многочлен $Q_n(x)$ наилучшего приближения четен.

Задача. $f(x) = \exp(x^2)$ приблизить на $[-1, 1]$ многочленом наилучшего приближения третьей степени.

Из рассмотренного выше вытекает, что этот многочлен имеет вид $a_0 + a_2 x^2$. Задача эквивалентна задаче наилучшего приближения $f_1(y) = \exp y$ на $[0, 1]$ многочленом $a_0 + a_2 y$ (см. п. 2).

7. Довольно распространен случай, когда достаточно получить хорошее приближение к функции и не обязательно искать наилучшее равномерное приближение. Для простоты рассматриваем случай приближения на отрезке $[-1, 1]$. Разложим функцию $f(x)$ в ряд по ортогональной системе многочленов Чебышева

$$f(x) \sim \sum_{j=0}^{\infty} d_j T_j(x).$$

Часто отрезок этого ряда

$$\sum_{j=0}^n d_j T_j(x)$$

невысокой степени обеспечивает неплохое равномерное приближение.

Иногда бывает затруднительно вычислить явно коэффициенты d_j , но зато известно разложение Тейлора

$$f(x) = \sum_{j=0}^{\infty} a_j x^j,$$

сходящееся при $|x| \leq 1$. Тогда может быть полезен следующий метод (называемый *телескопическим*).

Выбирают некоторое n такое, что погрешность формулы

$$f(x) \approx P_n(x) = \sum_{j=0}^n a_j x^j$$

является достаточно малой. Затем приближают многочлен $P_n(x)$ многочленом его наилучшего равномерного приближения $P_{n-1}(x)$. Согласно формуле (3),

$$P_{n-1}(x) = P_n(x) - a_n T_n(x) 2^{1-n}.$$

Поскольку $|T_n(x)| \leq 1$ на $[-1, 1]$, то

$$|P_{n-1}(x) - P_n(x)| \leq |a_n| 2^{1-n},$$

далее приближают многочлен $P_{n-1}(x)$ его многочленом наилучшего равномерного приближения $P_{n-2}(x)$ и т. д., понижение продолжается до тех пор, пока погрешность от таких последовательных аппроксимаций остается малой.

Рассматриваемый прием можно описать еще следующим образом. Разложим многочлен $P_n(x)$ по многочленам Чебышева

$$P_n(x) = \sum_{j=0}^n d_j T_j(x).$$

Введем обозначения $Q_m(x) = \sum_{j=0}^m d_j T_j(x)$ при $m \leq n$. Всякий многочлен $Q_m(x)$ является многочленом наилучшего равномерного приближения степени m для многочлена $Q_{m+1}(x)$, при этом

$$E_m(Q_{m+1}) = \|Q_{m+1} - Q_m\| = |d_{m+1}|. \quad (4)$$

Это следует, например, из формулы (3) или непосредственно из теоремы Чебышева. Отсюда вытекает, что $Q_{n-1}(x) = P_{n-1}(x)$, $Q_{n-2}(x) = P_{n-2}(x)$ и т. д.

Таким образом, существо описанного метода заключается в следующем. Исходная функция приближается отрезком ее ряда Тейлора $P_n(x)$. Затем многочлен $P_n(x)$ раскладывается по многочленам Чебышева и отбрасывается несколько последних членов разложения. Так как

$$|P_n(x) - P_m(x)| \leq \sum_{j=m+1}^n |d_j|,$$

то общая оценка погрешности

$$|f(x) - P_m(x)| \leq \max |f(x) - P_n(x)| + \sum_{j=m+1}^n |d_j|.$$

Рассмотрим задачу приближения $\arctg x$ на $\left[-\operatorname{tg} \frac{\pi}{8}, \operatorname{tg} \frac{\pi}{8}\right]$ с точностью $0,5 \cdot 10^{-5}$. Для достижения такой точности при аппроксимации отрезком ряда Тейлора требуется положить

$$\arctg x \approx x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \frac{x^{11}}{11}.$$

После трех шагов приближения многочлена многочленом его наилучшего равномерного приближения на единицу меньшей степени получим многочлен

$$P_6(x) = 0,9999374x - 0,3303433x^3 + 0,1632823x^5,$$

также обеспечивающий требуемую точность. Отметим, что здесь из-за нечетности исходного многочлена показатель степени на каждом шаге уменьшался на 2.

8. Иногда вместо теоремы Валле—Пуссена для оценки величины $E_n(f)$ снизу удобнее использовать следующий результат:

Теорема. Пусть $f(x)$ приближается на $[-1, 1]$,

$$Q_{n+1}(x) = a_{n+1}x^{n+1} + \dots + a_0,$$

тогда

$$E_n(f) \geq |a_{n+1}| 2^{-n} - \|f - Q_{n+1}\|.$$

Доказательство. Если Q_m^f — многочлен наилучшего приближения к функции f , а Q_m^g к функции g , то

$$\begin{aligned} |(Q_m^f(x) + Q_m^g(x)) - (f(x) + g(x))| &\leq \\ &\leq |Q_m^f(x) - f(x)| + |Q_m^g(x) - g(x)| \leq E_m(f) + E_m(g). \end{aligned}$$

Отсюда следует, что

$$E_m(f + g) \leq E_m(f) + E_m(g).$$

Полагая $m = n$, $g = Q_{n+1} - f$, получим

$$E_n(Q_{n+1}) \leq E_n(f) + E_n(Q_{n+1} - f).$$

Вследствие (3) имеем

$$E_n(Q_{n+1}) = |a_{n+1}| 2^{-n}.$$

Таким образом,

$$E_n(f) \geq E_n(Q_{n+1}) - E_n(Q_{n+1} - f) \geq |a_{n+1}| 2^{-n} - \|f - Q_{n+1}\|. \quad (5)$$

Теорема доказана.

Рассмотрим пример получения двусторонней оценки для $E_n(f)$. Пусть

$$f(x) = \sum_{j=0}^{\infty} d_j T_j(x),$$

положим

$$Q_m(x) = \sum_{j=0}^m d_j T_j(x).$$

Очевидно,

$$\|f - Q_{n+1}\| \leq \sum_{j=n+2}^{\infty} |d_j|;$$

согласно (4), $E_n(Q_{n+1}) = |d_{n+1}|$. Обращаясь к первому из неравенств (5), получаем оценку

$$E_n(f) \geq |d_{n+1}| - \sum_{j=n+2}^{\infty} |d_j|.$$

В то же время очевидно, что

$$E_n(f) \leq \|f - Q_n\| \leq \sum_{j=n+1}^{\infty} |d_j| = |d_{n+1}| + \sum_{j=n+2}^{\infty} |d_j|.$$

При $|d_{n+1}| > \sum_{j=n+2}^{\infty} |d_j|$ из сравнения этих оценок можно получить двустороннюю оценку для $E_n(f)$.

§ 7. Итерационный метод построения многочлена наилучшего равномерного приближения

Обычно многочлен наилучшего равномерного приближения отыскивается каким-либо итерационным методом. Описываемый ниже алгоритм состоит в получении последовательных приближений к точкам чебышевского альтернанса и соответствующих этим точкам многочленов, сходящихся к многочлену наилучшего равномерного приближения.

В дальнейшем предполагается, что приближаемая функция непрерывна и не является многочленом степени n , т. е. $E_n(f) \neq 0$.

Пусть имеются некоторые приближения к точкам чебышевского альтернанса $x_0^k < \dots < x_{n+1}^k$. Построим многочлен

$$Q_n^k(x) = \sum_{l=0}^n a_l^k x^l,$$

удовлетворяющий в этих точках соотношениям

$$f(x_i^k) - Q_n^k(x_i^k) = (-1)^i E^k, \quad (1)$$

где E^k — некоторая неизвестная величина. Получились уравнения

$$\sum_{l=0}^n a_l^k (x_i^k)^l + (-1)^i E^k = f(x_i^k) \quad (2)$$

относительно $n+2$ неизвестных $a_0^k, \dots, a_n^k, E^k$.

Покажем, что определитель этой системы отличен от нуля.

Предположим противное; тогда однородная система уравнений

$$Q_n^k(x_i) + (-1)^i E^k = \sum_{l=0}^n a_l^k (x_i^k)^l + (-1)^i E^k = 0$$

имеет ненулевое решение. Если у этого решения $E^k = 0$, то хотя бы одно из a_l^k отлично от нуля; ненулевой многочлен $Q_n^k(x)$ степени n имеет $n+2$ нуля; пришли к противоречию. Если $E^k \neq 0$, то хотя бы одно из a_l^k также отлично от нуля, тогда ненулевой многочлен $Q_n^k(x)$ меняет знак между каждой парой точек x_i^k, x_{i+1}^k , т. е. имеет не менее $n+1$ -го нуля; опять получается противоречие.

Введем обозначения

$$x_{-1}^k = a, \quad x_{n+2}^k = b, \quad R_n^k(x) = f(x) - Q_n^k(x).$$

В процессе итераций фиксируется некоторое q_0 в пределах $0 < q_0 < 1$.

За x_i^k при $k > 0$ принимаем точки, принадлежащие отрезкам $[x_{i-1}^{k-1}, x_{i+1}^{k-1}]$, $i = 0, \dots, n+1$, и удовлетворяющие в совокупности условиям

- а) $\max_{0 \leq i \leq n+1} |R_n^{k-1}(x_i^k)| \geq |E^{k-1}| + q_0 (\|R_n^{k-1}(x)\| - |E^{k-1}|),$
- б) $R_n^{k-1}(x_i^k) R_n^{k-1}(x_{i+1}^k) \leq 0 \quad \text{при} \quad i = 0, \dots, n,$
- в) $|R_n^{k-1}(x_i^k)| \geq |E^{k-1}| \quad \text{при} \quad i = 0, \dots, n+1.$

Множество таких точек не пусто и может быть построено, например, следующим образом.

Возьмем произвольную точку x^k такую, что

$$|R_n^{k-1}(x^k)| \geq |E^{k-1}| + q_0(\|R_n^{k-1}(x)\| - |E^{k-1}|);$$

так как правая часть этого выражения не превосходит $\|R_n^{k-1}(x)\|$, то такая точка существует. Если $E^{k-1} = 0$, то требуемым условиям удовлетворяет совокупность, состоящая из точки x^k и любых $n+1$ точек из $x_0^{k-1}, \dots, x_{n+1}^{k-1}$

При $E^{k-1} \neq 0$ все $|R_n^{k-1}(x_i^{k-1})| \neq 0$; рассмотрим отдельно случаи:

а) $x^k \in [x_{-1}^{k-1}, x_0^{k-1}]$,

б) $x^k \in [x_{n+1}^{k-1}, x_{n+2}^{k-1}]$,

в) $x^k \in [x_0^{k-1}, x_{n+1}^{k-1}]$.

В случае а) полагаем $x_0^k = x^k$,

$$x_i^k = \begin{cases} x_i^{k-1} & \text{при } R_n^{k-1}(x_0^{k-1}) R_n^{k-1}(x_0^k) > 0, \\ x_{i-1}^{k-1} & \text{при } R_n^{k-1}(x_0^{k-1}) R_n^{k-1}(x_0^k) < 0 \end{cases} \quad \text{для } i = 1, \dots, n+1.$$

В случае б) полагаем $x_{n+1}^k = x^k$,

$$x_i^k = \begin{cases} x_i^{k-1} & \text{при } R_n^{k-1}(x_{n+1}^{k-1}) R_n^{k-1}(x_{n+1}^k) > 0, \\ x_{i+1}^{k-1} & \text{при } R_n^{k-1}(x_{n+1}^{k-1}) R_n^{k-1}(x_{n+1}^k) < 0 \end{cases} \quad \text{для } i = 0, \dots, n.$$

Обратимся к случаю в); пусть

$$x^k \in [x_{i_0}^{k-1}, x_{i_0+1}^{k-1}], \quad \text{где } 0 \leq i_0 \leq n;$$

если $R_n^{k-1}(x^k) R_n^{k-1}(x_{i_0}^{k-1}) > 0$, то положим

$$x_{i_0}^k = x^k, \quad x_i^k = x_i^{k-1} \quad \text{при } i \neq i_0;$$

если же $R_n^{k-1}(x^k) R_n^{k-1}(x_{i_0+1}^{k-1}) > 0$, положим

$$x_{i_0+1}^k = x^k, \quad x_i^k = x_i^{k-1} \quad \text{при } i \neq i_0 + 1.$$

Во всех случаях полученные совокупности точек x_i^k удовлетворяют условиям а), б), в).

Если не известна дополнительная информация о положении искомым точек чебышевского альтернанса, то в качестве x_i^0 часто берут точки экстремума многочлен Чебышева

$$x_i^0 = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\pi \frac{n+1-i}{n+1}\right), \quad i = 0, \dots, n+1.$$

Задача. Пусть $f(x)$ — многочлен степени $n+1$. Показать, что при таких начальных приближениях x_i^0 многочлен $Q_n^0(x)$ окажется искомым многочленом наилучшего равномерного приближения.

Если функция $f(x)$ близка к четной или нечетной относительно середины отрезка, то может оказаться удобным взять в качестве x_i^0 точки экстремума многочленов Чебышева соответствующей степени. В случае построения многочленов наилучшего приближения для большой серии функций $f(A, x)$, зависящих от параметра A , иногда оказывается удобной следующая процедура, основанная на предположении, что точки чебышевского альтернанса x_i являются гладкими функциями параметра A : после нахождения этих точек для некоторых значений $A_1, \dots, A_m$ начальные приближения при других значениях A определяются интерполяцией.

Теорема. Пусть

$$\max_{[a, b]} |f^{(l)}(x)| = M_l < \infty \quad \text{при } l = 0, \dots, n+2.$$

Тогда для описанного итерационного процесса выполняется соотношение

$$\|f - Q_n^k\| \leq E_n(f) + \text{const } q^k$$

при $q < 1$.

При этом $\|Q_n^k - Q_n\| \rightarrow 0$, где Q_n — многочлен наилучшего равномерного приближения.

Доказательство. Разность $f - Q_n^k$ обращается в нуль в некоторых точках $y_1^k, \dots, y_{n+1}^k$ вследствие перемены знака при переходе от каждой точки x_i^k к точке x_{i+1}^k . Поэтому $Q_n^k(x)$ является интерполяционным многочленом для $f(x)$ с узлами интерполяции $y_1^k, \dots, y_{n+1}^k \in [a, b]$. Согласно (2.3.1) и (2.17.5), при $x \in [a, b]$ имеем

$$R_n^k(x) = f(x) - Q_n^k(x) = f^{(n+1)}(\xi_1) \frac{\Omega^k(x)}{(n+1)!},$$

$$(R_n^k(x))' = f'(x) - (Q_n^k(x))' = f^{(n+1)}(\xi_1) \frac{(\Omega^k(x))'}{(n+1)!} + f^{(n+2)}(\xi_2) \frac{\Omega^k(x)}{(n+2)!};$$

здесь $\Omega^k(x) = (x - y_1^k) \dots (x - y_{n+1}^k)$, $\xi_1, \xi_2 \in [a, b]$. Так как $|(\Omega^k(x))'| \leq (n+1)(b-a)^n$, то получаем оценку

$$|R_n^k(x)| \leq M_{n+1} \frac{(b-a)^{n+1}}{(n+1)!} = D_1,$$

$$|(R_n^k(x))'| \leq M_{n+1} \frac{(b-a)^n}{n!} + M_{n+2} \frac{(b-a)^{n+1}}{(n+2)!} = D_2$$

при всех k .

Запишем систему соотношений (1) в виде

$$R_n^k(x_i^k) = E^k \psi^k(x_i^k),$$

где $\psi^k(x_i^k) = (-1)^i$ в точке x_i^k . Образует разделенную разность по точкам $x_0^k, \dots, x_{n+1}^k$ от обеих частей:

$$R_n^k(x_0^k; \dots; x_{n+1}^k) = E^k \psi^k(x_0^k; \dots; x_{n+1}^k).$$

Поскольку разделение разности по $(n+2)$ -м узлам от многочлена степени n равна 0, то вместо R_n^k можно поставить R_n^{k-1} ; получим

$$R_n^{k-1}(x_0^k; \dots; x_{n+1}^k) = E^k \psi^k(x_0^k; \dots; x_{n+1}^k),$$

или

$$E^k = \frac{R_n^{k-1}(x_0^k; \dots; x_{n+1}^k)}{\psi^k(x_0^k; \dots; x_{n+1}^k)}.$$

Раскрывая выражение разделенной разности, имеем

$$E^k = \sum_{i=0}^{n+1} \frac{d_i^k R_n^{k-1}(x_i^k)}{\sum_{i=0}^{n+1} (-1)^i d_i^k}, \quad (3)$$

где

$$d_i^k = \prod_{j \neq i} (x_i^k - x_j^k)^{-1}.$$

Величины $R_n^{k-1}(x_i^k)$ при $i=0, \dots, n+1$, k — фиксированном имеют чередующиеся знаки из-за способа выбора точек x_i^k ; величины d_i^k также имеют чередующиеся знаки. Поэтому соотношение (3) записывается в виде

$$|E^k| = \sum_{i=0}^{n+1} h_i^k |R_n^{k-1}(x_i^k)|, \quad (4)$$

где

$$h_i^k = |d_i^k| / \left(\sum_{i=0}^{n+1} |d_i^k| \right) > 0.$$

Очевидно, $\sum_{i=0}^{n+1} h_i^k = 1$. Из выполнения условия γ) при выборе точек x_i^k следует, что

$$|R_n^{k-1}(x_i^k)| \geq |E^{k-1}| \quad \text{при} \quad i=0, \dots, n+1.$$

Поэтому из (4) получаем

$$|E^k| \geq \sum_{i=0}^{n+1} h_i^k |E^{k-1}| = |E^{k-1}|. \quad (5)$$

Если Q_n^{k-1} не является искомым многочленом наилучшего равномерного приближения, то точки $x_0^{k-1}, \dots, x_{n+1}^{k-1}$ не образуют чебышевский альтернанс и, согласно условию α),

$$|R_n^{k-1}(x_i^k)| > |E^{k-1}|,$$

хотя бы при одном i , лежащем в пределах $0 \leq i \leq n+1$. Тогда вместо (5) получим $|E^k| > |E^{k-1}|$. Могло оказаться, что $E^0 = 0$, однако вследствие последнего неравенства при всех $k \geq 1$ имеем $|E^k| \geq |E^1| > |E^0| \geq 0$. По-

тому при любых $k \geq 1$ и $0 \leq i \leq n$ можно написать цепочку соотношений

$$\begin{aligned} |E^k| &\leq 2 |R_n^k(x_{i+1}^k) - R_n^k(x_i^k)| \leq \\ &\leq \max_{[a, b]} |(R_n^k(x))'| (x_{i+1}^k - x_i^k) \leq D_2 (x_{i+1}^k - x_i^k). \end{aligned}$$

Отсюда получаем, что при $k \geq 1$, $0 \leq i < j \leq n+1$ справедливо

$$x_j^k - x_i^k \geq x_{i+1}^k - x_i^k \geq \varepsilon > 0, \quad \text{где} \quad \varepsilon = 2 \frac{|E^1|}{D_2};$$

кроме того, $|x_i^k - x_j^k| \leq b - a$, поэтому

$$(b - a)^{-(n+1)} \leq |d_i^k| \leq e^{-(n+1)}.$$

Отсюда следует, что

$$h_i^k \geq \frac{(b - a)^{-(n+1)}}{\sum_{i=0}^{n+1} e^{-(n+1)}} = \frac{1}{n+2} \left(\frac{\varepsilon}{b - a} \right)^{n+1} = d, \quad (6)$$

где d не зависит от k и i . Из выполнения условия α) следует, что хотя бы при одном p , $0 \leq p \leq n+1$, выполняются неравенства

$$|R_n^{k-1}(x_p^k)| \geq |E^{k-1}| + q_0 (\|R_n^{k-1}\| - |E^{k-1}|).$$

Обращаясь к (4), с учетом (6) получим

$$\begin{aligned} |E^k| &\geq \sum_{i \neq p} h_i^k |E^{k-1}| + h_p^k (|E^{k-1}| + q_0 (\|R_n^{k-1}\| - |E^{k-1}|)) = \\ &= |E^{k-1}| + h_p^k q_0 (\|R_n^{k-1}\| - |E^{k-1}|) \geq |E^{k-1}| + d q_0 (\|R_n^{k-1}\| - |E^{k-1}|), \quad (7) \end{aligned}$$

Так как $E_n(f) \leq \|R_n^{k-1}\|$, то отсюда следует неравенство

$$E_n(f) - |E^k| \leq (1 - d q_0) (E_n(f) - |E^{k-1}|),$$

и поэтому

$$0 \leq E_n(f) - |E^k| \leq (1 - d q_0)^k (E_n(f) - |E^0|), \quad |E^k| = E_n(f) + O(q^k).$$

Из (7) также следует, что

$$\|R_n^{k-1}\| - |E^k| \leq (d q_0)^{-1} (|E^k| - |E^{k-1}|);$$

отсюда имеем

$$\|R_n^k\| = E_n(f) + O(q^k). \quad (8)$$

Мы доказали справедливость первого из утверждений теоремы.

При практическом отыскании многочлена наилучшего приближения обычно производится дополнительное смещение из точек x_i^k , предложенных нами ранее, с тем, чтобы каждая из величин $|R^{k-1}(x_i^k)|$ была по возможности больше, а условия α), β), γ) не нарушались.

Далее будет доказано второе утверждение теоремы. Заметим, что по существу сам факт сходимости $Q_n^k(x)$ к $Q_n(x)$ не имеет практического интереса. Многочлен наилучшего приближения строится с целью применения

его в дальнейшем для вычисления значений функции. Если вместо него получается многочлен $Q_n^k(x)$, для которого $\|Q_n^k - f\| \approx E_n(f) = \|Q_n - f\|$, то по отношению к поставленной цели совершенно неважно, близки ли многочлены $Q_n^k(x)$ и $Q_n(x)$.

Фиксируем $n+1$ различную точку $x_0, \dots, x_n \in [a, b]$. Пусть S^k — точка с координатами $(Q_n^k(x_0), \dots, Q_n^k(x_n))$. Покажем, что

$$S^k \rightarrow S = (Q_n(x_0), \dots, Q_n(x_n)) \quad \text{при } k \rightarrow \infty.$$

Предположим противное, т. е. предположим, что существует подпоследовательность $k(j) \rightarrow \infty$, для которой

$$\rho(S^{k(j)}, S) \geq \gamma > 0 \quad (9)$$

при всех j . Здесь ρ — обычное евклидово расстояние. Величины $|Q_n^k(x_i)|$ равномерно ограничены вследствие неравенства

$$|f(x) - Q_n^k(x)| \leq D_1.$$

Следовательно, множество точек $S^{k(j)}$ лежит в конечной части $(n+1)$ -мерного пространства и является компактом; поэтому из последовательности $S^{k(j)}$ можно выделить подпоследовательность $S^{k(j(m))}$, сходящуюся к некоторому пределу

$$S_0 = (b_0, \dots, b_n) \quad \text{при } m \rightarrow \infty.$$

Интерполяционный многочлен степени n для $Q_n^k(x)$ при узлах интерполяции $x_0, \dots, x_n$ совпадает с ним самим; обращаясь к формуле для интерполяционного многочлена Лагранжа, имеем

$$Q_n^k(x) = \sum_{i=0}^n Q_n^k(x_i) \omega_i(x), \quad (10)$$

где

$$\omega_i(x) = \prod_{l \neq i} \frac{x - x_l}{x_i - x_l}.$$

Так как $Q_n^{k(j(m))} \rightarrow b_i$ при $m \rightarrow \infty$, то из соотношения (10) получаем

$$Q_n^{k(j(m))} \rightarrow \sum_{i=0}^n b_i \omega_i(x) = \tilde{Q}_n(x) \quad (11)$$

равномерно на $[a, b]$ при $m \rightarrow \infty$. Имеем

$$\|\tilde{Q}_n - f\| \leq \|\tilde{Q}_n - Q_n^{k(j(m))}\| + \|Q_n^{k(j(m))} - f\|.$$

При $m \rightarrow \infty$ первое слагаемое стремится к нулю, согласно (11), а второе к $E_n(f)$, согласно (8). Поэтому

$$\|\tilde{Q}_n - f\| \leq E_n(f).$$

По теореме об единственности многочлена наилучшего равномерного приближения имеем

$$\tilde{Q}_n(x) = Q_n(x).$$

Следовательно,

$$Q_n^{k(l(m))}(x_i) \rightarrow \tilde{Q}_n(x_i) = Q_n(x_i)$$

при всех i . Это означает, что $S^{k(l(m))} \rightarrow S$, и приходим к противоречию с (9). Таким образом, установлено, что $S^k \rightarrow S$ при $k \rightarrow \infty$. Тогда из (11) следует, что $Q_n^k(x) \rightarrow Q_n(x)$ равномерно на $[a, b]$. Теорема доказана.

Может возникнуть вопрос: зачем берется $q_0 < 1$. Может быть, лучше всего брать $q_0 = 1$; тогда и оценка (8) будет лучше и, по-видимому, итерационный процесс будет сходиться быстрее. Вопрос о выборе значения q_0 для каждого конкретного класса требует проведения экспериментальных расчетов. Трудоемкость рассматриваемого алгоритма складывается из работы по решению системы (2) и работы по выбору точек x_i^k . Работа по определению точек x_i^k зависит от сложности вычисления значений функции $f(x)$. Если значения $f(x)$ вычисляются относительно просто, то разумно взять q_0 побольше. Также следует иметь в виду, что на начальных этапах приближений мы все равно находимся далеко от искомого решения, и поэтому допустимы некоторые «загрубления алгоритма» с целью уменьшения трудоемкости первых итераций.

Например, описанный выше алгоритм иногда модифицируется следующим образом. Поставим в соответствие каждому натуральному m некоторое ω_m : $0 < \omega_m < 1$ и совокупность M_m точек отрезка $[a, b]$: $y_1^m, \dots, y_{N(m)}^m$. При этом налагается требование, чтобы при $m \rightarrow \infty$ число точек y_i^m , попадающих в любой отрезок $[a_1, b_1] \subset [a, b]$, стремилось к бесконечности и чтобы $\omega_m \rightarrow 1$. Алгоритм приближенного нахождения многочлена наилучшего приближения состоит в последовательном построении многочленов $P_n^m(x)$, удовлетворяющих условиям: среди точек y_i^m можно выбрать $n+2$ точки $y_{i_0}^m < \dots < y_{i_{n+1}}^m$ такие, что знаки $f(y_{i_k}^m) - P_n^m(y_{i_k}^m)$ в последующих точках различны и

$$0 \leq k \leq n+1 \quad |f(y_{i_k}^m) - P_n^m(y_{i_k}^m)| \geq \omega_m \quad 1 \leq i \leq N(m) \quad |f(y_i^m) - P_n^m(y_i^m)|.$$

Многочлен $P_n^m(x)$ берется в качестве начального приближения для нахождения многочлена $P_n^{m+1}(x)$. Алгоритм итераций для нахождения $P_n^m(x)$ в точности такой же, как описанный нами алгоритм нахождения многочлена наилучшего равномерного приближения, когда многочлены рассматриваются на всем отрезке $[a, b]$. Отличие состоит лишь в том, что точки x_i^k выбираются среди точек множества M_m . Одним из распространенных является случай, когда M_m — множество точек максимума многочленов Чебышева M_m^0 : $y_i^m = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\pi \frac{i-1}{N(m)-1}\right)$. Обычно $N(m) = sp^m + 1$, где s, p — целые; тогда выполняется соотношение $M_m^0 \subset M_{m+1}^0$; при этом чаще всего берут $s = n+1$ и $p = 2$.

Конечно, мы еще не дали полного описания стандартной программы отыскания многочлена наилучшего равномерного приближения. В частности, точки x_i^k можно определять или путем рассмотрения таблицы всех значений $f(x) - Q_n^k(x)$ в точках множеств M_m или каким-то более быстрым, но менее надежным способом; не указан также способ выбора значений ω_m .

§ 8. О форме записи многочлена

Одна из стандартных программ наилучшего равномерного приближения была отлажена в случае приближения многочленами невысокой степени и включена в пакет стандартных программ. Однако при практическом приближении многочленами высокой степени в ряде случаев оказывалось, что программа останавливается, не дав нужного приближения, или работает недопустимо долго, так что вычисления приходится прекращать.

На первый взгляд казалось, что причина неудачи в большой вычислительной погрешности при решении системы уравнений (7.2). Поэтому вместо непосредственного решения этой системы было решено определять E^k из (7.3). Теперь, вследствие (7.2), становятся известными значения $Q_n^k(x_i^k)$, и значения $Q_n^k(x)$, необходимые для получения последующего приближения, можно получить по интерполяционной формуле Лагранжа

$$Q_n^k(x) = \sum_{i=0}^n Q_n^k(x_i^k) \omega_i(x), \quad \omega_i(x) = \prod_{j \neq i, 0 \leq j \leq n} \frac{x - x_j^k}{x_i^k - x_j^k}. \quad (1)$$

После окончания процесса последний из полученных многочленов $P_n(x) = Q_n^k(x)$ расписывался в обычной форме

$$P_n(x) = a_0 + a_1x + \dots + a_nx^n$$

и на печать выдавались значения

$$\{a_j\} \quad \text{и} \quad e_n = \|P_n(x) - f(x)\|$$

Однако оказалось, что такое решение проблемы не является удовлетворительным.

В ряде случаев обнаруживалось большое расхождение значений $f(x)$ и значений $P_n(x)$, вычисляемых по схеме Горнера.

$$P_n(x) = a_0 + x(a_1 + x(a_2 + \dots + x(a_{n-1} + xa_n) \dots)).$$

Причина этого заключалась в том, что величины $|a_j|$ были очень велики и поэтому имели большую абсолютную погрешность.

Оказывается, что для многих функций рост суммы $A_n = \sum_{j=0}^n |a_j|$ является неизбежным при повышении точности приближения функции многочленами. Ниже показано, что скорость этого роста связана с аналитическими свойствами приближаемой функции.

Предположим, существует последовательность многочленов

$$P^m(x) = \sum_{j=0}^{n(m)} a_j^m x^j,$$

удовлетворяющих условию

$$|f(x) - P^m(x)| \leq 2^{-m} \quad \text{на} \quad [-1, 1], \quad (2)$$

при этом $P^m(x)$ — не обязательно многочлены наилучшего равномерного приближения; предположим, что коэффициенты этих многочленов растут не очень сильно:

$$l_m = \sum_{j=0}^{n(m)} |a_j^m| \leq M 2^{qm}, \quad 0 \leq q, \quad 1 \leq M. \quad (3)$$

Теорема. Пусть последовательность многочленов $P^m(x)$ удовлетворяет соотношениям (2), (3). Тогда функция $f(x)$ может быть аналитически продолжена в открытую область комплексного переменного G_Φ , $\Phi_0 = \pi \frac{2q+1}{2q+2}$ (определение этой области будет дано ниже).

Доказательство. Рассмотрим многочлены

$$Q^m(z) = P^m(z) - P^{m-1}(z);$$

вследствие (2) имеем

$$|Q^m(x)| \leq |f(x) - P^m(x)| + |f(x) - P^{m-1}(x)| \leq 3 \cdot 2^{-m},$$

$$|P^m(z)| \leq \sum_{j=0}^{n(m)} |a_j^m| = l_m \quad \text{при} \quad |z| \leq 1.$$

Поэтому

$$|Q^m(z)| \leq |P^m(z)| + |P^{m-1}(z)| \leq l_{m-1} + l_m \leq 2M 2^{qm} \quad \text{при} \quad |z| \leq 1.$$

Лемма. Пусть $g(z)$ аналитическая в круге $|z| \leq 1$, причем $|g(z)| \leq M_2$ на границе круга и $|g(z)| \leq M_1$ на отрезке $[-1, 1]$. Тогда в этом круге

$$|g(z)| \leq M_1^{(2/\pi)\Phi(z)-1} M_2^{2-(2/\pi)\Phi(z)}, \quad (4)$$

здесь $\Phi(z)$ — угол (ненаправленный), под которым из точки $z = x + iy$ виден отрезок $[-1, 1]$.

Доказательство проведем для точек полукруга K_1 : $\text{Im } z \geq 0$. Рассмотрим функцию

$$\omega(z) = \left(\frac{2}{\pi} \Phi(z) - 1\right) \ln M_1 + \left(2 - \frac{2}{\pi} \Phi(z)\right) \ln M_2 - \ln |g(z)|.$$

На отрезке $[-1, 1]$ $\Phi(z) = \pi$ и поэтому

$$\omega = \ln M_1 - \ln |g(z)| \geq 0; \quad (5)$$

на дуге круга $\Phi(z) = \frac{\pi}{2}$ и поэтому

$$\omega = \ln M_2 - \ln |g(z)| \geq 0. \quad (6)$$

Вследствие равенств $\ln |g(z)| = \text{Re } \ln g(z)$,

$$\Phi(z) = \arg \left(\frac{z-1}{z+1} \right) = \text{Im } \ln \left(\frac{z-1}{z+1} \right)$$

функции $\ln|g(z)|$ и $\varphi(z)$ — гармонические по переменным x, y ; следовательно, $\omega(x, y)$ — гармоническая, и поэтому ее нижняя грань в K_1 достигается на границе. В соответствии с (5) и (6) это означает $\omega(x, y) \geq 0$ в K_1 . Отсюда

$$\ln|\tilde{g}(z)| \leq \left(\frac{2}{\pi}\varphi(z) - 1\right) \ln M_1 + \left(2 - \frac{2}{\pi}\varphi(z)\right) \ln M_2.$$

Потенцируя это неравенство, получим (4).

Применяя эту лемму для оценки $|Q^m(z)|$, получим

$$|Q^m(z)| \leq (3 \cdot 2^{-m})^{((2/\pi)\varphi - 1)} (2M \cdot 2^{qm})^{2 - (2/\pi)\varphi} \leq 3M2^{((2q+1) - (2q+2)\frac{\varphi}{\pi})m}. \quad (7)$$

Обозначим через G_φ открытую область, ограниченную дугами, из точек которых отрезок $[-1, 1]$ виден под углом φ (см. рис. 4.8.1). Поскольку функция $2q+1 - (2q+2)\frac{\varphi}{\pi}$ убывает с ростом φ и из всех точек области G_φ отрезок $[-1, 1]$ виден под углом, не меньшим φ , то неравенство (7) выполняется во всех точках множества G_φ .

Пусть φ_0 — решение уравнения

$$(2q+1) - (2q+2)\frac{\varphi_0}{\pi} = 0, \quad (8)$$

т. е. $\varphi_0 = \pi \frac{2q+1}{2q+2}$. Возьмем произвольное $\varphi_1 > \varphi_0$. Верно очевидное неравенство

$$\begin{aligned} |P^m(z)| &\leq |P^0(z)| + \\ &+ \sum_{k=1}^m |P^k(z) - P^{k-1}(z)| = \\ &= |P^0(z)| + \sum_{k=1}^m |Q^k(z)|. \end{aligned}$$

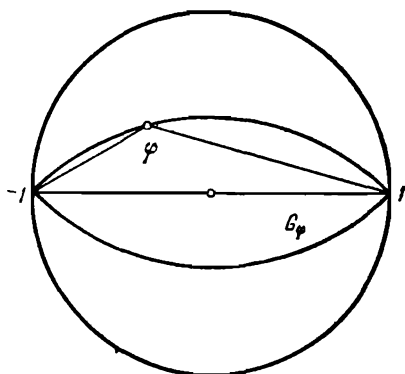


Рис. 4.8.1.

Вследствие (7) для всех точек области $\bar{G}_{\varphi_1}$ получаем

$$|P^m(z)| \leq \max_{\bar{G}_{\varphi_1}} |P^0(z)| + \sum_{k=1}^m 3M \cdot 2^{((2q+1) - (2q+2)\frac{\varphi_1}{\pi})m} = M(\varphi_1).$$

Величина $M(\varphi_1)$ конечна, поскольку при $\varphi_1 < \varphi_0$ коэффициент при m в показателе отрицателен. Таким образом, при $z \in \bar{G}_{\varphi_1}$ равномерно по m выполняется оценка $|P^m(z)| \leq M(\varphi_1)$. Напомним одно известное утверждение: пусть функции $f_m(z)$ аналитические в односвязной открытой области G и равномерно ограничены по модулю: $|f_m(z)| \leq C$, пусть последовательность значений $f_m(z)$ сходится на некотором множестве значений $z \in G$, имеющем предельную точку $z_0 \in G$, тогда последовательность $f_m(z)$ сходится к функции $f(z)$, аналитической в G , $|f(z)| \leq C$. В рассматриваемом случае это означает, что последовательность многочленов $P^m(z)$ сходится равномерно к некоторой функции $h(z)$, аналитической в G_{φ_1} . В свою очередь отсюда следует, что функция $f(x)$, определенная на $[-1, 1]$, продолжаема как

аналитическая во все точки открытой области G_{φ_1} . Вследствие произвольности $\varphi_1 > \varphi_0$ она также продолжается во все точки открытой области G_{φ_0} . Теорема доказана.

Какие практические следствия вытекают из этой теоремы? Если число a записывается в системе с плавающей запятой с t двоичными разрядами, то его погрешность может оказаться больше $|a|2^{-t-1}$; в худшем случае, когда погрешности коэффициентов a_j^m одного знака, погрешность значения

$$P^m(1) = \sum_{j=0}^{n(m)} a_j^m,$$

являющаяся следствием этих погрешностей, может превзойти величину

$$\left(\sum_{j=0}^{n(m)} |a_j| \right) 2^{-t-1} = l_m 2^{-t-1}.$$

Чтобы качественно представить характер влияния этой погрешности, будем считать, что число m , входящее в (2), (3), и число разрядов t велики. Пусть функция $f(x)$ аналитична в некоторой области G_{φ_0} и не аналитична ни в какой области G_{φ} при $\varphi < \varphi_0$. Пусть q_0 соответствует этому φ_0 , согласно (8), т. е. $q_0 = \frac{2\varphi_0 - \pi}{2(\pi - \varphi_0)}$. Возьмем произвольное $q < q_0$; если предположить, что для всех m выполняется неравенство $l_m \leq 2^{qm}$, то, согласно теореме, $f(z)$ будет аналитична в соответствующей области G_{φ} при $\varphi < \varphi_0$ и получается противоречие с предположением о свойствах функции $f(z)$. Таким образом, будут встречаться сколь угодно большие m , для которых $l_m \geq 2^{qm}$; следовательно, при неблагоприятном стечении обстоятельств ошибка от округления значений a_j^m может оказаться более $2^{qm}2^{-t-1}$. В то же время нельзя рассчитывать на оценку погрешности, лучшую (2). Для суммарной погрешности от замены $f(x)$ на $P^m(x)$ и от ошибок в коэффициентах многочлена мы не можем предложить оценки погрешности, лучшей $\varepsilon(m) = 2^{qm}2^{-t-1} + 2^{-m}$. Уравнение для точки экстремума функции $\varepsilon(m)$:

$$\varepsilon'_m(\bar{m}) = q \ln 2 \cdot 2^{q\bar{m}} 2^{-t-1} - \ln 2 \cdot 2^{-\bar{m}} = 0.$$

Отсюда $2^{\bar{m}} = \left(\frac{2}{q}\right)^{\frac{1}{q+1}} 2^{\frac{t}{q+1}}$; подставив это значение в выражение $\varepsilon(\bar{m})$, получим

$$\begin{aligned} \varepsilon(\bar{m}) &= \left(\frac{2}{q}\right)^{\frac{q}{q+1}} 2^{-\frac{t}{q+1}} 2^{-1} + \left(\frac{2}{q}\right)^{-\frac{1}{q+1}} 2^{-\frac{t}{q+1}} = \\ &= \frac{q+1}{2} \left(\frac{2}{q}\right)^{\frac{q}{q+1}} 2^{-\frac{t}{q+1}}. \end{aligned}$$

Следовательно, при вычислениях на машине с t разрядами нельзя надеяться на получение приближений к значениям функции существенно более, чем с $\frac{t}{q+1}$ разрядами. Например, для функции $f(x) = \frac{1}{1+25x^2}$, нижней гранью φ , при которых $f(z)$ аналитична в области G_φ , будет $\varphi^0 = 2 \operatorname{arctg} 0,2$; соответствующее значение $q+1 \approx 3,2$. Таким образом, на машине с 60 двоичными разрядами мы в лучшем случае можем рассчитывать на получение приближений к значениям функции примерно с 20 верными двоичными разрядами.

Эти рассуждения побуждают к отысканию иных простых форм записи и вычисления значений многочленов.

Можно привести следующий довод об исключительности подобной обстановки: обычно приходится приближать функции целые или такие, особенно-сти которых лежат далеко за единичным кругом; тогда из проведенных выше рассмотрений такие удручающие выводы не следуют. Формально это верно; однако такого рода функции часто, являясь целыми аналитическими, имеют очень быстрый рост при росте мнимой части у аргумента. Хотя для таких функций величины I_n остаются равномерно ограниченными по n , они могут принимать столь большие значения, что величина $|I_n|2^{-t-1}$ окажется существенно больше требуемой точности. Примером подобных функций может

служить функция $\varphi(x) = \int_0^1 \exp(i\lambda x t^2) dt$ в случае больших λ .

• Обратимся к форме записи многочленов наилучшего приближения в виде суммы значений многочленов Чебышева

$$P_n(x) = \sum_{j=0}^n d_j T_j(x); \quad (9)$$

имеем

$$\frac{2}{\pi} \int_{-1}^1 \frac{T_n(x) T_m(x)}{\sqrt{1-x^2}} dx = \begin{cases} 2 & \text{при } n=m=0, \\ \delta_{n-m} & \text{при } n^2+m^2 > 0. \end{cases}$$

Поэтому

$$\frac{2}{\pi} \int_{-1}^1 \frac{P_n^2(x)}{\sqrt{1-x^2}} dx = \frac{2}{\pi} \sum_{j=0}^n |d_j|^2 \int_{-1}^1 \frac{T_j^2(x)}{\sqrt{1-x^2}} dx = 2d_0^2 + \sum_{j=1}^n d_j^2 \geq \sum_{j=0}^n d_j^2.$$

Воспользуемся неравенством Коши — Буняковского

$$\sum_{j=0}^n |d_j| = \sum_{j=0}^n 1 \cdot |d_j| \leq \sqrt{\left(\sum_{j=0}^n 1\right) \cdot \left(\sum_{j=0}^n d_j^2\right)} = \sqrt{(n+1) \sum_{j=0}^n d_j^2}.$$

Отсюда заключаем

$$\sum_{j=0}^n |d_j| \leq \sqrt{\frac{2}{\pi}(n+1) \int_{-1}^1 \frac{P_n^2(x)}{\sqrt{1-x^2}} dx}.$$

Для последовательности аппроксимирующих многочленов

$$\frac{2}{\pi} \int_{-1}^1 \frac{P_n^2(x)}{\sqrt{1-x^2}} dx \rightarrow \frac{2}{\pi} \int_{-1}^1 \frac{f^2(x)}{\sqrt{1-x^2}} dx.$$

Таким образом, величина $\sum_{j=0}^n |d_j|$ растет не быстрее $\sqrt{n}$.

Поскольку $|T_n(x)| \leq 1$, то ошибка в значении многочлена, являющаяся следствием ошибок в a_j , не превзойдет величины порядка $\left(\sum_{j=0}^n |d_j|\right) 2^{-t} = O(\sqrt{n} 2^{-t})$. При реальных вычислениях, как правило, не используются приближения функций многочленами очень высоких степеней и поэтому такая оценка погрешности оказывается приемлемой. Конечно, не обязательно представлять многочлены, приближающие функцию в виде линейной комбинации (9) многочленов Чебышева. В зависимости от конкретной обстановки иногда удобнее представлять многочлен как линейную комбинацию многочленов какой-либо другой ортогональной системы, например, многочленов Лежандра. Мы рассмотрели систему многочленов Чебышева вследствие наибольшей простоты рекуррентных соотношений для вычисления их значений.

Остановимся на вопросе практического получения представления приближающего многочлена в форме (9). Можно с самого начала искать приближения к многочлену наилучшего приближения в виде

$$Q_n^k(x) = \sum_{j=0}^n d_j^k T_j(x)$$

и определять коэффициенты d_j^k из линейной системы уравнений, соответствующей (7.2):

$$f(x_i^k) - Q_n^k(x_i^k) = (-1)^i E^k,$$

пользуясь каким-либо устойчивым алгоритмом решения линейных алгебраических систем.

Другой путь, при котором многочлен наилучшего приближения получается в виде (1), мы уже описали. Однако не следует думать, что после получения многочлена в форме (1) можно записать многочлен в обычной форме, а затем перейти к представлению (9). Если величина A_n велика, то после записи в форме $a_0 + \dots + a_n x^n$ многочлен $P_n(x)$ приобретает

большую погрешность. По отношению к дальнейшим преобразованиям она является неустранимой, и поэтому представление многочлена в форме (9) также будет содержать эту погрешность. Переход от (1) к (9) можно осуществить непосредственно, минуя обычную форму записи, следующим образом. Вследствие свойства ортогональности системы функций $T_n(x)$ с весом $2/(\pi \sqrt{1-x^2})$ имеем

$$a_0 = \frac{1}{\pi} \int_{-1}^1 \frac{Q_n(x)}{\sqrt{1-x^2}} dx, \quad a_j = \frac{2}{\pi} \int_{-1}^1 \frac{Q_n(x) T_j(x)}{\sqrt{1-x^2}} dx.$$

В рассматриваемом случае $Q_n T_j(x)$ многочлены степени не выше $2n$. Поэтому формула Гаусса — Эрмита с $(n+1)$ -м узлом для вычисления этих интегралов будет точной:

$$a_0 = \frac{1}{n+1} \sum_{q=1}^{n+1} Q_n(y_{n+1}^q), \quad a_j = \frac{2}{n+1} \sum_{q=1}^{n+1} Q_n(y_{n+1}^q) T_j(y_{n+1}^q) \quad (10)$$

при $j > 0$; здесь

$$y_{n+1}^q = \cos \left(\pi \frac{2q-1}{2(n+1)} \right),$$

необходимые значения $Q_n(y_{n+1}^q)$ вычисляем по формуле (1).

Заметим, что в процессе итераций возможна и такая комбинированная процедура: по заданным узлам $x_0^k, \dots, x_{n+1}^k$ многочлены $Q_n^k(x)$ ищутся ранее описанным способом в виде (1). Затем при помощи соотношений (10) они представляются в виде (9), который используется с целью более экономного отыскания точек x_i^{k+1} .

Дадим окончательные рекомендации по вычислению значений многочлена $P_n(x)$ на $[-1, 1]$ безотносительно к вопросу о наилучшем приближении функций.

Пусть в обычной форме этот многочлен записывается в виде $\sum_{i=0}^n a_i x^i$ и погрешность порядка $\left(\sum_{i=0}^n |a_i| \right) 2^{-t}$ является допустимой. Тогда имеет смысл воспользоваться схемой Горнера

$$P_n(x) = a_0 + x(a_1 + x(\dots + x(a_{n-1} + x a_n) \dots)).$$

В противном случае этот многочлен следует получать в виде

$$P_n(x) = \sum_{j=0}^n d_j T_j(x).$$

Для уменьшения числа действий при вычислении значений $P_n(x)$ целесообразно поступить следующим образом: полагаем

$$D_0 = d_0 - \frac{d_2}{2}, \quad D_j = \frac{d_j - d_{j+2}}{2}$$

при $j = 1, \dots, n-2$, $D_{n-1} = \frac{d_{n-1}}{2}$, $v_n = \frac{d_n}{2}$.

При необходимости вычислить $P_n(x)$ находим $y = 2x$, $v_{n-1} = yv_n + D_{n-1}$, затем последовательно $v_{n-2}, \dots, v_0 = P_n(x)$ по рекуррентной формуле $v_k = yv_{k+1} - v_{k+2} + D_k$. Известно, что вычислительная погрешность такого алгоритма есть

$$O\left(n\left(\min\left(n, \frac{1}{\sqrt{1-x^2}}\right)\right)\left(\max_{[-1,1]}(P_n(x))\right)2^{-t}\right).$$

З а д а ч а. Проверить, что действительно $v_0 = P_n(x)$.

§ 9. О способах вычисления элементарных функций

Многочлены наилучшего приближения часто используются для приближения функций, возникающих из эксперимента или в результате сложных вычислений в случае, когда не имеется никакой дополнительной информации об их свойствах. Если такая информация имеется, то иногда целесообразнее другой вид приближающей функции.

Например, в случае элементарных функций имеется большая информация об их аналитических свойствах, позволяющая резко сократить объем вычислений.

Рассмотрим одну из стандартных программ вычисления $\ln x$. Число x представляется в виде $x = 2^p X$, где $\frac{1}{2} \leq X \leq 1$, т. е. p -двоичный порядок числа; далее $\ln x = p \ln 2 + \ln X$. Для вычисления $\ln X$ используется ряд

$$\ln X = 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \left(\frac{X-1}{X+1}\right)^{2k+1}.$$

При $\frac{1}{2} \leq X \leq 1$ его члены убывают быстрее 3^{-k} , в то время как для простейшего ряда

$$\ln X = \sum_{k=0}^{\infty} \frac{(-1)^{k-1}}{k} (X-1)^k$$

убывание порядка $2^{-k}/k$. Для ускорения сходимости полагают $X = y/\lambda$ и вычисляют $\ln y$ с помощью ряда

$$\ln y = 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \left(\frac{y-1}{y+1}\right)^{2k+1};$$

множитель λ подбирается из условия, чтобы величина $u = \max_{\frac{1}{2} \leq X \leq 1} \left| \frac{y-1}{y+1} \right|$

была наименьшей. Это дает $\lambda = \sqrt{2}$ и $u < 0,172$; расчетная формула приобретает вид

$$\ln x = \left(p - \frac{1}{2}\right) \ln 2 + 2v \sum_{k=0}^{\infty} \frac{1}{2k+1} v^{2k},$$

$$v = \frac{\sqrt{2} X - 1}{\sqrt{2} X + 1}.$$

Для уменьшения числа учитываемых слагаемых в выражении $\sum_{k=0}^{\infty} \frac{1}{2k+1} v^{2k}$ оно заменяется многочленом наилучшего приближения или близким к нему многочленом на соответствующем участке изменения v . В рассматриваемом случае формула

$$\ln x \approx (p - 0,5) \ln 2 + v(2,000000815 + 0,666445069v^2 + 0,415054254v^4)$$

имеет погрешность не более $3 \cdot 10^{-8}$.

В ряде машин в основу стандартных программ вычисления значений элементарных функций положены представления этих функций в виде цепных дробей или аппроксимация дробно-рациональными функциями. Например, функцию $\operatorname{tg} x$ можно вычислять, пользуясь представлением в виде цепной дроби

$$\operatorname{tg} x = \frac{x}{1 - \frac{x^2}{3 - \frac{x^2}{5 - \dots}}}$$

Другой распространенный способ — дробно-рациональное приближение вида

$$\frac{P_n(x) + Q_m(x)}{P_n(x) - Q_m(x)}.$$

Например, приближение

$$\exp x \approx \frac{12(x^2 + 10) + x(x^2 + 60)}{12(x^2 + 10) - x(x^2 + 60)}$$

обеспечивает точность $9 \cdot 10^{-9}$ при $|x| \leq \frac{\ln 2}{2}$.

Трудно представить все многообразие приемов, элементы которых используются при создании стандартных программ вычисления значений элементарных функций. В частности, используемые методы сильно зависят от разрядности машин, предъявляемых требований точности, длительности выполнения элементарных арифметических операций.

Рассмотрим один прием такого рода. Этот прием выделен в основном с целью еще раз обсудить вопрос о вычислительной погрешности

Простейшие элементарные функции удовлетворяют рекуррентным соотношениям вида

$$f(x) = F(f(G(x))), \quad (1)$$

где функции F и G легко вычисляются, $0 \leq G'(x)$, $0 \leq G(x) \leq qx$ при $0 \leq x \leq x_0$, $q < 1$. Воспользовавшись n раз этой записью, получим

$$f(x) = \underbrace{F(F(\dots(F(f(G(G(\dots(G(x))\dots))))))}_{n \text{ раз}}. \quad (2)$$

Положим

$$G_n(x) = \underbrace{G(G(\dots(G(x))\dots))}_{n \text{ раз}},$$

$$F_n(x) = \underbrace{F(F(\dots(F(x))\dots))}_{n \text{ раз}}.$$

При помощи записи (2) задача приближения $f(x)$ на отрезке $[0, x_0]$ сводится к задаче приближения на меньшем отрезке $[0, G_n(x_0)]$, где $G_n(x_0) \leq q^n x_0$. За счет малой длины отрезка высокая точность приближения достигается при значительно меньшей степени аппроксимирующего многочлена.

Укажем ряд соотношений вида (1)

$$\exp x = \left(\exp \frac{x}{2} \right)^2, \quad \cos x = 2 \left(\cos \frac{x}{2} \right)^2 - 1,$$

$$\sin x = 2 \left(\sin \frac{x}{2} \right) \sqrt{1 - \left(\sin \frac{x}{2} \right)^2}, \quad \operatorname{tg} x = \frac{2 \operatorname{tg} \frac{x}{2}}{1 - \left(\operatorname{tg} \frac{x}{2} \right)^2},$$

$$\operatorname{ctg} x = \frac{1}{2} \left(\operatorname{ctg} \frac{x}{2} - \frac{1}{\operatorname{ctg} \frac{x}{2}} \right), \quad \operatorname{arctg} x = 2 \operatorname{arctg} \frac{x}{1 + \sqrt{1 + x^2}},$$

$$\arcsin x = 2 \arcsin \frac{x}{\sqrt{2(1 + \sqrt{1 - x^2})}}.$$

Наиболее привлекательной по своей простоте выглядит рекуррентная формула для $\operatorname{ctg} x$. После отыскания $u = \operatorname{ctg} \frac{x}{2}$ можно вычислить

$$\sin x = \frac{2}{u + \frac{1}{u}}, \quad \cos x = \frac{u - \frac{1}{u}}{u + \frac{1}{u}}, \quad \operatorname{tg} x = \frac{2}{u - \frac{1}{u}}.$$

Некоторые из этих формул довольно сложны и их применение не всегда будет разумно. Из приведенных построений вытекает следующая схема вычисления значений функций. Находим некоторую аппроксимацию $f(x)$ на $[0, G_n(x_0)]$, например, многочленом или отношением двух многочленов. Последовательно вычисляем

$$G_1(x) = G(x), \quad G_2 = G(G_1(x)), \quad \dots, \quad G_n(x) = G(G_{n-1}(x)),$$

приближенное значение $f^* \approx f(G_n(x))$,

$$F_1(f^*) = F(f^*), \quad \dots, \quad F_n(f^*) = F(F_{n-1}(f^*))$$

и полагаем $f(x) \approx F_n(f^*)$.

Заданная. Предположим, что вычисления производятся без округлений и что на отрезке $[0, G_n(x_0)]$ функция $f(x)$ приближается отрезком ряда Тейлора из q членов. Доказать, что при малых ε можно выбрать взаимно согласованные n и q порядка $O(\sqrt{\ln(1/\varepsilon)})$, при которых погрешность приближений к значениям $f(x)$ на $[0, x_0]$ не превосходит ε .

На примере функции $\exp x$ обратим внимание на опасность нежелательного роста вычислительной погрешности при вычислении значения $F_n(f^*)$. При помощи равенства $\exp x = 2^p \exp X$, где

$$p = \left[\frac{x}{\ln 2} \right], \quad X = \left\{ \frac{x}{\ln 2} \right\} \ln 2,$$

задача сводится к вычислению значений функции $\exp x$ на отрезке $[0, \ln 2]$; здесь $\{ \}$ — знак дробной доли. Имеем равенство

$$\exp X = \left(\exp \frac{X}{2^n} \right)^{2^n} = \left(\dots \left(\exp \left(\frac{X}{2^n} \right) \right)^2 \dots \right)^2.$$

Как бы ни вычислялись значения $\exp \frac{X}{2^n}$, они будут округлены и их погрешность может превзойти число 2^{-t-1} (t — разрядность чисел в ЭВМ).

Вместо точного значения $\left(\exp \frac{X}{2^n} \right)^{2^n}$ мы получаем приближенное значение $\left(\left(\exp \frac{X}{2^n} \right) + \varepsilon \right)^{2^n}$. Согласно формуле Лагранжа $y_1^N - y_2^N = N \bar{y}^{N-1} (y_1 - y_2)$ где $\bar{y}$ лежит между y_1 и y_2 . Поэтому в рассматриваемом случае погрешность равна

$$2^n \varepsilon \left(\left(\exp \frac{X}{2^n} \right) + \bar{\varepsilon} \right)^{2^n-1}, \quad \text{где } |\bar{\varepsilon}| \leq \varepsilon.$$

Последний множитель не менее 1, и таким образом, погрешность может оказаться более 2^{n-t-1} . К стандартным программам обычно предъявляется требование точности значений функций порядка 2^{-t} . Поэтому при больших n такая величина погрешности является неприемлемой.

В рассматриваемом примере уже с первого взгляда видны резервы повышения точности.

Число $u_k = \left(\exp \frac{X}{2^n} \right)^{2^k}$ в двоичной системе записывается в виде $1, \underbrace{0 \dots 0}_{n-k \text{ раз}} \dots$. Таким образом, из t разрядов $n - k + 1$ тратятся на запоминание информации, которая и так известна. Поэтому есть надежда, заменив эту информацию на существенную, повысить точность вычислений. Чтобы избавиться от запоминания этих известных двоичных знаков, введем в рассмотрение функцию $v(X) = \exp X - 1$. Соотношение (1) для этой функции имеет вид $v(X) = v\left(\frac{X}{2}\right) \left(2 + v\left(\frac{X}{2}\right) \right)$. Значения $v\left(\frac{X}{2^n}\right)$ не превосходят 2^{1-n} . Если они будут известны с погрешностью $O(2^{-n-t})$, то погрешность значения $v(X)$ а, следовательно, $\exp X = 1 + v(X)$ окажется величиной порядка $O(n2^{-t})$. Мы предоставляем читателю проверку справедливости этого утверждения.

По той же причине непригодна рекуррентная формула $\cos x = 2 \left(\cos \frac{x}{2} \right)^2 - 1$. Вместо непосредственного вычисления по этой формуле следует положить $u(x) = 1 - \cos x$ и вычислять сначала значения функции $u(x)$, пользуясь рекуррентной формулой $u(x) = 2u\left(\frac{x}{2}\right) \left(2 - u\left(\frac{x}{2}\right) \right)$. В случае функции $\ln x$ вместо простейшего соотношения $\ln h = 2 \ln \sqrt{h}$ следует взять

$$\ln(1+x) = 2 \ln \left(1 + \frac{x}{1 + \sqrt{1+x}} \right).$$

Подобная потеря точности возникает всегда, если одно из чисел $f(0)$ (как в случае $\exp x$) или $\lim_{n \rightarrow \infty} G_n(x)$ (как в случае $\ln x$) конечно и отлично от нуля.

- Задачи. 1. Провести подобные построения в случае $f(x) = \arccos x$.
 2. Убедиться, что рекуррентные соотношения для вычисления

$$\sin x, \lg x, \operatorname{ctg} x, \operatorname{arctg} x, \operatorname{arcsin} x$$

не нуждаются в подобных преобразованиях.

3. Пусть значения элементарных функций вычисляются в машине с плавающей запятой с t двоичными разрядами, t — большое. Предположим, что порядок чисел, употребляемых в машине, не ограничен (таких машин не бывает, но ограничения на порядок чисел в современных ЭВМ таковы, что конечностью максимального порядка часто можно пренебречь). Доказать, что в случае аппроксимации $f(x)$ на $[0, G_n(x_0)]$ отрезком ряда Тейлора длины q , при подходящем выборе n и q (порядка $\sqrt{t}$) и подходящей форме записи соотношений (1), можно указать алгоритмы вычисления указанных выше элементарных функций со следующими характеристиками: объем вычислений $O(\sqrt{t})$, ошибка приближенных значений $O(\sqrt{t}/2^t)^*$.

§ 10. О скорости приближения функций различных классов

Приведем сведения о скорости приближения многочленами и дробно-рациональными функциями

$$R_n(x) = \frac{a_0 + \dots + a_n x^n}{b_0 + \dots + b_n x^n}$$

в зависимости от аналитических и дифференциальных свойств приближаемых функций.

Речь пойдет о приближении функций, заданных на $[-1, 1]$ многочленами и дробно-рациональными функциями, и приближении периодических с периодом 2π функций тригонометрическими многочленами и тригонометрическими дробями. Соответственно, определяем норму

$$\|f\|_1 = \sup_{[-1, 1]} |f(x)|,$$

$$\|f\|_2 = \sup_{[0, 2\pi]} |f(x)|.$$

В теоретических исследованиях обычно рассматривают следующие типичные классы функций: $C_r(M, [-1, 1])$, (r — целое) — класс функций, у которых на отрезке $[-1, 1]$ производные $f^{(l)}$, $l = 0, \dots, r-1$, непрерывны, производная $f^{(r)}$ кусочно-непрерывна и $|f^{(l)}| \leq M$ при $l = 0, \dots, r$; $H_r(M, 2\pi)$ — класс функций, имеющих период 2π и удовлетворяющих тем же условиям на производные; $A(M, G)$ — класс функций, допускающих аналитическое продолжение с отрезка $[-1, 1]$ в открытую односвязную область комплексного переменного G и удовлетворяющих в G условию $|f(z)| \leq M$; $A(M, h, 2\pi)$ — класс функций, периодических с периодом 2π , допускающих аналитическое продолжение в область $|\operatorname{Im} z| < h$ и удовлетворяющих там условию $|f(z)| \leq M$.

Через $\mathcal{E}_H$ будем обозначать эллипсы с суммой полуосей $H > 1$ и с фокусами в точках вещественной оси $[-1, 1]$. Положим $\alpha(G) = \sup_{\mathcal{E}_H \in G} H$,

*) Часть рассуждений этого параграфа принадлежит С. А. Усову.

иначе $\mathfrak{z}(G)$ — это сумма полюсов наибольшего из эллипсов $\mathfrak{E}_H$, принадлежащих замкнутой области G (рис. 4.10.1).

Существует одно и только одно конформное отображение с неподвижными точками -1 и 1 , при котором область G отображается на внутренность эллипса из семейства $\mathfrak{E}_H$. Значение H , соответствующее этому эллипсу, обозначаем $\beta(G)$.

Введем обозначения:

$t_n(f)$ — величина погрешности наилучшего равномерного приближения тригонометрическими многочленами

$$T_n(x) = a_0 + \sum_{j=1}^n (a_j \cos jx + b_j \sin jx),$$

$$t_n = \inf_{a_j, b_j} \|f - T_n\|_2;$$

$r_n(f)$ — величина погрешности наилучшего равномерного приближения рациональными функциями $R_n(x) = \frac{a_0 + \dots + a_n x^n}{1 + b_1 x + \dots + b_n x^n}$, $r_n(f) = \inf_{a_j, b_j} \|f - R_n\|_1$;

$e_n(f) = \|f - L_n^f\|_1$, где $L_n^f(x)$ — интерполяционный многочлен степени n с чебышевскими узлами интерполяции

$$\cos\left(\pi \frac{2m+1}{2n+2}\right), \quad m = 0, \dots, n;$$

$\tau_n(f) = \|f - T_n^f\|_2$, где T_n^f — интерполяционный тригонометрический многочлен степени n с узлами интерполяции

$$x_j = \frac{2\pi j}{2n+1}, \quad j = 0, \dots, 2n;$$

$\rho_n(f)$ — величина погрешности наилучшего равномерного приближения тригонометрическими дробями

$$R_n^T(x) = \frac{a_0 + \sum_{j=1}^n (a_j \cos jx + b_j \sin jx)}{1 + \sum_{j=1}^n (c_j \cos jx + d_j \sin jx)}.$$

Справедливы следующие оценки. Для функций классов $C_r(M, [-1, 1])$:

$$E_n(f) = O\left(\frac{1}{n^r}\right), \quad e_n(f) = O\left(\frac{\ln n}{n^r}\right), \quad r_n(f) = O\left(\frac{1}{n^r}\right);$$

для функций классов $H_r(M, 2\pi)$:

$$t_n(f) = O\left(\frac{1}{n^r}\right), \quad \tau_n(f) = O\left(\frac{\ln n}{n^r}\right), \quad \rho_n(f) = O\left(\frac{1}{n^r}\right);$$

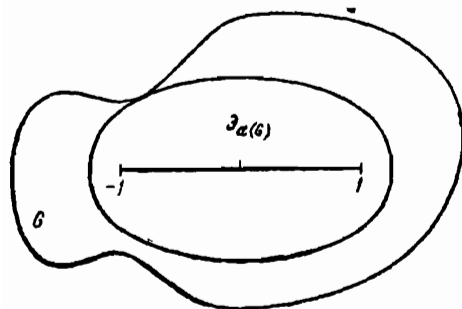


Рис. 4.10.1.

для функций классов $A(M, G)$.

$$E_n(f) = O(M(\alpha(G))^{-n}), \quad (1)$$

$$e_n(f) = O(M(\dot{\alpha}(G))^{-n}),$$

$$r_n(f) = O(M(\beta(G))^{-n});$$

для функций классов $A(M, h, 2\pi)$:

$$t_n(f) = O(Me^{-2hn}),$$

$$\tau_n(f) = O(Me^{-2hn}),$$

$$\rho_n(f) = O(Me^{-2hn}).$$

Аппроксимация тригонометрическими рациональными функциями

$$R_n^T(x) = \frac{a_0 + \sum_{j=1}^n (a_j \cos jx + b_j \sin jx)}{1 + \sum_{j=1}^n (c_j \cos jx + d_j \sin jx)}$$

бывает целесообразна, когда периодическая функция продолжается аналитически в область комплексного переменного, отличную от полосы $|\operatorname{Im} z| \leq h$. Большинство приближаемых функций оказываются аналитическими; из приведенных оценок видно, что наилучшее равномерное приближение таких функций не дает выигрыша по порядку погрешности по сравнению со случаем интерполяции по узлам Чебышева.

Если область аналитичности — один из эллипсов $\mathcal{E}_h$ или в периодическом случае полоса $|\operatorname{Im} z| \leq h$, то при наилучшем рациональном приближении, объем вычислений примерно вдвое больше, а выигрыша в порядке оценки нет. Если область аналитичности функции имеет существенно другой вид, то приближение рациональными или тригонометрическими рациональными функциями может дать выигрыш.

Особенно большой выигрыш получается в случае $\alpha(G) = 1$. Именно, пусть $f(x)$ аналитична во всех точках отрезка $[-1, 1]$, кроме конечного числа точек, и удовлетворяет условию Гельдера с некоторым $\lambda > 0$:

$$|f(x_1) - f(x_2)| \leq C |x_1 - x_2|^\lambda.$$

На рис. 4.10.2 приведен пример области аналитичности, соответствующей такому случаю. Тогда про величину $E_n(f)$ можно лишь сказать, что

$$E_n(f) = O(n^{-\sigma}). \quad \sigma = \sigma(\lambda) > 0.$$

В то же время

$$r_n(f) = O(e^{-c\sqrt[n]{n}}), \quad c > 0.$$

Наилучшее приближение рациональными функциями $R_n(x)$ практически получить трудно, так как $R_n(x)$ нелинейно зависит от большого числа параметров. Поэтому иногда целесообразно применять некоторые комбинированные способы, где число параметров, от которых функция зависит нелинейно,

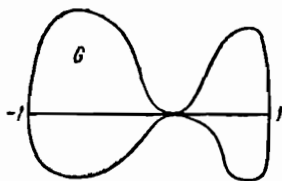


Рис. 4.10.2.

существенно меньше. Иногда целесообразно использовать приближения вида:

$$P_n(P_m(x)), \quad P_n, \quad P_m - \text{многочлены,}$$

$$\frac{P_n(x)}{(Q_m(x))^l}, \quad P_n \text{ и } Q_m - \text{многочлены, } n = ml,$$

$P_n(R_m(x)) - P_n$ — многочлен, R_m — рациональная дробь.

Погрешность наилучшего приближения такими функциями будем обозначать, соответственно,

$$E_{nm}(f), \quad r_{nm}(f), \quad e_{nm}(f).$$

Рассмотрим случай, когда граница области G — аналитическая кривая. Тогда для функций классов $A(M, G)$ при $\alpha(G) > 1$ справедливы оценки:

$$\text{при } m \geq \gamma(G) \ln n$$

$$e_{nm}(f) \leq E_{nm}(f) = O(M\beta(G)^{-n});$$

$$\text{при } m \geq \eta(G) \ln n$$

$$r_{nm}(f) = O(M\beta(G)^{-n}).$$

Если $\alpha(G) = 1$, но все точки интервала $(-1, 1)$ являются точками аналитичности (рис. 4.10.3), то при $m \geq \chi(G) \ln n$ справедлива оценка

$$r_{nm}(f) = O(M\delta(G)^{-V\sqrt{n}}), \quad \delta(G) > 1.$$

Рис. 4.10.3.

Если $\alpha(G) = 1$, но все точки интервала $(-1, 1)$ являются точками аналитичности (рис. 4.10.3), то при $m \geq \chi(G) \ln n$ справедлива оценка

§ 11. Интерполяция и приближение сплайнами

Способ приближения сплайнами интересен прежде всего отношением к нему. Одни считают его универсальным методом решения проблем, стоящих перед численным анализом, и ищут применения ему в самых различных направлениях, другие рассматривают его как очередную дань переменчивой математической моде. По-видимому, истина проходит где-то посередине; в настоящее время область применения этого метода непрерывно растет.

Для определенности будем говорить о приближении функции $f(x)$ на отрезке $[0, 1]$. Отрезок разбивается на части:

$$[x_0, x_1], \dots, [x_{N-1}, x_N], \quad x_0 = 0, \quad x_N = 1.$$

Обозначим это разбиение через Δ . Назовем *сплайном* $S_\Delta^m(f, x)$ порядка m функцию, являющуюся многочленом степени m на каждом из отрезков $[x_{n-1}, x_n]$:

$$S_\Delta^m(f, x) = P_{nm}(x) = a_{n0} + \dots + a_{nm}x^m \quad \text{при } x_{n-1} \leq x \leq x_n, \quad (1)$$

и удовлетворяющую условиям непрерывности производных до порядка $m-1$ в точках $x_1, \dots, x_{N-1}$:

$$P_{nm}^{(k)}(x_n) = P_{n+1, m}^{(k)}(x_n) \quad (2)$$

$$\text{при } k = 0, \dots, m-1, \quad n = 1, \dots, N-1.$$

Всего имеется в распоряжении $Q = N(m+1)$ неизвестных a_{nm} , соотношения (2) образуют систему из $(N-1)m$ уравнений. Другие уравнения для коэффициентов получаются из условия близости к приближаемой функции и из некоторых дополнительных условий.

Рассмотрим простейшую задачу приближения линейными сплайнами ($m=1$). Тогда общее число Q свободных параметров равно $2N$. Поставим вопрос построения сплайна $S_\Delta^1(f, x)$, совпадающего с $f(x)$ в точках $x_0, \dots, x_N$; получится система уравнений

$$P_{n1}(x_{n-1}) = f(x_{n-1}), \quad n = 1, \dots, N,$$

$$P_{n1}(x_n) = f(x_n), \quad n = 1, \dots, N.$$

Эта система распадается на системы уравнений относительно коэффициентов отдельных многочленов

$$P_{n1}(x_{n-1}) = a_{n0} + a_{n1}x_{n-1} = f(x_{n-1}),$$

$$P_{n1}(x_n) = a_{n0} + a_{n1}x_n = f(x_n);$$

отсюда находим

$$a_{n1} = (f(x_n) - f(x_{n-1})) / (x_n - x_{n-1}),$$

$$a_{n0} = f(x_{n-1}) - a_{n1}x_{n-1};$$

многочлен $P_{n1}(x)$ является многократно рассматривавшимся интерполяционным многочленом первой степени с узлами интерполяции x_{n-1}, x_n .

Этот сплайн $S_\Delta^1(f, x)$ возникает также из рассмотрения следующей задачи. Рассмотрим множество S_1 непрерывных функций $s(x)$, удовлетворяющих условиям

$$s(x_n) = f(x_n), \quad n = 0, \dots, N, \quad I_1(s) = \int_0^1 (s'(x))^2 dx < \infty.$$

Поставим задачу найти функцию $s_1(x) \in S_1$, реализующую

$$\inf_{s \in S_1} I_1(s).$$

Мы ищем функцию с наименьшей нормой $\|s'\|_{L_2}$, т. е. в каком-то смысле наиболее гладкую функцию семейства S_1 , совпадающую с $f(x)$ в точках x_n . Уравнение Эйлера для этого функционала имеет вид $s''(x) = 0$. Таким образом, $s_1(x)$ линейна на каждом из отрезков $[x_{n-1}, x_n]$ и, следовательно,

$$s_1(x) = S_\Delta^1(f, x).$$

Этот подход к приближению функции, при котором естественным образом возник сплайн $S_\Delta^1(f, x)$, может быть развит в следующем направлении.

Приближающая функция ищется среди пространства функций большей гладкости. Рассмотрим, например, множество S_2 функций $s(x)$ с непрерывной первой производной, удовлетворяющих условиям

$$s(x_n) = f(x_n), \quad n = 0, \dots, N, \quad I_2(s) = \int_0^1 (s''(x))^2 dx < \infty.$$

Ищется функция $s_2(x) \in S_2$, реализующая $\inf_{s \in S_2} I_2(s)$. Уравнение Эйлера для этого функционала $s^{(4)}(x) = 0$, т. е. $s_2(x) = P_{n3}(x) = a_{n0} + a_{n1}x + a_{n2}x^2 + a_{n3}x^3$ на отрезках $[x_{n-1}, x_n]$. Число параметров a_{nk} , подлежащих определению, равно $4N$, число условий $3N - 1$:

$2N$ условий

$$P_{n3}(x_n) = P_{n+1,3}(x_n) = f(x_n) \quad (3)$$

и $N - 1$ условие непрерывности первой производной

$$P'_{n3}(x_n) = P'_{n+1,3}(x_n) \quad (4)$$

в точках $x_1, \dots, x_{N-1}$. Недостающие условия получаются как естественные краевые условия в точках x_n .

Чтобы не отсылать к соответствующему разделу вариационного исчисления, выведем эти условия независимо. Предположим, что минимизирующая функция $s_2(x)$ существует и является многочленом третьей степени на отрезках $[x_{n-1}, x_n]$. Пусть $\eta(x)$ бесконечно дифференцируемая,

$$\eta(x_0) = \dots = \eta(x_N) = 0. \quad (5)$$

Положим

$$\begin{aligned} F(t) &= \int_0^1 (s_2''(x) + t\eta''(x))^2 dx = \\ &= \int_0^1 (s_2''(x))^2 dx + 2t \int_0^1 s_2''(x) \eta''(x) dx + t^2 \int_0^1 (\eta''(x))^2 dx. \end{aligned}$$

Поскольку, по предположению,

$$F(t) \geq F(0), \quad \text{то} \quad F'(0) = 2 \int_0^1 s_2''(x) \eta''(x) dx = 0.$$

Представим величину $B = F'(0)/2$ в виде $B = b_1 + \dots + b_N$,

$$b_n = \int_{x_{n-1}}^{x_n} s_2''(x) \eta''(x) dx = \int_{x_{n-1}}^{x_n} P'_{n3}(x) \eta''(x) dx,$$

и произведем в выражении для b_n дважды интегрирование по частям. Получим

$$b_n = \int_{x_{n-1}}^{x_n} P_{n3}^{(4)}(x) \eta(x) dx + (P_{n3}''(x) \eta'(x) - P_{n3}'''(x) \eta(x)) \Big|_{x_{n-1}}^{x_n}.$$

Вследствие (5) и равенства $P_{n3}^{(4)}(x) = 0$ имеем

$$b_n = P_{n3}''(x) \eta'(x) \Big|_{x_{n-1}}^{x_n}.$$

После суммирования по n получим

$$B = P_{N3}''(1) \eta'(1) + \sum_{n=1}^{N-1} (P_{n3}''(x_n) - P_{n+1,3}''(x_n)) \eta'(x_n) - P_{13}''(0) \eta'(0).$$

Заметим, что мы не имели права производить интегрирование по частям непосредственно в исходном интеграле B , поскольку неясен вопрос о дифференцируемости $s_2(x)$ в точках $x_1, \dots, x_{N-1}$.

Так как $B = 0$ при произвольной $\eta(x)$, удовлетворяющей (5), то отсюда получаются недостающие $N+1$ уравнения

$$P_{13}''(0) = P_{N3}''(1) = 0, \quad P_{n3}''(x_n) - P_{n+1,3}''(x_n) = 0, \quad (6) \\ n = 1, \dots, N-1.$$

Рассмотрим вопрос о разрешимости и о практическом решении системы уравнений (3), (4), (6). Для удобства введем в рассмотрение величины $M_n = s_2''(x_n)$. Поскольку функция $s_2''(x)$ линейна на $[x_{n-1}, x_n]$, то

$$s_2''(x) = M_{n-1} \frac{x_n - x}{h_n} + M_n \frac{x - x_{n-1}}{h_n} \quad \text{на } [x_{n-1}, x_n];$$

здесь $h_n = x_n - x_{n-1}$. Из этого соотношения и условий

$$s_2(x_{n-1}) = f(x_{n-1}), \quad s_2(x_n) = f(x_n)$$

можно получить, что

$$s_2(x) = P_{n3}(x) = M_{n-1} \frac{(x_n - x)^3}{6h_n} + M_n \frac{(x - x_{n-1})^3}{6h_n} + \\ + \left(f(x_{n-1}) - \frac{M_{n-1} h_n^2}{6} \right) \frac{x_n - x}{h_n} + \\ + \left(f(x_n) - \frac{M_n h_n^2}{6} \right) \frac{x - x_{n-1}}{h_n} \quad \text{на } [x_{n-1}, x_n]. \quad (7)$$

Условия

$$P_{n3}'(x_n) = P_{n+1,3}'(x_n), \quad n = 1, \dots, N-1,$$

образуют уравнения

$$\begin{aligned} \frac{h_n}{6} M_{n-1} + \frac{h_n + h_{n+1}}{3} M_n + \frac{h_{n+1}}{6} M_{n+1} = \\ = \frac{f(x_{n+1}) - f(x_n)}{h_{n+1}} - \frac{f(x_n) - f(x_{n-1}))}{h_n}. \end{aligned} \quad (8)$$

Кроме того, имеем условия

$$P''_{13}(x_0) = 0, \quad P''_{N3}(x_N) = 0,$$

иначе $M_0 = 0$, $M_N = 0$. После подстановки $M_0 = 0$ и $M_N = 0$, соответственно, в первое и последнее уравнение (7) получим систему

$$CM = d \quad (9)$$

$N - 1$ -го уравнения с $N - 1$ -м неизвестным:

$$M = (M_1, \dots, M_{N-1})^T, \quad d = (d_1, \dots, d_{N-1})^T.$$

Элементы c_{ij} , $i, j = 1, \dots, N - 1$ матрицы C , согласно (8), задаются соотношениями

$$c_{ij} = \begin{cases} \frac{h_i}{6} & \text{при } j = i - 1, \\ \frac{h_i + h_{i+1}}{3} & \text{при } j = i, \\ \frac{h_{i+1}}{6} & \text{при } j = i + 1, \\ 0 & \text{при } |i - j| > 1, \end{cases}$$

а элементы d_i столбца d соотношениями

$$d_i = \frac{f(x_{i+1}) - f(x_i)}{h_{i+1}} - \frac{f(x_i) - f(x_{i-1}))}{h_i}.$$

В главе IX будет показано, что эта система решается «методом прогонки» за $O(N)$ арифметических действий. После нахождения M_j по формуле (7) определяем многочлены $P_{n3}(x)$.

Задачи. 1. Доказать, что решение системы (9) удовлетворяет неравенству

$$\max_{1 \leq n \leq N-1} |M_n| \leq \frac{3}{\min_{1 \leq n \leq N} h_n} \max_{1 \leq n \leq N-1} |d_n|;$$

получить отсюда однозначную разрешимость системы (3), (4), (6).

2. Пусть $S_\Delta^2(f, x)$ — сплайн, соответствующий решению системы (9). Доказать, что

$$I_2(f) = I_2(S_\Delta^2) + I_2(f - S_\Delta^2).$$

Из справедливости этого утверждения вытекает единственность минимизирующей функции $s_2(x)$.

При рассматриваемом подходе получаемый сплайн совпадает с $f(x)$ во всех узлах и естественно назвать такие сплайны *интерполяционными*.

Часто известна дополнительная информация о поведении $f(x)$ в точках $x = 0$, $x = 1$. Например, может случиться, что конечной целью является решение некоторого дифференциального уравнения с граничными условиями

$$f'(0) + \alpha f(0) = \beta, \quad f'(1) + \gamma f(1) = \delta.$$

Тогда вместо естественных граничных условий, порождающих уравнения $M_0 = M_N = 0$, надо поставить эти граничные условия.

Рассмотрим случай, когда значения функции $f(x_n)$ заданы с погрешностями. Существует ряд естественных подходов к приближению таких функций, порождающих сплайны.

а) Пусть известно, что заданные значения f_n и точные значения $f(x_n)$ удовлетворяют соотношениям

$$|f_n - f(x_n)| \leq \delta.$$

Можно поставить задачу следующим образом. Найти непрерывно дифферен-

цируемую функцию $s(x)$ с $\int_0^1 (s'')^2 dx < \infty$, реализующую $\inf I_2(s)$ в классе функций, удовлетворяющих условиям $|s(x_n) - f_n| \leq \delta$. Практическое получение решения этой задачи довольно трудоемко. Поэтому иногда удобнее поставить задачу отыскания $\inf I_2(s)$ среди функций, удовлетворяющих условию:

$$\sum_{n=0}^N (s(x_n) - f_n)^2 \leq c(N+1)\delta^2 \quad (10)$$

б) Известно, что разности $f_n - f(x_n)$ являются случайными величинами с дисперсией δ . В этом случае также можно минимизировать $I_2(s)$ при условии (10); $c = O(1)$ выбирают исходя из задаваемого доверительного уровня.

Часто левая и правая части (10) равны в точке минимума $I_2(s)$. Поэтому обычно вместо исходной задачи рассматривают задачу безусловной минимизации по λ и s функционала

$$J_2(\lambda, s) = I_2(s) + \lambda \left(\sum_{n=0}^N (s(x_n) - f_n)^2 - c(N+1)\delta^2 \right).$$

Рис. 4.11.1 соответствует некоторому конкретному случаю приближения при $N = 100$ и $c = 0,9$. При каждом фиксированном λ величина

$$J(\lambda) = \inf_s J_2(\lambda, s)$$

определяется с помощью решения системы линейных уравнений. Для нахождения $\inf_{\lambda} J(\lambda)$ применяется какой-либо метод минимизации функций одной переменной (см. гл. VII).

Иногда (из предшествующего опыта или из соображений здравого смысла) известно примерное значение λ , соответствующее точке безусловного экстремума, $\lambda \approx \lambda_0$; и тогда ограничиваются минимизацией функционала $J_2(\lambda_0, s)$.

в) Еще один подход к приближению сплайнами состоит в следующем. Заданы приближения f_j к значениям функции $f(z_j)$, $j = 1, \dots, M$. Исходя

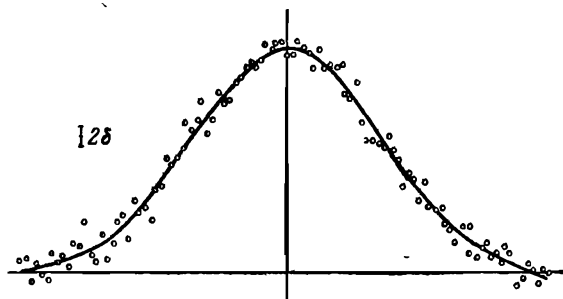


Рис. 4.11.1.

из какой-то априорной информации о приближаемой функции, разбивают отрезок $[0, 1]$ на части $[x_0, x_1], \dots, [x_{N-1}, x_N]$, где $N \ll M$. Приближение отыскивается в виде сплайна $S_\Delta^m(j, x)$, минимизирующего функционал

$$\sum_{j=1}^M p_j (s(x_j) - f(x_j))^2.$$

§ 12. Энтропия и ε -энтропия

Условно можно говорить, что всякий объект задается двоякого рода информацией:

- 1) информацией о некотором множестве, которому он принадлежит, и
- 2) информацией, указывающей данный объект среди элементов множества.

Например, конкретный многочлен степени m задается: 1) информацией о том, что это многочлен степени m , 2) таблицей его коэффициентов.

Нашей целью явится обсуждение вопроса о минимальном количестве информации, необходимой для задания элементов множеств. Вообще говоря, для задачи приближения функций более традиционным и приспособленным к реальным вычислениям является способ задания функции при помощи таблицы ее значений. Однако рассмотрение вопроса именно о минимальном количестве информации позволяет выработать более общий взгляд на рассматриваемую проблему.

Пусть множество G , элементы которого требуется задавать, состоит из N различных элементов $g_1, \dots, g_N$, и пусть эти элементы задаются совокупностью из n двоичных знаков

$y_1, \dots, y_n = 0, 1$. Для возможности однозначного задания $g_k \in G$ необходимо, чтобы число элементов N не превосходило числа таких совокупностей. Всего таких совокупностей 2^n ; поэтому необходимым условием однозначности является $N \leq 2^n$. С другой стороны, при $N \leq 2^n$ мы можем поставить в соответствие каждой совокупности $y_1, \dots, y_n$ натуральное число

$$j = 1 + \sum_{i=1}^n 2^{n-i} y_i. \quad (1)$$

При изменении $y_1, \dots, y_n$ число j пробегает все значения от 1 до 2^n ; элемент g_j будем задавать совокупностью $y_1, \dots, y_n$, определяющей j , согласно (1). Величину $\log_2 N$ называют *энтропией* $H(G)$ множества G . Минимальное количество двоичных знаков, необходимое для однозначного задания элементов множества G , будем обозначать $I(G)$; через $\bar{y}$ будем обозначать наименьшее целое, не меньшее y . Мы видели, что минимальное количество двоичных знаков, нужное для задания элементов G , равно наименьшему целому n такому, что $N \leq 2^n$, следовательно,

$$I(G) = \overline{\log_2 N} = \overline{H(G)}. \quad (2)$$

Задачи. 1. Доказать неравенства

$$I\left(\bigcup_{i=1}^q G_i\right) \leq \overline{\log_2 q} + \max_{1 \leq i \leq q} I(G_i), \quad (3)$$

$$I(G_1 \times \dots \times G_q) \leq \sum_{i=1}^q I(G_i), \quad (4)$$

здесь $\times$ — знак прямого произведения, т. е. элементами $G_1 \times \dots \times G_q$ являются совокупности $(g^1, \dots, g^q)$ при $g^i \in G_i$.

2. Привести конкретные примеры, где в (3), (4) имеют место равенство или строгое неравенство.

3. Загадывается целое число от 0 до 1000. Отгадывающий может задавать вопросы, на которые будет получать ответы «да» или «нет». Доказать, что 10 вопросов достаточно для угадывания, а 9 нет.

4. Имеется N ящиков, одинаковых по виду, из них $N-1$, равных по весу, а один тяжелее остальных. Разрешается производить n взвешиваний на весах без гирь. Доказать, что при $N \leq 3^n$ можно гарантировать отыскание более тяжелого ящика, а при $N > 3^n$ нет.

5. Имеется N ящиков, одинаковых по виду, любые два из которых различны по весу. Оценить порядок минимального числа n взвешиваний на весах без гирь, при котором возможно упорядочивание ящиков по весу.

Ответ. $n \sim \log_2 N!$ при $N \rightarrow \infty$.

Пусть теперь G состоит из бесконечного множества элементов. В частности, в соответствии с (2) никакого фиксированного конечного числа двоичных знаков недостаточно для однозначного задания элементов из G .

Однако можно ставить вопрос о минимальном количестве информации, необходимом для задания элементов из G с точностью ε в метрике некоторого метрического пространства F , которому принадлежит G . Предполагается следующий способ задания элементов $g \in G$. Элемент g определяется некоторой совокупностью из двоичных знаков $y_1, \dots, y_n$; совокупности $y_1, \dots, y_n$ поставлен в соответствие элемент $f_{y_1, \dots, y_n} \in F$, для которого $\rho(f_{y_1, \dots, y_n}, g) \leq \varepsilon$. Через $J_\varepsilon(G; F)$ будем обозначать минимальное количество двоичных знаков, достаточное для задания элементов из G таким способом. Пусть Φ — метрическое расширение F , т. е. метрическое пространство, содержащее F своим подмножеством, такое, что для любой пары элементов $f_1, f_2 \in F$ выполняется равенство

$$\rho_F(f_1, f_2) = \rho_\Phi(f_1, f_2);$$

здесь нижний индекс означает пространство, которому соответствует метрика. Поскольку всякий способ приближения элементов g элементами из F является также способом приближения элементами из Φ , то

$$J_\varepsilon(G; \Phi) \leq J_\varepsilon(G; F). \quad (5)$$

В дальнейшем считаем пространство F фиксированным и иногда опускаем в обозначениях соответствующий индекс.

Лемма 1. Пусть элементы $f_1, \dots, f_N \in F$ образуют ε -сеть для множества G . Тогда

$$J_\varepsilon(G; F) \leq \overline{\log_2 N}.$$

Доказательство. В соответствии с условием леммы каждому элементу $g \in G$ можно поставить в соответствие некоторый элемент $f_j \in F$, для которого $\rho(g, f_j) \leq \varepsilon$. Пусть $y_1, \dots, y_n$ — совокупность двоичных символов, соответствующих j , согласно (1). Задание совокупности $y_1, \dots, y_n$ определяет элемент f_j , для которого $\rho(g, f_j) \leq \varepsilon$.

Лемма 2. Пусть $G' \subset G$. Тогда $J_\varepsilon(G'; \Phi) \leq J_\varepsilon(G; \Phi)$ при любом Φ .

Доказательство очевидно.

Множество элементов G' называют η -различимым, если расстояние между двумя любыми его элементами более η .

Лемма 3. Пусть $G' \subset G$, G' — 2ε -различно и состоит из N элементов. Тогда

$$\overline{\log_2 N} \leq J_\varepsilon(G; \Phi),$$

где Φ — любое расширение F .

Доказательство. Для доказательства достаточно показать, что

$$\overline{\log_2 N} \leq J_\varepsilon(G'; \Phi). \quad (6)$$

Тогда, воспользовавшись леммой 2, получим требуемое неравенство. Согласно определению способа задания функций g , неравенство $J_\varepsilon(G'; \Phi) \leq n$ означает следующее. Существует $M \leq 2^n$ элементов $\varphi_1, \dots, \varphi_M \in \Phi$, удовлетворяющих условию: любому $g \in G'$ соответствует некоторое φ_j , для которого $\rho_\Phi(g, \varphi_j) \leq \varepsilon$. Различным $g_i, g_k \in G'$ не может соответствовать один и тот же элемент φ_j ; предположив противное, получим противоречие, вследствие цепочки неравенств:

$$2\varepsilon < \rho_F(g_i, g_k) = \rho_\Phi(g_i, g_k) \leq \rho_\Phi(g_i, \varphi_j) + \rho_\Phi(g_k, \varphi_j) \leq 2\varepsilon.$$

Таким образом, число элементов N множества G' не более M , т. е.

$$N \leq M \leq 2^n = 2^{J_\varepsilon(G'; \Phi)}.$$

Мы получили (6) и доказали лемму 3.

Введем в рассмотрение величину

$$J_\varepsilon(G) = \inf_{\Phi} J_\varepsilon(G; \Phi);$$

нижняя грань берется по множеству всевозможных метрических расширений Φ пространства F .

Беря нижнюю грань по Φ в утверждениях лемм 1—3, получаем, соответственно:

- 1) $J_\varepsilon(G) \leq \log_2 N,$
- 2) $J_\varepsilon(G') \leq J_\varepsilon(G),$
- 3) $\overline{\log_2 N} \leq J_\varepsilon(G).$

При помощи этих лемм и следствий можно получать оценки сверху и снизу величин $J_\varepsilon(G; F)$.

ε -энтропией $H_\varepsilon(G)$ множества G называется $\log_2 N_\varepsilon(G)$, где $N_\varepsilon(G)$ — число элементов покрытия G минимальным числом множеств диаметра ε . Если G — компакт, то известно, что

$$H_\varepsilon(G) \leq J_\varepsilon(G) \leq [H_\varepsilon(G)] + 1,$$

где нижняя грань берется по всевозможным метрическим расширениям Φ .

Рассмотрим примеры.

1. F — вещественная ось, G — множество точек

$$x \in [0, 1], \quad \rho(x_1, x_2) = |x_1 - x_2|.$$

Положим $N_\varepsilon = \left(\frac{1}{2\varepsilon}\right)$. Точки $\varepsilon, 3\varepsilon, \dots, (2N_\varepsilon - 1)\varepsilon$ образуют ε -сеть. Поэтому

$$J_\varepsilon(G; F) \leq \overline{\log_2 N_\varepsilon}.$$

Согласно определению N_ε , имеем $(N_\varepsilon - 1)2\varepsilon < 1$, поэтому $\varepsilon' = \frac{1}{2}(N_\varepsilon - 1)^{-1} > \varepsilon$. N_ε точек $0, 2\varepsilon', \dots, 1$ образуют $2\varepsilon'$ -различимое множество и, согласно лемме 3,

$$J_\varepsilon(G; \Phi) \geq \overline{\log_2 N_\varepsilon}.$$

В итоге мы получили равенство

$$J_\varepsilon(G) = J_\varepsilon(G; F) = \overline{\log_2 \left(\frac{1}{2\varepsilon}\right)}.$$

2. Задача. F — s -мерное евклидово пространство точек $P(x_1, \dots, x_s)$, $\rho(P_1, P_2)$ равно некоторой норме вектора $P_1 P_2$, G — множество точек, принадлежащее шару $\rho(P, 0) \leq A_2 < \infty$ и содержащее внутри себя шар $\rho(P, 0) \leq A_1 \neq 0$. Доказать асимптотическое равенство

$$J_\varepsilon(G) = s \log_2 \frac{1}{\varepsilon} + O(1).$$

3. G — множество непрерывных функций $g(x)$, определенных на $[0, 1]$ и удовлетворяющих условию $g(\cdot) = 0$ и условию $|g(x_2) - g(x_1)| \leq |x_2 - x_1|$. Положим

$$\rho(f_1, f_2) = \sup_{[0, 1]} |f_1(t) - f_2(t)|.$$

Наиболее широкое пространство, в котором можно искать приближения к функциям рассматриваемого класса, — это пространство всевозможных функций, определенных на $[0, 1]$ Обозначим его через F .

Положим, $n = \left(\frac{1}{\varepsilon}\right) - 1$, что означает $n\varepsilon < 1 \leq (n+1)\varepsilon$.

Покажем, что $J_\varepsilon(G, F) = n$.

Сначала получим оценку $J_\varepsilon(G; F) \geq n$. Очевидно, $\varepsilon' = \frac{1}{n} > \varepsilon$; поставим в соответствие каждой совокупности двоичных знаков $y_1, \dots, y_n$ непрерывную функцию $g_{y_1, \dots, y_n}(x)$, определяемую соотношениями

$$1) \ g_{y_1, \dots, y_n}(0) = 0,$$

$$2) \text{ на отрезке } [(j-1)\varepsilon', j\varepsilon'] \ g_{y_1, \dots, y_n}(x) \text{ линейная, причем}$$

$$(g_{y_1, \dots, y_n}(x))' = \begin{cases} 1, & \text{если } y_j = 1, \\ -1, & \text{если } y_j = 0. \end{cases} \quad (7)$$

Поскольку производная $(g_{y_1, \dots, y_n}(x))'$ кусочно-непрерывна, то

$$g_{y_1, \dots, y_n}(x_2) - g_{y_1, \dots, y_n}(x_1) = \int_{x_1}^{x_2} (g_{y_1, \dots, y_n}(x))' dx;$$

воспользовавшись (7), получим

$$|g_{y_1, \dots, y_n}(x_2) - g_{y_1, \dots, y_n}(x_1)| \leq |x_2 - x_1|.$$

Итак, функции $g_{y_1, \dots, y_n}(x)$ принадлежат рассматриваемому классу. Возьмем две различные совокупности

$$(y_1^1, \dots, y_n^1) \quad \text{и} \quad (y_1^2, \dots, y_n^2).$$

Очевидно, существует $j < n$, для которого $y_i^1 = y_i^2$ при $i < j$, а $y_j^1 \neq y_j^2$. Функции $g_{y_1^1, \dots, y_n^1}(x)$, $l = 1, 2$, обращаются в 0 при $x = 0$; при $0 \leq x \leq (j-1)e'$

их производные совпадают вследствие (7). Поэтому они совпадают всюду при $0 \leq x \leq (j-1)e'$, в частности, в точке $(j-1)e'$. На отрезке $[(j-1)e', je']$, согласно (7), производная разности этих функций постоянна и равна по модулю 2. Поэтому

$$g_{y_1^1, \dots, y_n^1}(je') - g_{y_1^2, \dots, y_n^2}(je') = \int_{(j-1)e'}^{je'} (\pm 2) dx = \pm 2e'.$$

Это означает, что

$$\rho(g_{y_1^1, \dots, y_n^1}, g_{y_1^2, \dots, y_n^2}) \geq 2e' > 2e.$$

Таким образом, мы имеем в распоряжении 2^n $2e$ -различимых функций. Следовательно,

$$n \leq J_e(G; F). \quad (8)$$

Перейдем к оценке сверху. Положим $e'' = \frac{1}{n+1}$ и рассмотрим всевозможные непрерывные функции $f_{y_1, \dots, y_n}(x)$, $y_1, \dots, y_n = 0, 1$, определяемые соотношениями:

$$f_{y_1, \dots, y_n}(x) = 0 \quad \text{при} \quad 0 \leq x \leq e'';$$

на отрезках $[(j-1)e'', je'']$ функция линейная, причем

$$(f_{y_1, \dots, y_n}(x))' = \begin{cases} 1, & \text{если } y_{j-1} = 1, \\ -1, & \text{если } y_{j-1} = 0. \end{cases}$$

Поскольку $e'' + ne'' = 1$, то эти соотношения определяют $f_{y_1, \dots, y_n}(x)$ на всем отрезке $[0, 1]$. Покажем, что множество этих функций образует e'' -сеть. С этой целью для каждой функции $g(x) \in G$ укажем некоторую функцию $f_{y_1, \dots, y_n}(x)$, удовлетворяющую условию $\rho(g, f_{y_1, \dots, y_n}) \leq e''$. Будем определять значения $y_1, \dots, y_n$, соответствующие такой функции, последовательно: заметим, что значения $f_{y_1, \dots, y_n}(x)$ на $[0, je'']$ определяются только значениями $y_1, \dots, y_{j-1}$. Пусть мы уже указали $y_1, \dots, y_{j-1}$ такие, что

$$|g(x) - f_{y_1, \dots, y_n}(x)| \leq e'' \quad (9)$$

для $0 \leq x \leq je''$. При $j = 1$ указывать такие значения не требуется, поскольку для всех функций рассматриваемого класса на $[0, e'']$ имеем

$$|g(x) - f_{y_1, \dots, y_n}(x)| = |g(x)| = |g(x) - g(0)| \leq e'' \quad \text{при } 0 \leq x \leq e''.$$

Таким образом, исходное предположение для проведения построения по индукции выполнено.

Покажем, что при выборе

$$y_j = \begin{cases} 1, & \text{если } g((j+1)e'') \geq f_{y_1, \dots, y_{j-1}, \dots, y_n}(je''), \\ 0, & \text{если } g((j+1)e'') < f_{y_1, \dots, y_{j-1}, \dots, y_n}(je''), \end{cases}$$

соотношение (9) выполняется и на отрезке $[je'', (j+1)e'']$.

Рассмотрим случай, когда y_j положено равным 1; мы можем написать соотношение

$$|g(x) - g((j+1)e'')| \leq |x - (j+1)e''|.$$

Отсюда

$$g(x) \geq g((j+1)e'') + (x - (j+1)e'') \geq \{f_{y_1, \dots, y_{j-1}, \dots, y_n}(je'') + (x - je'')\} - e''$$

при $je'' \leq x \leq (j+1)e''$. Выражение в фигурных скобках, по определению есть $f_{y_1, \dots, y_j, \dots, y_n}(x)$; таким образом, мы получили

$$g(x) \geq f_{y_1, \dots, y_j, \dots, y_n}(x) - e'' \quad \text{при } je'' \leq x \leq (j+1)e''.$$

С другой стороны, имеем

$$g(x) \leq g(je'') + (x - je''),$$

подставляя $x = je''$ в (9), получим

$$|f_{y_1, \dots, y_{j-1}, \dots, y_n}(je'') - g(je'')| \leq e'',$$

следовательно,

$$g(x) \leq (f_{y_1, \dots, y_{j-1}, \dots, y_n}(je'') + (x - je'')) + e''$$

и

$$g(x) \leq f_{y_1, \dots, y_j, \dots, y_n}(x) + e'' \quad \text{при } je'' \leq x \leq (j+1)e'';$$

в итоге имеем

$$|g(x) - f_{y_1, \dots, y_j, \dots, y_n}(x)| \leq e'' \quad \text{на } [je'', (j+1)e''].$$

Аналогично разбирается случай, когда $y_j = 0$. Мы продвинулись еще на один шаг в построении требуемой функции $f_{y_1, \dots, y_n}(x)$. Таким образом мы сможем дойти до конца отрезка.

Так как $e'' = \frac{1}{n+1} \leq e$, то построенное множество из 2^n функций образует e -сеть. Эти функции принадлежат исходному множеству G . Поэтому на основании леммы 1 имеем

$$J_e(G; F) \leq n. \quad (10)$$

Объединяя эту оценку с (8), получаем

$$J_e(G; F) = n = \frac{1}{e} - 1.$$

Заметим, что элементы построенной выше e -сети принадлежали самому исходному множеству G .

Задачи. 1. Пользуясь этим замечанием, показать, что в рассматриваемом случае

$$J_e(G) = n.$$

2. G — пространство функций с $(r-1)$ -й непрерывной производной, определенных на $[0, a]$ и удовлетворяющих условиям

$$f(0) = \dots = f^{(r-1)}(0) = 0, \quad |f^{(r-1)}(x_2) - f^{(r-1)}(x_1)| \leq \\ \leq A |x_2 - x_1| \quad \text{при} \quad x_1, x_2 \in [0, a];$$

пусть пространство F и метрика те же, что и в предшествующем примере. Величина $J_\varepsilon(G; F)$ является функцией от параметров ε, A, a . Доказать, что она является функцией от безразмерной комбинации $a \sqrt[r]{A/\varepsilon}$, т. е.

$$J_\varepsilon(G; F) = \psi \left(a \sqrt[r]{A/\varepsilon} \right).$$

3. Доказать, что

$$\ln \psi(y) = \ln y + O(1)$$

при $y \rightarrow \infty$.

4. G — пространство аналитических функций $A(M, \Omega)$,

$$\rho(f, g) = \sup_{|z|=1} |f - g|,$$

$\alpha(\Omega) > 1$ (определение $A(M, \Omega)$ см. в § 10). Доказать, что

$$\log_2 H_\varepsilon(G) \sim 2 \log_2 \log_2 \frac{1}{\varepsilon}.$$

Дополнительная литература к главе IV

§§ 3, 4: II [12], §§ 5—7: II [38], § 10: II [11, 43], III [23]; § 11: II [1], III [63]; § 12: II [4, 6], III [30].

МНОГОМЕРНЫЕ ЗАДАЧИ

Вопросы приближения, интерполирования, численного интегрирования и дифференцирования функций одной переменной, как видно из предшествующего, разработаны достаточно подробно; на основе результатов теоретических исследований в настоящее время созданы довольно развитые системы стандартных программ решения одномерных задач. Значительная часть результатов теоретических исследований для одномерного случая переносится автоматически на случай функций двух и более переменных, однако при этом часто появляются практически недостаточно эффективные методы.

При теоретических построениях задачи разбивают на классы, например, класс задач вычисления интегралов от функций с ограниченными производными, и затем проводят исследования, связанные с этими классами задач. Конечно, принимаемое описание не всегда (скорее даже редко) хорошо описывает класс реально встречающихся задач, однако скорости современных ЭВМ таковы, что, несмотря на грубость описания одномерных задач, удастся решать большинство из них, пользуясь стандартными методами, разработанными в результате теоретических исследований.

В случае двух и большего числа переменных принимаемое в теоретических исследованиях описание задач также часто оказывается неудачным. Трудоемкость решения задач резко возрастает с ростом размерности решаемых задач, и поэтому, как правило, не удастся разработать стандартные методы решения широких классов многомерных задач со столь же высокой точностью, как в одномерном случае.

Некоторым положительным фактором является следующее обстоятельство. Многомерные математические задачи обычно возникают из описания сложных процессов. Обычно эти описания уже являются довольно грубыми и поэтому значительно реже предъявляется требование решения этих задач с такой же высокой точностью, как в одномерном случае. Например, требо-

вания к точности решения уравнений газовой динамики существенно ниже требований, предъявляемых к решению уравнений баллистики и небесной механики.

Не следует думать, что решение многомерных задач является почти безнадежным делом. В настоящее время основное отличие от одномерного случая, пожалуй, состоит в том, что многообразие многомерных задач требует привлечения к их исследованию специалистов более высокой квалификации. Изучение узких классов многомерных задач и повышение скоростей ЭВМ, конечно, будут иметь следствием создание стандартных методов решения задач определенных классов и снижение требований к квалификации.

§ 1. Метод неопределенных коэффициентов

Решение ряда многомерных задач часто сводится к решению следующих элементарных задач.

В некоторой области s -мерного пространства G заданы точки $P_1, \dots, P_N$ и значения функции f в этих точках. Требуется:

- 1) получить приближение к значению функции $f(P)$;
- 2) получить приближение к значению некоторой производной $\mathcal{D}f$ функции в точке P ;
- 3) вычислить интеграл

$$I(f) = \int_G f(P) p(P) dP;$$

здесь $p(P)$ — весовая функция.

Простейшим способом решения этих задач является способ *неопределенных коэффициентов*. Пусть из каких-то соображений нам известно, что функция $f(P)$ хорошо приближается линейными комбинациями

$$\sum_{j=1}^N B_j \omega_j(P).$$

Потребуем, чтобы такая линейная комбинация совпадала с $f(P)$ в заданных узлах:

$$\sum_{j=1}^N B_j \omega_j(P_q) = f(P_q), \quad q = 1, \dots, N. \quad (1)$$

Предположим, оказалось $\det \|\omega_j(P_q)\| \neq 0$; тогда матрица $\|\omega_j(P_q)\|$ имеет обратную $A = \|a_{jq}\|$ и решение (1) записывается в виде

$$B_j = \sum_{q=1}^N a_{jq} f(P_q). \quad (2)$$

Функция

$$g(P) = \sum_{j=1}^N B_j \omega_j(P)$$

совпадает с $f(P)$ в точках P_q . Подставляя B_j из (2), получим иное представление $g(P)$:

$$g(P) = \sum_{q=1}^N z_q(P) f(P_q), \quad \text{где} \quad z_q(P) = \sum_{j=1}^N a_{jq} \omega_j(P). \quad (3)$$

Такая форма записи интерполирующей функции является аналогом записи интерполяционного многочлена в форме Лагранжа. Как и в одномерном случае, можно надеяться, что при удачном выборе узлов P_q и функций $\omega_j(P)$ будет мала погрешность в приближенных равенствах:

$$f(P) \approx g(P) = \sum_{q=1}^N z_q(P) f(P_q), \quad (4)$$

$$\mathcal{D}f(P) \approx \mathcal{D}g(P) = \sum_{q=1}^N \mathcal{D}z_q(P) f(P_q), \quad (5)$$

$$I(f) \approx I(g) = \sum_{q=1}^N C_q f(P_q), \quad (6)$$

где

$$C_q = I(z_q);$$

здесь $\mathcal{D}$ — некоторый оператор дифференцирования.

Как отмечалось в главе II для одномерного случая, при неудачном выборе большого числа узлов интерполирования погрешность приближенного равенства (4) может оказаться катастрофически большой. Поскольку приближенные равенства (5), (6) являются следствием приближенного равенства (4), то может быть большой и ошибка в приближенных равенствах (5), (6). Поэтому применение соотношений (4), (5), (6) требует обоснования и выяснения условий законности использования.

§ 2. Метод наименьших квадратов

При построении $g(P)$ мы добивались равенства с функцией $f(P)$ в N точках. Функция $g(P)$ содержит большое число параметров, и поэтому у нее есть определенная свобода, чтобы сильно отличаться от $f(P)$ вне этих точек.

Повышение качества приближения может достигаться различными способами. Рассмотрим первый из них, называемый *методом наименьших квадратов*. Приближение ищется в виде

$$g(P) = \sum_{j=1}^n B_j \omega_j(P),$$

где $n \ll N$; параметры B_l определяются из условия

$$\min_{B_1, \dots, B_n} \Phi(g),$$

где, например,

$$\begin{aligned} \Phi(g) = \Phi(B_1, \dots, B_n) &= \sum_{q=1}^N p_q (g(P_q) - f(P_q))^2 = \\ &= \sum_{q=1}^N p_q \left(\sum_{j=1}^n B_j \omega_j(P_q) - f(P_q) \right)^2. \end{aligned} \quad (1)$$

В основе метода наименьших квадратов лежит следующее соображение. Малость величины $\Phi(g)$ обеспечивает близость $g(P)$ и $f(P)$ в точках P_q ; при $n \ll N$ функция $g(P)$ содержит относительно мало параметров и поэтому у нее меньше возможностей отличаться от $f(P)$ вне узлов по сравнению со случаем $n = N$.

Числа $p_q > 0$, называемые весами, подбирают в зависимости от плотности распределения точек P_q ; если значения $f(P_q)$ содержат случайную ошибку, то их выбирают также в зависимости от дисперсии ошибок измеряемых значений. Там, где точки P_q распределены плотнее, числа p_q берутся меньше; значениям $f(P_q)$ с большей дисперсией ошибки ставят в соответствие меньшие значения p_q . Наши рекомендации выглядят довольно неопределенными, поскольку нельзя предложить общего правила, пригодного для всех задач. Для конкретных классов задач принципы выбора p_q и n вырабатываются с учетом специфических свойств задач на основе статистических критериев и численного эксперимента.

Приравнивая нулю производные $\frac{\partial \Phi}{\partial B_l}$, получим систему линейных уравнений для определения B_j :

$$\begin{aligned} \frac{1}{2} \frac{\partial \Phi}{\partial B_l} &= \sum_{j=1}^n d_{lj} B_j - d_l = 0, \\ d_{lj} &= \sum_{q=1}^N p_q \omega_l(P_q) \omega_j(P_q), \\ d_l &= \sum_{q=1}^N p_q \omega_l(P_q) f(P_q). \end{aligned} \quad (2)$$

Числа B_j можно находить, непосредственно решая указанную систему или минимизируя каким-то образом функцию Φ . Преобразуем (1)

$$\Phi(B_1, \dots, B_n) = \sum_{i,j=1}^n d_{ij} B_i B_j - 2 \sum_{j=1}^n d_j B_j + d_0,$$

где

$$d_0 = \sum_{q=1}^N p_q (f(P_q))^2.$$

Так как $\Phi(B_1, \dots, B_n) \geq 0$, то квадратичная форма $\sum_{i,j=1}^n d_{ij} B_i B_j$ не отрицательна; это свойство используется в ряде стандартных программ метода наименьших квадратов.

Выражая B_j из (2) через d_i , а затем через $f(P_q)$, получим

$$B_j = \sum_{q=1}^N a_{jq} f(P_q),$$

следовательно,

$$g(P) = \sum_{q=1}^N z_q(P) f(P_q), \quad (3)$$

где

$$z_q(P) = \sum_{i=1}^n a_{iq} \omega_i(P).$$

Воспользовавшись (3), можно получить также формулу численного дифференцирования $\mathcal{D}f \approx \mathcal{D}g$ и квадратурную формулу

$$I(f) \approx I(g).$$

§ 3. Метод регуляризации

В основе *метода регуляризации* непосредственно лежат соображения о сглаживании аппроксимирующей функции. Наиболее распространенной формой метода регуляризации является следующая. Приближение отыскивается в виде

$$g(P) = \sum_{j=1}^n B_j \omega_j(P),$$

коэффициенты B_j выбираются из условия минимума выражения

$$\Phi(\lambda, g) = \Phi(g) + \lambda \Psi(g), \quad \lambda > 0. \quad (1)$$

Функционал $\Psi(g)$ подбирается из следующего условия: если значение этого функционала невелико, то функция g обладает определенной гладкостью. Например, $\Psi(g)$ может быть некоторым приближением к интегралу $\int |\text{grad } g(P)|^2 dP$. В приложениях распространен случай $n = \overset{G}{N}$, на котором мы далее и остановимся. Пусть минимум $\Phi(\lambda, g)$ достигается при некоторых $B_1^\lambda, \dots, B_N^\lambda$ и

$$g^\lambda(P) = \sum_{j=1}^N B_j^\lambda \omega_j(P).$$

Рассмотрим крайние случаи: $\lambda = 0$ и λ очень большое. Имеем равенство

$$\Phi(0, g) = \sum_{j=1}^N p_q \left(\sum_{j=1}^N B_j \omega_j(P_q) - \hat{f}(P_q) \right)^2.$$

Если $\det \|\omega_j(P_q)\| \neq 0$, система (1.1) имеет решение, и на ее решении правая часть этого равенства обращается в нуль. В то же время выражение $\Phi(0, g)$ всегда неотрицательно; таким образом, нижняя грань достигается на значениях B_j , являющихся решениями (1.1). Тогда $g^0(P)$ совпадает с интерполяционным многочленом с узлами интерполяции P_j . При больших λ в функционале $\Phi(\lambda, g)$ определяющим является второе слагаемое, нижняя грань которого достигается на гладкой функции. Следовательно, есть какие-то основания ожидать, что при промежуточных значениях λ функции $g^\lambda(P)$ будут гладкими и в то же время не очень сильно отличающимися от приближаемой функции в заданных узлах.

Наши рассуждения выглядят довольно расплывчато, однако при такой общей постановке задачи приближения функции без конкретного указания системы функций ω_j , распределения точек P_j и класса рассматриваемых функций вряд ли можно сказать что-либо более определенное. Подобные грубые «физические» соображения часто помогают при конструировании новых методов решения задачи в условиях недостаточной информированности о самой задаче.

Ниже на конкретном примере будет объяснено существо эффекта, достигаемого за счет применения регуляризации.

§ 4. Пример регуляризации

Пусть $f(x)$ — действительная, периодическая функция с периодом 1; известны некоторые приближенные значения f_q величин $f(x_q)$ при $x_q = qh$, $q = 0, \dots, N-1$; $Nh = 1$. Ошибки приближенных значений $f_q - f(x_q) = d_q$ — независимые случайные величины с математическим ожиданием, равным нулю, и дисперсией d^2 . Требуется получить таблицу приближенных значений производных $f'(x_q)$. Погрешность оценивается в норме, соответствующей скалярному произведению

$$(f, g) = \left(\sum_{q=0}^{N-1} f_q \bar{g}_q \right) h.$$

Далее f_q рассматривается как периодически продолженная.

Для нахождения производной используем простейшую формулу численного дифференцирования

$$f'(x_q) \approx \mathcal{D}f_q = \frac{f_{q+1} - f_{q-1}}{2h}. \quad (1)$$

Предположим, что функция $f(x)$ достаточно гладкая и шаг h настолько мал, что при отсутствии ошибок в значениях функции погрешность формулы численного дифференцирования

$$f'(x_q) \approx \frac{f(x_{q+1}) - f(x_{q-1}))}{2h} \quad (2)$$

пренебрежимо мала. Ошибка в значении производной представляема в виде

$$\begin{aligned} R_q &= \frac{f_{q+1} - f_{q-1}}{2h} - f'(x_q) = \\ &= \left(\frac{f(x_{q+1}) - f(x_{q-1}))}{2h} - f'(x_q) \right) + \frac{d_{q+1} - d_{q-1}}{2h} \approx \frac{d_{q+1} - d_{q-1}}{2h} = r_q. \end{aligned}$$

Здесь и далее знак приближенного равенства означает равенство с точностью до величин порядка погрешности формулы (2). Имеем равенство

$$\begin{aligned} M \|R\|^2 &\approx M \|r\|^2 = M \left(\sum_{q=0}^{N-1} \left(\frac{d_{q+1} - d_{q-1}}{2h} \right)^2 h \right) = \\ &= \frac{N}{4} \sum_{q=0}^{N-1} (M d_{q+1}^2 - 2M(d_{q-1}d_{q+1}) + M d_{q-1}^2). \end{aligned}$$

Согласно указанным выше свойствам случайных величин d_q , имеем

$$M d_q d_p = d^2 \delta_{p-q}. \quad (3)$$

Тогда получаем $M \|R\|^2 = N d^2 / 2$, иначе среднеквадратичное значение нормы ошибки равно $N d / \sqrt{2}$. Мы видим, что погрешность приближенной формулы (1) растет с уменьшением $h = 1/N$; в то же время для малости погрешности в предположении отсутствия ошибок в исходной информации (2) требуется достаточная малость h .

Чтобы выяснить пути уменьшения погрешности, произведем более детальное исследование. Пусть A_j^0 — дискретные коэффициенты Фурье (д. к. Ф.) функции $f(x_q)$:

$$f(x_q) = \sum_{-N/2 < j \leq N/2} A_j^0 \exp(2\pi i j x_q),$$

A_j — д. к. Ф. функции f_q . Величины $a_j = A_j - A_j^0$ будут д. к. Ф. функции $d_q = f_q - f(x_q)$. Имеем

$$\begin{aligned} d_q &= \sum_{-N/2 < j \leq N/2} a_j \exp(2\pi i j x_q), \\ a_j &= h \sum_{q=0}^{N-1} d_q \exp(-2\pi i j x_q). \end{aligned} \quad (4)$$

Чтобы вычислить д. к. Ф. для r_q , применим оператор $\mathcal{D}$ к функции $\exp(2\pi i j x_q)$:

$$\mathcal{D} \exp(2\pi i j x_q) = \frac{\exp(2\pi i j x_{q+1}) - \exp(2\pi i j x_{q-1})}{2h} = \frac{i \sin 2\pi j h}{h} \exp(2\pi i j x_q).$$

Применяя оператор $\mathcal{D}$ к обеим частям (4), получим

$$r_q = \sum_{-N/2 < j \leq N/2} \frac{i \sin 2\pi j h}{h} a_j \exp(2\pi i j x_q).$$

Воспользуемся равенством Парсеваля:

$$\|r\|^2 = \sum_{-N/2 < j \leq N/2} \left(\frac{\sin 2\pi j h}{h} \right)^2 |a_j|^2.$$

Отсюда следует, что

$$M \|r\|^2 = \sum_{-N/2 < j \leq N/2} \left(\frac{\sin 2\pi j h}{h} \right)^2 M |a_j|^2.$$

Имеем:

$$|a_j|^2 = a_j \bar{a}_j = h^2 \sum_{q, p=0}^{N-1} d_q d_p \exp(2\pi i j (x_p - x_q));$$

мы заменили $\bar{d}_p$ на d_p , воспользовавшись действительностью ошибок d_p . Таким образом,

$$M |a_j|^2 = h^2 \sum_{q, p=0}^{N-1} \exp(2\pi i j (x_p - x_q)) M (d_q d_p) = h^2 \sum_{q, p=0}^{N-1} d^2 \delta_{p-q} = h d^2.$$

Таким образом,

$$M \|r\|^2 = \sum_{-N/2 < j \leq N/2} \left(\frac{\sin 2\pi j h}{h} \right)^2 h d^2. \quad (5)$$

Возьмем некоторую функцию $\mu(j)$ и положим

$$f^\mu(x) = \sum_{-N/2 < j \leq N/2} A_j^\mu(j) \exp(2\pi i j x),$$

$$f_q^\mu = \sum_{-N/2 < j \leq N/2} A_j^\mu(j) \exp(2\pi i j x_q).$$

При заданных значениях f_q можно вычислить коэффициенты A_j , а следовательно, и величины f_q^μ . Пусть r_q^μ — составляющая погрешности формулы численного дифференцирования, являющаяся следствием наличия ошибок d_q :

$$r_q^\mu = \mathcal{D} f_q^\mu - \mathcal{D} f^\mu(x_q).$$

Аналогично (5) имеем

$$M \|r^\mu\|^2 = \sum_{-N/2 < j \leq N/2} \left(\frac{\sin 2\pi j h}{h} \right)^2 \mu^2(j) h d^2;$$

если

$$\mathcal{D} f(x_q) \approx \mathcal{D} f^\mu(x_q), \quad M \|r^\mu\|^2 \ll M \|r\|^2, \quad (6)$$

то имеет смысл положить $f'(x_q) \approx \mathcal{D} f_q^\mu$. Для выполнения первого из соотношений (6) существенна близость коэффициентов Фурье функций $f(x)$ и $f^\mu(x)$; для выполнения второго — малость $\mu(j)$ при больших j .

Попробуем прибегнуть к сглаживанию функции f_q при помощи метода регуляризации. Рассмотрим функционал

$$\Phi_1^h(\lambda, g) = h \sum_{q=0}^{N-1} (g_q - f_q)^2 + \lambda^2 h \sum_{q=0}^{N-1} \left(\frac{g_{q+1} - g_q}{h} \right)^2,$$

где $g_N = g_0$. При $g_q = g(x_q)$, $g(x_q)$ гладкой, величина

$$h \sum_{q=0}^{N-1} \left(\frac{g_{q+1} - g_q}{h} \right)^2$$

стремится к интегралу $\int_0^1 (g'(x))^2 dx$; условие, что этот интеграл

невелик, гарантирует определенную гладкость функции $g(x)$. Таким образом, есть какие-то основания считать, что функционал $\Phi_1^h(\lambda, g)$ удовлетворяет требованиям, накладываемым на функционалы метода регуляризации. Определим сеточную функцию g_q^λ из условия минимума функционала $\Phi_1^h(\lambda, g)$ и положим

$$f'(x_q) \approx \frac{g_{q+1}^\lambda - g_{q-1}^\lambda}{2h}.$$

Посмотрим, что является аналогом такой регуляризации для случая функций непрерывного аргумента. Пусть

$$f(x) = \sum_{j=-\infty}^{\infty} A_j^0 \exp(2\pi i j x),$$

$g_1^\lambda(x)$ — функция, реализующая минимум функционала

$$\Phi_1(\lambda, g) = \int_0^1 (g(x) - f(x))^2 dx + \lambda^2 \int_0^1 (g'(x))^2 dx.$$

Уравнение Эйлера для этого функционала

$$\lambda^2 g'' - (g - f) = 0.$$

Непосредственной проверкой убеждаемся, что функция

$$g_1^\lambda(x) = \sum_{j=-\infty}^{\infty} \frac{A_j^0}{1 + (\lambda 2\pi j)^2} \exp(2\pi i j x)$$

является решением этого уравнения. Произведем сравнение $g_1^\lambda(x)$ и $f(x)$ для малых λ . Если $\lambda 2\pi j \ll 1$, то коэффициенты

Фурье этих функций $\frac{A_j^0}{1 + (\lambda 2\pi j)^2}$ и A_j^0 близки между собой. Если $\lambda 2\pi j \gg 1$, то коэффициенты Фурье функции $g_1^\lambda(x)$ много

меньше коэффициентов Фурье функции $f(x)$. Таким образом, регуляризация равносильна на языке техники некоторой «филтрации»: несущественно искажая гармоники с малой частотой колебания, она сильно ослабляет гармоники с большой частотой. Если требуется еще меньше исказить амплитуды A_j^0 гармоник с малой частотой и сильнее «отфильтровать» высокочастотные колебания, то следует рассмотреть функционал

$$\Phi_n(\lambda, g) = \int_0^1 (g(x) - f(x))^2 dx + \lambda^{2n} \int_0^1 (g^{(n)}(x))^2 dx.$$

Уравнение Эйлера для этого функционала

$$\lambda^{2n} g^{(2n)} + (-1)^n (g - f) = 0,$$

отсюда

$$g_n^\lambda(x) = \sum_{j=-\infty}^{\infty} \frac{A_j^0}{1 + (\lambda 2\pi j)^{2n}} \exp(2\pi i j x).$$

Рассмотрим графики множителей $\frac{1}{1 + (\lambda 2\pi j)^{2n}}$; при $|j| < \frac{1}{2\pi\lambda}$ и $n \rightarrow \infty$ эти множители стремятся к 1 и, таким образом, при больших n амплитуды соответствующих гармоник искажаются все

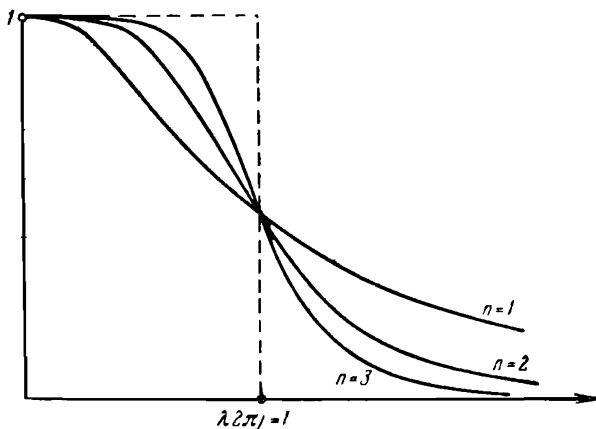


Рис. 5.4.1.

меньше и меньше. В то же время при $|j| > \frac{1}{2\pi\lambda}$ эти множители стремятся к 0 и, таким образом, соответствующие гармоники умножаются на все меньшие множители (рис. 5.4.1). Таким образом, при $n \rightarrow \infty$ и $\frac{1}{2\pi\lambda}$ не целом

$$g_n^\lambda(x) \rightarrow \tilde{g}^\lambda(x) = \sum_{|j| < 1/(2\pi\lambda)} A_j^0 \exp(2\pi i j x).$$

Вернемся к дискретному случаю. В выражение $\Phi_1^h(\lambda, g)$ величина g_q входит только в слагаемые

$$h(g_q - f_q)^2 + \lambda^2 h \left(\left(\frac{g_{q+1} - g_q}{h} \right)^2 + \left(\frac{g_q - g_{q-1}}{h} \right)^2 \right).$$

Имеем систему уравнений относительно значений g_q в точке экстремума

$$\frac{1}{2h} \frac{\partial \Phi_1^h}{\partial g_q} = (g_q - f_q) - \frac{\lambda^2}{h^2} (g_{q+1} - 2g_q + g_{q-1}) = 0. \quad (7)$$

Матрица системы (7) относительно неизвестных g_q совпадает с матрицей положительно определенной квадратичной формы

$$\sum_{q=0}^{N-1} g_q^2 + \frac{\lambda^2}{h^2} \sum_{q=0}^{N-1} (g_{q+1} - g_q)^2;$$

поэтому определитель ее отличен от нуля и система (7) имеет, и притом единственное, решение.

Пусть

$$f_q = \sum_{-N/2 < j \leq N/2} A_j \exp(2\pi i j x_q);$$

будем отыскивать периодическую функцию g_q^λ в виде

$$g_q^\lambda = \sum_{-N/2 < j \leq N/2} A_j^\lambda \exp(2\pi i j x_q).$$

Сначала вычислим оператор второй разности от функции $\exp(2\pi i j x_q)$:

$$\begin{aligned} \delta^2 \exp(2\pi i j x_q) &= \exp(2\pi i j x_{q+1}) - 2 \exp(2\pi i j x_q) + \exp(2\pi i j x_{q-1}) = \\ &= (\exp(2\pi i j h) - 2 + \exp(-2\pi i j h)) \exp(2\pi i j x_q) = \\ &= (2 \cos(2\pi j h) - 2) \exp(2\pi i j x_q), \end{aligned}$$

откуда

$$\delta^2 \exp(2\pi i j x_q) = -4 \sin^2(\pi j h) \exp(2\pi i j x_q). \quad (8)$$

Подставим представление f_q и g_q в виде суммы Фурье в (7), преобразуем вторую разность $\delta^2 g_q$ с учетом (8) и приведем подобные члены:

$$\sum_{-N/2 < j \leq N/2} \left(A_j^\lambda - A_j + 4\lambda^2 \left(\frac{\sin \pi j h}{h} \right)^2 A_j^\lambda \right) \exp(2\pi i j x_q) = 0.$$

Это равенство будет удовлетворяться при

$$A_j^\lambda = \frac{A_j}{1 + 4\lambda^2 \left(\frac{\sin \pi j h}{h} \right)^2}.$$

Таким образом,

$$g_q^\lambda = \sum_{-N/2 < j \leq N/2} \frac{A_j}{1 + 4\lambda^2 \left(\frac{\sin \pi j h}{h} \right)^2} \exp(2\pi i j x_q).$$

Множитель

$$\frac{1}{1 + 4\lambda^2 \left(\frac{\sin \pi j h}{h} \right)^2}$$

убывает с ростом j . Мы видим, что регуляризация с использованием функционала $\Phi_1^h(\lambda, g)$ также приводит к ослаблению высших гармоник функции. По аналогии с функционалами $\Phi_n(\lambda, g)$ можно осуществлять регуляризацию с помощью функционалов

$$\Phi_n^h(\lambda, g) = h \sum_{q=0}^{N-1} (g_q - f_q)^2 + \lambda^{2n} h \sum_{q=0}^{N-1} \left(\frac{\Delta^n g_q}{h^n} \right)^2.$$

Решение системы (7) и систем линейных уравнений, возникающих при минимизации функционалов $\Phi_n^h(\lambda, g)$, можно получить за $O(n^2 N)$ арифметических операций «методом прогонки» решения периодической сеточной краевой задачи (§ 3 гл. IX). По сравнению с регуляризацией с помощью функционалов $\Phi_n^h(\lambda, g)$ при $n > 1$ часто более удобна n -кратная регуляризация с помощью функционала $\Phi_1^h(\lambda, g)$: гармоники $\exp(2\pi i j x)$ в этом случае умножаются на

$$\frac{1}{\left(1 + 4\lambda^2 \left(\frac{\sin \pi j h}{h} \right)^2 \right)^n}.$$

Преимущество такой последовательной регуляризации состоит в простоте алгоритма, несколько меньшем объеме вычислений (порядка $O(nN)$) и меньшем влиянии вычислительной погрешности.

В случае рассматриваемого примера можно пойти и по такому пути регуляризации: произвести дискретное преобразование Фурье (если возможно — быстрое), найти коэффициенты Фурье A_j , заменить нулями A_j , соответствующие гармоникам с высокой частотой, и затем произвести обратное преобразование Фурье.

§ 5. Сведение многомерных задач к одномерным

Выше рассматривались способы решения многомерных задач, в которых не используется дополнительная информация о распределении точек, где известны значения функции.

Во многих случаях расположение этих точек носит регулярный характер; если рассматриваемая функция задана

аналитически, эти точки можно выбирать по желанию. При удачном расположении точек осуществление операций приближения, интерполирования, численного дифференцирования и интегрирования функций многих переменных может быть сведено к последовательному осуществлению этих операций над функциями одной переменной.

Большинство таких случаев укладывается в следующую схему. Заданы значения функции на некоторой совокупности узлов. Можно произвести некоторую замену переменных, после которой значения функции $f(y_1, \dots, y_s)$ будут заданы на совокупности узлов Ω , обладающей следующими свойствами: совокупность узлов Ω распадается на n_1 совокупностей узлов $\Omega_1, \dots, \Omega_{n_1}$, так что большинство совокупностей Ω_{m_1} содержит достаточно много узлов, Ω_{m_1} лежит в гиперплоскости $y_1 = y_1^{m_1}$, $m_1 = 1, \dots, n_1$, совокупность узлов Ω_{m_1} в свою очередь распадается на совокупности узлов Ω_{m_1, m_2} , $m_2 = 1, \dots, n_2(m_1)$, со свойствами: большинство совокупностей Ω_{m_1, m_2} содержит достаточно много узлов, совокупность Ω_{m_1, m_2} лежит в плоскости $y_1 = y_1^{m_1}$, $y_2 = y_2^{m_1, m_2}$, в свою очередь Ω_{m_1, m_2} распадается на совокупности узлов Ω_{m_1, m_2, m_3} и т. д. Иначе можно сказать, что все узлы могут быть записаны в виде

$$(y_1^{m_1}, y_2^{m_1 m_2}, \dots, y_s^{m_1 \dots m_s}), \quad (1)$$

где

$$1 \leq m_1 \leq n_1, \quad 1 \leq m_k \leq n_k(m_1, \dots, m_{k-1}) \quad \text{при } k > 1.$$

Очевидно, в виде (1) может быть представлена произвольная совокупность узлов, однако нами накладывается дополнительное условие, чтобы совокупности $\Omega_{m_1} \dots \Omega_{m_k}$ состояли из достаточно большого числа узлов. Примерами подобных совокупностей могут служить совокупности, представленные на рис. 5.5.1, 5.5.2. Случай, соответствующий рис. 5.5.2, удовлетворяет наложенным выше требованиям относительно полярной системы координат. Рассмотрим задачу вычисления значения некоторого оператора

$$\mathcal{D}f(p^0) = f^{(r_1, \dots, r_s)}(y_1^0, \dots, y_s^0),$$

частный случай $r_1 = \dots = r_s = 0$ соответствует задаче интерполирования. Возьмем какую-либо формулу вычисления производной $g^{(r_1)}(y_1^0)$ по значениям функции в узлах $y_1^{m_1}$:

$$g^{(r_1)}(y_1^0) \approx \sum_{m_1=1}^{n_1} D^{m_1} g(y_1^{m_1}).$$

Подставляя

$$g(y_1) = f^{(0, r_2, \dots, r_s)}(y_1, y_2^0, \dots, y_s^0),$$

получаем

$$f^{(r_1, \dots, r_s)}(y_1^0, \dots, y_s^0) \approx \sum_{m_1=1}^{n_1} D^{m_1} f^{(0, r_2, \dots, r_s)}(y_1^{m_1}, y_2^0, \dots, y_s^0). \quad (2)$$

Задача дифференцирования функции s переменных $f(y_1, \dots, y_s)$ при помощи этого равенства сводится к задаче дифференцирования функций $f(y_1^{m_1}, y_2, \dots, y_s)$ на единицу меньшего числа переменных. Значения этих функций известны на сетках Ω_{m_1} , удовлетворяющих тем же условиям, что и исходная сетка.

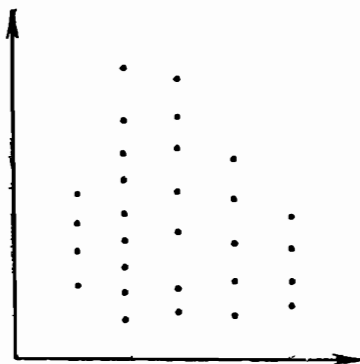


Рис. 5.5.1.



Рис. 5.5.2.

При каждом m_1 выберем некоторую формулу вычисления производной $g^{(r_2)}(y_2^0)$ по значениям функции в узлах $y_2^{m_1 m_2}$:

$$g^{(r_2)}(y_2^0) \approx \sum_{m_2=1}^{n_2(m_1)} D^{m_2} g(y_2^{m_1 m_2}).$$

Подставляя сюда

$$g = f^{(0, 0, r_3, \dots, r_s)}(y_1^{m_1}, y_2, y_3^0, \dots, y_s^0),$$

аналогично (2) получаем

$$\begin{aligned} f^{(0, r_2, \dots, r_s)}(y_1^{m_1}, y_2^0, \dots, y_s^0) &\approx \\ &\approx \sum_{m_2=1}^{n_2(m_1)} D^{m_2} f^{(0, 0, r_3, \dots, r_s)}(y_1^{m_1}, y_2^{m_1 m_2}, y_3^0, \dots, y_s^0). \end{aligned}$$

Воспользовавшись этим соотношением, из (2) получаем

$$f(r_1, \dots, r_s)(y_1^0, \dots, y_s^0) \approx \\ \approx \sum_{m_1=1}^{n_1} D^{m_1} \left(\sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} f^{(0, 0, r_3, \dots, r_s)}(y_1^{m_1}, y_2^{m_1 m_2}, y_3^0, \dots, y_s^0) \right)$$

и т. д. Последовательно осуществляя этот процесс, получим в итоге некоторую формулу численного дифференцирования

$$f(r_1, \dots, r_s)(y_1^0, \dots, y_s^0) \approx \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \dots \\ \dots \sum_{m_s=1}^{n_s(m_1, \dots, m_{s-1})} D^{m_1 \dots m_s} f^{(0, \dots, 0)}(y_1^{m_1}, \dots, y_s^{m_1 \dots m_s}), \quad (3)$$

где $f^{(0, \dots, 0)} \equiv f$. Не следует пугаться, что в правой части стоит сумма слагаемых по всем узлам: на самом деле одномерные формулы численного дифференцирования и интерполирования, на основе которых получена формула (3), содержат относительно мало узлов.

Точно таким же образом строятся формулы численного интегрирования. Пусть вычисляется интеграл

$$I_0 = \int_G f(y_1, \dots, y_s) dy_1, \dots, dy_s.$$

Его можно представить в виде

$$I_0 = \int_{G_0} I_1(y_1) dy_1, \quad I_1(y_1) = \int_{G_1(y_1)} I_2(y_1, y_2) dy_2, \\ I_2(y_1, y_2) = \int_{G_2(y_1, y_2)} I_3(y_1, y_2, y_3) dy_3, \dots, I_{s-1}(y_1, \dots, y_{s-1}) = \\ = \int_{G_{s-1}(y_1, \dots, y_{s-1})} I_s(y_1, \dots, y_s) dy_s.$$

Здесь

$$I_s(y_1, \dots, y_s) = f(y_1, \dots, y_s),$$

G_0 — проекция области G на ось y_1 , $G_k(\bar{y}_1, \dots, \bar{y}_k)$ — сечение множества G плоскостью $y_1 = \bar{y}_1, \dots, y_k = \bar{y}_k$. Воспользуемся какой-либо квадратурной формулой для вычисления первого из этих

выражения $I_1, I_2, \dots$, получаем цепочку соотношений

$$\begin{aligned} I_0 &= E^0 + \sum_{m_1=1}^{n_1} D^{m_1} I_1(y_1^{m_1}) = \\ &= E^0 + \sum_{m_1=1}^{n_1} D^{m_1} \left(E_{m_1}^1 + \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} I_2(y_1^{m_1}, y_2^{m_1 m_2}) \right) = \dots \\ &\dots = R + \sum_{m_1=1}^{n_1} D^{m_1} \left(\sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \times \right. \\ &\times \left(\dots \left(\sum_{m_s=1}^{n_s(m_1, \dots, m_{s-1})} D^{m_1, \dots, m_s} f(y_1^{m_1}, \dots, y_s^{m_1, \dots, m_s}) \right) \dots \right) \right), \end{aligned}$$

где

$$\begin{aligned} R &= E^0 + \sum_{m_1=1}^{n_1} D^{m_1} E_{m_1}^1 + \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} E_{m_1 m_2}^2 + \dots \\ &\dots + \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \dots \\ &\dots \sum_{m_{s-1}=1}^{n_{s-1}(m_1, \dots, m_{s-2})} D^{m_1, \dots, m_{s-1}} E_{m_1, \dots, m_{s-1}}^{s-1}. \quad (6) \end{aligned}$$

Из последнего равенства видно, что погрешность аппроксимации может оказаться существенно больше, чем в одномерном случае, если коэффициенты $D^{m_1}, \dots, m_k$ большие.

Формально сведение многомерной задачи к одномерной носит одинаковый характер и в случае задачи численного дифференцирования (интерполирования) и в случае численного интегрирования. Однако между этими задачами есть такое существенное различие: задача численного дифференцирования (интерполирования) чаще ставится как задача нахождения оператора от функции по значениям на некоторой заданной совокупности узлов Ω . Для задачи интегрирования более типично наличие возможности распоряжаться выбором узлов.

Рассмотрим один пример подобных построений. Цель этого рассмотрения — указать на необходимость внимательности при выборе вспомогательных одномерных формул численного дифференцирования и на практическую важность оценки остаточного члена. Пусть значения функции $f(x_1, x_2)$ известны в узлах сетки $(m_1 h, m_2 h)$. Требуется вычислить значение $f_{x_1 x_2}(0, 0)$. Выпишем простейшую формулу численного дифференцирования функции одной переменной

$$f'(x) \sim \frac{f(x+h) - f(x)}{h}.$$

Вследствие формулы Тейлора

$$f(x+h) = f(x) + f'(x)h + f''(x+\theta h)\frac{h^2}{2}, \quad 0 \leq \theta \leq 1,$$

имеем

$$f'(x) = \frac{f(x+h) - f(x)}{h} - f''(x+\theta h)\frac{h}{2}.$$

Поэтому можно написать

$$f_{x_1 x_2}(0, 0) = \frac{f_{x_2}(h, 0) - f_{x_2}(0, 0)}{h} - f_{x_1 x_2}(\theta h, 0)\frac{h}{2}. \quad (7)$$

В свою очередь возьмем какие-либо аппроксимации производных $f_{x_2}(h, 0)$, $f_{x_2}(0, 0)$; имеем равенства:

$$f_{x_2}(h, 0) = \frac{f(h, h) - f(h, 0)}{h} - f_{x_1 x_2}(h, \theta_1 h)\frac{h}{2},$$

$$f_{x_2}(0, 0) = \frac{f(0, 0) - f(0, -h)}{h} + f_{x_1 x_2}(0, \theta_2 h)\frac{h}{2}.$$

После подстановки в (7) получим

$$f_{x_1 x_2}(0, 0) = \frac{f(h, h) - f(h, 0) - f(0, 0) + f(0, -h)}{h^2} - f_{x_1 x_2}(\theta h, 0)\frac{h}{2} - \frac{1}{2}f_{x_1 x_2}(0, \theta_1 h) - \frac{1}{2}f_{x_1 x_2}(h, \theta_2 h). \quad (8)$$

При построении аппроксимации без одновременного учета остаточного члена можно было бы прийти к неверному заключению, что величина

$$\frac{f(h, h) - f(h, 0) - f(0, 0) + f(0, -h)}{h^2}$$

является хорошей аппроксимацией значения $f_{x_1 x_2}(0, 0)$. Соотношение (8) можно переписать в виде

$$f_{x_1 x_2}(0, 0) = \frac{f(h, h) - f(h, 0) - f(0, 0) + f(0, -h)}{h^2} - f_{x_1 x_2}(0, 0) + O(h).$$

Таким образом, если не обращать внимания на возможность возрастания погрешности за счет умножения на большие множители $D^{m_1} \dots m_k$, то можно получить аппроксимацию с конечной погрешностью, в данном случае аппроксимирующую не требуемую производную, а выражение $f_{x_1 x_2}(0, 0) + f_{x_1 x_2}(0, 0)$.

Важным для приложений является случай, когда узлы образуют некоторую сетку в узком понимании этого слова, иначе множество точек вида

$$(y_1^{m_1}, \dots, y_s^{m_s}) \quad \text{при} \quad m_1 = 1, \dots, N_1; \dots; m_s = 1, \dots, N_s.$$

На рис. 5.5.3 изображено такое расположение точек в случае $s = 2$. Тогда при осуществлении многомерных операций численного дифференцирования (интерполирования и интегрирования) разумно использовать одинаковые формулы для аппроксимации

промежуточных величин. Имеется в виду, что, например, формулы (4) должны иметь вид

$$I_{k-1}(y_1^{m_1}, \dots, y_{k-1}^{m_{k-1}}) \approx \sum_{m_k=1}^{N_k} D^{m_k} I_k(y_1^{m_1}, \dots, y_k^{m_k}), \quad (9)$$

Рис. 5.5.3.

т. е. $D^{m_1} \dots D^{m_k}$ зависят только от m_k . Правая часть (5) тогда приобретет вид

$$\sum_{m_1=1}^{N_1} \dots \sum_{m_s=1}^{N_s} D^{m_1} \dots D^{m_s} f(y_1^{m_1}, \dots, y_s^{m_s}).$$

При построении такой квадратуры мы неявно предполагаем, что область интегрирования — прямоугольный параллелепипед с ребрами, параллельными координатным осям.

Такие формулы численного интегрирования (интерполирования, дифференцирования) называют *прямым произведением соответствующих одномерных формул* численного интегрирования (интерполирования, дифференцирования).

В § 13 будет приведен более сложный пример прямого произведения квадратуры по отрезку на квадратуру по сфере. В случае применения таких аппроксимаций оказывается, что остаточные члены аппроксимации в правой части (6) в значительной мере компенсируются. В том же самом примере при применении формул вида (9) получится аппроксимация

$$f_{x_1 x_2}(0, 0) \approx \frac{f(h, h) - f(0, h) - f(h, 0) + f(0, 0)}{h^2}$$

с погрешностью $O(h)$.

В случае задачи интегрирования возможна такая ситуация: при гладкой подынтегральной функции может оказаться, что промежуточные интегралы $I_k(y_1, \dots, y_k)$ не обладают достаточной гладкостью.

Пусть вычисляется интеграл по единичному кругу

$$I = \int_{-1}^1 \left(\int_{-V_{1-y_1}^2}^{V_{1-y_1}^2} f(y_1, y_2) dy_2 \right) dy_1$$

при $f(y_1, y_2)$ гладкой; если $f(\pm 1, 0) \neq 0$, то функция

$$I_1(y_1) = \frac{\sqrt{1-y_1^2}}{-\sqrt{1-y_1^2}} \int f(y_1, y_2) dy_2$$

имеет неограниченные производные в точках ± 1 , и поэтому при численном интегрировании по переменной y_1 следует применять специальные приемы вычисления интегралов от таких функций — переменный шаг, выделение особенностей и т. д.; целесообразнее записать этот интеграл в виде

$$I = \int_0^1 r I_1(r) dr, \quad I_1(r) = \int_0^{2\pi} f(r \cos \varphi, r \sin \varphi) d\varphi.$$

Подынтегральная функция внутреннего интеграла периодическая и имеет смысл применить квадратуру (3.3.16)

$$\int_0^{2\pi} g(\varphi) d\varphi \approx \frac{2\pi}{n} \sum_{m=0}^{n-1} g\left(\frac{2\pi m}{n}\right).$$

При r малом производные функции f по переменной φ содержат малый множитель r ; поэтому при малых φ иногда выгоднее взять число n поменьше. Для вычисления интеграла по переменной r можно применить любые из рассматривавшихся ранее квадратурных формул, в частности формулы Гаусса, соответствующие весовой функции r .

§ 6. Оценка погрешности численного интегрирования по равномерной сетке

Чтобы проиллюстрировать конкретные способы получения оценок погрешности описываемых выше методов, рассмотрим случай вычисления интеграла по единичному кубу:

$$I = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s) dx_1 \dots dx_s.$$

Предположим, что подынтегральная функция принадлежит классу

$$C_r(A) = C_{r_1 \dots r_s}(A_1, \dots, A_s)$$

функций $f(x_1, \dots, x_s)$, у которых производные $f_{x_k^{r_k-1}}(x_1, \dots, x_s)$ непрерывные, а производные $f_{x_k^{r_k}}(x_1, \dots, x_s)$ кусочно-непрерывны и удовлетворяют условиям $|f_{x_k^{r_k}}(x_1, \dots, x_s)| \leq A_k$. Будут рассматриваться квадратуры, являющиеся прямым произведением

одномерных квадратур. Положим

$$I_s(x_1, \dots, x_s) = f(x_1, \dots, x_s)$$

и

$$I_k(x_1, \dots, x_k) = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s) dx_{k+1} \dots dx_s$$

при $k = 0, \dots, s-1$.

Имеем неравенства

$$\left| (I_k)_{x_k^{r_k}} \right| \leq \int_0^1 \dots \int_0^1 \left| f_{x_k^{r_k}}(x_1, \dots, x_s) \right| dx_{k+1} \dots dx_s \leq A_k.$$

Для вычисления интегралов

$$I_{k-1}(x_1, \dots, x_{k-1}) = \int_0^1 I_k(x_1, \dots, x_k) dx_k$$

применим какую-либо составную квадратурную формулу (3.6.4) с $n = r_k$ при постоянном шаге $a_q - a_{q-1} \equiv H_k = N_k^{-1}$:

$$\begin{aligned} \int_0^1 I_k(x_1, \dots, x_k) dx_k &\approx \\ &\approx \sum_{q_k=1}^{N_k} \frac{H_k}{2} \sum_{j=1}^{m_k} D_j^k I_k(x_1, \dots, x_{k-1}, (q_k - \frac{1}{2})H_k + d_j^k \frac{H_k}{2}). \end{aligned} \quad (1)$$

Согласно (3.6.6), ее погрешность не превосходит

$$D(r_k) \max_{0 \leq x_k \leq 1} \left| (I_k(x_1, \dots, x_k))_{x_k^{r_k}} \right| N_k^{-r_k}$$

и, следовательно, не превосходит $D(r_k) A_k N_k^{-r_k}$. Для такой квадратурной формулы сумма модулей коэффициентов весов квадратуры оценивается величиной

$$\sum_{q_k=1}^{N_k} \frac{H_k}{2} \sum_{j=1}^{m_k} |D_j^k| = \frac{1}{2} \left(\sum_{j=1}^{m_k} |D_j^k| \right) = B^k,$$

не зависящей от N_k . Обращаясь к (5.6), можно написать оценку остаточного члена

$$R = E^0 + \sum_{q_1=1}^{N_1} D^{q_1} E_{q_1}^1 + \sum_{q_1=1}^{N_1} D^{q_1} \sum_{q_2=1}^{N_2} D^{q_2} E_{q_1 q_2}^2 + \dots$$

Отсюда

$$|R| \leq |E^0| + B^1 \max |E_{q_1}^1| + B^1 B^2 \max |E_{q_1 q_2}^2| + \dots \quad (2)$$

В рассматриваемом случае $E_{q_1 \dots q_{k-1}}^{k-1}$ — погрешность квадратуры (1) при некоторых значениях $x_1, \dots, x_{k-1}$, оцениваемая через $D(r_k) A_k N_k^{-r_k}$. Подставляя эту оценку в правую часть (2), получим оценку

$$|R| \leq D_0(r) \sum_{k=1}^s A_k N_k^{-r_k},$$

$D_0(r)$ определяется через величины B^k и $D(r_k)$. Полученная оценка погрешности зависит от выбора конкретных квадратурных формул и числа отрезков разбиения по различным осям.

Займемся минимизацией множителя $\sum_{k=1}^s A_k N_k^{-r_k}$ за счет выбора значений N_k при фиксированном общем числе узлов, пропорциональном величине $N_1 \dots N_s = Q$. Составим функцию Лагранжа

$$\Phi = \sum_{k=1}^s A_k N_k^{-r_k} + \lambda (N_1 \dots N_s - Q).$$

Приравнявая нулю производные Φ_{N_k} , получим систему уравнений

$$-A_k r_k N_k^{-r_k-1} + \lambda N_1 \dots N_s N_k^{-1} = 0,$$

откуда

$$A_k r_k N_k^{-r_k} = \lambda N_1 \dots N_s = \lambda Q, \quad N_k = \left(\frac{A_k r_k}{\lambda Q} \right)^{1/r_k}.$$

Уравнение $N_1 \dots N_s = Q$ переписывается в виде

$$\prod_{k=1}^s \left(\frac{A_k r_k}{\lambda Q} \right)^{1/r_k} = Q$$

или

$$(\lambda Q)^{1/r} = \frac{A}{Q},$$

где

$$A = \prod_{k=1}^s (A_k r_k)^{1/r_k}, \quad \frac{1}{r} = \frac{1}{r_1} + \dots + \frac{1}{r_s};$$

отсюда $\lambda Q = \left(\frac{A}{Q} \right)^r$. Подставляя значение λQ в выражение для N_k , получим

$$N_k = \left(\frac{A_k r_k Q^r}{A^r} \right)^{1/r_k}.$$

Тогда

$$A_k N_k^{-r_k} = \frac{1}{r_k} \left(\frac{A}{Q} \right)^r,$$

и, следовательно, величина $\sum_{k=1}^s A_k N_k^{-r_k}$ есть $O\left(\left(\frac{A}{Q}\right)^r\right)$.

Рассмотрим частный случай: $r_1 = \dots = r_s = r_0$, $A_1 = \dots = A_s = A_0$. Тогда $\frac{1}{r} = \frac{1}{r_1} + \dots + \frac{1}{r_s} = \frac{s}{r_0}$ и $A^r = A_0^r$. Таким образом, рассматриваемые квадратуры обеспечивают оценку погрешности $O(A_0 N^{-r_0/s})$. У реальных подынтегральных функций порядок ограниченных производных r_0 часто оказывается не очень большим и поэтому при больших s такая оценка скорости сходимости оказывается плохой. Иногда удается несколько улучшить порядок оценки, если, как в одномерном случае, оценивать погрешность не через ограниченные производные, а через те производные, которые интегрируемы. Однако, как правило, и у таких оценок скорость стремления к нулю оказывается малой.

Возникает вопрос об оптимальных квадратурах на классах многомерных функций. Так как ни для каких реальных классов функций такие квадратуры неизвестны, мы ограничимся оценкой снизу погрешности оптимальных квадратур.

§ 7. Оценка снизу погрешности численного интегрирования

Напомним постановку задачи оптимизации квадратурных формул на классе функций. Пусть приближенное значение интеграла вычисляется по формуле

$$I(f) = \int_a^b f(P) p(P) dP \approx S_N(f) = \sum_{j=1}^N D_j f(P_j).$$

Величина

$$R_N(f) = I(f) - S_N(f)$$

называется *погрешностью квадратуры*, величина

$$R_N(F) = \sup_{f \in F} |R_N(f)|$$

— *погрешностью квадратуры на классе функций F* , величина

$$W_N(F) = \inf_{D_j, P_j} R_N(F)$$

— *оптимальной оценкой погрешности квадратур на классе F* ; квадратура (если такая существует), на которой эта нижняя грань достигается, называется *оптимальной*. Мы будем предпо-

лагать выполненным условие: существует некоторый куб $\Delta \in G$, в котором $p(P) \geq \gamma > 0$.

Теорема.

$$W_N(C_r(A)) \geq d(\Delta, r, A) \gamma N^{-r},$$

где $d(\Delta, r, A) > 0$.

Доказательство теоремы осуществляется следующим образом. Будет показано, что для любой совокупности узлов $P_1, \dots, P_N$ можно построить функцию $f_{P_1, \dots, P_N}(P)$ рассматриваемого класса $C_r(A)$, обращающуюся в нуль во всех этих узлах и такую, что $I(f) \geq d(\Delta, r, A) \gamma N^{-r}$. При этом постоянная $d > 0$ не зависит от точек $P_1, \dots, P_N$. Тогда для любой квадратуры с этими узлами

$$\begin{aligned} R_N(C_r(A)) &\geq \left| I(f_{P_1, \dots, P_N}) - \sum_{j=1}^N D_j f_{P_1, \dots, P_N}(P_j) \right| = \\ &= |I(f_{P_1, \dots, P_N})| \geq d(\Delta, r, A) \gamma N^{-r}. \end{aligned}$$

Величина $R_N(C_r(A))$ оценена снизу постоянной, не зависящей от узлов квадратур P_j и весов D_j ; поэтому и ее нижняя грань $W_N(C_r(A))$ по множеству всевозможных квадратур также оценивается снизу этой постоянной $d(\Delta, r, A) \gamma N^{-r}$. Таким образом, доказательство теоремы сводится к построению соответствующей функции $f_{P_1, \dots, P_N}(P)$ для каждой совокупности точек $P_1, \dots, P_N$.

Для простоты выкладок будем проводить построения для случая, когда Δ — единичный куб $0 \leq x_1, \dots, x_s \leq 1$, $A_1 = \dots = A_s = A_0$. Положим $n_k = \lceil (2N)^{r/r_k} \rceil + 1$ и разобьем куб Δ на $n_1 \dots n_s$ прямоугольных параллелепипедов $\Pi_m = \Pi_{m_1} \dots m_s$:

$$\frac{m_k - 1}{n_k} \leq x_k \leq \frac{m_k}{n_k}, \quad 0 < m_k \leq n_k, \quad k = 1, \dots, s.$$

Пусть $\varphi(x)$ бесконечно дифференцируема на прямой $(-\infty, \infty)$, причем

$$\begin{aligned} \varphi(x) &= 0 \quad \text{при } x \leq 0 \text{ и } x \geq 1, \\ 0 < \varphi(x) &\leq 1 \quad \text{при } 0 < x < 1. \end{aligned}$$

Пусть

$$\begin{aligned} \bar{r} &= \max_{1 \leq k \leq s} r_k, \\ g &= \max_{0 \leq l \leq \bar{r}} (2^l \max_{[0, 1]} |\varphi^{(l)}(x)|). \end{aligned}$$

Из определения чисел n_k имеем

$$(2N)^{r/r_k} \leq n_k \leq 2 \cdot (2N)^{r/r_k};$$

перемножив эти неравенства и воспользовавшись равенством $\frac{r}{r_1} + \dots + \frac{r}{r_s} = 1$, получим

$$2N \leq n_1 \dots n_s \leq 2^{s+1} N. \quad (1)$$

Построим функцию

$$f_{P_1, \dots, P_N}(P) = f_0(P)$$

следующим образом. В тех параллелепипедах Π_m , которые не содержат внутри ни одной из точек $P_1, \dots, P_s$, положим

$$f_0(P) = \frac{A_0}{g(2N)^r} \prod_{k=1}^s \Phi(n_k x_k - (m_k - 1)). \quad (2)$$

Во всех остальных точках положим $f_0(P) = 0$. Функция f_0 бесконечно дифференцируема внутри каждого параллелепипеда Π_m и обращается в нуль вместе со всеми производными на границе каждого такого параллелепипеда. Следовательно, она бесконечно дифференцируема во всем пространстве $-\infty < x_1, \dots, x_s < \infty$. Из ее построения также следует, что она обращается в нуль со всеми производными во всех узлах $P_1, \dots, P_N$. Оценим снизу значение $I(f_0)$. Воспользовавшись неравенством

$p(P) \geq \gamma$, после замены переменных $x_k = \frac{m_k - 1 + y_k}{n_k}$ получим цепочку соотношений

$$\begin{aligned} & \int_{\Pi_m} p(x_1, \dots, x_s) A_0 (g(2N)^r)^{-1} \prod_{k=1}^s \Phi(n_k x_k - (m_k - 1)) dx_1 \dots dx_s \geq \\ & \geq \frac{A_0 \gamma}{g(2N)^r n_1 \dots n_s} \int_0^1 \dots \int_0^1 \prod_{k=1}^s \Phi(y_k) dy_1 \dots dy_s = \frac{d_1(r) A_0 \gamma}{N^r n_1 \dots n_s}; \quad (3) \end{aligned}$$

здесь

$$d_1(r) = \frac{1}{g \cdot 2^r} \left(\int_0^1 \Phi(y) dy \right)^s.$$

Каждая точка P_j может находиться внутри только одного из параллелепипедов Π_m . Следовательно, по крайней мере в $n_1 \dots n_s - N$ параллелепипедах Π_m функция f_0 отлична от тождественного нуля и определяется равенством (2). Согласно (1), имеем $N \leq (n_1 \dots n_s)/2$, и поэтому таких параллелепипедов не менее $(n_1 \dots n_s)/2$. С учетом (3) получаем оценку

$$I(f_0) \geq \frac{n_1 \dots n_s}{2} \frac{d_1(r) A_0 \gamma}{N^r \cdot n_1 \dots n_s} = \frac{d(\Delta, r, A) \gamma}{N^r}; \quad (4)$$

здесь

$$d(\Delta, r, A) = \frac{d_1(r) A_0}{2}.$$

Поскольку функция f_0 бесконечно дифференцируема, то для доказательства ее принадлежности рассматриваемому классу достаточно оценить ее производные в тех точках параллелепипедов, где она отлична от нуля. В этих точках имеем

$$f_{x_l}^{l_l} = \frac{A_0}{g(2N)^r} \varphi^{(l_l)}(n_l x_l - (m_l - 1)) n_l^{l_l} \prod_{k \neq l} \varphi(n_k x_k - (m_k - 1)).$$

Поскольку

$$|\varphi| \leq 1, \quad |\varphi^{(l_l)}| \leq g \cdot 2^{-l_l} \quad \text{при} \quad l_l \leq \bar{r},$$

то отсюда следует оценка

$$|f_{x_l}^{l_l}| \leq \frac{A_0}{(2N)^r} \left(\frac{n_l}{2}\right)^{l_l}. \quad (5)$$

Согласно определению n_l , имеем

$$2 \leq n_l \leq 2(2N)^{r_l/r_l}.$$

Поэтому

$$\left(\frac{n_l}{2}\right)^{l_l} \leq (2N)^{r_l l_l / r_l} \leq (2N)^r$$

при $0 \leq l_j \leq r_j$. Обращаясь к (5), получаем оценку $|f_{x_l}^{l_l}| \leq A_0$

при $0 \leq l_j \leq r_j$. Таким образом, построенная функция принадлежит рассматриваемому классу, для нее выполняется неравенство (4) с постоянной $d(\Delta, r, A)$, не зависящей от узлов интегрирования, она обращается в нуль во всех узлах P_j ; таким образом, способ построения требуемой функции для любой совокупности узлов $P_1, \dots, P_N$ указан.

З а м е ч а н и е. Построенная нами функция обращается в нуль в узлах $P_1, \dots, P_N$ вместе со всеми ее производными. Поэтому ту же самую оценку снизу мы получили бы для класса всевозможных квадратурных формул по значениям функции и ее производных в N точках.

§ 8. Об оптимизации оценки погрешности на более широких классах способов интегрирования

В § 7 производилась оценка снизу погрешности квадратур, т. е. формул интегрирования

$$I(f) \approx \sum_{j=1}^N D_j f(P_j),$$

где узлы интегрирования P_j и веса D_j не зависят от конкретной подынтегральной функции. Для многих задач практически более эффективны способы интегрирования, где последующие узлы интегрирования выбираются в зависимости от информации об уже вычисленных значениях функции, например, способы интегрирования с автоматическим выбором шага (§§ 17, 18 гл. III).

Пусть имеется какой-то способ интегрирования, где информация о подынтегральной функции учитывается лишь в виде информации об ее значениях в отдельных точках. Этот способ определяется заданием первого узла интегрирования, правила, по которому отыскиваются последующие узлы, и способа вычисления приближенного значения интеграла. Таким образом, всякий такой способ укладывается в следующую схему: задается некоторый узел P_1 и функции

$$Q_q = \Phi_q(Q_1, \dots, Q_{q-1}; y_1, \dots, y_{q-1}), \quad q = 2, \dots, N,$$

$$S_N(Q_1, \dots, Q_N; y_1, \dots, y_N).$$

Здесь Q_j — точки области G , y_j , S_N — числа. При приближенном вычислении конкретного интеграла последовательно вычисляют величины

$$f(P_1), \quad P_2 = \Phi_2(P_1; f(P_1)), \quad f(P_2),$$

$$P_3 = \Phi_3(P_1, P_2; f(P_1), f(P_2)), \quad f(P_3), \dots, f(P_N)$$

и затем полагают

$$I(f) \approx S_N(P_1, \dots, P_N; f(P_1), \dots, f(P_N)).$$

Поскольку точка P_1 задается вместе с заданием функций Φ_j и S_N , зависимость этих функций от P_1 можно было бы опустить.

Так же, как в случае квадратурных формул, можно определить погрешность метода при вычислении данного интеграла

$$\tilde{R}_N(f) = I(f) - S_N(P_1, \dots, P_N; f(P_1), \dots, f(P_N)),$$

погрешность метода на классе подынтегральных функций

$$\tilde{R}_N(F) = \sup_{f \in F} |\tilde{R}_N(f)|$$

и оптимальную оценку погрешности на множестве всевозможных методов интегрирования

$$Z_N(F) = \inf_{P_1, \Phi_2, \dots, \Phi_N, S_N} \tilde{R}_N(F).$$

Теорема. Пусть класс функций F — выпуклый, центрально-симметричный компакт с центром симметрии $f \equiv 0$ и все функции этого класса равномерно ограничены. Тогда оптимальные оценки погрешности на множествах всевозможных квадратурных

формул и всевозможных способов интегрирования совпадают:

$$W_N(F) = Z_N(F).$$

Условие выпуклости класса F означает, что вместе с любыми $f_1, f_2 \in F$ этому классу принадлежат также все функции $f_1\theta + f_2(1-\theta)$ при $0 < \theta < 1$. Условие, что класс является центрально симметричным с центром симметрии $f \equiv 0$, означает, что вместе со всякой функцией f этому классу принадлежит также функция $-f$. В частности, все классы $C_r(A)$ являются классами, удовлетворяющими условиям теоремы.

Наметим ход доказательства теоремы. Возьмем произвольное $\eta > 0$. Согласно определению $Z_N(F)$, можно указать способ интегрирования, для которого

$$\tilde{R}_N(F) \leq Z_N(F) + \eta.$$

При $P_1, \Phi_2, \dots, \Phi_N$, соответствующих этому способу, положим

$$P_1^0 = P_1, \quad P_2^0 = \Phi_2(P_1^0; 0), \dots,$$

$$P_N^0 =$$

$$= \Phi_N(P_1^0, \dots, P_{N-1}^0; 0, \dots, 0).$$

Обозначим через Y множество точек $(N+1)$ -мерного пространства

$$(y_0, y_1, \dots, y_N) = \\ = (I(f), f(P_1^0), \dots, f(P_N^0))$$

при $f \in F$. Согласно свойствам класса F , множество точек Y — центрально симметричное с центром симметрии в начале координат, замкнутое и лежит в конечной части пространства (рис. 5.8.1). Пусть

$$D_0 = \sup_{(y_0, \dots, y_N) \in Y} y_0.$$

Вследствие замкнутости Y точка $(D_0, 0, \dots, 0)$ принадлежит Y и существует соответствующая ей функция $f_0 \in F$, для которой

$$I(f_0) = D_0, \quad f_0(P_1^0) = \dots = f_0(P_N^0) = 0.$$

Из теории выпуклых множеств известно, что через всякую точку границы множества можно провести опорную плоскость; напомним, что опорной плоскостью называется плоскость, проходящая через некоторую точку границы множества и такая, что по одну сторону от этой плоскости нет точек множества. Пусть уравнение этой опорной плоскости

$$\alpha_0(y_0 - D_0) + \sum_{j=1}^N \alpha_j y_j = 0.$$

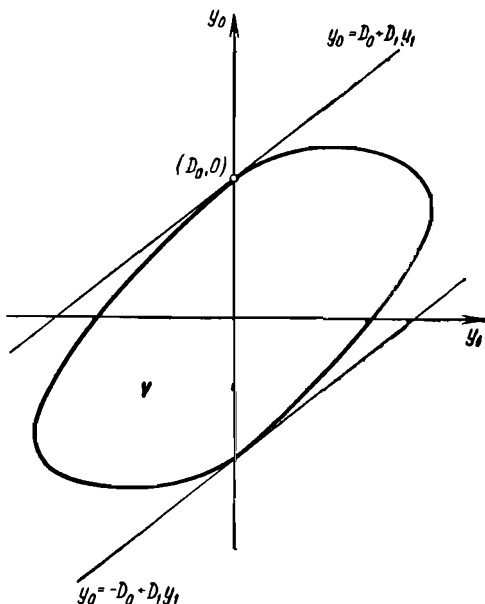


Рис. 5.8.1.

Предоставляем читателю самому убедиться, что среди всех опорных плоскостей, проходящих через точку $(D_0, 0, \dots, 0)$, существует такая, для которой $\alpha_0 \neq 0$. Уравнение этой опорной плоскости можно записать в виде

$$y_0 = D_0 + \sum_{l=1}^N D_l y_l, \quad D_l = -\frac{\alpha_l}{\alpha_0}.$$

Плоскостью, симметричной этой плоскости относительно начала координат $(0, \dots, 0)$, будет плоскость

$$y_0 = -D_0 + \sum_{l=1}^N D_l y_l.$$

Вследствие симметрии тела Y относительно начала координат эта плоскость также будет опорной к телу Y . Тело Y находится между этими плоскостями, и поэтому для всех точек тела

$$\left| y_0 - \sum_{l=1}^N D_l y_l \right| \leq D_0;$$

вспоминая смысл координат y_j , перепишем это соотношение в виде

$$\left| I(f) - \sum_{j=1}^N D_j f(P_j^0) \right| \leq D_0$$

при всех $f \in F$.

Мы получили квадратурную формулу с оценкой остаточного члена $R_N(F) \leq D_0$. Следовательно, выполняется неравенство

$$W_N(F) = \inf R_N(F) \leq D_0.$$

Применим исходный алгоритм к вычислению интегралов $I(\pm f_0)$. Вследствие условия $f_0(P_j^0) = 0$ при всех j последовательно будем получать значения узлов и значения функций в них:

$$\begin{aligned} \pm f_0(P_1^0) &= 0, \quad P_2 = \Phi_2(P_1^0; \pm f(P_1^0)) = \\ &= \Phi_2(P_1^0; 0) = P_2^0, \quad \pm f_0(P_2^0) = 0, \dots, \quad \pm f_0(P_{N-1}^0) = 0, \\ P_N &= \Phi_N(P_1^0, \dots, P_{N-1}^0; \pm f_0(P_1^0), \dots, \pm f_0(P_{N-1}^0)) = \\ &= \Phi_N(P_1^0, \dots, P_{N-1}^0; 0, \dots, 0) = P_N^0, \quad \pm f_0(P_N^0) = 0. \end{aligned}$$

В результате в обоих случаях приближенное значение интеграла будет положено равным величине

$$S^0 = S(P_1^0, \dots, P_N^0; \pm f_0(P_1^0), \dots, \pm f_0(P_N^0)) = S(P_1^0, \dots, P_N^0; 0, \dots, 0).$$

В то же время

$$I(\pm f_0) = \pm D_0,$$

поэтому

$$|I(f_0) - S^0| + |S^0 - I(-f_0)| \geq |(I(f_0) - S^0) + (S^0 - I(-f_0))| = 2D_0.$$

Следовательно, для одной из функций $\pm f_0$ ошибка приближенного значения

не менее D_0 . Это означает, что $D_0 \leq \tilde{R}_N(F)$; мы имеем цепочку соотношений

$$\mathcal{W}_N(F) \leq R_N(F) \leq D_0 \leq \tilde{R}_N(F) \leq Z_N(F) + \eta,$$

вследствие произвольности η отсюда заключаем, что

$$\mathcal{W}_N(F) \leq Z_N(F).$$

Множество всевозможных способов интегрирования содержит в себе множество квадратурных формул, поэтому нижняя грань погрешности по множеству всевозможных способов не больше, чем нижняя грань по множеству квадратурных формул

$$Z_N(F) \leq \mathcal{W}_N(F).$$

Таким образом,

$$\mathcal{W}_N(F) = Z_N(F).$$

Классы функций, рассматриваемые в различных направлениях теории численного интегрирования, всегда удовлетворяют условиям доказанной теоремы; на основании этого, казалось бы, можно сделать вывод, что рассмотрение способов интегрирования более общего вида не дает преимуществ по сравнению с рассмотрением квадратурных формул и поэтому является нецелесообразным. С позиций прикладной науки такой вывод является неправильным. Приведем соображения в обоснование этого высказывания.

1. Возможно, что реально встречающиеся классы функций просто не являются выпуклыми. По-видимому, типичным классом функций, встречающихся в приложениях, является класс кусочно-аналитических функций. Обратимся к одномерному случаю. В случае выпуклости класса функций полусумма двух функций класса также принадлежит этому классу. Полусумма функций, имеющих l точек, где нарушается аналитичность, может иметь $2l$ точек с нарушением аналитичности. Таким образом, класс аналитических функций, имеющих не более заданного числа l точек нарушения аналитичности, не является выпуклым. Если число точек нарушения аналитичности не ограничено сверху, то класс функций не является замкнутым: предел последовательности кусочно-аналитических функций с неограниченно растущим числом точек нарушения аналитичности может оказаться функцией, не являющейся кусочно-аналитической.

2. Выше производилось сравнение методов интегрирования по верхней грани погрешности на классе. Однако возможна следующая ситуация. Два метода имеют одну и ту же погрешность на классе, в то же время на большинстве функций класса один из методов имеет меньшую погрешность. Ясно, что этот метод является более предпочтительным, и сравнение методов по верхней грани погрешности на классе в данном случае не дает общей картины.

§ 9. Метод Монте-Карло

При построении квадратурных формул мы всегда получали вместе с формулой оценку ее погрешности на некотором классе функций. Например, для формулы трапеций была получена оценка погрешности вида $\text{const} \cdot A_2 N^{-2}$ на классе функций со второй производной, ограниченной постоянной A_2 ; здесь N — число узлов интегрирования. Такого рода оценки погрешности называют *гарантированными оценками погрешности на классе функций*. На основании этой оценки можно гарантировать, что погрешность приближенного значения интеграла не превосходит определенной величины для всех подынтегральных функций из рассматриваемого класса. При таком подходе, как оценка погрешности метода на классе, в случае оценки погрешности конкретного интеграла мы ориентируемся на величину погрешности в случае интегрирования «худшей» функции рассматриваемого класса. Для ряда классов функций эта оценка погрешности на классе настолько плоха, что не дает никакой надежды на получение приближенного значения интеграла с требуемой точностью. Например, согласно теоремам из §§ 7, 8, не существует методов с оценкой погрешности на классе функций $C_{1, \dots, 1}(A, \dots, A)$, лучшей $dAN^{-1/s}$ (здесь s — кратность вычисляемого интеграла).

Предположим, что требуется гарантировать оценку погрешности, меньшую $0,01 dA$. Тогда число узлов должно удовлетворять неравенству $dAN^{-1/s} \leq 0,01 dA$, иначе $N \geq 100^s$. Поскольку вычисление каждого значения подынтегральной функции требует обычно большого числа арифметических операций, уже при $s = 6$ такое требование на число узлов оказывается практически невыполнимым.

Мы оказались в положении, когда нет возможности вычислить значение интеграла с гарантированной оценкой погрешности $0,01 dA$. Один из способов разрешения создавшегося противоречия состоит в более детальном описании классов подынтегральных функций. В частности, при рассмотрении задачи вычисления интегралов, возникающих при решении интегральных уравнений с помощью разложения в ряд Неймана, возникли новые классы подынтегральных функций и соответствующие им теоретико-числовые методы интегрирования, остающиеся вне нашего рассмотрения. Другим выходом из создавшейся обстановки является отказ от получения строгой, гарантированной оценки погрешности и получение оценки погрешности лишь с определенной степенью достоверности. В частности, при конструировании стандартных программ интегрирования в § 17 гл. III мы пошли по пути отказа от строгой оценки погрешности: погрешность оценивалась через разность результатов вычисления

приближенного значения интеграла при различных значениях параметров.

Одним из методов приближенного вычисления значений интегралов, при котором погрешность оценивается не гарантированно, а лишь с некоторой степенью достоверности, является метод Монте-Карло.

Пусть требуется вычислить приближенное значение интеграла

$$I(f) = \int_G f(P) dP.$$

Предположим, что каким-то образом удастся получить N случайных, попарно независимых точек $P_1, \dots, P_N$, одинаково распределенных в G с плотностью распределения $p(P)$, $\int_G p(P) dP = 1$. Положим

$$s_I(f) = \frac{f(P_I)}{p(P_I)}.$$

Случайные величины $s_I(f)$ попарно независимы и одинаково распределены:

$$Ms_I(f) = \int_G \frac{f(P)}{p(P)} p(P) dP = I(f)$$

и

$$\begin{aligned} D(s_I(f)) &= Ms_I^2(f) - (Ms_I(f))^2 = \\ &= \int_G \left(\frac{f(P)}{p(P)} \right)^2 p(P) dP - (Ms_I(f))^2 = D(f), \end{aligned}$$

где

$$D(f) = I\left(\frac{f^2}{p}\right) - (I(f))^2.$$

Положим

$$S_N(f) = \frac{1}{N} \sum_{I=1}^N s_I(f).$$

Вследствие указанных свойств величин $s_I(f)$ имеем

$$MS_N(f) = \frac{1}{N} \sum_{I=1}^N Ms_I(f) = I(f),$$

$$D(S_N(f)) = \frac{1}{N^2} \sum_{I=1}^N D(s_I(f)) = \frac{1}{N} D(f).$$

С вероятностью $1 - \eta$ выполняется неравенство (неравенство Чебышева)

$$|S_N(f) - I(f)| \leq \sqrt{\frac{D(f)}{\eta N}}. \quad (1)$$

Полагая $\eta = 0,01$, получаем: с вероятностью 99% выполняется неравенство

$$|S_N(f) - I(f)| \leq 10 \sqrt{\frac{D(f)}{N}}.$$

Еще лучшая оценка получится в предположении, что точки P_j не только попарно независимы, но и независимы в совокупности. Тогда, согласно центральной предельной теореме, случайная величина

$$\frac{S_N(f) - I(f)}{\sqrt{D(f)/N}}$$

распределена асимптотически нормально с функцией распределения

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y \exp\left(-\frac{t^2}{2}\right) dt.$$

Вероятность того, что случайная величина с такой функцией распределения не превосходит по модулю значение $y > 0$, асимптотически равна

$$p_0(y) = 1 - \frac{\sqrt{2}}{\sqrt{\pi}} \int_y^{\infty} \exp\left(-\frac{t^2}{2}\right) dt.$$

Таким образом, при N больших с вероятностью, близкой к $p_0(y)$, выполняется неравенство

$$|S_N(f) - I(f)| \leq y \sqrt{\frac{D(f)}{N}}.$$

Полагая $y = 3$ и $y = 5$, получаем, что неравенства

$$|S_N(f) - I(f)| \leq 3 \sqrt{\frac{D(f)}{N}} \quad \text{и} \quad |S_N(f) - I(f)| \leq 5 \sqrt{\frac{D(f)}{N}}$$

выполняются, соответственно, с вероятностями 0,997 и 0,99999. Сформулированные выше утверждения называются иногда правилами «трех сигм» и «пяти сигм», соответственно.

В правой части этих оценок стоит неизвестная величина

$$D(f) = I\left(\frac{f^2}{p}\right) - (I(f))^2;$$

попытаемся оценить ее на основании информации о вычисленных значениях $f(P_j)$. Согласно определению $D(f)$, имеем

$$D(f) = \mathbf{M}(s_I(f) - I(f))^2.$$

Поскольку $S_N(f) \approx I(f)$, то $D(f) \approx l(f)$, где

$$l(f) = \mathbf{M}(s_I(f) - S_N(f))^2.$$

Проведем более аккуратные рассуждения.

Так как

$$\mathbf{M}(s_I(f) - S_N(f)) = \mathbf{M}s_I(f) - \mathbf{M}S_N(f) = I(f) - I(f) = 0,$$

то $l(f) = \mathbf{D}(s_I(f) - S_N(f))$. Имеем

$$s_I(f) - S_N(f) = s_I(f) - \frac{1}{N} \sum_{i=1}^N s_i(f) = s_I(f) \left(1 - \frac{1}{N}\right) - \frac{1}{N} \sum_{i \neq I} s_i(f),$$

отсюда получаем

$$\begin{aligned} \mathbf{D}(s_I(f) - S_N(f)) &= \left(1 - \frac{1}{N}\right)^2 \mathbf{D}(s_I(f)) + \frac{1}{N^2} \sum_{i \neq I} \mathbf{D}(s_i(f)) = \\ &= \left(\left(1 - \frac{1}{N}\right)^2 + \frac{N-1}{N^2}\right) D(f) = \left(1 - \frac{1}{N}\right) D(f). \end{aligned}$$

В итоге $l(f) = \left(1 - \frac{1}{N}\right) D(f)$, и таким образом,

$$D(f) = \frac{N}{N-1} \mathbf{M}(s_I(f) - S_N(f))^2.$$

На основании закона больших чисел можно показать, что с большой вероятностью выполняется приближенное равенство

$$\mathbf{M}(s_I(f) - S_N(f))^2 \approx \frac{1}{N} \sum_{i=1}^N (s_i(f) - S_N(f))^2.$$

Из последнего соотношения делаем вывод, что с большой вероятностью выполняется приближенное равенство

$$D(f) \approx D^{(N)}(f),$$

где

$$D^{(N)}(f) = \frac{1}{N-1} \sum_{i=1}^N (s_i(f) - S_N(f))^2.$$

При $\mu(G) = 1$, $p(P) = 1$ приведем схему применения метода Монте-Карло. Пусть $1 - \eta$ — уровень надежности, с которым желательно получить приближенное значение интеграла, E —

заданная точность. Определяют y из равенства

$$\eta = \frac{1}{\sqrt{\pi}} \int_y^{\infty} \exp\left(-\frac{t^2}{2}\right) dt.$$

Последовательно, при $n = 1, \dots$ получают случайные точки P_n и вычисляют величины $t_n(f)$, $S_n(f)$, $d_n(f)$, $D_n(f)$, пользуясь рекуррентными соотношениями

$$t_n(f) = t_{n-1}(f) + f(P_n)$$

$$S_n(f) = \frac{t_n(f)}{n},$$

$$d_n(f) = d_{n-1}(f) + \frac{n}{n-1} (f(P_n) - S_n(f))^2,$$

$$D_n(f) = \frac{d_n(f)}{n-1},$$

и величину

$$\lambda_n = y \sqrt{\frac{D_n(f)}{n}}.$$

Начальные условия рекурсии:

$$t_1(f) = S_1(f) = f(P_1), \quad d_1(f) = D_1(f) = 0.$$

Задача. Проверить, что $D_n(f) = D^{(n)}(f)$.

Если оказалось, что $\lambda_n \leq E$, то вычисления прекращаются; полагают $I(f) \approx S_N(f)$ и считают, что с вероятностью $1 - \eta$ выполняется неравенство $|I(f) - S_N(f)| \leq E$.

Заметим, что в действительности неравенство $|I(f) - S_N(f)| \leq E$ выполняется с вероятностью, несколько меньшей $1 - \eta$, хотя и очень близкой к ней.

Часто в руководствах по использованию метода Монте-Карло говорится следующее. Метод Монте-Карло является универсальным методом вычисления интегралов высокой кратности. Порядок оценки скорости сходимости метода Монте-Карло $O(1/\sqrt{N})$ не зависит от размерности интеграла, в то время как порядок гарантированных оценок скорости сходимости существенно ухудшается с ростом размерности. Для метода Монте-Карло при каждом n имеется эффективная оценка погрешности через величину $\sqrt{D_n(f)/n}$, и таким образом, вычисления прекращаются ровно в тот момент, когда достигнута требуемая точность.

Отнесемся осторожно к этому заявлению и рассмотрим различные соображения «за» и «против» метода Монте-Карло.

§ 10. Обсуждение правомерности использования недетерминированных методов решения задач

Часть пользователей предубеждена против метода Монте-Карло вследствие того, что малость погрешности метода обеспечивается лишь с некоторой вероятностью, и вследствие этого отрицает правомерность его использования.

Выше мы уже обрисовали картину, складывающуюся при вычислении интегралов очень высокой кратности, как картину почти полной безнадёжности получения приближенного значения интеграла с гарантией малой оценки погрешности. Поэтому применение метода Монте-Карло вызвано обстановкой, когда трудно надеяться на получение приближенного значения интеграла с гарантированно малой оценкой погрешности.

Заметим, что даже описанные в главе III стандартные программы вычисления однократных интегралов не дают полной гарантии малости погрешности в случае произвольных гладких подынтегральных функций; кроме того, при решении всякой задачи возможны ошибки в постановке задачи, в программе и т. д. В силу этих и ряда других причин редко можно дать стопроцентную гарантию малости погрешности результата расчета по отношению к реальной модели; некоторая вероятность ошибочности результатов вычислений имеется в любом случае. Все это подчеркивает, что полный отказ от метода Монте-Карло только из-за его вероятностной природы является неоправданным.

С другой стороны, при использовании метода Монте-Карло нужно учитывать следующие отрицательные эффекты.

Для применения способа Монте-Карло необходимо иметь в распоряжении последовательность независимых точек P_j с заданным законом распределения. Обычно пользователь располагает датчиками случайных или так называемых псевдослучайных чисел, которые выдают последовательности случайных чисел, равномерно распределенных на $[0, 1]$; при помощи преобразований таких случайных величин получают случайные числа с требуемыми законами распределения. На первых ЭВМ датчиками случайных чисел были некоторые приборы, например, использующие явление радиоактивного распада, которые выдавали последовательности случайных величин, иногда даже удовлетворявшие требованию независимости в совокупности. Однако такие приборы работают с малой скоростью, и поэтому с увеличением скорости работы ЭВМ от них отказались. Вместо датчиков случайных чисел используются датчики псевдослучайных чисел — некоторые программы, выдающие последовательности чисел, которые рекомендуется рассматривать как случайные. Использование датчиков псевдослучайных чисел явилось

прогрессивным шагом, позволившим широко применять вероятностные методы. Однако при использовании этих датчиков нужно всегда иметь в виду, какими свойствами обладают последовательности чисел, выдаваемые этими датчиками. Например, некоторые датчики псевдослучайных чисел вырабатывают последовательности чисел, которые можно рассматривать лишь как попарно независимые, а не как независимые в совокупности. В этом случае будет неправомерно пользоваться оценками погрешности, основанными на центральной предельной теореме. Пусть методом Монте-Карло вычисляется интеграл

$$I(f) = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s) dx_1 \dots dx_s.$$

Предположим, что в качестве узлов мы хотим выбрать последовательность независимых равномерно распределенных точек единичного куба. Если датчик псевдослучайных чисел выдает последовательность $\xi_1, \dots$ чисел, равномерно распределенных на $[0, 1]$, то естественно попытаться в качестве узлов интегрирования взять точки $(\xi_1, \dots, \xi_s)$, $(\xi_{s+1}, \dots, \xi_{2s})$, $\dots$

Для законности применения неравенства Чебышева нужно выполнение предположения о независимости распределения любых точек P_j , т. е. независимости распределения совокупностей

$$(\xi_{(j-1)s+1}, \dots, \xi_{js}), (\xi_{(i-1)s+1}, \dots, \xi_{is}).$$

При увеличении s это условие накладывает все более жесткие требования на датчик псевдослучайных чисел. Известно много реальных примеров неудачного применения метода Монте-Карло в случае больших s , вызванных следующей причиной. При использовании метода и оценке погрешности делались допущения о тех или иных статистических свойствах псевдослучайных чисел, в то время как эти предположения на самом деле не выполнялись. В результате делался вывод о малости значения погрешности, который на самом деле не был справедлив.

Таким образом, опасность применения метода Монте-Карло заключается больше не в вероятностном характере оценки погрешности, а в том, что вероятностная оценка погрешности производится зачастую в предположении о свойствах датчиков случайных чисел, которые на самом деле не имеют места.

К числу достоинств метода Монте-Карло относят независимость порядка оценки погрешности от размерности вычисляемого интеграла. Однако рассуждая только о порядке сходимости метода, можно не заметить следующую немаловажную деталь. Мы получали оценку погрешности приближенного значения интеграла вида

$$|S_N(f) - I(f)| \leq \text{const} \cdot \sqrt{D(f)/N}.$$

Типичным для практики является требование малости относительной погрешности приближенного значения интеграла, что в данном случае означает требование малости величины

$$\frac{\sqrt{D(f)}}{|I(f)|\sqrt{N}}.$$

Статистика реально предъявляемых к вычислению интегралов показывает, что величина $\sqrt{D(f)}/|I(f)|$ имеет тенденцию к резкому росту с ростом размерности интегралов. В качестве иллюстрации приведем интеграл

$$I(f_s) = \int_0^1 \dots \int_0^1 \exp(-32(x_1^2 + \dots + x_s^2)) dx_1 \dots dx_s,$$

для которого

$$\frac{\sqrt{D(f_s)}}{|I(f_s)|} > 10^{s/2} - 1.$$

Поэтому практическая трудоемкость метода Монте-Карло существенно возрастает с ростом размерности интегралов (при одинаковой относительной погрешности). При действительном вычислении многократных интегралов методом Монте-Карло перед непосредственным применением метода зачастую с целью уменьшения величины $\sqrt{D(f)}/|I(f)|$ проводится довольно кропотливое исследование свойств подынтегральной функции, преобразование интеграла с помощью замен переменных и других приемов.

§ 11. Ускорение сходимости метода Монте-Карло

Рассмотрим некоторые приемы повышения практической эффективности метода Монте-Карло.

1. Сюда относится, в частности, представление $f(P)$ в виде

$$f(P) = F(P) + g(P),$$

где $F(P)$ интегрируется в явном виде и содержит в себе все резко меняющиеся компоненты $f(P)$, а $g(P)$ — плавно меняющаяся функция с небольшой дисперсией. Иногда область интегрирования разбивают на малые подобласти и в каждой части в качестве $F(P)$ берут некоторый интерполяционный полином с узлами в этой подобласти.

2. Описанный в § 9 вариант метода Монте-Карло содержал как свободный параметр функцию $p(P)$ — плотность распределения узлов интегрирования. Из оценки погрешности видно, что погрешность тем меньше, чем меньше разброс значений функции $f(P)/p(P)$. Таким образом, подходящий подбор

функции $p(P)$ также приводит к уменьшению дисперсии. Мы рассматривали случай, когда все узлы P_j имеют одинаковую функцию распределения $p(P)$; в ряде случаев оказывается целесообразным выбирать узлы интегрирования каждый со своей функцией распределения $p_j(P)$.

Широко распространенным частным случаем этого приема является следующий. Исходный интеграл представляется в виде суммы интегралов

$$I(f) = \int_G f(P) dP = \sum_{i=1}^m I_i(f), \quad I_i(f) = \int_{G_i} f(P) dP,$$

N представляется в виде

$$N = N_1 + \dots + N_m,$$

и каждый интеграл $I_i(f)$ вычисляется по методу Монте-Карло с N_i узлами интегрирования. Обратимся к случаю вычисления интеграла от функции, изображенной на рис. 5.11.1. При непосредственном вычислении исходного интеграла в случае $p \equiv 1$ имеем $D(f) = 1/4$. Если $G_1 = [0, 1/2]$, $G_2 = [1/2, 1]$, то при любых $N_1, N_2 \neq 0$ оба интеграла $I_i(f)$, а следовательно, и исходный интеграл будут вычислены точно. Конечно, это очень редкий случай, когда исходную область интегрирования удалось разбить на части, где подынтегральная функция постоянна; однако если все-таки удастся разбить ее на части, где функция меняется мало, то можно получить существенное увеличение точности при том же объеме вычислений.

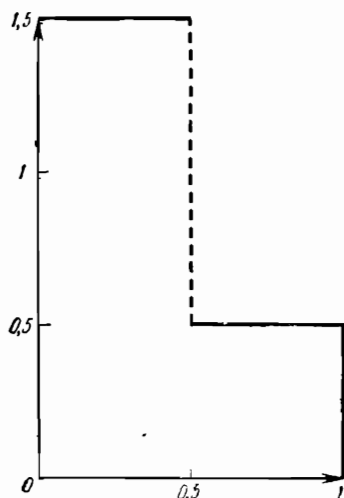


Рис. 5.11.1.

Рассмотрим случай, когда при вычислении каждого из интегралов $I_i(f)$ узлы интегрирования $P_j^i, j = 1, \dots, N_i$, имеют равномерную плотность распределения $p_i = \mu_i^{-1}$ в области G_i , причем любые два узла (как принадлежащие одной области, так и принадлежащие различным областям) распределены независимо; здесь μ_i — мера области G_i . Произведем сравнение описанного метода с исходной схемой метода Монте-Карло, если в последней узлы интегрирования распределены равномерно и попарно независимы. Тогда для исходной схемы метода

Монте-Карло

$$p(P) = \mu^{-1}, \quad \mathbf{D}S_N(f) = D(f) N^{-1},$$

где μ — мера области G ,

$$D(f) = \mu I(f^2) - (I(f))^2.$$

За приближенные значения интегралов $I_l(f)$ принимаются величины

$$S^l(f) = \frac{\mu_l}{N_l} \sum_{i=1}^{N_l} f(P_i).$$

Соответственно, имеем

$$\mathbf{M}S^l(f) = I_l(f),$$

$$\mathbf{D}S^l(f) = D_l(f) N_l^{-1},$$

где

$$D_l(f) = \mu_l I_l(f^2) - (I_l(f))^2.$$

Положим

$$\bar{S}_N(f) = \sum_{l=1}^m S^l(f);$$

при попарной независимости распределений любых двух узлов P_n^i и P_q^j величины $S^l(f)$ также попарно независимы, и поэтому

$$\mathbf{M}\bar{S}_N(f) = \sum_{l=1}^m \mathbf{M}S^l(f) = I(f),$$

$$\mathbf{D}\bar{S}_N(f) = \sum_{l=1}^m \mathbf{D}(S^l(f)) = \sum_{l=1}^m D_l(f) N_l^{-1}.$$

Рассмотрим задачу минимизации дисперсии $\mathbf{D}(\bar{S}_N(f))$ за счет выбора значений $N_1, \dots, N_m$ при фиксированном общем числе узлов $N = N_1 + \dots + N_m$. Введем в рассмотрение функцию Лагранжа

$$\Psi = \mathbf{D}\bar{S}_N(f) - \lambda \left(\sum_{l=1}^m N_l - N \right),$$

приравнивая нулю ее производные по N_l , получим систему уравнений $-D_l(f) N_l^{-2} - \lambda = 0$; решая эту систему уравнений совместно с уравнением $N = N_1 + \dots + N_m$, определяем требуемые значения N_l . Здесь и далее, как обычно, мы пренебрегаем тем обстоятельством, что значения N_l должны быть целыми.

Покажем, что при $N_l = N\mu_l\mu^{-1}$ имеет место неравенство

$$D(\bar{S}_N(f)) \leq D(S_N(f)).$$

Лемма. Пусть ξ — случайная величина. Тогда при любом a

$$M(\xi - a)^2 \geq M(\xi - M\xi)^2 = D\xi.$$

Доказательство предоставляется читателю.

При $N_l = N\mu_l\mu^{-1}$ можно написать цепочку соотношений

$$\begin{aligned} DS_N(f) &= \frac{D(f)}{N} = \frac{1}{N} \int_G \frac{1}{\mu} (\mu f(P) - I(f))^2 dP = \\ &= \sum_{l=1}^m \frac{\mu}{N} \int_{G_l} \left(f(P) - \frac{I(f)}{\mu} \right)^2 dP = \sum_{l=1}^m \frac{\mu}{N\mu_l} \int_{G_l} \mu_l \left(f(P) - \frac{I(f)}{\mu} \right)^2 dP = \\ &= \sum_{l=1}^m \frac{1}{N_l} \int_{G_l} \frac{1}{\mu_l} \left(\mu_l f(P) - \frac{\mu_l}{\mu} I(f) \right)^2 dP = \\ &= \sum_{l=1}^m \frac{1}{N_l} M_l \left(\mu_l f(P) - \frac{\mu_l}{\mu} I(f) \right)^2; \quad (1) \end{aligned}$$

здесь через $M_l g(P)$ обозначается математическое ожидание $g(P)$ при условии, что точка P распределена равномерно в G_l . Обращаясь к (1), на основании леммы получаем неравенство

$$DS_N(f) \geq \sum_{l=1}^m \frac{1}{N_l} M_l \left(\mu_l f(P) - M_l(\mu_l f(P)) \right)^2,$$

но

$$M_l(\mu_l f(P)) = I_l(f), \quad M_l(\mu_l f(P) - I_l(f))^2 = D_l(f),$$

и поэтому правая часть последнего неравенства совпадает с $D(\bar{S}_N(f))$. Таким образом, требуемое соотношение $D(S_N(f)) \geq D(\bar{S}_N(f))$ доказано.

Разбиение области интегрирования на части с целью уменьшения дисперсии метода Монте-Карло широко используется, в частности, при обработке естественнонаучной информации. Пусть требуется определить водосодержание снега в бассейне некоторой реки. При непосредственном применении метода Монте-Карло выбиралось бы независимо, с равномерной плотностью распределения несколько точек, где производилось бы измерение количества водосодержания на единицу поверхности. Участки поверхности с однородными природными условиями (высота над уровнем моря, уровень облачности, залесенность, ориентация склонов гор, осадки, господствующее направление ветра) характеризуются примерно одинаковым водосодержанием. Поэтому удастся добиться существенного повышения точности, разбивая бассейн на части с однородными природными условиями.

§ 12. Квадратурные формулы повышенной точности со случайными узлами

В случае гладкой подынтегральной функции разбиение области интегрирования на части приводит к увеличению порядка скорости сходимости. Пусть вычисляется интеграл

$$I(f) = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s) dx_1 \dots dx_s.$$

Положим $N = n^s$ и разобьем исходную область интегрирования на равные кубы Π_m :

$$\frac{m_1 - 1}{n} \leq x_1 \leq \frac{m_1}{n}, \dots, \frac{m_s - 1}{n} \leq x_s \leq \frac{m_s}{n}.$$

В каждом кубе выберем случайную точку P_m . Считаем, что ее плотность распределения — постоянная, равная n^s , и случайные точки в любых двух кубах выбираются независимо. Положим

$$\bar{S}_N(f) = \frac{1}{n^s} \sum_{m_1, \dots, m_s=1}^n f(P_m).$$

В данном случае условие $N_l = N \mu_l \mu^{-1}$ выполнено. Поэтому, согласно доказанному выше соотношению, имеем

$$D\bar{S}_N(f) \leq DS_N(f) = \frac{D(f)}{N},$$

причем

$$M\bar{S}_N(f) = I(f).$$

Уточним оценку дисперсии в предположении, что функция $f(P)$ удовлетворяет условию Липшица с постоянной A по каждой из переменных. Справедливо равенство

$$D\bar{S}_N(f) = \sum_{m_1, \dots, m_s=1}^n D\left(\frac{1}{n^s} f(P_m)\right) = \sum_{m_1, \dots, m_s=1}^n \frac{1}{n^{2s}} Df(P_m). \quad (1)$$

Проведем грубую оценку величин $Df(P_m)$; имеем равенство

$$Mf(P_m) = \sigma_m = \int_{\Pi_m} n^s f(P) dP;$$

на основании теоремы о среднем

$$\sigma_m = f(\bar{P}_m), \quad \text{где } \bar{P}_m \in \Pi_m,$$

и поэтому

$$Mf(P_m) = f(\bar{P}_m).$$

В то же время

$$\begin{aligned} f(x_1 + \Delta_1, \dots, x_s + \Delta_s) - f(x_1, \dots, x_s) = \\ = (f(x_1 + \Delta_1, \dots, x_s + \Delta_s) - f(x_1 + \Delta_1, \dots, x_{s-1} + \Delta_{s-1}, x_s)) + \\ + (f(x_1 + \Delta_1, \dots, x_{s-1} + \Delta_{s-1}, x_s) - \\ - f(x_1 + \Delta_1, \dots, x_{s-2} + \Delta_{s-2}, x_{s-1}, x_s)) + \dots \\ \dots + (f(x_1 + \Delta_1, x_2, \dots, x_s) - f(x_1, \dots, x_s)). \end{aligned}$$

Отсюда следует, что для функций рассматриваемого класса

$$|f(x_1 + \Delta_1, \dots, x_s + \Delta_s) - f(x_1, \dots, x_s)| \leq A(|\Delta_1| + \dots + |\Delta_s|).$$

Если точка P_m принадлежит Π_m , то каждая из ее координат отличается от соответствующей координаты точки $\bar{P}_m$ не более чем на n^{-1} . Поэтому из последнего неравенства следует оценка

$$|f(P_m) - f(\bar{P}_m)| \leq Asn^{-1}$$

при $P_m \in \Pi_m$ и, следовательно,

$$|f(P_m) - \sigma_m| \leq Asn^{-1}. \quad (2)$$

Всякая случайная величина ξ удовлетворяет неравенству

$$|M\xi| \leq \sup |\xi|.$$

Следовательно,

$$D(f(P_m)) = M(f(P_m) - \sigma_m)^2 \leq \sup_{P_m \in \Pi_m} (f(P_m) - \sigma_m)^2 \leq \left(\frac{As}{n}\right)^2.$$

С помощью этой оценки заключаем, что правая часть (1) не превосходит величины

$$n^s n^{-2s} \left(\frac{As}{n}\right)^2 = \frac{(As)^2}{n^{s+2}}.$$

Обозначив общее число узлов n^s через N , получаем оценку

$$D\bar{S}_N(f) \leq \frac{A^2 s^2}{N^{1+2/s}}. \quad (3)$$

Отсюда на основании неравенства Чебышева (9.1) заключаем: с вероятностью $1 - \eta$ выполняется неравенство

$$|\bar{S}_N(f) - I(f)| \leq \frac{As}{N^{1/s} \sqrt{\eta N}}.$$

Полученная оценка погрешности по порядку лучше, чем оценка погрешности $O(1/\sqrt{N})$ метода Монте-Карло. Мы получили оценку погрешности по вероятности; оказывается, для рассмат-

риваемого метода можно получить и гарантированную оценку погрешности. Умножая (2) на n^{-s} , получаем неравенство

$$\left| \frac{1}{n^s} f(P_m) - \int_{\Pi_m} f(P) dP \right| \leq \frac{A_s}{n^{s+1}}.$$

Величина $\bar{S}_N(f) - I(f)$ может быть представлена в виде суммы таких слагаемых

$$\sum_{m_1, \dots, m_s=1}^n \left(\frac{1}{n^s} f(P_m) - \int_{\Pi_m} f(P) dP \right).$$

Суммируя оценки для этих слагаемых, получим

$$|\bar{S}_N(f) - I(f)| \leq \frac{A_s}{n} = \frac{A_s}{N^{1/s}}.$$

Сопоставляя с теоремой из § 7, заключаем, что рассматриваемый метод имеет гарантированную оценку погрешности, оптимальную по порядку на рассматриваемом классе.

Возникает вопрос, можно ли улучшить на этом классе оценку дисперсии (3). Метод Монте-Карло и другие способы интегрирования, подобные рассматриваемому, где приближенное значение интеграла зависит от некоторых случайных параметров, называют *недетерминированными*. Пусть $S_N(F)$ — приближенное значение интеграла $I(f)$, получаемое при применении некоторого недетерминированного способа вычисления интеграла.

Теорема (без доказательства). *Существуют $d_1(r, A)$, $d_2 > 0$, удовлетворяющие следующему соотношению. Для любого способа вычисления интеграла, где информация о подынтегральной функции используется лишь как информация об ее значениях в N точках, найдется функция $f \in C_r(A)$, для которой*

$$|S_N(f) - I(f)| > \frac{d_1(r, A)}{N^{r+1/2}} \quad (4)$$

с вероятностью d_2 ; здесь, как и ранее, $\frac{1}{r} = \frac{1}{r_1} + \dots + \frac{1}{r_s}$, $S_N(f)$ — приближенное значение интеграла.

Очевидно, что следствием неравенства (4) является неравенство

$$DS_N(f) \geq \frac{d_3(r, A)}{N^{2r+1}},$$

которое, в частности, означает, что оценка (3) не может быть улучшена по порядку.

Теорема (без доказательства). *Можно указать способы интегрирования, для которых имеет место гарантированная*

оценка погрешности

$$|S_N(f) - I(f)| \leq \frac{d_4(r, A)}{N^r}$$

и одновременно

$$MS_N(f) = I(f), \quad \sqrt{DS_N(f)} \leq \frac{d_5(r, A)}{N^{r+1/2}}$$

для всех $f \in C(A)$.

Мы построили выше такой способ при $r_1 = \dots = r_s = 1$. Идея построения таких способов состоит в разбиении исходной области интегрирования на малые части и вычислении интеграла по малой части при помощи некоторой «случайной квадратурной формулы».

Обратимся к классу функций $C_{(2, \dots, 2)}(A, \dots, A)$. Так же, как и выше, в каждом кубе Π_m возьмем по случайной точке P_m . Обозначим через P_m^* точку, симметричную P_m относительно центра куба; положим

$$s_m(f) = \frac{1}{2n^s} (f(P_m) + f(P_m^*)),$$

$$S_N(f) = \sum_{m_1, \dots, m_s=1}^n s_m(f).$$

Точки P_m, P_m^* распределены равномерно в рассматриваемом кубе; поэтому

$$M\left(\frac{1}{n^s} f(P_m)\right) = M\left(\frac{1}{n^s} f(P_m^*)\right) = \int_{\Pi_m} f(P) dP.$$

Отсюда следует, что

$$Ms_m(f) = \int_{\Pi_m} f(P) dP;$$

суммируя эти соотношения, получаем равенство

$$MS_N(f) = I(f).$$

Как и выше, имеем

$$DS_N(f) = \sum_{m_1, \dots, m_s=1}^n DS_m(f) = \frac{1}{n^{2s}} \sum_{m_1, \dots, m_s=1}^n D \frac{(f(P_m) + f(P_m^*))}{2}. \quad (5)$$

Обозначим через $\bar{P}_m$ центр куба Π_m ; вычитая постоянную под знаком дисперсии, получаем равенство

$$D \frac{(f(P_m) + f(P_m^*))}{2} = D \frac{(f(P_m) - 2f(\bar{P}_m) + f(P_m^*))}{2}. \quad (6)$$

При фиксированной точке $P_m = (x_1, \dots, x_s)$ введем в рассмотрение функцию

$$g(t) = f\left(\frac{m_1 - 1/2}{n} + \frac{x_1 - (m_1 - 1/2)}{n} t, \dots, \frac{m_s - 1/2}{n} + \frac{x_s - (m_s - 1/2)}{n} t\right).$$

Очевидно,

$$g(1) = f(P_m), \quad g(0) = f(\bar{P}_m), \quad g(-1) = f(P_m^*).$$

Выразим вторую разность через вторую производную:

$$f(P_m) - 2f(\bar{P}_m) + f(P_m^*) = g(1) - 2g(0) + g(-1) = g''(\theta), \quad |\theta| \leq 1.$$

Непосредственным дифференцированием убеждаемся, что

$$g''(t) = \sum_{i,j=1}^s f_{x_i x_j} \left(\frac{m_i - 1/2}{n} + \frac{n x_i - (m_i - 1/2)}{n} t, \dots \right) \times \\ \times \left(x_i - \frac{m_i - 1/2}{n} \right) \left(x_j - \frac{m_j - 1/2}{n} \right).$$

Так как $(x_1, \dots, x_s) \in \Pi_m$, то

$$\left| x_i - \frac{m_i - 1/2}{n} \right| \leq \frac{1}{2n}.$$

Поэтому

$$|g''(t)| \leq \frac{As^2}{4n^2} \quad \text{при} \quad |t| \leq 1, \\ \frac{|f(P_m) - 2f(\bar{P}_m) + f(P_m^*)|}{2} \leq \frac{As^2}{8n^2}. \quad (7)$$

На основании этого неравенства имеем

$$Ds_m(f) \leq \frac{1}{n^{2s}} \left(\frac{As^2}{8n^2} \right)^2.$$

Подставив эту оценку в правую часть (5), получим

$$DS_N(f) \leq \frac{(As^2)^2 2^{4/s}}{32N^{1+4/s}}, \quad N = 2n^s.$$

Отсюда опять с помощью неравенства Чебышева (9.1) имеем

$$|S_N(f) - I(f)| \leq \frac{As^2 2^{1/2+2/s}}{8N^{2/s} \sqrt{\eta N}}$$

с вероятностью $1 - \eta$.

Для рассматриваемого способа интегрирования можно получить также гарантированную оценку погрешности $O(AN^{-2/s})$. В самом деле, соотношение (7) означает, что значение величины $s_m(f)$ при всех P_m принадлежит отрезку

$$\left[\frac{1}{n^s} \left(f(\bar{P}_m) - \frac{As^2}{8n^2} \right), \frac{1}{n^s} \left(f(\bar{P}_m) + \frac{As^2}{8n^2} \right) \right].$$

Следовательно, этому же отрезку принадлежит и ее математическое ожидание, равное

$$\int_{\Pi_m} f(P) dP.$$

Отсюда заключаем, что при всех P_m величина

$$\left| s_m(f) - \int_{\Pi_m} f(P) dP \right|$$

не превосходит длины этого отрезка, равной $\frac{As^2}{4n^{2+s}}$. Таким образом,

$$\left| s_m(f) - \int_{\Pi_m} f(P) dP \right| \leq \frac{As^2}{4n^{2+s}}.$$

Суммируя эти неравенства по m , получим оценку погрешности

$$|S_N(f) - I(f)| \leq \frac{As^2}{4n^2} = \frac{As^2 2^{2/s}}{4N^{2/s}}.$$

Из сравнения с формулировками теорем об оценках снизу гарантированных и недетерминированных оценок погрешности видно, что в случае класса $C_{(2, \dots, 2)}(A, \dots, A)$ построенный способ имеет оптимальные по порядку оценки скорости сходимости и гарантированные, по вероятности.

§ 13. О выборе метода решения задачи

Обсудим схему решения одной задачи, где оказалось выгодным использование методов Монте-Карло такого рода. Требовалось вычислить серию интегралов

$$I(\alpha_1, \alpha_2, \alpha_3) = \int_G f(P; \alpha_1, \alpha_2, \alpha_3) dP$$

при различных значениях параметров $\alpha_1, \alpha_2, \alpha_3$. Область интегрирования G содержалась в единичном трехмерном кубе; при этом условие принадлежности точки к области G задавалось громоздкой системой неравенств. Было ясно, что нельзя произвести замену независимых переменных так, чтобы область интегрирования G приобрела стандартный вид, где было бы возможным применение квадратурных формул высокого порядка точности. Поэтому пришлось вне G продолжить f нулем и рассматривать исходную задачу как задачу вычисления интеграла

$$\bar{I}(\alpha_1, \alpha_2, \alpha_3) = \int_0^1 \int_0^1 \int_0^1 \bar{f}(P; \alpha_1, \alpha_2, \alpha_3) dx_1 dx_2 dx_3,$$

где

$$\bar{f}(P; \alpha_1, \alpha_2, \alpha_3) = \begin{cases} f(P; \alpha_1, \alpha_2, \alpha_3) & \text{при } P \in G, \\ 0 & \text{при } P \notin G. \end{cases}$$

Подынтегральная функция оказалась теперь разрывной. Для вычисления интеграла были применены формулы прямоугольников и Гаусса. С целью контроля точности проводились расчеты с различным числом узлов интегрирования; оказалось, что результаты расчетов медленно устанавливаются (сильно меняются при изменении числа узлов), что указывает на малую точность получаемых приближенных значений. При непосредственном применении ме-

тогда Монте-Карло установление получаемых приближенных значений было еще хуже. Поэтому было принято решение применить для вычисления описанные выше способы уменьшения дисперсии путем разбиения области на части. Были опробованы оба описанных выше способа. Область интегрирования $0 \leq x_1, x_2, x_3 \leq 1$ разбивалась на равные кубы со стороной $1/n$ и далее применялись оба метода, рассмотренные в предшествующем параграфе. Как и в случае других квадратур, исследовалось установление результатов вычислений. Оказалось, что оба метода могут обеспечить требуемую точность вычислений при приемлемых затратах машинного времени.

Во многих случаях вычисление интегралов высокой кратности или большой серии интегралов удается осуществить лишь за счет устранения повторения одинаковых вычислений. Решение рассматриваемой задачи представлялось сначала бесперспективным вследствие большой трудоемкости проверки условия принадлежности узла P области G ; нахождение каждого значения $f(P; \alpha_1, \alpha_2, \alpha_3)$ требовало существенно меньшего объема вычислений. Так как все интегралы вычислялись по одной и той же области, то было принято решение применить следующий алгоритм. Куб $0 \leq x_1, x_2, x_3 \leq 1$ разбивался на кубики со стороной $1/n$; в каждом кубике выбиралась случайная точка P_m и проверялось условие принадлежности $P_m \in G$, координаты всех точек $P_m \in G$ были записаны в памяти ЭВМ. Далее все интегралы серии заменяли суммами

$$\sum_{P_m \in G} \frac{1}{n^3} f(P_m; \alpha_1, \alpha_2, \alpha_3)$$

с одними и теми же узлами. Отказ от проверки условия $P_m \in G$ при вычислении каждого интеграла привел к снижению требований на затраты машинного времени примерно в 100 раз. Другим фактором, позволившим резко снизить затраты машинного времени, оказался следующий. Подынтегральная функция имела вид

$$g(x_1, \alpha_1) h(x_1, x_2, x_3; \alpha_1, \alpha_2, \alpha_3),$$

где вычисление каждого значения функции h требовало относительно малого числа элементарных операций ЭВМ по сравнению с вычислением каждого значения функции g . Поэтому была применена следующая схема вычислений. Все интегралы $I(\alpha_1, \alpha_2, \alpha_3)$, соответствующие одному и тому же значению параметра α_1 , вычислялись одновременно; благодаря этому каждое значение $g(x_1, \alpha_1)$ вычислялось при вычислении всей серии интегралов только один раз. Такая организация работы позволила довести затраты машинного времени при решении рассматриваемой задачи до приемлемых размеров. Анализ хода решения задачи, проведенный после окончания вычисления серии, показал на ряд неиспользованных возможностей, которые предоставили бы дополнительные удобства как исполнителю, так и заказчику. С целью уменьшения затрат машинного времени заказчик старался уменьшить число M совокупностей значений параметров $(\alpha_1^i, \alpha_2^i, \alpha_3^i)$, при которых требовалось вычислить значение интеграла. Наиболее трудоемкую часть при вычислениях составляло вычисление значений функции $g(x_1, \alpha_1)$. Поэтому общие затраты машинного времени были пропорциональны не числу M , а числу различных значений α_1^i . Таким образом общие затраты времени могли бы быть снижены также за счет удачного выбора совокупности $(\alpha_1^i, \alpha_2^i, \alpha_3^i)$.

В рассматриваемой задаче имелся еще один неиспользованный резерв повышения точности. Трудоемкость рассматриваемого метода пропорциональна произведению числа различных значений α на число различных значений координат x_1 узлов интегрирования. Узлы интегрирования выбирались

случайно, и поэтому число различных значений x_1 было очень большим. Для уменьшения этого числа можно было бы пойти, например, по следующему пути. Аппроксимировать исходный интеграл по формуле прямоугольников

$$\bar{I}(\alpha_1, \alpha_2, \alpha_3) \approx \sum_{q=1}^m \frac{1}{m} I\left(\alpha_1, \alpha_2, \alpha_3, \frac{q-1/2}{m}\right),$$

где

$$I\left(\alpha_1, \alpha_2, \alpha_3, \frac{q-1/2}{m}\right) = \int_0^1 \int_0^1 \bar{f}\left(\alpha_1, \alpha_2, \alpha_3, \frac{q-1/2}{m}, x_2, x_3\right) dx_2 dx_3,$$

и метод Монте-Карло с разбиением области интегрирования применять только для вычисления интегралов $I(\alpha_1, \alpha_2, \alpha_3; x_1)$. Другой подход к решению задачи: разбить отрезок $[0, 1]$ на отрезки $\left[\frac{q-1}{m}, \frac{q}{m}\right]$, на каждом отрезке выбрать случайную точку ξ_q и положить

$$\bar{I}(\alpha_1, \alpha_2, \alpha_3) \approx \frac{1}{m} \sum_{q=1}^m I(\alpha_1, \alpha_2, \alpha_3; \xi_q).$$

Оба указанных выше метода имеют худшую скорость сходимости по сравнению с примененным методом, однако у этих методов вычисляется существенно меньше различных значений функций $g(x_1, \alpha_1)$.

Из последних рассуждений видно, что при вычислении больших серий интегралов (как, впрочем, и в случае больших серий других задач) часто больший эффект достигается не за счет повышения качества метода при решении каждой из задач серии, а за счет лучшей организации вычислений.

При применении процедуры, описанной в § 5, вычисление интеграла высокой кратности как бы сводится к вычислению большого числа интегралов меньшей кратности. Поэтому при вычислении интегралов высокой кратности иногда можно использовать указанные выше резервы повышения эффективности в случае вычисления серий интегралов.

Обратим внимание на следующую опасность, возникающую при одновременном получении результата сразу по всей серии задач.

Предъявляемая к решению совокупность задач соответствует какому-то реальному явлению. Может случиться, что ранее явление не «обсчитывалось» на ЭВМ и предъявленная математическая модель не является удовлетворительной. Тогда в случае одновременного решения задач серии все результаты вычислений окажутся бесполезными. При последовательном решении задач можно уже после получения первых результатов обнаружить рассогласование математической модели с общей картиной явления. Такое рассогласование типично при решении новых задач и обычно оно устраняется не сразу, а после большого числа

пробных просчетов и совместного обсуждения модели заказчиком и исполнителем.

Исполнитель имеет основания быть заинтересованным в обсуждении постановки задачи.

Если исходная постановка задачи окажется неразумной, исполнителю придется потратить много времени на выбор нового алгоритма и написание новой программы. Участвуя в обсуждении, исполнитель может уяснить себе возможные варианты изменения постановок задач и предусмотреть их при составлении программы решения задачи. В связи с этим при решении задач новых типов особенно важно, чтобы программа подразделялась на отдельные функциональные блоки, поддающиеся независимому изменению.

Часто столкновение исходных позиций заказчика (обсчет подробной модели) и исполнителя (минимальный объем работ) приводит к построению упрощенной модели явления, быстрый обсчет которой позволяет решить вопрос о целесообразности рассмотрения более сложной модели.

Рассмотрим другой пример, относящийся к организации работы по выбору метода интегрирования в случае кратных интегралов. Обычно вычисление интегралов по сложной области сводят к вычислению интегралов по области, являющейся прямым произведением областей простейшего вида: отрезков, параллелепипедов, лучей, бесконечных прямых, кругов, сфер, шаров и т. д.

Для таких стандартных областей имеется достаточно совершенных методов интегрирования, и после такого сведения можно применить процедуру, описанную в § 5, или какую-либо аналогичную ей.

Пусть вычисляется интеграл

$$I(f) = \int \int \int_{1 \leq x_1^2 + x_2^2 + x_3^2 \leq 4} f(x_1, x_2, x_3) dx_1 dx_2 dx_3.$$

Его удобно записать в виде

$$I(f) = \int_1^2 \int_{S_1} g(l, \omega) d\omega dl,$$

где S_1 — единичная сфера, $d\omega$ — элемент ее площади,

$$g(l, \omega) = l^2 f(l\omega_1, l\omega_2, l\omega_3),$$

$$\omega = (\omega_1, \omega_2, \omega_3), \quad \omega_1^2 + \omega_2^2 + \omega_3^2 = 1.$$

Предположим, что решено вычислять интеграл при помощи квадратур, являющихся «прямым произведением» квадратур по отрезку $[1, 2]$ и по сфере S_1 .

Под *прямым произведением* квадратуры

$$\int_1^2 h(l) dl \approx \sum_{j=1}^n d_j h(l_j) \quad (1)$$

и квадратуры

$$\int_{S_1} p(\omega) d\omega \approx \sum_{q=1}^m k_q p(\omega_q) \quad (2)$$

здесь понимается квадратура

$$\int_1^2 \int_{S_1} g(l, \omega) d\omega dl \approx \sum_{j=1}^n \sum_{q=1}^m d_j k_q g(l_j, \omega_q).$$

Исследуем зависимость погрешности интегрирования от способа интегрирования и от числа узлов. Для исследования поведения погрешности численного интегрирования по оси l выберем какую-то «базовую» квадратуру по единичной сфере

$$\int_{S_1} p(\omega) d\omega \approx \sum_{q=1}^{m_0} k_q^0 p(\omega_q^0); \quad (3)$$

требуется, чтобы число узлов m_0 было мало и в то же время выражения

$$G(l) = \int_{S_1} g(l, \omega) d\omega \quad \text{и} \quad G^0(l) = \sum_{q=1}^{m_0} k_q^0 g(l, \omega_q^0)$$

имели одинаковый качественный характер поведения по l . Например, можно попытаться взять в качестве (3) квадратуру

$$\int_{S_1} p(\omega) d\omega \approx \frac{2\pi}{3} \sum p(\omega_1, \omega_2, \omega_3),$$

где суммирование производится по 6 точкам пересечения единичной сферы с координатными осями x_1, x_2, x_3 .

Предположим, что для вычисления интеграла по оси l принято решение применить или формулу Гаусса, или формулу Симпсона. Последовательно применяя формулы Гаусса

$$\int_1^2 h(l) dl \approx \sum_{j=1}^n d_j^n h(l_j^n)$$

при $n = n_1, n_2, \dots$ узлах, получаем некоторые величины

$$G_{n_i}^0 = \sum_{j=1}^{n_i} d_j^{n_i} G^0(I_j^{n_i}).$$

Точно так же при числе узлов $n = n'_1, n'_2, \dots$ получаем приближения по формуле Симпсона $S_{n'_i}^0$.

Из рассмотрения поведения всех этих величин можно усмотреть значение предела I^0 , к которому они стремятся. Далее при каждом n среди всех приближений $G_{n_i}^0$ и $S_{n'_i}^0$, с $n_i, n'_i \leq n$ вы-

берем приближение Γ_n , обеспечивающее лучшую точность, и введем функцию

$$\varphi_l(n) = |\Gamma_n - I^0|$$

погрешности численного интегрирования по оси l .

Точно так же фиксируется некоторая базовая квадратура по переменной l (часто это квадратура Гаусса с двумя узлами) и строится функция $\varphi_\omega(m)$ погрешности численного интегрирования по сфере ω . Предположим, что суммарная погрешность есть $R = \varphi_l(n) + \varphi_\omega(m)$. Если значения функции f вычисляются независимо, то трудоемкость метода пропорциональна nm . Минимизируя nm при заданном требовании на точность $\varphi_l(n) + \varphi_\omega(m) \leq \varepsilon$, получаем искомые значения числа узлов n_0 и m_0 . В зависимости от того, какой квадратуре — Гаусса или Симпсона отвечает данное n , выбираем соответствующую квадратуру. В случае сомнений в правомерности использования этой методики остается возможность проверить правильность результата, проведя дополнительное интегрирование с несколько отличными от n_0 и m_0 значениями n и m .

Приведем типичный пример подобной организации выбора способа интегрирования по каждой из осей в случае большой серии интегралов по единичному s -мерному кубу. По каждой из осей применяется формула Гаусса, при выборе числа ее узлов по каждой из осей в качестве базовых квадратур по остальным осям берутся квадратуры Гаусса с двумя узлами. Дробление числа узлов по каждой рассматриваемой оси продолжается до тех пор, пока разность между двумя последующими приближениями не станет менее $\varepsilon \cdot s^{-1}$. После определения таким образом нужного числа узлов $n^{(i)}$, соответствующего каждой оси, производится вычисление интеграла с $n^{(i)}$ узлами по каждой из осей.

Вследствие некоторой ненадежности описанного алгоритма обычно производится проверка правомерности его применения в случае данной серии интегралов: выбирается достаточно представительная подсерия интегралов, и для нее результаты

расчетов по данному алгоритму сравнивают с результатами расчетов при несколько измененных значениях числа узлов по осям.

В одной из стандартных программ многомерного метода Гаусса число узлов $n^{(i)}$ по каждой оси выбирается по описанному выше способу, а затем производится просчет, соответствующий этим $n^{(i)}$, и контрольный просчет с несколько измененными значениями $n^{(i)}$.

Дополнительная литература к главе V

I [8], II [20, 21]; §§ 1—4: II [23, 24, 45], III [24, 25, 33, 34, 35, 44, 45]; § 4: III [4]; § 7: III [10]; § 8: III [13]; §§ 9—12: II [5, 13, 26], III [10, 61]; § 13: II [14].

Часть II

ЗАДАЧИ АЛГЕБРЫ И ОПТИМИЗАЦИИ

Глава VI

ЧИСЛЕННЫЕ МЕТОДЫ АЛГЕБРЫ

К численным методам алгебры традиционно относят численные методы решения систем линейных уравнений, обращения матриц, вычисления определителей, нахождения собственных векторов и собственных значений матриц и корней многочлена.

При формальном подходе решение этих задач не встречает затруднений: решение системы можно найти, раскрыв определители в формуле Крамера; для нахождения собственных значений матрицы достаточно выписать характеристическое уравнение и найти его корни. Однако эти рекомендации встречают возражения со многих сторон.

Так, при непосредственном раскрытии определителей решение системы с m неизвестными требует порядка $m!m$ арифметических операций; уже при $m = 30$ такое число операций недоступно для современных ЭВМ. При сколько-нибудь больших m применение методов с таким порядком числа операций, как показывает следующий подсчет, будет невозможно и в обозримом будущем.

Рассмотрим гипотетическую ЭВМ с параллельным выполнением операций, занимающую объем V и состоящую из элементов, каждый из которых представляет собой куб со стороной Δ . Естественно предположить, что время, затрачиваемое элементом для выполнения какой-либо элементарной операции, не менее Δ/c , где $c = 3 \cdot 10^8$ м/сек — скорость распространения света, максимальная из известных скоростей передачи информации. Тогда общее число элементарных операций, которое данная ЭВМ может осуществлять за 1 сек, оценивается сверху через Vc/Δ^3 . При $V = 1$ км³ и $\Delta = 10^{-8}$ см (порядка радиуса атома) эта величина равна $3 \cdot 10^{57}$. В то же время $100! = 10^{157,9} \dots$

Таким образом, вычисление определителя порядка 100 по явной формуле будет невозможно и в дальнейшем.

Другой причиной, по которой эти классические способы неприменимы даже при малых m , является сильное влияние округлений при вычислениях на окончательный результат. Уже

при $m = 20$ при вычислениях на современных ЭВМ типична аварийная остановка по причине переполнения порядка чисел; даже если такая остановка не происходит, результат вычислений часто далек от истинного значения из-за влияния вычислительной погрешности. Точно так же обстоит дело при нахождении собственных значений матриц с использованием явного выражения характеристического многочлена.

Методы решения алгебраических задач разделяются на точные, итерационные и вероятностные. Классы задач, для решения которых обычно применяются методы этих групп, можно условно назвать, соответственно, классами задач с малым, средним и большим числом неизвестных.

Изменение объема и структуры памяти ЭВМ, увеличение быстродействия ЭВМ и развитие численных методов приводит к смещению границ применения методов в сторону систем более высоких порядков. В настоящее время точные методы обычно применяются для решения систем до порядка 10^3 , итерационные — до порядка 10^6 .

§ 1. Методы последовательного исключения неизвестных

Рассмотрим задачу решения системы $Ax = b$; здесь $A = [a_{ij}]$ — матрица размерности $m \times m$, $\det A \neq 0$

$$b = (a_{1, m+1}, \dots, a_{m, m+1})^T.$$

Метод решения задачи относят к классу точных, если в предположении отсутствия округлений он дает точное решение задачи после конечного числа арифметических и логических операций. Если число ненулевых элементов матрицы системы имеет порядок m^2 , то для используемых в настоящее время точных методов решения таких систем требуемое число операций имеет порядок m^3 . Поэтому для применимости точных методов необходимо, чтобы такой порядок числа операций был приемлем для данной ЭВМ; другие ограничения накладываются объемом и структурой памяти ЭВМ.

Оговорка об «используемых в настоящее время методах» имеет следующий смысл. Существуют методы решения таких систем с меньшим порядком числа операций; однако они не используются из-за сильной чувствительности результата к вычислительной погрешности.

Наиболее известным из точных методов решения систем линейных уравнений является *метод исключения Гаусса*. Рассмотрим одну из его возможных реализаций. В предположении, что

$a_{11} \neq 0$, первое уравнение системы

$$\sum_{j=1}^m a_{1j} x_j = a_{1, m+1}, \quad i = 1, \dots, m, \quad (1)$$

делим на коэффициент a_{11} , в результате чего получаем уравнение

$$x_1 + \sum_{j=2}^m a_{1j}^1 x_j = a_{1, m+1}^1.$$

Затем из каждого из остальных уравнений вычитается первое уравнение, умноженное на соответствующий коэффициент a_{ii} . В результате эти уравнения преобразуются к виду

$$\sum_{j=2}^m a_{ij}^1 x_j = a_{i, m+1}^1, \quad i = 2, \dots, m.$$

Первое неизвестное оказалось исключенным из всех уравнений, кроме первого. Далее в предположении, что $a_{22}^1 \neq 0$, делим второе уравнение на коэффициент a_{22}^1 и исключаем неизвестное x_2 из всех уравнений, начиная со второго, и т. д. В результате последовательного исключения неизвестных система уравнений преобразуется к треугольному виду

$$x_i + \sum_{j=i+1}^m a_{ij}^i x_j = a_{i, m+1}^i, \quad i = 1, \dots, m. \quad (2)$$

Исключение x_i происходит в результате следующих операций: 1) деления i -го уравнения на a_{ii}^{i-1} , 2) вычитания i -го уравнения, умноженного на a_{ik}^{i-1} , из уравнений с номерами $k=i+1, \dots, m$. Первая операция равносильна умножению системы уравнений слева на диагональную матрицу

$$C_i = \begin{vmatrix} 1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & (a_{ii}^{i-1})^{-1} \\ & & & & & \ddots \\ & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{vmatrix};$$

вторая операция равносильна умножению слева на матрицу

$$\tilde{C}_i = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & \alpha_{i+1} & 1 & \\ & & & \vdots & & \ddots \\ & & & \vdots & & & \ddots \\ & & & \alpha_m & & & & 1 \end{pmatrix}.$$

где $\alpha_k = -a_{i,k}^{i-1}$ при $i < k \leq m$.

Таким образом, система (2), получаемая в результате этих преобразований, запишется в виде

$$CAx = Cb,$$

где

$$C = C_m \tilde{C}_{m-1} C_{m-1} \dots \tilde{C}_1 C_1.$$

Произведение левых (правых) треугольных матриц является левой (правой) треугольной матрицей; поэтому матрица C — левая треугольная.

Из формулы для элементов обратной матрицы

$$(A^{-1})_{ij} = \frac{A_{ji}}{\det A}$$

следует, что матрица, обратная к левой (правой) треугольной, является левой (правой) треугольной. Следовательно, матрица $B = C^{-1}$ — левая треугольная.

Введем обозначение $CA = D$; согласно построению, все $d_{ii} = 1$ и матрица D — правая треугольная. Имеем

$$A = C^{-1}D = BD.$$

Равенство $A = BD$ вместе с условиями $d_{ii} = 1$, $i = 1, \dots, m$, образует систему уравнений относительно элементов треугольных матриц B и D : $\sum_{j=1}^m b_{ij}d_{jk} = a_{ik}$. Поскольку $b_{ij} = 0$ при $i < j$ и $d_{jk} = 0$ при $k < j$, то эта система может быть записана в виде

$$\sum_{j=1}^{\min(i, k)} b_{ij}d_{jk} = a_{ik} \quad (3)$$

или, что то же самое,

$$\sum_{j=1}^k b_{ij} d_{jk} = a_{ik} \quad \text{при} \quad k \leq i,$$

$$\sum_{j=1}^i b_{ij} d_{jk} = a_{ik} \quad \text{при} \quad i < k;$$

воспользовавшись условием, что все $d_{ii} = 1$, получаем систему рекуррентных соотношений для определения элементов b_{ij} и d_{ij} :

$$b_{ik} = a_{ik} - \sum_{j=1}^{k-1} b_{ij} d_{jk} \quad \text{при} \quad k \leq i, \quad (4)$$

$$d_{ik} = \frac{a_{ik} - \sum_{j=1}^{i-1} b_{ij} d_{jk}}{b_{ii}} \quad \text{при} \quad i < k.$$

Вычисления проводятся последовательно для совокупностей

$$(i, k) = (1, 1), \dots, (1, m), (2, 1), \dots, (2, m), \dots$$

$$\dots, (m, 1), \dots, (m, m).$$

Здесь и далее в случае, когда верхний предел суммирования меньше нижнего, считается, что вся сумма равна нулю.

Таким образом, вместо последовательных преобразований системы (1) к виду (2) можно непосредственно произвести вычисление матриц B и D с помощью формул (4). Эти вычисления можно осуществить, если только все элементы b_{ii} скажутся отличными от нуля. Пусть A_k, B_k, D_k — матрицы главных миноров k -го порядка матриц A, B, D . Согласно (3), $A_k = B_k D_k$; поскольку $\det D_k = 1$, $\det B_k = b_{11} \dots b_{kk}$, то $\det A_k = b_{11} \dots b_{kk}$. Следовательно,

$$b_{kk} = \frac{\det A_k}{\det A_{k-1}}, \quad \det A_0 = 1.$$

Итак, для осуществимости вычислений по формулам (4) необходимо и достаточно выполнение условий

$$\det A_k \neq 0, \quad k = 1, \dots, m-1. \quad (5)$$

Осуществление этих вычислений потребует $(2m^3 - 3m^2 + m)/6$ умножений, столько же сложений и $m(m-1)/2$ делений. Введем обозначение $Cb = d = (d_{1, m+1}, \dots, d_{m, m+1})^T$. Поскольку $C^{-1} = B$ и $CA = D$, то справедливы равенства $Bd = b$, $Dx = d$. Таким образом, после разложения матрицы исходной системы на произведение левой и правой треугольных матриц $A = BD$ решение исходной системы сводится к последовательному

решению двух систем $Bd = b$, $Dx = d$ с треугольными матрицами; это потребует $2m^2 - m$ арифметических операций.

Последовательность операций по разложению матрицы A на произведение треугольных и по определению вектора d часто удобно объединить. Уравнения

$$\sum_{j=1}^i b_{ij} d_{j, m+1} = a_{i, m+1}$$

системы $Bd = b$ можно записать в виде (3)

$$\sum_{j=1}^{\min(i, m+1)} b_{ij} d_{j, m+1} = a_{i, m+1}.$$

Следовательно, значения $d_{i, m+1}$ могут вычисляться одновременно с остальными значениями d_{ij} по формулам (4). Совокупность операций, приводящую исходную систему к виду $Dx = d$, называют обычно *прямым ходом метода Гаусса*, а совокупность операций по решению системы $Dx = d$ — *обратным ходом метода Гаусса*.

При вычислениях без помощи ЭВМ велика вероятность случайных ошибок. Для устранения таких ошибок вводят *контрольный столбец системы* $a_{m+2} = (a_{1, m+2}, \dots, a_{m, m+2})^T$, состоящий из контрольных элементов уравнений системы

$$a_{i, m+2} = \sum_{j=1}^{m+1} a_{ij}.$$

При преобразовании уравнений системы над контрольными элементами производятся те же операции, что и над свободными членами уравнений. В результате этого контрольный элемент каждого нового уравнения должен равняться сумме коэффициентов этого уравнения; большое расхождение между ними указывает на ошибки в вычислениях или на неустойчивость алгоритма вычислений по отношению к вычислительной погрешности.

К примеру, в случае приведения системы уравнений $Ax = b$ к виду $Dx = d$ с помощью формул (4) контрольный элемент $d_{i, m+2}$ каждого из уравнений системы $Dx = d$ вычисляется по тем же формулам (4). После вычисления всех элементов d_{ij} при фиксированном i контроль осуществляется проверкой равенства

$$\sum_{j=1}^{m+1} d_{ij} = d_{i, m+2}.$$

Обратный ход метода Гаусса также сопровождается вычислением контрольных элементов строк системы.

В ряде случаев заранее известно, что условие (5) выполнено; например, многие задачи математической физики сводятся к решению систем с положительно определенной матрицей A .

Однако в общем случае этого заранее сказать нельзя; возможна и такая ситуация: все $\det A_{hk} \neq 0$, но среди величин b_{hk}

есть очень малые; при делении на них будут получаться большие числа с большими абсолютными погрешностями. В результате этого решение сильно исказится.

Чтобы избежать катастрофического влияния вычислительной погрешности, применяют *метод Гаусса с выбором главного элемента*. Его отличие от описанной выше схемы метода Гаусса состоит в следующем. Пусть по ходу исключения неизвестных получена система уравнений

$$x_i + \sum_{j=i+1}^m a_{ij}^i x_j = a_{i, m+1}^i, \quad i = 1, \dots, k,$$

$$\sum_{j=k+1}^m a_{ij}^k x_j = a_{i, m+1}^k, \quad i = k+1, \dots, m.$$

Найдем l такое, что $|a_{k+1, l}^k| = \max_j |a_{k+1, j}^k|$, и переобозначим $x_{k+1} = x_l$ и $x_l = x_{k+1}$; далее произведем исключение неизвестной x_{k+1} из всех уравнений, начиная с $k+2$ -го. Такое переобозначение приводит к изменению порядка исключения неизвестных и во многих случаях существенно уменьшает чувствительность решения к округлениям при вычислениях.

Большинство распространенных точных методов решения систем линейных уравнений можно рассматривать как варианты метода Гаусса, отличающиеся между собой лишь некоторыми деталями. Количество арифметических операций для всех таких методов оказывается примерно одним и тем же, однако некоторые варианты метода Гаусса требуют несколько меньшего объема используемой памяти ЭВМ; так при используемой памяти в N ячеек методами *оптимального исключения и окаймления* можно решать системы порядка $\sim 2\sqrt{N}$. *Метод отражений*, несколько отличный от метода Гаусса, удобен однородной схемой вычислений.

Часто требуется решить несколько систем уравнений $Ax = b_q$, $q = 1, \dots, p$, с одной и той же матрицей A . Удобно поступить следующим образом: введя обозначения

$$b_q = (a_{1, m+q}, \dots, a_{m, m+q})^T,$$

произведем вычисления по формулам (4), причем элементы d_{ik} вычислим при $i < k \leq m+p$.

В результате будут получены p систем уравнений с треугольной матрицей, соответствующих исходной задаче

$$Dx = d_q, \quad d_q = (d_{1, m+q}, \dots, d_{m, m+q})^T, \quad q = 1, \dots, p;$$

затем решаем эти системы каждую в отдельности. Оказывается, что общее число арифметических действий при решении p систем уравнений таким способом $\sim \frac{2}{3} m^3 + 2pm^2$.

Описанный выше прием иногда используется для того, чтобы без существенных дополнительных затрат получить суждение о погрешности решения, являющейся следствием округлений при вычислениях. Задаются вектором z с компонентами, имеющими по возможности тот же порядок и знак, что и компоненты искомого решения; часто из-за отсутствия достаточной информации берут $z = (1, \dots, 1)^T$. Вычисляется вектор $c = Az$ и наряду с исходной системой уравнений решается система $Az = c$.

Пусть x' и z' — реально получаемые решения этих систем. Суждение о погрешности $x' - x$ искомого решения можно получить, основываясь на гипотезе: относительные погрешности при решении методом исключения систем с одной и той же матрицей и различными правыми частями

$$\frac{\|x - x'\|}{\|x'\|} \quad \text{и} \quad \frac{\|z - z'\|}{\|z'\|}$$

отличаются не в очень большое число раз.

Другой прием для получения суждения о реальной величине погрешности, вследствие округлений при вычислениях, состоит в *изменении масштабов*, меняющем картину накопления вычислительной погрешности. Наряду с исходной системой тем же методом решается система

$$(\alpha A)x' = \beta b, \quad \alpha \quad \text{и} \quad \beta - \text{числа};$$

при α и β , не являющихся целыми степенями двойки, сравнение векторов x и $\frac{\alpha}{\beta} x'$ дает представление о величине вычислительной погрешности. Например, можно взять $\alpha = \beta = \sqrt{2}$.

Для решения систем линейных уравнений с эрмитовой матрицей A удобен *метод квадратного корня*. Матрица A представляется в виде

$$A = S^* D S, \quad (6)$$

где S — правая треугольная, S^* — сопряженная к ней,

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots \\ 0 & s_{22} & \dots \\ \dots & \dots & \dots \end{pmatrix},$$

причем все $s_{ii} > 0$, D — диагональная матрица с элементами d_{ii} , равными $+1$ или -1 . Матричное равенство (6) образует систему уравнений

$$a_{ij} = \bar{s}_{1i} s_{1j} d_{11} + \dots + \bar{s}_{ii} s_{ij} d_{ii} \quad \text{при} \quad i \leq j.$$

Аналогичные уравнения при $i > j$ отброшены, так как уравнения, соответствующие парам (i, j) и (j, i) , эквивалентны. Отсюда получаем рекуррентные формулы для определения элемен-

тов d_{ii} и s_{ij} :

$$d_{ii} = \text{sign} \left(a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk} \right),$$

$$s_{ii} = \sqrt{\left| a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk} \right|},$$

$$s_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} \bar{s}_{ki} s_{kj} d_{kk}}{\bar{s}_{ii} d_{ii}} \quad \text{при } i < j.$$

Матрица S является правой треугольной, и таким образом, после получения представления (6) решение исходной системы также сводится к последовательному решению двух систем с треугольными матрицами. Заметим, что в случае $A > 0$ все $d_{ii} = 1$ и $A = S^*S$.

Если есть подозрение, что реально полученное решение x^1 сильно искажено вычислительной погрешностью, то можно поступить следующим образом. Определим вектор $b^1 = b - Ax^1$. Очевидно, ошибка $r^1 = x - x^1$ удовлетворяет системе уравнений

$$Ar^1 = Ax - Ax^1 = b^1. \quad (7)$$

Решая эту систему в условиях реальных округлений, получаем приближение $r^{(1)}$ к r^1 . Полагаем $x^2 = x^1 + r^{(1)}$. Если точность нового приближения представляется неудовлетворительной, то повторяем эту операцию. При решении системы (7) над компонентами правой части производятся те же линейные операции, что и над компонентами правой части при решении системы (1). Поэтому при вычислениях на машине с плавающей запятой естественно ожидать, что относительные погрешности решений этих систем будут одного порядка. Поскольку погрешности округлений обычно малы, то $\|b^1\| \ll \|b\|$; тогда $\|r^1\| \ll \|x^1\|$, и, по-видимому, решение (7) определится с существенно меньшей абсолютной погрешностью, чем решение системы (1). Таким образом, применение описанного приема приводит к повышению точности приближенного решения.

Особенно удобно применять этот прием, когда по ходу вычислений в памяти ЭВМ сохраняются матрицы B и D . Тогда для каждого уточнения требуется найти вектор $b^h = b - Ax^h$ и решить две системы с треугольными матрицами. Это потребует всего $\sim 4m^2$ арифметических операций, что составит малую долю от $\sim 2/3 m^3$ операций, требующихся для представления матрицы A в виде $A = BD$.

Идея описанного приема последовательного уточнения приближений к решению часто реализуется в такой форме. Пусть матрица B близка в каком-то смысле к матрице A , но решение системы $Bx = c$ требует существенно

меньшего объема вычислений по сравнению с решением системы $Ax = b$. Решение системы $Bx = b$ принимаем в качестве первого приближения x^1 к решению. Разность $x - x^1$ удовлетворяет системе уравнений

$$A(x - x^1) = b - Ax^1.$$

Вместо решения этой системы находим решение системы

$$Bx^1 = b - Ax^1$$

и полагаем $x^2 = x^1 + x^1$. Таким образом, каждое приближение находится из предыдущего по формуле

$$x^{n+1} = x^n + B^{-1}(b - Ax^n) = (E - B^{-1}A)x^n + B^{-1}b.$$

Если матрицы A и B достаточно близки, то матрица $E - B^{-1}A$ имеет малую норму и такой итерационный процесс быстро сходится (см. § 3).

Значительно более редкой, чем задача решения системы уравнений, является задача обращения матриц. Для обратной матрицы $X = A^{-1}$ имеем равенство $AX = BDX = E$. Таким образом, для нахождения матрицы X достаточно последовательно решить две матричные системы $BY = E$, $DX = Y$. Нетрудно подсчитать, что общий объем вычислений при нахождении матрицы $A^{-1} \sim 2m^3$ арифметических операций.

В случае необходимости уточнения приближения к обратной матрице могут производиться при помощи итерационного процесса $X_k = X_{k-1}(2E - AX_{k-1})$. Для исследования сходимости итерационного процесса рассмотрим матрицы $G_k = E - AX_k$. Имеем равенство

$$G_k = E - AX_k = E - AX_{k-1}(2E - AX_{k-1}) = (E - AX_{k-1})^2 = G_{k-1}^2.$$

Отсюда получаем цепочку равенств

$$G_k = G_{k-1}^2 = G_{k-2}^4 = \dots = G_0^{2^k}.$$

Поскольку

$$A^{-1} - X_k = A^{-1}(E - AX_k) = A^{-1}G_k = A^{-1}G_0^{2^k},$$

то имеем оценку

$$\|A^{-1} - X_k\| \leq \|A^{-1}\| \|G_0\|^{2^k};$$

таким образом, при достаточно хорошем начальном приближении, иначе, если $\|E - AX_0\| < 1$, этот итерационный процесс быстро сходится.

При решении систем больших порядков величины, необходимые в процессе решения системы, не могут одновременно уместиться в оперативной памяти ЭВМ; поэтому встает проблема размещения этих величин в элементах памяти различных уровней и организации процесса вычислений, при котором передача этих величин из одного места памяти в другое по возможности

$y = (y_1, \dots, y_m, 1)$, удовлетворяющий условиям $(a_i, y) = 0$, $i = 1, \dots, m$, в виде

$$y = \sum_{j=1}^{m+1} d_j v_j.$$

Соотношения $(a_i, y) = 0$ образуют систему уравнений

$$\sum_{j=1}^{m+1} d_j (a_i, v_j) = 0, \quad i = 1, \dots, m;$$

кроме того, условие, что $(m+1)$ -я координата вектора y равна 1, дает уравнение

$$\sum_{j=1}^{m+1} d_j (v_j)_{m+1} = 1;$$

здесь и далее $(v)_k$ — k -я координата вектора v . Если вычисления производятся без округлений, то, вследствие свойств построенной совокупности векторов, $(a_i, v_j) = 0$ при $i < j$ и рассматриваемая система имеет вид

$$\left. \begin{aligned} d_1 (a_1, v_1) &= 0, \\ d_1 (a_2, v_1) + d_2 (a_2, v_2) &= 0, \\ d_1 (a_m, v_1) + \dots + d_m (a_m, v_m) &= 0, \\ d_1 (v_1)_{m+1} + \dots + d_{m+1} (v_{m+1})_{m+1} &= 1. \end{aligned} \right\} \quad (1)$$

Решением этой системы является вектор $d^0 = (0, \dots, 0, ((v_{m+1})_{m+1})^{-1})$.

При наличии округлений в матрице системы (1) выше диагонали будут стоять малые, но отличные от нуля элементы. Запишем получающуюся систему в виде $Cd + Dd = g$, $g = (0, \dots, 0, 1)^T$; здесь C — матрица системы (1),

$$D = \begin{pmatrix} 0 & (a_1, v_2) & (a_1, v_3) & \dots & (a_1, v_{m+1}) \\ 0 & 0 & (a_2, v_3) & \dots & (a_2, v_{m+1}) \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

Задаваясь начальным приближением d^0 , будем определять уточненные значения d^k из итерационного процесса

$$Cd^{k+1} + Dd^k = g. \quad (2)$$

Это соотношение можно переписать в виде

$$d^{k+1} = -C^{-1}Dd^k + C^{-1}g.$$

Вследствие малости элементов матрицы D норма матрицы $-C^{-1}D$ мала; поэтому (см. § 3) последовательность приближений, получаемых из этого итерационного процесса, сходится очень быстро. Этот прием приводит к некоторому уменьшению влияния вычислительной погрешности. Чтобы еще более обезопасить себя от большого влияния вычислительной погрешности, подобный прием применяют на каждом шаге ортогонализации. При наличии округлений возможно, что $(v_i, v_j) \neq 0$ при $i \neq j$; поэтому условия $(u_k, v_j) = 0$ при $j < k$ образуют систему уравнений

$$(a_k, v_j) - \sum_{i=1}^{k-1} c_{ki} (v_i, v_j) = 0$$

с матрицей, близкой к диагональной. Значения c_{ki} находят из итерационного процесса вида (2).

§ 3. Метод простой итерации

Простейшим итерационным методом решения систем линейных уравнений является *метод простой итерации*. Система уравнений

$$Ax = b \quad (1)$$

преобразуется к виду

$$x = Bx + c, \quad (2)$$

и ее решение находится как предел последовательности

$$x^{n+1} = Bx^n + c. \quad (3)$$

Очевидно, всякая система

$$x = x - D(Ax - b) \quad (4)$$

имеет вид (2) и при $\det D \neq 0$ эквивалентна системе (1). В то же время всякая система (2), эквивалентная (1), записывается в виде (4) с матрицей $D = (E - B)A^{-1}$.

Напомним определения основных норм в пространствах векторов и матриц. Если в пространстве векторов $x = (x_1, \dots, x_m)^T$ введена норма $\|x\|$, то согласованной с ней нормой в пространстве матриц A называют норму

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Наиболее употребительны нормы

$$\|x\|_1 = \max_{1 \leq j \leq m} |x_j|,$$

$$\|x\|_2 = \sum_{j=1}^m |x_j|,$$

$$\|x\|_3 = \sqrt{\sum_{j=1}^m |x_j|^2} = \sqrt{(x, x)}.$$

Согласованные с ними нормы в пространстве матриц:

$$\|A\|_1 = \max_{1 \leq i \leq m} \left(\sum_{j=1}^m |a_{ij}| \right), \quad (5)$$

$$\|A\|_2 = \max_{1 \leq j \leq m} \left(\sum_{i=1}^m |a_{ij}| \right), \quad (6)$$

$$\|A\|_3 = \sqrt{\max_{1 \leq i \leq m} \lambda_{A^*A}^i}; \quad (7)$$

здесь и далее λ_D^i — собственные значения матрицы D .

Приведем вывод этих соотношений:

$$\begin{aligned}\|Ax\|_1 &= \max_i \left| \sum_j a_{ij} x_j \right| \leqslant \\ &\leqslant \max_i \left(\sum_j |a_{ij}| \max_j |x_j| \right) \leqslant \max_i \left(\sum_j |a_{ij}| \right) \max_j |x_j|,\end{aligned}$$

т. е.

$$\frac{\|Ax\|_1}{\|x\|_1} \leqslant \max_i \left(\sum_j |a_{ij}| \right).$$

Пусть $\max_i \left(\sum_j |a_{ij}| \right)$ достигается при $i=l$; для вектора

$$x = (\text{sign}(a_{l1}), \dots, \text{sign}(a_{lm}))^T$$

имеем $\|x\|_1 = 1$,

$$\|Ax\|_1 \geqslant \left| \sum_j a_{lj} x_j \right| = \sum_j |a_{lj}| = \left(\max_i \sum_j |a_{ij}| \right) \|x\|_1.$$

Из этих соотношений следует (5).

Точно так же

$$\begin{aligned}\|Ax\|_2 &= \sum_i \left| \sum_j a_{ij} x_j \right| \leqslant \sum_i \sum_j |a_{ij}| |x_j| = \\ &= \sum_j \left(\sum_i |a_{ij}| \right) |x_j| \leqslant \left(\max_j \sum_i |a_{ij}| \right) \sum_j |x_j|,\end{aligned}$$

т. е.

$$\frac{\|Ax\|_2}{\|x\|_2} \leqslant \max_j \left(\sum_i |a_{ij}| \right).$$

Пусть $\max_j \sum_i |a_{ij}|$ достигается при $j=l$. Для вектора x , у которого лишь одна компонента x_l отлична от нуля, имеем

$$\begin{aligned}\|Ax\|_2 &= \sum_i \left| \sum_j a_{ij} x_j \right| = \sum_i |a_{il}| |x_l| = \\ &= \left(\sum_i |a_{il}| \right) |x_l| = \left(\max_j \sum_i |a_{ij}| \right) \sum_j |x_j| = \left(\max_j \sum_i |a_{ij}| \right) \|x\|_2;\end{aligned}$$

отсюда следует (6).

Согласно определению $\|A\|_3$, имеем

$$\|A\|_3 = \sup_x \frac{\|Ax\|_3}{\|x\|_3} = \sup_x \sqrt{\frac{(Ax, Ax)}{(x, x)}} = \sqrt{\sup_x \frac{(A^*Ax, x)}{(x, x)}}.$$

Матрица A^*A — симметричная, поскольку $(A^*A)^* = A^*(A^*)^* = A^*A$.

Пусть матрица B — симметричная, $e_1, \dots, e_m$ — ортонормированная система ее собственных векторов, $\lambda_1, \dots, \lambda_m$ — соответ-

ствующие собственные значения. Всякий вектор x представим в виде $\sum_i c_i e_i$. Очевидно,

$$(Bx, x) = \left(\sum_i \lambda_i c_i e_i, \sum_i c_i e_i \right) = \sum_i \lambda_i |c_i|^2;$$

поэтому

$$(Bx, x) \leq (\max_i \lambda_i) \sum_i |c_i|^2 = (\max_i \lambda_i) (x, x) \quad (8)$$

и

$$(Bx, x) \geq (\min_i \lambda_i) (x, x). \quad (9)$$

В то же время $\frac{(Be_i, e_i)}{(e_i, e_i)} = \lambda_i$. Из этих соотношений следует, что

$$\sup_x \frac{|(Bx, x)|}{(x, x)} = \max_i |\lambda_i|. \quad (10)$$

Поскольку $(A^*Ax, x) = (Ax, Ax) \geq 0$, то все $\lambda_{A^*A}^i \geq 0$; полагая в (10) $B = A^*A$, получаем

$$\sup_x \frac{|(A^*Ax, x)|}{(x, x)} = \max_i |\lambda_{A^*A}^i| = \max_i \lambda_{A^*A}^i.$$

Из полученных соотношений следует (7).

Отметим важный частный случай. Если A симметричная, то $\lambda_{A^*A}^i = \lambda_A^i = |\lambda_A^i|^2$, поэтому для симметричной матрицы

$$\|A\|_2 = \max_i |\lambda_A^i|. \quad (11)$$

Если $Ax = \lambda x$, то $\|A\| \|x\| \geq \|\lambda x\| = |\lambda| \|x\|$; таким образом, модуль любого собственного значения матрицы A не больше любой ее нормы.

Теорема (о достаточном условии сходимости метода простой итерации). Если $\|B\| < 1$, то система уравнений (2) имеет единственное решение и итерационный процесс (3) сходится к решению со скоростью геометрической прогрессии.

Доказательство. Для всякого решения системы (2) имеем $\|x\| \leq \|B\| \|x\| + \|c\|$; отсюда $\|x\| (1 - \|B\|) \leq \|c\|$, или $\|x\| \leq (1 - \|B\|)^{-1} \|c\|$. Отсюда следует существование и единственность решения однородной системы $x = Bx$, а, следовательно, и системы (2). Пусть X — решение системы (2). Из (2) и (3) получаем уравнение относительно ошибки $r^n = x^n - X$:

$$r^{n+1} = Br^n. \quad (12)$$

Очевидно равенство

$$r^n = B^n r^0. \quad (13)$$

Из (13) следует, что $\|r^n\| \leq \|B\|^n \|r^0\| \rightarrow 0$. Теорема доказана.

Качество итерационного процесса удобно характеризовать скоростью убывания отношения ошибки после n итераций к начальной ошибке

$$s_n = \sup_{x^0 \neq x} \frac{\|r^n\|}{\|r^0\|} = \sup_{r^0 \neq 0} \frac{\|B^n r^0\|}{\|r^0\|} = \|B^n\|.$$

Можно гарантировать, что величина $s_n \leq \varepsilon$, если $\|B\|^n \leq \varepsilon$, т. е. при

$$n \geq n_\varepsilon = \ln(\varepsilon^{-1}) / \ln(\|B\|^{-1}). \quad (14)$$

Если существуют постоянные $\gamma_{\alpha\beta}$, $\gamma_{\beta\alpha}$ такие, что при $x \neq 0$

$$\frac{\|x\|_\beta}{\|x\|_\alpha} \leq \gamma_{\alpha\beta}, \quad \frac{\|x\|_\alpha}{\|x\|_\beta} \leq \gamma_{\beta\alpha},$$

то нормы $\|x\|_\alpha$ и $\|x\|_\beta$ называют эквивалентными. Очевидно,

$$\|r^n\|_\beta \leq \gamma_{\alpha\beta} \|r^n\|_\alpha \leq \gamma_{\alpha\beta} \|B\|_\alpha^n \|r^0\|_\alpha \leq \gamma_{\alpha\beta} \gamma_{\beta\alpha} \|B\|_\alpha^n \|r^0\|_\beta.$$

Таким образом, если условие доказанной теоремы выполнено для какой-то нормы, то утверждение справедливо относительно любой эквивалентной ей нормы.

Любые две нормы в конечномерном пространстве являются эквивалентными; в частности, нормы $\|x\|_1$, $\|x\|_2$, $\|x\|_3$ эквивалентны между собой вследствие цепочки очевидных неравенств

$$\|x\|_1 \leq \|x\|_3 \leq \|x\|_2 \leq m \|x\|_1.$$

Лемма. Пусть все собственные значения λ_i матрицы B лежат в круге $|\lambda| \leq q$, причем собственным значениям, по модулю равным q , соответствуют жордановы клетки размерности 1. Тогда существует матрица $\Lambda = D^{-1}BD$ с нормой $\|\Lambda\|_1 \leq q$.

Доказательство. Положим $\eta = q - \max_{|\lambda_i| < q} |\lambda_i|$. Собственными значениями матрицы $\eta^{-1}B$ будут $\eta^{-1}\lambda_i$. Преобразуем матрицу $\eta^{-1}B$ к жордановой форме

$$D^{-1}(\eta^{-1}B)D = \begin{pmatrix} \eta^{-1}\lambda_1 & \alpha_{12} & 0 & \dots \\ 0 & \eta^{-1}\lambda_2 & \alpha_{21} & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix},$$

где $\alpha_{i, i+1}$ принимают значения 0 или 1.

После умножения на η получим

$$\Lambda = D^{-1}BD = \begin{pmatrix} \lambda_1 & \alpha_{12}\eta & 0 & \dots \\ 0 & \lambda_2 & \alpha_{21}\eta & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}.$$

Если $|\lambda_i| = q$, то, согласно условию леммы, $\alpha_{i, i+1} = 0$, и поэтому $|\lambda_i| + |\alpha_{i, i+1}\eta| = q$; если $|\lambda_i| < q$, то

$$|\lambda_i| + |\alpha_{i, i+1}\eta| \leq \max_{|\lambda_i| < q} |\lambda_i| + \eta = q.$$

Таким образом, $\|A\|_1 = \max_i (|\lambda_i| + |\alpha_{i, i+1}\eta|) \leq q$.

Утверждение леммы потребуется нам в главе VIII. При доказательстве последующей теоремы мы воспользуемся частным случаем леммы.

Теорема (о необходимом и достаточном условии сходимости метода простой итерации). Пусть система (2) имеет единственное решение. Итерационный процесс (3) сходится к решению системы (2) при любом начальном приближении тогда и только тогда, когда все собственные значения матрицы B по модулю меньше 1.

Доказательство достаточности. Возьмем произвольное q в пределах $\max_i |\lambda_i| < q < 1$. Условие леммы выполнено

по отношению к этому q , поэтому существует матрица D такая, что $\|A\|_1 \leq q$ при $A = D^{-1}BD$. Поскольку $B = DAD^{-1}$, то

$$B^n = DAD^{-1}D \dots D^{-1}DAD^{-1} = DA^nD^{-1}.$$

Поэтому

$$\|B^n\|_1 \leq \|D\|_1 \|D^{-1}\|_1 q^n \rightarrow 0$$

и

$$\|x^n - X\|_1 \leq \|D\|_1 \|D^{-1}\|_1 q^n \|x^0 - X\|_1 \rightarrow 0 \quad (15)$$

при $n \rightarrow \infty$.

Если λ_i — координатные орты, $x = (x_1, \dots, x_m)^T$, то $x = \sum_i x_i \lambda_i$

и

$$\|x\| \leq \sum_i |x_i| \|\lambda_i\| \leq \left(\sum_i \|\lambda_i\| \right) \|x\|_1.$$

Поэтому при любой норме имеем

$$\|x^n - X\| \leq \left(\sum_i \|\lambda_i\| \right) \|D\|_1 \|D^{-1}\|_1 q^n \|x^0 - X\|_1 \rightarrow 0. \quad (16)$$

Соотношения (15), (16) означают также, что любые нормы погрешности убывают быстрее любой геометрической прогрессии со знаменателем, большим $\max_i |\lambda_i|$.

Необходимость. Пусть $|\lambda_i| \geq 1$ и e_i — соответствующий собственный вектор матрицы B . Тогда при начальном приближении $x^0 = X + ce_i$, $c \neq 0$ имеем

$$r^0 = ce_i \text{ и } r^n = \lambda_i^n ce_i \not\rightarrow 0 \text{ при } n \rightarrow \infty.$$

Задача. Пусть все собственные значения матрицы B , за исключением простого $\lambda_1 = 1$, лежат внутри единичного круга и система (2) имеет решение X ; очевидно, решением системы будут также все $x = X + ce_1$. Доказать, что итерационный процесс (3) сходится к одному из таких решений.

§ 4. Исследование реального итерационного процесса

Если все собственные значения матрицы B лежат внутри единичного круга, то может показаться, что не возникает никаких проблем относительно поведения метода в реальных условиях ограниченности порядков чисел в машине и присутствия округлений. В обоснование этого впечатления иногда приводят следующий довод — возмущения приближений вследствие округлений равносильны возмущениям начальных условий итерационного процесса. Поскольку процесс сходящийся, «самоисправляющийся», эти возмущения в конце концов затухнут и будет получено хорошее приближение к решению исходной задачи.

Однако при решении некоторых систем возникала следующая ситуация. Все собственные значения матрицы B лежали в круге $|\lambda| \leq 1/2$, а итерационный процесс останавливался после некоторого числа итераций по причине переполнения порядков чисел в ЭВМ. В других случаях такого переполнения не происходило, но векторы x^n , получаемые при вычислениях, не имели тенденции к установлению. Последний случай наиболее опасен. Неопытный пользователь может необоснованно решить, что при условии $\max |\lambda_i| \leq 1/2$ какое-то определенное число итераций, например 100, заведомо достаточно для получения решения с требуемой точностью. Затем он производит эти 100 итераций и выдает полученный результат за требуемый. Наличие подобных явлений послужило толчком к более детальному исследованию итерационных процессов и формированию новых понятий в теории операторов.

Чтобы понять существо явления, полезно построить пример, где это явление прослеживается в явном виде. В качестве модели выберем итерационный процесс, соответствующий двудиагональной матрице

$$B_0 = \begin{bmatrix} \alpha & \beta & 0 & \dots & 0 \\ 0 & \alpha & \beta & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & \beta \\ 0 & 0 & 0 & \dots & \alpha \end{bmatrix}.$$

При возведении B_0 в степень n получается треугольная матрица

$$B_0^n = [b_{ij}^{(n)}] = \begin{pmatrix} \alpha^n & C_n^1 \alpha^{n-1} \beta & C_n^2 \alpha^{n-2} \beta^2 & \dots \\ 0 & \alpha^n & C_n^1 \alpha^{n-1} \beta & \dots \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

с элементами $b_{ij}^{(n)} = C_n^{j-i} \alpha^{n-(j-i)} \beta^{j-i}$. Если $r^0 = (0, \dots, 0, 1)^T$, то

$$r^n = B_0^n r^0 = (b_{1m}^{(n)}, \dots, b_{mm}^{(n)})^T, \quad \|r^n\|_2 = \sum_{i=1}^m |b_{im}^{(n)}|.$$

При $n < m$ последнее выражение упрощается:

$$\begin{aligned} \|r^n\|_2 &= \sum_{i=1}^m C_n^{m-i} |\alpha|^{n-(m-i)} |\beta|^{m-i} = \\ &= \sum_{k=0}^{m-1} C_n^k |\alpha|^{n-k} |\beta|^k = \sum_{k=0}^n C_n^k |\alpha|^{n-k} |\beta|^k = (|\alpha| + |\beta|)^n. \end{aligned}$$

Рассмотрим случай $|\alpha| < 1$, $|\alpha| + |\beta| > 1$, $|\beta|/(1-\alpha) < 1$. Пусть $c = c^0 = (0, \dots, 0, 1)^T$; непосредственно проверяется, что при таком c решением рассматриваемой системы будет

$$X^0 = \left(\frac{1}{1-\alpha} \left(\frac{\beta}{1-\alpha} \right)^{m-1}, \dots, \frac{1}{1-\alpha} \right)^T,$$

очевидно,

$$\|X^0\|_2 \leq \frac{1}{|1-\alpha|} \sum_{k=0}^{\infty} \left| \frac{\beta}{1-\alpha} \right|^k = \frac{1}{|1-\alpha| \left(1 - \left| \frac{\beta}{1-\alpha} \right| \right)} = \omega.$$

При начальном приближении $x^0 = X^0 + c^0$ имеем $r^0 = c^0$ и, согласно проводившимся выше построениям,

$$\|r^n\|_2 = (|\alpha| + |\beta|)^n \quad \text{для } n < m.$$

Выберем m таким, чтобы число $\sigma = \frac{(|\alpha| + |\beta|)^{m-1} - \omega}{m}$ превосходило пределы, допустимые в ЭВМ. Из полученных ранее соотношений следует, что $\|x^{m-1}\|_1 \geq \frac{\|x^{m-1}\|_2}{m} \geq \frac{\|r^{m-1}\|_2 - \|X^0\|_2}{m} \geq \sigma$; поэтому построенный пример обладает следующими свойствами: норма начального приближения невелика, итерационный процесс сходится при отсутствии округлений и ограничения на порядки чисел в ЭВМ, в реальных условиях итерационный процесс остановится не позднее, чем при $n = m - 1$ из-за недопустимо больших значений компонент приближений. Итак, требуемый модельный пример построен.

Конкретный числовой пример, удовлетворяющий наложенным условиям по отношению к большинству современных ЭВМ: $\alpha = -1/2$, $\beta = 5/4$, $m = 120$.

Пусть $|\beta/(1-\alpha)| > 1$, $c_1 - c_2 = \delta c^0$. Тогда для решений x_1, x_2 систем $x_i = B_0 x_i + c_i$, $i = 1, 2$, выполнены соотношения

$$\|x_1 - x_2\|_1 = \delta \|X^0\|_1 = |\beta|^{-1} |\beta/(1-\alpha)|^m \delta.$$

Если $|\beta/(1-\alpha)|^m$ — очень большое число, то ошибка округления правой части, пропорциональная ϵ^0 , будет иметь следствием недопустимо большую погрешность решения.

При $|\beta/(1-\alpha)| > 1$ и m большим рассматриваемая система является одним из важных модельных примеров так называемых плохо обусловленных систем, о которых пойдет речь в § 13.

Обратимся к реальному случаю, когда на каждом шаге вычислений происходят округления. Маловероятно, что присутствие округлений избавит от переполнения там, где переполнение есть при отсутствии округлений; поэтому рассмотрим подробнее случай, когда переполнение не происходит. Вместо x^n реально получаются векторы x^{*n} , связанные соотношениями

$$x^{*n+1} = Bx^{*n} + c + \rho^n,$$

где ρ^n — суммарное округление на шаге итерации.

Отсюда и из (3.2) получается уравнение относительно реальной ошибки $r^{*n} = x^{*n} - X$:

$$r^{*n+1} = \rho^n + Br^{*n}. \quad (1)$$

Выражая каждое r^{*n} через предыдущее, получаем

$$\begin{aligned} r^{*n} &= \rho^{n-1} + Br^{*n-1} = \rho^{n-1} + B(\rho^{n-2} + Br^{*n-2}) = \\ &= \rho^{n-1} + B\rho^{n-2} + \dots + B^{n-1}\rho^0 + B^n r^0. \end{aligned} \quad (2)$$

Как мы видели, норма $\|B_0^n\|$ при $|\alpha| < 1$, $|\alpha| + |\beta| > 1$ имеет следующий характер поведения: при малых n она имеет тенденцию к возрастанию, при больших n стремится к нулю. Можно показать, что максимальное значение $\phi(B_0) = \max_n \|B_0^n\|$ достигается при значении $n = n_0$ порядка m . При таком характере поведения норм $\|B^n\|$ может возникнуть следующая ситуация. Величина $\max_n \|x^{*n}\|$ не настолько велика, чтобы происходило переполнение и остановка ЭВМ; в то же время $\phi(B)2^{-t} \gg \omega$, где ω — максимально допустимая погрешность решения. Поэтому при $n > n_0$ среди слагаемых в правой части (2) всегда имеется слагаемое $B^{n_0}\rho^{n-1-n_0}$ с нормой $\gg \omega$. В результате установление приближений x^n с приемлемой точностью не происходит.

Суммарная вычислительная погрешность $\rho_n = \sum_{j=0}^{n-1} B^{n-1-j}\rho^j$ может оказаться большей не только из-за большой величины отдельных слагаемых, но и из-за того, что их много.

Пусть B симметрична и $\|B\|_3 = \max_i |\lambda_B^i| = \lambda_B^1 < 1$, e^1 — соответствующий λ_B^1 нормированный собственный вектор. Рассмотрим модель округлений, когда на каждом шаге $\rho^j = \rho e^1$, ρ по-

рядка 2^{-i} . Имеем равенство

$$\rho_n = \rho \sum_{j=0}^{n-1} (\lambda_B^1)^j e^1 = \rho \frac{1 - (\lambda_B^1)^n}{1 - \lambda_B^1} e^1.$$

Поскольку число итераций берется таким, что $\|B^n\| \ll 1$, а $\|B^n\| = (\lambda_B^1)^n$, то можно считать $\|\rho_n\| \approx \frac{\rho}{1 - \lambda_B^1}$. Таким обра-

зом, если λ_B^1 близко к 1, суммарное влияние округлений на шагах итерирования может оказаться довольно большим.

Покажем, что такой порядок вычислительной погрешности является неизбежным. Предположим, что вместо системы (3.2) решается система $x = Bx + c + \rho e_1$. Разность $\bar{X} - X$ решений этих систем удовлетворяет соотношению $(\bar{X} - X) = B(\bar{X} - X) + \rho e_1$; отсюда $\bar{X} - X = (E - B)^{-1} \rho e_1 = (1 - \lambda_B^1)^{-1} \rho e_1$. Поэтому погрешность порядка $(1 - \lambda_B^1)^{-1} \rho$ является неустранимой и возмущение приближений, создаваемое в ходе итераций, сравнимо с неустранимой погрешностью. С целью уменьшения числа итераций при переходе к системе $x = Bx + c$ стараются получить систему с возможно меньшим значением $\|B\|$; поэтому можно привести довольно мало примеров решения задач, где λ_B , например, больше 0,9999.

§ 5. Спектр семейства матриц

Пусть имеется семейство квадратных матриц $\{A_m\}$; m — размерность матрицы A_m — пробегает некоторое подмножество множества натуральных чисел. Пусть в пространствах векторов x_m размерности m определены нормы $\|x\|_{(m)}$.

Точку $\lambda \neq \infty$ назовем *точкой спектра* $S = S(\{A_m\})$ семейства матриц $\{A_m\}$, если выполняется следующее условие: при любом $\varepsilon > 0$ найдется матрица A_m и вектор $x_m \neq 0$ такие, что

$$\|A_m x_m - \lambda x_m\|_{(m)} \leq \varepsilon \|x_m\|_{(m)}. \quad (1)$$

Соответственно, точку $\lambda = \infty$ назовем *точкой спектра семейства матриц*, если при любом K найдутся матрица A_m и вектор $x_m \neq 0$ такие, что

$$\|A_m x_m\|_{(m)} \geq K \|x_m\|_{(m)}. \quad (2)$$

Совокупность точек спектра назовем *спектром семейства матриц*.

Если семейство $\{A_m\}$ состоит только из одной матрицы A_m , то, очевидно, тогда спектр семейства совпадает с множеством S_m собственных значений матрицы A_m .

Задача. Доказать, что спектр любого семейства матриц — замкнутое множество.

Примечание. Часто в литературе спектр семейства матриц определяется несколько иным способом. Точку $\lambda \neq \infty$ называют *точкой спектра семейства матриц*, если выполняется следующее условие: при любом $\varepsilon > 0$ найдется бесконечная последовательность матриц A_m и соответствующих им векторов $x_m \neq 0$ таких, что

$$\|A_m x_m - \lambda x_m\|_{(m)} \leq \varepsilon \|x_m\|_{(m)};$$

$\lambda = \infty$ называют *точкой спектра*, если при любом K найдется бесконечная последовательность матриц A_m и векторов $x_m \neq 0$ таких, что

$$\|A_m x_m\|_{(m)} \geq K \|x_m\|_{(m)}.$$

Величину $\rho_m = \sup_{z \in S_m} |z|$ называют *спектральным радиусом матрицы*; величину

$$\rho = \rho(\{A_m\}) = \sup_{z \in S(\{A_m\})} |z|$$

— *спектральным радиусом семейства матриц* $\{A_m\}$. Очевидно, что $\rho_m \leq \|A_m\|_{(m)}$.

Задача. Доказать, что

$$S(\{A_m\}) = \overline{\bigcup_m S_m}, \quad (3)$$

если $A_m = A_m^*$ и $\|x\|_{(m)} = \|x\|_3$ при всех m .

Из справедливости сформулированного утверждения вытекает, что спектр семейства симметричных матриц при норме $\|x\|_3$ определяется спектрами матриц семейства.

В случае несимметричных матриц соотношение (3) может не выполняться. В качестве модели выберем простейшее семейство несимметричных матриц A_m :

$$A_m = \begin{pmatrix} \alpha & \beta & 0 & \dots & 0 \\ 0 & \alpha & \beta & \dots & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & \dots & \beta \\ 0 & 0 & 0 & \dots & \alpha \end{pmatrix}, \quad \alpha, \beta \neq 0.$$

При любом m множество S_m состоит из точки $z = \alpha$ и поэтому множество $\overline{\bigcup_m S_m}$ совпадает с точкой α . Покажем, что мно-

жеству S принадлежат также все точки круга $|z - \alpha| \leq |\beta|$. Возьмем произвольное комплексное ω , $|\omega| \leq 1$. Пусть x_m — вектор с компонентами

$$(x_m)_j = \omega^{j-1} \left(1 - \frac{j-1}{m}\right).$$

Справедливы равенства

$$\begin{aligned}(A_m x_m)_j &= \alpha (x_m)_j + \beta (x_m)_{j+1} = \\ &= \alpha \omega^{j-1} \left(1 - \frac{j-1}{m}\right) + \beta \omega^j \left(1 - \frac{j}{m}\right) = \\ &= (\alpha + \beta \omega) \omega^{j-1} \left(1 - \frac{j-1}{m}\right) - \beta \frac{\omega^j}{m} \quad \text{при } j < m,\end{aligned}$$

$$\begin{aligned}(A_m x_m)_m &= \alpha (x_m)_m = \\ &= \alpha \omega^{m-1} \left(1 - \frac{m-1}{m}\right) = (\alpha + \beta \omega) \omega^{m-1} \left(1 - \frac{m-1}{m}\right) - \beta \frac{\omega^m}{m}.\end{aligned}$$

Таким образом,

$$(A_m x_m)_j = (\alpha + \beta \omega) (x_m)_j - \beta \frac{\omega^j}{m} \quad \text{при всех } j.$$

Отсюда получаем

$$\|A_m x_m - (\alpha + \beta \omega) x_m\|_3 = \sqrt{\sum_{j=1}^m \left|\beta \frac{\omega^j}{m}\right|^2} \leq \sqrt{m \left(\frac{\beta}{m}\right)^2} = \frac{|\beta|}{\sqrt{m}}.$$

В то же время

$$\|x_m\|_3 = \sqrt{\sum_{j=1}^m \left|\omega^{j-1} \left(1 - \frac{j-1}{m}\right)\right|^2} \geq 1.$$

Таким образом,

$$\|A_m x_m - (\alpha + \beta \omega) x_m\|_3 \leq \left(\frac{|\beta|}{\sqrt{m}}\right) \|x_m\|_3.$$

Так как для любого $\varepsilon > 0$ найдется m такое, что $\frac{|\beta|}{\sqrt{m}} < \varepsilon$, то полученное неравенство означает принадлежность точки $\alpha + \beta \omega$ спектру семейства матриц $\{A_m\}$. Поскольку всякая точка круга $|z - \alpha| \leq |\beta|$ записывается в таком виде с $|\omega| \leq 1$, то мы доказали принадлежность этого круга спектру S .

Положим $\varphi_n = \sup_m \|(A_m)^n\|_{(m)}$.

Теорема 1. Если $\rho(\{A_m\}) = \rho < \infty$, то при любом $\sigma > \rho$ выполняется оценка

$$\varphi_n \leq k(\sigma) \sigma^n \quad (4)$$

при всех n ; здесь $k(\sigma)$ не зависит от n .

Доказательство. Пусть D — произвольная матрица с $\|D\| < d < 1$. Рассмотрим матричное уравнение

$$X = DX + C. \quad (5)$$

Очевидно,

$$\|X\| \leq \|D\| \|X\| + \|C\| \leq d \|X\| + \|C\|,$$

откуда

$$\|X\| \leq (1 - d)^{-1} \|C\|. \quad (6)$$

Система (5) является системой линейных уравнений относительно элементов матрицы X . Поскольку $X=0$ при $C=0$ вследствие (6), то эта система однозначно разрешима при любой матрице C . При $C=E$ имеем $X=(E-D)^{-1}$ и, согласно (6),

$$\|(E-D)^{-1}\| \leq (1-d)^{-1}. \quad (7)$$

При любом n справедливо равенство

$$E - D^n = (E - D)(E + D + \dots + D^{n-1}),$$

после умножения на $(E-D)^{-1}$ имеем

$$(E-D)^{-1} - (E-D)^{-1} D^n = E + D + \dots + D^{n-1},$$

откуда

$$\begin{aligned} \|(E-D)^{-1} - (E + \dots + D^{n-1})\| &\leq \\ &\leq \|(E-D)^{-1}\| \|D^n\| \leq (1-d)^{-1} d^n. \end{aligned} \quad (8)$$

Пусть B — произвольная матрица, β — некоторое число, большее ее нормы: $\|B\| < \beta$; при $|z| \geq \beta$ имеем $\|z^{-1}B\| \leq \frac{\|B\|}{\beta} = d < 1$.

Положив в (8) $D = z^{-1}B$, получим

$$\left\| (E - z^{-1}B)^{-1} - \sum_{j=0}^{n-1} (z^{-1}B)^j \right\| \leq \left(1 - \frac{\|B\|}{\beta}\right)^{-1} \left(\frac{\|B\|}{\beta}\right)^n.$$

Таким образом, матричный ряд $\sum_{j=0}^{\infty} z^{-1}B^j$ равномерно сходится к матрице $(E - z^{-1}B)^{-1}$ в области $|z| \geq \beta$. На основании формулы для коэффициентов ряда Лорана имеем

$$B^n = \frac{1}{2\pi i} \oint_{|z|=\beta} z^{n-1} (E - z^{-1}B)^{-1} dz,$$

иначе

$$B^n = \frac{1}{2\pi i} \oint_{|z|=\beta} z^n (zE - B)^{-1} dz. \quad (9)$$

Элементы матрицы $(zE - B)^{-1}$ с точностью до знака равны отношениям миноров матрицы $zE - B$ к ее определителю. Поэтому они аналитичны всюду, кроме точек z , являющихся нулями $\det(zE - B)$, т. е. вне спектра матрицы B . Следовательно, в (9) можно перейти к интегрированию по любому контуру, содержащему внутри себя спектр матрицы B . Возьмем произвольное $\sigma > \rho$. Спектры S_m всех матриц A_m лежат внутри круга

$|z| \leq \sigma$. Поэтому

$$A_m^n = \frac{1}{2\pi i} \oint_{|z|=\sigma} z^n (zE - A_m)^{-1} dz. \quad (10)$$

Для доказательства теоремы 1 потребуется следующая лемма.

Лемма. Если $\sigma > \rho$, то

$$k(\sigma) = \sup_{|z|=\sigma, m} \|(zE - A_m)^{-1}\|_{(m)} < \infty.$$

Доказательство. Поскольку $\sigma > \rho$, то при $|z| = \sigma$ и любом m $\det(zE - A_m) \neq 0$ и $\|(zE - A_m)^{-1}\|_{(m)} < \infty$.

Допустим, что утверждение леммы неверно. Тогда найдется последовательность $(z(q), m(q))$ такая, что

$$|z(q)| = \sigma, \quad w(q) = \|(z(q)E - A_{m(q)})^{-1}\|_{(m(q))} \rightarrow \infty$$

при $q \rightarrow \infty$. Пусть z_0 , $|z_0| = \sigma$ — предельная точка последовательности $z(q)$. Чтобы не производить дополнительных переобозначений, предположим, что $z(q) \rightarrow z_0$. Вследствие определения нормы матрицы при любом q можно указать вектор $x_{m(q)}$, для которого

$$\|(z(q)E - A_{m(q)})^{-1} x_{m(q)}\|_{(m(q))} \geq \frac{w(q)}{2} \|x_{m(q)}\|_{(m(q))},$$

или, все равно, что

$$\|(z(q)E - A_{m(q)}) y_{m(q)}\|_{(m(q))} \leq \frac{2}{w(q)} \|y_{m(q)}\|_{(m(q))},$$

где

$$y_{m(q)} = (z(q)E - A_{m(q)})^{-1} x_{m(q)}.$$

Отсюда следует неравенство

$$\|A_{m(q)} y_{m(q)} - z_0 y_{m(q)}\|_{(m(q))} \leq \left(|z(q) - z_0| + \frac{2}{w(q)} \right) \|y_{m(q)}\|_{(m(q))}.$$

Так как $w(q) \rightarrow \infty$ согласно предположению, а $|z(q) - z_0| \rightarrow 0$, то отсюда вытекает, что точка z_0 с $|z_0| = \sigma$ является точкой спектра; мы пришли к противоречию. Лемма доказана.

Оценивая норму правой части (10), получаем неравенство

$$\|A_m^n\|_{(m)} \leq \frac{1}{2\pi} \oint_{|z|=\sigma} \sigma^n k(\sigma) |dz| = k(\sigma) \sigma^n.$$

Отсюда следует справедливость неравенства (4).

Теорема 2. Если $\rho > 1$, то

$$\varphi = \sup_n \varphi_n = \sup_{n, m} \|A_m^n\|_{(m)} = \infty.$$

Доказательство. Если для данного семейства матриц $\varphi_1 = \infty$, то для него утверждение теоремы выполнено.

Рассмотрим случай, когда $\varphi_1 < \infty$. Для доказательства теоремы 2 достаточно показать, что при любом K существуют m и n , для которых $\|A_m^n\|_{(m)} \geq K$.

При $\rho > 1$ существует некоторое $z \in S$ с $|z| > 1$. Возьмем произвольное K и выберем n из условия $|z|^n \geq 2K$. Положим $\delta = K/(nb^{n-1})$, где $b = \max(\varphi_1, |z|)$. Поскольку z — точка спектра, то существуют m и вектор x_m , для которых

$$\|A_m x_m - z x_m\|_{(m)} \leq \delta \|x_m\|_{(m)}.$$

Очевидно,

$$\begin{aligned} \|A_m^{n-1} z^l (A_m x_m - z x_m)\|_{(m)} &\leq \|A_m\|_{(m)}^{n-1} |z|^l \delta \|x_m\|_{(m)} \leq \\ &\leq \varphi_1^{n-1} |z|^l \delta \|x_m\|_{(m)} \leq b^{n-1} \delta \|x_m\|_{(m)} = \frac{K}{n} \|x_m\|_{(m)}. \end{aligned} \quad (11)$$

Справедливо тождество

$$A_m^n x_m - z^n x_m = \sum_{i=0}^{n-1} A_m^{n-i-1} z^i (A_m x_m - z x_m).$$

После оценки нормы каждого слагаемого в правой части при помощи неравенства (11) получим

$$\|A_m^n x_m - z^n x_m\|_{(m)} \leq K \|x_m\|_{(m)},$$

откуда

$$\begin{aligned} \|A_m^n x_m\|_{(m)} &\geq |z|^n \|x_m\|_{(m)} - K \|x_m\|_{(m)} = \\ &= (|z|^n - K) \|x_m\|_{(m)} \geq K \|x_m\|_{(m)}. \end{aligned}$$

Таким образом, указаны m и n , для которых $\|A_m^n\|_{(m)} \geq K$; теорема доказана.

Из утверждений теорем следует, что спектральный радиус семейства операторов во многом определяет свойства итерационного процесса. В самом деле, если $\rho < 1$, то, вследствие теоремы 1, нормы $\|A_m^n\|_{(m)}$ равномерно ограничены и следует ожидать устойчивости итерационного процесса к округлениям; если же $\rho > 1$, то $\sup_{n, m} \|A_m^n\|_{(m)} = \infty$ вследствие утверждения

теоремы 2 и для некоторых из матриц A_m итерационный процесс $x^{n+1} = A_m x^n + b$ будет или расходящимся или в достаточной мере сильно чувствительным к округлениям при вычислениях.

На основании доказанных теорем нельзя высказать суждения об устойчивости конкретного итерационного процесса к округлениям; при отсутствии четко выраженных связей между элементами матрицы A даже неясно, к какому семейству матриц следует отнести данную матрицу; если проблема выбора семейства $\{A_m\}$ и отыскания его спектра решена, все равно вопрос об устойчивости итерационного процесса остается откры-

тым. При реальном числе двоичных знаков и фиксированной точности вычислений может оказаться, что медленный рост норм $\|A_m^n\|_{(m)}$ с ростом m и n предпочтительнее их равномерной ограниченности большой постоянной. Таким образом, мы не можем сделать конкретных практических выводов. Однако мы достигли понимания того, что несимметричные матрицы большой размерности обладают новыми, неизвестными нам ранее свойствами. Кроме собственных значений и векторов у этих матриц имеются «почти собственные» значения и векторы, т. е. такие λ и y , что $By \approx \lambda y$. Из последнего равенства следует, что $B^n y \approx \lambda^n y$; следовательно, норма ошибки, пропорциональной y , после n шагов итерации возрастает примерно в $|\lambda|^n$ раз. Наличие таких «почти собственных» значений и явилось причиной разбиравшихся нами неприятных эффектов, обнаруженных при расчетах.

§ 6. δ²-процесс практической оценки погрешности и ускорения сходимости

Рассмотрим вопрос об оценке погрешности приближенного решения системы уравнений. Если X^* — приближенное решение системы $AX = b$, то можно написать равенство $A(X^* - X) = AX^* - b$. Отсюда следует оценка погрешности

$$\|X^* - X\| = \|A^{-1}(AX^* - b)\| \leq \|A^{-1}\| \|AX^* - b\|;$$

эта оценка редко применима из-за сложности оценки $\|A^{-1}\|$; поэтому при практическом анализе погрешности приближений, получаемых итерационными методами, обычно вместо этой оценки используется рассматриваемая далее нестрогая, но более простая оценка погрешности.

Напомним критерий разумности практической оценки погрешности: v^n принимается за практическую погрешность приближения x^n , стремящегося к X при $n \rightarrow \infty$, если

$$\|v^n - (x^n - X)\| / \|x^n - X\| \rightarrow 0 \quad \text{при} \quad n \rightarrow \infty. \quad (1)$$

Очевидно, что тогда $\|v^n\| \sim \|x^n - X\|$.

Рассмотрим метод простой итерации $x^{n+1} = Bx^n + c$; для простоты ограничимся случаем, когда матрица B простой структуры (ее жорданова форма диагональна, и поэтому она обладает полной системой собственных векторов).

Пусть λ_i , $i = 1, \dots, m$, — собственные значения матрицы B , занумерованные в порядке убывания $|\lambda_i|$, причем

$$1 > |\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_m|,$$

e_i , $\|e_i\| = 1$ — соответствующие собственные векторы, образующие полную систему,

Разложим вектор r^0 по базису e_i : $r^0 = \sum c_i e_i$. Тогда

$$r^n = x^n - X = B^n r^0 = \sum c_i \lambda_i^n e_i = c_1 \lambda_1^n e_1 + O(|\lambda_2|^n). \quad (2)$$

Здесь и далее выражение $x_n = y_n + O(e_n)$ имеет следующий смысл

$$\|x_n - y_n\| = O(e_n) \quad \text{при} \quad n \rightarrow \infty.$$

Укажем способ построения приближения к вектору $w^n = c_1 \lambda_1^n e_1$ на основании информации, получающейся в ходе вычислений. Согласно (2), имеем

$$x^{n-2} - X = w^n \lambda_1^{-2} + O(|\lambda_2|^n),$$

$$x^{n-1} - X = w^n \lambda_1^{-1} + O(|\lambda_2|^n),$$

$$x^n - X = w^n + O(|\lambda_2|^n).$$

Вычитая друг из друга соседние соотношения, получим

$$\begin{aligned} x^{n-1} - x^{n-2} &= w^n (1 - \lambda_1^{-1}) \lambda_1^{-1} + O(|\lambda_2|^n), \\ x^n - x^{n-1} &= w^n (1 - \lambda_1^{-1}) + O(|\lambda_2|^n). \end{aligned} \quad (3)$$

Отсюда

$$\begin{aligned} (x^n - x^{n-1}, x^n - x^{n-1}) &= \|w^n\|^2 |1 - \lambda_1^{-1}|^2 + O(\|w^n\| |\lambda_2|^n), \\ (x^{n-1} - x^{n-2}, x^n - x^{n-1}) &= \|w^n\|^2 |1 - \lambda_1^{-1}|^2 \lambda_1^{-1} + O(\|w^n\| |\lambda_2|^n). \end{aligned} \quad (4)$$

Положим

$$\lambda_1^{(n)} = \frac{(x^n - x^{n-1}, x^n - x^{n-1})}{(x^{n-1} - x^{n-2}, x^n - x^{n-1})}.$$

Воспользуемся соотношениями (4) и в предположении $c_1 \neq 0$ поделим числитель и знаменатель на $\|w^n\|^2 |1 - \lambda_1^{-1}|^2 \lambda_1^{-1}$:

$$\lambda_1^{(n)} = \frac{\lambda_1 + O\left(\frac{|\lambda_2|^n}{\|w^n\|}\right)}{1 + O\left(\frac{|\lambda_2|^n}{\|w^n\|}\right)}.$$

Поскольку

$$\|w^n\| = |c_1| |\lambda_1|^n, \quad (5)$$

то

$$\lambda_1^{(n)} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right). \quad (6)$$

Поделив второе из соотношений (3) на $1 - (\lambda_1^{(n)})^{-1}$, получим

$$\begin{aligned} \frac{x^n - x^{n-1}}{1 - (\lambda_1^{(n)})^{-1}} &= w^n \frac{1 - \lambda_1^{-1}}{1 - (\lambda_1^{(n)})^{-1}} + O(|\lambda_2|^n) = \\ &= w^n + w^n \frac{\lambda_1 - \lambda_1^{(n)}}{\lambda_1 (\lambda_1^{(n)} - 1)} + O(|\lambda_2|^n). \end{aligned}$$

Из (5), (6) следует $\|w^n(\lambda_1 - \lambda_1^{(n)})\| = O(|\lambda_2|^n)$, поэтому

$$\frac{x^n - x^{n-1}}{1 - (\lambda_1^{(n)})^{-1}} = w^n + O(|\lambda_2|^n).$$

Отсюда и из (2) следует, что

$$x^n - X = v^n + O(|\lambda_2|^n),$$

где $v^n = \frac{x^n - x^{n-1}}{1 - (\lambda_1^{(n)})^{-1}}$; заметим, что, согласно (3), (6) $\|v^n\| = |c_1| |\lambda_1|^n + O(|\lambda_2|^n)$. Из этих равенств вытекает, что v^n удовлетворяет (1), и поэтому его можно принять за практическую погрешность приближения x^n .

В случае $c_1 = \dots = c_l = 0$, $c_{l+1} \neq 0$ проведенные рассуждения останутся в силе, если $|\lambda_{l+1}| > |\lambda_{l+2}|$; во всех соотношениях следует лишь заменить λ_i , c_i , e_i при $i = 1, 2$ на λ_{l+1} , c_{l+1} , e_{l+1} . Описанный способ получения оценки приближенного решения называется δ^2 -процессом.

Если положить $y^n = x^n - v^n$, то $y^n - X = O(|\lambda_2|^n)$, и поэтому y^n , вообще говоря, является лучшим начальным условием для последующих итераций по сравнению с x^n . Присизводя время от времени такие уточнения, иногда удается существенно уменьшить общее число итераций.

Для справедливости приближенного равенства

$$x^n - X \approx v^n \quad (7)$$

необходимо, чтобы в правой части равенства

$$x^n - X = \sum_i c_i \lambda_i^n e_i$$

слагаемое $c_1 \lambda_1^n e_1$ преобладало над остальными. Если это так, то векторы $x^n - x^{n-1}$, $x^{n-1} - x^{n-2}$ приблизительно пропорциональны, и поэтому

$$\mu_n = \frac{|(x^{n-1} - x^{n-2}, x^{n-1} - x^n)|}{\|x^{n-1} - x^{n-2}\| \|x^{n-1} - x^n\|} \approx 1.$$

Таким образом, условие $\mu_n \approx 1$ является необходимым, чтобы приведенные ранее построения были справедливы. Поэтому его можно принять за условие практической применимости (7).

Например, будет разумной следующая схема метода простой итерации с применением δ^2 -процесса ускорения сходимости. Задаются некоторым малым $\eta > 0$; если по ходу итераций оказалось, что $\mu_n \geq 1 - \eta$, то вычисляется v^n и вектор y^n принимается за начальное приближение для последующих итераций. Итерационный процесс прекращается, если $\mu_n \geq 1 - \eta$ и $\|v^n\| \leq \varepsilon$, где ε — требуемая точность.

Если η выбрано очень малым, то условие $\mu_n \geq 1 - \eta$ будет выполняться только после большого числа итераций, ускорение сходимости не будет иметь места. При большом η соотношения, положенные в основу наших построений, выполняются грубо, и поэтому не исключено, что применение

δ^2 -процесса сходимости замедлит итерационный процесс. Картина итераций также осложняется наличием округлений, и поэтому описанная выше схема требует практической отработки на большом числе примеров с целью выбора оптимального η и указания нижней границы значений ε , при которых алгоритм применим. Если однородный итерационный процесс подвергается перестройке (в нашем случае, при переходе от x^n к y^n), то иногда полезно проверить, не ведет ли эта перестройка к ухудшению. В качестве критерия целесообразности перестройки можно взять некоторое соотношение, связывающее нормы невязок для x^n , y^n , например неравенство вида

$$\|(E - B)y^n - c\| \leq q \|(E - B)x^n - c\|.$$

Наше замечание о необходимости указания нижней грани значений ε вызывается следующим обстоятельством. Пусть для определенности $\lambda_1 > 0$. Уже при вычислении x^n по заданному x^{n-1} округления могут возмутить результат на величину δx^n с нормой порядка ρ . Следствием этого может явиться возмущение δy^n , имеющее норму порядка $(1 - \lambda_1)^{-1}\rho$. Отсюда следует, что в случае $\varepsilon < (1 - \lambda_1)^{-1}\rho$ итерационный процесс может никогда не закончиться.

Проведенные построения показывают, что при реализации метода возникает много таких мест, разбор которых требует серьезной математической подготовки и проведения большого числа численных экспериментов. Поэтому, несмотря на «простоту» метода простой итерации, будет вполне оправданным создание стандартной программы этого метода.

§ 7. Оптимизация скорости сходимости итерационных процессов

Рассмотрим простейший итерационный способ решения системы уравнений $Ax = b$:

$$x^{n+1} = x^n - \alpha(Ax^n - b).$$

Мы видели, что скорость сходимости такого итерационного процесса существенно зависит от максимального модуля собственных значений матрицы $B = E - \alpha A$. Если $\lambda_1, \dots, \lambda_n$ — собственные значения матрицы A , то $\max_i |\lambda_i(B)| = \max_i |1 - \alpha\lambda_i|$.

Из рис. 6.7.1 видно, что при наличии действительных собственных значений различных знаков этот максимум > 1 и итерационный процесс расходится.

Обратимся к рассмотрению частного случая, когда все $\lambda_i > 0$. Значения λ_i известны крайне редко, однако довольно типичен случай, когда известна оценка для этих чисел:

$$0 < \mu \leq \lambda_i \leq M < \infty \quad \text{при всех } i.$$

Скорость сходимости итерационного процесса можно характеризовать величиной

$$\rho(\alpha) = \max_{\mu \leq \lambda \leq M} |1 - \alpha\lambda|.$$

Рассмотрим задачу минимизации $\rho(\alpha)$ за счет выбора α .

Для нахождения $\min_{\alpha} \rho(\alpha)$ удобно обратиться к геометрической картине (рис. 6.7.2). Очевидно, $\rho(\alpha) \geq 1$ при $\alpha \leq 0$. При $0 < \alpha \leq M^{-1}$ функция $1 - \alpha\lambda$ неотрицательна и монотонно убывает на $[\mu, M]$ и поэтому $\rho(\alpha) = 1 - \alpha\mu$. При $M^{-1} < \alpha$ величина $1 - \alpha\lambda$ становится отрицательной и модуль ее растет с ростом α ; при некотором $\alpha = \alpha_0$ наступит момент, когда $1 - \alpha_0\mu = = -(1 - \alpha_0M)$, и тогда $\rho(\alpha_0) = |1 - \alpha_0\mu|$. Если $\alpha < \alpha_0$, то $\rho(\alpha) = 1 - \alpha\mu > 1 - \alpha_0\mu = \rho(\alpha_0)$; если $\alpha_0 < \alpha$, то $\rho(\alpha) \geq |1 - \alpha M| = M\alpha - 1 \geq M\alpha_0 - 1 = \rho(\alpha_0)$. Таким образом, значение $\alpha = \alpha_0$ является искомым. Решая уравнение относительно α_0 , получим $\alpha_0 = \frac{2}{M + \mu}$. Отсюда

$$\rho(\alpha_0) = \frac{M - \mu}{M + \mu}.$$

Мы произвели оптимизацию итерационного процесса частного вида, исходя из наглядных соображений.

На примере систем с матрицей $A > 0$ (здесь и далее неравенство $A > 0$ означает, что матрица A — самосопряженная, положительно определена) рассмотрим другие постановки проблем оптимизации скорости сходимости итерационных процессов. Часть полученных результатов может быть применена и к несимметричным матрицам.

Всякая система $Ax = b$ с $\det A \neq 0$, вообще говоря, может быть приведена к виду, где $A > 0$, умножением на матрицу A^* (как говорят, симметризована). В самом деле, система $A^*Ax = A^*b$

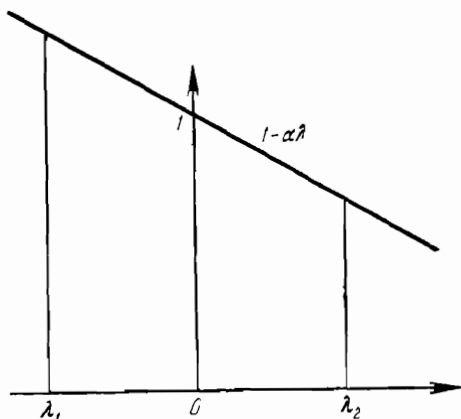


Рис. 6.7.1.

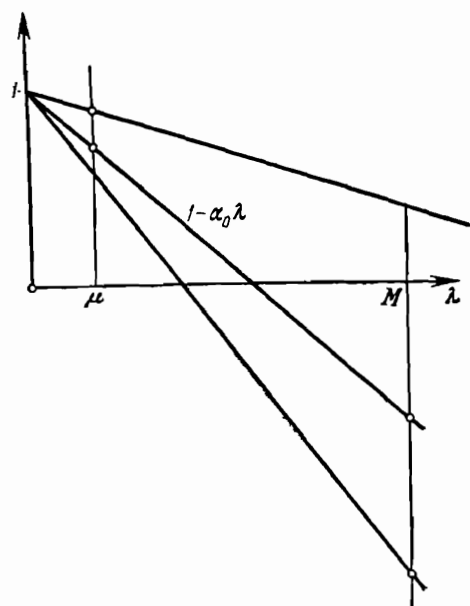


Рис. 6.7.2.

эквивалентна исходной, матрица A^*A — самосопряженная, так как $A^*A = (A^*A)^*$, и положительно определена, так как $(A^*Ax, x) = \|Ax\|^2 > 0$ при $x \neq 0$. Стараются избегать симметризации, поскольку, как мы увидим в конце параграфа, она часто приводит к ухудшению свойств матриц.

Пусть x^0 — начальное приближение, и каждое следующее приближение x^k построено по формуле

$$x^k = x^0 - Q_{k-1}(A)(Ax^0 - b), \quad (1)$$

где $Q_{k-1}(A)$ — многочлен степени $k-1$ от матрицы A .

Предположим, известно, что спектр матрицы A принадлежит отрезку $[\mu, M]$, где $0 < \mu \leq M < \infty$. Для решения X исходной системы имеем

$$X = X - Q_{k-1}(A)(AX - b).$$

Вычитая это равенство из (1), получим уравнение для ошибки

$$r^k = r^0 - Q_{k-1}(A)Ar^0 = P_k(A)r^0,$$

где $P_k(\lambda) = 1 - \lambda Q_{k-1}(\lambda)$. Матрица $P_k(A)$ симметрична, как функция симметричной матрицы; поэтому

$$\sup_{r^0} \frac{\|r^k\|_3}{\|r^0\|_3} = \|P_k(A)\|_3 = \max_{\lambda \in S(A)} |P_k(\lambda)|; \quad (2)$$

введем обозначение $\sigma(g) = \max_{\mu \leq \lambda \leq M} |g(\lambda)|$. Положим

$$\sigma_k = \inf_{Q_{k-1}} \sigma(P_k).$$

Лемма. *Справедливо равенство*

$$\sigma_k = \max_{\mu \leq \lambda \leq M} \left| 1 - \lambda Q_{k-1}^0(\lambda) \right| = \frac{1}{|t_k|}; \quad (3)$$

здесь

$$1 - \lambda Q_{k-1}^0(\lambda) = P_k^0(\lambda) = T_k\left(\frac{2\lambda - (M + \mu)}{M - \mu}\right) \left(T_k\left(-\frac{M + \mu}{M - \mu}\right)\right)^{-1}, \quad (4)$$

как обычно, T_k — многочлены Чебышева, $t_k = T_k\left(-\frac{M + \mu}{M - \mu}\right)$.

Доказательство. Воспользовавшись соотношением $\max_{|x| \leq 1} |T_k(x)| = 1$, получаем $\sigma(P_k^0) = |t_k|^{-1}$. Предположим, что существует многочлен $P_k(\lambda) = 1 - \lambda Q_{k-1}(\lambda)$ такой, что

$$\sigma(P_k) < \sigma(P_k^0) = |t_k|^{-1}. \quad (5)$$

Рассмотрим многочлен $S_k(\lambda) = P_k^0(\lambda) - P_k(\lambda)$. Пусть

$$\lambda^j = \frac{M + \mu}{2} + \frac{M - \mu}{2} \cos \frac{\pi j}{k}, \quad j = 0, \dots, k.$$

Вследствие равенства $T_k(x) = \cos(k \arccos x)$ имеем

$$P_k^0(\lambda^j) = (-1)^j t_k^{-1}.$$

Поскольку $\lambda^j \in [\mu, M]$, то, согласно (5),

$$|P_k(\lambda^j)| < |t_k|^{-1},$$

отсюда следует, что

$$\text{sign } S_k(\lambda^j) = \text{sign } P_k^0(\lambda^j).$$

Точки $\lambda^0, \dots, \lambda^k$ расположены монотонно на отрезке $[\mu, M]$; поскольку $S_k(\lambda)$ меняет знак при переходе от каждой из этих точек к следующей, то $S_k(\lambda)$ имеет на $[\mu, M]$ k корней. Кроме того,

$$S_k(0) = P_k^0(0) - P_k(0) = 1 - 1 = 0.$$

Мы получили многочлен степени k , который имеет $k + 1$ нуль; следовательно, $S_k(\lambda) \equiv 0$, $P_k(\lambda) \equiv P_k^0(\lambda)$,

$$\max_{[\mu, M]} |P_k(\lambda)| = |t_k|^{-1}.$$

Мы пришли к противоречию с (5). Лемма доказана.

У п р а ж н е н и е. Пользуясь формулой

$$T_k(x) = \frac{(x + \sqrt{x^2 - 1})^k + (x - \sqrt{x^2 - 1})^k}{2},$$

проверить справедливость равенства

$$t_k = \frac{(-1)^k}{2} (\lambda_0^k + \lambda_0^{-k}),$$

$$\lambda_0 = \frac{\sqrt{M} + \sqrt{\mu}}{\sqrt{M} - \sqrt{\mu}}.$$

Из этого представления t_k следует, в частности, что

$$\sigma_k = \frac{1}{|t_k|} = \frac{1}{\text{ch}(k \ln \lambda_0)} < 1 \quad \text{при } k > 0.$$

На основе проведенных построений можно предложить несколько типов итерационных процессов.

В одном случае задаются последовательностью значений $n_0 = 0 < n_1 < n_2 < \dots$, приближения x^{n_i} определяют по рекуррентной формуле

$$x^{n_{i+1}} = x^{n_i} - Q_{q_i-1}^1(A)(Ax^{n_i} - b), \quad (6)$$

где $q_i = n_{i+1} - n_i$. Очевидно,

$$r^{n_{i+1}} = P_{q_i}^0(A) r^{n_i},$$

$$\|r^{n_{i+1}}\|_3 \leq \sigma_{q_i} \|r^{n_i}\|_3$$

и в итоге

$$\|r^{n_p}\|_3 \leq \sigma_{q_0} \dots \sigma_{q_{p-1}} \|r^0\|_3.$$

Рассмотрим случай $q_i \equiv k$, т. е. $n_i = ik$. Тогда, обозначив $x^{n_i} = y^i$, можно записать итерационный процесс (6) в виде

$$y^{i+1} = y^i - Q_{k-1}^0(A)(Ay^i - b);$$

соответствующая оценка погрешности

$$\|y^i - X\|_3 \leq (\sigma_k)^i \|y^0 - X\|_3.$$

Такой итерационный процесс называют *оптимальным линейным k -шаговым итерационным процессом*. В частном случае $k = 1$ имеем, согласно (4),

$$P_1^0(\lambda) = 1 - \lambda Q_0^0(\lambda) = \frac{2\lambda - M - \mu}{-\frac{M - \mu}{M + \mu}} = 1 - \lambda \frac{2}{M + \mu},$$

$$t_1 = \frac{1}{-\frac{M + \mu}{M - \mu}}, \quad \sigma_1 = \frac{1}{|t_1|} = \frac{M - \mu}{M + \mu}.$$

Таким образом, *оптимальный линейный одношаговый итерационный процесс* имеет вид

$$y^{i+1} = y^i - \frac{2}{M + \mu} (Ay^i - b);$$

оценка погрешности

$$\|y^i - X\|_3 \leq \left(\frac{M - \mu}{M + \mu}\right)^i \|y^0 - X\|_3. \quad (7)$$

Можно проверить, что коэффициенты многочленов $Q_{k-1}^0(A)$ быстро растут с k ; вследствие этого при больших k алгоритм вычисления x^n , использующий информацию о значениях этих коэффициентов, сильно чувствителен к вычислительной погрешности. Поэтому для вычисления x^k прибегают к окольным путям.

Запишем соотношение (1) в виде

$$x^k = P_k(A)x^0 - (P_k(A) - E)A^{-1}b$$

и будем рассматривать его как преобразование $x^0 \rightarrow x^k$.

Пусть $P^{(1)}(\lambda)$ и $P^{(2)}(\lambda)$ — произвольные многочлены и $P(\lambda) = P^{(1)}(\lambda)P^{(2)}(\lambda)$; произведем последовательно преобразования,

соответствующие многочленам $P^{(1)}$ и $P^{(2)}$:

$$\begin{aligned} \mathbf{x}^{(1)} &= P^{(1)}(A) \mathbf{x}^0 - (P^{(1)}(A) - E) A^{-1} \mathbf{b}, \\ \mathbf{x}^{(2)} &= P^{(2)}(A) \mathbf{x}^{(1)} - (P^{(2)}(A) - E) A^{-1} \mathbf{b}. \end{aligned} \quad (8)$$

Имеем равенства

$$\begin{aligned} \mathbf{x}^{(2)} &= P^{(2)}(A) P^{(1)}(A) \mathbf{x}^0 - P^{(2)}(A) (P^{(1)}(A) - E) A^{-1} \mathbf{b} - \\ &- (P^{(2)}(A) - E) A^{-1} \mathbf{b} = P^{(2)}(A) P^{(1)}(A) \mathbf{x}^0 - (P^{(2)}(A) P^{(1)}(A) - E) A^{-1} \mathbf{b}. \end{aligned}$$

Таким образом, последовательное осуществление преобразований (8) равносильно преобразованию, соответствующему многочлену $P(A)$. Отсюда следует, что при $P(\lambda) = P^{(1)}(\lambda) \dots P^{(k)}(\lambda)$ последовательное осуществление преобразований, соответствующих многочленам $P^{(i)}(\lambda)$, $i = 1, \dots, k$, приводит к тому же результату, что и преобразование, соответствующее многочлену $P(\lambda)$.

Для вычисления $\mathbf{x}^k = \mathbf{x}^0 - Q_{k-1}^0(A)(A\mathbf{x}^0 - \mathbf{b})$ поступим следующим образом. Разложим многочлен

$$P_k^0(\lambda) = 1 - \lambda Q_{k-1}^0(\lambda)$$

на множители

$$P_k^0(\lambda) = (1 - \alpha_1^k \lambda) \dots (1 - \alpha_k^k \lambda)$$

и произведем последовательно преобразования, соответствующие этим множителям. Если $P(\lambda) = 1 - \alpha_i^k \lambda$, то $Q(\lambda) = (1 - P(\lambda)) \lambda^{-1} = \alpha_i^k$, и поэтому эти преобразования имеют вид

$$\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} - \alpha_i^k (A\mathbf{x}^{(i-1)} - \mathbf{b}), \quad i = 1, \dots, k. \quad (9)$$

В результате получим $\mathbf{x}^{(k)}$, равное $\mathbf{x}^k$. Нулями многочленов $P_k^0(\lambda)$ являются $\lambda_i = (\alpha_i^k)^{-1}$, нулями многочленов $T_k\left(\frac{2\lambda - (M + \mu)}{M - \mu}\right)$ являются

$$\lambda_i = \frac{M + \mu}{2} + \frac{M - \mu}{2} \cos \frac{\pi(2i - 1)}{2k}, \quad i = 1, \dots, k.$$

Отсюда

$$\alpha_i^k = \frac{2}{M + \mu + (M - \mu) \cos \frac{\pi(2i - 1)}{2k}}. \quad (10)$$

При больших k алгоритм вычисления по формулам (9), (10) также неустойчив к округлениям. Вычитая из (9) равенство $\mathbf{X} = \mathbf{X} - \alpha_i^k (A\mathbf{X} - \mathbf{b})$, получим уравнение для ошибки

$r^{(i)} = (E - \alpha_i^k A) r^{(i-1)}$. При наличии округлений оно запишется в виде

$$r^{(i)} = (E - \alpha_i^k A) r^{(i-1)} + \rho^{i-1}.$$

Последовательно выражая $r^{(i)}$ через предыдущие, имеем равенство

$$r^{(k)} = \prod_{i=1}^k (E - \alpha_i^k A) r^{(0)} + \sum_{i=0}^{k-1} \prod_{i+2 \leq l \leq k} (E - \alpha_l^k A) \rho^i.$$

В этой формуле к $r^{(0)}$ применяется оператор $P_k^j(A)$ с нормой $\sigma_k = \left| \frac{1}{t_k} \right| < 1$, в то время как операторы, применяемые к ρ^i , могут быть с очень большими нормами.

При реальных вычислениях для обеспечения устойчивости алгоритма к округлениям осуществляют «перемешивание» чисел α_i^k . Алгоритм перемешивания в случае $k = 2^l$ заключается в следующем. Последовательно при $j = 1, \dots, l$ строится «наиболее перемешанная» перестановка чисел $1, \dots, 2^j$. При $j = 1$ она состоит из чисел $1, 2$. Пусть уже построена перестановка $(b_1^{j-1}, \dots, b_{2^{j-1}}^{j-1})$; следующая перестановка берется в виде $(b_1^{j-1}, 2^j + 1 - b_1^{j-1}, b_2^{j-1}, 2^j + 1 - b_2^{j-1}, \dots)$. Например, при $l = 4$ эта перестановка имеет вид $(1, 16, 8, 9, 4, 13, 5, 12, 2, 15, 7, 10, 3, 14, 6, 11)$. Задаваясь $k = 2^l$, строим таблицу чисел $b_1^l, \dots, b_{2^l}^l$ и производим итерации (9) при значениях

$$\alpha_i^k = \frac{2}{M + \mu + (M - \mu) \cos \frac{\pi(2b_i^l - 1)}{2k}}.$$

Зададимся целью построить итерационный процесс, в котором после каждых k операций умножения матрицы A на вектор получается приближение

$$y^k = x^0 - Q_{k-1}^0(A)(Ax^0 - b); \quad (11)$$

согласно (4), имеем

$$Q_{k-1}^0(A) = \left(E - T_k \left(\frac{2A - (M + \mu)E}{M - \mu} \right) t_k^{-1} \right) A^{-1}.$$

Заметим, что $Q_{-1}^0(A) = (E - E)A^{-1} = 0$ и $y^0 = x^0$. Подставив выражение $Q_{k-1}^0(A)$ в (11) и умножив затем обе части на t_k , получим

$$t_k y^k = T_k \left(\frac{2A - (M + \mu)E}{M - \mu} \right) (x^0 - A^{-1}b) + t_k A^{-1}b.$$

Из справедливости этого соотношения при $k = n - 1$, $k = n$, $k = n + 1$ следует равенство

$$\begin{aligned} t_{n+1}y^{n+1} - 2 \frac{2A - (M + \mu)E}{M - \mu} t_n y^n + t_{n-1}y^{n-1} = \\ = \left\{ T_{n+1} \left(\frac{2A - (M + \mu)E}{M - \mu} \right) - 2 \frac{2A - (M + \mu)E}{M - \mu} T_n \left(\frac{2A - (M + \mu)E}{M - \mu} \right) + \right. \\ \left. + T_{n-1} \left(\frac{2A - (M + \mu)E}{M - \mu} \right) \right\} (x^0 - A^{-1}b) + \\ + \left(t_{n+1}E - 2 \frac{2A - (M + \mu)E}{M - \mu} t_n - t_{n-1}E \right) A^{-1}b \quad (12) \end{aligned}$$

Многочлены Чебышева связаны рекуррентным соотношением

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0. \quad (13)$$

Если подставить вместо x некоторую матрицу, в частности матрицу $\frac{2A - (M + \mu)E}{M - \mu}$, то это соотношение также будет выполняться. Поэтому матрица в фигурной скобке $\{ \}$ нулевая. Вследствие (13) и равенства $t_n = T_n \left(-\frac{M + \mu}{M - \mu} \right)$ имеем

$$t_{n+1} - 2 \left(-\frac{M + \mu}{M - \mu} \right) t_n + t_{n-1} = 0. \quad (14)$$

Поэтому соотношение (12) может быть преобразовано к виду

$$t_{n+1}y^{n+1} - 2 \left(-\frac{M + \mu}{M - \mu} \right) t_n y^n + t_{n-1}y^{n-1} = 4 \frac{t_n}{M - \mu} (Ay^n - b). \quad (15)$$

Мы получили требуемое рекуррентное соотношение. Приведем его к виду, более удобному для практического использования. Вследствие (14) равенство (15) можно переписать как

$$\begin{aligned} t_{n+1}y^{n+1} - (t_{n+1} + t_{n-1})y^n + t_{n-1}y^{n-1} = \\ = -\frac{2}{M + \mu} (t_{n+1} + t_{n-1})(Ay^n - b), \end{aligned}$$

или

$$y^{n+1} = y^n + \omega_n \omega_{n-1} (y^n - y^{n-1}) - \frac{2}{M + \mu} (1 + \omega_n \omega_{n-1}) (Ay^n - b), \quad (16)$$

где

$$\omega_n = -\frac{t_n}{t_{n+1}}.$$

Разделив обе части (14) на t_{n+1} , получим

$$1 - 2 \frac{M + \mu}{M - \mu} \omega_n + \omega_n \omega_{n-1} = 0,$$

откуда

$$\omega_n = \frac{1}{2 \frac{M + \mu}{M - \mu} - \omega_{n-1}}. \quad (17)$$

Таким образом, можно рекуррентно вычислять величины ω_n из (17) и затем векторы y^{n+1} из (16). Для получения совокупности векторов $y^1, \dots, y^n$ потребуются произвести n умножений матрицы на вектор, $O(n)$ умножений векторов на числа, сложений векторов и операций с числами. При этом для всех k , вследствие (2) — (4), выполняется оценка

$$\|y^k - X\|_3 \leq \|t_k\|^{-1} \|x^0 - X\|_3. \quad (18)$$

Итерационный процесс (16), (17) называют *оптимальным линейным итерационным процессом*. Это определение имеет следующие основания. Можно показать, что всякое разумное приближение к решению, в ходе получения которого матрица A участвует только в k операциях умножения на вектор, записывается в виде (1). При реализации процесса (16), (17) после каждого k применений матрицы A мы получаем оптимальный результат в смысле (2) — (3). Из сказанного видно, что по скорости сходимости итерационный процесс (16), (17) предпочтительнее, чем (6). Однако иногда от него следует отказываться в пользу (6) из-за соображений экономии памяти ЭВМ.

Получим более наглядную оценку скорости сходимости построенных итерационных процессов.

Согласно явной формуле для многочленов Чебышева, имеем $T_n(x) = \frac{\lambda_1^n + \lambda_2^n}{2}$, где $\lambda_{1,2} = x \pm \sqrt{x^2 - 1}$. Если $x > 1$, то $\lambda_1 = x + \sqrt{x^2 - 1} > 1$, а $0 < \lambda_2 = \frac{1}{\lambda_1} < 1$. Поэтому при больших n

$$T_n(x) \sim \frac{\lambda_1^n}{2}, \quad (19)$$

причем $T_n(x) > \frac{\lambda_1^n}{2}$. Отсюда следует неравенство $\|t_n\|^{-1} \leq 2\lambda_0^{-n}$, где $\lambda_0 = \frac{M + \mu}{M - \mu} + \sqrt{\left(\frac{M + \mu}{M - \mu}\right)^2 - 1}$, и, согласно (18), для оптимального итерационного процесса

$$\|y^n - X\|_3 \leq 2\lambda_0^{-n} \|x^0 - X\|_3.$$

Норма ошибки уменьшится по крайней мере в ϵ^{-1} раз, если $2\lambda_0^{-n} \leq \epsilon$. Отсюда получаем оценку нужного для такого уменьшения числа итераций

$$n \geq n_1 = \log_{\lambda_0} \frac{2}{\epsilon} = (\ln \lambda_0)^{-1} \ln \frac{2}{\epsilon}.$$

Нетрудно проверить, что

$$\lambda_0 = \frac{\sqrt{M} + \sqrt{\mu}}{\sqrt{M} - \sqrt{\mu}} = \frac{1 + \sqrt{\frac{\mu}{M}}}{1 - \sqrt{\frac{\mu}{M}}}.$$

Для многих задач число M/μ оказывается очень большим; при $M/\mu \rightarrow \infty$ имеем

$$\lambda_0 = 1 + 2\sqrt{\frac{\mu}{M}} + O\left(\frac{\mu}{M}\right);$$

таким образом,

$$\ln \lambda_0 \sim 2\sqrt{\frac{\mu}{M}}, \quad n_1 \sim \frac{1}{2}\sqrt{\frac{M}{\mu}} \ln \frac{2}{\varepsilon}.$$

Для сравнения рассмотрим оптимальный линейный одношаговый итерационный процесс, имеющий, согласно (7), оценку погрешности

$$\|x^n - X\|_3 \leq \left(\frac{M-\mu}{M+\mu}\right)^n \|x^0 - X\|_3;$$

отсюда получаем оценку числа итераций

$$n \geq n_2 = \left(\ln \frac{M+\mu}{M-\mu}\right)^{-1} \ln \frac{1}{\varepsilon}.$$

При $\frac{M}{\mu} \rightarrow \infty$ имеем

$$\ln \frac{M+\mu}{M-\mu} \sim 2\frac{\mu}{M} \quad \text{и} \quad n_2 \sim \frac{1}{2}\frac{M}{\mu} \ln \frac{1}{\varepsilon}.$$

Таким образом, в этом сравнении оптимальный итерационный процесс дает выигрыш в числе операций в $\sim \sqrt{\frac{M}{\mu}}$ раз.

Задачи 1. Рассматривается итерационный процесс

$$x^i = x^{i-1} - \alpha_{i-1}(Ax^{i-1} - b), \quad i = 1, 2, \dots$$

Пусть p — нечетное и при всех k совокупности $\alpha_0, \dots, \alpha_{p k-1}$ совпадают с совокупностями

$$\gamma_{i-1} = 2 \left(M + \mu + (M - \mu) \cos \frac{\pi(2i-1)}{2p^k} \right)^{-1}, \quad i = 1, \dots, p^k.$$

Проверить, что при всех $i = p^k$ приближения x^i совпадают с приближениями, получаемыми по оптимальному линейному итерационному процессу.

2. Пусть

$$\alpha_0 = \frac{2}{M + \mu}, \quad \alpha_1 = \frac{1}{M}, \quad \alpha_2 = \frac{1}{\mu}$$

и совокупности

$$\alpha_{2k+1}, \dots, \alpha_{2k+1}$$

при каждом $k > 1$ совпадают с совокупностями величин

$$\gamma_i = 2 \left(M + \mu + (M - \mu) \cos \frac{\pi(2i-1)}{2^{k+1}} \right)^{-1}, \quad i = 1, \dots, 2^k.$$

Показать, что для такого итерационного процесса при всех $n = 2^{k+1}$ справедлива оценка

$$\|x^n - X\|_3 \leq \frac{4}{(\lambda_0^{n-1} - \lambda_0^{1-n})(\lambda_0 - \lambda_0^{-1})} \|x^0 - X\|_3,$$

т. е.

$$\|x^n - X\|_3 = \|x^0 - X\|_3 O(\lambda_0^{-n}).$$

Рассмотрим типичную задачу математической физики, сводящуюся к решению системы уравнений с большим отношением M/μ . Пусть в квадрате $0 \leq x_1, x_2 \leq 1$ решается уравнение Пуассона $-\Delta u = f$ при нулевых условиях на границе. Зададимся сеткой с шагами $h = 1/l$ и напишем систему уравнений, аппроксимирующих дифференциальную задачу:

$$-\frac{u_{j_1+1, j_2} - 2u_{j_1, j_2} + u_{j_1-1, j_2}}{h^2} - \frac{u_{j_1, j_2+1} - 2u_{j_1, j_2} + u_{j_1, j_2-1}}{h^2} = f_{j_1, j_2} \quad (20)$$

при $0 < j_1, j_2 < l$;

$$u_{j_1, j_2} = 0, \quad \text{если} \quad j_1(l - j_1)j_2(l - j_2) = 0.$$

Матрица этой системы положительно определенная и для нее

$$\mu = \min \lambda_i = 8l^2 \sin^2 \left(\frac{\pi h}{2} \right) \sim 2\pi^2,$$

$$M = \max \lambda_i = 8l^2 \cos^2 \left(\frac{\pi h}{2} \right) \sim 8l^2,$$

т. е. $\sqrt{\frac{M}{\mu}} \sim \frac{2l}{\pi}$. При реально применяемом шаге $l = 30$ выигрыш в числе итераций довольно большой.

В начале этого параграфа было упомянуто, что при симметризации системы ее свойства могут ухудшаться. Действительно, пусть этот процесс применяется к уже симметричной матрице A , т. е. переходим от системы $Ax = b$ к системе $A^2x = Ab$. Если у старой системы отношение максимального и минимального собственных значений равнялось M/μ , то у новой оно будет $(M/\mu)^2$ и скорость сходимости итерационных процессов будет меньше.

Примечание. Вследствие (19) при $n \rightarrow \infty$ имеем

$$\omega_n = \frac{T_n \left(\frac{M + \mu}{M - \mu} \right)}{T_{n+1} \left(\frac{M + \mu}{M - \mu} \right)} \rightarrow \frac{1}{\lambda_0}.$$

образом: все корни уравнения

$$\det \begin{pmatrix} a_{11}\lambda & a_{12} & a_{13} & \dots & a_{1m} \\ a_{21}\lambda & a_{22}\lambda & a_{23} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1}\lambda & a_{m2}\lambda & a_{m3}\lambda & \dots & a_{mm}\lambda \end{pmatrix} = 0$$

по модулю меньше 1.

Получим более удобные для применения достаточные критерии сходимости метода Зейделя.

Теорема. Пусть при всех i

$$\sum_{j \neq i} |a_{ij}| \leq q |a_{ii}|,$$

$q < 1$. Тогда

$$\|x^{n+1} - X\|_1 \leq q \|x^n - X\|_1.$$

Доказательство. Вычитая из (2) равенство $BX + CX = b$, которому удовлетворяет точное решение, получим уравнение для ошибки

$$Br^{n+1} + Cr^n = 0. \quad (4)$$

Пусть $\|r^{n+1}\|_1 = |r_l^{n+1}|$; выпишем l -е уравнение системы (4)

$$\sum_{j=1}^{l-1} a_{lj} r_j^{n+1} + a_{ll} r_l^{n+1} + \sum_{j=l+1}^m a_{lj} r_j^n = 0,$$

разрешим его относительно r_l^{n+1} :

$$r_l^{n+1} = - \sum_{j=1}^{l-1} \frac{a_{lj}}{a_{ll}} r_j^{n+1} - \sum_{j=l+1}^m \frac{a_{lj}}{a_{ll}} r_j^n.$$

Отсюда

$$|r_l^{n+1}| \leq \alpha \|r^{n+1}\|_1 + \beta \|r^n\|_1, \quad (5)$$

где

$$\alpha = \sum_{j=1}^{l-1} \left| \frac{a_{lj}}{a_{ll}} \right|, \quad \beta = \sum_{j=l+1}^m \left| \frac{a_{lj}}{a_{ll}} \right|;$$

иначе это соотношение запишется в виде

$$\|r^{n+1}\|_1 \leq \alpha \|r^{n+1}\|_1 + \beta \|r^n\|_1,$$

откуда

$$\|r^{n+1}\|_1 \leq \frac{\beta}{1-\alpha} \|r^n\|_1.$$

Согласно условию теоремы, $\alpha + \beta \leq q < 1$, поэтому

$$\frac{\beta}{1-\alpha} \leq \frac{q-\alpha}{1-\alpha} = q - \frac{\alpha(1-q)}{1-\alpha} \leq q \quad \text{и} \quad \|r^{n+1}\|_1 \leq \frac{\beta}{1-\alpha} \|r^n\|_1 \leq q \|r^n\|_1.$$

Отметим некоторые обстоятельства, помогающие иногда при исследовании сходимости метода Зейделя. Если соотношения (1) умножить на не-

которые множители $\alpha^i \neq 0$, векторы x^n , определяемые из этих соотношений, не изменятся. Таким образом, сходимость метода Зейделя сохраняется при умножении строк матрицы A на какие-либо числа. Если j -й столбец матрицы A умножить на $\beta_j = 0$, а x_j^0 поделить на β_j , то все величины x_j^n разделятся на β_j . Следовательно, сходимость метода Зейделя сохраняется и при умножении столбцов матрицы A на какие-либо числа.

Задача решения систем линейных уравнений является модельной относительно более сложных задач решения систем не-

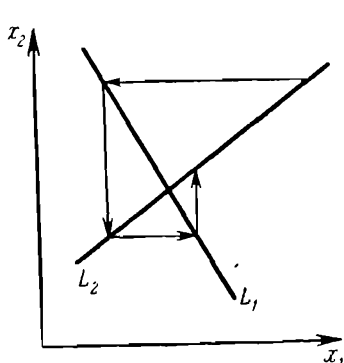


Рис. 6.8.1.

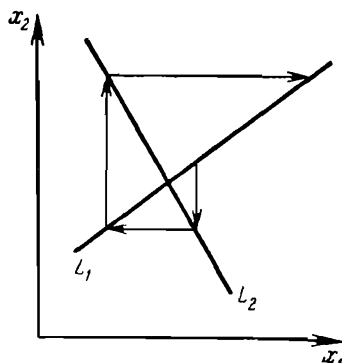


Рис. 6.8.2.

линейных уравнений и минимизации функций многих переменных. Для перенесения метода на более сложные задачи важно понять его наиболее грубые, качественные свойства, обеспечивающие сходимость; с этой целью наиболее желательно получить геометрическую интерпретацию метода. Обозначим через L_i плоскость $\sum_{j=1}^m a_{ij}x_j - b_i = 0$. При получении приближения $(x_1^{n+1}, \dots, x_i^{n+1}, x_{i+1}^n, \dots, x_m^n)$ из приближения $(x_1^n, \dots, x_{i-1}^n, x_i^n, \dots, x_m^n)$ происходит перемещение приближения параллельно оси x_i до пересечения с плоскостью L_i ; таким образом, геометрически метод Зейделя состоит в циклическом перемещении параллельно координатным осям x_i до пересечения с плоскостями L_i . Рис. 6.8.1—6.8.3 иллюстрируют при $m=2$ случаи, когда метод Зейделя сходится, расходится или имеет цикл (как говорят, «зацикливается»). Сравнение первых двух рисунков показывает, что сходимость

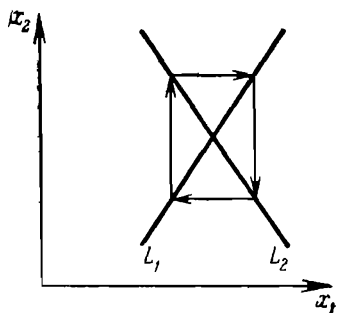


Рис. 6.8.3.

метода Зейделя может изменить характер при перестановке уравнений.

Особенно интересная геометрическая картина возникает в случае, когда матрица A — симметричная.

Теорема Пусть A — вещественная, симметричная, положительно определенная матрица. Тогда метод Зейделя сходится.

Доказательство. При A симметричной имеем

$$\begin{aligned} F(y) &= (A(y - X), y - X) - (AX, X) = \\ &= (Ay, y) - 2(AX, y) = (Ay, y) - 2(b, y). \end{aligned}$$

Если $A > 0$, то $(A(y - X), y - X) > 0$ при $y \neq X$, и поэтому функция $F(y)$ имеет минимум, и притом единственный, при $y = X$. Таким образом, задача отыскания решения системы $Ax = b$ оказалась равносильной задаче отыскания единственного минимума функции $F(y)$.

Одним из методов минимизации функций многих переменных является метод покоординатного спуска.

Пусть имеется приближение $(x_1^0, \dots, x_m^0)$ к точке экстремума функции $F(x_1, \dots, x_m)$. Рассмотрим функцию $F(x_1, x_2^0, \dots, x_m^0)$ как функцию переменной x_1 и найдем точку x_1^1 ее минимума. Затем, исходя из приближения $(x_1^1, x_2^0, \dots, x_m^0)$ путем

минимизации функции $F(x_1^1, x_2, x_3^0, \dots, x_m^0)$, находим следующее приближение $(x_1^1, x_2^1, x_3^0, \dots, x_m^0)$. Процесс циклически повторяется. При уточнении компоненты x_k происходит смещение по прямой, параллельной оси x_k до точки с наименьшим на этой прямой значением $F(x) = c$. Очевидно, эта точка будет точкой касания рассматриваемой прямой и линии уровня $F(x) = c$. Поэтому в двумерном случае картина приближений выглядит как на рис. 684

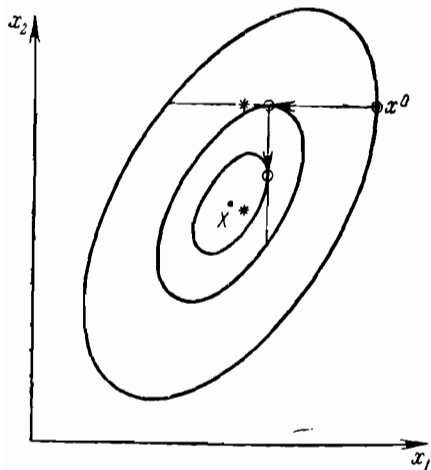


Рис. 684.

Применим метод покоординатного спуска для отыскания экстремума функции F . Обозначим $F(y) + (AX, X) = (A(y - X), y - X)$ через $F_0(y)$. При минимизации по переменной x_k происходит перемещение параллельно оси x_k до точки, где $F'_{x_k} = 0$.

Следовательно, новое значение x_k определяется из уравнения

$$F'_{x_k} = 2 \left(\sum_{j=1}^m a_{kj} x_j - b_k \right) = 0,$$

того же, что и в случае метода Зейделя. Таким образом, приближения покоординатного спуска минимизации функции F и метода Зейделя решения исходной системы совпадают.

Если $x^n \neq X$, то хотя бы одно из уравнений системы не удовлетворяется, и соответствующее значение $F'_{x_k}(x^n) \neq 0$. Выберем среди таких k наименьшее. Тогда при уточнении компонент $x_1, \dots, x_{k-1}$ мы остаемся в точке x^n , а при уточнении компоненты x_k происходит смещение в сторону меньших значений $F(x)$, при уточнении остальных компонент значение $F(x)$ не возрастет. Таким образом,

$$F(x^{n+1}) < F(x^n), \quad F_0(x^{n+1}) < F_0(x^n),$$

и поэтому

$$\frac{F_0(x^{n+1})}{F_0(x^n)} < 1 \quad (6)$$

при $x^n \neq X$. Вследствие (4) имеем равенство $r^{n+1} = -B^{-1}Cr^n$. Соотношение (6) можно переписать в виде

$$\varphi(r^n) = \frac{(AB^{-1}Cr^n, B^{-1}Cr^n)}{(Ar^n, r^n)} < 1 \quad (7)$$

при $r^n \neq 0$. На сфере $\|r^n\| = 1$ величина $\varphi(r^n)$ непрерывна и поэтому достигает своего наибольшего значения φ_0 . Так как $A > 0$, то всегда $\varphi(r^n) > 0$ и поэтому $\varphi_0 > 0$; положим $\sqrt{\varphi_0} = \lambda$. Вследствие (7) имеем $\lambda^2 < 1$. Очевидно, $\varphi(cr^n) = \varphi(r^n)$ при любом $c \neq 0$; поэтому $\varphi(r^n) = \varphi(r^n/\|r^n\|) \leq \lambda^2$ при любом r^n . Отсюда получаем неравенство

$$\frac{F_0(x^{n+1})}{F_0(x^n)} \leq \lambda^2, \quad (8)$$

и таким образом,

$$F_0(x^n) \leq \lambda^{2n} F_0(x^0).$$

Из (3.8), (3.9) следуют неравенства

$$(\min \lambda_A^i) \|y - X\|_3^2 \leq F_0(y) \leq (\max \lambda_A^i) \|y - X\|_3^2;$$

отсюда получаем оценку скорости сходимости

$$\begin{aligned} \|x^n - X\|_3 &\leq \sqrt{\frac{F_0(x^n)}{\min \lambda_A^i}} \leq \lambda^n \sqrt{\frac{F_0(x^0)}{\min \lambda_A^i}} \leq \lambda^n \sqrt{\frac{\max \lambda_A^i}{\min \lambda_A^i}} \|x^0 - X\|_3. \end{aligned} \quad (9)$$

Теорема доказана.

Из рис. 6.8.5 можно усмотреть, что метод Зейделя сходится быстрее, если направление осей эллипсоидов близко к направлению координатных осей, т. е. матрица A близка к диагональной.

В случае, изображенном на рис. 6.8.4, последовательные приближения смещаются все время монотонно влево и вниз. Такая картина — монотонное смещение отдельных компонент все время в одном и том же направлении характерна для ряда классов матриц. При этом довольно часто монотонное смещение на-

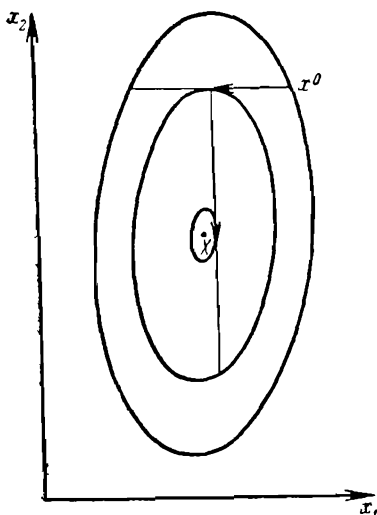


Рис. 6.8.5.

блюдается именно у компонент решения, скорость сходимости которых наиболее плохая. В этих случаях для ускорения сходимости прибегают к *методу релаксации*, который заключается в следующем. После уточнения каждой координаты по методу Зейделя производится дополнительное смещение в том же направлении на p -ю часть этого смещения (p необязательно положительно). Таким образом, приближения метода релаксации отыскиваются из соотношения

$$B \left(\frac{x^{n+1} + p x^n}{1 + p} \right) + C x^n = b. \quad (10)$$

Как показала практика вычислений, при $A > 0$ целесообразно

брать показатель релаксации p в пределах $0 < p < 1$; в случае $0 < p < 1$ метод релаксации обычно называют *методом сверхрелаксации*. На рис. 6.8.4 * — изображения приближений метода Зейделя, * — метода сверхрелаксации при $p = 1/4$. Например, для уменьшения ошибки в ε^{-1} раз при решении системы (7.20) методом Зейделя требуется $\sim ch^{-2} \ln(\varepsilon^{-1})$ итераций. Если применить метод сверхрелаксации при $p = 1 - \omega h$, $\omega > 0$, то потребуется $\sim c(\omega) h^{-1} \ln(\varepsilon^{-1})$ итераций. Подбор параметра ω в этой задаче требует детального рассмотрения. В частности, во многих случаях оптимальное значение параметра релаксации p определяется экспериментально.

Иногда параметр релаксации p выбирают зависящим от n и i .

Для случая $A > 0$ еще раз обратимся к геометрической картине (см. рис. 6.8.4). После уточнения компоненты x_1 по методу сверхрелаксации при $-1 < p < 1$ мы попадаем в точку, лежащую внутри или на границе эллипсоида $F(x) = F(x^0)$. Рассуждая так

же, как при обосновании сходимости метода Зейделя, заключаем, что при $-1 < \rho < 1$ всегда $F_0(x^{n+1}) < F_0(x^n)$, если $x^n \neq X$.

Задача. Доказать, что при условии $|\rho_n| \leq q < 1$ метод релаксации сходится со скоростью геометрической прогрессии.

§ 9. Метод наискорейшего градиентного спуска

Распространенным методом минимизации функций большого числа переменных является *метод градиентного спуска*. Последующее приближение получается из предыдущего смещением в направлении градиента функции $F(x)$. Таким образом, каждое следующее приближение ищется в виде

$$x^{n+1} = x^n - \delta_n \text{grad } F(x^n). \quad (1)$$

Приведенное описание не определяет алгоритм однозначно, поскольку ничего не сказано о выборе параметра δ_n . Например, его можно определять из условия минимума величины

$$F(x^n - \delta_n \text{grad } F(x^n)). \quad (2)$$

В этом случае рассматриваемый метод называют *методом наискорейшего градиентного спуска*, или просто *методом наискорейшего спуска*.

Для функции $F(x) = (Ax, x) - 2(b, x)$, соответствующей системе линейных уравнений с матрицей $A > 0$, задача нахождения минимума решается в явном виде.

В этом конкретном случае

$$\text{grad } F = 2(Ax - b)$$

и

$$x^{n+1} = x^n - \delta_n 2(Ax^n - b).$$

Обозначим $2\delta_n$ через Δ_n , т. е. положим

$$x^{n+1} = x^n - \Delta_n(Ax^n - b). \quad (3)$$

Пусть $\varphi(\Delta_n) = F(x^{n+1})$. Вспоминая, что $A = A^*$, вычислим $\varphi'(\Delta_n)$:

$$\begin{aligned} \varphi'(\Delta_n) &= \left(A \frac{dx^{n+1}}{d\Delta_n}, x^{n+1} \right) + \left(Ax^{n+1}, \frac{dx^{n+1}}{d\Delta_n} \right) - 2 \left(\frac{dx^{n+1}}{d\Delta_n}, b \right) = \\ &= 2 \left(Ax^{n+1} - b, \frac{dx^{n+1}}{d\Delta_n} \right) = -2(Ax^{n+1} - b, Ax^n - b). \end{aligned}$$

С учетом (3) уравнение $\varphi'(\Delta_n) = 0$ имеет вид

$$(Ax^n - b - \Delta_n A(Ax^n - b), Ax^n - b) = 0,$$

откуда

$$\Delta_n = \frac{(Ax^n - b, Ax^n - b)}{(A(Ax^n - b), Ax^n - b)}. \quad (4)$$

На рис. 6.9.1 изображено расположение приближений метода наискорейшего спуска и линий уровня функции F_0 . Итерационный процесс (3), (4) называют *методом наискорейшего спуска* решения рассматриваемой системы линейных уравнений.

Пусть $S_A \subset [\mu, M]$.

Теорема. Приближения метода наискорейшего спуска удовлетворяют соотношению

$$F_0(x^n) \leq \left(\frac{M - \mu}{M + \mu} \right)^{2n} F_0(x^0). \quad (5)$$

Доказательство. При $y^n = x^n$ произведем одну итерацию оптимального одношагового итерационного процесса

$$y^{n+1} = y^n - \frac{2}{M + \mu} (Ay^n - b). \quad (6)$$

Ошибки итераций $r^n = y^n - X$ связаны соотношением

$$r^{n+1} = \left(E - \frac{2}{M + \mu} A \right) r^n.$$

Пусть $e_1, \dots$ — ортонормированная система собственных векторов матрицы A : $Ae_i = \lambda_i e_i$.

Поскольку $\mu \leq \lambda_i \leq M$, то при всех i выполняются соотношения

$$-\frac{M - \mu}{M + \mu} \leq 1 - \frac{2}{M + \mu} \lambda_i \leq \frac{M - \mu}{M + \mu}$$

и, таким образом,

$$\left| 1 - \frac{2}{M + \mu} \lambda_i \right| \leq \frac{M - \mu}{M + \mu}. \quad (7)$$

Пусть $r^n = \sum c_i e_i$. Справедливы соотношения

$$(Ar^n, r^n) = \left(\sum c_i \lambda_i e_i, \sum c_i e_i \right) = \sum \lambda_i c_i^2,$$

$$r^{n+1} = \sum b_i e_i, \quad \text{где} \quad b_i = \left(1 - \frac{2\lambda_i}{M + \mu} \right) c_i,$$

$$(Ar^{n+1}, r^{n+1}) = \sum \lambda_i b_i^2 = \sum \lambda_i \left(1 - \frac{2\lambda_i}{M + \mu} \right)^2 c_i^2.$$

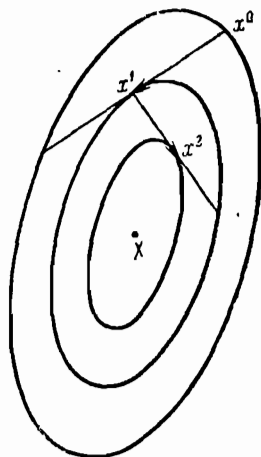


Рис. 6.9.1.

С учетом (7) получаем

$$(Ar^{n+1}, r^{n+1}) \leq \left(\frac{M - \mu}{M + \mu} \right)^2 (Ar^n, r^n).$$

Поскольку $F_0(y^n) = (Ar^n, r^n)$, то это означает, что

$$F_0(y^{n+1}) \leq \left(\frac{M - \mu}{M + \mu} \right)^2 F_0(y^n) = \left(\frac{M - \mu}{M + \mu} \right)^2 F_0(x^n).$$

Приближение y^{n+1} можно записать в виде (1)

$$y^{n+1} = x^n - \alpha \operatorname{grad} F(x^n), \quad \alpha = \frac{1}{M + \mu}.$$

Поскольку x^{n+1} минимизирует функционал $F(x)$ среди всех приближений вида (1), то

$$F(x^{n+1}) \leq F(y^{n+1}).$$

Отсюда следует оценка

$$F_0(x^{n+1}) \leq F_0(y^{n+1}) \leq \left(\frac{M - \mu}{M + \mu} \right)^2 F_0(x^n),$$

а, следовательно, и справедливость утверждения теоремы. Аналогично (8.9) можно получить неравенство

$$\|x^n - X\|_3 \leq \left(\frac{M - \mu}{M + \mu} \right)^n \sqrt{\frac{M}{\mu}} \|x^0 - X\|_3.$$

Хотя на каждом шаге метода наискорейшего спуска уменьшение величины $F_0(x)$ заведомо не меньше, чем у итерационного процесса (6), мы получили примерно одинаковые оценки скорости сходимости. Однако есть принципиальная разница в этих методах и оценках. Для написания итерационного процесса (6) требуется информация о границах спектра μ, M . Для применения метода (3), (4) такой информации не требуется.

Однако у метода наискорейшего спуска есть следующий недостаток по сравнению с простейшим итерационным процессом (6). При нахождении каждого следующего приближения он требует не одной, а двух трудоемких операций умножения матрицы на вектор.

Двукратного умножения матрицы на вектор при каждой итерации можно избежать следующим образом. Обозначим $w^n = Ax^n - b$ и перепишем (3) в виде

$$x^{n+1} = x^n - \Delta_n w^n. \quad (8)$$

Умножая слева на A и вычитая b , получим

$$w^{n+1} = w^n - \Delta_n A w^n. \quad (9)$$

Формулу (4) для определения Δ_n можно записать в виде

$$\Delta_n = \frac{(w^n, w^n)}{(A w^n, w^n)}. \quad (10)$$

В процессе итерации запоминаются векторы x^n, w^n и на каждом шаге последовательно вычисляются $A w^n, \Delta_n, x^{n+1}, w^{n+1}$. В исходном методе

наискорейшего спуска (3), (4) ошибка на шаге итерации равносильна возмущению начального приближения и, поскольку процесс сходящийся, ее влияние должно иметь тенденцию к затуханию. В итерационном процессе (8)—(10) имеется более сложная зависимость между погрешностями последующих приближений, приводящая часто к существенному росту суммарной вычислительной погрешности с ростом n . Поэтому разумно время от времени прерывать ход вычислений по формулам (8)—(10) и производить одну итерацию с помощью формул (3), (4). После такой операции погрешность, являющаяся следствием округлений на предшествующих шагах, по отношению к последующим вычислениям играет роль погрешности начального условия и поэтому имеет тенденцию к убыванию.

Реальный выбор итерационного процесса должен производиться с учетом имеющейся информации о границе спектра, объема и состава памяти ЭВМ. Например, при решении сеточных аппроксимаций дифференциальных уравнений в частных производных иногда идут по следующему пути. Проводят вспомогательную вычислительную работу по возможно более точному определению значений μ и M , а затем применяют оптимальный линейный итерационный процесс.

§ 10. Метод сопряженных градиентов

Этот метод также предназначен для решения систем уравнений

$$Ax = b$$

с положительно определенной матрицей A .

Пусть x^0 — начальное приближение к решению. Образуем вектор

$$x^n = x^0 + \sum_{i=0}^{n-1} \alpha_i A^i r^0, \quad r^0 = Ax^0 - b. \quad (1)$$

Поставим задачу минимизации величины

$$F(x^n) = (Ax^n, x^n) - 2(b, x^n)$$

за счет выбора параметров α_i . Имеем равенство

$$\frac{dF(x^n)}{d\alpha_j} = 2 \left(Ax^n - b, \frac{dx^n}{d\alpha_j} \right).$$

Имеем $\frac{dx^n}{d\alpha_j} = A^j r_0$ и

$$Ax^n - b = r^0 + \sum_{i=0}^{n-1} \alpha_i A^{i+1} r^0, \quad (2)$$

условия $F'_{\alpha_j} = 0$ образуют систему уравнений

$$(Ax^n - b, A^j r^0) = 0, \quad j = 0, \dots, n-1. \quad (3)$$

Подставляя сюда выражение $Ax^n - b$, получим систему линейных уравнений

$$(r^0, A^j r^0) + \sum_{i=0}^{n-1} \alpha_i (A^{i+1} r^0, A^j r^0) = 0, \quad j=0, \dots, n-1, \quad (4)$$

относительно параметров $\alpha_0, \dots, \alpha_{n-1}$.

При заданном начальном приближении x^0 можно выбрать какое-то n , решить систему уравнений (4) относительно неизвестных α_i и получить приближение x^n . Возможны различные пути использования описанного алгоритма.

Можно последовательно осуществлять подобные шаги, беря в качестве начального приближения на каждом шаге приближение, полученное из предыдущего шага.

Другой путь состоит в построении последовательности векторов

$$x^n = x^0 + \sum_{i=0}^{n-1} \alpha_i^* A^i (Ax^0 - b),$$

минимизирующих $F(x)$ при каждом n .

Поставим задачу построить рекуррентное соотношение, связывающее эти векторы аналогично рекуррентным соотношениям оптимального линейного итерационного процесса.

Пусть $e_1, \dots, e_m$ — полная ортонормированная система собственных векторов матрицы A . Представим вектор $r^0 = Ax^0 - b$ в виде $r^0 = \sum_{j=1}^m c_j e_j$. Линейная комбинация собственных векторов, соответствующих одному и тому же собственному значению, также является собственным вектором, соответствующим этому собственному значению. Поэтому, объединяя слагаемые, соответствующие одинаковым собственным значениям, и выбрасывая нулевые слагаемые, можно написать равенство $r^0 = \sum_{j=1}^q g_j$; здесь $g_j \neq 0$,

$$Ag_j = \lambda_j g_j,$$

$\lambda_1, \dots, \lambda_q$ — различные собственные значения матрицы A . Поскольку собственные векторы, соответствующие различным собственным значениям, ортогональны, то $(g_i, g_j) = 0$ при $i \neq j$.

Покажем, что векторы $r^0, \dots, A^{q-1}r^0$ линейно независимы. Пусть это неверно; тогда $\sum_{k=0}^{q-1} \gamma_k A^k r^0 = 0$. Иначе это соотношение можно переписать в виде

$$0 = \sum_{k=0}^{q-1} \gamma_k \left(\sum_{j=1}^q (\lambda_j)^k g_j \right) = \sum_{j=1}^q P(\lambda_j) g_j,$$

где $P(\lambda) = \sum_{k=0}^{q-1} \gamma_k \lambda^k$. Вследствие линейной независимости векторов g_j имеем равенства

$$P(\lambda_j) = 0, \quad j = 1, \dots, q.$$

Приходим к противоречию, получив, что многочлен степени, меньшей q , имеет q различных нулей.

Из доказанного следует, что вектор

$$Ax^n - b = r^0 + \sum_{i=1}^{n-1} \alpha_i^q A^i r^0$$

отличен от нуля при $n < q$.

Покажем, что $x^q = X$. Действительно, определим многочлен

$$Q_q(\lambda) = \prod_{j=1}^q \left(1 - \frac{\lambda}{\lambda_j}\right) = 1 + \sum_{k=1}^q \beta_k \lambda^k.$$

При

$$\bar{x}^q = x^0 + \sum_{k=1}^q \beta_k A^{k-1} (Ax^0 - b),$$

согласно (2) и определению $Q_q(\lambda)$, имеем

$$A\bar{x}^q - b = \sum_{j=1}^q Q_q(\lambda_j) g_j = 0;$$

следовательно, $\bar{x}^q$ реализует минимум функционала $F(x)$, и поэтому $\bar{x}^q = x^q = X$.

Лемма. Величина α_{n-1}^n отлична от 0 при $n \leq q$.

Доказательство. Если $\alpha_{q-1}^q = 0$, то

$$0 = Ax^q - b = r^0 + \sum_{i=0}^{q-2} \alpha_i^q A^{i+1} r^0$$

и приходим к противоречию с утверждением о линейной независимости векторов $r^0, \dots, A^{q-1} r^0$. Пусть $\alpha_{j-1}^j = 0$ и $j < q$; тогда вектор $Ax^j - b$, согласно (2), запишется в виде линейной комбинации векторов $r^0, \dots, A^{j-1} r^0$; в то же время, вследствие (3), он ортогонален ко всем этим векторам. Отсюда следует, что $Ax^j - b = 0$, и, следовательно, $j \geq q$. Утверждение леммы доказано.

Имеем совокупность равенств

$$x^1 = x^0 + \alpha_0^1 r^0,$$

$$x^2 = x^0 + \alpha_0^2 r^0 + \alpha_1^2 A r^0,$$

$$\dots \dots \dots$$

$$x^q = x^0 + \dots + \alpha_{q-1}^q A^{q-1} r^0.$$

Вследствие условия $\alpha'_{j-1} \neq 0$ эта система с треугольной матрицей может быть разрешена относительно неизвестных $A^j r^0$, $j=0, \dots, q-1$; в результате этого получатся соотношения

$$A^{j-1} r^0 = \sum_{i=1}^j b'_i (x^i - x^0), \quad j=1, \dots, q.$$

Подставляя выражения $A^{j-1} r^0$ в правую часть (2), получим

$$Ax^n - b = r^0 + \sum_{i=0}^{n-1} \alpha_i^n A^{i+1} r^0 = \sum_{i=1}^{n+1} \gamma_i^n (x^i - x^0) \quad \text{при } n < q.$$

Применив к обеим частям оператор A , имеем

$$A(Ax^n - b) = \sum_{i=1}^{n+1} \gamma_i^n ((Ax^i - b) - (Ax^0 - b)) = \sum_{i=0}^{n+1} \sigma_i^n (Ax^i - b). \quad (5)$$

При этом, очевидно, $\sum_{i=0}^{n+1} \sigma_i^n = 0$. Соотношения (3) можно переписать в виде

$$(A(Ax^n - b), A^j r^0) = 0, \quad j=0, \dots, n-2;$$

отсюда и из (2) следуют равенства

$$(A(Ax^n - b), Ax^k - b) = 0 \quad \text{при } k \leq n-2.$$

Точно так же получаем

$$(Ax^n - b, Ax^k - b) = 0 \quad \text{при } k \leq n-1.$$

Умножая (5) скалярно на векторы $(Ax^k - b)$, $k=0, \dots, n-2$, получим равенства

$$0 = \sigma_k^n (Ax^k - b, Ax^k - b),$$

откуда следует, что все $\sigma_k^n = 0$ при $k < n-1$. Таким образом, (5) перепишется в виде

$$A(Ax^n - b) = \sum_{i=n-1}^{n+1} \sigma_i^n (Ax^i - b);$$

применяя оператор A^{-1} , получим равенство

$$Ax^n - b = \sum_{i=n-1}^{n+1} \sigma_i^n (x^i - X).$$

Поскольку $\sum \sigma_i^n = 0$, то его можно записать в виде

$$Ax^n - b = \sigma_{n-1}^n x^{n-1} + \sigma_n^n x^n + \sigma_{n+1}^n x^{n+1}. \quad (6)$$

Отметим, что проведенные построения имеют много общего с построениями из § 15 гл. II.

Для рассматриваемых значений $n < q$ коэффициент σ_{n+1}^n всегда отличен от нуля. В самом деле, предположим противное. Тогда, подставляя в равенство

$$Ax^n - b - \sum_{i=n-1}^{n+1} \sigma_i^n (x^i - x^0) = 0$$

выражения

$$Ax^n - b = r^0 + \sum_{i=0}^{n-1} \alpha_i^n A^{i+1} r^0,$$

$$x^k - x^0 = \sum_{i=0}^{k-1} \alpha_i^k A^i r^0, \quad k = n-1, n,$$

имеем соотношение вида

$$\alpha_{n-1}^n A^n r^0 + \dots = 0,$$

противоречащее линейной независимости векторов $r^0, \dots, A^n r^0$. Умножая (5) скалярно на $Ax^k - b$, $k = n-1, n$, получаем выражения для коэффициентов σ_{n-1}^n и σ_n^n :

$$\sigma_k^n = \frac{(A(Ax^n - b), Ax^k - b)}{(Ax^k - b, Ax^k - b)},$$

поскольку $Ax^k - b \neq 0$ и $A > 0$, то $\sigma_{n-1}^n, \sigma_n^n \neq \infty$. Воспользовавшись тем, что $\sigma_{n+1}^n = -\sigma_n^n - \sigma_{n-1}^n$, мы можем переписать (6) в виде

$$Ax^n - b = \sigma_n^n (x^n - x^{n+1}) + \sigma_{n-1}^n (x^{n-1} - x^{n+1}). \quad (7)$$

Определяя при каждом n значения коэффициентов σ_n^n и σ_{n-1}^n , можно последовательно находить из этого соотношения векторы $x^2, \dots$.

Вектор x^1 определяется непосредственно с помощью решения системы (4) и, оказывается, совпадает с первым приближением, в методе наискорейшего спуска. Для нахождения x^{n+1} из (7) необходимы две операции умножения матрицы на вектор: Ax^n и $A(Ax^n)$, требующие каждая $\sim 2m^2$ арифметических операций, и вычисление нескольких скалярных произведений. Таким образом, этот итерационный процесс примерно вдвое более трудоемок по сравнению с оптимальным линейным итерационным процессом. Трудоемкость метода можно уменьшить следующим образом. Из (7) следует равенство

$$A(Ax^n - b) = \sigma_n^n ((Ax^n - b) - (Ax^{n+1} - b)) + \\ + \sigma_{n-1}^n ((Ax^{n-1} - b) - (Ax^{n+1} - b)). \quad (8)$$

При одновременном использовании формул (7), (8) операция умножения матрицы на вектор на каждом шаге производится однажды — при вычисле-

нии произведения $A(Ax^n - b)$; векторы $Ax^{n+1} - b$ определяются из рекуррентного соотношения (8).

Однако как при таком алгоритме, так и при непосредственном использовании формулы (7) происходит существенное накопление вычислительной погрешности. Чтобы избежать этого, применяется следующая модификация алгоритма. Расчетную формулу (7) можно записать в виде

$$x^{n+1} = x^n + \alpha_n (x^n - x^{n-1}) + \beta_n (Ax^n - b), \quad (9)$$

где

$$\beta_n = \frac{-1}{\sigma_n^2 + \sigma_{n-1}^2}, \quad \alpha_n = \beta_n \sigma_{n-1}^2.$$

Вместо того чтобы вычислять α_n и β_n по этим формулам, будем каждый раз определять их, минимизируя по α_n и β_n функционал $F(x^{n+1})$. Для вычисления α_n и β_n получаются расчетные формулы

$$\begin{aligned} \alpha_n &= ((Ax^n - b, Ax^n - b)(Ax^n - b, A(x^n - x^{n-1})) - (Ax^n - b, \\ &\quad x^n - x^{n-1})(A(Ax^n - b), Ax^n - b)) \sigma_n, \\ \beta_n &= ((Ax^n - b, x^n - x^{n-1})(A(x^n - x^{n-1}), Ax^n - b) - (Ax^n - b, \\ &\quad Ax^n - b)(x^n - x^{n-1}, A(x^n - x^{n-1}))) \sigma_n, \\ \sigma_n &= \frac{1}{(x^n - x^{n-1}, A(x^n - x^{n-1}))(Ax^n - b, A(Ax^n - b)) - (Ax^n - b, A(x^n - x^{n-1}))}. \end{aligned} \quad (10)$$

Каждый шаг итераций требует уже трех операций умножения матрицы на вектор. Очевидно, при отсутствии округлений эти итерационные процессы выдают одну и ту же последовательность векторов x^n . Однако последний итерационный процесс является самоисправляющимся — влияние ошибок округления, допущенных на предшествующих шагах, постепенно затухает. Так же, как в случае метода наискорейшего спуска, может оказаться наиболее разумным чередование итераций по формулам (7)–(8) и (9)–(10). Заметим, что при отсутствии округлений $(Ax^n - b, x^n - x^{n-1}) = 0$; однако если в (10) эту величину заменить нулем, то процесс может утратить самоисправляемость.

Построенный итерационный процесс имеет много различных названий, в частности *метод Ланцоша* или *сопряженных градиентов*.

Произведем оценку скорости сходимости. Для приближений оптимального линейного итерационного процесса y^n имеем:

если

$$x^0 - X = y^0 - X = \sum_i c_i e_i,$$

то

$$y^n - X = \sum_i c_i g_i e_i,$$

где

$$|g_i| \leq |t_n|^{-1} = \left| T_n \left(-\frac{M+\mu}{M-\mu} \right) \right|^{-1}.$$

Отсюда следует оценка*

$$\begin{aligned} F_0(y^n) &= (A(y^n - X), y^n - X) = \\ &= \sum_i \lambda_i c_i^2 g_i^2 \leq |t_n|^{-2} \sum_i \lambda_i c_i^2 = |t_n|^{-2} (A(x^0 - X), x^0 - X). \end{aligned}$$

Приближение метода сопряженных градиентов минимизирует функционалы F и F_0 среди всех приближений вида (1), которому принадлежит также y^n . Поэтому

$$F_0(x^n) \leq F_0(y^n) \leq |t_n|^{-2} F_0(x^0).$$

Как и ранее, получаем оценку скорости сходимости в обычной норме

$$\|x^n - X\|_3 \leq \sqrt{\frac{M}{\mu}} |t_n|^{-1} \|x^0 - X\|_3.$$

Рассматриваемый процесс имеет ту же по порядку оценку скорости сходимости, что и оптимальный линейный итерационный процесс. Его существенным преимуществом оказывается то, что не требуется знания границ спектра.

Поскольку после $q \leq m$ итераций в предположении отсутствия округлений получается точное решение задачи, этот метод относится также к классу точных методов. Пусть матрица A имеет l ненулевых элементов, естественно $l \geq m$. Для проведения одного шага итерации требуется $O(l)$ арифметических операций, и поэтому для получения точного решения потребуется $qO(l) = O(ml)$ арифметических операций. Таким образом, в случае матриц с малым числом ненулевых элементов этот метод оказывается также и эффективным точным методом. При этом он предъявляет относительно малые требования к памяти ЭВМ. Если размерность системы m велика, часто хорошее приближение к решению получается при числе итераций $n \ll q$.

Рассмотрим случай, когда не выполнено условие $A > 0$. Система $A^*Ax = A^*b$ эквивалентна исходной, но может оказаться, что число ненулевых элементов матрицы A^*A много больше l . Тогда будет разумнее не выписывать эту систему в явном виде, а каждый раз вычислять векторы вида A^*Au последовательным умножением u на матрицы A и A^* .

§ 11. Метод Монте-Карло решения систем линейных уравнений

Рассмотренные выше итеративные методы решения систем линейных уравнений при своей реализации требуют хранения в памяти машины всех компонент решения, что при очень больших m является обременительным. В случае, когда порядок системы очень большой, а коэффициенты системы вычисляются по простым формулам, может быть применен метод Монте-Карло.

Рассмотрим общую схему метода. Пусть требуется вычислить сумму $S = \sum_{k=1}^N b_k$. Представим каждое b_k в виде $b_k = p_k f_k$, где $p_k \geq 0$, $\sum_{k=1}^N p_k = 1$. Пусть ξ — случайная величина, принимающая значение k с вероятностью p_k .

Выберем M независимых значений $k_1, \dots, k_M$ этой величины и положим

$$S \approx S_M = \frac{1}{M} \sum_{l=1}^M f_{k_l}.$$

Имеем равенства

$$Mf_k = \sum_{k=1}^N p_k f_k = \sum_{k=1}^N b_k = S, \quad (1)$$

$$d^2 = Df_k = Mf_k^2 - (Mf_k)^2 = \sum_{k=1}^N p_k f_k^2 - S^2 = \sum_{k=1}^N \frac{b_k^2}{p_k} - S^2.$$

Вследствие независимости значений f_k справедливо равенство $D(S_M) = \frac{d^2}{M}$.

Применяя неравенство Чебышева, получаем оценку: при $0 < \eta < 1$ с вероятностью $1 - \eta$ выполняется соотношение $|S_M - S| \leq d/\sqrt{\eta M}$. Таким образом, при больших M величина S_M с большой вероятностью близка к S .

Для получения значений случайной величины ξ можно поступать следующим образом. Обычно на машине имеется датчик случайных чисел или программа псевдослучайных чисел, выдающие значения $\eta_1, \dots$ некоторой случайной величины, равномерно распределенной на $[0, 1]$.

Для каждого значения η_l определяют число k_l такое, что η_l принадлежит полусегменту

$$p_{k_l} = \left[\sum_{i=1}^{k_l-1} p_i, \sum_{i=1}^{k_l} p_i \right)$$

и полагают $\xi_l = k_l$. Вероятность попадания η_l на полусегмент p_k равна p_k ; следовательно, полученная таким образом величина ξ принимает каждое значение s с вероятностью p_s . Проще всего было бы взять $p_k \equiv \frac{1}{N}$, тогда

$\sum_{i=1}^l p_i = \frac{l}{N}$, и поэтому величина ξ_l определялась бы по простой формуле.

$$\xi_l = [\eta_l N] + 1.$$

Однако, как видно из (1), величина d , характеризующая точность результата, существенно зависит от выбора значений p_k . Очень часто такой простой алгоритм не дает удовлетворительного результата и при решении конкретных задач методом Монте-Карло приходится затрачивать большие усилия на оптимальный выбор значений p_k применительно к данной совокупности чисел b_k .

Рассмотрим практически важный случай, когда все $b_k \geq 0$. Будем минимизировать величину d^2 методом Лагранжа при условии $\sum_{k=1}^N p_k = 1$. Функция Лагранжа имеет вид

$$\Phi(p_1, \dots, p_N) = \sum_{k=1}^N \frac{b_k^2}{p_k} - S^2 + \lambda \left(\sum_{k=1}^N p_k - 1 \right).$$

Из условия $\Phi'_{p_k} = 0$, $k = 1, \dots, N$, имеем $-\frac{b_k^2}{p_k^2} + \lambda = 0$. Отсюда $p_k = \frac{b_k}{\sqrt{\lambda}}$.

Из условия $\sum_{k=1}^N p_k = 1$ имеем $\sum_{k=1}^N b_k = \sqrt{\lambda}$, т. е. $\sqrt{\lambda} = S$. Таким образом,

разумнее всего взять $p_k = \frac{b_k}{S}$. При таком выборе p_k

$$f_k = \frac{h_k}{r_k} \equiv S, \quad d^2 = 0.$$

Для реализации оптимального распределения p_k нужно знать искомую величину S . Иногда поступают следующим образом: подбирают функцию $B_k \geq 0$

такую, что $b_k \approx B_k$, и суммы $\sigma_n = \sum_{k=1}^n B_k$ известны при всех n ; затем по-

лагают $p_k = \frac{B_k}{\sigma_N}$.

Перейдем непосредственно к применению метода Монте-Карло для решения систем линейных уравнений. Пусть нам нужно найти компоненту $(x)_q$ решения системы $x = Bx + c$. Пусть матрица B такова, что итерационный процесс $x^{n+1} = Bx^n + c$ сходится к решению системы X при любом начальном приближении x^0 . При $x^0 = 0$ имеем

$$x^1 = c, \quad x^2 = c + Bc, \dots, x^{n+1} = \sum_{j=0}^n B^j c.$$

Вследствие сходимости процесса при достаточно большом n величина $(x^{n+1})_q$ достаточно хорошо приближает $(X)_q$. Из соображений близости x^{n+1} к X выберем значение n и займемся отысканием $(x^{n+1})_q$. Имеем

$$(x^{n+1})_q = c_q + \sum_{j_1=1}^m b_{qj_1} c_{j_1} + \dots + \sum_{j_1=1}^m \dots \sum_{j_m=1}^m b_{qj_1} \dots b_{j_{n-1}j_n} c_{j_n}. \quad (2)$$

Таким образом, формально задача отыскания значения $(x^{n+1})_q$ равносильна только что рассмотренной задаче отыскания величины S . Однако при очень больших N непосредственное применение рассмотренного ранее способа бессмысленно из-за конечности числа разрядов чисел, используемых в машине. Числа в машине округляются до некоторых величин вида $k2^{-t}$, где k — целое, t — число разрядов чисел в машине. Поэтому длины полусегментов π_i , получаемых при реальных вычислениях, будут кратны 2^{-t} и при значениях p_i , сравнимых с 2^{-t} , искажение в значениях случайной функции ξ будет очень большим. В частности, число ненулевых полусегментов π_i из-за конечности разрядной сетки всегда не превосходит 2^t . В формуле (2) число слагаемых $N > m^n$, и поэтому этот метод непосредственно неприменим уже при не очень больших значениях m и n .

Чтобы избежать этого, поступим следующим образом. Пусть $p_{ij} \geq 0$ и удовлетворяют условиям

$$\sum_{j=1}^m p_{ij} = 1 \quad \text{при } i = 1, \dots, m, \quad p_{ij} > 0, \quad \text{если } b_{ij} \neq 0. \quad (3)$$

Пусть ξ_i , $i = 1, \dots, m$, — независимые случайные величины, причем случайная величина ξ_i принимает с вероятностью p_{ij} значение j . Разыграем значение случайной величины ξ_q , пусть оно будет равно j_i ; затем разыграем зна-

чение случайной величины $\xi_{j_1}^{j_1}$, пусть оно будет равно j_2 и т. д.; разыграем значение случайной величины $\xi_{j_{n-1}}^{j_{n-1}}$, пусть оно будет равно j_n . Положим

$$\bar{S}_{j_1, \dots, j_n} = c_q + f_{qj_1} c_{j_1} + \dots + f_{qj_1} \dots f_{j_{n-1}j_n} c_{j_n},$$

где $f_{ij} = b_{ij}/p_{ij}$, причем $f_{ij} = 0$, если $b_{ij} = 0$, $p_{ij} = 0$. Осуществляя такой розыгрыш M раз, получим M последовательностей $j_1^1, \dots, j_n^1$ и M значений $\bar{S}_{j_1^1, \dots, j_n^1}$. Положим

$$(x^{n+1})_q \approx \bar{S}^{(M)} = \frac{1}{M} \sum_{l=1}^M \bar{S}_{j_1^l, \dots, j_n^l}.$$

Вычислим $M\bar{S}_{j_1, \dots, j_n}$. Вероятность того, что получится последовательность $j_1, \dots, j_n$, равна $p_{qj_1} \dots p_{j_{n-1}j_n}$. Отсюда

$$\begin{aligned} M\bar{S}_{j_1, \dots, j_n} &= \\ &= \sum_{j_1=1}^m \dots \sum_{j_n=1}^m (c_q + f_{qj_1} c_{j_1} + \dots + f_{qj_1} \dots f_{j_{n-1}j_n} c_{j_n}) p_{qj_1} \dots p_{j_{n-1}j_n}. \end{aligned}$$

Правую часть этого равенства представим в виде $\sum_{k=0}^n T_k$, где

$$T_0 = \sum_{j_1=1}^m \dots \sum_{j_n=1}^m c_q p_{qj_1} \dots p_{j_{n-1}j_n},$$

$$T_k = \sum_{j_1=1}^m \dots \sum_{j_n=1}^m f_{qj_1} \dots f_{j_{k-1}j_k} c_{j_k} p_{qj_1} \dots p_{j_{n-1}j_n} \quad \text{при } k > 0.$$

Пусть

$$\sigma_j^{k,n} = \sum_{j_{k+1}=1}^m \dots \sum_{j_n=1}^m p_{jj_{k+1}} \dots p_{j_{n-1}j_n},$$

тогда

$$T_0 = c_q \sigma_q^{0,n},$$

$$T_k = \sum_{j_1=1}^m \dots \sum_{j_k=1}^m (f_{qj_1} p_{qj_1}) \dots (f_{j_{k-1}j_k} p_{j_{k-1}j_k}) c_{j_k} \sigma_{j_k}^{k,n}.$$

Осуществляя суммирование последовательно по переменным $j_n, \dots, j_{k+1}$ при помощи равенства (3), получим $\sigma_j^{k,n} = 1$. Отсюда и из равенств $f_{ij} p_{ij} = b_{ij}$ следует, что

$$T_0 = c_q,$$

$$T_k = \sum_{j_1=1}^m \dots \sum_{j_k=1}^m b_{qj_1} \dots b_{j_{k-1}j_k} c_{j_k}$$

при $k > 0$. Таким образом, $T_k = (B^k c)_q$; с учетом равенств

$$M\bar{S}_{j_1, \dots, j_n} = \sum_{k=0}^n T_k$$

и (2) получаем

$$M\bar{S}_{j_1, \dots, j_n} = (x^{n+1})_q.$$

Отсюда

$$M\bar{S}^{(M)} = (x^{n+1})_q.$$

Так как $D \bar{S}^{(M)} = \frac{D \bar{S}_{j_1, \dots, j_n}}{M}$, то при больших M величина $\bar{S}^{(M)}$ с большой вероятностью близка к $(x^{n+1})_q$, и поэтому можем считать $(X)_q \approx \bar{S}^{(M)}$. Значения $\bar{S}_{j_1, \dots, j_n}$ можно вычислять в процессе розыгрыша чисел $j_1, \dots, j_n$ пользуясь рекуррентными формулами

$$\begin{cases} \bar{S}_{j_1, \dots, j_l} = \bar{S}_{j_1, \dots, j_{l-1}} + z_{j_1, \dots, j_l} c_{j_l} \\ z_{j_1, \dots, j_l} = z_{j_1, \dots, j_{l-1}} f_{j_{l-1}, j_l} \text{ при } l > 1, \end{cases}$$

причем

$$\bar{S}_{j_l} = c_{j_l} + z_{j_l} c_{j_l}, \quad z_{j_l} = f_{q, j_l}.$$

Если при реализации алгоритма пользоваться этими формулами, то вычисление значения $S^{(M)}$ потребует хранения в памяти машины лишь конечного (не растущего при $M, n \rightarrow \infty$) множества чисел.

Так же как при вычислении интегралов высокой кратности, применение метода Монте-Карло для решения систем линейных уравнений требует большой осторожности. В частности, особое внимание следует уделить вопросу удачного выбора совокупности величин p_{ij} .

Проиллюстрируем это высказывание на примере простейшей модельной задачи. Пусть на отрезке $0 \leq t \leq 1$ решается уравнение

$$x''(t) = 0, \quad x(0) = c_0, \quad x(1) = c_1. \quad (4)$$

Зададимся шагом $h = 1/L$, $L = 2l$, l — целое, и напишем сеточную аппроксимацию

$$x_0 = c_0, \quad x_{k+1} - 2x_k + x_{k-1} = 0, \quad k = 1, \dots, L-1, \quad x_L = c_1 \quad (5)$$

(в данном случае исходная задача (4) аппроксимируется без ошибки). Перепишем (5) в виде

$$\begin{aligned} x_0 &= c_0, \\ x_k &= \frac{x_{k-1} + x_{k+1}}{2}, \quad k = 1, \dots, L-1, \\ x_L &= c_1. \end{aligned} \quad (6)$$

Пусть $x = Bx + c$ — матричная запись соотношений (6). Будем искать компоненту x_l . Матрица B симметричная и

$$\lambda_B^l = \cos \frac{\pi l}{L}, \quad l = 1, \dots, L-1.$$

Поэтому

$$\max |\lambda_B^i| = \cos \frac{\pi}{L}.$$

$$\|B^n\|_3 = \left(\cos \frac{\pi}{L}\right)^n = \exp\left(n \ln \cos \frac{\pi}{L}\right) = \exp\left(-n \left(\frac{\pi^2}{2L^2} + O\left(\frac{\pi^4}{L^4}\right)\right)\right). \quad (7)$$

Для малости погрешности приближенного значения $(x^{n+1})_i$ существенно условие $\|B^n\|_3 \ll 1$, поэтому следует взять

$$n \gg L^2. \quad (8)$$

Пусть, например, все $p_{ij} = \frac{1}{L+1}$ и для определенности n — четное. Тогда

$$f_{ij} = \begin{cases} \frac{L+1}{2} & \text{при } |i-j|=1, \text{ если } (i, j) \neq (0, 1), (L, L-1), \\ 0 & \text{при остальных } (i, j). \end{cases}$$

Если $n > 1$, то величина $\bar{S}_{j_1, \dots, j_n}$ может содержать слагаемое

$$f_{1j_1} f_{j_1 j_2} \dots f_{j_{n-1} j_n} c_0$$

где $|j_{p+1} - j_p| = 1$ при $p = 1, \dots, n-1$. Величина этого слагаемого $\left(\frac{L+1}{2}\right)^n c_0$, а вероятность появления $(L+1)^{-n}$. Если ξ — случайная величина, принимающая значения y_i с вероятностями q_i , то

$$D\xi = \sum_i q_i (y_i - M\xi)^2 \geq q_i (y_i - M\xi)^2$$

при любом i . Поэтому

$$D\bar{S}_{j_1, \dots, j_n} \geq (L+1)^{-n} \left(\left(\frac{L+1}{2} \right)^n c_0 - M\bar{S}_{j_1, \dots, j_n} \right)^2.$$

Мы выбрали такое n , что $M\bar{S}_{j_1, \dots, j_n} = (x^{n+1})_i \approx x_i$. Так как $x_i = (c_0 + c_1)/2$, то главный член правой части последнего неравенства имеет порядок $((L+1)/4)^n$. Сопоставляя это соотношение с (8), видим, что уже при небольших L мало надежды на получение значения x_i с приемлемой точностью при разумном числе испытаний M .

Пусть теперь

$$p_{ij} = \begin{cases} \frac{1}{2} & \text{при } i \neq 0, L, |i-j|=1, \\ 1 & i=j=0, L, \\ 0 & \text{во всех остальных случаях.} \end{cases} \quad (9)$$

Рассмотрим случайное блуждание частицы по точкам $j\hbar$: $j = 0, \dots, L$. Пусть блуждание начинается в момент времени $\tau = 0$ в точке $l\hbar$, и пусть из каждой точки $j\hbar$ при $j \neq 0, L$ за единицу времени частица переходит в одну из точек $(j \pm 1)\hbar$ в каждую с вероятностями $1/2$; если частица попадает в одну из точек $t = 0, t = L$, то она там и остается. Пусть $lh, j_1\hbar, \dots, j_n\hbar, \dots$ — последовательные положения частицы; положим

$$\tilde{S}_{j_1, \dots, j_n} = \begin{cases} c_0 & \text{при } j_n = 0, \\ 0 & \text{при } j_n \neq 0, L, \\ c_1 & \text{при } j_n = L. \end{cases}$$

Нетрудно проверить, что вероятность такой траектории частицы совпадает с вероятностью появления последовательности $j_1, \dots, j_n$ при розыгрыше метода Монте-Карло для решения исходной системы с таблицей вероятностей (9), причем $\bar{S}_{j_1, \dots, j_n} = \bar{S}_{j_1, \dots, j_n}$. Мы видим, что в рассматриваемом случае

$$|\bar{S}_{j_1, \dots, j_n}| \leq \bar{c} = \max(|c_0|, |c_1|)$$

и, следовательно, $D_{j_1, \dots, j_n} \bar{S}_{j_1, \dots, j_n} \leq \bar{c}^2$ при всех L и n .

Здесь уже есть возможность получить решение с умеренной точностью при приемлемой затрате машинного времени. При анализе трудоемкости описанного выше метода следует иметь в виду, что в соответствии с (8) время, затрачиваемое на розыгрыш одной совокупности $j_1, \dots, j_n$, растет с L . Поэтому, например, в случае аппроксимаций краевых задач для обыкновенных дифференциальных уравнений метод Монте-Карло не выдерживает конкуренции с точными методами.

Еще одно замечание общего характера. Пусть q_n — вероятность события $f_{qj_1} \dots f_{j_{n-1}} j_n = 0$, тогда величина $q_n - q_{n-1}$ равна вероятности события $f_{qj_1} \dots f_{j_{n-2}} j_{n-1} \neq 0$ и $f_{j_{n-1}} j_n = 0$. Если $q_n \rightarrow 1$ при $n \rightarrow \infty$, то величина

$$Q = \sum_{n=1}^{\infty} (q_n - q_{n-1}) n$$

равна математическому ожиданию значения n , при котором произведение $f_{qj_1} \dots f_{j_{n-1}} j_n$ впервые окажется равным нулю.

В приложениях часто удается выбрать величины p_{ij} так, что дисперсии $DS_{j_1, \dots, j_n}$ равномерно ограничены по n , а число Q не очень велико; тогда поступают следующим образом: число n заранее не фиксируется, а розыгрыш совокупности $j_1, \dots, j_n$ продолжается до тех пор, пока не окажется

$$f_{j_{n-1}} j_n = 0.$$

З а д а ч а. Доказать, что для такого алгоритма

$$MS_{j_1, \dots, j_n} = (x)_q.$$

При решении задачи методом Монте-Карло следует также уделить существенное внимание записи системы в виде $x = Bx + c$, для которого итерационный процесс $x^{n+1} = Bx^n + c$ сходится по возможности быстрее (это приводит к уменьшению числа n , т. е. трудоемкости одного испытания). Кроме удачного выбора величин p_{ij} для целей уменьшения $D_{j_1, \dots, j_n}$ может послужить знание хорошего приближения x^0 к решению. При наличии такого приближения следует перейти к решению методом Монте-Карло системы

$$r = Br + (c + Bx^0 - x^0)$$

относительно $r = x - x^0$, в этом случае можно ожидать уменьшения величины дисперсии $DS_{j_1, \dots, j_n}$, а следовательно, и требуемого объема вычислений.

§ 12. Итерационные методы с использованием спектрально эквивалентных операторов

Кроме методов простой итерации вида

$$x^{n+1} = x^n - \alpha (Ax^n - b) \quad (1)$$

часто применяются итерационные методы

$$Bx^{n+1} = Bx^n - \alpha (Ax^n - b), \quad (2)$$

где матрица $B \neq E$ легко обратима. Если (2) умножить на B^{-1} , то получится

$$x^{n+1} = x^n - \alpha B^{-1} (Ax^n - b).$$

Таким образом, итерационный процесс (2) равносильен методу простой итерации с матрицей $E - \alpha B^{-1}A$.

Рассмотрим случай, когда $A > 0$ и отношение максимального M и минимального μ из собственных значений матрицы A велико. Тогда рассматривавшиеся ранее итерационные процессы сходятся медленно. Пусть $B > 0$ и

$$M_1 = \sup_x \frac{(Ax, x)}{(Bx, x)}, \quad \mu_1 = \inf_x \frac{(Ax, x)}{(Bx, x)}. \quad (3)$$

Предположим, что матрица B легко обратима и $M_1/\mu_1 \ll \ll M/\mu$.

В случае, когда отношение M_1/μ_1 не очень велико, итерационные методы типа рассматриваемых в настоящем параграфе принято называть *итерационными методами с использованием спектрально эквивалентных операторов*. Покажем, что при удачном подборе α метод итераций (2) обладает лучшей сходимостью по сравнению с простейшим методом (1).

Точное решение X удовлетворяет равенству

$$BX = BX - \alpha (AX - b),$$

вычитая его из (2), получим уравнение относительно ошибки

$$Br^{n+1} = Br^n - \alpha Ar^n. \quad (4)$$

Приведем матрицу B при помощи ортогонального преобразования к диагональному виду. Пусть $B = U^* \Lambda U$, где

$$\Lambda = \begin{bmatrix} \lambda_1 & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & \lambda_m \end{bmatrix},$$

U — ортогональная матрица.

Через $\sqrt{B}$ принято обозначать матрицу

$$U^* \begin{bmatrix} +\sqrt{\lambda_1} & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & +\sqrt{\lambda_m} \end{bmatrix} U.$$

Очевидно, $\sqrt{B} > 0$, $(\sqrt{B})^* = \sqrt{B}$, $\sqrt{B} \sqrt{B} = B$. Умножим обе части (4) слева на $(\sqrt{B})^{-1}$ и введем новые величины $\sqrt{B} r^n = v^n$. Получим равенство

$$v^{n+1} = v^n - \alpha C v^n, \quad (5)$$

где $C = (\sqrt{B})^{-1} A (\sqrt{B})^{-1}$.

Вследствие соотношений $A = A^*$, $(\sqrt{B}) = (\sqrt{B})^*$ матрица C симметрична. Рассмотрим выражение

$$w(x) = \frac{(Cx, x)}{(x, x)}.$$

Положив $(\sqrt{B})^{-1} x = y$, можем написать равенство

$$w(x) = w(\sqrt{B} y) = \frac{(Ay, y)}{(By, y)}.$$

Вследствие (3) имеем $w(x) \in [\mu_1, M_1]$; поэтому все собственные значения матрицы C также принадлежат отрезку $[\mu_1, M_1]$.

При $\alpha = \frac{2}{M_1 + \mu_1}$ собственные значения матрицы $E - \alpha C$ по модулю не превосходят $\frac{M_1 - \mu_1}{M_1 + \mu_1}$. Поэтому $\|E - \alpha C\|_3 \leq \frac{M_1 - \mu_1}{M_1 + \mu_1}$, и аналогично (7.7) имеем

$$\|v^n\|_3 \leq \left(\frac{M_1 - \mu_1}{M_1 + \mu_1} \right)^n \|v^0\|_3. \quad (6)$$

Напомним, что

$$\|v^n\|_3 = \sqrt{(\sqrt{B} r^n, \sqrt{B} r^n)} = \sqrt{(B r^n, r^n)}.$$

Если $0 < \mu_2 \leq \frac{(By, y)}{(y, y)} \leq M_2$ при всех y , то из (6) следует оценка

$$\|r^n\|_3 \leq \frac{\|v^n\|_3}{\sqrt{\mu_2}} \leq \left(\frac{M_1 - \mu_1}{M_1 + \mu_1} \right)^n \frac{\|v^0\|_3}{\sqrt{\mu_2}} \leq \left(\frac{M_1 - \mu_1}{M_1 + \mu_1} \right)^n \sqrt{\frac{M_2}{\mu_2}} \|r^0\|_3.$$

Так как функция

$$\frac{M - \mu}{M + \mu} = \frac{1 - (\mu/M)}{1 + (\mu/M)}$$

монотонно убывает с уменьшением M/μ , то скорость сходимости нового итерационного процесса лучше.

При окончательном решении вопроса о переходе от итераций по формулам (1) к итерациям по формулам (2) следует учесть относительную трудоемкость этих итераций; если итерации по формулам (2) существенно более трудоемки, такой переход может оказаться нецелесообразным.

Задача. Пусть $T(A)$ — трудоемкость одной итерации по формуле (1), $T(B)$ — трудоемкость одной итерации по формуле (2). Привести соображения в пользу того, что переход к итерациям по формуле (2) целесообразен, если

$$\frac{T(B)}{\ln \frac{M_1 + \mu_1}{M_1 - \mu_1}} < \frac{T(A)}{\ln \frac{M + \mu}{M - \mu}}.$$

По аналогии с тем, как по итерационному методу (1) был построен метод (2), можно построить соответствующий итерационный метод для любого из рассматривавшихся в §§ 7, 9, 10.

Пусть имеется итерационный процесс, где ошибки последующих приближений связаны равенством

$$r^{n+1} = \left(E - \sum_{i=1}^k \alpha_i A^i \right) r^n.$$

Напишем равенство

$$v^{n+1} = \left(E - \sum_{i=1}^k \alpha_i C^i \right) v^n, \quad (7)$$

где $C = (\sqrt{B})^{-1} A (\sqrt{B})^{-1}$, положим $v^n = \sqrt{B} r^n$ и умножим (7) слева на $\sqrt{B}$. Получим уравнение

$$B r^{n+1} = B r^n - \sum_{i=1}^k \alpha_i (AB^{-1})^{i-1} A r^n.$$

Такое уравнение для ошибки соответствует итерационному процессу

$$B x^{n+1} = B x^n - \sum_{i=1}^k \alpha_i (AB^{-1})^{i-1} (A x^n - b). \quad (8)$$

Если произвести оптимизацию параметров α_i так, чтобы оценка отношения норм $\|v^{n+1}\|_3 / \|v^n\|_3$ была наилучшей, то, так же как в § 7, получится наилучший итерационный метод вида (8).

Аналогом оптимального линейного итерационного метода (7.16), (7.17) будет итерационный процесс

$$B y^{n+1} = B y^n + \bar{\omega}_n \bar{\omega}_{n-1} B (y^n - y^{n-1}) - \frac{2}{M_1 + \mu_1} (1 + \bar{\omega}_n \bar{\omega}_{n-1}) (A y^n - b),$$

где $\bar{\omega}_n$ строятся по M_1 и μ_1 так же, как ω_n по M и μ .

Объединением метода (2) с методом наискорейшего спуска является следующий метод. Приближение x^{n+1} ищется из соотношения

$$B x^{n+1} = B x^n - \alpha_n (A x^n - b), \quad (9)$$

при этом коэффициент α_n определяется из условия минимума

$$F(x) = (Ax, x) - 2(b, x),$$

из (9) получаем

$$B \frac{dx^{n+1}}{d\alpha_n} = -(Ax^n - b).$$

Поскольку

$$\begin{aligned} \frac{dF(x^{n+1})}{d\alpha_n} &= 2 \left(Ax^{n+1} - b, \frac{dx^{n+1}}{d\alpha_n} \right) = \\ &= 2 (A(x^n - \alpha_n B^{-1}(Ax^n - b)) - b, -B^{-1}(Ax^n - b)), \end{aligned}$$

то условие $\frac{dF(x^{n+1})}{d\alpha_n} = 0$ является линейным уравнением относительно α_n .

Аналог итерационного процесса (10.9), (10.10) имеет вид

$$Bx^{n+1} = Bx^n + \alpha_n (Bx^n - Bx^{n-1}) + \beta_n (Ax^n - b).$$

Методы, рассмотренные в этом параграфе, получили широкое применение при решении систем уравнений $Ax = b$ с большим разбросом собственных значений матрицы $A > 0$.

§ 13. Погрешность приближенного решения системы уравнений и обусловленность матриц. Регуляризация

Предположим, что матрица и правая часть системы заданы неточно и вместо предъявленной к решению системы

$$Ax = b \tag{1}$$

в действительности должна была решаться некоторая система

$$A_1 x = b_1, \quad A_1 = A + \Delta, \quad b_1 = b + \eta. \tag{2}$$

Пусть известны оценки $\|\Delta\|$ и $\|\eta\|$. Займемся оценкой погрешности решения.

Сначала выделим главный член погрешности. Будем обозначать решения (1) и (2) через X и X^* и разность $X^* - X$ через r . Подставим выражения A_1 , b_1 и X^* в (2)

$$(A + \Delta)(X + r) = b + \eta.$$

Вычитая (1), получим

$$Ar + \Delta X + \Delta r = \eta,$$

откуда

$$Ar = \eta - \Delta X - \Delta r$$

и

$$r = A^{-1}(\eta - \Delta X - \Delta r). \tag{3}$$

Если $\|\Delta\|$ и $\|\eta\|$ малы, то следует ожидать и малости $\|r\|$. Тогда слагаемое Δr имеет более высокий порядок малости; отбрасывая

это слагаемое, получаем

$$r \approx A^{-1}(\eta - \Delta X).$$

Отсюда следует оценка погрешности

$$\|r\| \leq \sigma \approx \|A^{-1}\|(\|\eta\| + \|\Delta\| \|X\|). \quad (4)$$

Строгая оценка погрешности получается следующим образом. Вследствие (3) выполняется неравенство

$$\|r\| \leq \|A^{-1}\| \|\eta\| + \|A^{-1}\| \|\Delta\| \|X\| + \|A^{-1}\| \|\Delta\| \|r\|.$$

Предположим, что $\|A^{-1}\| \|\Delta\| < 1$. Перенеся последнее слагаемое в левую часть и поделив неравенство на коэффициент при $\|r\|$, получим оценку

$$\|r\| \leq \frac{\|A^{-1}\|(\|\eta\| + \|\Delta\| \|X\|)}{1 - \|A^{-1}\| \|\Delta\|}. \quad (5)$$

Довольно распространен случай, когда погрешность матрицы системы существенно меньше погрешности правой части; в качестве модели этой ситуации будем рассматривать случай точного задания матрицы системы. Тогда, полагая в (5) $\Delta = 0$, имеем

$$\|r\| \leq \|A^{-1}\| \|\eta\|.$$

Для качественной характеристики связи между погрешностями правой части и решения вводятся понятия *обусловленности системы и обусловленности матрицы системы*. Абсолютные погрешности правой части и решения системы зависят от масштабов, которыми измеряются эти величины и матрица системы. Поэтому правильнее характеризовать свойства системы через связь между относительными погрешностями правой части и решения.

Соответственно этому в качестве *меры обусловленности системы* принимается число

$$\tau = \sup_{\eta} \left(\frac{\|r\|}{\|X\|} : \frac{\|\eta\|}{\|b\|} \right) = \frac{\|b\|}{\|X\|} \sup_{\eta} \frac{\|r\|}{\|\eta\|}.$$

Согласно этому определению, имеем оценку относительной погрешности решения через меру обусловленности и относительную погрешность правой части:

$$\frac{\|r\|}{\|X\|} \leq \tau \frac{\|\eta\|}{\|b\|}. \quad (6)$$

Так как $r = A^{-1}\eta$, то

$$\sup_{\eta} \frac{\|r\|}{\|\eta\|} = \|A^{-1}\|$$

и

$$\tau = \frac{\|b\|}{\|x\|} \|A^{-1}\|.$$

Иногда удобнее иметь более грубую характеристику свойств системы только через свойства матрицы A . Эту характеристику $\nu(A) = \sup_b \tau$ называют *мерой обусловленности матрицы A* . Согласно этому определению и (6), имеем оценку

$$\frac{\|r\|}{\|x\|} \leq \nu(A) \frac{\|\eta\|}{\|b\|},$$

связывающую относительные погрешности правой части и решения только через свойства матрицы системы. Так как

$$\sup_b \frac{\|b\|}{\|x\|} = \sup_x \frac{\|Ax\|}{\|x\|} = \|A\|,$$

то

$$\nu(A) = \|A\| \|A^{-1}\|.$$

Любая норма матрицы не меньше ее наибольшего по модулю собственного значения, т. е. $\|A\| \geq \max |\lambda_A|$. Собственные значения матрицы A и A^{-1} взаимно обратны; поэтому

$$\|A^{-1}\| \geq \max \frac{1}{|\lambda_A|} = \frac{1}{\min |\lambda_A|}.$$

Таким образом,

$$\nu(A) \geq \frac{\max |\lambda_A|}{\min |\lambda_A|} \geq 1.$$

В частности, при $A = A^*$ имеем $\|A\|_3 = \max |\lambda_A|$ и

$$\|A^{-1}\|_3 = \max \frac{1}{|\lambda_A|} = \frac{1}{\min |\lambda_A|},$$

тогда

$$\nu(A) = \frac{\max |\lambda_A|}{\min |\lambda_A|}.$$

Рассмотрим вопрос о погрешности решения вследствие округления в ЭВМ правой части. Пусть, как обычно, t — двоичная разрядность чисел в ЭВМ. Каждый элемент b_i правой части округляется с относительной погрешностью $O(2^{-t})$, т. е. абсолютной $O(|b_i| 2^{-t})$, поэтому

$$\|\eta\| = O(\|b\| 2^{-t}) \quad \text{и} \quad \frac{\|\eta\|}{\|b\|} = O(2^{-t}).$$

Следовательно,

$$\frac{\|r\|}{\|x\|} \leq \nu(A) O(2^{-t}).$$

При практической работе вопрос о строгой оценке погрешности полученного приближенного решения системы линейных уравнений с помощью полученных неравенств или каким-либо иным способом возникает редко. Однако информация о порядке погрешности решения часто полезна для получения качественных выводов о том, с какой точностью разумно решать задачу. Соотношения (4), (5) оценивают сверху погрешность решения, являющуюся следствием погрешности исходных данных. Из равенства (3) видно, что оценки (4), (5) довольно точны, и поэтому обычно не имеет смысла стремиться получать решение задачи с погрешностью, существенно меньшей σ .

Системы уравнений и матрицы с большими значениями мер обусловленности принято называть *плохо обусловленными*, а с малыми — *хорошо обусловленными*. Если правая часть (4), оценивающая погрешность решения через погрешность исходных данных, или оценка вычислительной погрешности недопустимо велики, полезно принять во внимание какую-то дополнительную информацию о решении рассматриваемой задачи. Подход к решению такой задачи должен быть таким же, как в случае других *некорректных задач*.

Рассмотрим простейший случай, когда A симметричная. Пусть $\lambda_1, \dots, \lambda_m \neq 0$ — ее собственные значения, упорядоченные в порядке убывания $|\lambda_i|$, $e_1, \dots, e_m$ — соответствующая ортонормированная система собственных векторов. Решением системы $Ax = b$ является вектор

$$X = \sum_{i=1}^m c_i e_i, \quad c_i = \frac{(b, e_i)}{\lambda_i}.$$

При реально заданной правой части $\tilde{b} = b + \eta$ решением будет

$$\tilde{X} = \sum_{i=1}^m \frac{(b, e_i) + (\eta, e_i)}{\lambda_i} e_i.$$

Коэффициент $\frac{(\eta, e_i)}{\lambda_i}$ может оказаться очень большим при малых λ_i , что приведет к сильному искажению решения. Иногда заранее известно, что в разложении искомого решения задачи $\sum_{i=1}^m c_i e_i$ коэффициенты c_i , соответствующие малым по модулю λ_i , малы. В этом случае следует принять какие-то меры с тем, чтобы «отфильтровать» эти составляющие решения.

При небольших m для решения этой задачи иногда применяется следующий способ: задаются некоторым $q > 0$, находят

все λ_i и e_i при $i \leq q$ и полагают

$$X \approx \sum_{i=1}^q \frac{(\tilde{b}, e_i)}{\lambda_i} e_i;$$

q следует подобрать, исходя из дополнительной информации о задаче.

Другие два способа проиллюстрируем на примере матриц, где все $\lambda_i > 0$. Первый способ. Задавая некоторое α , находим решение x^α системы

$$(\alpha E + A)x^\alpha = \tilde{b};$$

очевидно, оно записывается в виде

$$x^\alpha = \sum_{i=1}^m \frac{(\tilde{b}, e_i)}{\lambda_i + \alpha} e_i.$$

Так как

$$\frac{1}{\lambda_i} - \frac{1}{\lambda_i + \alpha} = \frac{\alpha}{\lambda_i(\lambda_i + \alpha)},$$

то наличие малого параметра α несущественно изменит слагаемые, соответствующие большим λ_i . В то же время при $\lambda_i \ll \alpha$ имеем

$$\left| \frac{(\tilde{b}, e_i)}{\lambda_i + \alpha} \right| \ll \left| \frac{(\tilde{b}, e_i)}{\lambda_i} \right|;$$

это означает, что введение параметра α приводит к существенному уменьшению роли слагаемых, соответствующих малым λ_i . Подбор оптимального значения α обычно осуществляют экспериментально, сравнивая результаты расчетов при различных α .

Второй способ заключается в следующем. Будем решать систему уравнений каким-либо итерационным способом. Рассмотрим случай итераций по формуле

$$x^{n+1} = x^n - \alpha(Ax^n - \tilde{b}) \quad (7)$$

при некотором начальном приближении x^0 . Пусть

$$\tilde{b} = \sum_{i=1}^m \beta_i e_i, \quad x^n = \sum_{i=1}^m z_i^n e_i;$$

подставим эти выражения в (7); приравнявая коэффициенты при e_i , получим соотношения

$$z_i^{n+1} = \alpha \beta_i + (1 - \alpha \lambda_i) z_i^n.$$

Последовательно выражая каждое z_i^k через предыдущие, имеем

$$\begin{aligned} z_i^n &= \alpha \beta_i + (1 - \alpha \lambda_i) z_i^{n-1} = \\ &= \alpha \beta_i + (1 - \alpha \lambda_i)(\alpha \beta_i + (1 - \alpha \lambda_i) z_i^{n-2}) = \dots \\ &\dots = \alpha \beta_i \sum_{k=0}^{n-1} (1 - \alpha \lambda_i)^k + (1 - \alpha \lambda_i)^n z_i^0. \end{aligned}$$

Если $|1 - \alpha \lambda_i| < 1$, то при $n \rightarrow \infty$

$$\sum_{k=0}^{n-1} (1 - \alpha \lambda_i)^k \rightarrow \sum_{k=0}^{\infty} (1 - \alpha \lambda_i)^k = \frac{1}{\alpha \lambda_i}, \quad (1 - \alpha \lambda_i)^n \rightarrow 0,$$

и поэтому $z_i^n \rightarrow \frac{\beta_i}{\lambda_i}$. Пусть $\alpha \approx (\max \lambda_i)^{-1}$, т. е. относительно мало.

При больших значениях λ_i величина $(1 - \alpha \lambda_i)^n$ быстро стремится к нулю с ростом n и z_i^n близко к своему предельному значению $\frac{\beta_i}{\lambda_i}$. В то же время иногда удается подобрать начальное приближение, для которого величины z_i^0 относительно малы при малых λ_i . Тогда при небольшом n коэффициенты z_i^n , соответствующие таким λ_i , еще не будут недопустимо большими и получаемое приближение может оказаться приемлемым.

В других случаях решение задачи находят, минимизируя некоторый функционал, близкий к $F(x) = (Ax, x) - 2(b, x)$, например, функционал $F(x) + \alpha(x, x)$ с малым $\alpha > 0$.

Успешность применения описанных приемов в случае несимметричных матриц A в существенной степени зависит от структуры жордановой формы и от ряда других свойств матрицы. Здесь часто решение находят, минимизируя функционал

$$(Ax - b, Ax - b) + \alpha(x, x)$$

при малых $\alpha > 0$; значение α опять-таки подбирается экспериментально из сравнения результатов расчетов при различных α .

Другая группа методов основана на представлении матрицы системы A в виде

$$A = G\Lambda P,$$

где G и P — ортогональные, а Λ — двудиagonalная, у которой могут быть отличными от нуля элементы λ_{ij} при $j = i$ и $j = i + 1$.

Большинство из описанных выше методов решения систем уравнений с плохо обусловленной матрицей относится к методам регуляризации.

Задача. Пусть $A^{(m)} = [a_{ij}^m]$ — матрица размерности $m \times m$ с элементами

$$a_{ij}^m = \begin{cases} p & \text{при } j = i, \\ q & \text{при } j = i + 1, \\ 0 & \text{при } j \neq i, i + 1. \end{cases}$$

1) Вычислить матрицу $(A^{(m)})^{-1}$ и доказать утверждение: при $|q| < |p|$ матрицы $A^{(m)}$ хорошо обусловлены, а при $|q| > |p|$ и m больших — плохо обусловлены.

2) Выписать явно решение системы $A^{(m)}x = b$ через правую часть.

3) Выписать явно через правую часть b вектор x_a , минимизирующий функционал

$$(A^{(m)}x - b, A^{(m)}x - b) + \alpha(x, x).$$

4) Попытаться качественно описать эффект, достигаемый за счет применения такой регуляризации.

Объясним еще одну причину, по которой стараются избегать симметризации матриц, предложенной в § 7. Сначала посмотрим, что происходит, когда операция симметризации применяется формально в случае уже симметричной матрицы. Тогда $A = A^*$ и $A^*A = A^2$. При возведении матрицы в квадрат собственные значения возводятся в квадрат; поэтому при $\|A\| = \|A\|_3$ имеем

$$\nu(A^2) = \frac{\max |\lambda_{A^2}|}{\min |\lambda_{A^2}|} = \frac{(\max |\lambda_A|)^2}{(\min |\lambda_A|)^2} = (\nu(A))^2.$$

Если $\nu(A) = 1$, то $\max |\lambda_A| = \min |\lambda_A|$ и все собственные значения матрицы A равны между собой по модулю; тогда $A = \text{const } E$. За исключением этого частного случая, $\nu(A) > 1$, и поэтому $\nu(A^2) > \nu(A)$. Таким образом, в случае симметричной матрицы применение симметризации увеличивает число обусловленности. Число обусловленности является непрерывной функцией от матрицы; поэтому для матриц, близких к симметричным, применение симметризации также увеличивает число обусловленности матриц. Таким образом, имеется непустой класс несимметричных матриц, по отношению к которым алгоритм симметризации увеличивает число обусловленности.

Задача. Существуют ли несимметричные матрицы, для которых $\nu(A^2) < (\nu(A))^2$?

§ 14. Проблема собственных значений

В различных случаях возникают разные требования к информации о собственных значениях и собственных векторах матриц и это порождает многообразие проблем и приемов решения этой задачи.

1. Для решения ряда задач механики, физики, химии требуется получение всех собственных значений, а иногда и всех собственных векторов некоторых матриц. Эту задачу называют *полной проблемой собственных значений*. Стандартные программы современных ЭВМ решают общую проблему собственных значений для матриц размерности нескольких сотен.

2. В ряде случаев требуется найти максимальное или минимальное по модулю собственное значение матрицы. Проблемы подобного сорта возникают, в частности, при решении некоторых ядерных задач. Здесь приходится решать задачи, эквивалентные задачам отыскания собственных значений матриц размерности , порядка $10^3 - 10^6$, а также и существенно большей. При меньших размерностях матриц для решения этих задач чаще применяют итерационные методы, при больших — вероятностные.

3. При исследовании колебательных процессов иногда требуется найти два максимальных по модулю собственных значения матрицы, причем меньшее из них обычно достаточно определить с меньшей точностью.

4. Там же возникает задача отыскания собственного значения матрицы, наиболее близкого к данному числу λ^0 , или отыскания расстояния от данного числа λ^0 до спектра матрицы.

Формально рассуждая, можно было бы сказать, что все остальные задачи, называемые *частичными проблемами* собственных значений, являются частным случаем общей проблемы собственных значений, и достаточно ограничиться рассмотрением методов решения общей проблемы. Однако такой подход приведет к неоправданно большому объему вычислений. При обсуждении постановок конкретных задач, связанных с отысканием собственных значений матриц, зачастую значительные усилия тратятся именно на установление минимального объема информации о спектре матрицы, которым можно ограничиться.

Решение задач 2—4 обычно сводят к отысканию максимального по модулю собственного значения некоторой матрицы $B = g(A)$ такой, что это собственное значение соответствует отыскиваемому собственному значению матрицы A .

Рассмотрим случай, когда все собственные значения матрицы вещественны. Если требуется найти максимальное или минимальное собственное значение матрицы A , то следует взять $g(A) = A + cE$; очевидно, при достаточно больших положительных (отрицательных) c максимальному по модулю собственному значению матрицы $A + cE$ соответствует максимальное (минимальное) собственное значение матрицы A . В случае задачи 4 при некотором c максимальное по модулю собственное значение матрицы $E - c(A - \lambda^0 E)^2$ соответствует отыскиваемому собственному значению матрицы A . Иногда в качестве такой матрицы $g(A)$ может использоваться матрица $(A - \lambda^0 E)^{-1}$. При этом матрица $(A - \lambda^0 E)^{-1}$ не выписывается в явном виде, а необходимые по ходу вычислений векторы $(A - \lambda^0 E)^{-1}y$ находят, решая систему уравнений $(A - \lambda^0 E)x = y$.

Рассмотрим типичную задачу отыскания двух максимальных по модулю собственных значений матрицы A . Для простоты

предполагаем наличие полной системы собственных векторов e_i : $Ae_i = \lambda_i e_i$, $\|e_i\| = 1$; пусть $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_m|$. Зададимся некоторым вектором x^0 и будем последовательно вычислять векторы $x^{n+1} = Ax^n$. Представим x^0 в виде $x^0 = \sum_{i=1}^m c_i e_i$; очевидно,

$$x^n = \sum_{i=1}^m c_i A^n e_i = \sum_{i=1}^m c_i \lambda_i^n e_i.$$

Отсюда следуют соотношения

$$\begin{aligned} x^n &= c_1 \lambda_1^n e_1 + O(|\lambda_2|^n), \\ (x^n, x^n) &= (c_1 \lambda_1^n e_1 + O(|\lambda_2|^n), c_1 \lambda_1^n e_1 + O(|\lambda_2|^n)) = \\ &= |c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n), \\ (x^{n+1}, x^n) &= (c_1 \lambda_1^{n+1} e_1 + O(|\lambda_2|^{n+1}), c_1 \lambda_1^n e_1 + O(|\lambda_2|^n)) = \\ &= \lambda_1 |c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n). \end{aligned}$$

Положим

$$\lambda_1^{(n)} = \frac{(x^{n+1}, x^n)}{(x^n, x^n)};$$

из последних соотношений при $c_1 \neq 0$ получаем

$$\begin{aligned} \lambda_1^{(n)} &= \frac{\lambda_1 |c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n)}{|c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n)} = \\ &= \frac{\lambda_1 \left(1 + O\left(\frac{1}{|c_1|^2} \left|\frac{\lambda_2}{\lambda_1}\right|^n\right)\right)}{1 + O\left(\frac{1}{|c_1|^2} \left|\frac{\lambda_2}{\lambda_1}\right|^n\right)} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right), \quad (1) \end{aligned}$$

Задача. Доказать, что в случае симметричной матрицы

$$\lambda_1^{(n)} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2n}\right).$$

Кроме (1) имеем

$$\|x^n\| = \sqrt{(x^n, x^n)} = |c_1| |\lambda_1|^n + O(|\lambda_2|^n),$$

$$\begin{aligned} e_1^{(n)} &= \frac{x^n}{\|x^n\|} = \frac{\sum_{i=1}^m c_i \lambda_i^n e_i}{|c_1 \lambda_1^n| + O(|\lambda_2|^n)} = \\ &= \frac{\frac{c_1 \lambda_1^n}{|c_1 \lambda_1^n|} e_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)}{1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)} = e^{i\varphi_n} e_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right); \end{aligned}$$

здесь $\varphi_n = \arg(c_1 \lambda_1^n)$.

Таким образом, этот итерационный процесс также выдает собственный вектор, соответствующий λ_1 .

Может случиться, что у матрицы A имеется два максимальных по модулю собственных значения

$$\lambda_1 \neq \lambda_2, \quad |\lambda_1| = |\lambda_2| > |\lambda_3| \geq \dots$$

Очевидно, тогда величина $\lambda_1^{(n)}$ будет устанавливаться только в частном случае, когда c_1 или c_2 равно нулю. Если заранее известно, что таких собственных значений два, то их и соответствующие им собственные векторы можно также определить, анализируя поведение $\lambda_1^{(n)}$ и $e_1^{(n)}$.

Рассмотрим типичный случай, когда λ_1, λ_2 — вещественные, $\lambda_1 > 0, \lambda_1 = -\lambda_2$. Тогда

$$x^n = c_1 \lambda_1^n e_1 + c_2 (-\lambda_1)^n e_2 + O(|\lambda_3|^n),$$

$$x^{n+2} = c_1 \lambda_1^{n+2} e_1 + c_2 (-\lambda_1)^{n+2} e_2 + O(|\lambda_3|^{n+2}) =$$

$$= \lambda_1^2 (c_1 \lambda_1^n e_1 + c_2 (-\lambda_1)^n e_2) + O(|\lambda_3|^{n+2}).$$

Отсюда получаем, что

$$\tilde{\lambda}^{(n)} = \frac{(x^{n+2}, x^n)}{(x^n, x^n)} = \lambda_1^2 + O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^n\right),$$

и определяем приближенные значения $\lambda_1^{(n)} > 0$ и $\lambda_2^{(n)} < 0$. Имеем:

$$z_1^{n+1} = x^{n+1} + \lambda_1^{(n)} x^n = 2c_1 \lambda_1^{n+1} e_1 + O(|\lambda_3|^{n+1}),$$

поэтому

$$e_1^{(n)} = \frac{z_1^{n+1}}{\|z_1^{n+1}\|} = e^{i\psi_n} e_1 + O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^n\right).$$

Точно так же,

$$z_2^{n+1} = x^{n+1} + \lambda_2^{(n)} x^n = 2c_2 \lambda_1^{n+1} e_2 + O(|\lambda_3|^{n+1})$$

и

$$e_2^{(n)} = \frac{z_2^{n+1}}{\|z_2^{n+1}\|} = e^{i\omega_n} e_2 + O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^n\right).$$

Если λ_i — собственные значения матрицы A , то у сопряженной матрицы A^* собственным значением будет $\bar{\lambda}_i$; при этом, если $Ae_i = \lambda_i e_i$, $A^*g_j = \bar{\lambda}_j g_j$, $\lambda_i \neq \lambda_j$, то $(e_i, g_j) = 0$.

Поэтому при $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \dots$ для нахождения собственного значения λ_2 можно поступить так. Получив приближения $\tilde{e}_1 \approx e_1$, аналогично определяем приближение $\tilde{g}_1$ к вектору g_1 ; перенормируем $\tilde{g}_1$ из условия $(\tilde{e}_1, \tilde{g}_1) = 1$. Далее ведем итерации по формуле $y^{n+1} = Ay^n$; для исключения компоненты, пропорциональной e_1 , время от времени векторы y^n ортогонализируем по отношению к вектору $\tilde{g}_1$, т. е. начальным для последующих

итераций вместо y^n берем вектор

$$\tilde{y}_n = y^n - (y^n, \tilde{g}_1) \tilde{e}_1.$$

Естественно, сходимость итерационного процесса лучше, если в начальном приближении $y_0 = \sum_{i=1}^m d_i e_i$ слагаемое $d_2 e_2$ преобладает над остальными слагаемыми. Возьмем из описанного выше итерационного процесса отыскания λ_1 и e_1 некоторое приближение x^l ; проортогонализируем его по отношению к g_1 . Получившийся вектор $y^0 = x^l - (x^l, \tilde{g}_1) \tilde{e}_1$ примем за начальное приближение. Не следует брать l малым, тогда компоненты $c_j \lambda_j^l e_j$ при $j > 2$ еще не будут малы по сравнению с $c_2 \lambda_2^l e_2$; в то же время не следует брать l очень большим, тогда компонента $c_2 \lambda_2^l e_2$ будет малой по сравнению с вычислительной погрешностью.

Можно встретить следующее высказывание. Если $c_1 = 0$, то описанный итерационный процесс, казалось бы, не должен давать приближения, сходящегося к максимальному по модулю собственному значению. Однако в действительности из-за присутствия округлений в процессе итераций появится компонента, пропорциональная e_1 , и вследствие этого требуемый результат все равно будет получен. По поводу этого высказывания можно отметить следующее. При ручном счете с малой точностью обычно дело обстоит так, как сказано выше. В случае современных ЭВМ с большой разрядностью может случиться, что после некоторого числа итераций влияние вычислительной погрешности еще не будет существенно, в то время как величина $\sum_{i=3}^m c_i \lambda_i^n e_i$ будет мала по сравнению с $c_2 \lambda_2^n e_2$. Тогда $Ax^n \approx \text{const } x^n$, и

можно сделать неверный вывод, что искомое первое собственное значение найдено. В ряде случаев появление составляющей, пропорциональной e_1 , не является неизбежным даже в случае присутствия округлений. При решении проблемы собственных значений для дифференциальных и интегральных операторов иногда возникают матрицы A со специфическими свойствами; например, часто встречается случай, когда при всех i, j выполняется равенство

$$a_{ij} = a_{m+1-i, m+1-j} \quad (2)$$

Для определенности рассмотрим случай четного m . Назовем четными векторы x , компоненты которых связаны равенством $x_i = x_{m+1-i}$, $i = 1, \dots, m$, и нечетными — равенством $x_i = -x_{m+1-i}$. При условии (2) вектор Ax четен, если x четен, и нечетен, если x — нечетен. Поэтому подпространства четных и нечетных векторов являются собственными для оператора A . Следовательно, существует полная система собственных векторов, принадлежащих этим подпространствам, т. е. являющихся или четными, или нечетными. Это обстоятельство может быть существенно использовано; если вектор x^0 четен или нечетен, то этим же свойством обладают все векторы x^n , поэтому при отыскании каждого последующего вектора x^n достаточно ограничиться определением первой половины его компонент. Кроме того, можно объединить коэффициенты, соответствующие компонентам x_i и x_{m+1-i} . Например, при четном x^0 вычисления компонент x^n можно производить по формулам

$$x_i^{n+1} = \sum_{j=1}^{m/2} (a_{ij} + a_{i, m+1-j}) x_j^n.$$

В случае итераций такого вида мы не выходим за пределы подпространств четных или нечетных векторов, соответственно. Поэтому если x^0 и e_1 принадлежат различным подпространствам, составляющая пропорциональная e_1 так и не появится. Кроме непосредственного уменьшения вычислений при отыскании каждого вектора x^n использование свойства (2) полезно также по следующей причине. Довольно типичен случай, когда собственные векторы с небольшими нечетными номерами являются четными, а с небольшими четными номерами — нечетными; предполагается, что $|\lambda_1| > |\lambda_2| > |\lambda_3| > \dots$. Тогда при x^0 четном всегда

$$x^n = c_1 \lambda_1^n e_1 + c_3 \lambda_3^n e_3 + \dots = c_1 \lambda_1^n e_1 + O(|\lambda_3|^n),$$

следовательно,

$$\lambda_1^{(n)} - \lambda_1 = O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^n\right) \quad \text{при} \quad \lambda^{(n)} = \frac{(x^{n+1}, x^n)}{(x^n, x^n)}.$$

Соответственно, при x^0 нечетном имеем

$$x^n = c_2 \lambda_2^n e_2 + O(|\lambda_4|^n)$$

и

$$\lambda_2^{(n)} - \lambda_2 = O\left(\left|\frac{\lambda_4}{\lambda_2}\right|^n\right) \quad \text{при} \quad \lambda_2^{(n)} = \frac{(x^{n+1}, x^n)}{(x^n, x^n)}.$$

При этом не возникает никакой проблемы подавления составляющей, пропорциональной e_1 .

Если $|\lambda_1| > 1$, то $\|x^n\| \rightarrow \infty$ при $n \rightarrow \infty$, и поэтому при достаточно большом n произойдет переполнение разрядности чисел и остановка ЭВМ. Если $|\lambda_1| < 1$, то $\|x^n\| \rightarrow 0$, и, вследствие конечности порядков чисел в машине, может случиться, что, начиная с некоторого n , $x^n \equiv 0$. Чтобы избежать этих явлений, полезно время от времени умножать вектор x^n на подходящий множитель.

Для практической оценки погрешности и ускорения сходимости итерационных процессов может быть применен δ^2 -процесс и другие приемы, аналогичные методам ускорения сходимости при решении систем линейных уравнений. Например, могут применяться итерации вида $x^{n+1} = g_h(A)x^n$ со специальным подбором многочлена $g_h(A)$ в зависимости от известной информации о спектре матрицы A .

Поскольку

$$\max \lambda_A' = \sup_x \frac{(Ax, x)}{(x, x)}, \quad \min \lambda_A' = \inf_x \frac{(Ax, x)}{(x, x)}$$

при $A = A^*$, то некоторые приемы отыскания максимального и минимального собственного значения матрицы A основываются на идее отыскания стационарных точек функционала

$$\Phi(x) = \frac{(Ax, x)}{(x, x)}.$$

Задача. Пусть $\lambda_1 \approx 5$, $1 \leq \lambda_i \leq 3$ при $i = 2, \dots, m$. Построить итерационный процесс вида $x^{n+1} = (A + \varepsilon E)x^n$ для получения λ_1 с наилучшей при данной информации скоростью сходимости.

Сделать, то же самое, если $\lambda_1 \approx 1$, $2 \leq |\lambda_i| \leq 3$ при $i = 2, \dots, m$.

§ 15. Решение полной проблемы собственных значений для симметричной матрицы методом вращений

Всякая симметричная матрица A может быть записана в виде

$$A = U^* \Lambda U, \quad (1)$$

где U — ортогональная матрица, Λ — диагональная с диагональными элементами $\lambda_1, \dots, \lambda_m$ — собственными значениями матрицы A . Из (1) получаем, что

$$AU^* = U^* \Lambda.$$

Если расписать это равенство по столбцам, то окажется, что каждый i -й столбец матрицы U^* является собственным вектором, соответствующим собственному значению λ_i .

Преобразуя (1), имеем

$$U A U^* = \Lambda.$$

Построим последовательность ортогональных матриц $V_1 \dots V_n$ такую, что при $W_n = V_n \dots V_1$, $n \rightarrow \infty$, справедливо

$$W_n A W_n^* \rightarrow \Lambda.$$

Обозначим через $V_{kl}(\varphi)$ матрицу с элементами $v_{kk} = v_{ll} = \cos \varphi$, $v_{kl} = -v_{lk} = \sin \varphi$, $v_{ii} = 1$ при $i \neq k$, $i \neq l$, $v_{ij} = 0$ при остальных (i, j) . Очевидно, что она ортогональная. Пусть B — симметричная матрица. Рассмотрим матрицу

$$B(\varphi) = V_{kl}(\varphi) B V_{kl}^*(\varphi). \quad (2)$$

Лемма. Для всякой симметричной матрицы B можно указать матрицу $V_{kl}(\varphi_0)$ такую, что

$$\sum_{i \neq l} |b_{il}^2(\varphi_0)| \leq \left(1 - \frac{2}{m(m-1)}\right) \sum_{i \neq l} |b_{il}|^2.$$

Доказательство. Очевидно,

$$B(0) = B.$$

Матрицы $B(0)$ и $B(\varphi)$ подобны, и поэтому их характеристические многочлены

$$\lambda^m - s_1(\varphi) \lambda^{m-1} + s_2(\varphi) \lambda^{m-2} - \dots = 0$$

не зависят от φ .

Здесь

$$s_1(\varphi) = \sum_{i=1}^m b_{ii}(\varphi), \quad s_2(\varphi) = \sum_{i < j} (b_{ij}(\varphi) b_{ji}(\varphi) - (b_{ij}(\varphi))^2);$$

при выписывании явного вида $s_2(\varphi)$ мы воспользовались симметричностью матрицы $B(\varphi)$. Отсюда следует, что величина

$$s(\varphi) = s_1^2(\varphi) - 2s_2(\varphi) = \sum_{i,j} (b_{ij}(\varphi))^2$$

не зависит от φ :

$$s(\varphi) = s(0). \quad (3)$$

Пусть $|b_{kl}| = \max_{i \neq j} |b_{ij}|$, $k < l$.

Введем обозначение

$$B_{kl}(\varphi) = \begin{pmatrix} b_{kk}(\varphi) & b_{kl}(\varphi) \\ b_{lk}(\varphi) & b_{ll}(\varphi) \end{pmatrix}.$$

Непосредственно перемножая матрицы в правой части (2) и сравнивая элементы, стоящие на пересечении k -й и l -й строк с k -м и l -м столбцами, получаем соотношения

$$B_{kl}(\varphi) = G(\varphi) B_{kl}(0) G^*(\varphi), \quad (4)$$

где

$$G(\varphi) = \begin{pmatrix} \cos \varphi & \sin \varphi \\ -\sin \varphi & \cos \varphi \end{pmatrix}.$$

Равенство (3) выполнено и по отношению к матрицам $B_{kl}(\varphi)$, $B_{kl}(0)$; поэтому

$$(b_{kk}(\varphi))^2 + 2(b_{kl}(\varphi))^2 + (b_{ll}(\varphi))^2 = (b_{kk})^2 + 2(b_{kl})^2 + (b_{ll})^2. \quad (5)$$

Складывая это равенство с равенствами

$$(b_{qq}(\varphi))^2 = (b_{qq})^2, \quad q \neq k, l,$$

получим

$$\sum_{q=1}^m (b_{qq}(\varphi))^2 + 2(b_{kl}(\varphi))^2 = \sum_{q=1}^m (b_{qq})^2 + 2(b_{kl})^2.$$

Вычтем это равенство из (3)

$$\sum_{i \neq j} (b_{ij}(\varphi))^2 - 2(b_{kl}(\varphi))^2 = \sum_{i \neq j} (b_{ij})^2 - 2(b_{kl})^2. \quad (6)$$

Непосредственно из (4) следует равенство

$$b_{kl}(\varphi) = b_{kl} \cos 2\varphi - \frac{1}{2}(b_{kk} - b_{ll}) \sin 2\varphi.$$

Пусть

$$\varphi_0 = \frac{1}{2} \operatorname{arctg} \frac{2b_{kl}}{b_{kk} - b_{ll}};$$

тогда $b_{kl}(\varphi_0) = 0$, и равенство (6) приобретет вид

$$\sum_{i \neq j} (b_{ij}(\varphi_0))^2 = \sum_{i \neq j} (b_{ij})^2 - 2(b_{kl})^2.$$

В сумме $\sum_{i \neq j} b_{ij}^2$ содержится $m(m-1)$ слагаемых b_{ij}^2 ; поскольку $(b_{kl})^2 = \max_{i \neq j} (b_{ij})^2$, то

$$2(b_{kl})^2 \geq \frac{2}{m(m-1)} \sum_{i \neq j} (b_{ij})^2.$$

Отсюда следует требуемая оценка

$$\sum_{i \neq j} (b_{ij}(\varphi_0))^2 \leq \left(1 - \frac{2}{m(m-1)}\right) \sum_{i \neq j} (b_{ij})^2.$$

Лемма доказана.

Построим последовательность матриц V_n и $A^{(n)} = [a_{ij}^{(n)}]$ по следующему правилу:

$$A^{(1)} = V_1 A V_1^*, \dots, A^{(n)} = V_n A^{(n-1)} V_n^*.$$

Матрица $V_n = B_{kl}(\varphi_0)$ строится в соответствии с леммой в зависимости от матрицы $A^{(n-1)}$ так, чтобы выполнялось неравенство

$$\sum_{i \neq j} (a_{ij}'^{(n)})^2 \leq \left(1 - \frac{2}{m(m-1)}\right) \sum_{i \neq j} (a_{ij}'^{(n-1)})^2. \quad (7)$$

Выражая каждую матрицу $A^{(n)}$ через предыдущую, получаем

$$A^{(n)} = V_n \dots V_1 A V_1^* \dots V_n^*,$$

иначе $A^{(n)} = W_n A W_n^*$, где $W_n = V_n \dots V_1$. Согласно оценке (7), имеем

$$\sum_{i \neq j} (a_{ij}^{(n)})^2 \leq \left(1 - \frac{2}{m(m-1)}\right)^n \sum_{i \neq j} (a_{ij})^2.$$

Пусть $\Lambda^{(n)}$ — матрица, получающаяся из $A^{(n)}$ заменой всех внедиагональных элементов на нулевые. Поскольку матрицы $\Lambda^{(n)}$ и $A^{(n)}$ близки при больших n , то близки коэффициенты их характеристических многочленов. Корни многочлена непрерывно зависят от коэффициентов в области, где старший коэффициент отличен от нуля; старший коэффициент характеристического многочлена равен 1; поэтому собственные значения матриц $\Lambda^{(n)}$ и $A^{(n)}$ близки между собой. Собственные значения матрицы $\Lambda^{(n)}$ — это диагональные элементы матрицы $A^{(n)}$, а собственные значения подобных матриц $A^{(n)}$ и A совпадают. Таким образом, величины $a_{ii}^{(n)}$ являются приближениями к собственным значениям матрицы A .

Приведем другое доказательство этого же утверждения, но уже с оценкой погрешности. Применяя дважды неравенство

$$|\sum_i \alpha_i \beta_i| \leq \sqrt{\left(\sum_i |\alpha_i|^2\right) \left(\sum_i |\beta_i|^2\right)},$$

получаем, что для любой матрицы D выполняется неравенство

$$\begin{aligned} |(Dx, x)| &= \left| \sum_i \left(\sum_j d_{ij} x_j \right) x_i \right| \leq \sum_i \left| \sum_j d_{ij} x_j \right| \cdot |x_i| \leq \\ &\leq \sum_i \sqrt{\left(\sum_j |d_{ij}|^2 \right) \left(\sum_j |x_j|^2 \right)} |x_i| = \sqrt{\sum_j |x_j|^2} \left(\sum_i \sqrt{\sum_j |d_{ij}|^2} |x_i| \right) \leq \\ &\leq \sqrt{\sum_j |x_j|^2} \sqrt{\left(\sum_i \left(\sum_j |d_{ij}|^2 \right) \right) \sum_i |x_i|^2} = \sqrt{\sum_{ij} |d_{ij}|^2} (x, x). \quad (8) \end{aligned}$$

Напомним формулировку минимаксного свойства собственных значений. Пусть матрица A симметрична, $\lambda_1, \dots, \lambda_m$ — ее собственные значения, упорядоченные по убыванию. Тогда

$$\lambda_k = \inf_{y_1, \dots, y_{k-1}} \sup_{(x, y_1) = \dots = (x, y_{k-1}) = 0, x \neq 0} \frac{(Ax, x)}{(x, x)}.$$

При любом x , согласно (8), имеем

$$|(A^{(n)}x, x) - (\Lambda^{(n)}x, x)| = |((A^{(n)} - \Lambda^{(n)})x, x)| \leq \sqrt{\sum_{i \neq j} |a_{ij}^{(n)}|^2} (x, x). \quad (9)$$

Пусть λ_k — собственные значения матрицы A , а следовательно, и $A^{(n)}$, занумерованные в порядке убывания, а $\lambda_k^{(n)}$ — собственные значения матрицы $\Lambda^{(n)}$, т. е. ее диагональные элементы, также занумерованные в порядке убывания. На основании (9) имеем

$$\left| \frac{(A^{(n)}x, x)}{(x, x)} - \frac{(\Lambda^{(n)}x, x)}{(x, x)} \right| \leq \sigma_n = \sqrt{\sum_{i \neq j} |a_{ij}^{(n)}|^2}.$$

Если две функции отличаются не более чем на ε , то их верхние (нижние) грани по некоторому множеству тоже отличаются не более чем на ε . Отсюда последовательно вытекают неравенства

$$\left| \sup_{(x, y_1) = \dots = (x, y_{k-1}) = 0, x \neq 0} \frac{(A^{(n)}x, x)}{(x, x)} - \sup_{(x, y_1) = \dots = (x, y_{k-1}) = 0, x \neq 0} \frac{(\Lambda^{(n)}x, x)}{(x, x)} \right| \leq \sigma_n \quad \text{при всех } y_1, \dots, y_{k-1}$$

и

$$|\lambda_k^{(n)} - \lambda_k| \leq \sigma_n \quad \text{при всех } k.$$

Столбцы матрицы W_n оказываются приближениями к собственным векторам матрицы A .

Все столбцы произведения матриц $V_{kl}(\varphi_0)$ и $A^{(n)}$, кроме k -го и l -го, совпадают с соответствующими столбцами матрицы $A^{(n)}$; k -й и l -й столбцы новой матрицы являются линейными комбинациями этих же столбцов матрицы A . Точно такого же числа действий потребует дальнейшее умножение на матрицу $V_{kl}(\varphi)$, когда происходит изменение k -й и l -й строк. Общее число арифметических операций для получения каждой последующей матрицы $A^{(n)}$ оказывается $\sim 6m$.

Предложенный алгоритм неприятен тем, что рекомендуется на каждом шаге выбирать наибольший по модулю элемент матрицы $A^{(n)}$; эта процедура требует $\sim m^2$ операций. Оказывается, условие выбора максимального по модулю элемента не является обязательным.

При некоторых реализациях алгоритма запоминают суммы $s_i = \sum_{j \neq i} |a_{ij}|^2$ и выбирают наибольший по модулю элемент в строке с максимальным значением s_i ; при этом оценка (7) также остается в силе. При других реализациях аналогичные преобразования, соответствующие различным элементам матрицы $a_{ij}^{(n)}$ с $j > i$, производятся в циклическом порядке. Для ряда алгоритмов имеет место так называемая квадратичная сходимость. После цикла из $m(m-1)/2$ шагов получается матрица $A^{(n+(m(m-1)/2))}$ такая, что

$$\sum_{i \neq j} |a_{ij}^{(n+(m(m-1)/2))}|^2 \leq c \left(\sum_{i \neq j} |a_{ij}^{(n)}|^2 \right)^2.$$

Это обеспечивает очень быстрое уточнение результата после получения достаточно хорошего приближения.

Для решения полной проблемы собственных значений в случае несимметричной матрицы применяются более сложные, но столь же однородные по своей структуре алгоритмы.

Дополнительная литература к главе VI

I [1, 2, 14, 15], II [8, 25, 42]; §§ 4, 5: I [4], II [10]; § 7: II [7, 8, 25, 39, 40], § 10: III [31]; § 11: II [5, 26]; § 12: II [8], III [26]; § 13: II [22], III [46]; §§ 14, 15: I [2], II [42].

РЕШЕНИЕ СИСТЕМ НЕЛИНЕЙНЫХ УРАВНЕНИЙ
И ЗАДАЧ ОПТИМИЗАЦИИ

Решение задач оптимизации, будь то оптимизация производственных или экономических процессов, оптимизация конструкций или оптимизация численных алгоритмов, сводится в математическом плане к отысканию точек экстремума функционалов.

В наиболее типичных случаях возникает задача минимизации функции большого числа переменных в области G , задаваемой большим числом ограничений типа неравенств или равенств: ищется

$$\inf_{x_1, \dots, x_m} \Phi(x_1, \dots, x_m)$$

при условиях

$$\varphi_i(x_1, \dots, x_m) \geq 0, \quad i = 1, \dots, l.$$

$$\psi_i(x_1, \dots, x_m) = 0, \quad i = 1, \dots, q.$$

Задача минимизации функций большого числа переменных возникает также в случае применения вариационных методов к решению задач математической физики и в других разделах прикладной математики. Системы уравнений

$$F(x) = 0: f_i(x_1, \dots, x_m) = 0, \quad i = 1, \dots, m,$$

возникают в случае многих задач указанных выше типов; например, в главе IX будет идти речь о подобных системах, возникающих при решении краевых задач.

Иногда оказывается полезным сведение одной из этих задач к другой.

Если $\Psi(y_1, \dots, y_m) > 0$ при

$$(y_1, \dots, y_m) \neq (0, \dots, 0), \quad \Psi(0, \dots, 0) = 0,$$

то решение системы $F(x) = 0$ равносильно минимизации функции

$$\Psi(f_1(x_1, \dots, x_m), \dots, f_m(x_1, \dots, x_m)).$$

С другой стороны, пусть $\inf_G \Phi(x_1, \dots, x_m)$ достигается в точке X , внутренней по отношению к G , и функция Φ дифференцируема в этой точке. Тогда точка минимума является решением системы уравнений

$$\Phi'_i = 0, \quad i = 1, \dots, m.$$

Возможность сведения одной из этих задач к другой не означает, что достаточно ограничиться рассмотрением только одной из них, родство этих задач скорее подчеркивает, что они одинаково трудны. О трудности этих задач свидетельствует то, что в этой главе не приведено ни одного универсального алгоритма решения этих задач в случае больших m ; отсутствие таких алгоритмов вызывается существом дела. При решении задач каждого типичного в приложениях класса приходится заниматься теоретической и экспериментальной «доводкой» методов применительно к этому классу задач.

Сведение одной задачи к другой производится на практике с целью снижения трудоемкости решения.

Например, при решении систем нелинейных уравнений иногда поступают следующим образом. Строится функционал, минимум которого достигается на решении системы. Затем, задавшись начальным приближением к точке минимума, проводят итерации каким-либо из методов спуска (см. § 4) и таким путем получают удовлетворительное приближение к решению системы. Исходя из этого приближения, производят уточнения при помощи какого-либо итерационного метода, специфического для задачи решения системы уравнений, например метода Ньютона (см. § 2).

Поясним причины, вызывающие такое комбинированное применение методов. Назовем *областью сходимости метода* множество начальных условий, при которых итерации по данному методу сходятся к решению задачи. Применение методов спуска на первоначальном этапе вызвано тем, что обычно они имеют более широкую область сходимости, чем методы, специфические для задачи решения системы уравнений. В то же время последние методы обычно обладают лучшей скоростью сходимости при наличии достаточно хорошего начального приближения; это и обуславливает их применение на заключительном этапе итераций.

На примере решения системы линейных уравнений было также видно, что сведение одной задачи к другой приводит к конструированию новых методов решения исходной задачи.

§ 1. Метод простой итерации и смежные вопросы

Так же, как в случае линейных уравнений, начнем изучение итерационных методов с *метода простой итерации*.

Этот метод состоит в следующем: система преобразуется к виду

$$x = g(x), \quad (1)$$

иначе

$$x_i = g_i(x_1, \dots, x_m), \quad i = 1, \dots, m,$$

и итерации проводятся по формуле

$$x^{n+1} = g(x^n), \quad (2)$$

иначе

$$x_i^{n+1} = g_i(x_1^n, \dots, x_m^n), \quad i = 1, \dots, m.$$

Подойдем к изучению этого метода с более общих позиций. Пусть $\mathcal{X}$ — полное метрическое пространство, а оператор $y = g(x)$ отображает $\mathcal{X}$ в себя.

Рассмотрим итерационный процесс

$$x^{n+1} = g(x^n) \quad (3)$$

решения уравнения

$$x = g(x). \quad (4)$$

Если при некотором $q < 1$ отображение $y = g(x)$ удовлетворяет условию

$$\rho(g(x_1), g(x_2)) \leq q\rho(x_1, x_2) \quad (5)$$

при всех x_1, x_2 , то такое отображение называют *сжатым*.

Теорема. Если отображение $y = g(x)$ — сжатое, то уравнение $x = g(x)$ имеет решение X и

$$\rho(X, x^n) \leq \frac{q^n a}{1 - q};$$

здесь $a = \rho(x^1, x^0)$.

Доказательство. Согласно (5), имеем

$$\rho(x^{n+1}, x^n) = \rho(g(x^n), g(x^{n-1})) \leq q\rho(x^n, x^{n-1}),$$

поэтому $\rho(x^{n+1}, x^n) \leq q^n \rho(x^1, x^0) = q^n a$. При $p > n$ имеем цепочку неравенств

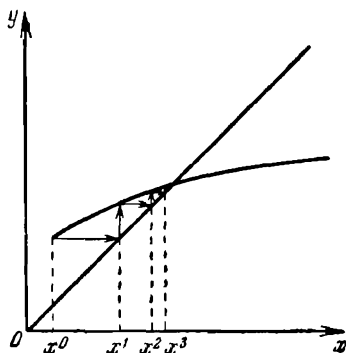
$$\begin{aligned} \rho(x^p, x^n) &\leq \rho(x^p, x^{p-1}) + \dots + \rho(x^{n+1}, x^n) \leq \\ &\leq q^{p-1}a + \dots + q^na \leq q^na \sum_{i=0}^{\infty} q^i = \frac{q^na}{1-q}. \end{aligned} \quad (6)$$

Согласно критерию Коши, последовательность x^n имеет некоторый предел X . Переходя к пределу в (6) при $p \rightarrow \infty$,

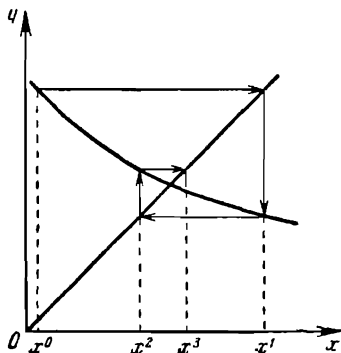
Нахождение каждого нового значения x_i^{n+1} требует решения, вообще говоря, нелинейного уравнения

$$f_i(x_1^{n+1}, \dots, x_{i-1}^{n+1}, x_i^{n+1}, x_{i+1}^n, \dots, x_m^n) = 0$$

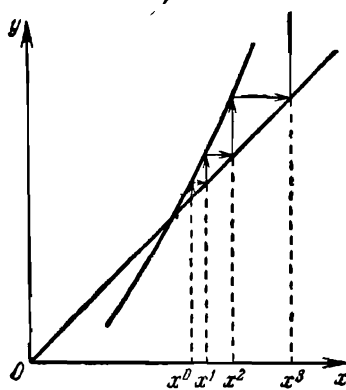
с одним неизвестным.



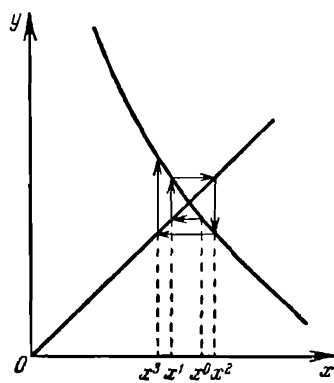
а)



б)



в)



г)

Рис. 7.1.1.

Промежуточное место между итерационными методами (2) и (7) занимает метод, где компоненты приближений определяются из соотношений

$$\begin{aligned} x_1^{n+1} &= g_1(x_1^n, \dots, x_m^n), \\ x_2^{n+1} &= g_2(x_1^{n+1}, x_2^n, \dots, x_m^n), \\ &\dots \dots \dots \\ x_m^{n+1} &= g_m(x_1^{n+1}, \dots, x_{m-1}^{n+1}, x_m^n). \end{aligned} \quad (8)$$

Методы (7) и (8) особенно широко используются в различных моделирующих устройствах.

В достаточно малой окрестности решения X системы для приближений методом простой итерации имеем

$$x^{n+1} - X = g(x^n) - g(X) \approx B(x^n - X), \quad (9)$$

где

$$B = \left[\frac{\partial g_i}{\partial x_j} \right] \bigg|_X.$$

Таким образом, при приближениях, находящихся в малой окрестности решения, ошибки приближений итерационного процесса (2) (а также и процессов (7) и (8)) подчиняются примерно тем же законам, что и ошибки итерационных методов решения систем линейных уравнений. В результате этого удается показать, что при достаточно хорошем начальном приближении итерационный процесс (2) сходится при тех же условиях на матрицу B , что и метод простой итерации. Наличие соотношения (9) позволяет также производить ускорение сходимости итерационных процессов.

Рассмотрим случай $m = 1$ и построим аналог δ^2 -процесса. При имеющемся приближении x^n обозначим $x^{n,1} = g(x^n)$, $x^{n,2} = g(x^{n,1})$. Согласно (9),

$$\begin{aligned} x^{n,1} - X &\approx g'(X)(x^n - X), \\ x^{n,2} - X &\approx g'(X)(x^{n,1} - X). \end{aligned}$$

Из этих соотношений получаем

$$\begin{aligned} g'(X) &\approx \frac{x^{n,2} - x^{n,1}}{x^{n,1} - x^n}, \\ X &\approx \frac{x^{n,2} - g'(X)x^{n,1}}{1 - g'(X)} \approx \frac{x^{n,2} - \frac{x^{n,2} - x^{n,1}}{x^{n,1} - x^n} x^{n,1}}{1 - \frac{x^{n,2} - x^{n,1}}{x^{n,1} - x^n}} = \frac{x^{n,2}x^n - (x^{n,1})^2}{x^{n,2} - 2x^{n,1} + x^n}. \end{aligned}$$

За следующее после x^n приближение примем

$$x^{n+1} = \frac{x^{n,2}x^n - (x^{n,1})^2}{x^{n,2} - 2x^{n,1} + x^n} = \frac{x^n g(g(x^n)) - (g(x^n))^2}{g(g(x^n)) - 2g(x^n) + x^n}. \quad (10)$$

Для характеристики методов решения уравнений вводится понятие *порядка метода*. Говорят, что метод имеет k -й порядок, если существуют $c_1 > 0$, $c_2 < \infty$ такие, что

$$\rho(x^{n+1}, X) \leq c_2(\rho(x^n, X))^k$$

при условии $\rho(x^n, X) \leq c_1$. Очевидно, чем больше k , тем быстрее сходится процесс итераций при малых значениях $\rho(x^n, X)$.

В вычислительной практике наиболее распространены методы первого (например, метод простой итерации) и второго порядка (например, метод Ньютона, рассматриваемый в следующем параграфе).

Задача. Доказать, что при гладкой $g(x)$ метод итераций (10) имеет второй порядок.

Примечание. Иногда в литературе встречается другое определение порядка метода: говорят, что метод решения системы уравнений $F(x) = 0$ имеет порядок k , если при его реализации вычисляются производные функций f_i до порядка $k - 1$ включительно.

В § 12 гл. VI были рассмотрены итерационные методы решения линейных систем с помощью спектрально эквивалентных операторов. Аналогичные методы применяются и для решения нелинейных систем. Выбирается оператор $G(x)$ такой, что $x = 0$ является единственным решением уравнения $G(x) = 0$. Приближения x^{n+1} к решению системы $F(x) = 0$ определяются из соотношения

$$G(x^{n+1} - x^n) = F(x^n). \quad (11)$$

Наиболее распространен случай, когда G — линейный оператор. Тогда соотношение (11) приобретает вид

$$A(x^{n+1} - x^n) = F(x^n); \quad (12)$$

в случае конечной размерности m векторов x A является матрицей. В ряде случаев оператор G выбирается зависящим от n , а также от приближения x^n . Тогда в схему (12) укладывается также рассматриваемый ниже метод Ньютона решения нелинейных задач.

§ 2. Метод Ньютона решения нелинейных уравнений

Если известно достаточно хорошее начальное приближение к решению системы уравнений

$$F(x) = 0, \quad (1)$$

то эффективным методом повышения точности является метод Ньютона.

Не усложняя хода рассуждений, мы сможем рассмотреть более общий случай — решение нелинейного функционального уравнения.

Пусть $F(x)$ — оператор, отображающий линейное нормированное пространство $\mathcal{X}$ на линейное нормированное пространство $\mathcal{Y}$, может быть и совпадающее с $\mathcal{X}$. Нормы в этих пространствах, соответственно, обозначаем $\|\cdot\|_{\mathcal{X}}$ и $\|\cdot\|_{\mathcal{Y}}$. Линейный оператор P , действующий из пространства $\mathcal{X}$ в пространство $\mathcal{Y}$, назовем *производной оператора $F(x)$ в точке x* , если

$$\|F(x + \eta) - F(x) - P\eta\|_{\mathcal{Y}} = o(\|\eta\|_{\mathcal{X}}) \quad (2)$$

при $\|\eta\|_{\mathcal{X}} \rightarrow 0$.

В дальнейшем будем обозначать такой оператор P через $F'(x)$. Пусть, например,

$$x = (x_1, \dots, x_m)^T, \quad F = (f_1, \dots, f_m)^T.$$

Если функции f_i непрерывно дифференцируемы в окрестности данной точки x , то

$$\begin{aligned} f_i(x_1 + \eta_1, \dots, x_m + \eta_m) &= \\ &= f_i(x_1, \dots, x_m) + \sum_{j=1}^m \frac{\partial f_i(x_1, \dots, x_m)}{\partial x_j} \eta_j + o(\|\eta\|) \end{aligned}$$

Совокупность этих соотношений можно переписать в виде (2), если за P принять оператор умножения слева на матрицу

$$F'(x) = \left[\frac{\partial f_i}{\partial x_j} \right].$$

В простейшем случае $m = 1$ оператор P превращается в оператор умножения на производную f'_x .

Пусть X — решение уравнения $F(X) = 0$, x^n — некоторое приближение к X . В предположении существования производной F' , согласно (2), имеем

$$\|F(X) - F(x^n) - F'(x^n)(X - x^n)\|_y = o(\|X - x^n\|_x). \quad (3)$$

Если величина $\|X - x^n\|_x$ мала, то можно написать приближенное равенство

$$F(x^n) + F'(x^n)(X - x^n) \approx F(X).$$

Поскольку $F(X) = 0$, то

$$F(x^n) + F'(x^n)(X - x^n) \approx 0.$$

Возьмем в качестве следующего приближения решение уравнения

$$F(x^n) + F'(x^n)(x^{n+1} - x^n) = 0$$

(если такое решение существует). Между прочим, последнее уравнение имеет вид (1.12). В предположении, что оператор F' обратим, это решение можно записать в виде

$$x^{n+1} = x^n - (F'(x^n))^{-1} F(x^n). \quad (4)$$

Такой итерационный процесс называют *методом Ньютона*.

Пусть $\Omega_a = \{x: \|x - X\| < a\}$.

Пусть при некоторых $a, a_1, a_2, 0 < a, 0 \leq a_1, a_2 < \infty$ выполнены условия:

$$1. \quad \|(F'(x))^{-1}\| \leq a_1 \quad \text{при} \quad x \in \Omega_a. \quad (5)$$

$$2. \quad \|F(U_1) - F(U_2) - F'(U_2)(U_1 - U_2)\|_y \leq a_2 \|U_2 - U_1\|_x^2 \quad (6)$$

при $U_1, U_2 \in \Omega_a$. Обозначим $c = a_1 \cdot a_2, b = \min(a, c^{-1})$.

Теорема (о сходимости метода Ньютона). При условиях 1, 2 и $x^0 \in \Omega_b$ итерационный процесс Ньютона (4) сходится

с оценкой погрешности

$$\|x^n - X\|_{\mathcal{X}} \leq c^{-1} (c \|x^0 - X\|_{\mathcal{X}})^{2^n}. \quad (7)$$

Примечание. Если в рассматривавшемся выше примере функции f_i имеют ограниченные вторые производные, то, согласно формуле Тейлора, имеем

$$f_i(y) = f_i(x) + \sum_{j=1}^n \frac{\partial f_i(x_1, \dots, x_n)}{\partial x_j} (y_j - x_j) + O(\|y - x\|^2),$$

и, таким образом, условие 2 выполнено.

Доказательство. Пусть $x^0 \in \Omega_b$. Индукцией по n докажем, что все $x^n \in \Omega_b$. Пусть это утверждение доказано при некотором n ; так как $b \leq a$, то тогда $x^n \in \Omega_a$. Подставив в (6) $U_1 = X$ и $U_2 = x^n$, получим

$$\|F(X) - F(x^n) - F'(x^n)(X - x^n)\|_y \leq a_2 \|x^n - X\|_{\mathcal{X}}^2.$$

Поскольку $F(x^n) = -F'(x^n)(x^{n+1} - x^n)$, а $F(X) = 0$, то это соотношение может быть переписано в виде

$$\|F'(x^n)(x^{n+1} - X)\|_y \leq a_2 \|x^n - X\|_{\mathcal{X}}^2.$$

Воспользовавшись (5), получаем неравенство

$$\|x^{n+1} - X\|_{\mathcal{X}} \leq c \|x^n - X\|_{\mathcal{X}}^2. \quad (8)$$

Отсюда следует, что

$$\|x^{n+1} - X\|_{\mathcal{X}} < cb^2 = (cb)b \leq b,$$

и поэтому x^{n+1} также принадлежит Ω_b . Таким образом, при $x^0 \in \Omega_b$ все x^n принадлежат Ω_b и, следовательно, для них выполняется (8).

Пусть $q_n = c \|x^n - X\|_{\mathcal{X}}$. После умножения на c неравенство (8) запишется в виде $q_{n+1} \leq q_n^2$.

Индукцией по n докажем справедливость неравенства

$$q_n \leq q_0^{2^n}.$$

При $n=0$ оно очевидно. Предположив его верным при $n=k$, получаем

$$q_{k+1} \leq q_k^2 \leq (q_0^{2^k})^2 = q_0^{2^{k+1}}.$$

Таким образом, $q_n \leq q_0^{2^n}$ при всех n . Это означает, что

$$c \|x^n - X\|_{\mathcal{X}} \leq (c \|x^0 - X\|_{\mathcal{X}})^{2^n};$$

отсюда следует (7). Согласно определению c и b ,

$$c \|x^0 - X\|_{\mathcal{X}} < cb \leq 1,$$

и поэтому $x^n \rightarrow X$.

Теорема доказана.

Оценка скорости сходимости метода Ньютона (7) получена как следствие неравенства (8), выполняющегося для любого метода второго порядка. Таким образом, оценка (7) при некотором значении постоянной c выполнена для любого метода второго порядка.

Иногда вместо (6) выполняется более слабая оценка

$$\|F(U_1) - F(U_2) - F'(U_2)(U_1 - U_2)\|_{\mathcal{Y}} \leq a_3 \|U_2 - U_1\|_{\mathcal{X}}^{\rho},$$

где $1 < \rho < 2$. Тогда вместо (7) имеет место оценка

$$\|x^n - X\|_{\mathcal{X}} \leq c_1^{-1} (c_1 \|x^0 - X\|_{\mathcal{X}})^{\rho^n}.$$

Обращение оператора $F'(x^n)$ зачастую оказывается более трудоемкой операцией, чем вычисление значения $F(x^n)$. Поэтому метод Ньютона часто модифицируется следующим образом. По ходу вычислений выбирают или заранее задаются некоторой

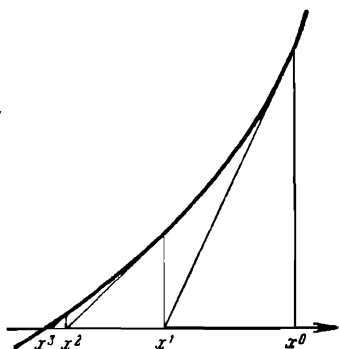


Рис. 7.2.1.

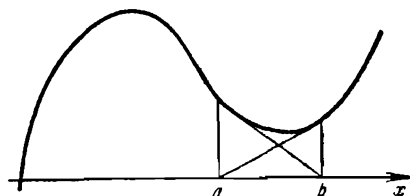


Рис. 7.2.2.

возрастающей последовательностью чисел $n_0=0, n_1, n_2, \dots$. При $n_k \leq n < n_{k+1}$ итерации производят по формуле

$$x^{n+1} = x^n - (F'(x^{n_k}))^{-1} F(x^n).$$

Увеличение числа итераций, сопровождающее часто такую модификацию, компенсируется большей «дешевизной» одного шага итерации. Выбор последовательности $\{n_k\}$ нужно производить с обоюдным учетом этих факторов.

Рассмотрим геометрическую интерпретацию метода Ньютона в случае решения скалярного уравнения $f(x)=0$. Тогда расчетная формула (4) приобретает вид

$$x^{n+1} = x^n - \frac{f(x^n)}{f'(x^n)}. \quad (9)$$

Для получения x^{n+1} геометрически надо найти абсциссу точки пересечения с осью x касательной к кривой $y = f(x)$ в точке $(x^n, f(x^n))$ (рис. 7.2.1). Уже в случае, когда $f(x)$ — многочлен третьей степени, может случиться, что последовательность $\{x^n\}$ не сходится к корню при плохом начальном приближении. Например, в случае, изображенном на рис. 7.2.2, все четные приближения совпадают с a , а нечетные — с b ; метод, как говорят, «зациклился». Для более сложных задач реальное поведение приближений x^n при плохом начальном приближении становится существенно более запутанным и трудно поддающимся анализу.

Сравним асимптотическую скорость сходимости метода Ньютона и метода простой итерации. Для последнего мы имели оценку погрешности

$$\|x^n - X\| \leq q^n \|x^0 - X\|, \quad q < 1.$$

Чтобы ошибка стала меньше ε , согласно этой оценке, достаточно взять

$$n \geq \log_{q^{-1}} \frac{\|x^0 - X\|}{\varepsilon} \sim \log_{q^{-1}} \frac{1}{\varepsilon}.$$

В случае метода Ньютона правая часть (7) будет меньше ε , если

$$n \geq -\log_2 \frac{\log_2(c \|x^0 - X\|)}{\log_2(c\varepsilon)} \sim \log_2 \log_2 \frac{1}{\varepsilon}. \quad (10)$$

Таким образом, асимптотически, при $\varepsilon \rightarrow 0$, метод Ньютона требует меньшего числа итераций.

Обратим внимание, что метод Ньютона, записанный в форме (4), сам является разновидностью метода простой итерации. В случае скалярного уравнения $f(x) = 0$ хорошо видна еще одна особенность метода Ньютона. Производная правой части (9) $g(x) = x - \frac{f(x)}{f'(x)}$ по x равна $\frac{f(x)f''(x)}{(f'(x))^2}$. Таким образом, $g'(X) = 0$, если $f'(X) \neq 0$, и рис. 7.1.1 в этом случае приобретает следующий вид (рис. 7.2.3).

Метод Ньютона оказывается удобным способом извлечения корней целой степени. Задача извлечения корня $\sqrt[p]{a}$, p — целое, равносильна задаче решения уравнения $x^p - a = 0$. Расчетная формула метода Ньютона в этом случае приобретает вид

$$x_{n+1} = \frac{p-1}{p} x_n + \frac{a}{px_n^{p-1}}.$$

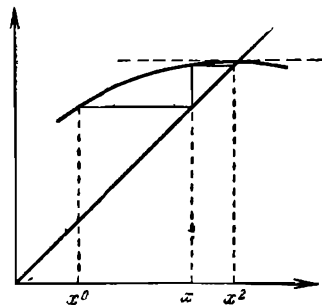


Рис. 7.2.3.

В качестве начального приближения к $\sqrt[n]{a}$ можно взять значение некоторого многочлена, приближающего эту функцию.

Задача. Рассматривается алгоритм вычисления $\sqrt[n]{a}$ при $1 \leq a \leq 4$, x_0 полагается равным значению многочлена наилучшего равномерного приближения для $\sqrt[n]{a}$ на $[1, 4]$: $x_0 = p_1(a) = \frac{17}{24} + \frac{a}{3}$. Убедиться в справедливости неравенства $|x_4 - \sqrt[n]{a}| \leq 0,5 \cdot 10^{-25}$.

§ 3. Другие методы решения одного уравнения

Если имеется хорошее приближение к корню X уравнения $f(x) = 0$, то, наряду с методом Ньютона, употребителю *метод секущих*.

Простейший вариант этого метода заключается в следующем. В процессе итераций фиксируется некоторая точка x^0 . Приближение x^{n+1} находится как абсцисса точки пересечения прямой, проходящей через точки $(x^0, f(x^0))$ и $(x^n, f(x^n))$ с осью x (рис. 7.3.1).

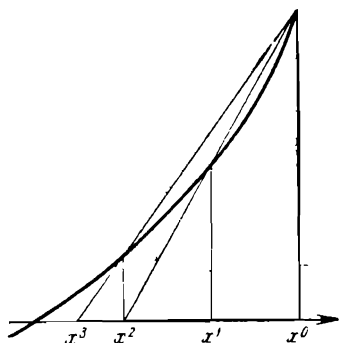


Рис. 7.3.1.

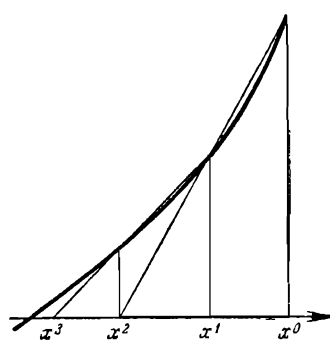


Рис. 7.3.2.

Обычно более эффективен способ, где за x^{n+1} принимается абсцисса точки пересечения с осью x прямой, проходящей через точки $(x^{n-1}, f(x^{n-1}))$ и $(x^n, f(x^n))$ (рис. 7.3.2). Уравнение этой прямой

$$y_n(x) = f(x^n) + (x - x^n) \frac{f(x^n) - f(x^{n-1})}{x^n - x^{n-1}}.$$

Из условия $y_n(x^{n+1}) = 0$ получаем

$$x^{n+1} = x^n - \frac{f(x^n)(x^n - x^{n-1})}{f(x^n) - f(x^{n-1})}. \quad (1)$$

Для оценки требуемого числа итераций установим связь между погрешностями $r_n = x^n - X$ последующих приближений. Справедливо равенство

(разложение Тейлора):

$$0 = f(X) = f(x^n) + f'(x^n)(X - x^n) + \frac{f''(x^n)}{2}(X - x^n)^2 + O((X - x^n)^3).$$

Отсюда

$$f(x^n) = f'(x^n)r_n - \frac{f''(x^n)}{2}r_n^2 + O(r_n^3).$$

Точно так же

$$f(x^{n-1}) - f(x^n) = f'(x^n)(x^{n-1} - x^n) + \frac{f''(x^n)}{2}(x^{n-1} - x^n)^2 + O((x^{n-1} - x^n)^3).$$

Отсюда

$$f(x^n) - f(x^{n-1}) = f'(x^n)(r_n - r_{n-1}) - \frac{f''(x^n)}{2}(r_n - r_{n-1})^2 + O((r_n - r_{n-1})^3).$$

Вычитая X из (1), получим

$$r_{n+1} = r_n - \frac{\left(f'(x^n)r_n - \frac{f''(x^n)}{2}r_n^2 + O(r_n^3)\right)(r_n - r_{n-1})}{f'(x^n)(r_n - r_{n-1}) - \frac{f''(x^n)}{2}(r_n - r_{n-1})^2 + O((r_n - r_{n-1})^3)}.$$

В предположении $f'(x) \neq 0$ разделим числитель и знаменатель на $f'(x^n) \times (r_n - r_{n-1})$; имеем

$$r_{n+1} = r_n - \frac{r_n - \lambda r_n^2 + O(r_n^3)}{1 - \lambda(r_n - r_{n-1}) + O((r_n - r_{n-1})^2)},$$

где

$$\lambda = \frac{f''(x^n)}{2f'(x^n)}.$$

На основании равенства

$$\frac{1}{1 - \varepsilon} = 1 + \varepsilon + O(\varepsilon^2)$$

получаем отсюда, что

$$r_{n+1} = r_n - (r_n - \lambda r_n^2 + O(r_n^3))(1 + \lambda(r_n - r_{n-1}) + O((r_n - r_{n-1})^2)).$$

Иначе это соотношение можно представить в виде

$$r_{n+1} = \lambda r_n r_{n-1} + O((|r_n| + |r_{n-1}|)^3). \quad (2)$$

Мы не будем производить строгих оценок скорости сходимости итерационного процесса, а ограничимся лишь правдоподобными рассуждениями. Глядя на соотношение (2), естественно предположить, что в ходе итераций погрешности r_n довольно быстро убывают:

$$r_{n-1} \gg r_n \gg r_{n+1}; \quad (3)$$

в то же время нет оснований рассчитывать, что величина r_{n+1} будет (при $n \rightarrow \infty$) по порядку меньше, чем произведение $r_n r_{n-1}$. Поскольку $r_{n-1} \gg r_n$, то $r_{n-1} r_n \gg r_n^2$, отсюда и из (2) следует, что $r_{n+1} \gg r_n^2$. Уменьшая значение n на 1, получаем $r_n \gg r_{n-1}^2$; следовательно, $r_{n-1}^3 \ll r_n r_{n-1}$; отсюда и из (3) вытекает соотношение $r_n^3 \ll r_n r_{n-1}$. Поэтому при больших n (2) можно записать в виде

$$r_{n+1} \sim \lambda r_n r_{n-1}.$$

После замены $|\lambda r_n| = R_n$ получим

$$R_{n+1} \sim R_n R_{n-1}.$$

Логарифмируя это соотношение, имеем

$$\ln R_{n+1} \sim \ln R_n + \ln R_{n-1}.$$

Есть основания предположить, что при больших n величина $\ln R_n$ ведет себя как решение конечно-разностного уравнения

$$u_{n+1} = u_n + u_{n-1};$$

соответствующее характеристическое уравнение

$$\mu^2 - \mu - 1 = 0$$

имеет корни

$$\mu_1 = \frac{1 + \sqrt{5}}{2}, \quad \mu_2 = \frac{1 - \sqrt{5}}{2}, \quad \text{причем } |\mu_2| < 1.$$

Таким образом $u_n = c_1 \mu_1^n + c_2 \mu_2^n \sim c_1 \mu_1^n$ при $n \rightarrow \infty$. В предположении $\ln R_n \sim u_n$ можно написать асимптотическое равенство

$$\log_2 |\ln R_n| \sim \log_2 |c_1 \mu_1^n| \sim n \log_2 \mu_1.$$

Неравенства $|r_n| \leq \varepsilon$, $R_n \leq |\lambda| \varepsilon$ и $|\ln R_n| \geq |\ln(|\lambda| \varepsilon)|$ эквивалентны при $|\lambda| \varepsilon \leq 1$. Поэтому для наименьшего n , при котором достигается требуемая точность ε , имеем $\log_2 |\ln(|\lambda| \varepsilon)| \sim n \log_2 \mu_1$; отсюда получаем

$$n \sim \frac{1}{\log_2 \mu_1} \log_2 \log_2 \frac{1}{\varepsilon}.$$

Сравнивая это соотношение с (2.10), видим, что для рассматриваемого нами метода секущих требуемое число итераций при малых ε примерно в $\log_{\mu_1} 2 \approx 1,44$ раза больше, чем для метода Ньютона. Если d_1 — объем работы при вычислении значения функции в точке, а d_2 — дополнительный объем работы при вычислении производной в той же точке, то в рассматриваемом случае суммарные затраты работы будут порядка $1,44 d_1 \log_2 \log_2 \frac{1}{\varepsilon}$,

а в случае метода Ньютона — порядка $(d_1 + d_2) \log_2 \log_2 \frac{1}{\varepsilon}$.

Рассмотрим типичную задачу: имеется конкретная функция $f(x, \alpha_1, \dots, \alpha_n)$; требуется составить программу, которая по задаваемым ей значениям $\alpha_1, \dots, \alpha_n$ выдает корень уравнения $f(x, \alpha_1, \dots, \alpha_n) = 0$. Часто программы вычисления функций f и f_x имеют много общих блоков и в результате этого выполняется неравенство $d_2 < 0,44 d_1$. Тогда в соответствии с приведенными выше оценками при наличии хорошего начального приближения к корню целесообразнее использовать метод Ньютона.

При решении системы уравнений

$$F(x) = 0$$

в качестве аналога метода секущих можно рассматривать такой метод. Пусть определены приближения $x^{n-m}, \dots, x^n$ и известны значения

$$f_i(x^{n-m}), \dots, f_i(x^n).$$

Пусть $y = L_i(x)$ — уравнение плоскости, проходящей через точки

$$(x^{n-m}, f_i(x^{n-m})), \dots, (x^n, f_i(x^n));$$

за следующее приближение x^{n+1} принимаем решение системы уравнений

$$L_i(x) = 0, \quad i = 1, \dots, m.$$

При $m = 1$ этот метод действительно превращается в рассматривавшийся выше метод секущих. Для $m > 1$ этот метод применяется редко, обычно в случае, когда можно ограничиться невысокой точностью.

Дело в том, что для этого метода при $n \rightarrow \infty$ характерно «сплющивание» m -мерного симплекса с вершинами в точках $x^{n-m}, \dots, x^n$; следствием этого является быстрое ухудшение обусловленности системы уравнений $L_i(x) = 0$; в результате алгоритм вычислений становится неустойчивым к вычислительной погрешности, а также теряет сходимость.

Кроме описанных выше, существует большое число других методов подобного типа, где в окрестности корня функция $f(x)$ приближается некоторой функцией $g(x)$, для которой уравнение $g(x) = 0$ решается в явном виде. Однако для применения всех этих методов необходимо достаточно хорошее приближение к решению.

Для получения такого приближения иногда используется метод вилки. Сначала определяют a^0 и b^0 такие, что $f(x)$ непрерывна на $[a^0, b^0]$ и $f(a^0)f(b^0) < 0$. Например, в упомянутом выше случае уравнения $f(x, \alpha_1, \dots, \alpha_n) = 0$ часто удается получить явное выражение $a_0(\alpha_1, \dots, \alpha_n)$, $b_0(\alpha_1, \dots, \alpha_n)$ таких a_0, b_0 ; в других случаях их иногда находят, составляя некоторую таблицу значений $f(x)$ в случайных или упорядоченных точках. После нахождения такого отрезка $[a^0, b^0]$ выбирают некоторую точку $c^0 \in (a^0, b^0)$. После сравнения знаков величин $f(a^0)$, $f(c^0)$ и $f(b^0)$ принимают за $[a^1, b^1]$ тот из отрезков $[a^0, c^0]$, $[c^0, b^0]$, на концах которого $f(x)$ принимает противоположные знаки. Далее выбирается точка $c^1 \in (a^1, b^1)$ и т. д. В результате получается последовательность вложенных друг в друга отрезков $[a^0, b^0] \supset [a^1, b^1] \supset \dots$ таких, что $f(a^n)f(b^n) < 0$.

Например, можно представить такую линию поведения при выборе c^n . На первых шагах берется $c^n = (a^n + b^n)/2$, после вычисления $f(c^n)$ решается вопрос, можно ли считать точки

$$(a^n, f(a^n)), (c^n, f(c^n)), (b^n, f(b^n)) \quad (4)$$

лежащими на одной прямой.

Если указанные точки лежат на одной прямой, то

$$f(a^n) - 2f(c^n) + f(b^n) = 0;$$

поэтому условием принадлежности этих точек одной прямой считается выполнение неравенства

$$|f(a^n) - 2f(c^n) + f(b^n)| < \gamma |f(b^n) - f(a^n)|;$$

γ — подбирается экспериментально.

В случае положительного ответа, начиная с этого момента, в качестве c^{n+1} берут точку пересечения секущей, проходящей через точки

$$(a^{n+1}, f(a^{n+1})), (b^{n+1}, f(b^{n+1})),$$

с осью x .

В случае отрицательного ответа опять полагают

$$c^{n+1} = (a^{n+1} + b^{n+1})/2$$

и рассматривают вопрос о принадлежности точек

$$(a^{n+1}, f(a^{n+1})), (c^{n+1}, f(c^{n+1})), (b^{n+1}, f(b^{n+1}))$$

одной прямой и т. д.

В некоторых случаях целесообразно всегда принимать за c^{n+1} абсциссу точки интервала (a_n, b_n) , в которой парабола, проходящая через точки (4), пересекается с осью x .

Важной задачей является разработка эффективных методов решения уравнений отдельных типичных классов. Для нахождения корней многочлена $P(z) = a_0 z^m + \dots + a_m$ как с действительными, так и с комплексными коэффициентами таким методом является *метод парабол*. При заданных приближениях к корню z_{n-2}, z_{n-1}, z_n следующее приближение определяется следующим образом. Строится интерполяционный многочлен второй степени, совпадающий с $P(z)$ в точках z_{n-2}, z_{n-1}, z_n . За z_{n+1} принимается корень этого многочлена, наиболее близкий к z_n . В стандартных программах метода парабол эта схема подвергнута некоторой модификации, вызванной анализом результатов применения метода. Хотя сходимость этого метода для произвольного многочлена при произвольных начальных условиях не доказана, не зарегистрировано ни одного случая, когда этот метод не сходится или сходится медленно.

Если ищется несколько корней многочлена, то производится так называемое *выделение корней*: после нахождения очередного корня $\bar{z}$ от нахождения корней многочлена $P(z)$ переходят к нахождению корней многочлена $\frac{P(z)}{z - \bar{z}}$.

§ 4. Методы спуска

Для решения задачи минимизации функционала наиболее часто применяются *методы спуска*. При заданном приближении определяется какое-либо направление, в котором функционал убывает, и производится перемещение приближения в этом направлении. Если величина перемещения взята не очень большой, то значение функционала обязательно уменьшится.

Рассмотрим примеры методов спуска.

При исследовании сходимости метода Зейделя в случае системы уравнений $Ax = b$ при $A > 0$ мы описали циклический *метод покоординатного спуска* минимизации функции $\Phi(x_1, \dots, x_m)$. При заданном приближении $\Phi(x_1^0, \dots, x_m^0)$ отыскивается значение $x_1 = x_1^1$, при котором достигается $\inf_{x_1} \Phi(x_1, x_2^0, \dots, x_m^0)$, затем отыскивается значение $x_2 = x_2^1$, при котором достигается $\inf_{x_2} \Phi(x_1^1, x_2, x_3^0, \dots, x_m^0)$ и т. д.; процесс циклически повторяется.

Обозначим через $P_k x$ приближение, получаемое при спуске из x по координате x_k . Присваивая приближению, получающе-

муся при спуске по очередной координате, следующий номер, можно записать приближения метода циклического покоординатного спуска в виде

$$\begin{aligned}x^1 &= P_1 x, & x^2 &= P_2 P_1 x, \dots, \\x^m &= P_m \dots P_1 x, & x^{m+1} &= P_1 P_m \dots P_1 x, \dots\end{aligned}$$

При практической реализации этого метода возникает проблема минимизации функций одной переменной. Рассмотрим отдельно задачу минимизации функции одной переменной $\Phi(x)$ при начальном приближении к точке минимума $x = x^0$. Так как эта задача обычно не может быть решена точно, то часто поступают следующим образом: берут некоторые значения $\bar{x}^0, \bar{\bar{x}}^0$ и строят параболу $y = P_2(x)$, удовлетворяющую условиям

$$P_2(x^0) = \Phi(x^0), \quad P_2(\bar{x}^0) = \Phi(\bar{x}^0), \quad P_2(\bar{\bar{x}}^0) = \Phi(\bar{\bar{x}}^0).$$

Абсциссу x точки минимума $P_2(x)$ принимают за следующее приближение x^1 . Последовательность точек, получаемых по описываемому методу, не обязательно сходится к искомой точке экстремума функции $\Phi(x)$.

Чтобы уяснить возможность подобных ситуаций, поучительно рассмотреть некоторые примеры. Пусть x^n — приближения, получаемые по этому алгоритму. Возможны различные способы выбора точек $\bar{x}^n, \bar{\bar{x}}^n$ на каждом шаге итераций.

Можно выбирать эти точки так, чтобы было

$$a) \quad |\bar{x}^n - x^n| + |\bar{\bar{x}}^n - x^n| \rightarrow 0 \quad \text{при} \quad n \rightarrow \infty.$$

Тогда возможна следующая ситуация (рис. 7.4.1). Точки $x^0, \bar{x}^0, \bar{\bar{x}}^0$ находятся в малой окрестности точки локального экстремума x^* , все последующие точки

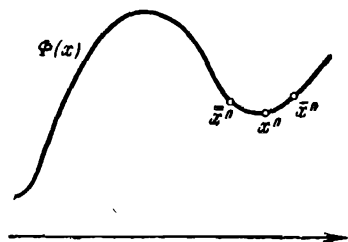


Рис. 7.4.1.

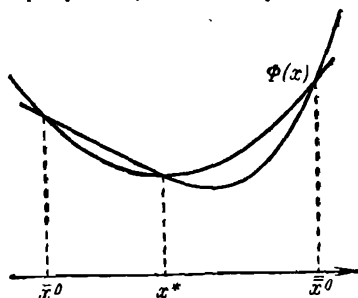


Рис. 7.4.2.

$x^n, \bar{x}^n, \bar{\bar{x}}^n$ тоже в этой окрестности и $x^n \rightarrow x^*$; мы получаем точку локального экстремума x^* вместо интересующей нас точки абсолютного экстремума X .

б) $|\bar{x}^n - x^n| + |\bar{\bar{x}}^n - x^n| \not\rightarrow 0$; нетрудно подобрать функцию $\Phi(x)$, удовлетворяющую следующим условиям (рис. 7.4.2):

1) существуют точки $x^*, \bar{x}^0, \bar{\bar{x}}^0$ такие, что параболa, совпадающая с $\Phi(x)$ в этих точках, имеет экстремум в точке x^* ;

2) если $\bar{x}^n = \bar{x}^0$, $\bar{x}^n = \bar{x}^0$, $|x^n - x^*| \leq \eta$, $\eta > 0$, то $|x^{n+1} - x^*| \leq q|x^n - x^*|$, где $q < 1$.

Тогда при начальном приближении x^0 таким, что $|x^0 - x^*| \leq \eta$, мы будем иметь полную иллюзию сходимости итерационного процесса к точке x^* , не являющейся даже точкой локального экстремума.

Задача. Получить явную формулу для такой функции $\Phi(x)$.

Даже если на каждом шаге отыскивается абсолютный экстремум функции $\Phi(x_1, \dots, x_m)$ по соответствующей координате, уже при $m=2$ можно построить пример, когда итерационный процесс сходится при некотором начальном приближении не к искомой точке абсолютного экстремума, а к некоторой точке локального экстремума. На рис. 7.4.3 изображены линии уровня такой функции и получаемые приближения.

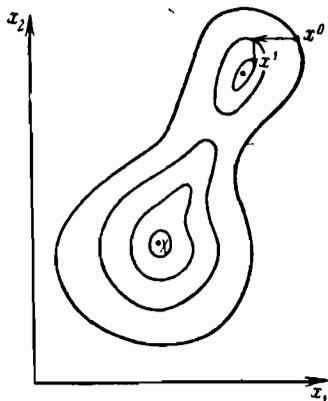


Рис. 7.4.3.

Спуск в циклическом порядке необязателен. Если из рассмотрения проводившихся ранее вычислений видно, что спуск по каким-либо координатам обеспечивает наибольшее убывание $\Phi(x)$, то иногда целесообразен более частый спуск по этим координатам.

В других случаях при каждом n после получения приближения x^n выбирается некоторая совокупность координат

$$x_{l(n,1)}, \dots, x_{l(n,q(n))};$$

производятся независимые спуски по координатам этой группы исходя из приближения x^n , т. е. находят точки $P_{l(n,k)}x^n$. Далее вычисляется

$$\min_{1 \leq k \leq q(n)} \Phi(P_{l(n,k)}x^n)$$

и соответствующая минимуму точка $P_{l(n,k_1)}x^n$ принимается за x^{n+1} .

Описанные методы называют *методами оптимального по координатного спуска*.

Иногда номер очередной координаты, по которой осуществляется спуск, выбирается недетерминированно. В этом случае говорят о *случайном по координатном спуске*.

Другой вариант метода спуска — *метод наискорейшего (градиентного) спуска*. Следующее приближение отыскивается в виде

$$x^{n+1} = x^n - \Delta_n \text{grad } \Phi(x^n)$$

(рис. 7.4.4). Значение Δ_n определяется из условия

$$\min_{\Delta_n} \Phi(x^n - \Delta_n \text{grad } \Phi(x^n)),$$

т. е. этот алгоритм опять состоит в последовательной минимизации функций одной переменной Δ_n .

Обратим внимание на то, что как в методе покоординатного спуска, так и в методе наискорейшего спуска нет необходимости полного решения задач минимизации. Иногда разумнее, достигнув лишь некоторого смещения в сторону минимума вспомогательной функции одной переменной, перейти к следующему шагу итерационного процесса. В окрестности точки своего минимума вспомогательная функция одной переменной меняется мало и слишком тщательное нахождение ее точки минимума не приводит к существенному уменьшению рассматриваемых значений функции.

В случае метода наискорейшего спуска вопрос об объеме вычислений при минимизации вспомогательных функций одной переменной должен решаться также с учетом относительной трудоемкости вычисления значений функции $\Phi(x)$ и ее градиента. Если вычисление значения функции существенно проще вычисления градиента, оправдано более точное нахождение минимума на каждом шаге итерации. В противоположном случае разумнее не затрачивать слишком много усилий на нахождение этих минимумов.

Для иллюстрации решения вопроса о выборе метода рассмотрим следующую типичную задачу: решается нелинейная краевая задача для системы обыкновенных дифференциальных уравнений. В главе IX будет показано, что эта задача сводится к решению нелинейной системы уравнений $F(x) = 0$:

$$f_i(x) = f_i(x_1, \dots, x_m) = 0, \quad i = 1, \dots, m,$$

обладающей следующими свойствами.

Объем вычислений при отыскании одного значения $f_i(x)$ и при одновременном отыскании всех значений $f_i(x)$, $i = 1, \dots, m$, в той же точке одинаков; обозначим его через A . Объем вычислений при непосредственном отыскании значения $\frac{\partial f_i(x)}{\partial x_j}$ и при одновременном отыскании всех значений $\frac{\partial f_i(x)}{\partial x_j}$, $i = 1, \dots, m$,

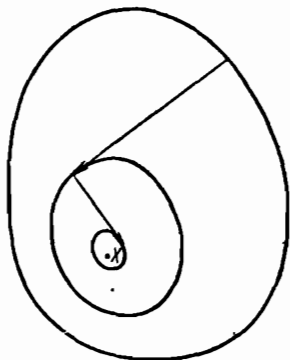


Рис. 7.4.4.

в той же точке одинаков; обозначим его через B . Обычно оказывается, что эти затраты A и B очень велики и $B \gg A$.

Поэтому целесообразнее вычислять производные

$$\frac{\partial f_i(x)}{\partial x_j}, \quad j = 1, \dots, m,$$

пользуясь приближенной формулой

$$\begin{aligned} \frac{\partial f_i(x_1, \dots, x_m)}{\partial x_j} &\approx \\ &\approx \frac{f_i(x_1, \dots, x_{j-1}, x_j + \Delta, x_{j+1}, \dots, x_m) - f_i(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_m)}{\Delta}. \end{aligned} \quad (1)$$

Тогда затраты по вычислению этих производных равны $2A$, т. е. затратам по вычислению двух совокупностей значений

$$\begin{aligned} &f_i(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_m) \\ &f_i(x_1, \dots, x_{j-1}, x_j + \Delta, x_{j+1}, \dots, x_m) \end{aligned}$$

(мы пренебрегаем вычислениями при вычитании этих величин и делении на Δ).

Если имеется хорошее приближение к решению задачи, то иногда неплохо воспользоваться методом Ньютона, имеющим быструю сходимость.

Область сходимости метода Ньютона обычно невелика, и поэтому, по крайней мере на начальном этапе итераций, может быть, целесообразно свести решение этой задачи к минимизации некоторого функционала и применить какой-либо из методов спуска.

Рассмотрим простейший случай функционала

$$\Phi(x) = \sum_{i=1}^m (\lambda_i f_i(x))^2;$$

множители $\lambda_i = \text{const} \neq 0$, называемые *масштабными*, обычно можно удачно подобрать из условий конкретной задачи.

Пусть $x^n = (x_1^n, \dots, x_m^n)$ — полученное приближение и решено сделать спуск по координате x_1 . Тогда вычисляются величины $f_i(x^n)$, $i = 1, \dots, m$ (если они ранее еще не были вычислены), и величины

$$\bar{f}_i(x^n, \Delta), \quad i = 1, \dots, m,$$

где

$$x^n, \Delta = (x_1^n + \Delta, x_2^n, \dots, x_m^n);$$

далее при помощи формулы (1) вычисляются приближения l_{i1} к производным $\frac{\partial f_i(x)}{\partial x_1}$.

На прямой, проходящей через x^n параллельно оси x_1 , в окрестности точки x^n имеем

$$f_i(x) \approx f_i(x^n) + l_{i1}(x_1 - x_1^n);$$

поэтому координату x_1 следующего приближения

$$x^{n+1} = (x_1^{n+1}, x_2^n, \dots, x_m^n)$$

определяют из условия

$$\min_{x_1} \sum_{i=1}^m (\lambda_i (f_i(x^n) + l_{i1}(x_1 - x_1^n)))^2.$$

В реальном алгоритме следует учесть ряд дополнительных обстоятельств, например, разумный выбор Δ : при больших Δ велика погрешность аппроксимации в формуле (1), в то же время погрешности в значениях f_i делятся на Δ , и поэтому при малых Δ погрешности формулы (1) также велика.

Предположим, что решено использовать метод градиентного спуска.

Вычисление $\text{grad } \Phi(x^n)$ требует знания всех производных $\frac{\partial f_i(x^n)}{\partial x_j}$; в случае их вычисления с помощью формул (1) объем вычислений равен $(m+1)A$; при больших m трудоемкость $(m+1)A$ вычисления $\text{grad } \Phi(x)$ много больше трудоемкости A вычисления $\Phi(x)$, и поэтому можно рекомендовать более тщательный выбор точки минимума в методе наискорейшего спуска.

§ 5. Другие методы сведения многомерных задач к задачам меньшей размерности

Иногда бывает полезно рассмотреть следующую формальную процедуру сведения многомерных задач к одномерным.

Пусть ищется минимум Φ^* функций

$$\Phi(x_1, \dots, x_m)$$

в области

$$\begin{aligned} \varphi_i(x_1, \dots, x_m) &\geq 0, & i = 1, \dots, l, \\ \psi_j(x_1, \dots, x_m) &= 0, & j = 1, \dots, q. \end{aligned} \quad (1)$$

Можно написать равенство

$$\Phi^* = \min_{x_1, \dots, x_m} \Phi(x_1, \dots, x_m) = \min_{x_m} \Phi_m(x_m),$$

где

$$\Phi_m(x_m) = \min_{x_1, \dots, x_{m-1}} \Phi(x_1, \dots, x_m);$$

минимум каждый раз берется по области определения минимизируемой функции в соответствии с условиями (1). Таким образом, исходная задача минимизации функции m переменных свелась к минимизации функции одной переменной, каждое значение которой определяется минимизацией функции $(m-1)$ -й переменной. В свою очередь минимизацию функции $\Phi_m(x_m)$ сведем к минимизации функции одной переменной, каждое значение которой определяется минимизацией функции

от $(m-2)$ -х переменных, и т. д. Получим цепочку соотношений

$$\Phi^* = \min_{x_m} \Phi_m(x_m),$$

$$\Phi_m(x_m) = \min_{x_{m-1}} \Phi_{m-1}(x_{m-1}, x_m),$$

$$\dots \dots \dots \Phi_3(x_3, \dots, x_m) = \min_{x_2} \Phi_2(x_2, \dots, x_m),$$

$$\Phi_2(x_2, \dots, x_m) = \min_{x_1} \Phi(x_1, \dots, x_m).$$

Кажущейся простоте метода сопутствует его большая трудоемкость. Предположим, что каждая минимизация функций одной переменной потребует вычисления s значений минимизируемой функции. Тогда минимизация $\Phi_m(x_m)$ требует минимизации $\min_{x_{m-1}} \Phi_{m-1}(x_{m-1}, x_m)$ при s значениях параметра x_m ,

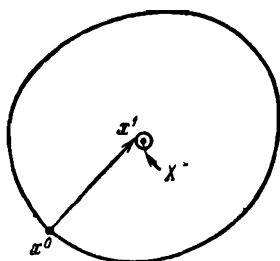


Рис. 7.5.1.

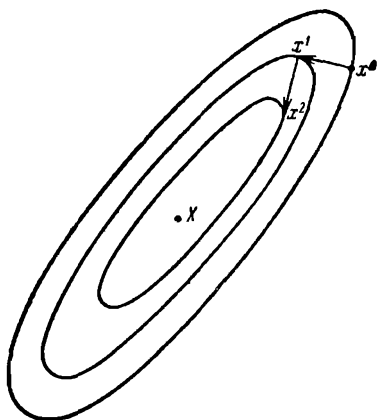


Рис. 7.5.2.

т. е. s^2 вычислений значений функции $\Phi_{m-1}(x_{m-1}, x_m)$. Это в свою очередь потребует вычисления s^3 значений

$$\Phi_{m-2}(x_{m-2}, x_{m-1}, x_m)$$

и т. д. В конечном счете потребуются вычисление s^m значений функции Φ . Уже при умеренных s и m , например $s=10$, $m=10$, такой объем вычислений окажется недопустимо большим. Однако при малых m некоторые модификации рассматриваемой идеи оказываются полезными. Например, возможен такой вариант. Задаются начальным приближением $(x_1^0, \dots, x_m^0)$. Реализуют указанный алгоритм при не очень большом значении s , например $s=3$ или $s=4$. При этом значения функции вычисляются в точках некоторого куба

$$|x_1 - x_1^0|, \dots, |x_m - x_m^0| \leq \Delta_0.$$

Получаемую точку минимума $(x_1^1, \dots, x_m^1)$ принимают за следующее приближение; приближение $(x_1^2, \dots, x_m^2)$ находят аналогичным образом, но значения функции Φ вычисляются в точках куба $|x_1 - x_1^1|, \dots, |x_m - x_m^1| \leq \Delta_1$ и т. д., где $\Delta_0 > \Delta_1 > \Delta_2 > \dots$

Рассмотрим еще один метод, общая структура которого похожа на структуру описанного выше.

Если линии уровня функции Φ похожи на сферы, то при применении методов спуска происходит быстрое смещение в направлении минимума (рис. 7.5.1). Однако практически более типичен случай, когда эти линии похожи на эллипсоиды с большим разбросом полуосей. Тогда при движении по градиенту смещение в направлении точки минимума будет довольно медленным (рис. 7.5.2).

Предположим, что оси этих «эллипсоидов» естественным образом разбиваются на две группы: первая группа состоит из m_1 осей одного порядка и относительно малых, вторая группа состоит из m_2 осей одного порядка и относительно больших.

Для некоторых задач такого рода хорошо зарекомендовал себя следующий метод, называемый *методом оврагов*. Задаются какими-то приближениями x^0 и x^1 и производят несколько шагов метода спуска, исходя из каждого из этих приближений. Будут получены приближения X^0 и X^1 . Из рис. 7.5.3 видно, что эти приближения будут лежать в многообразии, расположенном в окрестности осей второй группы. Процесс итераций состоит в получении последовательных приближений $X^0, X^1, \dots$, лежащих в окрестности этого многообразия. В случае $m_2 = 1$ приближение X^{l+1} отыскивается следующим образом. Проведем через X^{l-1} и X^l прямую и найдем приближение x^{l+1} к точке минимума $\Phi(x)$ на этой прямой. Таким образом, это приближение ищется в виде

$$x^{l+1} = X^l + \alpha(X^l - X^{l-1}).$$

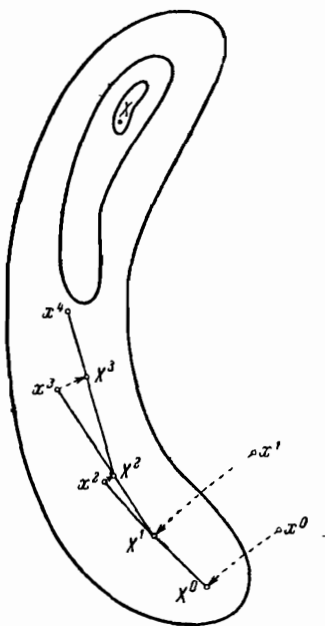


Рис. 7.5.3.

Далее проводим несколько итераций, исходя из x^{l+1} , и получаем приближение X^{l+1} , также лежащее в овраге. В случае $m_2 > 1$

приближение x^{l+1} иногда удобно отыскивать в виде

$$x^{l+1} = X^l + \alpha(X^l - X^{l-1}) + \beta \operatorname{grad} \Phi(X^l).$$

Из рис. 7.5.4 видно, что описанный способ оказывается эффективным и в ряде случаев, когда линии уровня функции Φ имеют более сложную структуру.

Решение системы уравнений

$$f_i(x_1, \dots, x_m) = 0, \quad i = 1, \dots, m, \quad (2)$$

также формально сводится к последовательному решению уравнений с одним неизвестным. Рассмотрим систему уравнений

$$f_i(x_1, \dots, x_m) = 0, \quad i = 2, \dots, m, \quad (3)$$

относительно неизвестных $x_2, \dots, x_m$. Пусть $x_2(x_1), \dots, x_m(x_1)$ — ее решение. Подставляя выражения $x_2(x_1), \dots, x_m(x_1)$ в первое из уравнений (2), получим уравнение

$$F_1(x_1) = f_1(x_1, x_2(x_1), \dots, x_m(x_1)) = 0 \quad (4)$$

относительно одной неизвестной x_1 .

Отыскание значений $x_2(x_1), \dots, x_m(x_1)$ и решение уравнения (4) можно проводить численно; выбирается какой-то метод решения (4) по значениям функции $F_1(x_1)$. При каждом требуемом значении x_1 в результате решения (3) получаются значения $x_2(x_1), \dots, x_m(x_1)$, которые подставляются затем в правую часть (4). Для решения системы уравнений (3) при каждом значении x_1 применим тот же прием.

Пусть $x_3(x_1, x_2), \dots, x_m(x_1, x_2)$ — решение системы уравнений

$$f_i(x_1, \dots, x_m) = 0, \quad i = 3, \dots, m, \quad (5)$$

относительно неизвестных $x_3, \dots, x_m$. Подставляя $x_3(x_1, x_2), \dots, x_m(x_1, x_2)$ во второе уравнение системы, получим уравнение

$$F_2(x_1, x_2) = f_2(x_1, x_2, x_3(x_1, x_2), \dots, x_m(x_1, x_2)) = 0.$$

При каждом x_1 это уравнение может быть разрешено относительно x_2 . Его решение $x_2(x_1)$, а также $x_3(x_1, x_2(x_1)), \dots, x_m(x_1, x_2(x_1))$ образуют решение системы (3). Систему (5)

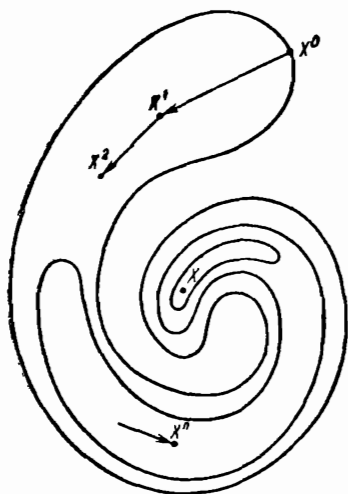


Рис. 7.5.4.

при каждом x_1, x_2 опять решаем, сводя к системе, где число неизвестных на единицу меньше и т. д.

Если для нахождения каждого вспомогательного уравнения с одной неизвестной потребуется s вычислений функций, то суммарно этот алгоритм потребует $s + \dots + s^m$ вычислений правых частей уравнений системы.

Так же как описанный в начале параграфа метод минимизации, этот алгоритм применяют относительно редко, обычно в случае небольших m при сложной структуре взаимного расположения поверхностей $f_i(x_1, \dots, x_m) = 0$.

Задача. Рассмотреть, во что переходит описанный метод в случае, когда система уравнений (2) линейная.

§ 6. Решение стационарных задач путем установления

Распространенным методом решения задач рассматриваемого нами типа является *метод* их решения путем *установления*. Строится некоторый нестационарный процесс, решение которого с течением времени устанавливается к решению исходной задачи. Рассмотрим систему дифференциальных уравнений

$$\frac{dx}{dt} + \text{grad } \Phi = 0. \quad (1)$$

Вектор $\frac{dx}{dt}$ пропорционален градиенту функции Φ , т. е. ортогонален ее линиям уровня и направлен в сторону убывания значений функции Φ . Таким образом, при перемещении вдоль траектории системы (1) значение Φ не возрастает. Формально справедливость этого утверждения следует из неравенства

$$\frac{d\Phi(x)}{dt} = \left(\text{grad } \Phi, \frac{dx}{dt} \right) = -(\text{grad } \Phi, \text{grad } \Phi), \quad (2)$$

означающего, что $\frac{d\Phi(x)}{dt} < 0$ всюду, за исключением стационарных точек функции $\Phi(x)$.

Другой нестационарный процесс, решение которого при весьма общих предположениях устанавливается к точке минимума функции $F(x)$, описывается системой дифференциальных уравнений

$$\frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} + \text{grad } \Phi = 0, \quad \gamma > 0. \quad (3)$$

Для решения этой системы имеем

$$\begin{aligned} \frac{d}{dt} \left(\frac{1}{2} \left(\frac{dx}{dt}, \frac{dx}{dt} \right) + \Phi(x) \right) &= \left(\frac{dx}{dt}, \frac{d^2x}{dt^2} \right) + \left(\text{grad } \Phi, \frac{dx}{dt} \right) = \\ &= -\gamma \left(\frac{dx}{dt}, \frac{dx}{dt} \right) < 0, \end{aligned} \quad (4)$$

если только $\frac{dx}{dt} \neq 0$. Функцию $\Phi(x)$ в первом случае и $\frac{1}{2} \left(\frac{dx}{dt}, \frac{dx}{dt} \right) + \Phi(x)$ — во втором можно рассматривать как энергию материальной системы, движение которой описывается системами уравнений (1) и (3). Соотношения (2) и (4) показывают, что рассматриваемые нестационарные процессы характеризуются оттоком, или, как говорят, *диссипацией* энергии.

Чтобы прояснить вопрос о разумном выборе γ , рассмотрим простейшую модель: x — скаляр, $\Phi(x) = \frac{a^2 x^2}{2}$; тогда (3) приобретает вид

$$x'' + \gamma x' + a^2 x = 0.$$

Соответствующее характеристическое уравнение

$$\lambda^2 + \gamma \lambda + a^2 = 0,$$

его корни $\lambda_{1,2} = -\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - a^2}$ и при $\lambda_1 \neq \lambda_2$, т. е. $\gamma \neq 2a$, общее решение $c_1 \exp(\lambda_1 t) + c_2 \exp(\lambda_2 t)$.

Скорость убывания решений рассматриваемого уравнения определяется величиной

$$\sigma(\gamma) = \max(\operatorname{Re} \lambda_1, \operatorname{Re} \lambda_2).$$

При $\gamma \leq 2a$ имеем $\frac{\gamma^2}{4} - a^2 \leq 0$, и поэтому

$$\operatorname{Re} \lambda_1 = \operatorname{Re} \lambda_2 = \sigma(\gamma) = -\frac{\gamma}{2} \geq -a.$$

При $\gamma > 2a$ λ_1 и λ_2 вещественны и, очевидно,

$$\operatorname{Re} \lambda_1 = -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2} > \operatorname{Re} \lambda_2 = -\frac{\gamma}{2} - \sqrt{\frac{\gamma^2}{4} - a^2}.$$

Тогда

$$\begin{aligned} \sigma(\gamma) = \operatorname{Re} \lambda_1 &= -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2} = \\ &= -\frac{a^2}{\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2}} > -\frac{a^2}{\frac{\gamma}{2}} > -a. \end{aligned}$$

Таким образом, график $\sigma(\gamma)$ имеет вид, изображенный на рис. 7.6.1, и $\min_{\gamma} \sigma(\gamma) = \sigma(2a) = -a$. Из результатов рассмотрения этой модельной задачи можно сделать следующие качественные выводы.

1. Если коэффициент трения γ очень мал (в нашем случае $\gamma \ll 2a$), то решение системы (3) медленно устанавливается

к положению равновесия; при этом (вследствие условия $\text{Im } \lambda_1, \text{Im } \lambda_2 \neq 0$) происходят колебания около положения равновесия.

2. Если γ велико ($\gamma \gg 2a$), то решение также медленно устанавливается; причина состоит в том, что при большом коэффициенте трения γ движение не может приобрести большой скорости.

3. Оптимальное значение γ лежит где-то посередине и зависит от свойств конкретной функции $\Phi(x)$.

Метод установления с помощью решения системы (3) иногда называют *методом тяжелого шарика*. Это название вызывается следующими соображениями.

Рассмотрим движение материальной точки по поверхности $y = \Phi(x)$ в поле тяжести, направленном в отрицательном направлении оси y ; предположим, что трение пропорционально скорости и точка не может отрываться от поверхности. Тогда ее движение опишется системой уравнений

$$\frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} + \frac{\text{grad } \Phi}{1 + \|\text{grad } \Phi\|^2} = 0.$$

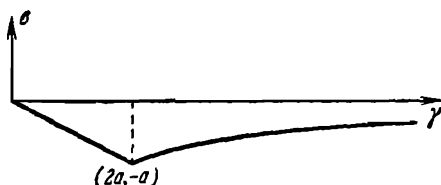


Рис. 761.

Ясно, что решение этой системы с течением времени установится к некоторой стационарной точке функции $\Phi(x)$. В окрестности стационарных точек $\|\text{grad } \Phi\| \ll 1$, и эта система близка к системе (3).

Большинство известных методов установления описывается уравнениями вида

$$A_0\left(x, \frac{dx}{dt}\right) \frac{dx}{dt} + A_1\left(x, \frac{dx}{dt}, \text{grad } \Phi\right) = 0,$$

или

$$B_0\left(x, \frac{dx}{dt}\right) \frac{d^2x}{dt^2} + B_1\left(x, \frac{dx}{dt}, \text{grad } \Phi\right) = 0,$$

где

$$A_0(X, 0) \neq 0, \quad A_1(X, 0, 0) = 0,$$

$$B_0(X, 0) \neq 0, \quad B_1(X, 0, 0) = 0$$

и выполнены условия диссипативности, обеспечивающие сходимость к точке экстремума. Формально говоря, можно обратить операторы A_0 и B_0 и преобразовать эти уравнения к виду, где A_0 и B_0 — тождественные операторы. Однако исходная форма записи уравнений часто практически удобнее.

Может показаться, что построение таких нестационарных процессов, устанавливающихся к решению, уже полностью решает проблему отыскания минимума функции. Осталось «лишь» проинтегрировать получившуюся систему

дифференциальных уравнений при помощи какой-либо стандартной программы.

В действительности сведение решения стационарной задачи к решению нестационарной не всегда дает удовлетворительное решение проблемы минимизации. Остается еще неясным существенный вопрос о выборе величин шагов численного интегрирования. Предположим, что решение нестационарной задачи устанавливается с требуемой точностью к решению стационарной за некоторый промежуток времени T . Если интегрирование производится с малым шагом Δ , то получаемые расчетные точки будут близки к рассматриваемой траектории и можно рассчитывать на попадание в малую окрестность точки минимума. Однако при этом число шагов T/Δ может оказаться недопустимо большим (рис. 7.6.2). Если шаг интегрирования берется очень

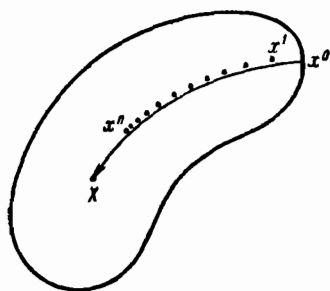


Рис. 7.6.2.

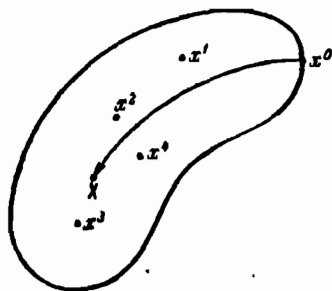


Рис. 7.6.3.

большим, то может случиться, что расчетные точки начнут сильно отклоняться от рассматриваемого решения и никогда не попадут в искомую окрестность точки минимума (рис. 7.6.3).

Чтобы понять механизм явления, полезно обратиться к случаю функционала

$$\Phi^0(x) = \frac{1}{2}(Ax, x) - (b, x), \quad (5)$$

минимум которого при $A > 0$ достигается на решении системы $Ax = b$. Тогда система уравнений (1) приобретает вид

$$\frac{dx}{dt} + (Ax - b) = 0.$$

Заменяя производную $\frac{dx}{dt}$ простейшим разностным отношением, получим

$$\frac{x(t_{n+1}) - x(t_n)}{t_{n+1} - t_n} + Ax(t_n) - b \approx 0;$$

соответствующая расчетная формула

$$x^{n+1} = x^n - \Delta_n (Ax^n - b), \quad \Delta_n = t_{n+1} - t_n.$$

Рассмотрим случай $\Delta_n \equiv \Delta = \text{const} \neq 0$, тогда

$$x^{n+1} = x^n - \Delta (Ax^n - b).$$

Перепишем это соотношение в виде $x^{n+1} = Bx^n + c$, где $B = E - \Delta A$, $c = \Delta b$; пусть $\lambda_1 \leq \dots \leq \lambda_m$ — собственные значения матрицы A , $\lambda_m/\lambda_1 \gg 1$. Собственными значениями матрицы B будут числа $1 - \Delta\lambda_i$. При Δ очень малых собственное число $1 - \Delta\lambda_1$ близко к 1, и поэтому число итераций будет большим. Если Δ велико, то собственное значение $1 - \Delta\lambda_m$ по модулю больше 1 и итерационный процесс расходится. В случае матрицы A с известными границами μ , M расположения собственных значений мы определили ранее оптимальное значение шага $\Delta = 2/(M + \mu)$ в таком итерационном процессе. Если значение Δ_n определяется на каждом шаге из условия $\min_{\Delta_n} \Phi^0(x^n -$

$\Delta_n \text{grad} \Phi^0(x^n))$, то в рассматриваемом случае метод установления совпадет с методом наискорейшего спуска. (Подробнее об особенностях интегрирования таких систем см. § 11 гл. VIII.)

Зададимся постоянным шагом интегрирования $\Delta_n = t_{n+1} - t_n \equiv \Delta$ и напомним одну из простейших аппроксимаций уравнения (3)

$$\frac{x^{n+1} - 2x^n + x^{n-1}}{\Delta^2} + \gamma \frac{x^{n+1} - x^{n-1}}{2\Delta} + \text{grad} \Phi(x^n) = 0.$$

В случае функционала (5) это равенство можно преобразовать к виду

$$x^{n+1} = x^n + \frac{1 - \frac{\gamma\Delta}{2}}{1 + \frac{\gamma\Delta}{2}} (x^n - x^{n-1}) - \frac{\Delta^2}{1 + \frac{\gamma\Delta}{2}} (Ax^n - b).$$

За счет подбора Δ и γ это соотношение можно записать в виде (6.7.21) и, таким образом, получится итерационный процесс, несущественно отличающийся от оптимального линейного итерационного процесса (6.7.16). Как мы видели в § 7 гл. VI, итерационный процесс (6.7.16) позволяет получить решение задачи с заданной точностью за существенно меньшее число итераций, чем метод простой итерации. Поэтому и метод установления с помощью численного интегрирования системы (3) обычно оказывается эффективнее метода установления, соответствующего численному интегрированию системы (1).

Зададимся теперь последовательностью промежутков времени $t_0, t_1, \dots$ и заменим производные в (3) разделенными разностями; предполагая коэффициент γ зависящим от t_n , получаем соотношения

$$2 \frac{\frac{x^{n+1} - x^n}{t_{n+1} - t_n} - \frac{x^n - x^{n-1}}{t_n - t_{n-1}}}{t_{n+1} - t_{n-1}} + \gamma_n \frac{x^{n+1} - x^{n-1}}{t_{n+1} - t_{n-1}} + (Ax^n - b) = 0.$$

При подходящем подборе параметров γ_n и t_n в таком виде могут быть записаны оптимальный линейный итерационный процесс (6.7.15) и метод сопряженных градиентов.

Как аналог метода сопряженных градиентов в случае минимизации функционала $\Phi(x)$ общего вида можно рассматривать следующий метод: последующие приближения ищутся в виде

$$x^{n+1} = x^n + \alpha_n (x^n - x^{n-1}) + \beta_n \text{grad } \Phi(x^n),$$

α_n и β_n определяются из условия $\min_{\alpha_n, \beta_n} \Phi(x^n)$.

Такое соответствие между методами решения стационарных задач путем установления и обычными итерационными методами позволяет обнаруживать новые итерационные методы или новые процессы установления.

Рассмотрим, например, расчетные формулы Ньютона

$$x^{n+1} - x^n = -(F'(x^n))^{-1} F(x^n) \quad (6)$$

решения системы нелинейных уравнений $F(x) = 0$. Этим формулам можно дать следующую интерпретацию: введем непрерывное время t и будем рассматривать величины x^n как значения некоторой функции в моменты времени $t_n = n$. Тогда соотношение (6), переписанное в виде

$$\frac{x^{n+1} - x^n}{t_{n+1} - t_n} = -(F'(x^n))^{-1} F(x^n),$$

можно интерпретировать как получившееся при аппроксимации системы

$$\frac{dx}{dt} = -(F'(x))^{-1} F(x). \quad (7)$$

Решение исходной задачи является стационарной точкой этой системы. При $x = X + \eta$ имеем

$$(F'(x))^{-1} = (F'(X))^{-1} + O(\|\eta\|),$$

$$F(x) = F(X) + F'(X)\eta + O(\|\eta\|^2) = F'(X)\eta + O(\|\eta\|^2),$$

таким образом,

$$\frac{d\eta}{dt} = \frac{dx}{dt} = -((F'(X))^{-1} + O(\|\eta\|))(F'(X)\eta + O(\|\eta\|^2)) = \\ = -\eta + O(\|\eta\|^2).$$

Отсюда следует, что $x = X$ является асимптотически устойчивым решением (7).

Заменяя производную $\frac{dx}{dt}$ разностным отношением по значениям функции в некоторых точках $t_0 = 0, t_1, \dots$, получаем соотношения

$$\frac{x^{n+1} - x^n}{\Delta_n} = -(F'(x^n))^{-1} F(x^n), \quad \Delta_n = t_{n+1} - t_n,$$

иначе,

$$x^{n+1} = x^n - \Delta_n (F'(x^n))^{-1} F(x^n). \quad (8)$$

Таким образом, построение нестационарного процесса, соответствующего методу Ньютона, привело нас к получению итерационного процесса (8) более общего вида.

Рассмотренный пример показывает, что переход от некоторого итерационного алгоритма к соответствующему ему нестационарному процессу имеет много общих черт с построением замыкания вычислительного алгоритма (понятие замыкания алгоритма будет введено в § 4, гл. IX). Нами была доказана сходимость метода Ньютона лишь при достаточно хорошем начальном приближении к решению. С целью расширения области сходимости иногда прибегают к следующей модификации метода Ньютона.

Задаются некоторым функционалом $\Phi(x)$, например $\|F(x)\|^2$, нижняя грань которого достигается на решении задачи. Последовательные приближения отыскиваются в виде (8), причем Δ_n определяется из условия

$$\min_{\Delta_n} \Phi(x^n - \Delta_n (F'(x^n))^{-1} F(x^n)).$$

Возникает вопрос о практическом нахождении величины Δ_n , при которой достигается этот минимум.

Одна из распространенных процедур определения Δ_n состоит в следующем. Задавая $\lambda \in (0, 1)$, $\theta \in (0, 1]$ и $l > 0$ — целое, при каждом n последовательно вычисляются

$$x^{n+1, i} = x^n - \lambda^i (F'(x^n))^{-1} F(x^n), \quad i = 0, \dots, l,$$

и $q^{n+1, i} = \Phi(x^{n+1, i})$. Если при некотором $k \leq l$ оказалось $q^{n+1, k} \leq \theta \Phi(x^n)$, то вычисления прекращают и полагают $x^{n+1} = x^{n+1, k}$; в других процедурах находят $\min_{0 < i \leq l} q^{n+1, i} = q^{n+1, m}$ и полагают $x^{n+1} = x^{n+1, m}$.

Если

$$q^{n+1,0}, \dots, q^{n+1,l} > \theta \Phi(x^n),$$

то в различных программах варьируются разные исходы:

- 1) временный переход к другому методу,
- 2) остановка,
- 3) изменение значений параметров l, λ, θ .

Отметим в заключение, что методы установления могут применяться и в случае минимизации функции в областях с ограничениями. Тогда уравнения (1), (3) следует дополнить какими-то уравнениями, которым будет подчиняться траектория точки, попавшей на границу.

§ 7. Что оптимизировать?

Выше рассмотрен ряд способов минимизации функций и решения систем нелинейных уравнений. Нами охвачена лишь небольшая часть известных методов решения этих задач. Вместо дальнейшего изучения методов обратим внимание на другие, не менее важные, вопросы.

Как уже упоминалось, к минимизации функционалов (функций) сводятся многие задачи управления отраслями промышленности, сельского хозяйства, транспорта, распределения ресурсов и других областей жизни общества. В этих случаях задачи минимизации принято называть *задачами оптимизации*, поскольку основной целью решения этих задач обычно является достижение какого-то наилучшего, оптимального, режима работы. При этом минимизируемую функцию обычно называют *целевой функцией*.

Решение задач оптимизации складывается из следующих элементов: создание математической модели явления, определение целевой функции и важнейших параметров, подлежащих оптимизации, непосредственная минимизация некоторой функции (обычно большого числа переменных), внедрение результатов исследования.

Естественно, что рассмотрение первых двух и последнего вопросов должно проводиться математиками совместно со специалистами конкретной отрасли.

Рассмотрим один реальный пример выбора модели. Производство сахара складывается из следующих этапов: выращивание сахарной свеклы, транспортировка, переработка на заводах. Математики, поставившие перед собой задачу оптимизировать производство сахарной свеклы, в первую очередь решили заняться минимизацией транспортных расходов. Эта задача была успешно решена с помощью ЭВМ, но хозяйственные руководители отказались принять полученный план перевозок к действию. Основой послужил довод о том, что при оптимизации транспортировки свеклы несколько ухудшались условия транспортировки других сельскохозяйственных культур, а также усложнялась организация переработки: некоторые заводы оказывались перегруженными, другие недогруженными.

По-видимому, за этими возражениями стояли следующие соображения.

Себестоимость сахара складывается из следующих составляющих (числа условные):

80% — стоимость сахарной свеклы,

5% — транспортные расходы,

15% — расходы по переработке.

Ожидаемая экономия от минимизации транспортных расходов составляет 10% этих расходов, т. е. всего 0,5% себестоимости сахара. В то же время эта перестройка требует больших хлопот и неизвестно, не произойдет ли более существенных потерь за счет каких-то неучтенных моментов.

Из описанной выше структуры себестоимости видно, что за счет оптимизации расходов по транспортировке и переработке трудно достигнуть существенного повышения эффективности. Внимательный анализ проблемы позволил обнаружить следующее обстоятельство. Выход сахара из тонны свеклы пропорционален ее сахаристости: сахаристость изменяется по определенному закону в процессе созревания и падает в процессе хранения. В частности, при позднем начале уборки происходят большие потери из-за того, что на сахарные заводы поступает одновременно большое количество свеклы и она там долго хранится. В результате этого анализа была поставлена новая целевая функция — оптимизация выхода сахара (при заданном оборудовании) и построена математическая модель, учитывающая изменение сахаристости в процессе копki и хранения свеклы.

Дальнейший анализ показал, что эта модель обладает следующим эффектом. При оптимальном (согласно этой модели) графике работы копki должна начинаться, когда свекла еще находится в периоде интенсивного роста, что экономически невыгодно для непосредственных производителей. Для преодоления этого противоречия нужно, чтобы модель учитывала вопросы заинтересованности производителей. Результатом решения оптимизационной задачи должен быть не только производственный план, но и рекомендации по наилучшей организации оплаты.

Точно такой же вопрос «Что оптимизировать?» возникнет и при решении математических задач. Типичная задача этого сорта: задана функция сложным аналитическим выражением или таблицей; требуется получить ее приближение с заданной точностью при помощи выражения наиболее простого вида.

Рассмотрим задачу получения такого приближения для функций

$$f(q) = 1 - \frac{8}{\pi^2} \sum_{k=1}^{\infty} \frac{\exp\left(-\frac{2k-1}{2}\pi q\right)}{(2k-1)^2} \quad \text{при} \quad 0 \leq q \leq \infty.$$

Для удобства произведем замену переменных

$$x = \exp\left(-\frac{\pi q}{2}\right),$$

т. е. рассмотрим задачу приближения функции

$$g(x) = 1 - \frac{8}{\pi^2} \sum_{k=1}^{\infty} \frac{x^{2k-1}}{(2k-1)^2} \quad \text{на} \quad [0, 1].$$

Попытки приближения многочленами дали неудовлетворительные результаты; анализ показал, что причиной является неограниченность производной $g'(x)$ в окрестности $x=1$. После исследования особенности производной в окрестности

точки $x = 1$ было принято решение отыскивать приближение в виде

$$h(x) = (1-x)(a_0 + a_1x + a_2x^2) \ln \frac{1+x}{1-x} + x(a_3 + a_4x) + 1.$$

За счет выбора параметров a_i удалось получить приближение с точностью 10^{-4} .

При других попытках приближения оказалось, что $g(x)$ также может быть приближена при $x \geq 10^{-4}$ с точностью $5 \cdot 10^{-4}$ выражением вида

$$\frac{(1-x)(1+x(a_1+a_2x))}{1+a_3x+a_4x^2+a_5x^3}.$$

После выбора класса объектов, в котором отыскивается решение задачи, возникает *проблема* разумной параметризации этого класса. Например, многочлен

$$Q(x) = \sum_{i=0}^n a_i x^i,$$

приближающий функцию $f(x)$ на отрезке $[-1, 1]$, можно записать также в виде

$$Q(x) = \frac{b_0}{\sqrt{2}} + \sum_{i=1}^n b_i T_i(x).$$

Как отмечалось, первая запись нежелательна вследствие возможной большой вычислительной погрешности; кроме этого, вторая форма записи имеет следующее преимущество. Пусть приближающий многочлен отыскивается из условия

$$\min_Q \Phi, \quad \text{где} \quad \Phi = \int_{-1}^1 \frac{(f(x) - Q(x))^2}{\sqrt{1-x^2}} dx.$$

При первой форме записи многочлена поверхностями уровня $\Phi(a_0, \dots, a_n) = \text{const}$ являются эллипсоиды с большим разбросом осей; поэтому итерационные методы минимизации этой функции обладают малой скоростью сходимости. При второй форме записи многочлена имеем

$$\begin{aligned} \Phi(b_0, \dots, b_n) = & \int_{-1}^1 \frac{(f(x))^2}{\sqrt{1-x^2}} dx - \sqrt{2} b_0 \int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx - \\ & - 2 \sum_{i=1}^n b_i \int_{-1}^1 \frac{f(x) T_i(x)}{\sqrt{1-x^2}} dx + \frac{\pi}{2} \sum_{i=0}^n b_i^2; \end{aligned}$$

поверхностями уровня функции $\Phi(b_0, \dots, b_n)$ являются сферы, и итерационные способы минимизации обеспечивают быструю сходимость к точке минимума (здесь эта точка известна).

Удачный выбор параметров и критериев их определения особенно важны при обработке результатов наблюдений.

Как одномерная модель движения снежной лавины может рассматриваться система дифференциальных уравнений

$$\begin{aligned} \frac{\partial h}{\partial t} + \frac{\partial hv}{\partial s} &= 0, \\ \frac{\partial v}{\partial t} + v \frac{\partial v}{\partial s} + \frac{g}{2h} \frac{\partial h^2 \cos \psi}{\partial s} &= g (\sin \psi - \mu \cos \psi) - k \frac{v^2}{h}; \end{aligned} \quad (1)$$

граничные условия на переднем фронте лавины

$$h(w - v) = h_0 w, \quad h_0 w v = \frac{1}{2} g h^2 \cos \psi - \sigma h;$$

здесь h — высота снега, v — его скорость, ψ — угол наклона склона, w — скорость переднего фронта, параметры μ , k , σ — неизвестные.

Была сделана попытка определения этих параметров из наблюдения за положением переднего фронта лавины, двигавшейся по склону с постоянным наклоном $\psi = \text{const}$

Параметры пытались определить из условия минимума функции

$$\Phi(\mu, k, \sigma) = \sum_i (s(\mu, k, \sigma, t_i) - s_i)^2;$$

здесь s_i — наблюдаемые положения переднего фронта в моменты времени t_i , $s(\mu, k, \sigma, t_i)$ — получаемые в результате численного интегрирования системы (1). При непосредственном применении большого числа методов минимизации не обнаруживалось никакой тенденции к установлению получаемых значений μ , k , σ . Анализ задачи указал на следующую причину этого явления. Оказывается, что при $\psi = \text{const}$ с большой точностью выполняется приближенное равенство $s(\mu, k, \sigma, t) \approx \lambda t$; здесь $\lambda = \left(1 + \frac{1}{\sqrt{\beta}}\right)^{1/2}$, $\beta = \frac{k}{\text{tg } \psi - \mu}$.

Поэтому функция $\Phi(\mu, k, \sigma)$ с большой точностью близка к некоторой функции $\Phi_0(\lambda)$ одной переменной λ . Таким образом, при минимизации функции Φ по переменным μ , k , σ можно рассчитывать не на получение значений всех параметров μ , k , σ , соответствующих точке минимума, а на получение лишь значения λ , соответствующего точке минимума. Пусть $\Phi_0(\lambda)$ достигается при $\lambda = \lambda_0$. Поверхность $\lambda = \lambda_0$ является двумерной, в данном случае цилиндрической, поверхностью в пространстве переменных μ , k , σ . На этой поверхности функция $\Phi(\mu, k, \sigma)$ оказывается примерно постоянной, и при ее численной минимизации происходит беспорядочное на вид перемещение точек (μ, k, σ) вдоль этой поверхности. Таким образом, в данном случае истинная причина неудачи состояла не в несовершенстве методов минимизации, а в недостаточности имеющейся информации по отношению к поставленной цели.

Вернемся к замечаниям общего характера.

При построении модели задачи возникает желание — создать подробную математическую модель, учтя многие детали задачи, и затем произвести полную оптимизацию проблемы за счет наилучшего выбора всех параметров. Этот путь чреват рядом неприятностей.

Первые трудности возникают при описании модели; стремление учесть слишком многое увеличивает возможность упустить

из внимания что-либо существенное и тем самым ухудшить модель. При оптимизации по очень большому числу параметров возникает задача минимизации функции большого числа переменных, численное решение которой встречает иногда непреодолимые трудности.

Предположим на мгновение, что такую минимизацию все-таки удалось осуществить. Возникает задача доведения информации до сведения заказчиков с целью реализации результатов математического анализа модели. Заказчик обычно принимает решение, исходя из своих собственных, качественных, критериев, не вдаваясь в детали модели. В случае получения рекомендаций относительно слишком большого числа параметров он не всегда сможет сделать для себя вывод о разумности этих рекомендаций и с большой вероятностью вообще откажется от них.

Поэтому на первых этапах оптимизации особенно важно построить простейшую модель, учитывающую лишь основные, определяющие, параметры. Это требование вызывается также следующим. Часто поначалу заказчик испытывает недоверие к математикам, вторгающимся в новую для них область. Только после завоевания доверия за счет достижения положительного эффекта при оптимизации важнейших параметров имеет смысл учесть влияние важнейших уже из неучтенных параметров. Следует также иметь в виду, что само построение уточненной модели явления обычно становится возможным лишь после совместного анализа заказчиком и исполнителем результатов обсчета упрощенной модели.

Очень важно также, чтобы результаты математических исследований при диалоге с нематематиками подавались в достаточно наглядной, привычной форме, по возможности с привлечением минимального математического аппарата. Неоправданное привлечение абстрактных понятий математики редко приносит реальную пользу.

§ 8. Как оптимизировать?

После выбора модели, целевой функции и параметризации задачи возникает задача минимизации функции обычно большого числа переменных в области, принадлежность к которой задается условием выполнения большого числа ограничений — равенств или неравенств. Наличие ограничений существенно увеличивает сложность задачи минимизации: как правило, точкой экстремума оказывается некоторая граничная точка области. Из-за сложности вопроса и ввиду наличия обширной специальной литературы оставим в стороне многочисленные тонкости, связанные с минимизацией функций при наличии ограничений, и рассмотрим общие вопросы.

Вследствие большой важности решения задач оптимизации для самых различных сторон жизни общества в настоящее время накопился большой багаж методов и стандартных программ решения задач оптимизации. Поэтому при решении единичной конкретной задачи часто наиболее оправдано обращение к одной из имеющихся стандартных программ.

При решении новых задач оптимизации полезно помнить о следующих моментах. Для рассматриваемых в этой главе методов сходимость приближений всегда доказывалась в предположении наличия достаточно хорошего начального приближения. Это ограничение на метод вызывается существом дела; так, поверхность уровня многочлена степени n от m переменных может содержать порядка n^m не связанных между собой компонент с весьма сложным взаимным расположением; поэтому, например, безнадежно рассчитывать на построение алгоритма, позволяющего быстро найти минимум любого многочлена восьмой степени от десяти переменных.

Вышесказанное следует учитывать при использовании стандартных программ, в описании которых указывается на возможность минимизации функций *очень* большого числа переменных. *В лучшем случае* оказывается, что:

- а) программа эффективно решает задачи минимизации из класса, с которыми автор программы обычно имеет дело, или
- б) программа, формально говоря, может решить любую задачу минимизации, но требуемое время решения в подавляющем числе случаев выходит за всякие разумные пределы.

При решении новых задач зачастую приходится разрабатывать специальные, приспособленные именно для этого класса задач, методы отыскания начального приближения к решению. Сложность отыскания приемлемого начального приближения можно проиллюстрировать следующим примером.

Существует ряд стандартных программ наилучшего приближения функций отношением многочленов

$$\frac{P_m(x)}{Q_n(x)} \quad (1)$$

заданных степеней m и n . При разработке стандартной программы вычисления интегралов Френеля проводились следующие эксперименты. Каждая из стандартных программ применялась для приближения рассматриваемой функции на некотором отрезке при всех m, n в пределах $0 < m + n \leq 12$. Оказалось, что не менее чем в 3 случаях из 90 возможных каждая программа выдавала ответ, что она не может решить рассматриваемую задачу. У некоторых программ доля таких отказов превосходила половину. В то же время для всех этих программ доказаны теоремы о сходимости при достаточно хорошем начальном приближении. Можно говорить, что до настоящего времени не существует эффективной и в то же время универсальной стандартной программы для отыскания наилучшего приближения функции рациональными дробями вида (1).

Чтобы не создалось впечатления полной безнадежности решения сколько-нибудь сложных задач оптимизации, рассмотрим тот же вопрос с оптимистических позиций.

С этих позиций приведенные выше доводы о сложности минимизации многочленов можно рассматривать как малоубедительные — ведь на самом деле нам никогда не потребуется минимизировать произвольный многочлен. Существует мнение, что задачи минимизации функций с *очень* сложной структурой линий уровня встречаются довольно редко.

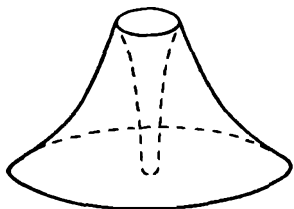


Рис 781.

Рассмотрим пример задачи, где сам факт необходимости ее решения может быть поставлен под сомнение.

Неудачи при попытках получения хорошего начального приближения могут быть вызваны следующей особенностью поведения рассматриваемой функции. Точка минимума находится в очень узкой «яме» — при удалении от нее во всех направлениях функция резко возрастает, а потом начинает убывать (рис. 7.8.1). Предположим, что точка минимума находится путем продвижения по градиенту функции Φ , иначе — численным интегрированием системы

$$\frac{dx}{dt} + \text{grad } \Phi = 0. \quad (2)$$

Тогда множество начальных условий, исходя из которых мы будем приходить в точку минимума, находится в небольшой окрестности этой точки. Можно говорить, что при рассмотрении окрестности точки минимума в микроскопическом масштабе эта точка оказывается точкой притяжения решений системы (2); при рассмотрении в более крупном масштабе она уже оказывается точкой отталкивания (рис. 7.8.2). Точно такой же характер поведения последовательных приближений к точке минимума будет и у других итерационных методов.

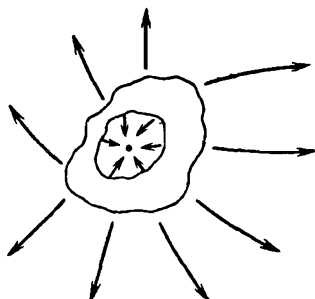


Рис. 782.

Если «яма», где находится точка минимума, очень узкая, иногда и не стоит искать эту точку минимума. В самом деле, пусть, например, параметры x , отвечают некоторой реальной работающей системе управления; в работе этой системы неизбежны некоторые сбои, т. е. изменения этих параметров. Если точка минимума функции Φ , находится в такой узкой «яме», то малые

сбои могут существенно испортить характеристики работы системы. В свете вышесказанного выбор точки экстремума в случае, изображенном на рис. 7.8.3, требует дополнительного изучения. Исследователи, обладающие опытом решения практических задач оптимизации, утверждают, что подобные целевые функции с узкими «ямами» возникают обычно в случае, когда математическая модель рассматриваемого явления построена неудачно.

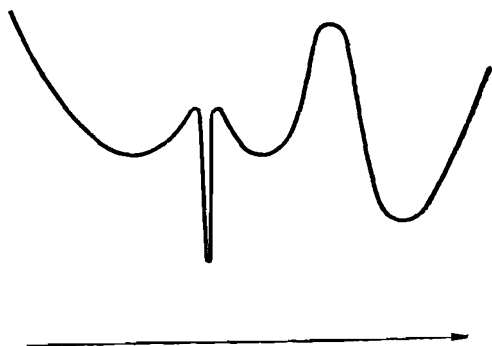


Рис. 7.8.3.

Более существенные причины для оптимизма состоят в следующем. Во многих случаях создание математической модели и ее оптимизация зачастую имеют целью улучшение работы уже существующей системы. В этих случаях параметры реальной системы часто являются хорошим приближением для дальнейшей оптимизации.

При разработке новой системы типична следующая линия поведения. Сначала строится и оптимизируется простейшая модель, учитывающая важнейшие факторы. Затем модель постепенно усложняется за счет учета все новых и новых факторов. Таким образом, последовательно возникают задачи минимизации функций все большего числа параметров. При удачном построении вспомогательных моделей решение каждой из этих задач оптимизации обычно оказывается хорошим начальным приближением для следующей по сложности задачи.

Это обстоятельство часто используется следующим образом. Пусть перед нами стоит задача минимизации функции большого числа параметров $x_1, \dots, x_n$. Построим функцию меньшего числа параметров $X_1, \dots, X_m$, $m < n$, приближающую рассматриваемую функцию, иначе — упрощенную модель с определяющими параметрами $X_1, \dots, X_m$. Произведем оптимизацию этой функции (модели) и получившееся приближение примем

за начальное для последующей оптимизации. Иногда оказывается полезным произвести несколько шагов такого упрощения функции (модели) и введения новых параметров.

Оптимизация функций (моделей) меньшего числа параметров оказывается более легкой из-за следующих двух причин: становится проще структура линий уровня минимизируемой функции, вычисления каждого значения функции обычно требуют меньшего объема вычислений.

Пример. Ищется $\min_{x_1, x_2, x_3} \Phi(x_1, x_2, x_3)$:

$$\Phi(x_1, x_2, x_3) = 1,1x_1^2 - 2x_1x_2 + x_2^2 + x_3^4 + \exp(x_1x_3).$$

Функция $\Phi(x_1, x_2, x_3)$ близка к

$$\Phi_1(x_1, x_2, x_3) = (x_1 - x_2)^2 + x_3^4 + \exp(x_1x_3).$$

Очевидно,

$$\min_{x_2} \Phi_1(x_1, x_2, x_3) = \Phi_2(x_1, x_3) = x_3^4 + \exp(x_1x_3).$$

Поиск начального приближения свелся к минимизации функции двух переменных.

Заметим, что требование уменьшения числа параметров при построении упрощенной модели не является обязательным.

Что такое важность факторов или параметров? Можно говорить, что важные параметры это те, от которых функция сильно зависит; на языке математики это означает, что производные функции по этим параметрам относительно велики. Второстепенные параметры это те, от которых рассматриваемая функция зависит слабо, т. е. те, производные по которым малы. Если производная функции $\Phi(x_1, \dots, x_n)$ по параметру x_n мала, то это означает близость функции $\Phi(x_1, \dots, x_n)$ к функции меньшего числа переменных $\Phi(x_1, \dots, x_{n-1}, \bar{x}_n)$, $\bar{x}_n = \text{const}$. Формально важность параметров можно определить, оценивая производные рассматриваемой функции. Однако для сложных задач такая оценка трудоемка; характеристика параметров по более наглядному критерию, важности параметров, дает возможность руководителю производственной системы или научному работнику подсказать математику первоочередность выбора параметров.

Иногда приемлемый метод оптимизации или способ отыскания хорошего начального приближения можно получить, изучая принципы, которыми руководствуются в своей работе опытный практический работник или руководитель, или приемы их работы.

Нижеследующий пример иллюстрирует полезность использования опыта практических работников.

На одном новом заводе долгое время не удавалось наладить ритмичное производство из-за недостаточного опыта работы операторов. Попытки создания модели производственного процесса, достаточно точной, но в то же время поддающейся анализу средствами математики с помощью ЭВМ, не приводили к успеху. В конце концов пришлось временно отказаться от разработки математической модели и пойти по следующему пути автоматизации и оптимизации производства. В память ЭВМ были записаны режимы работы лучших операторов на родственных предприятиях. Далее, в зависимости от имеющихся на данный момент условий, машина выбирала режим работы, наиболее близкий к режиму работы одного из лучших операторов. Такое мероприятие позволило устранить возникшие трудности.

Способы нахождения начального приближения и сами итерационные методы часто имеют аналогию с какими-то реальными процессами в других явлениях.

Выше упоминалось о методе тяжелого шарика и других методах установления. Работа мозга и зрительного аппарата при отыскании какого-либо предмета, по-видимому, организована по следующей схеме. Сначала производится беглый просмотр всего поля зрения в крупном масштабе, на основании полученной информации выбирается участок для дальнейшего просмотра, затем производится просмотр этого участка в крупном масштабе и т. д. Обращаем внимание на сходство с методами из §§ 17, 18 гл. III и § 5 данной главы.

При решении развивать исследования новых трудных задач важное значение имеет правильная организация научных исследований. Если в основном ясно, в каком направлении нужно развивать исследования, обычно производится концентрация научных усилий в этом направлении. Если наиболее рациональный путь к намеченной цели еще не определился, то часто прибегают к дублированию исследований. Несколько независимых организаций ищут решение, каждая на своем пути, иногда без постоянного обмена информацией. Хотя, на первый взгляд, кажется, что взаимный обмен информацией всегда полезен, постоянный обмен ею может и помешать возникновению и продвижению оригинальных решений проблемы.

Существует гипотеза (неообщепризнанная), что сходным образом работает мозг при решении какой-либо проблемы: получаемая информация фиксируется не вполне детерминированным образом в различных участках мозга; в то же время в каждый момент работы над проблемой эта информация извлекается лишь из локализованного участка мозга.

По аналогии со сказанным выше напрашивается следующий подход к решению задачи оптимизации (а также и любых других задач) в случае, если требуется срочное получение результата. Для решения задачи поочередно или независимо используются несколько известных методов решения подобных задач. В обоих случаях, при поочередном или при независимом использовании, алгоритмы A_p , $p = 1, \dots, l$, работают циклически; длительность t_{pq} промежутка времени, в течение которого

алгоритм A_p работает q -й раз, задается пользователем или определяется в процессе работы. При *поочередном использовании* каждый алгоритм начинает минимизацию с приближения, полученного предшествующим алгоритмом. При *независимом использовании* каждый алгоритм начинает минимизацию с приближений, полученного в результате предшествующего применения данного алгоритма. Таким образом, в этом случае алгоритмы работают по принципу «кто быстрее».

Режим независимого использования алгоритмов является аналогом параллельной работы организаций при отсутствии взаимного обмена информацией. Как аналог реальной организации научных исследований с обменом информацией в дискретные моменты времени может рассматриваться следующий режим работы. В начале каждого промежутка времени t_{pq} алгоритм A_p просматривает некоторую совокупность приближений из полученных всеми алгоритмами и выбирает наилучшее приближение, исходя из своих позиций. Например, он может просматривать все приближения последнего цикла или приближения, полученные алгоритмами в концах всех промежутков t_{ij} .

Иногда может принести пользу следующая организация работы: после получения алгоритмом A_p приближения, очень хорошего с позиций алгоритма A_s , предоставляется время для работы алгоритма A_s .

Конечно, не следует думать, что непосредственное копирование различных реальных систем всегда позволит наилучшим образом решить рассматриваемую оптимизационную задачу.

Подведем общий итог наших рассуждений. Обычно задачи оптимизации функций большого числа переменных очень трудны; при решении новых задач приходится затрачивать много, иногда бесплодных, усилий, производя пробные просчеты по различным известным и новым алгоритмам. Однако при наличии таких благоприятных факторов, как контакт с практическими работниками и возможность анализа упрощенных моделей, есть все основания сохранять уверенность в благоприятном исходе попыток решения задачи.

Заметим, что в настоящее время намечается существенный прогресс в области решения задач оптимизации за счет диалога человека, математика или руководителя системы, с ЭВМ, работающей в режиме разделения времени.

Дополнительная литература к главе VII

II [15, 16, 29, 34, 37, 41]; § 2. III [29]; §§ 3, 4. I [2], II [29, 41]; § 5. II [41], III [19], §§ 7, 8. II [29, 30, 31, 35], III [47].

См. также: Ф. П. Васильев, Лекции по методам решения экстремальных задач, М., изд-во МГУ, 1974.

Часть III

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Глава VIII

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ КОШИ

Задача решения обыкновенных дифференциальных уравнений сложнее задачи вычисления однократных интегралов, и доля задач, интегрируемых в явном виде, здесь существенно меньше.

В рассуждениях об интегрируемости в явном виде обычно имеется в виду, что решение может быть вычислено при помощи конечного числа «элементарных» операций сложения, вычитания, умножения, деления, возведения в степень, логарифмирования, потенцирования, вычисления синуса и косинуса. Еще в период, предшествующий появлению ЭВМ, понятие «элементарной» операции и «элементарной» функции претерпело изменение. Решения некоторых частных задач настолько часто встречались в приложениях, что пришлось составить таблицы их значений, например, таблицы интегралов Фрелея

$$\int_0^x \exp(it^2) dt,$$

интегрального синуса $\int_0^x \frac{\sin t}{t} dt$, решений дифференциального уравнения

$$x^2 y'' + xy' + (x^2 - v^2)y = 0,$$

называемых функциями Бесселя, и ряда других функций. Эти функции называют специальными функциями. При наличии таких таблиц исчезает принципиальная разница между вычислением функций $\sin x$, $\ln x$, ... и специальных функций. В том и другом случаях можно вычислять значения этих функций при помощи таблиц, и те и другие функции можно вычислять, приближаясь многочленами, рациональными дробями и т. д. Таким образом, в класс задач, интегрируемых в явном виде, включились задачи, решения которых выражаются через специальные функции. Однако и этот, более широкий, класс составляет относительно малую долю задач, предъявляемых к решению. Существенное расширение класса реально решаемых дифференциальных уравнений, а, следовательно, и расширение сферы применения математики произошло с созданием численных методов и распространением применения ЭВМ.

В настоящее время объем работы по подготовке к решению на ЭВМ дифференциального уравнения не так уж существенно превышает объем

работы по написанию этого уравнения; при желании можно непосредственно из машины получить график решения или изображение решения на экране. В результате этого отпала необходимость в изучении большого числа частных способов интегрирования уравнений в явном виде.

Эта глава посвящена обсуждению различных вопросов теории численных методов решения дифференциальных уравнений, которые, в частности, послужили основой стандартных программ решения обыкновенных дифференциальных уравнений.

§ 1. Разложение решения в ряд Тейлора

Одним из старейших методов решения дифференциальных уравнений является метод разложения в ряд Тейлора.

Пусть требуется найти на отрезке $[x_0, x_0 + X]$ решение дифференциального уравнения

$$y' = f(x, y) \quad (1)$$

при начальном условии $y(x_0) = y_0$; $f(x, y)$ аналитична в точке (x_0, y_0) . Дифференцируя (1) по x , имеем соотношения

$$y'' = f_x(x, y) + f_y(x, y) y',$$

$$y''' = f_{xx}(x, y) + 2f_{xy}(x, y) y' + f_{yy}(x, y) y'^2 + f_y(x, y) y'', \dots$$

Подставляя $x = x_0$ и $y = y_0$, получаем последовательно значения

$$y'(x_0), \quad y''(x_0), \quad y'''(x_0), \dots$$

Таким образом, можно написать приближенное равенство

$$y(x) \approx \sum_{i=0}^n \frac{y^{(i)}(x_0)}{i!} (x - x_0)^i. \quad (2)$$

Если $|x - x_0|$ больше радиуса сходимости ряда

$$\sum_i \frac{y^{(i)}(x_0)}{i!} (x - x_0)^i,$$

то погрешность (2) не стремится к нулю при $n \rightarrow \infty$. Иногда целесообразно поступить следующим образом.

Разобьем $[x_0, x_0 + X]$ на отрезки

$$[x_{j-1}, x_j], \quad j = 1, \dots, N;$$

последовательно получаем приближения y_j к значениям решения

$$y(x_j), \quad j = 1, \dots, N,$$

по следующему правилу: пусть значение y_j уже найдено, вычисляем значения в точке x_j производных $y_j^{(i)}$ решения исходного

дифференциального уравнения, проходящего через точку (x_j, y_j) ; на отрезке $[x_j, x_{j+1}]$ полагаем

$$y(x) \approx z_l(x) = \sum_{i=0}^n \frac{y_l^{(i)}}{i!} (x - x_j)^i \quad (3)$$

и, соответственно, полагаем

$$y_{l+1} = z_l(x_{j+1}). \quad (4)$$

Далее мы проведем подробную оценку погрешности одношаговых способов интегрирования, к которым относится рассматриваемый способ. На настоящем этапе обсуждения можно привести следующие соображения в пользу того, что погрешность этого метода может оказаться малой. Рассмотрим случай $x_{j+1} - x_j \equiv h$; если бы значение y_j совпадало со значением точного значения $y(x_j)$, то погрешность от замены y_{j+1} на $z_j(x_{j+1})$ имела бы порядок $O(h^{n+1})$. Поскольку мы вносим погрешность на $O(h^{-1})$ отрезках, то можно ожидать, что при измельчении шага сетки будет выполняться соотношение

$$\max_{0 \leq j \leq N} |y_j - y(x_j)| = O(h^n).$$

Это рассуждение является типичным примером правдоподобных рассуждений, приводящихся иногда в обоснование применения тех или иных численных методов; однако следует иметь в виду, что во многих случаях встречается такая ситуация: в обоснование сходимости метода такие рассуждения можно провести, а на самом деле сходимости приближенного решения к точному нет. Поэтому вопрос строгого обоснования сходимости методов при мельчении шага, а также получения оценки погрешности является вопросом, имеющим не только теоретическое, но и важнейшее практическое значение.

Описанный выше способ интегрирования дифференциальных уравнений в случае $n > 1$ был исключен из широкой практики более столетия назад. При использовании этого метода нужно вычислять значения функции f и всех ее производных $f_{x^j y^{m-1}}$ при $m < n$, т. е. вычислять $\frac{n(n+1)}{2}$ различ-

ных функций. Кроме других причин, при ручном счете это неприятно еще и тем, что рассеивается внимание расчетчика и возрастает вероятность ошибок. При работе с ЭВМ применение этого метода требует написания большого числа блоков вычисления производных, что противоречит основной тенденции упрощения отношений между пользователем и машиной.

В настоящее время имеются системы программ, которые осуществляют формульное дифференцирование функций; при заданной программе вычисления функции $f(x, y)$ эта система программ расписывает программы вычисления производных функции $f(x, y)$; таким образом, отпадает необходимость написания программ вычисления этих производных. Однако по затратам

машинного времени подобные алгоритмы, как правило, существенно уступают рассматриваемым ниже методам типа Рунге — Кутты и конечно-разностным методам.

Наши рассуждения не следует понимать как категорические рекомендации воздерживаться от применения методов, использующих разложение решения в ряд Тейлора. В ряде случаев ЭВМ используются для решения дифференциальных уравнений строго фиксированного типа. Например, при расчетах траекторий движения небесных тел приходится многократно интегрировать системы дифференциальных уравнений вполне определенного вида при различных начальных условиях и различных значениях параметров в правых частях. То обстоятельство, что все время решается одна и та же система дифференциальных уравнений, может дать следующее преимущество: конкретные формулы для производных правых частей системы обычно имеют много общих частей; в результате для отдельных классов систем может случиться, что одновременное вычисление всех этих производных потребует относительно малого числа арифметических операций и рассматриваемый метод окажется эффективнее других методов численного интегрирования. В дальнейшем мы, в основном, будем изучать методы численного интегрирования, требующие вычисления только значений правой части; на основе этих методов обычно и строятся стандартные программы решения широких классов типовых задач.

§ 2. Методы Рунге — Кутты

В частном случае $n = 1$ расчетная формула (1.3) имеет вид

$$y_{j+1} = y_j + hf(x_j, y_j), \quad h = x_{j+1} - x_j. \quad (1)$$

Этот метод называется *методом Эйлера*. Можно построить другой класс расчетных формул, к которому принадлежит метод Эйлера. Укажем сначала простейшие методы этого класса, получаемые из наглядных соображений. Пусть известно значение решения $y(x)$ и требуется вычислить значение $y(x+h)$. Рассмотрим равенство

$$y(x+h) = y(x) + \int_0^h y'(x+t) dt. \quad (2)$$

При замене интеграла в правой части на величину $hy'(x)$ погрешность имеет порядок $O(h^2)$, т. е.

$$y(x+h) = y(x) + hy'(x) + O(h^2);$$

поскольку $y'(x) = f(x, y(x))$, то отсюда имеем

$$y(x+h) = y(x) + hf(x, y(x)) + O(h^2).$$

Отбрасывая член порядка $O(h^2)$ и обозначая $x = x_j$, $x+h = x_{j+1}$, получим расчетную формулу Эйлера (1). Для получения более точной расчетной формулы нужно точнее аппроксимировать интеграл в правой части (2). Воспользовавшись фор-

мулой трапеций, получим

$$y(x+h) = y(x) + \frac{h}{2}(y'(x) + y'(x+h)) + O(h^3),$$

иначе,

$$y(x+h) = y(x) + \frac{h}{2}(f(x, y(x)) + f(x+h, y(x+h))) + O(h^3). \quad (3)$$

Соответствующая расчетная формула:

$$y_{i+1} = y_i + \frac{h}{2}(f(x_i, y_i) + f(x_{i+1}, y_{i+1})). \quad (4)$$

Обычно это уравнение неразрешимо явно относительно y_{i+1} . Поэтому произведем дальнейшее преобразование алгоритма. Заменим $y(x+h)$ в правой части (3) на некоторую величину

$$y^* = y(x+h) + O(h^2); \quad (5)$$

тогда правая часть изменится на величину

$$\frac{h}{2}(f(x+h, y^*) - f(x+h, y(x+h))) = \frac{h}{2}f_y(x+h, \bar{y})(y^* - y(x+h)),$$

где $\bar{y}$ находится между y^* и $y(x+h)$: вследствие предположения (5) эта величина имеет порядок $O(h^3)$. Таким образом, при условии (5)

$$y(x+h) = y(x) + \frac{h}{2}(f(x, y(x)) + f(x+h, y^*)) + O(h^3);$$

условию (5) удовлетворяет результат вычислений по расчетной формуле Эйлера

$$y^* = y(x) + hf(x, y(x)).$$

Последние соотношения определяют пару расчетных формул

$$\begin{aligned} y_{i+1}^* &= y_i + hf(x_i, y_i), \\ y_{i+1} &= y_i + \frac{h}{2}(f(x_i, y_i) + f(x_{i+1}, y_{i+1}^*)). \end{aligned} \quad (6)$$

Построим другую пару формул с погрешностью на шаге такого же порядка. Интеграл в правой части (2) заменим по формуле прямоугольников:

$$y(x+h) = y(x) + hy'(x + \frac{h}{2}) + O(h^3)$$

или, все равно, что

$$y(x+h) = y(x) + hf\left(x + \frac{h}{2}, y\left(x + \frac{h}{2}\right)\right) + O(h^3).$$

Если

$$y^* = y\left(x + \frac{h}{2}\right) + O(h^2),$$

то, как и в предшествующем случае, имеем

$$y(x+h) = y(x) + hf\left(x + \frac{h}{2}, y^*\right) + O(h^3).$$

В качестве y^* можно взять результат вычислений по формуле Эйлера с шагом $\frac{h}{2}$: $y^* = y(x) + \frac{h}{2} f(x, y(x))$. Этим соотношениям соответствует пара расчетных формул

$$\begin{aligned} y_{l+\frac{1}{2}} &= y_l + \frac{h}{2} f(x_l, y_l), \\ y_{l+1} &= y_l + hf\left(x_l + \frac{h}{2}, y_{l+\frac{1}{2}}\right). \end{aligned} \quad (7)$$

Полученные методы относятся к семейству методов Рунге—Кутты, имеющих следующий вид. В процессе вычислений фиксированы некоторые числа

$$\alpha_2, \dots, \alpha_q, \quad \rho_1, \dots, \rho_q, \quad \beta_{ij}, \quad 0 < j < i \leq q;$$

последовательно получаем

$$\begin{aligned} k_1(h) &= hf(x, y), \\ k_2(h) &= hf(x + \alpha_2 h, y + \beta_{21} k_1(h)), \\ &\dots \dots \dots \\ k_q(h) &= hf(x + \alpha_q h, y + \beta_{q,1} k_1(h) + \dots + \beta_{q,q-1} k_{q-1}(h)) \end{aligned}$$

и полагаем

$$y(x+h) \approx z(h) = y(x) + \sum_{i=1}^q \rho_i k_i(h).$$

Рассмотрим вопрос о выборе параметров $\alpha_i, \rho_i, \beta_{ij}$. Обозначим $\varphi(h) = y(x+h) - z(h)$. Если $f(x, y)$ — достаточно гладкая функция своих аргументов, то $k_1(h), \dots, k_q(h)$ и $\varphi(h)$ — гладкие функции параметра h . Предположим, что

$$\varphi(0) = \varphi'(0) = \dots = \varphi^{(s)}(0) = 0$$

при произвольных достаточно гладких функциях $f(x, y)$, а $\varphi^{(s+1)}(0) \neq 0$ для некоторой гладкой $f(x, y)$. Согласно формуле Тейлора, справедливо равенство

$$\varphi(h) = \sum_{i=0}^s \frac{\varphi^{(i)}(0)}{i!} h^i + \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1} = \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1}, \quad (8)$$

где $0 < \theta < 1$. Величина $\varphi(h)$ называется погрешностью метода на шаге, а s порядком погрешности метода. При $q = 1$

имеем

$$\begin{aligned}\varphi(h) &= y(x+h) - y - p_1 h f(x, y), \\ \varphi(0) &= 0, \quad \varphi'(0) = (y'(x+h) - p_1 f(x, y))|_{h=0} = \\ &= f(x, y)(1 - p_1), \quad \varphi''(h) = y''(x+h),\end{aligned}$$

здесь и далее $y = y(x)$. Очевидно, что равенство $\varphi'(0) = 0$ выполняется при всех $f(x, y)$ лишь в случае $p_1 = 1$. Этому значению p_1 соответствует метод Эйлера. Для погрешности этого метода на шаге, согласно (8), получаем выражение

$$\varphi(h) = \frac{y''(x) h^2}{2}.$$

Рассмотрим случай $q=2$, тогда

$$\varphi(h) = y(x+h) - y - p_1 h f(x, y) - p_2 h f(\bar{x}, \bar{y}),$$

где $\bar{x} = x + \alpha_2 h$, $\bar{y} = y + \beta_{21} h f(x, y)$.

Вычислим производные функции $\varphi(h)$:

$$\begin{aligned}\varphi'(h) &= y'(x+h) - p_1 f(x, y) - p_2 f(\bar{x}, \bar{y}) - p_2 h (\alpha_2 f_x(\bar{x}, \bar{y}) + \\ &\quad + \beta_{21} f_y(\bar{x}, \bar{y}) f(x, y)), \\ \varphi''(h) &= y''(x+h) - 2p_2 (\alpha_2 f_x(\bar{x}, \bar{y}) + \\ &\quad + \beta_{21} f_y(\bar{x}, \bar{y}) f(x, y)) - p_2 h (\alpha_2^2 f_{xx}(\bar{x}, \bar{y}) + \\ &\quad + 2\alpha_2 \beta_{21} f_{xy}(\bar{x}, \bar{y}) f(x, y) + \beta_{21}^2 f_{yy}(\bar{x}, \bar{y}) (f(x, y))^2), \\ \varphi'''(h) &= y'''(x+h) - 3p_2 (\alpha_2^2 f_{xx}(\bar{x}, \bar{y}) + 2\alpha_2 \beta_{21} f_{xy}(\bar{x}, \bar{y}) f(x, y) + \\ &\quad + \beta_{21}^2 f_{yy}(\bar{x}, \bar{y}) (f(x, y))^2) + O(h).\end{aligned}$$

Согласно исходному дифференциальному уравнению,

$$y' = f, \quad y'' = f_x + f_y f, \quad y''' = f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_y y''.$$

Подставим в выражения $\varphi(h)$, $\varphi'(h)$, $\varphi''(h)$, $\varphi'''(h)$ значение $h=0$ и воспользуемся этими соотношениями; получим

$$\begin{aligned}\varphi(0) &= y - y = 0, \\ \varphi'(0) &= (1 - p_1 - p_2) f(x, y), \\ \varphi''(0) &= (1 - 2p_2 \alpha_2) f_x(x, y) + (1 - 2p_2 \beta_{21}) f_y(x, y) f(x, y), \\ \varphi'''(0) &= (1 - 3p_2 \alpha_2^2) f_{xx}(x, y) + (2 - 6p_2 \alpha_2 \beta_{21}) f_{xy}(x, y) f(x, y) + \\ &\quad + (1 - 3p_2 \beta_{21}^2) f_{yy}(x, y) (f(x, y))^2 + f_y(x, y) y''(x).\end{aligned}\quad (9)$$

Соотношение $\varphi'(0) = 0$ выполняется при всех f , если

$$1 - p_1 - p_2 = 0, \quad (10)$$

соотношение $\varphi''(0) = 0$, если

$$1 - 2p_2\alpha_2 = 0 \quad \text{и} \quad 1 - 2p_2\beta_{21} = 0. \quad (11)$$

Таким образом $\varphi(0) = \varphi'(0) = \varphi''(0) = 0$ при всех $f(x, y)$, если выполнены три указанных выше соотношения ((10), (11)) относительно четырех параметров; произвольно задавая один из параметров, получим различные методы Рунге — Кутты с погрешностью второго порядка малости по h . Например, при $p_1 = 1/2$ получаем $p_2 = 1/2$, $\alpha_2 = 1$, $\beta_{21} = 1$, что соответствует паре расчетных формул (6). При $p_1 = 0$ получаем $p_2 = 1$; $\alpha_2 = 1/2$, $\beta_{21} = 1/2$, что соответствует паре расчетных формул (7). В случае уравнения $y' = y$, согласно (9), имеем $\varphi'''(0) = y$ независимо от значений p_1 , p_2 , α_2 , β_{21} . Отсюда следует, что формул Рунге — Кутты со значениями $q = 2$ и $s = 3$ не существует.

Вследствие (8) для методов с $q = s = 2$ главная часть погрешности на шаге есть $\frac{\varphi'''(0)}{6} h^3$. Пользуясь явным выражением $\varphi'''(0)$, в ряде случаев удастся уменьшить эту величину за счет удачного подбора параметров метода. Если для рассматриваемого класса уравнений величина $f_y y''$ мала, то разумно распорядиться этими параметрами так, чтобы в (9) обращались в нуль первые три слагаемые. Нетрудно проверить, что при $p_1 = 1/4$, $p_2 = 3/4$, $\alpha_2 = \beta_{21} = 2/3$ выполняются равенства (10) и (11) и одновременно $\varphi'''(0) = f_y(x, y)y''(x)$. Этой совокупности параметров отвечают расчетные формулы

$$\begin{aligned} k_1 &= hf(x, y), \\ k_2 &= hf\left(x + \frac{2}{3}h, y + \frac{2}{3}k_1\right), \\ \Delta y &= z(h) - y = \frac{1}{4}(k_1 + 3k_2). \end{aligned} \quad (12)$$

Приведем результаты аналогичных рассмотрений для случая $q = 3$. Расчетных формул, соответствующих значению $s = 4$, не существует. Чтобы s равнялось трем, необходимо выполнение соотношений

$$\begin{aligned} \alpha_2 &= \beta_{21}, \quad \alpha_3 = \beta_{31} + \beta_{32}, \quad \alpha_3(\alpha_3 - \alpha_2) - \beta_{32}\alpha_2(2 - 3\alpha_2) = 0, \\ p_3\beta_{32}\alpha_2 &= \frac{1}{6}, \quad p_2\alpha_2 + p_3\alpha_3 = \frac{1}{2}, \quad p_1 + p_2 + p_3 = 1. \end{aligned}$$

Эта система шести уравнений с восемью неизвестными имеет бесчисленное множество решений. Наиболее употребительная совокупность расчетных формул:

$$\begin{aligned} k_1 &= hf(x, y), \quad k_2 = hf\left(x + \frac{h}{2}, y + \frac{k_1}{2}\right), \\ k_3 &= hf(x + h, y - k_1 + 2k_2), \quad \Delta y = \frac{1}{6}(k_1 + 4k_2 + k_3). \end{aligned}$$

Если правая часть $f(x, y)$ не зависит от y , иначе, $f_y \equiv 0$, то эта расчетная формула превращается в формулу Симпсона

$$y(x+h) - y(x) \approx \frac{h}{6} \left(f(x) + 4f\left(x + \frac{h}{2}\right) + f(x+h) \right). \quad (13)$$

Согласно оценке погрешности формулы Симпсона, погрешность этого приближенного равенства имеет порядок $O(h^5)$ и, следовательно, в этом частном случае порядок погрешности повышается. Поэтому аналогично (12) эту формулу разумно использовать при малых значениях f_y .

При $q = 4, 5$ не удается построить расчетных формул рассматриваемого вида со значением $s = 5$; при $q = 4$ имеется двухпараметрическое множество расчетных формул, соответствующих $s = 4$. Для примера приведем одно однопараметрическое семейство таких формул.

$$\begin{aligned} k_1 &= hf(x, y), \quad k_2 = hf\left(x + \frac{h}{2}, y + \frac{k_1}{2}\right), \\ k_3 &= hf\left(x + \frac{h}{2}, y + \left(\frac{1}{2} - \frac{1}{2t}\right)k_1 + \frac{1}{2t}k_2\right), \\ k_4 &= hf(x+h, y + (1-t)k_2 + tk_3), \\ \Delta y &= \frac{1}{6}(k_1 + (4-2t)k_2 + 2tk_3 + k_4). \end{aligned}$$

Наиболее употребительная совокупность формул этого семейства соответствует $t = 1$:

$$\begin{aligned} k_1 &= hf(x, y), \quad k_2 = hf\left(x + \frac{h}{2}, y + \frac{k_1}{2}\right), \\ k_3 &= hf\left(x + \frac{h}{2}, y + \frac{k_2}{2}\right), \quad k_4 = hf(x+h, y + k_3), \\ \Delta y &= \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4). \end{aligned} \quad (14)$$

Среди других совокупностей формул со значениями $q = s = 4$ отметим совокупность

$$\begin{aligned} k_1 &= hf(x, y), \quad k_2 = hf\left(x + \frac{h}{3}, y + \frac{k_1}{3}\right), \\ k_3 &= hf\left(x + \frac{2h}{3}, y - \frac{k_1}{3} + k_2\right), \quad k_4 = hf(x+h, y + k_1 - k_2 + k_3), \\ \Delta y &= \frac{1}{8}(k_1 + 3k_2 + 3k_3 + k_4). \end{aligned} \quad (15)$$

Мы использовали выше формулировку «наиболее употребительный». Эта формулировка отражает исторически сложившуюся тенденцию в использовании численных методов. Казалось бы, в руководстве по численным методам следовало не просто отражать тенденцию, а указать, какая формула из данного семейства расчетных формул является наилучшей. Однако ответ на такой вопрос не прост.

Обратимся сначала к задаче вычисления интеграла, где обстановка для сравнения методов более благоприятна. При одинаковом порядке погрешности квадратурных формул всегда оказывается, что главные члены погрешности этих формул пропорциональны. Например, при вычислении интеграла

$\int_0^1 f(x) dx$ с разбиением $[0, 1]$ на отрезки длины h погрешность формулы трапеций есть

$$\frac{(f'(1) - f'(0)) h^2}{12} + O(h^3),$$

а погрешность формулы прямоугольников

$$- \frac{(f'(1) - f'(0)) h^2}{24} + O(h^3);$$

при этом число точек, где вычисляются значения $f(x)$, примерно одинаково, соответственно, $1/h$ и $1/h$. Из сравнения главных членов можно сделать вывод о предпочтительности формулы прямоугольников. Однако этот вывод не является безусловным: в оценку остаточного члена входит член высшего порядка; поэтому на основании имеющейся информации мы имеем право делать вывод о преимуществе формулы прямоугольников лишь при очень малых значениях h . Выше мы сравнивали методы при их однократном использовании. В случае вычисления интегралов типична ситуация, когда для вычисления одного и того же интеграла проводятся расчеты с различными шагами. Узлы формулы трапеций с шагом $h' = hq^{-1}$, q — целое, являются узлами формулы трапеций с шагом h , и поэтому при интегрировании с шагом hq^{-1} могут быть использованы результаты вычислений с шагом h . В случае формулы прямоугольников такое использование возможно лишь при шаге $h' = hq^{-1}$, где q нечетно. Вследствие определенных удобств использования $q = 2$ преимущество формулы прямоугольников не является уже столь бесспорным.

В случае дифференциальных уравнений возникает ряд дополнительных моментов. У формул одинакового порядка точности по h главные члены погрешности на шаге часто оказываются непропорциональными. Например, вследствие (8), (9) главный член погрешности формулы (6) равен $(B - A)h^3$, где $B = \frac{1}{6} f_{yy''}$,

$$A = \frac{1}{12} (f_{xx} + 2f_{xy}y' + f_{yy}(y')^2),$$

а у формулы (7) равен $\left(B + \frac{A}{2}\right)h^3$. Поэтому для одних уравнений будет предпочтительнее один из методов (6), (7), для других — другой.

Рассмотрим пример. Пусть интегрируется уравнение $y' = y^q$, $q > 0$, при $y(0) > 1$. В этом случае главные члены погрешности формулы (6), (7), соответственно, равны

$$\frac{q(q+1)y^{3q-2}h^3}{12} \quad \text{и} \quad \frac{q(3q+1)y^{3q-2}h^3}{24};$$

при $q > 1$ главный член погрешности второй формулы оказывается больше, чем у первой, а при $q < 1$ меньше.

В подобной ситуации рекомендации в пользу того или иного метода должны основываться на «волевом решении», принятом с учетом статистических свойств рассматриваемых задач и традиций использования методов.

Приведем пример рассуждений, которые иногда применяются при сравнительной характеристике методов. Пусть решаются уравнения из некоторого множества дифференциальных уравнений и уравнения из этого множества рассматриваются как случайные. Предположим, что

$$\xi_1 = f_{xx} + 2f_{xy}y' + f_{yy}(y')^2 \quad \text{и} \quad \xi_2 = f_{yy}y''$$

являются независимыми случайными величинами с $M\xi_1 = M\xi_2 = 0$. Тогда математическое ожидание квадрата главного члена погрешности на шаге равно, соответственно,

$$\left(\frac{1}{144} M\xi_1^2 + \frac{1}{36} M\xi_2^2\right) h^6, \quad \left(\frac{1}{576} M\xi_1^2 + \frac{1}{36} M\xi_2^2\right) h^6.$$

При таком подходе формула (7) выглядит предпочтительнее по сравнению с формулой (6). Еще более предпочтительной в этом смысле следует считать формулу (12), где математическое ожидание квадрата главного члена погрешности равно $\frac{1}{35} M\xi_2^2 h^6$. Эти рассуждения опираются на гипотезу о случайности и независимости величин ξ_1 и ξ_2 , справедливость которой, вообще говоря, вызывает сомнения; поэтому к только что полученным выводам надо относиться с большой степенью осторожности.

Обратимся к формулам (14) и (15). В случае дифференциального уравнения $y' = f(x)$ погрешность формулы (14) есть

$$\rho_1 = y(x+h) - y(x) - \frac{h}{6} \left(f(x) + 4f\left(x + \frac{h}{2}\right) + f(x+h) \right).$$

Разлагая в ряд Тейлора, получим

$$\rho_1 = -\frac{f^{(4)}(x) h^5}{2880} + O(h^6).$$

В случае формулы (15), соответственно, имеем

$$\rho_2 = y(x+h) - y(x) - \frac{h}{8} \left(f(x) + 3f\left(x + \frac{h}{3}\right) + \right. \\ \left. + 3f\left(x + \frac{2}{3}h\right) + f(x+h) \right) = -\frac{f^{(4)}(x) h^5}{6480} + O(h^6).$$

Из сравнения главных членов ρ_1 и ρ_2 видно, что формула (15) имеет преимущество в случае уравнения $y' = f(x)$, иначе говоря для уравнений, где $f_y = 0$. Следовательно, эта формула имеет преимущество и для уравнений, где величина f_y близка к нулю.

Традиция применения методов рассматриваемого типа отдаст предпочтение формуле (14). К традиции применения вычислительных методов следует относиться двояко. Если исторически первый из методов рассматриваемого класса оказался приемлемым, то в дальнейшем пользователи привыкают к нему. Замена этого метода на другой, даже более эффективный метод, требует определенной психологической ломки у пользователей, определенных затрат времени на «привыкание» к новому методу. Чтобы широкий круг пользователей согласился на подобную перестройку, необходимо существенное преимущество нового метода по какой-либо из характеристик. В главе VII подобные вопросы уже возникали в связи с обсуждением проблемы внедрения математических методов решения экономических задач.

Часто оказывается, что указанная перестройка практики использования численных методов требует существенного времени. Численные методы

решения задач, соответственно, технологические процессы имеют устойчивую тенденцию к совершенствованию. Можно говорить о каком-то среднем проценте совершенствования методов за единицу времени (например, за год). Если предполагаемая перестройка обеспечит меньший эффект, но в то же время может помешать последующему применению новых методов или новой техники, то иногда следует от нее воздержаться. Как и в случае проблем планирования, переход к новым методам может встретиться с определенными трудностями, которые не были предусмотрены при первоначальном теоретическом анализе. Может оказаться, что метод с меньшей погрешностью на шаге имеет большую суммарную погрешность метода или более чувствителен к вычислительной погрешности. При сравнении методов интегрирования (6), (7), (12) мы отдали бы предпочтение способу интегрирования (12), ориентируясь на высказанные выше, вообще говоря, сомнительные соображения о величине квадрата математического ожидания погрешности этих формул. В то же время традиция использования методов интегрирования второго порядка точности отдает предпочтение методу (6) с большей величиной главного члена математического ожидания квадрата погрешности.

Проведенное нами сравнение двух формул четвертого порядка точности указывает на предпочтительность формулы (15) при малом значении hf_v . В то же время можно встретить такое высказывание: преимущество формулы (15) сказывается лишь при очень малых значениях hf_v ; если уравнение интегрируется с умеренной точностью, практика вычислений показывает предпочтительность употребления формулы (14). Понятие практики вычислительной работы является довольно неопределенным. Число различных классов реально встречающихся дифференциальных уравнений существенно превосходит число задач, где производится сравнение методов численного интегрирования; поэтому суждения «с позиций практики» не всегда объективны. Однако несмотря на такую неопределенность, критерий практики часто несет в себе определенную положительную информацию, которая зачастую на данном этапе развития науки не может быть формализована или обоснована. Есть определенные математические соображения в пользу того, что расчетные формулы (14) более предпочтительны при $f_v < 0$ и гладких решениях, поскольку они допускают интегрирование с более крупным шагом без потери слабой чувствительности метода к влиянию вычислительной погрешности; отметим, однако, что эти соображения появились позже, чем практика численного интегрирования обыкновенных дифференциальных уравнений сделала определенный выбор в пользу формул (14).

В ситуации, подобной рассматриваемой, обычно не удается предложить полностью обоснованное (с позиций математика) решение возникающей проблемы. В данном случае сложность задачи усугубляется невозможностью дать удовлетворительное описание класса дифференциальных уравнений, предъявляемых к решению.

При дальнейшем рассмотрении для нас будет существенно, что погрешность метода на шаге $\varphi(h)$ имеет главный член, а именно, справедливо представление вида

$$\varphi(h) = \psi(x, y)h^{s+1} + O(h^{s+2}). \quad (16)$$

Наметим основные этапы доказательства этого соотношения. Предположим, что правая часть и все ее производные до порядка $s \pm 1$ включительно ограничены равномерно в области G :

$$x_0 \leq x \leq x_0 + X.$$

Тогда также будут равномерно ограничены производные всех решений уравнения до порядка $s+2$ включительно. Согласно формуле Тейлора, равенство (8) можно записать в уточненной форме

$$\varphi(h) = \frac{\varphi^{(s+1)}(0)}{(s+1)!} h^{s+1} + \frac{\varphi^{(s+2)}(\theta h)}{(s+2)!} h^{s+2}.$$

Имеем равенство

$$\varphi^{(s+1)}(0) = y^{(s+1)}(0) - z^{(s+1)}(0).$$

Нетрудно убедиться, что обе величины $y^{(s+1)}(0)$ и $z^{(s+1)}(0)$ явно выражаются через значения в точке (x, y) функции f и ее производных порядка не выше s ; примеры таких явных выражений мы уже получали.

Поскольку правая часть дифференцируема $s+1$ раз, то отсюда следует, что функция $\psi(x, y)$ дифференцируема в области G и ее производные ψ_x и ψ_y равномерно ограничены в этой области. Аналогично устанавливается, что величина $\varphi^{(s+2)}(\theta h)$ равномерно ограничена при $x_0 \leq x < x+h \leq x_0+X$. Некоторые частные случаи этого утверждения также следуют из выписанных ранее соотношений.

§ 3. Методы с контролем погрешности на шаге

В главе III при оптимизации алгоритма численного интегрирования возникало требование распределить узлы интегрирования так, чтобы на каждый элементарный отрезок интегрирования приходилась примерно одинаковая погрешность; для этой цели нужно иметь представление о величине погрешности на каждом элементарном отрезке интегрирования. Точно такая же задача возникает и при численном интегрировании обыкновенных дифференциальных уравнений.

При ее решении можно рассуждать следующим образом. Главный член погрешности на шаге интегрирования есть

$$\frac{\varphi^{(s+1)}(0) h^{s+1}}{(s+1)!}.$$

Точка $(x+h, z(h))$ находится близко от точки (x, y) ; поэтому погрешность на следующем шаге интегрирования будет иметь такой же главный член. В результате двух шагов будет получено приближение $y^{(1)}$ к значению $y(x+2h)$ такое, что

$$y^{(1)} - y(x+2h) \sim 2 \frac{\varphi^{(s+1)}(0) h^{s+1}}{(s+1)!}.$$

Если, исходя из точки (x, y) , применить метод Рунге — Кутты с шагом $2h$, то получится приближенное значение $y^{(2)}$, для

которого

$$y^{(2)} - y(x + 2h) \sim \frac{\Phi^{(s+1)}(0)(2h)^{s+1}}{(s+1)!}.$$

Из этих соотношений вытекает представление главного члена погрешности на шаге

$$y^{(1)} - y(x + 2h) \sim \frac{y^{(2)} - y^{(1)}}{2^s - 1}.$$

При желании можно уточнить полученное приближенное значение $y^{(1)}$, прибавив к нему величину главного члена погрешности, т. е. положить

$$y(x + 2h) \approx y^{(1)} + \frac{y^{(1)} - y^{(2)}}{2^s - 1}. \quad (1)$$

В результате этих рассуждений мы имеем возможность через каждые два шага, получая приближенное значение $y^{(1)}$, иметь также выражение главного члена его погрешности. Если характер поведения решения меняется резко, то часто возникает следующая ситуация. После вычисления величины $y^{(1)} - y^{(2)}$ обнаруживается, что погрешность на шаге очень велика, и поэтому приходится отказываться от части проведенных вычислений и изменять шаг.

Для более гибкого управления выбором шага интегрирования иногда желательно иметь возможность совершать шаг интегрирования и оценивать погрешность при меньшем количестве вычисляемых значений правых частей.

Примером совокупностей формул, с теми же характеристиками точности при меньшем числе обращений к правой части, может служить совокупность формул

$$\begin{aligned} k_1 &= hf(x, y), \quad k_2 = hf\left(x + \frac{h}{2}, y + \frac{k_1}{2}\right), \\ k_3 &= hf\left(x + \frac{h}{2}, y + \frac{1}{4}(k_1 + k_2)\right), \quad k_4 = hf\left(x + h, y - k_2 + 2k_3\right), \\ k_5 &= hf\left(x + \frac{2h}{3}, y + \frac{1}{27}(7k_1 + 10k_2 + k_4)\right), \\ k_6 &= hf\left(x + \frac{h}{5}, y + \frac{1}{625}(28k_1 - 125k_2 + 546k_3 + 54k_4 - 378k_5)\right), \\ \Delta y &= \frac{1}{6}(k_1 + 4k_3 + k_4) \end{aligned} \quad (2)$$

с главным членом погрешности

$$\begin{aligned} y(x + h) - z(h) &= r + O(h^6), \\ r &= \frac{1}{336}(42k_1 + 224k_3 + 21k_4 - 162k_5 - 125k_6). \end{aligned}$$

Заметим, что значение $z(h)$ вычисляется по одной из формул Рунге — Кутты, указанных на стр. 455.

Если положить $y(x_0 + h) \approx z(h) + r$, то получим метод рассматриваемого типа со значением $s = 5$ и, соответственно, погрешностью на шаге порядка $O(h^6)$; при описанном выше непосредственном применении метода Рунге — Кутты и оценке погрешности через величину $y^{(1)} - y^{(2)}$ требуется не 6, а 11 вычислений правой части.

§ 4. Оценка погрешности одношаговых методов

Рассмотрим множество всевозможных методов интегрирования, где последовательно получаются приближения y_j к значениям $y(x_j)$, $x_0 < x_1 < \dots < x_N = x_0 + X$. Пусть в процессе численного интегрирования k фиксировано и при $j \geq k$ значения y_j определяются как значения некоторого функционала

$$y_j = F(j; x_j, \dots, x_{j-k}; y_{j-1}, \dots, y_{j-k}). \quad (1)$$

Такой способ численного интегрирования называют *k-шаговым*. Все рассматривавшиеся выше способы интегрирования имеют следующее общее свойство: приближенное значение решения в следующей точке определялось только в зависимости от значения решения в предыдущей точке; следовательно, расчетные формулы, соответствующие этим способам, представимы в виде (1) со значением $k = 1$; такие методы называют *одношаговыми*.

На первый взгляд может показаться, что к классу формул (1) с $k = 1$ не относится расчетная формула (3.1), поскольку приближение к $y(x + 2h)$ вычисляется через приближения к $y(x)$ и к $y(x + h)$. Однако само приближение к $y(x + h)$ является функцией от $y(x)$. Поэтому в конечном счете эта расчетная формула может быть представлена в виде

$$y(x + 2h) \approx \Phi(f, x, 2h, y).$$

Если при $x_j = x$ мы примем $x_{j+1} = x + 2h$, то получим частный случай формулы (1) при $k = 1$.

Рассмотрим специальный способ получения оценки погрешности, применимый лишь к одношаговым методам.

При $k = 1$ иногда удобнее записывать формулу (1) в виде

$$y_{j+1} = F(f, x_j, x_{j+1} - x_j, y_j). \quad (2)$$

Получаемые в процессе реальных вычислений приближения к значениям $y(x_j)$ связаны не соотношениями (2), а некоторыми соотношениями

$$y_{j+1} = F(f, x_j, x_{j+1} - x_j, y_j) + \delta_{j+1}. \quad (3)$$

Наличие слагаемого δ_{j+1} обусловлено следующими причинами:

1) округлением чисел при вычислениях;

2) ошибками в значениях правой части $f(x, y)$; эти ошибки вызваны тем, что рассматриваемая нами функция $f(x, y)$ является некоторым приближением к правой части реального дифференциального уравнения; кроме того, зачастую в процессе вычисления значений $f(x, y)$ в машине эта функция приближается функциями, вычислимыми относительно элементарных операций ЭВМ:

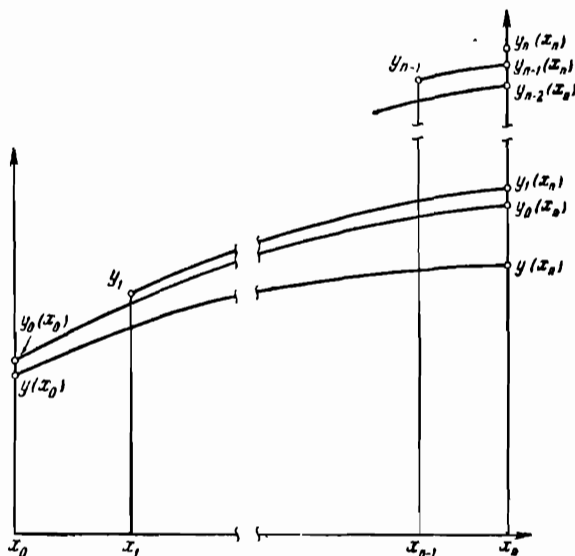


Рис. 8.4.1.

3) в некоторых случаях значение y_{j+1} определяется из уравнения, эквивалентного (1), но неразрешенного в явном виде относительно переменной y_{j+1} ; тогда величина δ_{j+1} содержит составляющую, являющуюся следствием приближенного решения этого уравнения.

Довольно часто бывает справедлива оценка

$$|\delta_{j+1}| \leq C(|y_j| + 1)2^{-t},$$

где t — разрядность чисел в машине, а постоянная C не очень большая.

Точно так же начальное условие y_0 отличается от значения отыскиваемого решения задачи $y(x_0)$ вследствие ошибки в определении исходных данных и округлений. Пусть $y(x)$ — искомое решение дифференциального уравнения, а $y_j(x)$ — решения, удовлетворяющие условиям $y_j(x_j) = y_j$ (рис. 8.4.1).

Ошибку $R_n = y_n - y(x_n)$ можно представить в виде

$$R_n = y_n(x_n) - y_0(x_n) + y_0(x_n) - y(x_n) = \\ = \sum_{j=1}^n (y_j(x_n) - y_{j-1}(x_n)) + (y_0(x_n) - y(x_n)). \quad (4)$$

Разность решений дифференциального уравнения в одной точке может быть выражена через разность их в другой точке следующим путем.

Лемма. Пусть $Y_1(x)$ и $Y_2(x)$ — решения дифференциального уравнения $y' = f(x, y)$. Тогда

$$Y_2(\beta) - Y_1(\beta) = (Y_2(\alpha) - Y_1(\alpha)) \exp \left(\int_{\alpha}^{\beta} f_y(x, \tilde{y}(x)) dx \right), \quad (5)$$

где $\tilde{y}(x)$ заключено между $Y_1(x)$ и $Y_2(x)$.

Доказательство. Вычтем друг из друга равенства

$$Y_2' = f(x, Y_2), \quad Y_1' = f(x, Y_1).$$

Согласно формуле Лагранжа, разность $f(x, Y_2) - f(x, Y_1)$ может быть представлена в виде $f_y(x, \tilde{y})(Y_2 - Y_1)$, где $\tilde{y}$ заключено между Y_1 и Y_2 . В результате получится линейное дифференциальное уравнение относительно $Y_2 - Y_1$:

$$(Y_2 - Y_1)' = f_y(x, \tilde{y})(Y_2 - Y_1),$$

откуда

$$(\ln(Y_2 - Y_1))' = f_y(x, \tilde{y}).$$

Подставляя пределы α и β и потенцируя, получим требуемое соотношение (5). Функция

$$f_y(x, \tilde{y}(x)) = \frac{f(x, Y_2(x)) - f(x, Y_1(x))}{Y_2(x) - Y_1(x)}$$

непрерывна при непрерывной $f(x, y)$, и поэтому все проведенные выше преобразования являются законными.

Пусть $\alpha = x_j$, $\beta = x_n$,

$$Y_1(x) = y_{j-1}(x), \quad Y_2(x) = y_j(x),$$

тогда

$$y_j(x_n) - y_{j-1}(x_n) = (y_j(x_j) - y_{j-1}(x_j)) \exp \left(\int_{x_j}^{x_n} f_y(x, \tilde{y}_j(x)) dx \right),$$

где $\tilde{y}_j(x)$ заключено между $y_{j-1}(x)$ и $y_j(x)$.

Точно так же

$$y_0(x_n) - y(x_n) = (y_0(x_0) - y(x_0)) \exp \left(\int_{x_0}^{x_n} f_y(x, \tilde{y}_0(x)) dx \right).$$

Теперь равенство (4) можно записать в виде

$$R_n = \sum_{j=1}^n \omega_j \exp \left(\int_{x_j}^{x_n} f_y(x, \tilde{y}_j(x)) dx \right) + R_0 \exp \left(\int_{x_0}^{x_n} f_y(x, \tilde{y}_0(x)) dx \right), \quad (6)$$

где $\omega_j = y_j(x_j) - y_{j-1}(x_j)$, $j = 1, \dots$

Из (3) вытекает соотношение

$$\omega_j = y_j - y_{j-1}(x_j) = \rho_j + \delta_j,$$

где

$$\rho_j = F(f, x_{j-1}, x_j - x_{j-1}, y_{j-1}) - y_{j-1}(x_j).$$

Посмотрим, какой смысл имеет величина ρ_j ; $F(f, x_{j-1}, x_j - x_{j-1}, y_{j-1})$ есть число, получаемое в результате вычислений по расчетной формуле (2), $y_{j-1}(x_j)$ — значение в точке x_j точного решения дифференциального уравнения, удовлетворяющего условию $y_{j-1}(x_{j-1}) = y_{j-1}$. Таким образом, ρ_j есть ошибка одного шага рассматриваемого метода, если вычисления начинаются с точки (x_{j-1}, y_{j-1}) и производятся без округлений, а шаг равен $x_j - x_{j-1}$.

Предположим, что при всех j , соответствующих рассматриваемому отрезку интегрирования $x_0 < x_j \leq x_0 + X$, выполняется неравенство

$$|\rho_j| \leq C_1 (x_j - x_{j-1})^{s+1}. \quad (7)$$

Пусть

$$L = \sup_{x_0 \leq x \leq x_0 + X} |f_y| < \infty$$

и

$$H = \max_{0 \leq j \leq N} (x_j - x_{j-1}).$$

Загрубляя (7), имеем

$$|\rho_j| \leq C_1 H^s (x_j - x_{j-1}). \quad (8)$$

При $x_0 \leq x_j \leq x_n \leq x_0 + X$ справедливы неравенства

$$\exp \left(\int_{x_j}^{x_n} f_y(x, \tilde{y}_j(x)) dx \right) \leq \exp(L(x_n - x_j)) \leq \exp(LX).$$

Воспользуемся этим неравенством для оценки правой части (6); получим

$$|R_n| \leq \exp(LX) \left(\sum_{j=1}^n (|\rho_j| + |\delta_j|) + |R_0| \right).$$

Применим теперь оценку (8):

$$\begin{aligned} |R_n| &\leq \exp(LX) \left(\sum_{j=1}^n (C_1 H^s (x_j - x_{j-1}) + |\delta_j|) + |R_0| \right) \leq \\ &\leq \exp(LX) (C_1 H^s (x_n - x_0) + n\delta + |R_0|) \leq \\ &\leq \exp(LX) (C_1 X H^s + N\delta + |R_0|); \quad (9) \end{aligned}$$

здесь $\delta = \max_j |\delta_j|$. Из этого соотношения следует, что

$$\max_{x_0 < x_n \leq x_0 + X} |R_n| \rightarrow 0 \quad \text{при} \quad H \rightarrow 0,$$

если одновременно $N\delta \rightarrow 0$, $|R_0| \rightarrow 0$. Мы показали, что при достаточно мелком шаге интегрирования и малой вычислительной погрешности приближенное решение, получаемое при употреблении метода Рунге — Кутты, будет близко к точному решению.

В ряде случаев решение дифференциального уравнения отыскивается на большом промежутке. Тогда в полученные нами оценки погрешности входит как множитель очень большое число $\exp(LX)$. Поэтому при LX большом может оказаться, что достижение требуемой точности требует столь мелких шагов и столь малой величины вычислительной погрешности на шаге, что это может поставить под сомнение целесообразность использования рассматриваемого метода. Поэтому характеристика методов только по признаку — сходится ли приближенное решение к точному при мельчении шага и при достаточно быстром уменьшении погрешности или не сходится — является еще недостаточной.

Если $f_y(x, y) \leq -b < 0$, то в оценке (9) можно избавиться от множителя, резко растущего с X . Рассмотрим случай постоянного шага $x_j - x_{j-1} \equiv h$. Тогда -

$$\exp \left(\int_{x_j}^{x_n} f_y(x, \tilde{y}(x)) dx \right) \leq \exp(-b(x_n - x_j)) \leq \exp(-b(n-j)h);$$

пусть

$$|\omega_j| \leq C_1 h^{s+1} + \delta.$$

Оценивая правую часть (6), получаем

$$|R_n| \leq \sum_{j=1}^n (C_1 h^{s+1} + \delta) \exp(-b(n-j)h) + |R_0| \exp(-bnh). \quad (10)$$

Имеем

$$\sum_{j=1}^n \exp(-b(n-j)h) \leq \sum_{k=0}^{\infty} \exp(-bkh) = \frac{1}{1 - \exp(-bh)}.$$

Таким образом, получаем окончательную оценку погрешности

$$|R_n| \leq \frac{C_1 h^{s+1} + \delta}{1 - \exp(-bh)} + |R_0| \exp(-bnh). \quad (11)$$

Поскольку $1 - \exp(-bh) \sim |b|h$, то также имеем

$$|R_n| \leq C_2 \left(h^s + \frac{\delta}{h} \right) + |R_0| \exp(-bnh).$$

Формально эта оценка не зависит от длины промежутка интегрирования X , однако длина промежутка интегрирования может неявно влиять на значение коэффициента C_2 через оценки производных. Мы еще вернемся к этому вопросу в общих замечаниях в § 12.

Задача. Пусть в области G равномерно ограничены f и все ее производные до порядка $s+1$ включительно; пусть

$$|R_0| = o(h^s) \quad \text{и} \quad \delta = \max_{x_0 \leq x_j \leq x_0 + X} |\delta_j| = o(h^s).$$

Доказать, что в случае интегрирования по методу Рунге—Кутты порядка точности s выполняется равенство

$$x_0 \leq x_n \leq x_0 + X \quad \left| R_n - h^s \int_{x_0}^{x_n} \psi(x, y(x)) \exp \left(\int_x^{x_n} f_y(t, y(t)) dt \right) dx \right| = o(h^s); \quad (12)$$

определение функции $\psi(x, y)$ см. (2.16). Соотношение (12) дает представление о главном члене погрешности метода Рунге—Кутта.

§ 5. Конечно-разностные методы

Среди k -шаговых методов наиболее употребительны методы интегрирования на сетке с постоянным шагом $x_j - x_{j-1} \equiv h = \text{const}$ при помощи соотношений вида

$$\sum_{i=0}^k a_i y_{j-i} - h \sum_{i=0}^k b_i (x_{j-i}, y_{j-i}) = 0; \quad (1)$$

здесь a_i — постоянные,

$$\left| \frac{\partial b_i}{\partial y_{j-i}} \right| \leq C \quad \text{при} \quad h \leq h_0.$$

В свою очередь среди таких методов традиционно наиболее употребительны методы

$$\sum_{i=0}^k a_{-i} y_{i-1} - h \sum_{i=0}^k b_{-i} f(x_{i-1}, y_{i-1}) = 0, \quad (2)$$

которые принято называть *конечно-разностными методами* или *конечно-разностными схемами*.

Простейшие и в то же время наиболее употребительные способы (2) получаются на основе квадратурных формул; всякая квадратурная формула

$$\int_{-ph}^0 f(x) dx \approx h \sum_{i=0}^m b_{-i} f(-ih), \quad m \geq 0, \quad (3)$$

порождает соответствующую формулу численного интегрирования обыкновенных дифференциальных уравнений. Действительно, подставляя в (3) $f(x) = y'(x_j + x)$, имеем

$$\int_{-ph}^0 f(x) dx = y(x_j) - y(x_{j-p}) \approx h \sum_{i=0}^m b_{-i} y'(x_{j-i}),$$

заменяя $y'(x)$ на $f(x, y(x))$, получим

$$y(x_j) - y(x_{j-p}) \approx h \sum_{i=0}^m b_{-i} f(x_{j-i}, y(x_{j-i})). \quad (4)$$

Соответствующая конечно-разностная схема:

$$y_j - y_{j-p} - h \sum_{i=0}^m b_{-i} f(x_{j-i}, y_{j-i}) = 0. \quad (5)$$

В вычислительной практике применяются формулы (1), (2) со значениями $a_0 \neq 0$, $b_0 = 0$, *явные*, или *экстраполяционные*, и формулы с $a_0 \neq 0$, $b_0 \neq 0$, *неявные*, или *интерполяционные*.

Формулы (1), (2) при $a_0 = 0$, $b_0 \neq 0$ называют *формулами с забеганием вперед*; в дальнейшем предполагается $a_0 \neq 0$.

Приведем ряд примеров формул (5); для краткости обозначаем

$$f_n = f(x_n, y_n).$$

Формуле трапеций

$$\int_{-h}^0 f(x) dx \approx \frac{h}{2} (f(0) + f(-h))$$

соответствует интерполяционная формула

$$y_I - y_{I-1} = \frac{h}{2} (f_I + f_{I-1}), \quad (6)$$

формуле Симпсона

$$\int_{-2h}^0 f(x) dx \approx \frac{h}{3} (f(0) + 4f(-h) + f(-2h))$$

— интерполяционная формула

$$y_I - y_{I-2} = \frac{h}{3} (f_I + 4f_{I-1} + f_{I-2}), \quad (7)$$

формуле

$$\int_{-3h}^0 f(x) dx \approx \frac{h}{8} (3f(0) + 9f(-h) + 9f(-2h) + 3f(-3h))$$

— интерполяционная формула

$$y_I - y_{I-3} = \frac{h}{8} (3f_I + 9f_{I-1} + 9f_{I-2} + 3f_{I-3}). \quad (8)$$

Формула прямоугольников

$$\int_{-h}^0 f(x) dx \approx hf\left(-\frac{h}{2}\right)$$

имеет вид, отличный от (3). Однако, заменяя h на $2h$, получим квадратуру

$$\int_{-2h}^0 f(x) dx \approx 2hf(-h)$$

и соответствующую экстраполяционную формулу

$$y_I - y_{I-2} = 2hf_{I-1}. \quad (9)$$

Если остаточный член квадратуры (3) оценивается через $D(q) \max |f^{(q)}(x)| h^{q+1}$, то погрешность равенства (4) будет оцениваться через $D(q) \max |y^{(q+1)}(x)| h^{q+1}$.

Среди формул вида (5) наиболее употребительными являются формулы, соответствующие $p = 1$; их называют *формулами Адамса*. Получим эти формулы вместе с оценкой остаточного члена. Согласно формуле Ньютона, имеем

$$f(x) = Q_m(x) + r_m(x),$$

где $Q_m(x)$ — интерполяционный многочлен с узлами интерполяции $(0, \dots, -mh)$

$$Q_m(x) = \sum_{i=0}^m \frac{\nabla^i f(0)}{h^i} \omega_{hi}(x), \quad \omega_{h0} = 1, \quad (10)$$

$$\omega_{hi} = \frac{x(x+h) \dots (x+(i-1)h)}{i!} \quad \text{при } i > 0,$$

$$r_m(x) = f(0, -h, \dots, -mh, x) (m+1)! \omega_{h, m+1}.$$

Соответственно, справедливо равенство

$$\int_0^h f(x) dx = \int_0^h Q_m(x) dx + \int_0^h r_m(x) dx. \quad (11)$$

Функция $\omega_{h, m+1}(x)$ не меняет знака на отрезке $[0, h]$. Поэтому на основании теоремы о среднем имеем равенство

$$\int_0^h r_m(x) dx = f(0; -h; \dots; -mh; \xi) (m+1)! \int_0^h \omega_{h, m+1}(x) dx,$$

где $0 \leq \xi \leq h$. Производя в интеграле $\int_0^h \omega_{hi}(x) dx$ замену переменной $x = ht$, получим

$$\int_0^h \omega_{hi}(x) dx = \gamma_i h^{i+1}, \quad (12)$$

где

$$\gamma_i = \int_0^1 \frac{t(t+1) \dots (t+i-1)}{i!} dt.$$

С учетом этих соотношений (11) можно переписать в виде

$$\int_0^h f(x) dx = h \sum_{i=0}^m \gamma_i \nabla^i f(0) + \gamma_{m+1} h^{m+2} (m+1)! f(0; -h; \dots; -mh; \xi).$$

Заменяя $(m+1)! f(0; -h; \dots; -mh; \xi)$ значением производной в средней точке, согласно (2.5.4), получим

$$\int_0^h f(x) dx = h \sum_{i=0}^m \gamma_i \nabla^i f(0) + \gamma_{m+1} h^{m+2} f^{(m+1)}(\eta), \quad (13)$$

где $-mh \leq \eta \leq h$. После замены $f(x)$ на $y'(x_1 + x - h)$ и отбрасывания остаточного члена получаем соответствующую

расчетную формулу

$$y_l - y_{l-1} = h \sum_{i=0}^m \gamma_i \nabla^i f_{l-1}. \quad (14)$$

Для получения интерполяционных формул Адамса в качестве узлов интерполяции возьмем точки $h, 0, \dots, -(m-1)h$; получим

$$f(x) = \bar{Q}_m(x) + \bar{r}_m(x),$$

где

$$\bar{Q}_m(x) = \sum_{i=0}^m \frac{\nabla^i f(h)}{h^i} \bar{\omega}_{hi}(x),$$

$$\bar{\omega}_{hi}(x) = \frac{(x-h)x(x+h)\dots(x+(i-2)h)}{i!},$$

$$\bar{r}_m(x) = f(h; 0; \dots; -(m-1)h; x)(m+1)! \bar{\omega}_{h, m+1}(x).$$

После аналогичных преобразований получаем равенство

$$\int_0^h f(x) dx = h \sum_{i=0}^m \bar{\gamma}_i \nabla^i f(h) + \bar{\gamma}_{m+1} h^{m+2} f^{(m+1)}(\eta), \quad (15)$$

где $-(m-1)h \leq \eta \leq h$, $\bar{\gamma}_i = \int_0^1 \frac{(t-1)t(t+1)\dots(t+i-2)}{i!} dt$.

Соответствующая расчетная формула

$$y_l - y_{l-1} = h \sum_{i=0}^m \bar{\gamma}_i \nabla^i f_l. \quad (16)$$

Исследуем поведение коэффициентов γ_i и $\bar{\gamma}_i$.

Очевидно, что $\gamma_0, \bar{\gamma}_0 > 0$ и $\gamma_i > 0, \bar{\gamma}_i < 0$ при $i > 0$. После замены переменной $t = 1 - u$ можно записать γ_i в виде

$$\gamma_i = \int_0^1 \frac{(1-u)\dots(l-u)}{i!} du = \int_0^1 \prod_{k=1}^i \left(1 - \frac{u}{k}\right) du.$$

Имеем равенство

$$1 - \frac{u}{k} = \exp\left(\ln\left(1 - \frac{u}{k}\right)\right) = \exp\left(-\frac{u}{k} + e_k\right),$$

где

$$0 \leq e_k < \frac{u^2}{2k^2}.$$

Таким образом,

$$\prod_{k=1}^i \left(1 - \frac{u}{k}\right) = \exp(-s(i)u) \exp\left(\sum_{k=1}^i e_k\right), \quad s(i) = \sum_{k=1}^i \frac{1}{k}.$$

Очевидно,

$$0 \leq \sum_{k=1}^i e_k < \sum_{k=1}^{\infty} \frac{1}{2k^2} = \frac{\pi^2}{12}.$$

В результате имеем оценку

$$v_i(u) \leq \prod_{k=1}^i \left(1 - \frac{u}{k}\right) \leq \exp\left(-\frac{\pi^2}{12}\right) v_i(u),$$

где $v_i(u) = \exp(-s(i)u)$, и, следовательно,

$$\int_0^1 v_i(u) du \leq \gamma_i \leq \exp\left(-\frac{\pi^2}{12}\right) \int_0^1 v_i(u) du. \quad (17)$$

Вспоминая, что $s(i) \sim \ln i$, получаем

$$\int_0^1 v_i(u) du = \frac{1 - \exp(-s(i))}{s(i)} \sim \frac{1}{\ln i}.$$

Отсюда и из (17) вытекает, что γ_i — величина порядка $\frac{1}{\ln i}$.

Справедливо очевидное тождество

$$(t-1)t \dots (t+i-2) + it(t+1) \dots (t+i-2) = t \dots (t+i-1).$$

Разделив его на $i!$ и проинтегрировав по t в пределах от 0 до 1, получим равенство

$$\bar{\gamma}_i + \gamma_{i-1} = \gamma_i;$$

поскольку все $\gamma_i > 0$ и $\bar{\gamma}_i < 0$ при $i > 0$, то отсюда заключаем, что γ_i монотонно убывают и всегда $|\bar{\gamma}_i| < \gamma_i$.

Таким образом, постоянный множитель в оценке погрешности на шаге у интерполяционных методов Адамса всегда меньше, чем у экстраполяционных того же порядка точности.

Для информации укажем значения

$$\{\gamma_p\} = 1, \frac{1}{2}, \frac{5}{12}, \frac{3}{8}, \frac{251}{720}, \dots, \{\bar{\gamma}_p\} = 1, -\frac{1}{2}, -\frac{1}{12}, -\frac{1}{24}, \dots$$

Задача. Показать, что $\gamma_i \sim \frac{\text{const}}{\ln i}$, $\bar{\gamma}_i \sim \frac{\text{const}}{i \ln i}$ при $i \rightarrow \infty$.

§ 6. Метод неопределенных коэффициентов

Для построения формул численного интегрирования можно также использовать метод неопределенных коэффициентов. Приблизим производную $y'(x)$ некоторым выражением

$$\sum_{i=0}^k \frac{a_{-i} y(x_{i-i})}{h},$$

а значение $f(x_l, y(x_l))$ выражением

$$\sum_{i=0}^k b_{-i} f(x_{l-i}, y(x_{l-i}))$$

(предполагается, что a_{-i} и b_{-i} не зависят от h). Отсюда получаем приближенное равенство

$$\sum_{i=0}^k \frac{a_{-i} y(x_{l-i})}{h} \approx \sum_{i=0}^k b_{-i} f(x_{l-i}, y(x_{l-i})). \quad (1)$$

Этому соотношению соответствует конечно-разностная схема

$$\sum_{i=0}^k \frac{a_{-i} y_{l-i}}{h} - \sum_{i=0}^k b_{-i} f_{l-i} = 0. \quad (2)$$

Величина

$$r_l = \sum_{i=0}^k \frac{a_{-i} y(x_{l-i})}{h} - \sum_{i=0}^k b_{-i} f(x_{l-i}, y(x_{l-i}))$$

называется *погрешностью аппроксимации исходного дифференциального уравнения схемой (2)*.

Определение. Говорят, что *разностная задача аппроксимирует дифференциальную* на отрезке $[x_0, x_0 + X]$, если

$$\max_{x_0 \leq x_l \leq x_0 + X} |r_l| \rightarrow 0 \quad \text{при} \quad h \rightarrow 0.$$

Вспомогая, что $f(x_{l-i}, y(x_{l-i})) = y'(x_{l-i})$ и $x_{l-i} = x_l - ih$, получим

$$r_l = \sum_{i=0}^k \frac{a_{-i} y(x_l - ih)}{h} - \sum_{i=0}^k b_{-i} y'(x_l - ih).$$

Предположим, что все производные решения до порядка q ограничены:

$$|y^{(p)}(x)| \leq M_p < \infty \quad \text{при} \quad x_0 \leq x \leq x_0 + X, \quad p = 0, \dots, q.$$

Представим все величины

$$y(x_l - ih) \quad \text{и} \quad y'(x_l - ih)$$

с помощью формулы Тейлора

$$y(x_l - ih) = \sum_{p=0}^{q-1} y^{(p)}(x_l) \frac{(-ih)^p}{p!} + \beta'_l,$$

$$y'(x_l - ih) = \sum_{p=1}^{q-1} y^{(p)}(x_l) \frac{(-ih)^{p-1}}{(p-1)!} + \gamma'_l;$$

согласно оценке остаточного члена ряда Тейлора, имеем

$$\begin{aligned} |\beta_j^i| &\leq M_q (ih)^q / q!, \\ |\gamma_j^i| &\leq M_q (ih)^{q-1} / (q-1)! \end{aligned}$$

Подставим выражения $y(x_j - ih)$ и $y'(x_j - ih)$ в правую часть представления r_j и соберем коэффициенты при $y^{(p)}(x_j)$, получим

$$r_j = E_0 h^{-1} y(x_j) + E_1 y'(x_j) + \dots + E_{q-1} h^{q-2} y^{(q-1)}(x_j) + \epsilon_j, \quad (3)$$

где

$$\begin{aligned} E_0 &= \sum_{i=0}^k a_{-i}, \\ E_p &= \sum_{i=0}^k \frac{a_{-i} (-i)^p}{p!} - \sum_{i=0}^k \frac{b_{-i} (-i)^{p-1}}{(p-1)!} \end{aligned} \quad (4)$$

при $p > 0$,

$$\epsilon_j = \sum_{i=0}^k \frac{a_{-i} \beta_j^i}{h} - \sum_{i=0}^k b_{-i} \gamma_j^i.$$

Очевидно,

$$|\epsilon_j| \leq D_q M_q h^{q-1},$$

где

$$D_q = \sum_{i=0}^k |a_{-i}| \frac{i^q}{q!} + \sum_{i=0}^k |b_{-i}| \frac{i^{q-1}}{(q-1)!}.$$

В ряде случаев, например, для рассмотренных в § 5 методов Адамса значение постоянной D_q в оценке $|\epsilon_j|$ удастся уменьшить. Если $E_0 = \dots = E_m = 0$, то $r_j = O(h^m)$, и говорят, что схема (2) имеет m -й порядок аппроксимации. Очевидно, всякая схема m -го порядка аппроксимации является схемой q -го порядка аппроксимации, если $q < m$. Если $E_0 = \dots = E_m = 0$, а $E_{m+1} \neq 0$, то говорят, что порядок аппроксимации строго равен m . При домножении (2) на произвольную постоянную величину r_j умножается на ту же постоянную. Поэтому принятое нами определение погрешности аппроксимации не является однозначным. Оно станет таковым, если потребовать дополнительных условий: для любой гладкой $y(x)$

$$\begin{aligned} \lim_{h \rightarrow 0} \sum_{i=0}^k \frac{a_{-i} y(x - ih)}{h} &= y'(x), \\ \lim_{h \rightarrow 0} \sum_{i=0}^k b_{-i} f(x - ih, y(x - ih)) &= f(x, y(x)). \end{aligned} \quad (5)$$

Лемма. Условия (5) выполнены тогда и только тогда, когда

$$E_0 = E_1 = 0, \quad b_0 + \dots + b_{-k} = 1. \quad (6)$$

Разложим выражения, стоящие под знаками пределов, в ряд Тейлора в точке x , удерживая члены порядка 1 и h в первом случае и порядка 1 во втором; получим

$$\lim_{h \rightarrow 0} \left(\left(\sum_{i=0}^k \frac{a_{-i}}{h} y(x) \right) + \sum_{i=0}^k a_{-i} (-i) y'(x) + O(h) \right) = y'(x),$$

$$\lim_{h \rightarrow \infty} \left(\left(\sum_{i=0}^k b_{-i} \right) f(x, y(x)) + O(h) \right) = f(x, y(x)).$$

Для справедливости этих соотношений необходимо и достаточно выполнение условий

$$\sum_{i=0}^k a_{-i} = 0, \quad - \sum_{i=0}^k i a_{-i} = 1, \quad \sum_{i=0}^k b_{-i} = 1.$$

Левая часть первого из этих равенств равна E_0 , разность левых частей второго и третьего равна E_1 . Отсюда следует справедливость утверждения леммы.

Если $E_0 = E_1 = 0$, $\omega = b_0 + \dots + b_{-k} \neq 0, 1$, то после домножения (2) на ω^{-1} мы получим разностную схему, для которой выполнены условия (6), а следовательно и условия нормировки (5). В дальнейшем всегда предполагается, что условия нормировки (5), (6) выполнены. Изучение схем с $\omega = 0$ не представляет практического интереса.

Задача. Доказать, что условие $E_0 = E_1 = 0$ является необходимым условием сходимости решения сеточной задачи к решению дифференциальной.

Уравнения $E_0 = \dots = E_m = 0$ образуют однородную линейную систему алгебраических уравнений относительно $2k + 2$ неизвестных. Если число неизвестных больше числа уравнений $2k + 2 > m + 1$ (или, все равно, что $2k \geq m$), то эта система имеет ненулевое решение. Таким образом, решая систему уравнений $E_0 = \dots = E_{2k} = 0$, получим схему $2k$ -го порядка аппроксимации. Иногда возникает необходимость построить аппроксимации со значением $b_0 = 0$. В системе уравнений $b_0 = 0$, $E_0 = \dots = E_{2k-1} = 0$ число неизвестных больше числа уравнений, и поэтому эта система имеет ненулевое решение, которому соответствует схема $(2k - 1)$ -го порядка аппроксимации.

Примеры. При $k = 2$ общее решение системы уравнений $E_0 = \dots = E_4 = 0$ имеет вид $a_0 = c$, $a_{-1} = 0$, $a_{-2} = -c$, $b_0 = b_{-2} = \frac{c}{3}$, $b_{-1} = \frac{4}{3}c$. Из условия нормировки (6) получаем

$c = \frac{1}{2}$; расчетная схема

$$\frac{y_I - y_{I-2}}{2h} - \left(\frac{1}{6} f_I + \frac{2}{3} f_{I-1} + \frac{1}{6} f_{I-2} \right) = 0.$$

Заметим, что мы уже получали эту схему из формулы Симпсона.

Если потребовать выполнения равенств $b_0 = E_0 = E_1 = E_2 = E_3 = 0$, то получим схему

$$\frac{y_I + 4y_{I-1} - 5y_{I-2}}{6h} - \left(\frac{2}{3} f_{I-1} + \frac{1}{3} f_{I-2} \right) = 0. \quad (7)$$

Отметим следующее обстоятельство. Если равенства

$$y'(x_j) \approx \sum_{i=0}^k \frac{a_{-i} y(x_{j-i})}{h}, \quad (8)$$

$$f(x_j, y(x_j)) \approx \sum_{i=0}^k b_{-i} f(x_{j-i}, y(x_{j-i}))$$

выполняются с точностью до членов порядка $O(h^m)$, то величина r_j , равная разности погрешностей этих соотношений, имеет порядок $O(h^m)$ и, следовательно, согласно (3), $E_0 = \dots = E_m = 0$. Однако для выполнения условия $r_j = O(h^m)$ необязательно требовать, чтобы погрешности этих приближенных соотношений имели порядок $O(h^m)$. Например, для последней схемы $r_j = O(h^3)$, в то время как погрешности соотношений (8) порядка $O(h)$.

Кроме построенных выше методов типа Рунге — Кутта и конечно-разностных методов следует отметить группу методов Батчера. В этих методах при нахождении каждого нового значения y_n используется несколько предыдущих значений y_{n-i} , как в конечно-разностных методах, но в то же время на каждом шаге производится несколько вычислений правой части, как в методах Рунге — Кутта.

Примеры расчетных формул методов этой группы:

1.

$$y_{n-\frac{1}{2}} = y_{n-2} + \frac{h}{8} (9f_{n-1} + 3f_{n-2}),$$

$$y_n^1 = \frac{1}{5} (28y_{n-1} - 23y_{n-2}) + \frac{h}{15} \left(32f_{n-\frac{1}{2}} - 60f_{n-1} - 26f_{n-2} \right),$$

$$y_n = \frac{1}{31} (32y_{n-1} - y_{n-2}) + \frac{h}{93} \left(64f_{n-\frac{1}{2}} + 15f_n^1 + 12f_{n-1} - f_{n-2} \right),$$

погрешность на шаге порядка $O(h^6)$; здесь $f_m^k = f(x_m, y_m^k)$.

2.

$$\begin{aligned}
y_{n-\frac{1}{2}} &= \frac{1}{128} (-225y_{n-1} + 200y_{n-2} + 153y_{n-3}) + \\
&\quad + \frac{h}{128} (225f_{n-1} + 300f_{n-2} + 45f_{n-3}) \\
y_n^I &= \frac{1}{31} (540y_{n-1} - 297y_{n-2} - 212y_{n-3}) + \\
&\quad + \frac{h}{155} \left(384f_{n-\frac{1}{2}} - 1395f_{n-1} - 2130f_{n-2} - 309f_{n-3} \right), \\
y_n &= \frac{1}{617} (783y_{n-1} - 135y_{n-2} - 31y_{n-3}) + \\
&\quad + \frac{h}{3085} \left(2304f_{n-\frac{1}{2}} + 465f_n^I - 135f_{n-1} - 495f_{n-2} - 39f_{n-3} \right),
\end{aligned}$$

погрешность на шаге $O(h^4)$.

§ 7. Исследование свойств конечно-разностных методов на модельных задачах

Чтобы понять, при каких условиях приближения, получаемые методом конечных разностей, сходятся к решению задачи, рассмотрим простейшие модельные задачи.

Рассмотрим поведение решения конечно-разностной схемы

$$\sum_{i=0}^k a_{-i} y_{j-i} - h \sum_{i=0}^k b_{-i} f_{j-i} = 0 \quad (1)$$

на примере уравнения $y' = My$, $M = \text{const.}$ В этом случае соотношение (1) приобретает вид

$$\sum_{i=0}^k a_{-i} y_{j-i} - h \sum_{i=0}^k b_{-i} M y_{j-i} = 0,$$

иначе,

$$\sum_{i=0}^k (a_{-i} - Mhb_{-i}) y_{j-i} = 0. \quad (2)$$

Далее предполагаем, что $a_0 \neq 0$ и Mh настолько мало, что коэффициент $a_0 - Mhb_0$ отличен от нуля.

Как мы видели в § 7 гл. II, решение разностного уравнения (2) с постоянными коэффициентами выражается через корни характеристического уравнения

$$\sum_{i=0}^k (a_{-i} - Mhb_{-i}) \mu^{k-i} = 0.$$

Это уравнение удобнее записать в виде

$$F(Q, \mu) = \sum_{i=0}^k (a_{-i} - Qb_{-i}) \mu^{k-i} = 0, \quad (3)$$

где $Q = Mh$. Корни многочлена являются непрерывными функциями его коэффициентов в области, где старший коэффициент многочлена не обращается в нуль. В свою очередь коэффициенты $a_{-i} - Qb_{-i}$ — непрерывные функции от Q . Следовательно, корни многочлена (3) непрерывно зависят от Q в области, где $a_0 - Qb_0 \neq 0$, в частности, при малых Q . Отсюда заключаем, что при малых Q корни (3) близки к корням уравнения

$$F(0, \mu) = F(\mu) = \sum_{i=0}^k a_{-i} \mu^{k-i} = 0, \quad (4)$$

которое принято называть *характеристическим уравнением разностной схемы* (1). Далее обозначаем корни (3) через $\mu_i(Q)$, $i = 1, \dots, k$, а значения $\mu_i(0)$, т. е. корни (4), через μ_i .

При условии $\max_i |\mu_i| > 1$ разностную схему следует забраковать. В самом деле, решение разностного уравнения записывается в виде

$$\sum_{i=1}^k c_i (\mu_i(Q))^n$$

(рассматривается случай, когда корни уравнения (3) простые). При наличии округлений коэффициенты c_i будут возмущены. Если шаг интегрирования мал, то $\mu_i(Q)$ близко к μ_i , а число n может принимать большие значения; тогда среди величин $(\mu_i(Q))^n$ будут очень большие. В результате этого малые возмущения коэффициентов c_i могут привести к катастрофически большому возмущению решения. Ниже приводится строгая форма этого рассуждения.

Реально получаемые при вычислениях величины y_j вследствие наличия округлений не удовлетворяют в точности уравнению (2). Рассмотрим «частично идеализированный случай», когда в процессе численного интегрирования по формуле (2) округлений не производится, но имеется погрешность в значении y_{k-1} .

Пусть y_j^q , $q = 1, 2$, — решения (2), совпадающие при

$$j = 0, \dots, k-2, \quad \text{а} \quad y_{k-1}^2 - y_{k-1}^1 = \delta.$$

Их разность $z_j = y_j^2 - y_j^1$ будет решением (2) при начальных условиях $z_0 = \dots = z_{k-2} = 0$, $z_{k-1} = \delta$. Обозначим через E оператор сдвига по узлам сетки, определяемый равенством $Ey_j = y_{j+1}$. Очевидно, $E^p y_j = y_{j+p}$ как при положительных, так и при отрицательных целых p , и поэтому уравнение (2)

можно переписать в виде

$$\sum_{i=0}^k (a_{-i} - Qb_{-i}) E^{-i} y_j = 0.$$

Пусть $\mu_l(Q)$ — один из корней (3), тогда

$$\sum_{i=0}^k (a_{-i} - Qb_{-i}) \mu^{k-i} = (\mu - \mu_l(Q)) \left(\sum_{i=0}^{k-1} c_{-i}(Q) \mu^{k-1-i} \right),$$

где $c_0 = a_0 - Qb_0$. Домножим это равенство на μ^{1-k} и вместо μ подставим E , получим

$$E \sum_{i=0}^k (a_{-i} - Qb_{-i}) E^{-i} = (E - \mu_l(Q)) \left(\sum_{i=0}^{k-1} c_{-i}(Q) E^{-i} \right).$$

Применяя операторы, стоящие в обеих частях, к функции z_j , имеем

$$0 = (E - \mu_l(Q)) \left(\sum_{i=0}^{k-1} c_{-i}(Q) E^{-i} \right) z_j.$$

Обозначим

$$\left(\sum_{i=0}^{k-1} c_{-i}(Q) E^{-i} \right) z_j \quad \text{через} \quad v_j;$$

поскольку $E^{-i} z_j = z_{j-i}$, то

$$v_{k-1} = \sum_{i=0}^{k-1} c_{-i}(Q) z_{k-1-i} = c_0(Q) z_{k-1} = c_0(Q) \delta.$$

Функция v_j удовлетворяет уравнению

$$(E - \mu_l(Q)) v_j = 0, \quad \text{иначе,} \quad v_{j+1} = \mu_l(Q) v_j.$$

Отсюда следует, что

$$v_j = (\mu_l(Q))^{j-(k-1)} v_{k-1} = (\mu_l(Q))^{j-(k-1)} c_0(Q) \delta,$$

т. е.

$$\sum_{i=0}^{k-1} c_{-i}(Q) z_{j-i} = (\mu_l(Q))^{j-(k-1)} c_0(Q) \delta.$$

Очевидно,

$$\left| \sum_{i=0}^{k-1} c_{-i}(Q) z_{j-i} \right| \leq \left(\sum_{i=0}^{k-1} |c_{-i}(Q)| \right) \max_{j-k < p \leq j} |z_p|.$$

Поэтому из последнего равенства следует оценка

$$\max_{j-k < p \leq j} |z_p| \geq |\mu_l(Q)|^{j+1-k} \frac{|c_0(Q)|}{\sum_{i=0}^{k-1} |c_{-i}(Q)|} |\delta|.$$

Мы видим, что возмущение начального условия y_{k-1} на величину δ приводит к пропорциональному возмущению решения с коэффициентом,

оцениваемым снизу через $\text{const} |\mu_l(Q)|^{l+1-k}$. Если при некотором l оказалось $|\mu_l| > 1$, то множитель $|\mu_l(Q)|^N = |\mu_l(Q)|^{X/h}$, соответствующий точке $x_N = x_0 + X$, резко растет с уменьшением h . Для получения решения с высокой точностью требуется уменьшать погрешность аппроксимации, т. е. уменьшать шаг интегрирования. Приведенные выше соображения показывают, что при уменьшении шага метод быстро становится практически неустойчивым к возмущениям исходных данных.

Например, для схемы (6.7) $\max_l |\mu_l| = 5$. Предположим, что при определении значения y_l допущена погрешность $\delta = 2^{-40}$. Тогда уже при $N = 18$ величина $|\mu_l(Q)|^N \delta$ имеет порядок 1 и результирующая ошибка сравнима с самим решением.

Точно так же можно опасаться, что в случае p -кратных μ_i , $p > 1$, на границе единичного круга приближенное решение будет иметь погрешность, пропорциональную $\delta N^{p-1} = \delta (Xh^{-1})^{p-1}$, которая при малых h очень велика. Такое влияние вычислительной погрешности еще не дает основания браковать эти схемы, поскольку при реальных δ и N величина погрешности порядка δN^{p-1} все-таки еще оказывается допустимой. Однако сейчас мы приведем более серьезные доводы, обосновывающие неприемлемость таких методов.

Пусть μ_1 — p -кратный корень уравнения (4), $|\mu_1| = 1$.

Разложим левую часть (3) в ряд Тейлора по степеням $\mu - \mu_1$ и Q ; поскольку $F(Q, \mu)$ линейна по Q , то это разложение имеет вид

$$a_{00} + \sum_{j=1}^{\infty} a_{0j} (\mu - \mu_1)^j + \sum_{j=0}^{\infty} a_{1j} Q (\mu - \mu_1)^j = 0, \quad (5)$$

где

$$a_{1j} = F_{Q\mu^j}(0, \mu_1)/j!$$

Если μ_1 — p -кратный корень (4), то $a_{00} = \dots = a_{0, p-1} = 0$. С учетом этого обстоятельства соотношение (5) можно переписать в виде

$$a_{0p} (\mu - \mu_1)^p + a_{10} Q + \psi(Q, \mu) = 0,$$

где $\psi(Q, \mu)$ состоит из слагаемых вида $\text{const} (\mu - \mu_1)^{p+l}$ или $\text{const} (\mu - \mu_1)^l Q$ при $l > 0$. Сделаем допущение, что $a_{10} \neq 0$; тогда последнее равенство можно представить как

$$(\mu - \mu_1)^p = AQ + \varphi(Q, \mu), \quad (6)$$

$$A = -\frac{a_{10}}{a_{0p}} \neq 0, \infty,$$

а $\varphi(Q, \mu)$ состоит из слагаемых того же вида, что и $\psi(Q, \mu)$, т. е.

$$\varphi(Q, \mu) = O(|Q| + |\mu - \mu_1|^p) |\mu - \mu_1|.$$

Из теории функций комплексного переменного следует, что уравнение (6) имеет p корней $\mu_1(Q), \dots, \mu_p(Q)$ вида

$$\mu_l(Q) = \mu_1 + (\sqrt[p]{AQ})_0 \exp\left(\frac{2\pi i l}{p}\right) + o(\sqrt[p]{|Q|^2}); \quad (7)$$

здесь $(\sqrt[p]{AQ})_0$ — какая-то фиксированная ветвь функции $\sqrt[p]{AQ}$.

Это утверждение можно доказать, следуя указываемой ниже схеме. Рассмотрим уравнение

$$\mu - \mu_1 = \varphi_j(Q, \mu), \quad (8)$$

где $\varphi_j(Q, \mu)$ — ветвь функции $\sqrt[p]{AQ + \varphi(Q, \mu)}$, совпадающая при $\mu = \mu_1$ с $(\sqrt[p]{AQ})_0 \exp(2\pi i j/p)$. Возьмем произвольное $a > \sqrt[p]{A}$ и рассмотрим область $|\mu - \mu_1| \leq a \sqrt[p]{|Q|}$; показываем, что при $|Q| \leq C_2$, $C_2 > 0$ в этой области выполняется оценка

$$\left| \frac{\partial \varphi_j}{\partial \mu} \right| \leq C_1 |Q|^{1/p} \leq 0,5.$$

Рассмотрим итерационный процесс решения (8):

$$\mu^{(m+1)} = \mu_1 + \varphi_j(Q, \mu^{(m)}),$$

при начальном приближении

$$\mu^{(0)} = \mu_1 + (\sqrt[p]{AQ})_0 \exp\left(\frac{2\pi i j}{p}\right);$$

показываем, что при $|Q| \leq C_3$, $C_3 > 0$ все приближения $\mu^{(m)}$ лежат в области $|\mu^{(m)} - \mu_1| \leq a \sqrt[p]{|Q|}$, и затем показываем, что при всех m выполняется неравенство

$$|\mu^{(m)} - \mu^{(0)}| \leq a_1 \sqrt[p]{|Q|^2}.$$

Вследствие сходимости итерационного процесса при условии

$$\left| \frac{\partial \varphi_j}{\partial \mu} \right| \leq 0,5$$

заключаем, что величина

$$\mu_j(Q) = \lim_{m \rightarrow \infty} \mu^{(m)}$$

удовлетворяет уравнению (8) и неравенству

$$|\mu_j(Q) - \mu^{(0)}| \leq a_1 \sqrt[p]{|Q|^2}.$$

Отсюда вытекает, что соотношение (7) справедливо.

Поскольку модуль комплексного числа не меняется от умножения на число, по модулю равное 1, то на основании (7) имеем

$$|\mu_j(Q)| = |\mu_j(Q) \mu^{-1}| = \left| 1 + (\sqrt[p]{AQ})_0 \mu^{-1} \exp\left(\frac{2\pi i j}{p}\right) + \varepsilon_j(Q) \right|,$$

где

$$\varepsilon_j(Q) = O(\sqrt[p]{|Q|^2}).$$

В свою очередь модуль комплексного числа не меньше его действительной части. Поэтому

$$\begin{aligned} |\mu_j(Q)| &\geq \operatorname{Re} \left(1 + (\sqrt[p]{AQ})_0 \mu^{-1} \exp\left(\frac{2\pi i j}{p}\right) + \varepsilon_j(Q) \right) = \\ &= 1 + \operatorname{Re} \left((\sqrt[p]{AQ})_0 \mu^{-1} \exp\left(\frac{2\pi i j}{p}\right) \right) + \operatorname{Re} \varepsilon_j(Q). \end{aligned}$$

Если вместо уравнения $y' = My$ взять уравнение $y' = -My$, то для этого уравнения, соответственно, справедливо неравенство:

$$|\mu_j(Q)| \geq 1 + \operatorname{Re} \left((V \pm AQ)_0 \mu^{-1} \exp \left(\frac{2\pi i j}{p} \right) \right) + \operatorname{Re} \varepsilon'_j(Q).$$

Рассмотрим $2p$ комплексных чисел

$$(V \pm AQ)_0 \mu^{-1} \exp \left(\frac{2\pi i j}{p} \right), \quad j = 1, \dots, p.$$

Точки комплексной плоскости, соответствующие этим числам, лежат на окружности $|z| = \sqrt[p]{|AQ|}$ и делят ее на $2p$ равных частей (рис. 8.7.1). Угол

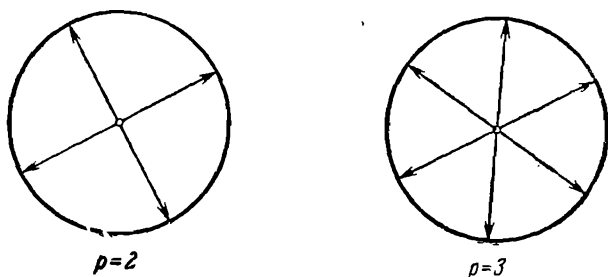


Рис. 8.7.1.

между направлениями на точки, соответствующие двум соседним числам, равен π/p . Поскольку $p \geq 2$, то он не более $\pi/2$. Следовательно, хотя бы одно из этих чисел имеет аргумент по модулю, не больший $\pi/4$. Если $z = \rho (\cos \varphi + i \sin \varphi)$ и $|\varphi| \leq \frac{\pi}{4}$, то

$$\operatorname{Re} z = \rho \cos \varphi \geq \rho \cos \left(\frac{\pi}{4} \right) = \frac{\rho}{\sqrt{2}}.$$

Поэтому для одного из этих чисел

$$\operatorname{Re} \left((V \pm AQ)_0 \mu^{-1} \exp \left(\frac{2\pi i j}{p} \right) \right) \geq \frac{\sqrt[p]{|AQ|}}{\sqrt{2}}.$$

Таким образом, или для уравнения $y' = My$ или для уравнения $y' = -My$ окажется, что

$$\max_j |\mu_j(Q)| \geq 1 + \frac{\sqrt[p]{|AQ|}}{\sqrt{2}} + \varepsilon(Q),$$

где

$$\varepsilon(Q) = o(\sqrt[p]{|Q|^2}).$$

Вспомнив, что $Q = Mh$, получаем

$$|\mu_j(Q)|^N \geq \left(1 + \frac{\sqrt[p]{|AQ|}}{\sqrt{2}} + \varepsilon(Q)\right)^{X/h} = \exp\left(\left(\frac{\sqrt[p]{|AQ|}}{\sqrt{2}} + o(\sqrt[p]{|Q|^2})\right) \frac{X}{h}\right) = \\ = \exp\left(\frac{\sqrt[p]{|AM|}}{\sqrt{2}} \frac{X}{h} + o\left(\frac{1}{h^{\frac{p-1}{p}}}\right)\right)$$

Последняя величина при измельчении шага h стремится к бесконечности быстрее любой степени h ; поэтому уже при небольшом числе шагов влияния возмущения от округлений окажется сравнимым с искомым решением задачи. Таким образом, конечно-разностные методы с кратными корням характеристического уравнения (4), по модулю равными 1, также следует забраковать.

По ходу рассуждений было сделано допущение $F_Q(0, \mu_1) \neq 0$. Если $F_Q(0, \mu_1) = 0$, то оказывается, что рассматриваемая k -шаговая схема соответствует некоторой $(k-1)$ -шаговой схеме. Например, схема

$$y_j - 2y_{j-1} + y_{j-2} - h(f_{j-1} - f_{j-2}) = 0$$

получается вычитанием уравнений

$$y_j - y_{j-1} - hf_{j-1} = 0,$$

$$y_{j-1} - y_{j-2} - hf_{j-2} = 0,$$

соответствующих схеме метода Эйлера.

Из проведенных выше рассуждений вытекает существенность выполнения следующего условия α :

все корни характеристического уравнения (4) лежат в единичном круге и на границе единичного круга нет кратных корней.

Можно было бы подумать, что все дело только в округлениях и ошибках исходных данных: если бы их не было, то, может быть, решение конечно-разностной задачи сходилась бы к решению дифференциальной? На самом деле, для любой разностной схемы, не удовлетворяющей условию α , можно построить пример дифференциального уравнения с бесконечно дифференцируемой правой частью, для которого и при отсутствии таких ошибок решение конечно-разностной задачи далеко от решения дифференциальной.

Мы не будем предлагать точных формулировок и доказательств соответствующих теорем. Строгое доказательство четко сформулированных утверждений всегда вносит определенную ясность, но вследствие своей законченности оно воспитывает также некоторый консерватизм в подходе к рассматриваемому явлению. Формулировка теоремы о том, что для любого метода, не удовлетворяющего условию α , найдется уравнение с гладкой правой частью, для которого при мельчении сетки решение сеточной задачи не сходится к решению дифференциальной в случае вычисления без округления, наводит на мысль, что от использования методов, не удовлетворяющих условию α , следует категорически отказаться. Именно этот вывод и останется

в памяти после доказательства соответствующей теоремы. Однако он не совсем оправдан. В настоящее время в вычислительную практику начинают проникать методы интегрирования, не удовлетворяющие условию α , но с совсем иным способом определения значений y_j , чем рекуррентный способ вычисления этих величин из разностной схемы. При решении одного класса задач требовалось получить решение с относительной погрешностью около 20%, но специфика задачи накладывала требование, чтобы число шагов было очень малым. Попытки математиков справиться с этой задачей при помощи методов, удовлетворяющих условию α , не привели к успеху. Инженерам по-настоящему удалось предложить метод, не удовлетворяющий условию α , для которого при достаточно крупном шаге удавалось получить решение с требуемой точностью. Здесь излишняя информация о методах решения мешала решить задачу. Впоследствии оказалось, что точность 20% не удовлетворяет новым требованиям, предъявляемым к приближенному решению. Дальнейшее повышение точности до порядка 1% потребовало применения в полном объеме результатов теоретических исследований и было достигнуто за счет применения методов, удовлетворяющих условию α , и более мощных ЭВМ.

На первый взгляд может показаться целесообразным строить схемы (1) с возможно большим порядком аппроксимации m . Однако оказывается, что все схемы с большим m не удовлетворяют условию α .

Теорема. В случаях:

схема явная $m > k$.

схема неявная, k — нечетно, $m > k + 1$,

схема неявная, k — четно, $m > k + 2$

среди корней характеристического уравнения (4) имеется корень, по модулю больший 1.

§ 8. Оценка погрешности конечно-разностных методов

Произведем оценку погрешности приближенных решений, которые получаются при помощи конечно-разностных методов вида (5.2), удовлетворяющих условию α . Получаемые в процессе реальных вычислений величины y_j , приближения к значениям $y(x_j)$, связаны на самом деле не соотношением (6.1), а соотношением

$$\sum_{i=0}^k a_{-i} y_{j-i} - h \sum_{i=0}^k b_{-i} f(x_{j-i}, y_{j-i}) = \delta_j, \quad (1)$$

где величина δ_j может быть отлична от нуля (причины появления величины δ_j уже указаны в § 4).

С другой стороны, в §§ 5, 6 установлено, что значения $y(x_n)$ точного решения дифференциальной задачи удовлетворяют соотношению

$$\sum_{i=0}^k a_{-i} y(x_{j-i}) - h \sum_{i=0}^k b_{-i} f(x_{j-i}, y(x_{j-i})) = h r_j; \quad (2)$$

при соответствующем подборе коэффициентов a_{-i} и b_{-i} оказывалось, что $r_j = O(h^m)$, где $m > 0$. Вычитая (2) из (1), получим уравнение для ошибки $R_j = y_j - y(x_j)$. На основании формулы Лагранжа имеем равенство

$$f(x_{j-i}, y_{j-i}) - f(x_{j-i}, y(x_{j-i})) = l_{j-i} R_{j-i}, \quad (3)$$

где

$$l_{j-i} = f_y(x_{j-i}, \bar{y}_{j-i}),$$

$\bar{y}_{j-i}$ лежит между y_{j-i} и $y(x_{j-i})$. С учетом (3) разность соотношений (1) и (2) запишется в виде

$$\sum_{i=0}^k a_{-i} R_{j-i} - h \sum_{i=0}^k b_{-i} l_{j-i} R_{j-i} = g_j, \quad (4)$$

где $g_j = \delta_j - h r_j$.

Теорема об оценке погрешности. Пусть разностная схема удовлетворяет условию α и $|f_y| \leq L$ при $x_0 \leq x \leq x_0 + X$. Тогда при $x_0 \leq x_n \leq x_0 + X$ выполняется неравенство

$$|R_n| \leq C(L, X) \left(\max_{0 \leq i \leq k} |R_i| + \sum_{j=k}^n |g_j| \right). \quad (5)$$

Доказательство. Сформулируем частный случай леммы из § 3 гл. VI: Пусть все собственные значения матрицы A лежат в круге $|\lambda| \leq q$ и на границе круга нет кратных корней; тогда можно указать матрицу C такую, что $\|D\|_1 \leq q$, где $D = C^{-1}AC$.

Для удобства оценки преобразуем уравнение (4) в одношаговое векторное уравнение. В дальнейшем для определенности предполагаем h настолько малым, что $|hb_0 L| \leq \left| \frac{a_0}{2} \right|$; тогда коэффициент $a_0 - hb_0 l_j$ при R_j в (4) по модулю не менее $\left| \frac{a_0}{2} \right|$. Переносим в (4) все слагаемые, не содержащие R_j , в правую часть и делим на коэффициент при R_j , получим равенство

$$R_j = \sum_{i=1}^k \frac{-a_{-i} + hb_{-i} l_{j-i}}{a_0 - hb_0 l_j} R_{j-i} + \frac{g_j}{a_0 - hb_0 l_j}. \quad (6)$$

Положим

$$\frac{-a_{-i} + hb_{-i} l_{j-i}}{a_0 - hb_0 l_j} - \left(\frac{-a_{-i}}{a_0} \right) = h v_{ij},$$

очевидно,

$$v_{ij} = \frac{b_{-i} a_0 l_{j-i} - a_{-i} b_0 l_j}{(a_0 - hb_0 l_j) a_0}, \quad |v_{ij}| \leq 2 \frac{|b_{-i} a_0| + |a_{-i} b_0|}{|a_0|^2} L.$$

Введем в рассмотрение векторы

$$\mathbf{Z}_j = (R_j, \dots, R_{j-k+1})^T.$$

Соотношение (6) равносильно равенству

$$\mathbf{Z}_j = A\mathbf{Z}_{j-1} + hV_j\mathbf{Z}_j + \mathbf{W}_j, \quad (7)$$

где

$$A = \begin{bmatrix} -\frac{a_{-1}}{a_0} & -\frac{a_{-2}}{a_0} & \dots & -\frac{a_{1-k}}{a_0} & -\frac{a_{-k}}{a_0} \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix},$$

$$V_j = \begin{bmatrix} v_{1j} & \dots & v_{kj} \\ 0 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & 0 \end{bmatrix}, \quad \mathbf{W}_j = \begin{bmatrix} \frac{g_j}{a_0 - hb_0 l_j} \\ 0 \\ \dots & \dots & \dots \\ 0 \end{bmatrix}.$$

Действительно, приравнявая первые компоненты векторов в правой и левой части (7), мы получим как раз (6), а, приравнявая остальные компоненты, получаем тождества.

$$R_{j-i} = R_{j-i}, \quad i = 1, \dots, k-1.$$

Вычислим характеристический многочлен матрицы A :

$$P(\lambda) = \det(A - \lambda E) =$$

$$= \det \begin{bmatrix} -\frac{a_{-1}}{a_0} - \lambda & -\frac{a_{-2}}{a_0} & \dots & -\frac{a_{1-k}}{a_0} & -\frac{a_{-k}}{a_0} \\ 1 & -\lambda & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & -\lambda \end{bmatrix}.$$

Для этого умножим первый столбец на λ и прибавим ко второму, затем — второй на λ и прибавим к третьему и т. д. В результате получим

$$P(\lambda) = \det \begin{bmatrix} p_1(\lambda) & p_2(\lambda) & \dots & p_{k-1}(\lambda) & p_k(\lambda) \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}.$$

где

$$p_1(\lambda) = -\frac{a_{-1}}{a_0} - \lambda, \quad p_2(\lambda) = -\frac{a_{-2}}{a_0} + \lambda \left(-\frac{a_{-1}}{a_0} - \lambda \right), \quad \dots$$

$$p_k(\lambda) = -\frac{a_{-k}}{a_0} + \lambda \left(-\frac{a_{1-k}}{a_0} + \lambda \left(-\frac{a_{2-k}}{a_0} + \dots + \lambda \left(-\frac{a_{-1}}{a_0} - \lambda \right) \dots \right) \right).$$

Раскрывая определитель по последнему столбцу, имеем

$$(-1)^k P(\lambda) = -p_k(\lambda) =$$

$$= \lambda^k + \frac{a_{-1}}{a_0} \lambda^{k-1} + \dots + \frac{a_{-k}}{a_0} = \frac{1}{a_0} (a_0 \lambda^k + a_{-1} \lambda^{k-1} + \dots + a_{-k}).$$

Характеристическое уравнение матрицы A оказалось пропорциональным характеристическому уравнению (7.5) разностной схемы. Согласно предположению α , все корни характеристического уравнения матрицы A лежат в круге $|z| \leq 1$ и на границе круга нет кратных. По отношению к матрице A условие леммы выполнено со значением $q = 1$. Следовательно, существует матрица C такая, что $C^{-1}AC = D$ и $\|D\|_1 \leq 1$. Произведем в уравнении (7) замену переменных $z_n = Cz_n$. После умножения слева на C^{-1} оно приведет к виду

$$z_j = Dz_{j-1} + hv_j z_{j-1} + w_j, \quad (8)$$

где

$$D = C^{-1}AC, \quad v_j = C^{-1}V_jC, \quad w_j = C^{-1}W_j.$$

В матрице V_j ненулевые элементы находятся только в первой строке; поэтому

$$\|V_j\|_1 \leq \left(2 \sum_{i=1}^k \frac{|b_{-i}a_0| + |a_{-i}b_0|}{|a_0|^2} \right) L = vL$$

и

$$\|v_j\|_1 \leq \|C^{-1}\|_1 \|V_j\|_1 \|C\|_1 \leq vL \|C^{-1}\|_1 \|C\|_1.$$

Очевидно,

$$\|w_j\|_1 \leq \|C^{-1}\|_1 \|W_j\|_1 = \|C^{-1}\|_1 \left| \frac{g_j}{a_0 - hb_0I_j} \right| \leq 2 \|C^{-1}\|_1 \frac{|g_j|}{|a_0|}.$$

После оценки норм слагаемых в правой части (8) можно написать неравенство

$$\|z_j\|_1 \leq \beta |g_j| + (1 + \gamma Lh) \|z_{j-1}\|_1, \quad (9)$$

где

$$\beta = 2 \frac{\|C^{-1}\|_1}{|a_0|}, \quad \gamma = v \|C^{-1}\|_1 \|C\|_1.$$

Для оценки $\|z_n\|_1$ через $\|z_{k-1}\|_1$ и величины $|g_j|$ поступим следующим образом. Выпишем неравенства (9) при $j = n, \dots, k$;

в правой части оценки $\|z_n\|_1$ через $\|z_{n-1}\|_1$ оценим $\|z_{n-1}\|_1$ через $\|z_{n-2}\|_1$, затем в правой части получившегося неравенства оценим $\|z_{n-2}\|_1$ через $\|z_{n-3}\|_1$ и т. д. В результате получим

$$\|z_n\|_1 \leq \beta |g_n| + (1 + \gamma Lh)(\beta |g_{n-1}| + (1 + \gamma Lh)(\beta |g_{n-2}| + \dots \\ \dots + (1 + \gamma Lh)(\beta |g_k| + (1 + \gamma Lh)\|z_{k-1}\|_1) \dots)),$$

или, все равно, что

$$\|z_n\|_1 \leq \beta \sum_{j=k}^n (1 + \gamma Lh)^{n-j} |g_j| + (1 + \gamma Lh)^{n-k+1} \|z_{k-1}\|_1. \quad (10)$$

Упростим эту оценку, одновременно несколько закругив ее. При $x_0 \leq x_f \leq x_n \leq x_0 + X$ имеем $(n-j)h \leq X$; поэтому

$$(1 + \gamma Lh)^{n-j} \leq \exp(\gamma Lh(n-j)) \leq \exp(\gamma LX).$$

Теперь из (10) получаем

$$\|z_n\|_1 \leq \exp(\gamma LX) \left(\beta \sum_{j=k}^n |g_j| + \|z_{k-1}\|_1 \right). \quad (11)$$

Справедливы неравенства

$$|R_n| \leq \|Z_n\|_1 \leq \|C\|_1 \|z_n\|_1, \\ \|z_{k-1}\|_1 \leq \|C^{-1}\|_1 \|Z_{k-1}\|_1 = \|C^{-1}\|_1 \max_{0 \leq i < k} |R_i|,$$

и поэтому

$$\|Z_n\|_1 \leq \|C\|_1 \exp(\gamma LX) \left(\beta \sum_{j=k}^n |g_j| + \|C^{-1}\|_1 \|Z_{k-1}\|_1 \right). \quad (12)$$

Теперь получаем оценку погрешности

$$|R_n| \leq \exp(\gamma LX) \left(M_1 \sum_{j=k}^n |g_j| + M_2 \max_{0 \leq i < k} |R_i| \right),$$

где $M_1 = \beta \|C\|_1$, $M_2 = \|C\|_1 \|C^{-1}\|_1$ и γ — некоторые постоянные, зависящие только от коэффициентов a_{-i} , b_{-i} исходной разностной схемы. В частности, M_1 и M_2 зависят только от коэффициентов a_{-i} . Утверждение теоремы доказано.

Подставляя в (12) значение $g_j = \delta_j - hr_j$, получим полную оценку погрешности

$$|R_n| \leq \exp(\gamma LX) \left(M_1 \sum_{j=k}^n (|\delta_j| + h|r_j|) + M_2 \max_{0 \leq i < k} |R_i| \right). \quad (13)$$

Из оценки (13) видно, что для сходимости решения разностного уравнения к решению дифференциального достаточно выполнения условий

$$\sum_{j=k}^n |\delta_j| \rightarrow 0, \quad h \sum_{j=k}^n |r_j| \rightarrow 0, \quad \max_{0 \leq i < k} |R_i| \rightarrow 0.$$

§ 9. Главный член погрешности

Мы получили оценки погрешности для всех конечно-разностных схем, удовлетворяющих условию α . Однако поскольку оценки являются завышенными, остается неясным, какому из методов следует отдать предпочтение. Для более подробной сравнительной характеристики разностных схем желательно выделить главный член погрешности приближенного решения.

Предположим, что порядок аппроксимации строго равен $m \geq 1$ и выполнены следующие условия:

1) разностная схема удовлетворяет условию α и нормирована, т. е. $b_0 + \dots + b_{-k} = 1$;

2) правая часть уравнения $f(x, y)$ достаточно гладкая, а именно,

$$|f_{x_1 y_2}^{l_1 + l_2}(x, y)| \leq L_{l_1 l_2} < \infty$$

в области $G: x_0 \leq x \leq x_0 + X$ для всех l_1, l_2 в пределах $0 \leq l_1 + l_2 \leq 2m$;

3) погрешность начальных условий имеет более высокий порядок малости, чем погрешность аппроксимации r_j , а именно,

$$|R_j| \leq \eta(h) h^m \quad \text{при} \quad j = 0, \dots, k-1,$$

где $\eta(h) \rightarrow 0$ при $h \rightarrow 0$;

4) погрешности от округлений имеют более высокий порядок малости по сравнению с величинами $\rho_j = h r_j$, называемыми *погрешностями на шаге*, а именно,

$$|\delta_j| \leq \eta(h) h^{m+1} \quad \text{при} \quad j = k, \dots, N.$$

Производная решения порядка $s+1$ выражается через правую часть уравнения и ее производные до порядка s . Поэтому при условии 2) имеем

$$|y^{(p)}(x)| \leq M_p < \infty \quad \text{при} \quad x_0 \leq x \leq x_0 + X \quad \text{и} \quad p = 0, \dots, 2m+1.$$

Тогда, согласно (6.3), справедливо неравенство:

$$\left| r_j - \sum_{p=0}^{2m} E_p h^{p-1} y^{(p)}(x_j) \right| \leq D_{2m+1} M_{2m+1} h^{2m}.$$

Поскольку порядок аппроксимации равен m , т. е. $E_0 = \dots = E_m = 0$, то отсюда следует соотношение

$$5) \quad r_j = E_{m+1} y^{(m+1)}(x_j) h^m + \omega_j,$$

где $|\omega_j| \leq Q h^{m+1}$ равномерно по j при $x_0 \leq x_j \leq x_0 + X$.

Хотя соотношение 5) является следствием условия 2), для удобства изложения будем формулировать его как независимое условие.

Теорема о главном члене погрешности. Если условия 1)–5) выполнены, то при $x_0 \leq x_n \leq x_0 + X$ справедливо равенство

$$y_n - y(x_n) = E_{m+1} z(x_n) h^m + \varepsilon(x_n), \quad (1)$$

где $|\varepsilon(x_n)| \leq C(\eta(h) + h) h^m$; здесь

$$z(x) = \int_{x_0}^x (-y^{(m+1)}(t)) \exp\left(\int_t^x f_y(\tau, y(\tau)) d\tau\right) dt$$

— решение уравнения

$$z' = f_y(x, y(x)) z - y^{(m+1)}(x)$$

при начальном условии $z(x_0) = 0$.

Основная идея доказательства теоремы заключается в следующем: уравнение (8.4) относительно погрешности $R_n = y_n - y(x_n)$ преобразуется к виду, из которого следует, что величина $R_n h^{-m}$ является решением разностного уравнения, аппроксимирующего указанное в условии теоремы дифференциальное уравнение относительно функции $z(x)$.

Доказательство. При выполнении условий теоремы соотношение (8.4) можно записать в виде

$$\begin{aligned} \sum_{i=0}^k a_{-i} R_{j-i} - h \sum_{i=0}^k b_{-i} l_{j-i} R_{j-i} &= \delta_j - h r_j = \\ &= -E_{m+1} y^{(m+1)}(x_j) h^{m+1} + q_j, \end{aligned} \quad (2)$$

где

$$q_j = \delta_j - (h r_j - E_{m+1} y^{(m+1)}(x_j) h^{m+1});$$

согласно условиям 4), 5), имеем

$$|q_j| \leq \eta(h) h^{m+1} + Q h^{m+2}.$$

Если подставить в правую часть (8.13) оценки для R_i , $0 \leq i < k$, δ_j , r_j , заданные в условиях теоремы, то получим оценку

$$|R_j| \leq d_1 h^m \quad \text{при} \quad x_0 \leq x_j \leq x_0 + X. \quad (3)$$

Вспомним, что $l_j = f_y(x_j, \tilde{y}_j)$, где $\tilde{y}_j$ лежит между $y(x_j)$ и $y_j = y(x_j) + R_j$. Применяя формулу Лагранжа по переменной y , имеем

$$\begin{aligned} l_j - f_y(x_j, y(x_j)) &= f_y(x_j, \tilde{y}_j) - f_y(x_j, y(x_j)) = \\ &= f_{yy}(x_j, \tilde{\tilde{y}}_j) (\tilde{y}_j - y(x_j)). \end{aligned}$$

Отсюда следует неравенство

$$|l_j - f_y(x_j, y(x_j))| \leq L_{02} |R_j| \leq d_1 L_{02} h^m \quad (4)$$

при $x_0 \leq x_j \leq x_0 + X$. Воспользовавшись равенством

$$\sum_{i=0}^k b_{-i} = 1,$$

запишем равенство (2) в виде

$$\sum_{i=0}^k a_{-i} R_{j-i} - h \sum_{i=0}^k b_{-i} (f_y(x_{j-i}, y(x_{j-i})) R_{j-i} - E_{m+1} y^{(m+1)}(x_{j-i}) h^m) = \alpha_j, \quad (5)$$

где

$$\begin{aligned} \alpha_j = & q_j + h \sum_{i=0}^k b_{-i} (l_{j-i} - f_y(x_{j-i}, y(x_{j-i}))) R_{j-i} + \\ & + h E_{m+1} \left(\sum_{i=0}^k b_{-i} (y^{(m+1)}(x_{j-i}) - y^{(m+1)}(x_j)) \right) h^m. \end{aligned}$$

Согласно оценкам (3), (4), имеем

$$|(l_{j-i} - f_y(x_{j-i}, y(x_{j-i}))) R_{j-i}| \leq d_2 h^{2m}.$$

Имеем

$$|y^{(m+1)}(x_{j-i}) - y^{(m+1)}(x_j)| \leq M_{m+1} i h;$$

из приведенных выше оценок следует

$$\begin{aligned} \max_{x_0 \leq x_j \leq x_0 + X} |\alpha_j| &= O(\eta(h) h^{m+1}) + O(h^{m+2}) + O(h^{2m+1}) = \\ &= O((\eta(h) + h) h^{m+1}). \end{aligned}$$

Разделив обе части равенства (5) на h^{m+1} и обозначив $v_j = R_j h^{-m}$, имеем

$$\sum_{i=0}^k a_{-i} v_{j-i} - h \sum_{i=0}^k b_{-i} (f_y(x_{j-i}, y(x_{j-i})) v_{j-i} - E_m y^{(m+1)}(x_{j-i})) = \beta_j, \quad (6)$$

$$\beta_j = \alpha_j h^{-m}, \quad |\beta_j| \leq d_3 (\eta(h) + h) h.$$

Сравним значения v_n со значениями решения дифференциального уравнения

$$z' = f_y(x, y(x)) z - E_m y^{(m+1)}(x) \quad (7)$$

при начальном условии $z(x_0) = 0$; имеем

$$z(x) = - \int_{x_0}^x E_m y^{(m+1)}(t) \exp \left(\int_t^x f_y(\tau, y(\tau)) d\tau \right) dt. \quad (8)$$

Соотношение (6) означает, что сеточная функция v_j приближенно удовлетворяет конечно-разностному уравнению, соответствующему этому линейному дифференциальному уравнению. Воспользуемся оценкой (8.13), подставляя β_j вместо δ_j , получим

$$|v_n - z(x_n)| \leq \exp(\gamma L X) \left(M_1 \left(\sum_{j=k}^n |\beta_j| + h |\tilde{r}_l| \right) + \right. \\ \left. + M_2 \max_{0 \leq l < k} |v_l - z(x_l)| \right); \quad (9)$$

здесь $\tilde{r}_l$ — погрешность аппроксимации:

$$\tilde{r}_l = \sum_{i=0}^k \frac{a_{-i} z(x_{l-i})}{h} - \sum_{i=0}^k b_{-i} z'(x_{l-i}),$$

соответствующая уравнению

$$z' = f_y(x, y(x)) z - y^{(m+1)}(x).$$

Решение $y(x)$ дифференцируемо по x $(2m+1)$ раз; отсюда следует, что правая часть уравнения (7) дифференцируема по x m раз. Тогда функция $z(x)$ дифференцируема $m+1$ раз; согласно построениям из § 5, имеем $|\tilde{r}_j| \leq d_4 h^m$ при $x_0 \leq x_j \leq x_0 + X$. Отсюда получаем

$$\sum_{j=k}^n h |\tilde{r}_j| \leq X \max_j |\tilde{r}_j| = XO(h^m).$$

Вследствие соотношения $|\beta_j| \leq d_3(\eta(h) + h)h$ справедливо

$$\sum_{j=1}^n |\beta_j| \leq d_3(\eta(h) + h)nh \leq d_3X(\eta(h) + h)$$

при $x_0 \leq x_n \leq x_0 + X$.

По условию теоремы $v_l = R_l h^{-m} = O(\eta(h))$ при $0 \leq l < k$; из явного вида $z(x)$ также следует, что $z(x_l) = O(h)$ при $0 \leq l < k$, поэтому

$$\max_{0 \leq l < k} |v_l - z(x_l)| = O(\eta(h) + h).$$

Используя эти неравенства для оценки правой части (9), получим

$$\max_n |v_n - z(x_n)| = O(\eta(h) + h).$$

После умножения этого соотношения на h^m имеем требуемое утверждение теоремы.

Теорема о главном члене погрешности дает формальные основания для применения правила Рунге практической оценки погрешности (см. §§ 13, 15 гл. III) при достаточно малых h и погрешностях округления δ_j . Однако в рассматриваемом случае к его применению следует относиться с большей осторожностью, чем в случае вычисления интегралов.

Дело заключается в следующем. Иногда соотношение

$$|E_{m+1/2}(x) h^m| \gg |\varepsilon(x)|, \quad (10)$$

необходимое для того, чтобы $E_{m+1/2}(x) h^m$ было реально главным членом погрешности, выполняется лишь при очень малых h и при невыполняемом в действительности условии на малость погрешностей округления δ_j .

Вышесказанное относится не ко всем методам. В случае одношаговых методов главный член погрешности (4.12) довольно часто реально преобладает над членами высшего порядка. Среди конечно-разностных методов следует выделить методы, у которых один из корней характеристического уравнения (7.4) равен 1, а остальные лежат строго внутри единичного круга; сюда относятся методы Адамса. Для методов Рунге — Кутты и Адамса соотношение (10) выполняется уже при относительно приемлемых шагах h . В то же самое время надо отметить, что при реальных вычислениях по методам Рунге — Кутты и Адамса довольно типичны крупные шаги, при которых соотношение (10) не всегда выполняется.

В случае, когда условие (10) не выполняется, иногда можно говорить, что реально у погрешности имеется главный член, но он не совпадает с

$E_{m+1/2}(x) h^m$. Например, для разностной схемы $\frac{y_j - y_{j-2}}{2h} - f(x_{j-1}, y_{j-1}) = 0$ реальным главным членом иногда оказывается некоторая составляющая $\varepsilon(x)$, имеющая порядок h^{m+1} . Некоторое подтверждение наших высказываний можно найти в следующем параграфе.

В целом можно сказать, что сравнение результатов расчетов с различными шагами является основным способом для получения суждения о величине реальной погрешности и используется относительно часто.

Из проведенного нами в §§ 8, 9 анализа, казалось бы, следует, что все схемы одного и того же порядка аппроксимации, удовлетворяющие условию α , равносильны между собой в отношении практического использования: условия сходимости приближенных решений, согласно построениям из § 8, имеют одинаковый вид, главные члены погрешности, согласно только что доказанной теореме, оказываются пропорциональными.

Однако проведенный нами анализ еще оказывается недостаточным; мы исследовали асимптотическое поведение погрешности при фиксированном отрезке интегрирования $[x_0, x_0 + X]$ и при стремлении шага сетки к нулю. Во многих задачах типично очень большое значение промежутка интегрирования, и в этом случае представляет интерес не только поведение погрешности при уменьшении шага интегрирования, но и поведение погрешности при большом отрезке интегрирования. Чтобы разобраться подробнее в этом вопросе, обратимся опять к рассмотрению модельных примеров.

§ 10. Изучение свойств конечно-разностных методов на более точных моделях

Произведем сравнение двух конечно-разностных схем: *метода Эйлера*

$$\frac{y_I - y_{I-1}}{h} - f(x_{I-1}, y_{I-1}) = 0 \quad (1)$$

и одного из вариантов так называемого *метода Милна*

$$\frac{y_I - y_{I-2}}{2h} - f(x_{I-1}, y_{I-1}) = 0. \quad (2)$$

Так же, как в § 5, подставим вместо y_j значения точного решения $y(x_j)$ и, разлагая левую часть в ряд Тейлора по степеням h , получим величину погрешности аппроксимации

$$r_I^{(1)} = \frac{y(x_I) - y(x_{I-1})}{h} - f(x_{I-1}, y(x_{I-1})) = y''(x_I) \frac{h}{2} + O(h^2), \quad (3)$$

$$r_I^{(2)} = \frac{y(x_I) - y(x_{I-2})}{2h} - f(x_{I-1}, y(x_{I-1})) = y'''(x_I) \frac{h^2}{6} + O(h^3), \quad (4)$$

соответственно, для первой и второй схем. Для начала вычислений по первой схеме необходимо располагать приближенным значением величины $y(x_0)$, а для начала вычислений по второй схеме еще и приближенным значением величины $y(x_1)$. Предположим, что эти значения задаются без ошибок, а каких-либо округлений при вычислениях вообще не происходит. Характеристическое уравнение первой схемы $\mu - 1 = 0$, характеристическое уравнение второй схемы $\mu^2 - 1 = 0$. Оба эти уравнения удовлетворяют условию α . Предположим, что правая часть $f(x, y)$ удовлетворяет всем условиям теоремы о главном члене погрешности из § 9; тогда, согласно теореме, выполняются соотношения

$$y_n^1 - y(x_n) = -\frac{h}{2} \int_{x_0}^{x_n} y''(t) \exp\left(\int_t^{x_n} f_y(\tau, y(\tau)) d\tau\right) dt \quad (5)$$

для первой схемы,

$$y_n^2 - y(x_n) = -\frac{h^2}{6} \int_{x_0}^{x_n} y'''(t) \exp\left(\int_t^{x_n} f_y(\tau, y(\tau)) d\tau\right) dt \quad (6)$$

для второй схемы.

Обратимся опять к задаче интегрирования модельного уравнения $y' = My$ при начальном условии $y(0) = 1$. Правая часть этого уравнения удовлетворяет всем требованиям теоремы из

§ 9; очевидно, $f_y = M$, $y^{(h)}(x) = M^h \exp(Mx)$. Из (5) и (6), соответственно, получим

$$\begin{aligned} y_n^1 - y(x_n) &= -\frac{h}{2} \int_0^{x_n} M^2 \exp(Mt) \exp\left(\int_t^{x_n} M d\tau\right) dt = \\ &= -\frac{M^2 x_n h}{2} \exp(Mx_n) + O(h^2), \\ y_n^2 - y(x_n) &= -\frac{M^3 x_n h^2}{6} \exp(Mx_n) + O(h^3). \end{aligned}$$

Главный член погрешности второй схемы отличается от главного члена погрешности первой множителем $Mh/3$. Если сравнивать эти схемы по величине главного члена, то, казалось бы, при $|Mh/3| < 1$ применение второй схемы предпочтительнее.

Однако такой анализ оказывается слишком поверхностным. Решения конечно-разностных уравнений (1), (2) могут быть выписаны в явном виде. В случае первой схемы имеем

$$\frac{y_n^1 - y_{n-1}^1}{h} - M y_{n-1}^1 = 0,$$

тогда

$$y_n^1 = (1 + Mh) y_{n-1}^1,$$

и поэтому

$$y_n^1 = (1 + Mh)^n.$$

Во втором случае получаем уравнение

$$\frac{y_n^2 - y_{n-2}^2}{2h} - M y_{n-1}^2 = 0;$$

общее решение этого уравнения имеет вид

$$y_n^2 = C_1 \mu_1^n + C_2 \mu_2^n,$$

где $\mu_{1,2} = Mh \pm \sqrt{1 + M^2 h^2}$ — корни уравнения

$$\mu^2 - 2Mh\mu - 1 = 0.$$

Условия совпадения с точным решением в точках $x = x_0 = 0$ и $x = x_1 = h$ дают систему уравнений относительно C_1 и C_2 :

$$C_1 + C_2 = 1, \quad C_1 \mu_1 + C_2 \mu_2 = \exp(Mh);$$

отсюда

$$C_1 = \frac{\exp(Mh) - \mu_2}{\mu_1 - \mu_2}, \quad C_2 = \frac{\exp(Mh) - \mu_1}{\mu_2 - \mu_1};$$

корни μ_1, μ_2 , а следовательно, и коэффициенты C_1, C_2 являются функциями от величины $Mh = Q$; напомним их разложения по

степеням этой величины:

$$\sqrt{1 + M^2 h^2} = 1 + \frac{Q^2}{2} + O(Q^4),$$

$$\begin{aligned} \mu_1 &= 1 + Q + \frac{Q^2}{2} + O(Q^4) = \exp Q - \frac{Q^3}{6} + O(Q^4) = \\ &= \exp \left(Q - \frac{Q^3}{6} + O(Q^4) \right), \end{aligned}$$

$$\begin{aligned} \mu_2 &= -1 + Q - \frac{Q^2}{2} + O(Q^4) = -\exp(-Q) - \frac{Q^3}{6} + O(Q^4) = \\ &= -\exp \left(-Q + \frac{Q^3}{6} + O(Q^4) \right), \quad (7) \end{aligned}$$

$$\mu_1 - \mu_2 = 2 + Q^2 + O(Q^4),$$

$$\begin{aligned} C_1 &= \frac{1 + Q + \frac{Q^2}{2} + \frac{Q^3}{6} + 1 - Q + \frac{Q^2}{2} + O(Q^4)}{2 + Q^2 + O(Q^4)} = \\ &= \frac{2 + Q^2 + \frac{Q^3}{6} + O(Q^4)}{2 + Q^2 + O(Q^4)} = 1 + \frac{Q^3}{12} + O(Q^4), \end{aligned}$$

$$C_2 = 1 - C_1 = -\frac{Q^3}{12} + O(Q^4).$$

Таким образом,

$$\begin{aligned} y_n^2 &= \left(1 + \frac{Q^3}{12} + O(Q^4) \right) \left(\exp \left(Q - \frac{Q^3}{6} + O(Q^4) \right) \right)^n - \\ &\quad - \left(\frac{Q^3}{12} + O(Q^4) \right) (-1)^n \left(\exp \left(-Q + \frac{Q^3}{6} + O(Q^4) \right) \right)^n; \end{aligned}$$

далее предполагается: $|Q| \leq \text{const}$. Справедливо очевидное равенство

$$1 + Q = \exp \left(Q - \frac{Q^2}{2} + O(Q^3) \right).$$

Поэтому для первой схемы отношение приближенного решения y_n^1 к точному $y(x_n) = \exp(Mx_n) = \exp(Qn)$ есть

$$\begin{aligned} \frac{y_n^1}{y(x_n)} &= \frac{\exp \left(\left(Q - \frac{Q^2}{2} + O(Q^3) \right) n \right)}{\exp(Qn)} = \exp \left(\left(-\frac{Q^2}{2} + O(Q^3) \right) n \right) = \\ &= 1 - \frac{Q^2 n}{2} + O(Q^3 n) + O(Q^4 n^2) = \\ &= 1 - \frac{M^2 x_n}{2} h + O(M^3 x_n (1 + |M| x_n) h^2) \quad (8) \end{aligned}$$

Во втором случае

$$\begin{aligned} \frac{y_n^2}{y(x_n)} &= \left(1 + \frac{Q^3}{12} + O(Q^4)\right) \exp\left(\left(-\frac{Q^3}{6} + O(Q^4)\right)n\right) - \\ &\quad - \left(\frac{Q^3}{12} + O(Q^4)\right) \exp\left(\left(-2Q + \frac{Q^3}{6} + O(Q^4)\right)n\right) = \\ &= 1 + O(M^3 x_n h^2) - \frac{(Mh)^3}{12} (1 + O(Mh)) \exp(-2Mx_n(1 + O(Mh)^2)). \end{aligned} \quad (9)$$

При $M \geq 0$ отсюда имеем

$$\frac{y_n^2}{y(x_n)} = 1 + O(M^3 x_n h^2); \quad (10)$$

оценка (10) предпочтительнее оценки (8) при малых Mh независимо от длины отрезка интегрирования. Если $M < 0$, то второе слагаемое в правой части (9) содержит множитель, резко растущий с ростом x_n . Например, при $M = -1$, $h = 0,01$, $x_n = 10$ оказывается, что

$$\left| \frac{y_n^2}{y(x_n)} \right| > 30.$$

В то же время в случае расчетной формулы (1) оказывается

$$\frac{y_n^1}{y(x_n)} = 0,95 \dots$$

Посмотрим, чем объясняется в случае $M < 0$ причина превосходства формулы (1), имеющий лишь первый порядок точности.

Решения рассматриваемого дифференциального уравнения имеют вид

$$\text{const} \cdot \exp(Mx),$$

и при переходе от одной точки x_n к следующей их значения умножаются на $\exp(Mh) = \exp Q$. Конечно-разностное уравнение, соответствующее схеме (2), имеет частные решения двух видов $C_1 \mu_1^n$ и $C_2 \mu_2^n$. Согласно (7)

$$\mu_1 = \exp(Mh) + O(|Mh|^3) \quad (11)$$

и составляющая решения, соответствующая этому корню, умножается на каждом шаге на множитель, близкий множителю $\exp(Mh)$. Составляющая, соответствующая второму корню, на

каждом шаге умножается на множитель μ_2 , существенно отличающийся от множителя $\exp(Mh)$; можно сказать, что эта составляющая носит паразитический характер. Если $M > 0$, то $|\mu_2| < |\mu_1|$ и влияние паразитической составляющей убывает с каждым шагом. Если же $M < 0$, то $|\mu_2| > |\mu_1|$ и влияние паразитической составляющей быстро возрастает, приводя к катастрофическому изменению результата.

Реально приходится решать не одно уравнение, а систему уравнений. Если разность между любыми двумя решениями системы растет, то уравнение $y' = My$ при $M > 0$ является приемлемой моделью для отработки методов решения такой системы. Если любые два решения сближаются с течением времени, то хорошей моделью для отработки методов решения такой системы является уравнение $y' = My$ при $M < 0$. Часто решения дифференциальных уравнений не имеют резко выраженной тенденции расходиться или сближаться в каждый момент, а эта тенденция выявляется интегрально. Поэтому предъявляется требование, чтобы метод достаточно хорошо интегрировал уравнение $y' = My$ как при $M > 0$, так и при $M < 0$. Наиболее существенными являются требования, предъявляемые к случаю $M < 0$, поскольку характерное сближение решений является типичным для диссипативных задач с оттоком энергии, которые составляют существенную долю задач, встречающихся в приложениях. Довольно распространены задачи, где решения сближаются или расходятся относительно медленно; к этому классу задач относятся многие задачи без диссипации энергии.

Уравнение $y' = My$ как при $M > 0$, так и при $M < 0$ не может служить моделью для отработки методов решения таких задач в силу иного характера поведения их решений.

Здесь в качестве модели можно взять задачу интегрирования уравнения $y' = My$ относительно комплексно-значной функции y при M чисто мнимом: $M = is$, s действительно. Решения этого уравнения $y(x) = y(0)\exp(isx)$ и разности между решениями остаются постоянными по модулю.

В случае схемы (1)

$$y_n = (1 + ish)y_{n-1}, \quad |1 + ish| = \sqrt{1 + s^2h^2} > 1,$$

следовательно,

$$|y_n| = \sqrt{1 + s^2h^2} |y_{n-1}|$$

и решения разностного уравнения по модулю возрастают.

В случае схемы (2) величины y_n удовлетворяют разностному уравнению

$$y_n - y_{n-2} - 2shiy_{n-1} = 0;$$

соответствующее характеристическое уравнение

$$\mu^2 - 2shi\mu - 1 = 0$$

имеет корни

$$shi \pm \sqrt{1 - s^2 h^2},$$

по модулю равные 1, и, следовательно, решение разностного уравнения колеблется, не имея устойчивой тенденции ни к возрастанию, ни к убыванию. Такой характер поведения решения предпочтительнее, чем у схемы (1), не говоря уже о том, что схема (2) имеет более высокий порядок точности. Характер поведения решения сеточной задачи может существенно измениться при переходе к случаю $M = M(x)$. В случае нелинейных уравнений картина становится еще более запутанной. Поэтому обработка методов решения задач с сохранением энергии только на модельных задачах вида $y' = My$ оказывается недостаточной. Для интегрирования дифференциальных уравнений, соответствующих задачам с сохранением энергии, на больших промежутках времени иногда применяют специальные схемы, на решениях которых остаются постоянными некоторые сеточные аналоги интегралов энергии. При отборе методов для стандартных программ обычно ограничиваются рассмотрением поведения решений разностных уравнений, соответствующих уравнению $y' = My$ при $M > 0$ и $M < 0$.

Решение дифференциального уравнения $y' = My$ умножается за шаг на множитель $\exp(Mh)$. В рассматриваемых выше примерах всегда оказывалось, что у соответствующего разностного уравнения характеристическое уравнение имело корень, близкий к $\exp(Mh)$; наличие такого корня и создает возможность близости решений разностного и дифференциального уравнений.

Теорема. Чтобы нормированная конечно-разностная схема

$$\sum_{i=0}^k \frac{a_{-i} y_{n-i}}{h} - \sum_{i=0}^k b_{-i} f(x_{n-i}, y_{n-i}) = 0 \quad (12)$$

имела порядок точности, строго равный m , необходимо и достаточно выполнение следующего условия: среди корней уравнения

$$\sum_{i=0}^{k-1} (a_{-i} - M h b_{-i}) \mu^{k-i} = 0 \quad (13)$$

найдется корень

$$\mu = \exp(Mh) - E_{m+1}(Mh)^{m+1} + O((Mh)^{m+2}), \quad E_{m+1} \neq 0.$$

Доказательство. Предположим, что разностная схема (12) имеет порядок аппроксимации, строго равный m , т. е. для любого достаточно гладкого решения

$$\sum_{i=0}^k \frac{a_{-i} y(x_{n-i})}{h} - \sum_{i=0}^k b_{-i} f(x_{n-i}, y(x_{n-i})) = \\ = E_{m+1} y^{(m+1)}(x_n) h^m + O(h^{m+1}), \quad E_{m+1} \neq 0. \quad (14)$$

Положим $f(x, y) = My$ и подставим в (14) решение $y(x) = \exp(Mx)$; при $x_q = qh$ и $n = k$ получим

$$\sum_{i=0}^k \frac{a_{-i} \exp(M(k-i)h)}{h} - \sum_{i=0}^k b_{-i} M \exp(M(k-i)h) = \\ = E_{m+1} M^{m+1} h^m \exp(Mkh) + O(h^{m+1}).$$

После домножения на h имеем

$$\sum_{i=0}^k (a_{-i} - Mhb_{-i}) \exp(Mh(k-i)) = \\ = E_{m+1} (Mh)^{m+1} \exp(Mkh) + O(h^{m+2}) = E_{m+1} (Mh)^{m+1} + O(h^{m+2}).$$

Все явно выписанные слагаемые в правой и левой части есть функции от $Q = Mh$; поэтому величина, указанная как $O(h^{m+2})$, есть на самом деле функция от Mh , т. е. $O((Mh)^{m+2})$. Таким образом, последнее соотношение можно переписать в виде

$$\sum_{i=0}^k (a_{-i} - Qb_{-i}) (\exp Q)^{k-i} = E_{m+1} Q^{m+1} + O(Q^{m+2}),$$

или

$$P(\exp Q) = E_{m+1} Q^{m+1} + r_{m+2},$$

где

$$P(\mu) = \sum_{i=0}^k (a_{-i} - b_{-i}Q) \mu^{k-i}, \quad r_{m+2} = O(Q^{m+2}).$$

Рассмотрим отрезок $|\mu - 1| \leq 2|Q|$ и оценим на этом отрезке производную

$$P'(\mu) = \sum_{i=0}^k (a_{-i} - Qb_{-i}) (k-i) \mu^{k-i-1}.$$

Для точек этого отрезка

$$\mu^{k-i-1} = 1 + O(Q),$$

и следовательно,

$$P'(\mu) = \sum_{i=0}^k a_{-i} (k-i) + O(Q).$$

Поскольку, вследствие условий $E_0 = F_1 = 0$ и нормированности схемы,

$$\sum_{i=0}^k a_{-i} = 0 \quad \text{и} \quad \sum_{i=0}^k a_{-i} (-i) = 1,$$

то это равенство переписывается в виде

$$P'(\mu) = 1 + O(Q),$$

иначе,

$$1 - C|Q| \leq P'(\mu) \leq 1 + C|Q| \quad \text{при} \quad |\mu - 1| \leq 2|Q|.$$

Положим

$$\mu^{1,2} = \exp Q - \frac{E_{m+1}Q^{m+1} + r_{m+2}}{1 \mp CQ}.$$

Очевидно,

$$\mu^{1,2} = 1 + Q + O(Q^2).$$

Будем предполагать $|Q|$ настолько малым, что $|\exp Q - 1| \leq 2|Q|$,

$$|\mu^{1,2} - 1| \leq 2|Q|, \quad C|Q| < 1 \quad \text{и} \quad |r_{m+2}| < |E_{m+1}Q^{m+1}|.$$

Для определенности рассмотрим случай $E_{m+1} > 0$, $Q > 0$, тогда

$$\mu^1 < \mu^2 < \exp Q.$$

Очевидны неравенства

$$\begin{aligned} P(\exp Q) - P(\mu^2) &\leq \left(\max_{[\mu^2, \exp Q]} P'(\mu) \right) (\exp Q - \mu^2) \leq \\ &\leq (1 + C|Q|) \frac{E_{m+1}Q^{m+1} + r_{m+2}}{1 + C|Q|} = P(\exp Q), \end{aligned}$$

$$\begin{aligned} P(\exp Q) - P(\mu^1) &\geq \left(\min_{[\mu^1, \exp Q]} P'(\mu) \right) (\exp Q - \mu^1) \geq \\ &\geq (1 - C|Q|) \frac{E_{m+1}Q^{m+1} + r_{m+2}}{1 - C|Q|} = P(\exp Q). \end{aligned}$$

Отсюда заключаем, что

$$P(\mu^1) \leq 0 \leq P(\mu^2);$$

на основании теоремы Ролля уравнение $P(\mu) = 0$ имеет корень в пределах $\mu^1 \leq \mu \leq \mu^2$. Согласно построению, обе величины μ^1 , μ^2 могут быть записаны в виде

$$\exp Q - E_{m+1}Q^{m+1} + O(Q^{m+2}),$$

и следовательно,

$$\mu = \exp Q - E_{m+1}(Q)^{m+1} + O((Q)^{m+2}).$$

Таким образом, в одну сторону утверждение теоремы доказано.

Пусть теперь известно, что уравнение $P(\mu) = 0$ имеет корень вида

$$\mu = \exp Q - E_{m+1}Q^{m+1} + O(Q^{m+2}), \quad E_{m+1} \neq 0.$$

Проводя в обратном порядке аналогичные рассуждения (это мы предоставляем читателю), убеждаемся, что тогда

$$P(\exp Q) = E_{m+1}Q^{m+1} + O(Q^{m+2}).$$

Разложим левую часть в ряд Тейлора по степеням Q . Поскольку главный член правой части есть $E_{m+1}Q^{m+1}$, то коэффициенты при степенях Q^l при $0 \leq l \leq m$ обращаются в нуль, а коэффициент при Q^{m+1} равен E_{m+1} .

Получим соотношения

$$E_0 = \sum_{i=0}^k a_{-i} = 0,$$

$$E_l = \sum_{i=0}^k \frac{a_{-i} (-i)^l}{l!} - \sum_{i=0}^k \frac{b_{-i} (-i)^{l-1}}{(l-1)!} = 0, \quad l = 1, \dots, m,$$

$$\sum_{i=0}^k \frac{a_{-i} (-i)^{m+1}}{(m+1)!} - \sum_{i=0}^k \frac{b_{-i} (-i)^m}{m!} = E_{m+1} \neq 0.$$

Согласно построениям § 5, это как раз означает, что для гладких решений дифференциальных уравнений выполняется соотношение (14). Теорема доказана.

Если разностная схема аппроксимирует дифференциальное уравнение, то $m \geq 1$; следовательно, уравнение (13) всегда имеет корень $\mu = \exp(Mh) + O((Mh)^2)$. При $Mh \rightarrow 0$ этот корень стремится к корню $\mu = 1$ характеристического уравнения разностной схемы (7.4). В настоящее время сложилась традиция использовать разностные схемы, у которых все остальные корни характеристического уравнения (7.4) равны нулю.

Это требование означает, что $\sum_{i=0}^k a_{-i} \mu^{k-i} = (\mu - 1) \mu^{k-1}$, таким образом, $a_0 = 1$, $a_{-1} = -1$. Для таких схем при $Mh \rightarrow 0$ паразитические корни (13) стремятся к нулю. К классу этих методов относятся определенные нами выше явные и неявные формулы Адамса. На примере явных формул Адамса посмотрим более внимательно, при каких значениях Mh влияние паразитических корней разностной схемы будет мало.

Явные формулы Адамса записываются в виде

$$y_n - y_{n-1} - h \sum_{i=0}^{k-1} \gamma_i \nabla^i f_{n-1} = 0.$$

Подставляя $f(x, y) = My$, получим разностное уравнение

$$y_n - y_{n-1} - Mh \sum_{i=0}^{k-1} \gamma_i \nabla^i y_{n-1} = 0,$$

соответствующее характеристическое уравнение имеет вид

$$P(\mu) = \mu^k - \mu^{k-1} - Mh \sum_{i=0}^{k-1} \gamma_i (\mu - 1)^i \mu^{k-1-i} = 0.$$

Рассмотрим случай $M < 0$. Имеем

$$P(-1) = (-1)^k - (-1)^{k-1} - Mh \sum_{i=0}^{k-1} \gamma_i (-2)^i (-1)^{k-1-i} = \\ = (-1)^k \left[2 + Mh \sum_{i=0}^{k-1} \gamma_i 2^i \right].$$

При достаточно больших по модулю отрицательных μ многочлен $P(\mu)$ имеет знак $(-1)^k$. Если выражение в квадратной скобке отрицательно, то значение $P(-1)$ имеет противоположный знак, и следовательно, многочлен $P(\mu)$ имеет корень $\mu < -1$. Тогда решение разностного уравнения содержит растущую составляющую, в то время как решение дифференциального уравнения убывает. Поэтому для получения хороших результатов при интегрировании на большом промежутке времени необходимо выполнение условия

$$2 + Mh \sum_{i=0}^{k-1} \gamma_i 2^i > 0. \quad (15)$$

Ранее мы показали, что γ_i порядка $\frac{1}{\ln i}$, следовательно, вели-

чина $\sum_{i=1}^{k-1} \gamma_i 2^i$ растет как $\frac{2^k}{\ln k}$; в результате мы видим, что с ростом порядка аппроксимации k условие (15) на величину Mh при $M < 0$ становится все более жестким.

При интегрировании гладких решений по методу Адамса с автоматическим выбором шага наблюдается следующая интересная картина. Вследствие достаточной гладкости решения шаг начинает укрупняться. С укрупнением шага в случае $f_y < 0$ проявляется преобладание паразитических корней уравнения (13) над основным корнем. В результате этого получаемая сеточная функция становится негладкой, и поэтому шаг мельчится. После мельчения шага она становится гладкой и шаг укрупняется и т. д. Это явление было обнаружено при сравнении реального времени вычисления по программе метода Адамса с предполагаемым. Оказалось, что реальное время много больше предполагавшегося по теоретическим прикидкам, и причиной этого был большой объем вычислений при периодическом укрупнении и мельчении шага.

§ 11. Интегрирование систем уравнений

Большинство проводившихся выше построений переносится без всяких изменений на случай систем уравнений

$$y' = f(x, y).$$

Формальное отличие состоит лишь в том, что в соответствующих соотношениях вместо скалярных величин участвуют не-

которые матрицы или тензоры. Например, при интегрировании одного дифференциального уравнения с помощью разложения решения в ряд Тейлора требовалось вычислять производные

$$y'' = f_x + f_y y',$$

$$y''' = f_{xx} + 2f_{xy}y' + f_{yy}(y')^2 + f_y y'' \text{ и т. д.}$$

Пусть

$$y = (y_1, \dots, y_q)^T, \quad f = (f_1, \dots, f_q)^T.$$

В случае системы дифференциальных уравнений производные $y_i^{(k)}$ получаются из аналогичных соотношений, получающихся дифференцированием исходной системы

$$y_i'' = \frac{d}{dx} (f_i(x, y)) = \frac{\partial f_i}{\partial x} + \sum_{j=1}^q \frac{\partial f_i}{\partial y_j} \frac{dy_j}{dx},$$

$$y_i''' = \frac{d}{dx} (y_i'') = \frac{\partial^2 f_i}{\partial x^2} + 2 \sum_{j=1}^q \frac{\partial^2 f_i}{\partial x \partial y_j} \frac{dy_j}{dx} + \sum_{j=1}^q \sum_{l=1}^q \frac{\partial^2 f_i}{\partial y_j \partial y_l} \frac{dy_j}{dx} \frac{dy_l}{dx} + \sum_{j=1}^q \frac{\partial f_i}{\partial y_j} \frac{d^2 y_j}{dx^2}, \dots;$$

здесь q — порядок системы.

Совокупность этих соотношений можно записать в тензорной форме:

$$y'' = f_x + f_y y',$$

$$y''' = f_{xx} + 2f_{xy}y' + f_{yy}y'y' + f_y y'', \dots;$$

здесь $f_{x^n y^m}$ — тензор

$$\frac{\partial^{n+m} f_i}{\partial x^n \partial y_{j_1} \dots \partial y_{j_m}}.$$

Внешний вид этих соотношений такой же, как в случае одного уравнения. При оценке погрешности конечно-разностных методов вместо величины

$$f(x_n, y_n) - f(x_n, y(x_n)) = f_y(x_n, \tilde{y}_n) R_n$$

стоит выражение

$$f(x_n, y_n) - f(x_n, y(x_n)) = f_y^n R_n,$$

где f_y^n — матрица с элементами $\left. \frac{\partial f_i}{\partial y_j} \right|_{(x_n, y_n^i)}$; зависимость y_n^i от

индекса i вызвана тем, что формула Лагранжа применяется независимо в каждой строке последнего равенства.

Всякая ошибка, вносимая в процессе численного интегрирования, может моделироваться как возмущение начального условия

для дальнейшего процесса численного интегрирования; это обстоятельство было особенно наглядно видно в § 4 при получении выражения погрешности метода Рунге — Кутта. В случае одного дифференциального уравнения вариация решения (разность двух бесконечно близких решений) удовлетворяет уравнению в вариациях $\eta' = f_y \eta$. Поскольку у решений численных аппроксимаций нет особых причин сближаться больше, чем у близких решений дифференциальных уравнений, то умножение погрешности на множитель $\exp\left(\int f_y dx\right)$, имевшее место в методе Рунге — Кутта, является неизбежным. Все наши построения, проводившиеся на примерах модельных уравнений $y' = My$, имели целью отбраковать схемы с ростом погрешности, худшим этого «неустранимого» роста. В случае систем уравнений система уравнений в вариациях имеет вид $\eta' = f_y \eta$, f_y — матрица, и поведение ее решений соответственно определяет вклад погрешности, вносимой на шагах, в окончательную погрешность. Если в случае одного уравнения главный член погрешности конечно-разностных методов оказывался пропорциональным решению уравнения

$$z' = f_y z + y^{(m+1)}(x),$$

то в случае системы он будет пропорционален решению системы

$$z' = f_y z + y^{(m+1)}(x).$$

Формулы Рунге — Кутта и конечно-разностные схемы, формально переписанные для случая систем уравнений, имеют тот же порядок погрешности, что и в случае одного уравнения.

Рассмотрим модельную задачу интегрирования линейной системы с постоянными коэффициентами

$$y' = Ay. \quad (1)$$

Для простоты предполагаем, что жорданова форма матрицы A простая:

$$C^{-1}AC = \Lambda,$$

где Λ — диагональная матрица. Расчетные формулы метода Рунге — Кутта приобретают вид

$$k_1 = hAy,$$

$$\dots \dots \dots$$

$$k_q = hA\left(y + \sum_{i=1}^{q-1} \beta_{qi} k_i\right),$$

$$y(x+h) \approx z(h) = y + \sum_{i=1}^q \rho_i k_i.$$

Если последовательно выражать $k_2, \dots, k_q, z(h)$ непосредственно через y , то получатся соотношения

$$\begin{aligned} k_1 &= Ah y, \\ k_2 &= P_2(Ah) y, \\ &\dots \dots \dots \\ k_q &= P_q(Ah) y, \\ z(h) &= Q_q(Ah) y, \end{aligned}$$

где P_l, Q_l — многочлены степени l от матрицы Ah с коэффициентами, зависящими от параметров $\alpha_i, \beta_{ij}, p_i$ данного метода. Соотношение

$$y_l = Q_q(Ah) y_{l-1} \quad (2)$$

будет выполняться на каждом шаге интегрирования. В результате после n шагов будет получено значение $y_n = (Q_q(Ah))^n y_0$. В случае интегрирования одного уравнения $y' = \lambda y$ это равенство приобретает вид

$$y_n = (Q_q(\lambda h))^n y_0.$$

Пусть $e_1, \dots, e_m$ — собственные векторы матрицы A , а $\lambda_1, \dots, \lambda_m$ — соответствующие собственные значения. Поскольку, по предположению, матрица A простой структуры, то эти векторы образуют базис; всякий вектор y_j однозначно представится в этом базисе:

$$y_j = \sum_{p=1}^m c_j^p e_p.$$

Подставим выражения y_n, y_{n-1} через векторы базиса в правую часть (2). При умножении многочлена $S(A)$ от матрицы A на собственный вектор e_p матрицы A имеем

$$S(A) e_p = S(\lambda_p) e_p.$$

Таким образом,

$$Q_q(Ah) y_{n-1} = \sum_{p=1}^m Q_q(Ah) c_{n-1}^p e_p = \sum_{p=1}^m Q_q(\lambda_p h) c_{n-1}^p e_p;$$

следовательно,

$$\begin{aligned} c_n^p &= Q_q(\lambda_p h) c_{n-1}^p, \\ c_n^p &= (Q_q(\lambda_p h))^n c_0^p, \end{aligned} \quad (3)$$

$$y_n = \sum_{p=1}^m (Q_q(\lambda_p h))^n c_n^p e_p. \quad (4)$$

Из соотношений (3), (4) видна следующая картина, имеющая место при интегрировании линейной системы (1) методом Рунге — Кутты: каждая составляющая решения, пропорциональная

одному из собственных векторов матрицы A , интегрируется независимо от остальных компонент; при этом формулы преобразования коэффициентов c_n^p совпадают с формулами численного интегрирования методов Рунге — Кутта уравнения $y' = \lambda_p y$.

Обратимся к конечно-разностным методам. В случае системы (1) соответствующая система конечно-разностных уравнений имеет вид

$$\sum_{i=0}^k a_{-i} y_{n-i} - h \sum_{i=0}^k b_{-i} A y_{n-i} = 0; \quad (5)$$

положим, как и выше,

$$y_n = \sum_{p=1}^m c_n^p e_p;$$

подставляя в (5), получим

$$\sum_{p=1}^m \left(\sum_{i=0}^k a_{-i} c_{n-i}^p - h \sum_{i=0}^k b_{-i} \lambda_p c_{n-i}^p \right) e_p = 0.$$

Вследствие линейной независимости векторов e_p каждая скобка обращается в нуль. Таким образом, справедливы равенства

$$\sum_{i=0}^k a_{-i} c_{n-i}^p - h \sum_{i=0}^k b_{-i} \lambda_p c_{n-i}^p = 0.$$

Это уравнение относительно величины c_n^p совпадает с разностным уравнением, возникающим при численном интегрировании уравнения $y' = \lambda_p y$ рассматриваемым методом.

В обоих случаях (методов типа Рунге — Кутта и конечно-разностных методов) составляющие решения, пропорциональные собственным векторам e_p , интегрируются независимо друг от друга и расчетные формулы совпадают с расчетными формулами при интегрировании скалярных уравнений $y' = \lambda_p y$. Если данный шаг интегрирования обеспечит приемлемую точность при интегрировании всех уравнений $y'_p = \lambda_p y_p$, то он обеспечит приемлемую точность и при интегрировании всей системы.

Решения $\text{const exp}(\lambda_p x)$ уравнения $y' = \lambda_p y$ меняются существенно на отрезке длины порядка $|\lambda_p|^{-1}$. Поэтому для получения приемлемой точности при интегрировании этой составляющей решения необходимо выполнение соотношения $h \ll |\lambda_p|^{-1}$, иначе, $|\lambda_p| h \ll 1$. Если требуется высокая точность при интегрировании всех составляющих решения, то, очевидно, выполнение условия

$$\max_p |\lambda_p| h \ll 1 \quad (6)$$

является необходимым.

Обратимся к случаю нелинейной системы

$$y' = f(x, y).$$

В окрестности искомого решения $Y(x)$ правая часть системы близка к правой части линейной системы

$$y' = f(x, Y) + f_y(x, Y)(y - Y).$$

Чтобы решение этой системы интегрировалось с высокой точностью, нужно, чтобы с высокой точностью интегрировалось решение линейной системы

$$z' = f_y(x, Y)z.$$

В окрестности каждой точки X эта система аппроксимируется линейной системой с постоянными коэффициентами:

$$u' = f_y(X, Y(X))u.$$

Таким образом, возникает требование: чтобы решение исходной системы интегрировалось с достаточно высокой точностью, необходимо выполнение условия $\max |\lambda_p(X)| h \ll 1$, где $\lambda_p(X)$ — собственные значения матрицы $f_y^p(X, Y(X))$. Поскольку это условие получено при помощи различных упрощений исходной системы, оно не является достаточным условием, обеспечивающим близость решений сеточной и дифференциальной задач.

Однако при интегрировании некоторых классов задач требование выполнения условия (6) является слишком обременительным. Подобные классы задач качественно можно описать следующим образом. Рассматриваемая система имеет некоторое асимптотически устойчивое решение $y_0(x)$. Среди решений, находящихся в окрестности $y_0(x)$, существуют решения, стремящиеся к $y_0(x)$ относительно быстро и относительно медленно. Численно отыскивается одно из решений системы, медленно сближающееся с $y_0(x)$, причем представляет интерес его поведение лишь при достаточно больших x . В качестве модели можно принять систему двух линейных уравнений $y' = Ay$ со следующими свойствами. Оба собственных значения λ_h матрицы A отрицательные, причем $|\lambda_1| \ll |\lambda_2|$. Представляет интерес получить решение этой системы при x таких, что $|\lambda_2|x \gg 1$, а величина $|\lambda_1|x$ не очень велика. Решение рассматриваемой задачи имеет вид

$$C_1 \exp(\lambda_1 x) e_1 + C_2 \exp(\lambda_2 x) e_2.$$

Указанные условия на величину x означают, что вторая составляющая решения $C_2 \exp(\lambda_2 x) e_2$ ничтожно мала, а первая нет. Первая компонента будет интегрироваться с приемлемой

точностью, если $|\lambda_1|h \ll 1$. Поскольку величина $|\lambda_1|x$ не очень велика, то могло бы показаться, что решение рассматриваемой задачи не составит существенных трудностей.

Однако отсутствие требований к точности интегрирования второй компоненты решения еще не означает, что для ее интегрирования применим любой метод.

Обратимся, например, к явным методам Адамса. При нарушении условия (10.15) относительно $M = \lambda_2$ разностная схема имеет паразитический корень, по модулю больший 1, и вторая компонента разностного решения сильно растет. Поэтому является актуальным построение метода, для которого вторая компонента хотя, возможно, и интегрируется плохо, но быстро затухает. Обратимся к методам типа Рунге — Кутта, где в случае линейного уравнения $y' = \lambda y$ соответствующее характеристическое уравнение имеет только один характеристический корень. Будем обозначать величину $Q_4(\lambda h)$ через μ .

Для метода Эйлера

$$z(h) = y + k_1 = (1 + \lambda h) y.$$

Таким образом, $\mu = 1 + \lambda h$, и условие $|\mu| \leq 1$ при $\lambda < 0$ нарушается, если $|\lambda|h > 2$.

Для метода второго порядка точности (2.7)

$$k_1 = \lambda h y,$$

$$k_2 = hf \left(x + \frac{h}{2}, y + \frac{k_1}{2} \right) = \lambda h \left(y + \frac{k_1}{2} \right) = \lambda h \left(y + \frac{\lambda h y}{2} \right),$$

$$z(h) = y_0 + k_2 = \left(1 + \lambda h + \frac{\lambda^2 h^2}{2} \right) y,$$

$$\mu = 1 + \lambda h + \frac{\lambda^2 h^2}{2},$$

и условие $|\mu| \leq 1$ при $\lambda < 0$ нарушается при $|\lambda|h > 2$. Для метода второго порядка точности (2.6)

$$k_1 = \lambda h y,$$

$$k_2 = \lambda h (y + k_1) = \lambda h (y + \lambda h y),$$

$$z(h) = y + \frac{1}{2} (k_1 + k_2) = \left(1 + \lambda h + \frac{\lambda^2 h^2}{2} \right) y,$$

и условие, при котором $|\mu| \leq 1$, для $\lambda < 0$ то же самое.

В случае схем четвертого порядка точности (2.14) и (2.15)

$$\mu = 1 + \lambda h + \frac{\lambda^2 h^2}{2} + \frac{\lambda^3 h^3}{6} + \frac{\lambda^4 h^4}{24};$$

$|\mu| \leq 1$ при $\lambda < 0$, если $|\lambda h| \leq \kappa \approx 2.8$. Таким образом, применение методов Рунге — Кутта для решения рассматриваемой задачи возможно лишь при выполнении условия типа

$$|\lambda_2| h \leq \text{const.}$$

З а д а ч а. Доказать, что для методов Рунге — Кутты при $q = s$ всегда

$$z(h) = \left(\sum_{k=0}^s \frac{\lambda^k h^k}{k!} \right) y$$

(определение q и s см. § 2).

Более приемлемое решение стоящей перед нами проблемы обеспечивает применение неявного метода Адамса второго порядка точности:

$$y_n - y_{n-1} = \frac{h}{2} (f_n + f_{n-1}).$$

Подставляя λy вместо f , получаем $y_n = \mu y_{n-1}$, где $\mu = \frac{1 + \frac{\lambda h}{2}}{1 - \frac{\lambda h}{2}}$. При

$\lambda = \alpha + i\beta$ имеем

$$\mu = \frac{1 + \frac{\alpha h}{2} + i \frac{\beta h}{2}}{1 - \frac{\alpha h}{2} - i \frac{\beta h}{2}}, \quad |\mu| = \frac{\sqrt{\left(1 + \frac{\alpha h}{2}\right)^2 + \left(\frac{\beta h}{2}\right)^2}}{\sqrt{\left(1 - \frac{\alpha h}{2}\right)^2 + \left(\frac{\beta h}{2}\right)^2}}.$$

Если $\alpha = \operatorname{Re} \lambda < 0$, то

$$\left(1 + \frac{\alpha h}{2}\right)^2 + \left(\frac{\beta h}{2}\right)^2 < \left(1 - \frac{\alpha h}{2}\right)^2 + \left(\frac{\beta h}{2}\right)^2,$$

и поэтому $|\mu| < 1$. Таким образом, применение этой схемы обеспечивает невозрастание компонент решения, соответствующих всем λ с $\operatorname{Re} \lambda < 0$. Приемлемая точность интегрирования первой компоненты обеспечится, если $|\lambda_1| h \ll 1$. Однако невозрастание второй компоненты еще не означает ее малости.

Справедливо равенство

$$\mu_2 = \frac{1 + \frac{\lambda_2 h}{2}}{1 - \frac{\lambda_2 h}{2}} = - \frac{1 + \frac{2}{\lambda_2 h}}{1 - \frac{2}{\lambda_2 h}};$$

отсюда видно, что при $\lambda_2 < 0$ с ростом $|\lambda_2| h$ величина $|\mu_2|$, оставаясь меньше 1, начинает приближаться к 1, и поэтому затухание второй составляющей решения окажется медленным. Требование быстрого затухания этой составляющей решения наложит некоторое ограничение сверху на величину шага h , однако оно окажется более слабым, чем ограничение $|\lambda_2| h \ll 1$. Для линейной системы с переменными коэффициентами аналогом последней схемы будут схемы второго порядка точности:

$$\begin{aligned} \frac{y_n - y_{n-1}}{h} - \frac{A(\dot{x}_n) y_n + A(x_{n-1}) y_{n-1}}{2} - F_{n-\frac{1}{2}} &= 0, \\ \frac{y_n - y_{n-1}}{h} - A\left(x_{n-\frac{1}{2}}\right) \frac{y_n + y_{n-1}}{2} - F_{n-\frac{1}{2}} &= 0, \end{aligned} \quad (7)$$

где

$$F_{n-\frac{1}{2}} = \frac{f(x_n) + f(x_{n-1}))}{2},$$

иногда целесообразнее брать

$$F_{n-\frac{1}{2}} = f\left(x_{n-\frac{1}{2}}\right), \quad \text{или} \quad F_{n-\frac{1}{2}} = \frac{1}{h} \int_{x_{n-1}}^{x_n} f(x) dx.$$

При использовании обоих типов расчетных формул на каждом шаге требуется производить обращение матрицы, размерность которой равна размерности системы. Если размерность системы большая, то производятся дальнейшие усовершенствования метода с целью избежать обращения матриц на каждом шаге.

Рассмотрим еще один способ решения задач подобного рода. Сначала обратимся к случаю системы $y' = Ay$, $A = \text{const.}$ Для численного интегрирования этой системы можно воспользоваться равенством

$$y(x_n + h) = \varphi(h) y(x_n),$$

где

$$\varphi(h) = \exp(Ah).$$

Матричная функция $\varphi(H)$ удовлетворяет соотношению

$$\varphi(H) = \varphi\left(\frac{H}{2}\right) \varphi\left(\frac{H}{2}\right), \quad (8)$$

поэтому при ее вычислении поступают следующим образом. Задаются некоторым l , вычисляют матрицу $\varphi\left(\frac{h}{2^l}\right)$, а затем при помощи равенства (8) последовательно вычисляют матрицы

$$\varphi\left(\frac{h}{2^{l-1}}\right), \dots, \varphi\left(\frac{h}{2}\right), \varphi(h).$$

Матрицу $\varphi\left(\frac{h}{2^l}\right)$ можно заменить отрезком ее ряда Тейлора

$$\varphi\left(\frac{h}{2^l}\right) \approx \sum_{i=0}^s \frac{1}{i!} \left(\frac{Ah}{2^l}\right)^i. \quad (9)$$

Заметим, что при $q = s \leq 4$ вычисление по формуле (9) равносильно интегрированию матричной системы уравнений $Y' = AY$ методом Рунге — Кутты порядка s при шаге $h/2^l$ и начальном условии $Y(0) = E$ — единичная матрица.

Для малости погрешности в (9) при небольших s существенно выполнение условия $\|A\|h/2^l \ll 1$. Поэтому величина 2^l в ряде случаев может оказаться довольно большой.

Чтобы избежать нежелательного накопления вычислительной погрешности, введем в рассмотрение (по аналогии с § 9

гл. IV) матричную функцию $\psi(h) = \varphi(h) - E$. Вычисляем

$$\psi\left(\frac{h}{2^l}\right) \approx \sum_{i=1}^s \frac{1}{il} \left(\frac{Ah}{2^l}\right)^i,$$

затем матрицы

$$\psi\left(\frac{h}{2^{l-1}}\right), \dots, \psi\left(\frac{h}{2}\right), \psi(h),$$

пользуясь рекуррентной формулой

$$\psi(H) = \psi\left(\frac{H}{2}\right) \left(2E + \psi\left(\frac{H}{2}\right)\right);$$

затем находим

$$\varphi(H) = E + \psi(H).$$

Рассмотрим теперь случай неоднородной системы

$$y' = Ay + b, \quad A, b = \text{const.}$$

Разность $u(x) = y(x) - y(x_n)$ удовлетворяет системе уравнений $u' = Au + c$, где $c = Ay(x_n) + b$. Введем в рассмотрение матрицу

$$\omega(H) = \int_0^H \exp(A(H-t)) dt.$$

Нетрудно проверить, что $u(x_n + h) = \omega(h)c$, а матрица $\omega(H)$ удовлетворяет соотношению

$$\omega(H) = \omega\left(\frac{H}{2}\right) \left(2E + A\omega\left(\frac{H}{2}\right)\right); \quad (10)$$

кроме того,

$$\omega(H) = \sum_{i=1}^{\infty} \frac{1}{il} A^{i-1} H^i.$$

Поэтому вычисляем

$$\omega\left(\frac{h}{2^l}\right) \approx \sum_{i=1}^s \frac{1}{il} A^{i-1} \left(\frac{h}{2^l}\right)^i,$$

затем матрицы

$$\omega\left(\frac{h}{2^{l-1}}\right), \dots, \omega\left(\frac{h}{2}\right), \omega(h)$$

с помощью соотношения (10) и полагаем

$$y(x_n + h) \approx y(x_n) + \omega(h)(Ay(x_n) + b).$$

В случае систем общего вида, линейных или нелинейных, можно на каждом из отрезков $[x_n, x_{n+1}]$ аппроксимировать рассматриваемую систему системой

$$y' = Ay + b, \quad A, b = \text{const.}$$

§ 12. Ряд общих вопросов

Для начала вычислений по какой-либо конечно-разностной схеме порядка $k > 1$ необходимо иметь в распоряжении значения $y_0, \dots, y_{k-1}$ приближений к $y(x_0), \dots, y(x_{k-1})$. Значение y_0 задается начальным условием. При имеющейся в нашем распоряжении информации о численных методах проще всего применить для нахождения $y_1, \dots, y_{k-1}$ какой-либо из одношаговых методов типа Рунге — Кутты. Традиционно для нахождения значений $y_1, \dots, y_{k-1}$ принято применять несколько иные методы.

Возьмем фиксированное q и произвольное j в пределах $0 < j \leq q$. В правой части равенства

$$y(x_j) - y(x_0) = \int_{x_0}^{x_j} y'(x) dx$$

заменяем $y'(x)$ интерполяционным многочленом Лагранжа по узлам $x_0, \dots, x_q$:

$$y'(x) = \sum_{i=0}^q y'(x_i) \prod_{p \neq i} \frac{x - x_p}{x_i - x_p} + r(x),$$

$$r(x) = y'(x_0; \dots; x_q; x) \prod_{p=0}^q (x - x_p);$$

получим равенство

$$y(x_j) - y(x_0) = h \sum_{i=0}^q c_{ji} y'(x_i) + R_j, \quad (1)$$

где

$$c_{ji} = \frac{1}{h} \int_{x_0}^{x_j} \prod_{p \neq i} \frac{x - x_p}{x_i - x_p} dx = \int_0^1 \prod_{p \neq i} \frac{t - p}{i - p} dt,$$

$$R_j = \int_{x_0}^{x_j} r(x) dx.$$

Величина $|y'(x_0; \dots; x_q; x)|$ оценивается сверху через

$$\max_{[x_0, x_q]} \left| \frac{y^{(q+2)}(x)}{(q+1)!} \right|;$$

производя в выражении R , замену переменной $x = th$ и оценивая подынтегральную функцию по модулю, получим

$$|R_J| \leq c_J \max_{[x_0, x_q]} |y^{q+2}(x)| h^{q+2}, \quad (2)$$

$$c_J = \int_0^1 \frac{\left| \prod_{p=0}^q (t-p) \right|}{(q+1)!} dt.$$

За приближения к значениям $y(x_i)$ примем решение системы

$$y_J - y_0 = h \sum_{i=0}^q c_{Ji} f(x_i, y_i). \quad (3)$$

Вычтем (1) из (3); при этом воспользуемся тем, что разность

$$f(x_i, y_i) - y'(x_i) = f(x_i, y_i) - f(x_i, y(x_i))$$

на основании формулы Лагранжа может быть записана в виде $l_i r_i$, где

$$r_i = y_i - y(x_i), \quad l_i = f_{y_i}(x_i, \tilde{y}_i),$$

$\tilde{y}_i$ лежит между y_i и $y(x_i)$. Получим равенство

$$r_J - r_0 = h \sum_{i=0}^q c_{Ji} l_i r_i - R_J. \quad (4)$$

Положим

$$L = \max_{x_0 \leq x \leq x_q} |f_y|,$$

$$c = \max_J \sum_{i=0}^q |c_{Ji}|.$$

Из (4) следует неравенство

$$|r_J| \leq |r_0| + cLh \max_{0 \leq i \leq q} |r_i| + |R_J|, \quad (5)$$

откуда

$$\max_{0 \leq J \leq q} |r_J| \leq |r_0| + \max_{0 \leq J \leq q} |R_J| + cLh \max_{0 \leq i \leq q} |r_i|.$$

Переносим последнее слагаемое в правую часть, получим

$$(1 - cLh) \max_{0 \leq i \leq q} |r_i| \leq |r_0| + \max_{0 \leq i \leq q} |R_i|.$$

Предположим далее h настолько малым, что $cLh < 1$, тогда

$$\max_{0 \leq i \leq q} |r_i| \leq \frac{|r_0| + \max_{0 \leq i \leq q} |R_i|}{1 - cLh}.$$

Если считать, что заданное значение y_0 точное, т. е. $r_0 = 0$, то отсюда следует оценка погрешности

$$\max_{0 \leq i \leq q} |r_i| = O(h^{q+2}). \quad (6)$$

В отдельных случаях эту оценку удастся уточнить. Например, если q четно, то узлы интерполяции расположены симметрично относительно точки $x_0 + \frac{q}{2}h$ — середины отрезка $[x_0, x_q]$. Тогда, как мы видели в разделе численного интегрирования, погрешность R_q квадратурной формулы

$$y(x_q) - y(x_0) \approx h \sum_{k=0}^q c_{qk} y'(x_k)$$

есть величина порядка $O(h^{q+3})$. Подставляя эту оценку в правую часть (5) при $j = q$ и воспользовавшись (6), получим $r_q = O(h^{q+3})$. Это обстоятельство используется в случае интегрирования по формулам Адамса, где наиболее важна точность значения y_q с максимальным q : дело в том, что в случае интегрирования по формуле Адамса k -го порядка вклад от погрешности начальных условий оказывается

$$O\left(|y_{k-1} - y(x_{k-1})| + h \sum_{j=0}^{k-2} |y_j - y(x_j)|\right).$$

В нелинейном случае система уравнений (3), как правило, не может быть разрешена относительно неизвестных $y_1, \dots, y_q$. Для ее решения применяется метод простой итерации или метод типа (7.1.8):

$$y_j^{m+1} - y_0 = h \sum_{i=0}^{j-1} c_{ji} f(x_i, y_i^{m+1}) + h \sum_{i=j}^q c_{ji} f(x_i, y_i^m).$$

Производные правых частей по переменным y_j содержат множитель h ; в результате этого при h малых рассматриваемые итерационные процессы осуществляют сжатое отображение и поэтому процесс итераций сходится. Выбор начального приближения для итерационного процесса, числа q и числа итераций часто объясняется не чисто математическими факторами, а обстоятельствами, связанными с объемом оперативной памяти машины, типичной величиной порядка систем, предъявляемых к решению, и рядом других причин.

Остановимся подробнее на вопросе нахождения значений y_n на каждом шаге в случае употребления неявного метода Адамса. Обычно неявный метод Адамса

$$y_n - y_{n-1} - h \sum_{i=0}^k b_{-i} f(x_{n-i}, y_{n-i}) = 0 \quad (7)$$

используется в паре с явным методом Адамса

$$y_n - y_{n-1} - h \sum_{i=1}^{k+1} \bar{b}_{-i} f(x_{n-i}, y_{n-i}) = 0.$$

Как видно из их построения, оба этих метода имеют порядок аппроксимации $O(h^{k+1})$. Применение неявных методов требует решения нелинейного уравнения или системы уравнений и, казалось бы, приводит к ненужным осложнениям по сравнению с явными методами. Однако неявные методы имеют несколько меньшее значение постоянного множителя в главном члене погрешности на шаге; в случае линейного уравнения ограничение на шаг типа (10.15), обеспечивающее при $f_y < 0$ отсутствие корней уравнения (7.3), по модулю больших 1, для неявных методов более слабое. Последнее обстоятельство, по-видимому, явилось наиболее важным фактором, определившим распространность неявных методов.

После получения значения y_{n-1} определяется нулевое приближение к значению y_n :

$$y_n^0 = y_{n-1} + h \sum_{i=1}^{k+1} b_{-i} f(x_{n-i}, y_{n-i});$$

затем производится несколько итераций с помощью формулы

$$y_n^{m+1} = y_{n-1} + h \sum_{i=1}^k b_{-i} f(x_{n-i}, y_{n-i}) + h b_0 f(x_n, y_n^m). \quad (8)$$

Производная правой части (8) по y_n равна $h b_0 f_y$, поэтому на каждом шаге итерационного процесса погрешность значений y_n^m по отношению к точному решению (7) умножается на множитель порядка h . Мы не будем производить детального анализа изменения погрешности значений y_n^m вследствие его громоздкости. Оказывается, для того чтобы главный член погрешности был таким же, как в случае точного решения (7) относительно неизвестной y_n , достаточно ограничиться одной итерацией по формуле (8). Поскольку трудоемкость вычисления правой части обычно существенно больше трудоемкости остальных вспомогательных операций, то за оценку трудоемкости метода естественно принять число вычислений правой части. Если на каждом шаге производить одну итерацию, а затем вычислять значение $f(x_n, y_n^1)$, необходимое для последующих вычислений, то трудоемкость метода на шаге возрастет вдвое по сравнению с явным методом Адамса; чтобы избежать этого, можно каждый раз полагать

$$f(x_n, y_n^1) \approx f(x_n, y_n^0).$$

При этом трудоемкость метода оказывается такой же, как у явного метода Адамса, а главный член погрешности такой же, как у неявного. Практически такое видоизменение неявного метода может приводить к возрастанию (по сравнению с неявным методом) величины шага, при которой сказывается влияние паразитических корней. Поэтому иногда может потребоваться дополнительное дробление шага по сравнению со случаем точного нахождения величины y_n из неявного уравнения (7). Исследователи, составляющие стандартные программы метода Адамса, находят компромиссное решение между возможностью дополнительного дробления шага и увеличением числа итераций на шаге «волевым образом» с учетом статистических свойств решаемых задач.

Иногда возникает задача численного интегрирования дифференциальных уравнений, где производная решения не выражена явно через само решение и аргумент; для определенности рассмотрим скалярный случай

$$F(x, y, y') = 0.$$

Если бы это уравнение удалось разрешить относительно y' , то получилось бы некоторое уравнение

$$y' = f(x, y).$$

Первый возможный путь решения задачи состоит в формальном применении методов Рунге — Кутты или конечно-разностных методов; при каждом x, y значение f определяется

путем численного решения уравнения

$$F(x, y, f) = 0.$$

Хорошее начальное приближение к искомому значению f можно получить интерполяцией ранее найденных значений; поэтому число требуемых итераций (например, метода секущих) обычно оказывается небольшим.

В случае неявных методов бывает целесообразно сразу решать систему уравнений

$$F(x_n, y_n, f_n) = 0,$$

$$\sum_{i=0}^k a_{-i} y_{n-i} - h \sum_{i=0}^k b_{-i} f_{n-i} = 0$$

относительно неизвестных y_n, f_n .

Иногда (довольно редко) удобно воспользоваться разложением решения в ряд Тейлора; дифференцируя исходное уравнение по x , получаем соотношения

$$\begin{aligned} \frac{d}{dx}(F(x, y, y')) &= \\ &= F_x(x, y, y') + F_y(x, y, y') y' + F_{y'}(x, y, y') y'' = 0, \\ \frac{d^2}{dx^2}(F(x, y, y')) &= \dots = 0 \end{aligned}$$

и т. д. Если значение $y'(x_0)$ уже найдено, то из этих соотношений можно явным образом определить значения $y''(x_0), \dots$

Первое из этих соотношений можно переписать в виде

$$y'' = g(x, y, y').$$

Отсюда получается третий путь решения задачи. Из уравнения

$$F(x_0, y_0, y'(x_0)) = 0$$

определяется $y'(x_0)$, и далее численно интегрируется уравнение

$$y'' = g(x, y, y').$$

Коснемся вопроса о переменном шаге интегрирования. Мы исследовали конечно-разностные методы в предположении, что шаг интегрирования $h = x_n - x_{n-1}$ постоянен. Часто желательно производить интегрирование с переменным шагом. Например, стандартные программы интегрирования методом Адамса с автоматическим выбором шага устроены так, что время от времени, в зависимости от гладкости решения, шаг удваивается или дробится вдвое. Удвоение шага можно просто произвести, выбрасывая узлы интегрирования через один и принимая оставшиеся значения неизвестных за начальные. Однако, обычно как укрупнение, так и дробление шага осуществляют с помощью формул интегрирования на неравномерной сетке. Эти формулы можно построить методом неопределенных коэффициентов, а некоторые из

них также и следующим способом. Пусть нам уже известны приближения $y_{n-1}, \dots, y_{n-k}$ к значениям решения $y(x_{n-1}), \dots, y(x_{n-k})$; в равенстве

$$y(x_n) - y(x_{n-1}) = \int_{x_{n-1}}^{x_n} y'(x) dx$$

заменяем $y'(x)$ значением интерполяционного многочлена по узлам $x_{n-1}, \dots, x_{n-k}$. Таким образом, получим формулу

$$y(x_n) - y(x_{n-1}) \approx \sum_{i=1}^k b_{-i}(x_n, \dots, x_{n-k}) y'(x_{n-i}).$$

Ей соответствует явная расчетная формула

$$y_n - y_{n-1} = \sum_{i=1}^k b_{-i}(x_n, \dots, x_{n-k}) f(x_{n-i}, y_{n-i}).$$

Если интерполирование осуществляется по узлам $x_n, \dots, x_{n-k}$, то получим неявную формулу

$$y_n - y_{n-1} = \sum_{i=0}^k b_{-i}(x_n, \dots, x_{n-k}) f(x_{n-i}, y_{n-i}).$$

Обсудим еще один вопрос общего характера. При получении оценок погрешности всегда предполагалось выполненным очень жесткое условие: $|f_y| \leq L$ при $x_0 \leq x \leq x_0 + X$ и всех y .

На самом деле, все полученные ранее оценки погрешности справедливы при достаточно малых h и δh^{-1} , если $|f_y| \leq L$ в β -окрестности решения:

$$x_0 \leq x \leq x_0 + X, \quad |y - y(x)| \leq \beta, \quad \text{где } \beta > 0.$$

Если предположить, что $|f_y| \leq L$ при всех y , то мы получаем некоторую оценку погрешности, например (8.13), правая часть которой стремится к нулю при h и $\delta h^{-1} \rightarrow 0$. Пусть значения h и δh^{-1} настолько малы, что во всех точках отрезка интегрирования правая часть оценки погрешности не превосходит β . Тогда все точки y_n , а, следовательно, и все промежуточные точки $\bar{y}_n$ также лежат в β -окрестности решения и при проведении оценки погрешности нигде не используется информация о значениях функции вне этой окрестности. Таким образом, при столь малых h и δh^{-1} оценка погрешности (8.13) справедлива в предположении, что $|f_y| \leq L$ в β -окрестности решения. Точно так же можно освободиться от ограничения: $|f_y| \leq L$ при всех y и в остальных случаях, например, при исследовании сходимости итерационного процесса в случае неявных формул. Аналогичным образом можно освободиться от условия ограниченности при всех y величин, входящих в выражение остаточного члена.

Вернемся еще раз к вопросу интегрирования уравнений и систем на больших промежутках времени. Для решения системы уравнений в вариациях $\eta' = f_y \eta$ можно написать равенство

$$\eta(x_2) = \exp \left(\int_{x_1}^{x_2} f_y(x, y(x)) dx \right) \eta(x_1).$$

Для многих асимптотически устойчивых задач равномерно по x_1, x_2 в пределах $x_0 \leq x_1 \leq x_2 \leq \infty$ выполняется оценка

$$\left| \exp \left(\int_{x_1}^{x_2} f_y(x, y(x)) dx \right) \right| \leq B \exp(-b(x_2 - x_1)), \quad b > 0. \quad (9)$$

Предположим, что в области $\|y - y(x)\| \leq \beta$, $\beta > 0$, $x_0 \leq x < \infty$ производные правой части системы, выражения которых входят в оценку остаточного члена, в частности f_y , равномерно ограничены. Это условие выполнено для многих задач, решения которых при больших x выходят на периодический или почти периодический режим. Для таких задач в случае интегрирования методами Рунге—Кутты или Адамса при достаточно малом постоянном шаге интегрирования h и малой величине δh^{-1} выполняется оценка погрешности

$$|y_n - y(x_n)| \leq B_1(h^s + \delta h^{-1}) + B_2 R \exp(-b(x_n - x_0)); \quad (10)$$

здесь s — порядок аппроксимации, R — оценка погрешности начальных условий, δ — максимум норм погрешностей δ_n , с которыми удовлетворяются сеточные уравнения (4.1), $b > 0$, b, B_1, B_2 не зависят от x_n .

По отношению к методу Рунге—Кутта в случае одного уравнения мы получили ранее такую оценку (4.11).

На примере интегрирования методом Рунге—Кутта порядка s линейного уравнения

$$y' = a_0(x)y + a_1(x)$$

аккуратно покажем, почему во всех подобных оценках несущественно поведение правой части вдали от решения.

В этом случае $f_y = a_0(x)$ и условие (9) приобретает вид

$$\exp \left(\int_{x_1}^{x_2} a_0(x) dx \right) \leq B \exp(-b(x_2 - x_1)), \quad b > 0. \quad (11)$$

Кроме того, предполагаем, что

$$0 \leq t \leq s, \quad p \neq 0, 1 \quad \sup_{x_0 \leq x < \infty} |a_p^{(i)}(x)| \leq A < \infty.$$

При условии (11), согласно (4.6), имеем

$$\begin{aligned} |R_n| &\leq \sum_{j=1}^n B \exp(-b(n-j)h) |\omega_j| + B \exp(-bnh) |R_0| \leq \\ &\leq \frac{B}{1 - \exp(-bh)} \max_{0 \leq j \leq n} |\omega_j| + B |R_0| \exp(-bnh). \end{aligned} \quad (12)$$

Далее, из соотношения

$$y(x) = y(x_0) \exp \left(\int_{x_0}^x a_0(x) dx \right) + \int_{x_0}^x b(t) \exp \left(\int_t^x a_0(\tau) d\tau \right) dt$$

и неравенства (11) следует равномерная ограниченность $|y(x)| \leq Y$ при $x_0 \leq x < \infty$.

Фиксируем некоторое $\beta > 0$; покажем, что при $h \leq h_0(\beta)$, $\frac{\delta}{h} \leq \gamma(\beta)$, $|R_0| \leq R(\beta)$, $h(\beta)$, $\gamma(\beta)$, $R(\beta) > 0$ все значения y_n лежат в области $|y_n - y(x_n)| \leq \beta$ и выполнена требуемая оценка погрешности.

Произведем сначала оценку погрешности в предположении, что $|y_n - y(x_n)| \leq \beta$ при всех n . Тогда $|y_n| \leq Y + \beta$ при всех n ; пользуясь равномерной ограниченностью y_n и производных $a_p^{(i)}(x)$ заключаем, что все погрешности на шаге

$$\rho_{j+1} = y_{j+1} - y_j(x_{j+1})$$

удовлетворяют неравенству

$$|\rho_{j+1}| \leq B'(\beta) h^{s+1}.$$

Тогда

$$|\omega_j| \leq B'(\beta) h^{s+1} + \delta$$

при всех j и из (12) следует оценка

$$|R_n| \leq \frac{B}{1 - \exp(-bh)} (B'(\beta) h^{s+1} + \delta) + B |R_0| \exp(-bnh). \quad (13)$$

На основании (13) получаем равномерную по n оценку

$$|R_n| \leq \frac{Bh}{1 - \exp(-bh)} \left(B'(\beta) h^s + \frac{\delta}{h} \right) + B |R_0|. \quad (14)$$

Множитель $\frac{Bh}{1 - \exp(-bh)}$, монотонно убывая, стремится к пределу $\frac{B}{b} > 0$ при h , стремящемся к нулю. Поэтому можно указать $h(\beta)$, $\gamma(\beta)$, $R(\beta) > 0$ такие, что правая часть последнего неравенства не превосходит β при

$$h \leq h(\beta), \quad \frac{\delta}{h} \leq \gamma(\beta), \quad |R_0| \leq R(\beta). \quad (15)$$

Лемма. При условии (15) для всех n выполнена оценка (13).

Доказательство проводим индукцией по n . При $n = 0$ доказываемое утверждение очевидно вследствие самого выбора величин $h(\beta)$, $\gamma(\beta)$, $R(\beta)$. Пусть оценка (13) выполнена при всех $n \leq n_0$. Тогда $|y_j - y(x_j)| \leq \beta$ при $j \leq n_0$; согласно проводившимся выше построениям, получаем, что тогда $|\rho_j| \leq B'(\beta) h^{s+1}$ при $j \leq n_0 + 1$. Отсюда и из (12) вытекает, что оценка (13) выполнена и при $n = n_0 + 1$. Лемма доказана.

В рассматриваемом примере y_n равномерно ограничены, значения $a_0(x_n)$ и $a_1(x_n)$ также равномерно ограничены; отсюда следует, что погрешности округления δ_n равномерно ограничены и, таким образом, все проводимые выше оценки имеют реальный смысл.

§ 13. Формулы численного интегрирования уравнений второго порядка

Введением новых неизвестных функций дифференциальные уравнения порядка выше первого и их системы сводятся к системам уравнений первого порядка. Таким образом, при формальном подходе вопрос о численном решении задачи Коши

для уравнений высших порядков можно было бы считать исчерпанным.

Однако методы, приспособленные специально для решения уравнений высших порядков, часто более эффективны. При разработке таких методов нужно также иметь в виду широкое распространение систем уравнений высоких порядков специального вида, учет специфики которых может еще более повысить эффективность методов. Например, большинство задач небесной механики сводится к интегрированию систем уравнений

$$y'' = f(y).$$

Рассмотрим несколько более широкий класс уравнений

$$y'' = f(x, y).$$

Как обычно, рассуждения проводятся для одного уравнения, поскольку перенос результатов на случай системы осуществляется автоматически. Наиболее распространенными методами интегрирования таких уравнений являются *методы Штермера-явный*

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{i=0}^k b_{-i} f(x_{n-i}, y_{n-i}) \quad (1)$$

и *неявный*

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{i=-1}^k b_{-i} f(x_{n-i}, y'_{n-i}). \quad (2)$$

Эти методы можно было бы получить методом неопределенных коэффициентов, потребовав, чтобы разность

$$\frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2} - \sum_i b_{-i} y''(x_{n-i})$$

была величиной как можно более высокого порядка по h . Здесь, как обычно, предполагается $x_{j+1} - x_j \equiv h$.

Укажем явный способ получения таких формул. В выражение

$$\frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2}$$

подставим разложение $y(x)$ в ряд Тейлора в точке x_n с остаточным членом в интегральной форме, выраженным через вторую производную:

$$y(x) = y(x_n) + y'(x_n)(x - x_n) + \int_{x_n}^x (x - t) y''(t) dt;$$

получим равенство

$$\frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2} = \int_{x_{n-1}}^{x_{n+1}} G(x) y''(x) dx,$$

$$G(x) = \begin{cases} \frac{x_{n+1} - x}{h^2} & \text{при } x_n \leq x \leq x_{n+1}, \\ \frac{x - x_{n-1}}{h^2} & \text{при } x_{n-1} \leq x \leq x_n. \end{cases}$$

Заменим в правой части $y''(x)$ интерполяционным многочленом Ньютона с узлами интерполяции $x_n, \dots, x_{n-k}$:

$$y''(x) = y''(x_n) + \frac{\nabla y''(x_n)}{h} (x - x_n) + \dots + \frac{\nabla^k y''(x_n)}{h^k k!} (x - x_n) \dots \dots (x - x_{n-k+1}) + y''(x_n; \dots; x_{n-k}; x) (x - x_n) \dots (x - x_{n-k}).$$

После осуществления интегрирования имеем соотношения

$$\frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2} = \sum_{p=0}^k \beta_p \nabla^p y''(x_n) + r_n^k,$$

где

$$\beta_p = \frac{1}{p!} \left(\int_0^1 (1-t) t \dots (t + (p-1)) dt + \int_0^{-1} (-1-t) t \dots (t + (p-1)) dt \right), \quad (3)$$

$$r_n^k = \int_{x_{n-1}}^{x_{n+1}} G(x) y''(x_n; \dots; x_{n-k}; x) (x - x_n) \dots (x - x_{n-k}) dx.$$

В выражении остаточного члена вынесем из-под знака интеграла значение $y''(x_n; \dots; x_{n-k}; x)$ в промежуточной точке. В результате получим равенство

$$r_n^k = \bar{\beta}_{k+1} y''(x_n; \dots; x_{n-k}; \bar{x}) h^{k+1};$$

после замены $y''(x_n; \dots; x_{n-k}; \bar{x})$ на значение $\frac{1}{(k+1)!} y^{(k+3)}(\bar{x}_n)$, согласно выражению (2.5.4), разделенной разности через производную имеем

$$r_n^k = \frac{\bar{\beta}_{k+1}}{(k+1)!} y^{(k+3)}(\bar{x}_n) h^{k+1}. \quad (4)$$

Можно непосредственно проверить, что $\frac{\bar{\beta}_{k+1}}{(k+1)!} = \beta_{k+1}$, где β_{k+1} определяется равенством (3) при $p = k+1$. Приведем другой вывод равенства

$$\frac{\bar{\beta}_{k+1}}{(k+1)!} = \beta_{k+1}; \text{ имеем}$$

$$\frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2} = \sum_{p=0}^{k+1} \beta_p \nabla^p y''(x_n) + O(h^{k+2}).$$

Отсюда и из (4) следует равенство

$$\frac{\bar{\beta}_{k+1}}{(k+1)!} y^{(k+3)}(\bar{x}_n) h^{k+1} = \beta_{k+1} \nabla^{k+1} y''(x_n) + O(h^{k+2});$$

поделив обе части на $y^{(k+1)}(\bar{x}_n) h^{k+1}$ и перейдя к пределу при $h \rightarrow 0$, считая $x_n = \text{const}$, получим

$$\frac{\bar{\beta}_{k+1}}{(k+1)!} = \beta_{k+1}.$$

В итоге получено равенство

$$\frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2} = \sum_{p=0}^k \beta_p \nabla^p y''(x_n) + \beta_{k+1} y^{(k+3)}(\bar{x}_n) h^{k+1}$$

и соответствующая расчетная формула

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{p=0}^k \beta_p \nabla^p f_n, \quad f_m = f(x_m, y_m).$$

Если заменить $y''(x)$ интерполяционным многочленом по узлам $x_{n+1}, \dots, x_{n-k+1}$, получим равенство

$$\frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2} = \sum_{p=0}^k \gamma_p \nabla^p y''(x_{n+1}) + \gamma_{k+1} y^{(k+3)}(\bar{x}_n) h^{k+1},$$

где

$$\gamma_p = \frac{1}{p!} \left(\int_0^1 (1-t)(t-1)t \dots (t+(p-2)) dt + \int_0^{-1} (-1-t)(t-1)t \dots (t+(p-2)) dt \right),$$

и соответствующую интерполяционную расчетную формулу

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{p=0}^k \gamma_p \nabla^p f_{n+1}. \quad (5)$$

Для информации укажем значения

$$\{\beta_p\} = 1, 0, \frac{1}{12}, \frac{1}{12}, \frac{19}{240}, \frac{3}{40}, \dots,$$

$$\{\gamma_p\} = 1, -1, \frac{1}{12}, 0, -\frac{1}{240}, -\frac{1}{240}, \dots$$

Иногда построенные выше методы называют методами *Адамса — Штермера*.

Для целей численного интегрирования уравнения

$$y'' = f(x, y, y')$$

употребительным является следующий метод: значения y_n и z_n приближений к $y(x_n)$ и $y'(x_n)$ вычисляются из рекуррентных соотношений — явных — вида

$$\begin{cases} z_{n+1} = z_n + h \sum_{k=0}^l a_k \nabla^k f_n, \\ y_{n+1} = y_n + h \sum_{k=0}^l b_k \nabla^k z_{n+1} \end{cases}$$

или неявных вида

$$\begin{cases} z_{n+1} = z_n + h \sum_{k=0}^l c_k \nabla^k f_{n+1}, \\ y_{n+1} = y_n + h \sum_{k=0}^l d_k \nabla^k z_{n+1}. \end{cases}$$

Уравнения более высокого порядка, чем второй, обычно сводят к системам уравнений второго или первого порядка.

§ 14. Оценка погрешности численного решения задачи Коши для уравнения второго порядка

Как в случае конечно-разностных схем для уравнений первого порядка отбросим в (13.1) и (13.2) слагаемые, содержащие значения f , и рассмотрим получившееся разностное уравнение. В обоих случаях оно имеет вид

$$y_{n+1} - 2y_n + y_{n-1} = 0.$$

Его характеристическое уравнение имеет кратный корень $\mu = 1$. Не нужно пугаться того, что корень оказался кратным: если в разностных схемах, аппроксимирующих уравнения первого порядка, перед значениями f стоял коэффициент порядка h , то здесь стоит коэффициент порядка h^2 .

Все рассматривавшиеся ранее разностные схемы интегрирования уравнения $y'' = f(x, y)$ могут быть представлены в виде

$$\sum_{i=0}^k \frac{a_{-i} y_{n-i}}{h^2} - \sum_{i=0}^k b_{-i} f(x_{n-i}, y_{n-i}) = 0, \quad a_0 \neq 0. \quad (1)$$

Пусть r_n — погрешность аппроксимации (результат подстановки в левую часть решения дифференциальной задачи):

$$r_n = \sum_{i=0}^k \frac{a_{-i} y(x_{n-i})}{h^2} - \sum_{i=0}^k b_{-i} f(x_{n-i}, y(x_{n-i})) \quad (2)$$

и $|r_n| \leq c_0 h^m$ равномерно на всем рассматриваемом отрезке интегрирования $[x_0, x_0 + X]$. Как обычно, вследствие округлений при вычислениях и неточности решения уравнения относительно неизвестной y_n в случае неявной схемы ($b_0 \neq 0$) реально получаемое приближенное решение y_n удовлетворяет (1) с некоторой погрешностью. Таким образом, имеем

$$\sum_{i=0}^k a_{-i} y_{n-i} - h^2 \sum_{i=0}^k b_{-i} f(x_{n-i}, y_{n-i}) = \delta_n.$$

Перепишем (2) в виде

$$\sum_{i=0}^k a_{-i} y(x_{n-i}) - h^2 \sum_{i=0}^k b_{-i} f(x_{n-i}, y(x_{n-i})) = h^2 r_n;$$

вычитая это соотношение из предыдущего, получим уравнение для погрешности $R_n = y_n - y(x_n)$:

$$\sum_{i=0}^k a_{-i} R_{n-i} - h^2 \sum_{i=0}^k b_{-i} l_{n-i} R_{n-i} = g_n, \quad (3)$$

где

$$g_n = \delta_n - h^2 r_n, \quad l_i = f_y(x_i, \tilde{y}_i).$$

Пусть

$$L = \sup_{x_0 \leq x \leq x_0 + X} |f_y(x, y)| < \infty.$$

Теорема. Пусть все корни характеристического уравнения разностной схемы

$$\sum_{i=0}^k a_{-i} \mu^{k-i} = 0 \quad (4)$$

лежат в единичном круге и на границе круга нет кратных корней, за исключением двукратного корня, равного 1. Тогда при $x_0 \leq x_n \leq x_0 + X$ решение уравнения (3) удовлетворяет неравенству

$$\max_n \left(|R_n|, \left| \frac{R_n - R_{n-1}}{h} \right| \right) \leq C(L, X) \left(\frac{1}{h} \sum_{j=k}^n |g_j| + \max_{0 \leq j < k} |R_j| + \max_{0 < j < k} \left| \frac{R_j - R_{j-1}}{h} \right| \right). \quad (5)$$

Доказательство. Положим $R_j - R_{j-1} = h z_j$. Согласно условию леммы, характеристическое уравнение (4) имеет корень $\mu = 1$, и поэтому левая часть его записывается в виде

$$\left(\sum_{i=0}^{k-1} d_{-i} \mu^{k-1-i} \right) (\mu - 1), \quad \text{где} \quad d_0 = a_0.$$

Пусть E , как и ранее, оператор сдвига: $Ey_j = y_{j+1}$. Имеем цепочку равенств

$$\begin{aligned}\sum_{i=0}^k a_{-i} R_{n-i} &= \sum_{i=0}^k a_{-i} E^{k-i} R_{n-k} = \\ &= \left(\sum_{i=0}^{k-1} d_{-i} E^{k-1-i} \right) (E-1) R_{n-k} = \sum_{i=0}^{k-1} d_{-i} E^{k-1-i} h z_{n-k+1};\end{aligned}$$

в итоге получаем

$$\sum_{i=0}^k a_{-i} R_{n-i} = h \sum_{i=0}^{k-1} d_{-i} z_{n-i}.$$

Тогда соотношение (3) можно переписать в виде

$$\sum_{i=0}^{k-1} d_{-i} z_{n-i} - h \sum_{i=0}^k b_{-i} l_{n-i} R_{n-i} = g_n h^{-1}.$$

Перенесем все слагаемые, кроме $d_0 z_n$, в правую часть и разделим равенство на $d_0 \neq 0$; получим

$$z_n = \sum_{i=1}^{k-1} \left(-\frac{d_{-i}}{d_0} \right) z_{n-i} + h \sum_{i=0}^k \left(\frac{b_{-i}}{d_0} \right) l_{n-i} R_{n-i} + \frac{g_n}{d_0 h}. \quad (6)$$

Введем в рассмотрение векторы

$$u_n = \begin{pmatrix} R_n \\ \vdots \\ R_{n-k+1} \end{pmatrix}, \quad w_n = \begin{pmatrix} z_n \\ \vdots \\ z_{n-k+2} \end{pmatrix}, \quad v_n = \begin{pmatrix} u_n \\ w_n \end{pmatrix}.$$

Соотношения $R_n = R_{n-1} + h z_n$ и (6) запишем в эквивалентной матричной форме

$$v_n = A v_{n-1} + h D_n v_n + h B_n v_{n-1} + q_n; \quad (7)$$

матрицы и вектор q_n разбиты на клетки, соответствующие векторам u_n и w_n :

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad D_n = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix},$$

$$B_n = \begin{pmatrix} B_{11} & B_{12} \\ B_{21}^n & B_{22} \end{pmatrix}, \quad q_n = \begin{pmatrix} q_1^n \\ q_2^n \end{pmatrix},$$

$$A_{11} = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix},$$

$$A_{22} = \begin{bmatrix} -\frac{d_{-1}}{d_0} & . & . & . & . & -\frac{d_{2-k}}{d_0} & -\frac{d_{1-k}}{d_0} \\ 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix},$$

$$D_{12} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}, \quad D_{21}^n = \begin{bmatrix} \frac{b_0 l_n}{d_0} & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix},$$

$$B_{21}^n = \begin{pmatrix} \frac{b_{-1} l_{n-1}}{d_0} & \dots & \frac{b_{-k} l_{n-k}}{d_0} \\ 0 & \dots & 0 \\ . & . & . \\ 0 & \dots & 0 \end{pmatrix}, \quad q_2^n = \begin{pmatrix} \frac{g_n}{d_0 h} \\ 0 \\ . \\ 0 \end{pmatrix}.$$

Матрицы A_{12} , A_{21} , D_{11} , D_{22} , B_{11} , B_{12} , B_{22} и векторы q_1^n — нулевые. Матрица A имеет клеточную диагональную форму $\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}$. Матрица A_{11} имеет следующее характеристическое уравнение: $(\mu - 1)\mu^{k-1} = 0$; матрица A_{22} имеет тот же вид, что и матрица A в § 8, причем уравнение

$$-\sum_{i=0}^{k-1} d_{-i} \mu^{k-1-i} = 0$$

удовлетворяет требованиям, наложенным на уравнение

$$\sum_{i=0}^k a_{-i} \mu^{k-i} = 0,$$

соответствующее матрице A в § 8. Из вышесказанного следует, что матрица A обладает следующим свойством: среди корней ее характеристического уравнения нет корней по модулю, больших 1, и корням, по модулю равным 1, соответствуют жордановы ящики размерности 1.

Перенесем в левую часть (7) выражение $hD_n v_n$ и умножим получившееся уравнение слева на $(E - hD_n)^{-1}$; получим равенство

$$v_n = (E - hD_n)^{-1} (A + hB_n) v_{n-1} + (E - hD_n)^{-1} q_n;$$

его можно записать в виде

$$v_n = Av_{n-1} + h(E - hD_n)^{-1}(D_n A + B_n)v_{n-1} + (E - hD_n)^{-1}q_n. \quad (8)$$

Соотношение (8) имеет вид, аналогичный (8.7); матрица A удовлетворяет условиям, требуемым в § 8; если

$$\left| \frac{b_0 L h}{d_0} \right| \leq \frac{1}{2}, \quad \text{то} \quad \|hD_n\|_1 = \left| \frac{b_0 L h}{d_0} \right| \leq \frac{1}{2},$$

и, согласно (6.5.7), имеем

$$\|(E - hD_n)^{-1}\|_1 \leq 2.$$

Отсюда следуют неравенства

$$\|(E - hD_n)^{-1}(D_n A + B_n)\|_1 \leq \gamma_1 L,$$

$$\|(E - hD_n)^{-1}q_n\|_1 \leq 2\|q_n\|_1.$$

Таким образом, (8) удовлетворяет тем же условиям, что и (8.7). Подобно тому как из (8.7) было получено (8.12), имеем

$$\|v_n\|_1 \leq \|C\|_1 \exp(\gamma_1 L X) \left(\beta_1 \sum_{j=k}^n \|q_j\|_1 + \|C^{-1}\|_1 \|v_{k-1}\|_1 \right).$$

Поскольку

$$\|q_j\|_1 = \left| \frac{g_j}{h d_0} \right|,$$

$$\|v_n\|_1 \geq \max \left(|R_n|, \left| \frac{R_n - R_{n-1}}{h} \right| \right),$$

$$\|v_{k-1}\|_1 = \max \left(\max_{0 \leq j < k} |R_j|, \max_{0 \leq j < k} \left| \frac{R_j - R_{j-1}}{h} \right| \right),$$

то отсюда следует справедливость утверждения теоремы.

Воспользовавшись оценкой

$$|g_j| \leq \max_l |\delta_l| + c_0 h^{m+2},$$

получим

$$\sum_{j=k}^n |g_j| \leq n \max_l |\delta_l| + c_0 n h^{m+2} \leq \frac{X}{h} \max_l |\delta_l| + c_0 X h^{m+1}.$$

Оценивая теперь правую часть (5), имеем оценку погрешности

$$\begin{aligned} & \max \left(|R_n|, \left| \frac{R_n - R_{n-1}}{h} \right| \right) \leq \\ & \leq C(L, X) \left(\frac{X}{h^2} \max_l |\delta_l| + c_0 X h^m + \max_{0 \leq j < k} |R_j| + \max_{0 \leq j < k} \left| \frac{R_j - R_{j-1}}{h} \right| \right). \quad (9) \end{aligned}$$

В случае уравнения $y' = My$ мы установили, что среди корней уравнения

$$\sum_{i=0}^k (a_{-i} - Mhb_{-i}) \mu^{k-i} = 0$$

есть корень $\mu = \exp(Mh) + O((Mh)^{m+1})$, если эта схема аппроксимирует дифференциальное уравнение с погрешностью аппроксимации $O(h^m)$.

Для уравнения $y'' = My$ разностное уравнение (1) приобретает вид

$$\sum_{i=0}^k (a_{-i} - Mh^2b_{-i}) y_{n-i} = 0.$$

Аналогично, если схема (1) аппроксимирует дифференциальное уравнение $y'' = f(x, y)$ с погрешностью $O(h^m)$, можно показать, что среди корней уравнения

$$\sum_{i=0}^k (a_{-i} - Mh^2b_{-i}) \mu^{k-i} = 0$$

обязательно есть два корня

$$\mu_{1,2} = \exp(\pm \sqrt{M} h) + O((\sqrt{M} h)^{m+1}).$$

§ 15. Двусторонние методы

Целью применения двусторонних методов является получение численным путем границ области, в которой находится точное решение задачи. Рассмотрим ряд примеров двусторонних методов, для простоты изложения пренебрегая вопросом о влиянии вычислительной погрешности.

Пусть заданы $x_0 < \dots < x_n$, значения $f(x_0), \dots, f(x_n)$, и пусть известно, что на $[x_0, x_n]$ производная $f^{(n)}(x)$ не меняет знака. Поставим задачу получения интервала, которому принадлежит значение $f(x)$, $x \in (x_0, x_n)$. Положим

$$A_0 = L_0, \dots, n-1(x), \quad A_1 = L_1, \dots, n(x), \quad (1)$$

где $L_k, \dots, n-1+k(x)$, $k = 0, 1$, — интерполяционный многочлен степени $n-1$ с узлами интерполяции $x_k, \dots, x_{n-1+k}$. Имеем (2.3.1) представление остаточного члена

$$f(x) - A_0 = \frac{f^{(n)}(\xi_0)}{n!} (x - x_0) \dots (x - x_{n-1}),$$

$$f(x) - A_1 = \frac{f^{(n)}(\xi_1)}{n!} (x - x_1) \dots (x - x_n),$$

$$\xi_0, \xi_1 \in [x_0, x_n].$$

Множители $x - x_0$ и $x - x_n$ имеют различные знаки; отсюда следует, что величины $f(x) - A_0$ и $f(x) - A_1$ имеют различные знаки (если ни одна из них не обращается в 0). Поэтому $f(x)$ лежит между A_0 и A_1 , иначе,

$$\min(A_0, A_1) \leq f(x) \leq \max(A_0, A_1).$$

Рассмотрим задачу вычисления интеграла

$$I(f) = \int_0^1 f(x) dx.$$

Представим его в виде

$$I(f) = \sum_{q=1}^M I_q(f), \quad I_q(f) = \int_{\Delta_q} f(x) dx, \quad \Delta_q = \left[\frac{q-1}{M}, \frac{q}{M} \right];$$

для вычисления интегралов $I_q(f)$ применим формулы трапеций и прямоугольников:

$$I_q(f) \approx t_q(f) = \frac{1}{2M} \left(f\left(\frac{q-1}{M}\right) + f\left(\frac{q}{M}\right) \right),$$

$$I_q(f) \approx p_q(f) = \frac{1}{M} f\left(\frac{q-\frac{1}{2}}{M}\right).$$

В § 1 гл. III были получены выражения погрешности:

$$I_q(f) - t_q(f) = -\frac{1}{12M^3} f''(x_q^0), \quad (2)$$

$$I_q(f) - p_q(f) = \frac{1}{24M^3} f''(x_q^1),$$

$$x_q^0, x_q^1 \in \Delta_q.$$

Рассмотрим сначала случай, когда $f''(x)$ не меняет знак, для определенности $f''(x) \geq 0$. Тогда из (2) имеем

$$p_q(f) \leq I_q(f) \leq t_q(f),$$

после суммирования по q получим

$$P_M(f) \leq I(f) \leq T_M(f),$$

где

$$P_M(f) = \sum_{q=1}^M p_q(f), \quad T_M(f) = \sum_{q=1}^M t_q(f).$$

В случае, когда $f''(x)$ меняет знак на $[0, 1]$, возможны различные подходы к получению двусторонних оценок значения интеграла.

Первый из подходов, обладающий меньшей надежностью, заключается в следующем. После суммирования (2) по q получаем равенства

$$\begin{aligned} I(f) &= T_M(f) - \frac{1}{12M^2} S_M^0, \\ I(f) &= P_M(f) + \frac{1}{24M^2} S_M^1; \end{aligned} \quad (3)$$

здесь

$$S_M^k = \frac{1}{M} \sum_{q=1}^M f''(x_q^k), \quad k=0, 1.$$

Рассмотрим случай, когда третья производная $f'''(x)$ ограничена по модулю: $|f'''(x)| \leq B < \infty$. Предоставляем читателю самому убедиться

в справедливости соотношений

$$S_M^k = \int_0^1 f''(x) dx + \varepsilon_M^k, \quad |\varepsilon_M^k| \leq \frac{B}{2M}, \quad k = 0, 1. \quad (4)$$

Рассмотрим для определенности случай

$$A = \int_0^1 f''(x) dx > 0.$$

Тогда

$$I(f) = T_M(f) - \frac{1}{12M^2} \left(\int_0^1 f''(x) dx + \varepsilon_M^0 \right) \leq \\ \leq T_M(f) - \frac{1}{12M^2} \varepsilon_M^0 \leq T_M(f) + \frac{B}{24M^3};$$

точно так же

$$I(f) = P_M(f) + \frac{1}{24M^2} \left(\int_0^1 f''(x) dx + \varepsilon_M^1 \right) \geq P_M(f) - \frac{B}{48M^3}.$$

В совокупности получаем оценку

$$P_M(f) - \frac{B}{48M^3} \leq I(f) \leq T_M(f) + \frac{B}{24M^3}. \quad (5)$$

Оценка (5) непригодна для практического использования, так как численным путем могут быть определены только величины $P_M(f)$ и $T_M(f)$. Если шаг интегрирования настолько мал, что $B/(2M) \leq A$, то, согласно (4),

$$\text{sign } S_M^0 = \text{sign } S_M^1 = \text{sign } A.$$

Тогда из (3) следует уже эффективная оценка

$$P_M(f) \leq I(f) \leq T_M(f).$$

При $A < 0$, соответственно, имеем

$$T_M(f) \leq I(f) \leq P_M(f).$$

Объединяя эти неравенства, получаем

$$\min(P_M(f), T_M(f)) \leq I(f) \leq \max(P_M(f), T_M(f)) \quad (6)$$

при любом знаке A , если только

$$\frac{B}{2M} \leq |A|.$$

При практическом использовании двусторонних оценок считают, что оценка (6) справедлива при всех M , хотя, например, при $A = 0$ для этого нет никаких оснований.

Другая, более осторожная стратегия имеет формальные основания для применения и в случае $A = 0$. Рассмотрим те отрезки Δ_q , где $f''(x)$ сохра-

няет знак. При $f''(x) \geq 0$ из (2) следует, что

$$I_q(f) \leq t_q(f) - \frac{1}{12M^3} F_q^1,$$

$$I_q(f) \geq p_q(f) + \frac{1}{24M^3} F_q^1;$$

здесь и далее

$$F_q^1 = \min_{\Delta_q} |f''(x)|,$$

$$F_q^2 = \max_{\Delta_q} |f''(x)|.$$

Точно так же при $f''(x) \leq 0$ имеем

$$I_q(f) \geq t_q(f) + \frac{1}{12M^3} F_q^1,$$

$$I_q(f) \leq p_q(f) - \frac{1}{24M^3} F_q^1.$$

Из этих соотношений вытекает, что при любом постоянном на Δ_q знаке $f''(x)$

$$s_q^1(f) + \frac{1}{24M^3} F_q^1 \leq I_q(f) \leq s_q^2(f) - \frac{1}{24M^3} F_q^1, \quad (7)$$

где

$$s_q^1(f) = \min(p_q(f), t_q(f)),$$

$$s_q^2(f) = \max(p_q(f), t_q(f)).$$

Пусть $f''(x)$ меняет знак в некоторой точке $\bar{x}_q$ отрезка Δ_q . Для таких отрезков из (2) выведем оценку

$$s_q^1(f) - \frac{F_q^2}{12M^3} \leq I_q(f) \leq s_q^2(f) + \frac{F_q^2}{12M^3}. \quad (8)$$

Суммирование по отрезкам, где $f''(x)$ сохраняет знак, обозначаем Σ' , а суммирование по остальным отрезкам обозначаем Σ'' . Суммируя (7), (8) по всем q , получаем

$$S_M(f) + E \leq I(f) \leq S_M^2(f) - E,$$

где

$$S_M^k(f) = \sum_{q=1}^M s_q^k(f), \quad k = 1, 2,$$

$$E = \sum_q' \frac{1}{24M^3} F_q^1 - \sum_q'' \frac{F_q^2}{12M^3}.$$

Задача. Доказать неравенство

$$E \geq \frac{1}{24M^2} \int_0^1 |f''(x)| dx - \frac{B}{16M^3}.$$

Напомним, что в $\sum_q''$ входят слагаемые, соответствующие отрезкам, где $f''(x)$ меняет знак; поэтому в таких слагаемых $F_q^2 = O\left(\frac{1}{M}\right)$. Обычно точек перемены знака производной немного, и поэтому уже при небольших M окажется $E \geq 0$. Тогда можно написать оценку

$$S_M^1(f) \leq I(f) \leq S_M^2(f). \quad (9)$$

Справедливы очевидные неравенства

$$S_M^1(f) = \sum_{q=1}^M \min(p_q(f), t_q(f)) \leq \min\left(\sum_{q=1}^M p_q(f), \sum_{q=1}^M t_q(f)\right) = \min(P_M(f), T_M(f)),$$

аналогично

$$\max(P_M(f), T_M(f)) \leq S_M^2(f).$$

Следовательно, неравенство (9) задает более широкие пределы изменения $I(f)$. Естественно, что оно выполняется при более слабых ограничениях на величину шага.

Рассмотрим менее подробно вопрос о двусторонних методах решения задачи Коши для дифференциального уравнения $y' = f(x, y)$, $y(x_0) = y_0$. Ищется значение $y(x_0 + X)$. Возьмем произвольное целое m и два метода Адамса m -го порядка точности: явный

$$y_j - y_{j-1} = h \sum_{i=0}^m \gamma_i \nabla^i f_{j-1}$$

и неявный

$$y_j - y_{j-1} = h \sum_{i=0}^m \bar{\gamma}_i \nabla^i f_j.$$

Оказывается, что главные члены погрешности аппроксимации этих методов

$$\gamma_{m+1} y^{(m+2)}(x_j) h^{m+1} \quad \text{и} \quad \bar{\gamma}_{m+1} y^{(m+2)}(x_j) h^{m+1}$$

противоположны по знаку, так как $\gamma_{m+1} \bar{\gamma}_{m+1} < 0$.

Применяя эти оба метода для решений одной и той же задачи Коши, получим два приближения y_N^1 и y_N^2 к значению решения $y(x_0 + X)$. При условиях применимости теоремы о главном члене погрешности (§ 9 гл. VIII) имеем

$$\begin{aligned} y_N^1 - y(x_0 + X) &= -\gamma_{m+1} z(X) h^{m+1} + o(h^{m+1}), \\ y_N^2 - y(x_0 + X) &= -\bar{\gamma}_{m+1} z(X) h^{m+1} + o(h^{m+1}). \end{aligned} \quad (10)$$

Здесь

$$z(X) = \int_{x_0}^{x_0+X} y^{(m+1)}(t) \exp\left(\int_t^{x_0+X} f_y(\tau, y(\tau)) d\tau\right) dt.$$

Отсюда заключаем, что при $z(X) \neq 0$ и достаточно малых h

$$y_N^1 - y(x_0 + X) \quad \text{и} \quad y_N^2 - y(x_0 + X)$$

имеют противоположные знаки, и следовательно,

$$\min(y_N^1, y_N^2) \leq y(x_0 + X) \leq \max(y_N^1, y_N^2). \quad (11)$$

Такая стратегия получения границ для $y(x_0 + X)$ равносильна первой, более рискованной стратегии оценки границ для $I(f)$. Однако к применению неравенства (11) следует относиться с большей осторожностью, чем к применению неравенства (6), так как в случае дифференциальных уравнений член высшего порядка в (10) часто становится меньше главного члена лишь при очень малых шагах.

Практически рационального варианта второй, осторожной, стратегии получения границ решения с помощью применения методов Адамса неизвестно.

Однако такую стратегию можно предложить по отношению к методам Рунге — Кутты; для определенности рассмотрим случай методов второго порядка точности. Дело осложняется тем, что главные члены погрешности на шаге у методов Рунге — Кутта одинакового порядка точности не обязательно пропорциональны, и поэтому приходится прилагать специальные усилия для построения методов с пропорциональными и противоположными по знаку главными членами погрешности на шаге.

Требуемая пара методов в обоих случаях вычисляются

$$k_1 = hf(x, y), \quad k_2 = hf\left(x + \frac{1}{3}h, y + \frac{1}{3}k_1\right),$$

далее в случае первого метода

$$k_3^{(1)} = hf\left(x + \frac{1}{2}h, y + \frac{1}{2}k_2\right)$$

и

$$y(x+h) \approx y^{(1)} = y(x) + k_3^{(1)},$$

в случае второго метода

$$k_3^{(2)} = hf\left(x + \frac{5}{6}h, y + \frac{5}{6}k_2\right)$$

и

$$y(x+h) \approx y^{(2)} = y(x) + \frac{2}{5}k_1 + \frac{3}{5}k_3^{(2)}.$$

Погрешность на шаге первого метода (см. § 2 гл. VIII)

$$\varphi(h) = y(x+h) - y^{(1)} = Ah^3 + O(h^4),$$

второго —

$$\varphi(h) = y(x+h) - y^{(2)} = -Ah^3 + O(h^4),$$

где

$$A = \frac{f_{yy}f^2}{24}.$$

Получение двусторонней оценки по более осторожной стратегии осуществляется следующим образом. В процессе интегрирования на каждом шаге вычисляются две величины y_n^1 и y_n^2 , границы для решения, $y_n^1 < y_n^2$; принимается, что $y_0^1 = y_0^2 = y(x_0)$. Исходя из y_n^1 , делается шаг по обоим методам и наименьшее из получившихся значений $y_{n+1}^{1,1}$ и $y_{n+1}^{1,2}$ принимается за y_{n+1}^1 . Соответственно, делается шаг по обоим методам, исходя из значений y_n^2 , и наибольшее из получившихся значений $y_{n+1}^{2,1}$, $y_{n+1}^{2,2}$ принимается за y_{n+1}^2 .

В тех, практически крайне редких случаях, когда при некотором n случайно окажется $y_n^1 > y_n^2$, по-видимому, следует считать, что этот метод получения двусторонней оценки неприменим.

Так как точки y_n^1 и y_n^2 расположены довольно близко, то значения A , соответствующие этим точкам, близки, и там, где $A \neq 0$, обычно

$$\begin{aligned} y_{n+1}^2 &= y_{n+1}^{2,2}, & \text{если} & \quad y_{n+1}^1 = y_{n+1}^{1,1} \\ y_{n+1}^2 &= y_{n+1}^{2,1}, & \text{если} & \quad y_{n+1}^1 = y_{n+1}^{1,2}. \end{aligned} \quad (12)$$

Отметим следующее обстоятельство. Так как рассматриваемые методы второго порядка точности, то $y_n^1 - y(x_n)$ и $y_n^2 - y(x_n)$ — величины порядка h^2 .

Однако вследствие того, что главные члены погрешности на шаге у рассматриваемых методов равны по абсолютной величине и противоположны по знаку, оказывается, что

$$\left| y_n - \frac{y_n^1 + y_n^2}{2} \right| = O(h^3).$$

Применение двусторонних методов предоставляет некоторые удобства в смысле получения возможности с большей надежностью судить о границах изменения решения задачи. В то же время оно приводит к усложнению алгоритма и, соответственно, увеличению объема вычислений. При численном построении двусторонних оценок описанными способами достоверность получаемых границ может нарушиться из-за наличия вычислительной погрешности при реальных вычислениях. Учет влияния этой погрешности приводит к дополнительному усложнению алгоритмов получения двусторонних оценок. Поэтому на практике эти алгоритмы применяются очень редко.

Дополнительная литература к главе VIII

I [1, 4, 5, 13], II [2, 3, 46]; §§ 1—3: III [54, 58]; § 4: III [9, 49]; §§ 5, 6: III [57, 59]; § 7: II [1], III [7, 55], § 8: I [1], III [32]; § 9: II [39], III [6, 15, 48]; § 10: I [13], III [6, 15]; § 11: III [17, 37, 40*), 56, 60, 62], § 12: I [7], III [6, 18, 56]; § 14: II [46]; § 15: III [16, 22, 41].

*) См также Ракитский Ю. В «Новые численные методы решения систем обыкновенных дифференциальных и разностных уравнений», Труды ЛПИ, 332 (1973), 88—97.

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ КРАЕВЫХ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

При решении краевых задач возникают дополнительные трудности по сравнению со случаем решения задачи Коши: значительно сложнее исследуется вопрос о существовании решения; после написания сеточной задачи возникает система линейных или нелинейных уравнений, проблема решения которой требует дополнительного изучения.

§ 1. Простейшие методы решения краевой задачи для уравнения второго порядка

Среди краевых задач для обыкновенных дифференциальных уравнений существенную часть составляют задачи для уравнений и систем второго порядка. В частности, такие задачи возникают в баллистике, теории упругости и т. д.

Начнем изучение вопроса с одной частной, но довольно распространенной краевой задачи. Ищется решение уравнения

$$Ly = y'' - p(x)y = f(x) \quad \text{на } (0, X) \quad (1)$$

при граничных условиях

$$y(0) = a, \quad y(X) = b. \quad (2)$$

Зададимся шагом $h = XN^{-1}$, N — целое; точки $x_j = jh$ примем за узлы сетки; как обычно, y_j — приближения к значениям $y(x_j)$. После замены производной $y''(x_j)$ на разностное отношение

$$\frac{\delta^2 y_j}{h^2} = \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2}$$

получаем систему уравнений

$$l(y_j) = \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} - p_j y_j = f_j, \quad j = 1, \dots, N-1; \quad (3)$$

здесь $p_j = p(x_j)$, $f_j = f(x_j)$; граничные условия заменим соотношениями

$$y_0 = a, \quad y_N = b. \quad (4)$$

Покажем, что при $p(x) \geq 0$ система уравнений (3), (4) имеет решение, и дадим оценку погрешности.

Лемма 1. Пусть $p(x) \geq 0$, $l(z_j) \leq 0$, $z_0, z_N \geq 0$. Тогда $z_j \geq 0$ при всех j .

Доказательство. Обозначим $\min_{0 \leq j \leq N} z_j$ через d . Предположим, что $d < 0$ и, следовательно, $d \neq z_0, z_N$. Пусть q — наименьшее целое, такое, что $z_q = d$; из определения d и q имеем $z_{q-1} > d$, $z_{q+1} \geq d$. Тогда

$$l(z_q) = \frac{(z_{q+1} - d) + (z_{q-1} - d)}{h^2} - p_q z_q \geq \frac{z_{q-1} - d}{h^2} > 0,$$

и мы приходим к противоречию с предположением $d < 0$.

Лемма 2. Если $p(x) \geq 0$, то для любой функции z_j выполняется неравенство

$$\max_{0 \leq j \leq N} |z_j| \leq \max(|z_0|, |z_N|) + Z \frac{X^2}{8},$$

где

$$Z = \max_{0 < j < N} |l(z_j)|.$$

Доказательство. Введем в рассмотрение функцию

$$w_j = |z_0| \left(1 - \frac{jh}{X}\right) + |z_N| \frac{jh}{X} + Z \frac{jh(X - jh)}{2}.$$

Для многочлена второй степени величина $\frac{\delta^2 f}{h^2}$ совпадает со второй производной; поэтому $\frac{\delta^2 w_j}{h^2} = -Z$. Так как из определения w_j следует, что $w_j \geq 0$, то

$$l(w_j) = \frac{\delta^2 w_j}{h^2} - p_j w_j \leq -Z, \quad l(w_j \pm z_j) \leq -Z \pm l(z_j) \leq 0.$$

Очевидно,

$$w_0 \pm z_0 = |z_0| \pm z_0 \geq 0, \quad w_N \pm z_N = |z_N| \pm z_N \geq 0.$$

Функции $w_j \pm z_j$ удовлетворяют условиям леммы 1, и поэтому $w_j \pm z_j \geq 0$. Отсюда следует оценка $|z_j| \leq |w_j| \leq \max_{0 \leq j \leq N} |w_j|$.

Имеем неравенства

$$\begin{aligned} |z_0| \left(1 - \frac{jh}{X}\right) + |z_N| \frac{jh}{X} &\leq \max(|z_0|, |z_N|) \left(\left|1 - \frac{jh}{X}\right| + \left|\frac{jh}{X}\right|\right) = \\ &= \max(|z_0|, |z_N|), \\ jh(X - jh) &\leq \frac{X^2}{4}. \end{aligned}$$

Поэтому

$$\max_{0 \leq j \leq N} |\omega_j| \leq \max(|z_0|, |z_N|) + Z \frac{X^2}{8}.$$

Лемма 2 доказана. Из утверждения леммы следует, что однородная система, соответствующая (3), (4), имеет только нулевое решение. Так как число уравнений в (3), (4) равно числу неизвестных, то неоднородная система (3), (4) имеет и притом единственное решение.

Пусть r_j — погрешность аппроксимации, соответствующая конечно-разностной схеме (3):

$$\begin{aligned} r_j &= l(y(x_j)) - f_j = \\ &= \frac{y(x_{j+1}) - 2y(x_j) + y(x_{j-1}))}{h^2} - \rho(x_j)y(x_j) - f(x_j). \end{aligned} \quad (5)$$

Поскольку $\rho(x)y(x) + f(x) = y''(x)$, то

$$r_j = \frac{y(x_{j+1}) - 2y(x_j) + y(x_{j-1}))}{h^2} - y''(x_j).$$

Из оценок погрешности формул численного дифференцирования имеем

$$r_j = \frac{y^{(4)}(\bar{x}_j) h^2}{12}, \quad \text{где } x_{j-1} \leq \bar{x}_j \leq x_{j+1}. \quad (6)$$

Вследствие различных округлений получаемые в процессе вычислений приближения y_j к значениям $y(x_j)$ удовлетворяют системе (3), (4) с некоторыми погрешностями

$$l(y_j) - f_j = \delta_j. \quad (7)$$

Вычитая (5) из (7), получим уравнение

$$l(R_j) = \delta_j - r_j$$

относительно ошибки приближенного решения $R_j = y_j - y(x_j)$. Воспользуемся леммой 2; получим

$$|R_j| \leq \max(|R_0|, |R_N|) + \frac{X^2}{8} \left(\max_{0 \leq j \leq N} |r_j| + \max_{0 \leq j \leq N} |\delta_j| \right).$$

Согласно оценке (6), имеем

$$|r_j| \leq M_4 \frac{h^2}{12}, \quad M_4 = \max_{[0, X]} |y^{(4)}(x)|.$$

Таким образом, окончательная оценка погрешности

$$\max_{0 \leq j \leq N} |R_j| \leq \max(|R_0|, |R_N|) + \frac{M_4 X^2 h^2}{96} + \left(\max_{0 \leq j \leq N} |\delta_j| \right) \frac{X^2}{8}.$$

Мы видим, что при повышении точности, с которой удовлетворяются граничные условия и разностное уравнение, при одновременном стремлении шага к нулю решение сеточной задачи сближается с решением дифференциальной.

Описанный метод дает приближенное решение, сходящееся к точному со скоростью $O(h^2)$. Займемся построением более точных схем. Будем предполагать, что $p(x)$ и $f(x)$ непрерывно дифференцируемы 4 раза; тогда решение задачи непрерывно дифференцируемо 6 раз. Еще раз рассмотрим величину

$$r_j = \frac{y(x_{j+1}) - 2y(x_j) + y(x_{j-1}))}{h^2} - y''(x_j);$$

разлагая все величины в ряд Тейлора в точке x_j , получим

$$r_j = \frac{y^{(4)}(x_j) h^2}{12} + O(h^4). \quad (8)$$

Вычтем из $l(y(x_j))$ некоторое слагаемое, аппроксимирующее величину $\frac{y^{(4)}(x_j) h^2}{12}$; получаемой схеме соответствует погрешность аппроксимации более высокого порядка. Например, можно приблизить $y^{(4)}(x_j)$ величиной

$$\frac{\delta^4 y(x_j)}{h^4} = \frac{y(x_{j+2}) - 4y(x_{j+1}) + 6y(x_j) - 4y(x_{j-1}) + y(x_{j-2}))}{h^4},$$

получится конечно-разностная схема

$$\frac{\delta^2 y_j}{h^2} - \frac{\delta^4 y_j}{12h^2} - p_j y_j = f_j. \quad (9)$$

Эту схему можно также получить, непосредственно заменив производную $y''(x_j)$ выражением $h^{-2}(\delta^2 y(x_j) - (1/12)\delta^4 y(x_j))$, приближающим ее с погрешностью $O(h^4)$. Уравнение (9) содержит 5 неизвестных y_j с ненулевыми коэффициентами; решение системы, состоящей из уравнений (9) и уравнений, получающихся при аппроксимации граничных условий, более трудоемко, чем решение системы (3).

Исходя из других соображений, построим конечно-разностную схему с погрешностью аппроксимации $O(h^4)$ такую, что в каждое уравнение входит только три неизвестных. Дифференцируя дважды исходное уравнение, имеем $y^{(4)}(x) = (p(x)y + f)''$. Поэтому

$$\begin{aligned} y^{(4)}(x_j) &\approx \frac{\delta^2 (py + f)|_{x_j}}{h^2} = \\ &= \frac{(p_{j+1}y(x_{j+1}) + f_{j+1}) - 2(p_j y(x_j) + f_j) + (p_{j-1}y(x_{j-1}) + f_{j-1}))}{h^2}. \end{aligned}$$

Вычитая из исходной схемы приближение для $y^{(4)}h^2/12$, получим

схему $\frac{\delta^2 y_j}{h^2} - p_j y_j - \frac{1}{12} \delta^2 (p_j y_j + f_j) = f_j$ или

$$l^{(1)}(y_j) = \frac{\delta^2 y_j}{h^2} - p_j y_j - \frac{1}{12} \delta^2 (p_j y_j) = \bar{l}^{(1)}(f_j) = f_j + \frac{1}{12} \delta^2 f_j.$$

В предположении, что решение непрерывно дифференцируемо 8 раз, рассмотрим погрешность аппроксимации новой схемы $r_j^{(1)} = l^{(1)}(y(x_j)) - \bar{l}^{(1)}(f_j)$, учитывая, что $p_j y(x_j) + f_j = y''(x_j)$. Получим

$$r_j^{(1)} = l^{(1)}(y(x_j)) - \bar{l}^{(1)}(f_j) = \frac{\delta^2 y(x_j)}{h^2} - y''(x_j) - \frac{1}{12} \delta^2 y''(x_j).$$

После разложения всех величин в ряд Тейлора по степеням h в точке x_j имеем равенство

$$r_j^{(1)} = -\frac{y^{(6)}(x_j) h^4}{240} + O(h^6).$$

Построим приближение величины $y^{(6)}(x_j)$ посредством значений $y(x_{j-1})$, $y(x_j)$, $y(x_{j+1})$. Таких приближений можно написать очень много, например, следующим образом. Согласно уравнению (1),

$$y'' = py + f, \quad y''' = (py + f)', \quad y^{(4)} = (py + f)'', \\ y^{(6)} = (py + f)^{(4)} = py^{(4)} + 4p'y^{(3)} + 6p''y'' + 4p^{(3)}y' + p^{(4)}y + f^{(4)};$$

поэтому справедливо приближенное равенство

$$y^{(6)}(x_j) \approx p_j \frac{\delta^2 (p_j y(x_j) + f_j)}{h^2} + \\ + 4p'(x_j) \frac{(p_{j+1}y(x_{j+1}) + f_{j+1}) - (p_{j-1}y(x_{j-1}) + f_{j-1})}{2h} + \\ + 6p''(x_j) (p_j y(x_j) + f_j) + 4p^{(3)}(x_j) \frac{y(x_{j+1}) - y(x_{j-1}))}{2h} + \\ + p^{(4)}(x_j) y(x_j) + f^{(4)}(x_j).$$

Прибавляя к $l^{(1)}(y_j)$ выражение, приближающее $\frac{y^{(6)}(x_j) h^4}{240}$, получим конечно-разностную схему с погрешностью аппроксимации $O(h^6)$; при этом в каждое уравнение полученной алгебраической системы входят только три неизвестных y_j .

Для практической оценки погрешности решения краевой задачи может применяться правило Рунге. Законность его применения основывается на существовании главного члена погрешности.

Задача. Пусть $p(x)$ и $f(x)$ четырежды дифференцируемы. Доказать, что для решения задачи (3)–(4) справедливо соотношение

$$\max_n |y_n - y(x_n) - h^2 z(x_n)| = O(h^4),$$

где $z(x)$ — решение краевой задачи

$$Lz = -y'(x)/12, \quad z(0) = z(X) = 0.$$

У к з а н и е. Воспользоваться построениями из § 5 гл. VIII.

Рассмотрим еще одну группу методов повышения точности приближенных решений, наиболее известную под названием *метода уточнения разностями высших порядков*.

Предположим, что после решения разностной задачи (3) удастся вычислить значения некоторой величины r_j^1 такой, что

$$\max_{0 < j < N} |e_j^1| = O(h^4), \quad \text{здесь} \quad e_j^1 = r_j - r_j^1.$$

Найдем решение R_j^1 сеточной задачи

$$l(R_j^1) = r_j^1, \quad j = 1, \dots, N-1, \quad R_0^1 = R_N^1 = 0,$$

и положим $y_j^1 = y_j + R_j^1$. Имеем равенства

$$l(y(x_j) - y_j^1) = l(y(x_j)) - l(y_j) - l(R_j^1) = r_j - r_j^1 = e_j^1.$$

Воспользовавшись леммой 2, получим оценку

$$|y(x_j) - y_j^1| \leq \frac{X^2}{8} \max_{0 < j < N} |e_j^1| = O(h^4). \quad (10)$$

В соответствии с (8) требуемые величины r_j^1 должны удовлетворять соотношению

$$\max_j \left| r_j^1 - \frac{y^{(4)}(x_j) h^2}{12} \right| = O(h^4). \quad (11)$$

Если $p(x)$ и $f(x)$ — достаточно гладкие, то можно приблизить $y^{(4)}(x_j)$ величиной $\frac{\delta^4 y_j}{h^4}$ и, соответственно, положить $r_j^1 = \frac{\delta^4 y_j}{12h^2}$. Правая часть этого соотношения не определена при $j = 1$ и $j = N-1$. Поэтому для определения приближений к $y^{(4)}(x_1)$ и $y^{(4)}(x_{N-1})$ следует воспользоваться какими-то другими формулами численного дифференцирования точности $O(h^4)$.

Приведенные выше высказывания о возможности приближения производных $y^{(4)}(x_j)$ требуют дополнительного обоснования.

Для получения требуемых величин r_j^1 можно поступить и следующим образом. Вследствие уравнения (1) имеем

$$y^{(4)}(x) = (y''(x))'' = (p(x)y(x) + f(x))'' = \frac{\delta^2(p_j y_j + f_j)}{h^2} + O(h^2).$$

Это равенство наводит на мысль положить

$$r_j^1 = \frac{\delta^2(p_j y_j + f_j)}{12};$$

можно показать, что и в этом случае выполнено (11), а следовательно и (10). Практика вычислений указывает на предпочтительность второго способа получения r_j^1 .

Общая схема уточнения решения разностями высшего порядка может быть описана следующим образом.

Строится уравнение, которому приближенно удовлетворяет ошибка полученного приближения, и затем решается уравнение относительно этой ошибки.

Иначе, пусть имеется приближение y_j^m с погрешностью $O(h^{2m+2})$.

Строится выражение r_j^{m+1} , эффективно вычисляемое через значения приближения y_j^m и обладающее следующим свойством: если R_j^{m+1} — решение сеточной задачи

$$l(R_j^{m+1}) = r_j^{m+1}, \quad j = 1, \dots, N-1, \quad R_0^{m+1} = R_N^{m+1} = 0,$$

то приближение $y_j^{m+1} = y_j^m + R_j^{m+1}$ имеет погрешность $O(h^{2m+4})$.

В соответствии с вышесказанным вычисляются r_j^{m+1} , R_j^{m+1} и затем y_j^{m+1} .

К удобству пользования описанными методами относится то, что, независимо от порядка точности получаемого приближения, решаются системы линейных алгебраических уравнений с одной и той же матрицей.

Рассмотрим случай граничного условия $y'(0) - \alpha y(0) = a$. Дискретное приближение высокой точности к такому граничному условию можно получить непосредственно, заменив производную $y'(0)$ по какой-либо формуле численного дифференцирования высокой точности

$$y'(0) \approx \sum_{i=0}^l \frac{c_i y(x_i)}{h}.$$

Однако погрешность аппроксимации будет меньше и решение возникающей алгебраической системы представит меньше трудностей, если идти по описанному выше пути последовательного повышения порядка точности аппроксимации.

Заменив производную $y'(0)$ отношением $\frac{y(h) - y(0)}{h}$, получим уравнение $l_0^{(1)}(y_j) = \frac{y_1 - y_0}{h} - \alpha y_0 - a = 0$. Разложим выражение $r_0^{(1)} = l_0^{(1)}(y(x_i)) = \frac{y(h) - y(0)}{h} - \alpha y(0) - a$ в ряд Тейлора по степеням h ; получим

$$r_0^{(1)} = y'(0) + \frac{y''(0)h}{2} + O(h^2) - \alpha y(0) - a = \frac{y''(0)h}{2} + O(h^2).$$

Таким образом, погрешность аппроксимации граничного условия $O(h)$. Поскольку, согласно уравнению (1),

$$y''(0) = p_0 y(0) + f_0,$$

то уравнению

$$l_0^{(2)}(y_j) = \frac{y_1 - y_0}{h} - \alpha y_0 - a - (p_0 y_0 + f_0) \frac{h}{2} = 0$$

соответствует второй порядок аппроксимации. Разлагая в ряд Тейлора по степеням h выражение $r_0^{(2)} = l_0^{(2)}(y(x_j))$, получим

$$r_0^{(2)} = \frac{y^{(3)}(0) h^2}{6} + O(h^3). \text{ После дифференцирования исходного}$$

уравнения (1) имеем

$$y^{(3)}(0) - p_0 y'(0) - p'(0) y(0) - f'(0) = 0;$$

поэтому с учетом граничного условия справедливо равенство

$$y^{(3)}(0) = p_0(\alpha y(0) + a) + p'(0) y(0) + f'(0).$$

Разностному уравнению

$$l_0^{(3)}(y_j) = l_0^{(2)}(y_j) - ((p_0 \alpha + p'(0)) y_0 + p_0 a + f'(0)) \frac{h^2}{6} = 0$$

будет соответствовать уже третий порядок аппроксимации

Можно было бы сразу написать равенство

$$r_0^{(1)} = y''(0) \frac{h}{2} + y^{(3)}(0) \frac{h^2}{6},$$

затем выразить производные $y''(0)$ и $y^{(3)}(0)$ через $y(0)$ и, вычитая из разностной схемы соответствующее выражение, получить уравнение $l_0^{(3)}(y_j) = 0$. Однако мы обратили основное внимание именно на способ последовательного повышения порядка точности, поскольку его перенесение на случай уравнений в частных производных является наиболее простым и естественным.

§ 2. Функция Грина сеточной краевой задачи

Функция w_{τ} , введенная в § 1 при доказательстве леммы 2, носит название *мажорирующей функции*, и метод получения оценок погрешности с использованием этой функции называется *методом мажорант*, или *методом Гершгорина*.

В ряде случаев, когда метод мажорант неприменим, оценку погрешности приближенного решения удается получить, используя так называемую *сеточную функцию Грина*. Проводимые ниже построения функции Грина сеточной краевой задачи (1.3) — (1.4) кроме всего прочего интересны своей аналогией со случаем дифференциальной краевой задачи.

Функция Грина $G(x, s)$ дифференциальной краевой задачи

$$Ly = y'' - p(x)y = f(x), \quad y(0) = a, \quad y(X) = b$$

определяется как решение уравнения

$$L(G(x, s)) = G''_{xx}(x, s) - p(x)G(x, s) = \delta(x - s) \quad (1)$$

при граничном условии $G(0, s) = G(X, s) = 0$. Решение задачи (1.1) — (1.2) записывается с помощью функции Грина в виде

$$y(x) = \int_0^x G(x, s)f(s)ds + G_s(x, X)b - G_s(x, 0)a. \quad (2)$$

Функцию Грина можно задать следующими явными формулами. Пусть $W^1(x)$, $W^2(x)$ — решения уравнения $L(W) = 0$ при условиях

$$\begin{aligned} W^1(0) &= 0, & (W^1)'(0) &= 1, \\ W^2(X) &= 0, & (W^2)'(X) &= -1; \end{aligned}$$

тогда

$$G(x, s) = \begin{cases} \frac{W^2(s)W^1(x)}{V^0} & \text{при } 0 \leq x \leq s, \\ \frac{W^1(s)W^2(x)}{V^0} & \text{при } s \leq x \leq X, \end{cases}$$

где

$$V(x) = \begin{vmatrix} W^1(x) & W^2(x) \\ (W^1)'(x) & (W^2)'(x) \end{vmatrix} = \text{const} = V^0.$$

Перейдем к сеточной задаче

$$l(y_j) = \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} - p_j y_j = f_j, \quad y_0 = a, \quad y_N = b.$$

По аналогии определим функции W_n^1 , W_n^2 из соотношений

$$\begin{aligned} l(W_j^i) &= 0, & i &= 1, 2, & j &= 1, \dots, N-1, \\ W_0^1 &= 0, & W_1^1 &= h, & W_N^2 &= 0, & W_{N-1}^2 &= h. \end{aligned}$$

Аналогом определителя Вронского является определитель

$$l_n = \begin{vmatrix} W_n^1 & W_n^2 \\ \frac{W_n^1 - W_{n-1}^1}{h} & \frac{W_n^2 - W_{n-1}^2}{h} \end{vmatrix} = \frac{W_n^2 W_{n-1}^1 - W_n^1 W_{n-1}^2}{h}.$$

Из тождества

$$\begin{aligned} 0 &= W_n^1 l(W_n^2) - W_n^2 l(W_n^1) = \\ &= \frac{W_n^1 W_{n+1}^2 + W_n^1 W_{n-1}^2 - W_n^2 W_{n+1}^1 - W_n^2 W_{n-1}^1}{h^2} = \frac{V_{n+1} - V_n}{h^2} \end{aligned}$$

следует, что величина V_n не зависит от n ; будем обозначать ее $V(h)$.

Положим

$$G_n^k = \begin{cases} \frac{W_k^2 W_n^1}{V(h)} & \text{при } 0 \leq n \leq k, \\ \frac{W_k^1 W_n^2}{V(h)} & \text{при } k \leq n \leq N. \end{cases}$$

Будем обозначать далее

$$l(G_n^k) = \frac{G_{n+1}^k - 2G_n^k + G_{n-1}^k}{h^2} - p_n G_n^k.$$

Из определения W_n^1 следуют равенства:

$$l(G_n^k) = \frac{W_k^2}{V(h)} l(W_n^1) = 0$$

при $n < k$,

$$l(G_n^k) = \frac{W_k^1}{V(h)} l(W_n^2) = 0,$$

при $n > k$; при $k = n$ имеем

$$\begin{aligned} l(G_n^k) &= \frac{G_{n+1}^n - 2G_n^n + G_{n-1}^n}{h^2} - p_n G_n^n = \\ &= \left(\frac{W_{n+1}^2 W_n^1 - 2W_n^1 W_n^2 + W_{n-1}^1 W_n^2}{V(h) h^2} - \frac{p_n W_n^1 W_n^2}{V(h)} \right) = \\ &= \frac{W_n^1}{V(h)} \left(\frac{W_{n+1}^2 - 2W_n^2 + W_{n-1}^2}{h^2} - p_n W_n^2 \right) + \frac{(W_n^2 W_{n-1}^1 - W_n^1 W_{n-1}^2)}{V(h) h^2}. \end{aligned}$$

Согласно определению W_n^2 , первая круглая скобка равна 0; вторая скобка равна $V(h)h$. В итоге при $k = n$ получаем $l(G_n^k) = h^{-1}$; объединяя полученные соотношения, имеем

$$l(G_n^k) = \delta_{n-k} h^{-1}. \quad (3)$$

Функция $\delta_{n-k} h^{-1}$ является сеточным аналогом δ -функции, а равенство (3) — аналогом (1).

Произведем оценку близости функций Грина дифференциальной и сеточной задач. Поскольку $(W^1)''_{xx}(0) = p(0) W^1(0) = 0$, то

$$W^1(h) = W^1(0) + (W^1)'(0)h + \frac{(W^1)''(0)h^2}{2} + O(h^3) = h + O(h^3).$$

Функция W_n^1 является решением конечно-разностного уравнения, аппроксимирующего уравнение $W'' - p(x)W = 0$. Характеристическое уравнение разностной схемы $\mu^2 - 2\mu + 1 = 0$ имеет

двукратный корень $\mu = 1$. Таким образом, условия теоремы из § 14 гл. VIII выполнены. Положим $R_n = W_n^1 - W^1(nh)$. Воспользовавшись оценкой (8.14.5), в конкретных условиях рассматриваемой задачи имеем

$$\max_{0 \leq x_n \leq X} |R_n| \leq C(P, X) \left(h \sum_{i=1}^{N-1} |r_i| + \max(|R_0|, |R_1|) + \left| \frac{R_1 - R_0}{h} \right| \right);$$

здесь r_i — погрешность аппроксимации:

$$r_i = \frac{(W^1)^{(4)}(\bar{x}_i) h^2}{12}, \quad P = \max_{[0, X]} |p(x)|.$$

Согласно определению W_0^1 и W_1^1 , имеем $R_0 = 0$, $R_1 = O(h^3)$ и, следовательно, $\frac{R_1 - R_0}{h} = O(h^2)$. Поэтому из последней оценки следует неравенство

$$\max_{0 \leq x_n \leq X} |W_n^1 - W^1(nh)| \leq Mh^2; \quad (4)$$

здесь M — некоторая постоянная. Точно так же получается оценка

$$\max_{0 \leq x_n \leq X} |W_n^2 - W^2(nh)| \leq Mh^2. \quad (5)$$

Согласно свойствам V^0 и $V(h)$, имеем

$$V^0 = V(X) = \begin{vmatrix} W^1(X) & 0 \\ (W^1)'(X) & -1 \end{vmatrix} = -W^1(X),$$

$$V(h) = \begin{vmatrix} W_N^1 & 0 \\ \frac{W_N^1 - W_{N-1}^1}{h} & -1 \end{vmatrix} = -W_N^1.$$

Поэтому

$$|V(h) - V^0| \leq Mh^2$$

на основании оценки (4).

Далее предполагается, что $W^1(X) \neq 0$; в противном случае однородная краевая задача $y'' - p(x)y = 0$, $y(0) = y(X) = 0$ имеет ненулевое решение $W^1(x)$ а, следовательно, неоднородная краевая задача (1.1)–(1.2) или не имеет решения, или имеет неединственное решение. Будем также предполагать h настолько малым, что $2Mh^2 \leq |W^1(X)| = |V^0|$. В этом случае имеем

$$|V(h)| \geq |V^0| - Mh^2 \geq \left| \frac{V^0}{2} \right| > 0.$$

При таких h

$$\left| \frac{1}{V(h)} - \frac{1}{V^0} \right| = \left| \frac{V^0 - V(h)}{V(h)V^0} \right| \leq \frac{2Mh^2}{|V^0|^2}. \quad (6)$$

Сравнивая явные выражения функций Грина дифференциальной и сеточной задач с учетом оценок (4)–(6), получаем

$$|G_n^k - G(nh, kh)| \leq Qh^2 \quad (7)$$

при $0 \leq kh, nh \leq X$.

Лемма. Если $V(h) \neq 0$, то рассматриваемая сеточная задача (1.3)–(1.4) однозначно разрешима и ее решение записывается в виде

$$y_n = h \sum_{k=1}^{N-1} G_n^k f_k + \frac{W_n^2}{W_0^2} a + \frac{W_N^1}{W_N^1} b. \quad (8)$$

Формула (8) является сеточным аналогом формулы (2).

Доказательство. Поскольку $G_0^k = G_N^k = 0$, согласно определению функции G_n^k , то $y_0 = a$, $y_N = b$. Имеем равенство

$$l(y_n) = h \sum_{k=1}^{N-1} l(G_n^k) f_k + \frac{l(W_n^2)}{W_0^2} a + \frac{l(W_N^1)}{W_N^1} b;$$

коэффициенты при a и b равны нулю по определению функций W_n^i . Воспользовавшись равенством (3), получим $l(y_n) = f_n$. Таким образом, система линейных уравнений (1.3)–(1.4) при любых правых частях имеет решение, записываемое в виде (8). Но если система линейных уравнений разрешима при любых правых частях, то её решение единственно. Лемма доказана.

Соотношение (8) может быть двояким способом использовано для оценки близости решений сеточной и дифференциальной задач.

Первый путь состоит в непосредственном сравнении выражений (2) и (8). Для простоты продемонстрируем его в случае $a = b = 0$, $\delta_n \equiv 0$. Подставляя $x = nh$, перепишем (2) в виде

$$y(nh) = \int_0^{nh} G(nh, s) f(s) ds + \int_{nh}^X G(nh, s) f(s) ds.$$

Если $p(x)$ и $f(x)$ дважды непрерывно дифференцируемы, то можно показать, что подынтегральные функции в обоих интегралах также дважды непрерывно дифференцируемы. Заменим

оба эти интеграла по формуле трапеций с шагом h . Получим

$$y(nh) = h \left(\frac{G(nh, 0)f(0)}{2} + \sum_{k=1}^{n-1} G(nh, kh)f(kh) + \frac{G(nh, nh)f(nh)}{2} \right) + \\ + h \left(\frac{G(nh, nh)f(nh)}{2} + \sum_{k=n+1}^{N-1} G(nh, kh)f(kh) + \right. \\ \left. + \frac{G(nh, Nh)f(Nh)}{2} \right) + O(h^2).$$

Поскольку $G(nh, 0) = G(nh, Nh) = 0$, по определению функции Грина $G(x, s)$, то

$$y(nh) = h \sum_{k=1}^{N-1} G(nh, kh)f_k + O(h^2). \quad (9)$$

Разбиение интеграла на две части нам потребовалось в связи с тем, что погрешность формулы трапеций оценивается через вторую производную подынтегральной функции, а функция $G(nh, s)$ имеет разрыв первой производной в точке nh . Из равенств (8), (9) получаем

$$y_n - y(nh) = h \sum_{k=1}^{N-1} (G_n^k - G(nh, kh))f_k + O(h^2).$$

Воспользовавшись оценкой (7), получим $y_n - y(nh) = O(h^2)$. Нетрудно проследить, что эта оценка погрешности равномерна по n при $0 \leq x_n \leq X$, т. е.

$$\max_{0 \leq x_n \leq X} |y_n - y(nh)| = O(h^2).$$

Другой, более привычный для нас, путь состоит в получении уравнения для погрешности и дальнейшей оценки погрешности с помощью формулы Грина. В § 1 было показано, что ошибка R_n удовлетворяет уравнению

$$l(R_n) = \delta_n - r_n, \quad \text{где} \quad |r_n| \leq \frac{M_4 h^2}{12};$$

δ_n — ошибка, обусловленная неточностью решения системы (1.3) — (1.4). Пусть $R_0 = R_N = 0$. Из явного вида функции Грина дифференциальной задачи следует ее равномерная ограниченность; с учетом (7) получаем, что функция G_n^k также равномерно ограничена: $|G_n^k| \leq D$ при $0 \leq kh, nh \leq X$. Напишем явное представление R_n с помощью формулы Грина (9):

$$R_n = h \sum_{k=1}^{N-1} G_n^k (\delta_k - r_k). \quad (10)$$

Тогда из (10) следует оценка

$$|R_n| \leq Dh \sum_{k=1}^{N-1} (|\delta_k| + |r_k|).$$

Загрубляя, можно получить более наглядную оценку

$$|R_n| \leq DX \left(\frac{M_4}{12} h^2 + \max_{0 < k < N} |\delta_k| \right),$$

имеющую в случае $p(x) \geq 0$ тот же порядок по отношению к h и $\max |\delta_k|$, что и полученная ранее оценка. С помощью описанных способов иногда удается получить оценки погрешности при довольно слабых предположениях о гладкости коэффициентов задачи; особенно полезным в этом смысле является первый из этих способов.

§ 3. Решение простейшей краевой сеточной задачи

При численном решении задачи Коши значения решения определяются в последовательных узлах по рекуррентным формулам; в случае краевой сеточной задачи, например (1.3) — (1.4), такой возможности нет, поскольку значения решения зависят от граничных условий на обоих концах отрезка интегрирования.

Краевую задачу

$$y'' - p(x)y = f(x), \quad y(0) = a, \quad y(X) = b$$

можно было бы решать следующим способом. Возьмем частное решение $y_0(x)$ неоднородного уравнения

$$y_0'' - p(x)y_0 = f(x)$$

и два линейно независимых решения однородного уравнения $y_i'' - p(x)y_i = 0$, $i = 1, 2$. Общее решение неоднородного уравнения запишется в виде

$$y_0(x) + C_1 y_1(x) + C_2 y_2(x);$$

постоянные C_1 и C_2 определяем из граничных условий. Приближения к функциям $y_i(x)$, $i = 0, 1, 2$, находим каким-либо численным методом решения задачи Коши; затем определяем C_1 и получаем нужное решение.

Экономнее поступить следующим образом. Находим частное решение неоднородного уравнения $y_0'' - p(x)y_0 = f(x)$, удовлетворяющее условию $y_0(0) = a$, и частное решение однородного уравнения $y_1(x)$, удовлетворяющее условию $y_1(0) = 0$. Общее решение неоднородного уравнения, удовлетворяющее условию

$y(0) = a$, имеет вид $y_0(x) + Cy_1(x)$; значение C определяется из условия $y_0(X) + Cy_1(X) = b$.

Метод решения краевой задачи, следующий этой схеме, принято называть *методом стрельбы* или *методом пристрелки*. Сеточный аналог этого метода заключается в следующем. Задаемся $y_0^0 = a$, $y_0^1 = 0$, произвольными $y_1^1 \neq 0$ и y_1^0 , и из уравнений

$$\frac{y_{n+1}^0 - 2y_n^0 + y_{n-1}^0}{h^2} - p_n y_n^0 = f_n, \quad (1)$$

$$\frac{y_{n+1}^1 - 2y_n^1 + y_{n-1}^1}{h^2} - p_n y_n^1 = 0 \quad (2)$$

последовательно определяем $y_2^0, \dots, y_N^0$, $y_2^1, \dots, y_N^1$. Затем находим C из уравнения $y_N^0 + Cy_N^1 = b$ и полагаем $y_n = y_n^0 + Cy_n^1$; функция y_n является требуемым решением.

Если все указанные выше величины невозможно поместить в оперативной памяти ЭВМ, то значения y_n^i , $i = 0, 1$, не хранят в процессе вычислений, а после отыскания C находят $y_1 = y_1^0 + Cy_1^1$ и затем последовательно определяют $y_2, \dots, y_{N-1}$ из уравнения $l(y_n) = f_n$. Описанный алгоритм формально применим при любых значениях y_1^0 и y_1^1 ; однако для уменьшения влияния вычислительной погрешности разумнее взять $y_1^0 = a + O(h)$.

Приведем соображения в пользу этого утверждения. Получаемые при реальных вычислениях значения $\bar{y}_n^0$ связаны соотношениями

$$\bar{y}_{n+1}^0 = 2\bar{y}_n^0 - \bar{y}_{n-1}^0 + h^2(p_n \bar{y}_n^0 - f_n) + \delta_n;$$

результат вычислений округляется в машине с относительной погрешностью порядка 2^{-t} ; поэтому величина δ_n может достигать значений порядка $2^{-t} y_{n+1}^0$. Введем вспомогательные функции: y_n^2 — решение (1) при начальных условиях $y_0^2 = y_1^2 = a$, y_n^3 — решение (2) при начальных условиях $y_0^3 = 0$, $y_1^3 = h$; пусть $y^2(x)$ — решение уравнения $y'' - p(x)y = f(x)$ при начальных условиях $y(0) = a$, $y'(0) = 0$. Мы показали в § 2, что

$$\max_{0 \leq x_n \leq X} |y_n^3 - W^1(nh)| \leq Mh^2, \quad (3)$$

на том же пути оценки близости решения y_n^2 сеточной и решения $y^2(nh)$ дифференциальной задач с помощью теоремы из § 14 гл. VIII можно получить оценку

$$\max_{0 \leq x_n \leq X} |y_n^2 - y^2(nh)| = O(h). \quad (4)$$

Функция $y_n^2 + y_n^3(y_1^0 - a)h^{-1}$ совпадает с y_n^0 при $n=0, 1$ и удовлетворяет тому же конечно-разностному уравнению, поэтому $y_n^0 = y_n^2 + y_n^3(y_1^0 - a)h^{-1}$. С учетом оценок (3)–(4) имеем

$$y_n^0 = (y^2(nh) + O(h)) + (W^1(nh) + O(h))(y_1^0 - a)h^{-1}.$$

Отсюда видно, что при $|(y_1^0 - a)h^{-1}| \gg 1$ величины y_n^0 оказываются порядка $(y_1^0 - a)h^{-1}$. Поскольку значения y_n^0 и $\bar{y}_n^0$ отличаются на относительно малые величины, то значения $\bar{y}_n^0$ также порядка $(y_1^0 - a)h^{-1}$. Отсюда следует вывод о желательности выбора значения y_1^0 , при котором величина $(y_1^0 - a)h^{-1}$ невелика.

Многие краевые задачи характерны тем, что среди решений соответствующих им однородных дифференциальных уравнений есть решения с сильным ростом (при убывании или при возрастании x).

Рассмотрим модельное уравнение

$$y'' - py = 0, \quad p = \text{const} > 0. \quad (5)$$

Решения соответствующего однородного уравнения $\exp(\pm \sqrt{p}x)$ резко возрастают (убывают) на рассматриваемом участке, если величина $\sqrt{p}X$ достаточно велика. Таким образом параметрами, определяющими влияние вычислительной погрешности при различных способах решения сеточной задачи (1.3)–(1.4), являются $\sqrt{p}X$ и h . Рассмотрим модель накопления вычислительной погрешности на одном из этапов решения сеточной задачи. Пусть уже найдено значение y_1 , причем оно найдено без всякой погрешности, а далее вычисления происходят по рекуррентной формуле

$$y_{n+1} = 2y_n - y_{n-1} + ph^2y_n. \quad (5')$$

Предположим, что получаемые при реальных вычислениях величины y_n^* связаны равенством

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + ph^2y_n^* + \delta,$$

где δ порядка 2^{-t} . Вычитая отсюда (5'), получим уравнение для ошибки $\Delta_n = y_n^* - y_n$:

$$\Delta_{n+1} = 2\Delta_n - \Delta_{n-1} + ph^2\Delta_n + \delta, \quad (6)$$

причем, по предположению, $\Delta_0 = \Delta_1 = 0$. Характеристическое уравнение этого разностного уравнения

$$\mu^2 - (2 + ph^2)\mu + 1 = 0$$

имеет корни

$$\mu_{1,2} = 1 + \frac{ph^2}{2} \pm \sqrt{ph^2 + \frac{(ph^2)^2}{4}}.$$

Частное решение неоднородного уравнения отыскиваем в виде $\Delta_n = c$. Подставляя в (6), получим $c = -\frac{\delta}{ph^2}$. Таким образом, общее решение неоднородного уравнения имеет вид

$$\Delta_n = C_1 \mu_1^n + C_2 \mu_2^n - \frac{\delta}{ph^2}; \quad (7)$$

из условия $\Delta_0 = \Delta_1 = 0$ определяем значения C_1, C_2 . Окончательно имеем

$$\Delta_n = \frac{\delta}{ph^2} \left(\frac{1 - \mu_2}{\mu_1 - \mu_2} \mu_1^n + \frac{1 - \mu_1}{\mu_2 - \mu_1} \mu_2^n - 1 \right).$$

Будем считать, что величина $\sigma = \sqrt{ph^2}$ мала. Тогда

$$\mu_{1,2} = 1 \pm \sigma + O(\sigma^2),$$

и следовательно,

$$\frac{1 - \mu_1}{\mu_1 - \mu_2} = \frac{\sigma + O(\sigma^2)}{2\sigma + O(\sigma^2)} = \frac{1}{2} + O(\sigma);$$

точно так же

$$\frac{1 - \mu_2}{\mu_2 - \mu_1} = \frac{1}{2} + O(\sigma).$$

Поскольку в то же время $\mu_{1,2} = \exp(\pm \sigma + O(\sigma^2))$, то мы можем записать равенство (7) в виде

$$\Delta_n = \frac{\delta}{ph^2} \left(\left(\frac{1}{2} + O(\sigma) \right) \exp(\sigma n (1 + O(\sigma))) + \right. \\ \left. + \left(\frac{1}{2} + O(\sigma) \right) \exp(-\sigma n (1 + O(\sigma))) - 1 \right).$$

При $nh = x$ имеем $\sigma n = \sqrt{p} x$ и

$$\Delta_n = \frac{\delta}{ph^2} \left(\left(\frac{1}{2} + O(\sigma) \right) \exp(\sqrt{p} x (1 + O(\sigma))) + \right. \\ \left. + \left(\frac{1}{2} + O(\sigma) \right) \exp(-\sqrt{p} x (1 + O(\sigma))) - 1 \right). \quad (8)$$

Если величина $\sqrt{p} x$ велика, то при x , близких к X , влияние вычислительной погрешности, в соответствии с формулой (8), будет катастрофически большим. Примечательно, что во многих случаях это влияние вызвано не множителем h^{-2} ,

являющимся следствием малости шага, а множителем $\exp(\sqrt{p}x(1+O(\sigma)))$. Например, в случае краевой задачи для $y'' - y = f$ на отрезке $[0, 30]$ имеем $\exp(\sqrt{1} \cdot 30) = 10^{12}, \dots$, в то время как шаги интегрирования $h < 10^{-3}$ применяются редко.

В случае $p(x) \geq 0$ рассматриваемую систему уравнений (1.3)–(1.4) имеет смысл решать с помощью описываемого ниже метода прогонки. Представим граничное условие $y_0 = a$ в виде $y_0 = C_0 y_1 + \varphi_0$, где $C_0 = 0$, $\varphi_0 = a$. Подставляя $y_0 = C_0 y_1 + \varphi_0$ в первое уравнение системы (1.3)–(1.4)

$$\frac{y_0 - 2y_1 + y_2}{h^2} - p_1 y_1 = f_1,$$

получаем уравнение, связывающее значения y_1 и y_2 . Разрешая это уравнение относительно y_1 , имеем

$$y_1 = C_1 y_2 + \varphi_1, \quad (9)$$

где

$$C_1 = \frac{1}{2 + p_1 h^2}, \quad \varphi_1 = C_1(a - f_1 h^2);$$

подставляя полученное выражение y_1 через y_2 во второе уравнение системы, получим уравнение, связывающее y_2 и y_3 и т. д. Пусть уже получено соотношение

$$y_n = C_n y_{n+1} + \varphi_n; \quad (10)$$

подставим выражение y_n в $(n+1)$ -е уравнение системы (1.3)–(1.4):

$$\frac{y_n - 2y_{n+1} + y_{n+2}}{h^2} - p_{n+1} y_{n+1} = f_{n+1}.$$

Разрешая получившееся уравнение

$$\frac{(C_n y_{n+1} + \varphi_n) - 2y_{n+1} + y_{n+2}}{h^2} - p_{n+1} y_{n+1} = f_{n+1},$$

будем иметь

$$y_{n+1} = C_{n+1} y_{n+2} + \varphi_{n+1},$$

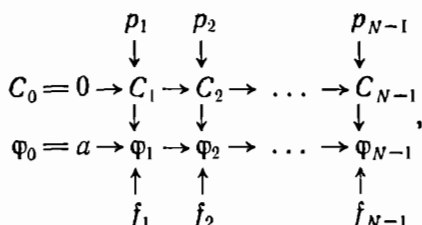
где

$$C_{n+1} = \frac{1}{2 + p_{n+1} h^2 - C_n}, \quad \varphi_{n+1} = C_{n+1}(\varphi_n - f_{n+1} h^2). \quad (11)$$

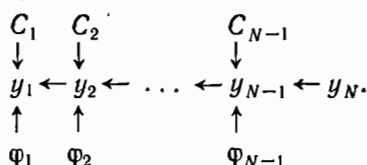
Таким образом, коэффициенты уравнений (10), связывающих последующие значения y_n и y_{n+1} , можно определять из рекуррентных соотношений (11) при начальных условиях $C_0 = 0$, $\varphi_0 = a$. Так как y_N известно, то после нахождения всех коэффициентов C_n , φ_n можно последовательно определять $y_{N-1}, \dots, y_1$ из соотношений (10). Процесс вычисления коэффициентов C_n , φ_n принято называть *прямым ходом прогонки*, а процесс вычисления неизвестных y_n — *обратным ходом прогонки*. Последователь-

ность производимых вычислений можно изобразить следующей схемой:

прямой ход прогонки



обратный ход прогонки



Названию «метод прогонки» иногда предлагается следующее объяснение. Уравнение

$$y_n = C_n y_{n+1} + \Phi_n$$

получено как следствие граничного условия в точке $x = 0$ и уравнений системы (1.3), соответствующих точкам $x_j \leq x_n$. Таким образом, это равенство выполнено для любого решения системы уравнений $l(y_j) = f_j$, $j = 1, \dots, n$, удовлетворяющего левому граничному условию; граничное условие в точке $x = 0$ как бы «перегоняется» в текущую точку $x = x_n$.

Рассмотренные выше методы стрельбы и прогонки являются частными случаями метода Гаусса исключения неизвестных при решении системы (1.3)–(1.4). Например, после приведения n уравнений системы (1.3)–(1.4) к виду (10) при прямом ходе метода прогонки система уравнений приобретает вид

$$\begin{array}{rcl}
 y_1 - C_1 y_2 & & = \Phi_1, \\
 \vdots & & \vdots \\
 y_n - C_n y_{n+1} & & = \Phi_n, \\
 y_n - (2 + p_{n+1} h^2) y_{n+1} + y_{n+2} & = & f_{n+1} h^2 \\
 \vdots & & \vdots \\
 \vdots & & \vdots
 \end{array}$$

тот же, что и после исключения n неизвестных в прямом ходе метода Гаусса. После окончания прямого хода метода прогонки матрица системы окажется приведенной к треугольному виду. Обратный ход метода прогонки совпадает с обратным ходом метода Гаусса.

З а д а ч а. Выписать расчетные формулы метода квадратного корня в случае решения рассматриваемой системы. Сравнить трудоемкость этого метода с трудоемкостью метода прогонки.

Рассмотрим другой вариант метода Гаусса решения системы (1.3) — (1.4). Из первого уравнения выражаем y_2 через y_1 и подставляем в остальные уравнения. После этого во второе уравнение входят только неизвестные y_1 и y_3 ; выражаем y_3 через y_1 и подставляем в остальные уравнения и т. д. Пусть мы уже выразили через y_1 неизвестные $y_2, \dots, y_n$ и получили соотношения $y_j = \alpha_j y_1 + \beta_j$ при $2 \leq j \leq n$; для единообразия добавим сюда равенства

$$y_0 = \alpha_0 y_1 + \beta_0, \quad \text{где} \quad \alpha_0 = 0, \quad \beta_0 = a,$$

$$y_1 = \alpha_1 y_1 + \beta_1, \quad \text{где} \quad \alpha_1 = 1, \quad \beta_1 = 0.$$

Подставляя выражения y_{n-1} и y_n в уравнение

$$y_{n-1} - (2 + p_n h^2) y_n + y_{n+1} = f_n h^2,$$

получим

$$(\alpha_{n-1} y_1 + \beta_{n-1}) - (2 + p_n h^2) (\alpha_n y_1 + \beta_n) + y_{n+1} = f_n h^2,$$

или

$$y_{n+1} = \alpha_{n+1} y_1 + \beta_{n+1},$$

где

$$\alpha_{n+1} = (2 + p_n h^2) \alpha_n - \alpha_{n-1},$$

$$\beta_{n+1} = (2 + p_n h^2) \beta_n - \beta_{n-1} + f_n h^2. \quad (12)$$

Таким образом, коэффициенты α_n, β_n можно последовательно вычислять по рекуррентным формулам (12). После получения соотношения $y_N = \alpha_N y_1 + \beta_N$ определяем значение y_1 , а затем все y_j по формулам

$$y_j = \alpha_j y_1 + \beta_j. \quad (13)$$

Вследствие (12) значения α_j удовлетворяют однородному конечно-разностному уравнению (2) и начальным условиям $\alpha_0 = 0, \alpha_1 = 1$, значения β_j — неоднородному конечно-разностному уравнению (1) и начальным условиям $\beta_0 = a, \beta_1 = 0$. Таким образом, α_n совпадает с y_n^1 , а β_n — с y_n^0 в методе стрельбы при определенном выборе начальных условий. Функция $z_n = \alpha_n C + \beta_n$ удовлетворяет неоднородному конечно-разностному уравнению и левому граничному условию при любых C , причем $z_1 = \alpha_1 C + \beta_1 = C$. Следовательно, решая уравнение $y_N = \alpha_N C + \beta_N$ относительно C , мы как раз находим значение $C = y_1$, которое отыскивалось в методе стрельбы. Вычисления по формуле (13) соответствуют вычислению значений y_n по формуле $y_n = y_n^1 C + y_n^0$. Построенный метод совпадает с одним из вариантов метода стрельбы; в то же время он совпадает с методом Гаусса при порядке исключения неизвестных: $y_0, y_2, \dots, y_N, y_1$.

В случае систем уравнений с полностью заполненной матрицей A размерности $N \times N$ обычный метод исключения Гаусса требует $\asymp N^3$ арифметических операций.

Матрицы $A = [a_{ij}]$, у которых $a_{ij} = 0$ при $|i - j| > m$, называют $(2m + 1)$ -диагональными; мы видели, что в случае трехдиагональных матриц метод Гаусса, совпадающий с методом прогонки, требует существенно меньшего числа операций. Точно так же число операций будет существенно меньше в случае $(2m + 1)$ -диагональных матриц при $m \ll N$.

Рассмотрим типичную задачу, сводящуюся к системе уравнений с матрицей вида, близкого к $(2n + 1)$ -диагональному.

При сглаживании функций методом регуляризации в § 4 гл. V возникла следующая задача. Дана периодическая функция f_q целочисленного аргумента q , период равен $N + 1$ (там он был N). Требуется найти периодическую с тем же периодом функцию u_q , удовлетворяющую системе соотношений

$$\frac{\delta^{2n} u_q}{h^{2n}} + (-\lambda^2)^n u_q = f_q \quad \text{при всех } q$$

Выпишем эти соотношения при $q = 0, \dots, N$. Вследствие условия периодичности заменим значения u_q при $q < 0$ и при $q > N$, входящие в эту систему, соответственно, на u_{q+N+1} и u_{q-N-1} . В результате этого получится система $N + 1$ уравнений относительно неизвестных $u_0, \dots, u_N$. Элементы матрицы $A = [a_{ij}]$ этой системы определяются соотношениями $a_{ij} = a(|i - j|)$, где

$$a(k) = \begin{cases} a(0) = (-1)^n (C_{2n}^n h^{-2n} + \lambda^{2n}), \\ (-1)^{n-k} C_{2n}^{n-k} h^{-2n} & \text{при } 0 < k \leq n, \\ 0 & \text{при } n < k \leq N - n, \\ (-1)^{k+n-N-1} C_{2n}^{k+n-N-1} h^{-2n} & \text{при } N - n < k \leq N \end{cases}$$

$$A = \begin{pmatrix} a(0) & a(1) & \dots & a(n) & 0 & \dots & 0 & a(N-n+1) & \dots & a(N) \\ a(1) & & & & & & & & & \vdots \\ \vdots & & & & & & & & & \\ a(n) & & & & & & & & & \vdots \\ 0 & & & & & & & & & \vdots \\ \vdots & & & & & & & & & \\ 0 & & & & & & & & & \vdots \\ a(N-n+1) & & & & & & & & & \vdots \\ \vdots & & & & & & & & & \\ a(N) & \dots & & & \dots & & \dots & & & \end{pmatrix}.$$

Матрица A симметричная и положительно определенная.

Одним из возможных методов решения такой системы является метод квадратного корня. Другой метод — последовательное исключение неизвестных в их естественном порядке $u_1, u_2, \dots$. В обоих случаях процесс преобразования системы можно записать в виде рекуррентных формул, аналогичных методу прогонки. Оба метода требуют $O(n^2N)$ арифметических операций. Поскольку используются операции разных типов, то выбор между этими методами должен производиться с учетом относительной трудоемкости операций ЭВМ. Вычислительная погрешность в обоих случаях может достигать значений порядка $2^{-1}h^{-2n}$.

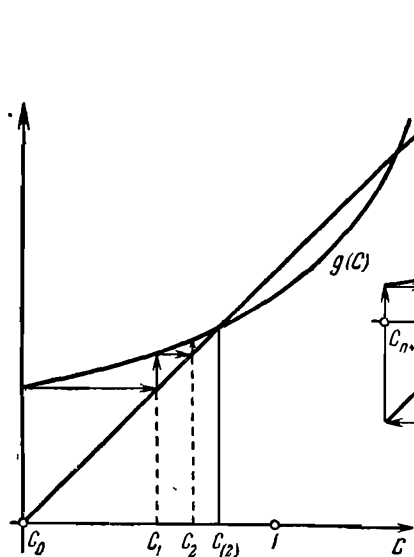


Рис. 9.3.1.

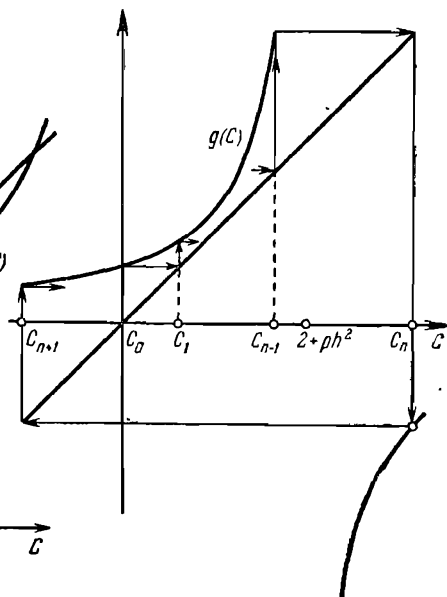


Рис. 9.3.2.

Задачи 1. Выписать расчетные формулы обоих методов в случае $n = 1$.

2. Подсчитать число арифметических действий различных видов для обоих методов при произвольном n .

В ряде случаев оказывается, что матрица A системы имеет $(2n+1)$ -диагональный клеточный вид. А именно, матрица разбивается на клетки A_{ij} такие, что $A_{ij} \equiv 0$ при $|i-j| > n$. Для решения систем с такими матрицами можно применять метод исключения по группам, соответствующим клеткам. В этом случае расчетные формулы имеют тот же вид, как в случае $(2n+1)$ -диагональной матрицы; отличие состоит лишь в замене скалярных операций на матричные.

На модельном примере $p(x) \equiv p = \text{const}$ рассмотрим поведение прогоночных коэффициентов C_n при различных знаках p . Соотношения

$$C_{n+1} = \frac{1}{2 + ph^2 - C_n},$$

которым удовлетворяют коэффициенты C_n метода прогонки совпадают с итерационными формулами решения уравнения

$$C = g(C) = \frac{1}{2 + ph^2 - C}$$

при начальном приближении $C_0 = 0$. Уравнение $C = g(C)$ равносильно квадратному уравнению

$$C^2 - (2 + ph^2)C + 1 = 0; \quad (14)$$

его корни

$$C_{(1)}, C_{(2)} = 1 + \frac{ph^2}{2} \pm \sqrt{ph^2 + \frac{p^2h^4}{4}}.$$

При $ph^2 > 0$, очевидно, $C_{(1)} > 1$, и поэтому $0 < C_{(2)} = C_{(1)}^{-1} < 1$. При $p < 0$ и h малом подкоренное выражение отрицательно и уравнение (14) не имеет вещественных корней. Совместное расположение графиков $y = C$ и $y = g(C)$ и точек (C_n, C_n) , (C_n, C_{n+1}) изображено на рисунках 9.3.1—9.3.2. Видно, что при $p \geq 0$ значения C_n лежат в пределах $0 < C_n \leq 1$ (рис. 9.3.1). При $p < 0$ не исключено (рис. 9.3.2), что при достаточно больших n некоторое значение C_n окажется близким к $2 + ph^2$; тогда следующее значение C_{n+1} будет очень большим. Это обстоятельство побуждает нас к более детальному изучению влияния вычислительной погрешности, по крайней мере при $p < 0$.

§ 4. Замыкания вычислительных алгоритмов

При предварительном анализе алгоритмов каждый раз удавалось многое выяснить, рассматривая случай решения модельной задачи. Другим эффективным приемом предварительного анализа является исследование замыканий вычислительных алгоритмов.

Чтобы не загромождать изложения, мы ограничимся рассмотрением существа проблемы; поэтому многие из проводимых в этом параграфе построений не всегда подробно обосновываются.

Пусть решается некоторая задача

$$Lu = f, \quad (1)$$

и пусть

$$L^h u^h = f^h \quad (2)$$

— последовательность задач, зависящих от параметра h (например, шага сетки), решения которых u^h сходятся при $h \rightarrow 0$ к решению исходной задачи. Предположим, что алгоритм решения задачи (2) $L^h u^h = f^h$ состоит в последовательном получении

некоторых соотношений

$$L_m^h u_m^h = f_m^h, \quad m = 1, \dots, M; \quad (3)$$

при этом $L_M^h = E$ — единичный оператор, $u_M^h = f_M^h = (L^h)^{-1} f^h = u^h$, т. е. на M -м шаге получается точное решение u^h . Пусть можно ввести параметр $z(m, h)$, монотонно зависящий от m такой, что при $h \rightarrow 0$, z фиксированном, соотношение (3) переходит в пределе в соотношение

$$L_z y = f_z, \quad 0 \leq z \leq z_0; \quad (4)$$

здесь

$$z_0 = \lim_{h \rightarrow 0} z(M, h).$$

Соотношение (4) называется *замыканием вычислительного алгоритма* (3).

В настоящий момент мы не рекомендуем читателю пытаться глубоко вникнуть в это нестрогое определение. Существо дела станет понятнее после рассмотрения замыканий алгоритмов решения краевой задачи.

Определение понятия вычислительного алгоритма не исключает возможности $M = \infty$; в этом случае равенства

$$L_M^h = E, \quad f_M^h = (L^h)^{-1} f^h$$

понимаются в том смысле, что

$$L_m^h \rightarrow E, \quad f_m^h \rightarrow (L^h)^{-1} f^h \quad \text{при} \quad m \rightarrow \infty.$$

Случай $M = \infty$ соответствует итерационным методам решения задачи (1). Если операторы L_z равномерно по z ограничены, то говорят, что алгоритм (3) имеет *регулярное замыкание*. В противном случае говорят, что алгоритм имеет *нерегулярное замыкание*.

Если замыкание алгоритма регулярно, то есть основания ожидать, что он будет устойчив к различным возмущениям, в частности к вычислительной погрешности. Поэтому исследование замыкания алгоритма является удобным способом получения предварительного суждения о свойствах нового метода на первоначальной стадии изучения вопроса. Однако такая предварительная оценка свойств метода не всегда оказывается окончательной. Может случиться, что операторы L_z равномерно ограничены, но очень большой постоянной, что равносильно практической неограниченности. С другой стороны, возможно, что неограниченность операторов L_z вызывается неудачной нормировкой уравнений (1.3) или неудачным выбором норм в пространствах функций u . Отсутствие равномерной ограниченности операторов L_z означает неограниченное возрастание $\rho(h) = \sup_m \|L_m^h\|$

при $h \rightarrow 0$. Наличие больших коэффициентов в уравнениях $L_m^h u_m^h = f_m^h$, казалось бы, должно быть равносильно большим возмущениям коэффициентов исходной задачи, а следовательно, большому влиянию вычислительной погрешности. Однако не исключена возможность, что величина $p(h)$ растет очень медленно и соответствующее возрастание влияния вычислительной погрешности при стремлении шага к нулю окажется допустимым. Несмотря на высказанные соображения, изучение замыканий вычислительных алгоритмов приносит большую пользу.

При прямом ходе прогонки получаются соотношения

$$y_n - C_n y_{n+1} = \varphi_n. \quad (5)$$

Попытаемся понять, во что переходят эти соотношения при $h \rightarrow 0$. Возьмем простейший случай $p \equiv 0$. Тогда $C_0 = 0$, $C_1 = 1/2$, $C_2 = 2/3$, ... и, как нетрудно установить методом индукции, $C_n = 1 - 1/(n+1)$. Пусть фиксирована какая-то точка $x = nh$. Тогда левая часть (5), соответствующая этой точке, равна

$$y_n - \left(1 - \frac{1}{\frac{x}{h} + 1}\right) y_{n+1}.$$

При подстановке вместо величин y_n значений гладкой функции $y(nh)$ предельная левая часть равна нулю, и нет смысла рассматривать подобное замыкание алгоритма. Чтобы получаемое замыкание было осмысленным, нужно подобрать такой множитель $\lambda(h)$, чтобы предел

$$\lim_{h \rightarrow 0} \lambda(h) \left(y(x) - \left(1 - \frac{1}{\frac{x}{h} + 1}\right) y(x+h) \right) \quad (6)$$

был конечен и отличен от нуля. Возьмем $\lambda(h) = h^{-1}$; тогда

$$\begin{aligned} \frac{1}{h} \left(y(x) - \left(1 - \frac{1}{\frac{x}{h} + 1}\right) y(x+h) \right) &= \\ &= \frac{1}{h} \left(y(x) - \left(1 - \frac{h}{x} + O(h^2)\right) (y(x) + y'(x)h + O(h^2)) \right) = \\ &= \frac{y(x)}{x} - y'(x) + O(h). \end{aligned}$$

При такой нормировке левая часть (6) имеет ненулевой предел. Попробуем применить такую же нормировку и в общем случае. Поскольку в рассмотренном примере существовал предел

$$\lim_{h \rightarrow 0} \frac{C_n - 1}{h} \quad \text{при} \quad nh = x,$$

введем вместо C_n новую переменную $\alpha_n = \frac{1-C_n}{h}$. Напомним, что прогоночные соотношения имеют вид

$$C_{n+1} = \frac{1}{2 + p_{n+1}h^2 - C_n}, \quad \Phi_{n+1} = C_{n+1}(\Phi_n - f_{n+1}h^2). \quad (7)$$

Умножим соотношение (5) на $-h^{-1}$ и положим $C_n = 1 - \alpha_n h$, а $\Phi_n = -\beta_n h$. Тогда (5) запишется в виде

$$\frac{y_{n+1} - y_n}{h} - \alpha_n y_{n+1} = \beta_n. \quad (8)$$

Подставив выражения C_n и Φ_n через α_n и β_n в (7), получим равенства

$$1 - \alpha_{n+1}h = \frac{1}{1 + p_{n+1}h^2 + \alpha_n h},$$

$$\beta_{n+1}h = (1 - \alpha_{n+1}h)(\beta_n h + f_{n+1}h^2),$$

отсюда

$$\alpha_{n+1} = \frac{\alpha_n + p_{n+1}h}{1 + \alpha_n h + p_{n+1}h^2}, \quad (9)$$

$$\beta_{n+1} = \beta_n + (f_{n+1} - \alpha_{n+1}\beta_n)h - \alpha_{n+1}f_{n+1}h^2. \quad (10)$$

Соотношение (9) можно преобразовать к виду

$$\frac{\alpha_{n+1} - \alpha_n}{h} = p_{n+1} - \alpha_n^2 - h \frac{2p_{n+1}\alpha_n - \alpha_n^3 + (p_{n+1}^2 - p_{n+1}\alpha_n^2)h}{1 + \alpha_n h + p_{n+1}h^2}, \quad (11)$$

а соотношение (10) к виду

$$\frac{\beta_{n+1} - \beta_n}{h} = f_{n+1} - \alpha_{n+1}\beta_n - f_{n+1}\alpha_{n+1}h. \quad (12)$$

Если предположить, что коэффициенты α_n и β_n при фиксированном $nh = x$ и $h \rightarrow 0$ стремятся к некоторым пределам $\alpha(x)$ и $\beta(x)$, то при подстановке в (8) вместо y_n значений $y(nh)$ в пределе получается дифференциальное уравнение

$$y' - \alpha(x)y = \beta(x),$$

являющееся замыканием вычислительного алгоритма. Соотношения (11), (12) показывают, что величины α_n , β_n удовлетворяют некоторым уравнениям, напоминающим метод Эйлера численного интегрирования дифференциальных уравнений

$$\alpha' = p(x) - \alpha^2, \quad (13)$$

$$\beta' = f(x) - \alpha\beta. \quad (14)$$

Доказательство того факта, что значения α_n , β_n при $nh = x$ фиксированном, $h \rightarrow 0$, стремятся к значениям решений этих дифференциальных уравнений, затрудняется следующим об-

стоятельством. Согласно определению,

$$\alpha_0 = \frac{1 - C_0}{h} = \frac{1}{h}, \quad \beta_0 = -\frac{\varphi_0}{h} = -\frac{a}{h}. \quad (15)$$

Таким образом, в окрестности точки $x = 0$ начальные условия для численного интегрирования по формулам (11), (12) имеют не совсем обычный характер и развитые нами методы оценки близости решений сеточных и дифференциальных задач в рассматриваемом случае неприменимы. Глядя на соотношения (15), можно было бы предположить, что функция $\alpha(x)$ является решением (13), ведущим себя вблизи нуля, как x^{-1} , а $\beta(x)$ — решением (14), ведущим себя вблизи нуля, как $-ax^{-1}$. Произведем обоснование сделанных высказываний.

Коэффициенты C_n , a , следовательно, и α_n получаются из рекуррентных соотношений (7), не зависящих ни от правой части дифференциального уравнения, ни от граничного условия в точке X ; начальное условие рекурсии $C_0 = 0$ также от них не зависит. Рассмотрим краевую задачу

$$y'' - p(x)y = 0, \quad y(0) = 0, \quad y(X) = 1$$

и соответствующую сеточную задачу

$$l(y_j) = 0, \quad y_0 = 0, \quad y_N = 1.$$

Решения этих задач можно записать в виде

$$y(x) = \frac{W^1(x)}{W^1(X)}, \quad y_n = \frac{W_n^1}{W_N^1}$$

(см. § 2). Поскольку для этой краевой задачи $\varphi_0 = 0$ и все $f_n = 0$, то $\beta_n \equiv 0$ и соотношение (8) имеет вид

$$\frac{y_{n+1} - y_n}{h} - \alpha_n y_{n+1} = 0.$$

Подставляя сюда $y_n = W_n^1 / W_N^1$, получим

$$\alpha_n = \left(\frac{W_{n+1}^1 - W_n^1}{h} \right) (W_{n+1}^1)^{-1}. \quad (16)$$

В § 2 получена оценка

$$|W_n^1 - W^1(nh)| \leq Mh^2 \quad \text{при} \quad 0 \leq nh \leq X, \quad (17)$$

величина $(W^1)''_{xx}$ равномерно ограничена на $[0, X]$; поэтому

$$\begin{aligned} \frac{W_{n+1}^1 - W_n^1}{h} &= \frac{W^1((n+1)h) - W^1(nh) + O(h^2)}{h} = \\ &= (W^1)'((n+1)h) + O(h) \end{aligned} \quad (18)$$

равномерно на $[0, X]$. После подстановки этих выражений в (16) получим равенство

$$\alpha_n = \frac{(W^1)'((n+1)h) + O(h)}{W^1((n+1)h) + O(h)}.$$

Таким образом, при $h \rightarrow 0$ значения α_n сближаются со значениями функции

$$\alpha(x) = \frac{(W^1)'(x)}{W^1(x)}.$$

Функция $\alpha(x)$ удовлетворяет дифференциальному уравнению (13), так как

$$\left(\frac{(W^1)'}{W^1}\right)' = \frac{(W^1)''}{W^1} - \left(\frac{(W^1)'}{W^1}\right)^2 = p(x) - \alpha^2.$$

Согласно определению $W^1(x)$, при малых x имеем $W^1(x) \sim x$, $(W^1)'(x) \sim 1$ и, таким образом, $\alpha(x) = \frac{(W^1)'(x)}{W^1(x)} \sim x^{-1}$, как и предполагалось. Отметим следующее обстоятельство. Функция $\alpha(x)$ обращается в бесконечность в точках, где $W^1(x) = 0$, в частности, при $x = 0$. В окрестности каждой такой точки y имеем

$$(W^1)'(x) \sim (W^1)'(y) \neq 0, \quad W^1(x) \sim ((W^1)'(y))(x - y)$$

и, таким образом,

$$\alpha(x) \sim (x - y)^{-1}. \quad (19)$$

Для точек этой окрестности можно написать цепочки соотношений

$$\frac{W_{n+1}^1 - W_n^1}{h} = (W^1)'((n+1)h)(1 + O(h)),$$

$$\begin{aligned} W_{n+1}^1 &= W^1((n+1)h) + O(h^2) = W^1((n+1)h) \left(1 + O\left(\frac{h^2}{W^1((n+1)h)}\right)\right) = \\ &= W^1((n+1)h) \left(1 + O\left(\frac{h^2}{(n+1)h - y}\right)\right) \end{aligned}$$

и, таким образом,

$$\alpha_n = \alpha((n+1)h)(1 + O(h)) \left(1 + O\left(\frac{h^2}{(n+1)h - y}\right)\right).$$

Относительная погрешность равенства $\alpha_n \approx \alpha((n+1)h)$ является величиной порядка $O(h)$ вплоть до окрестности порядка h точек, где $\alpha(x) = \infty$.

В случае неоднородного уравнения, согласно (8), справедливо равенство

$$\beta_n = \frac{y_{n+1} - y_n}{h} - \alpha_n y_{n+1}.$$

Вследствие (18) и оценки сходимости решения сеточной задачи к решению дифференциальной, полученной в § 2, имеем

$$\beta_n = \beta((n+1)h) + O(h)$$

во всех точках $x = (n+1)h$, где $\alpha(x) \neq 0$; здесь

$$\beta(x) = y'(x) - \alpha(x)y(x).$$

Таким образом, сеточная функция β_n действительно сближается с некоторой функцией непрерывного аргумента $\beta(x)$. Очевидно,

$$\beta' = y'' - \alpha'y - \alpha y' = (py + f) - (p - \alpha^2)y - \alpha y' = f - \alpha(y' - \alpha y) = f - \alpha\beta;$$

в окрестности точки 0, вследствие (19), имеем

$$\beta(x) \sim -y(0)x^{-1} \sim -\alpha x^{-1}.$$

Выпишем теперь соотношения (3) и соответствующие замыкания алгоритма. Возьмем $M = 2N - 1$ и при $m = 1, \dots, N$ за (3) примем совокупность соотношений

$$\begin{aligned} \frac{y_j - C_j y_{j+1}}{h} &= \frac{\varphi_j}{h}, & 0 \leq j < m, \\ \frac{y_{j-1} - 2y_j + y_{j+1}}{h^2} - p_j y_j &= f_j, & m \leq j < N, \\ y_N &= b, \end{aligned}$$

а при $N < m \leq M$ — соотношений

$$\begin{aligned} \frac{y_j - C_j y_{j+1}}{h} &= \frac{\varphi_j}{h}, & 0 \leq j \leq M - m, \\ y_j &= \psi_j, & M - m < j \leq N, \end{aligned}$$

где ψ_j — точное решение сеточной задачи. Положим $z = mh$. Тогда замыкание алгоритма (4) будет выглядеть следующим образом:

$$L_z y = f_z(x), \quad \text{где для } 0 \leq z < 1$$

$$\begin{aligned} L_z y &= \begin{cases} y' - \alpha(x)y & \text{при } 0 < x < z, \\ y'' - p(x)y & \text{при } z \leq x < 1, \\ y & \text{при } x = 1, \end{cases} \\ f_z(x) &= \begin{cases} \beta(x) & \text{при } 0 < x < z, \\ f(x) & \text{при } z \leq x < 1, \\ b & \text{при } x = 1 \end{cases} \end{aligned}$$

и для $1 \leq z \leq 2$

$$\begin{aligned} L_z y &= \begin{cases} y' - \alpha(x)y & \text{при } 0 < x < 2 - z, \\ y & \text{при } 2 - z \leq x \leq 1, \end{cases} \\ f_z(x) &= \begin{cases} \beta(x) & \text{при } 0 < x < 2 - z, \\ \psi(x) & \text{при } 2 - z \leq x \leq 1; \end{cases} \end{aligned}$$

здесь $\psi(x)$ — точное решение дифференциальной задачи (равенство $L_z y|_{x=0} = a$ опущено).

Рассмотрим примеры поведения функций $\alpha(x)$ и $\beta(x)$.

Пусть $p(x) > 0$; тогда, согласно уравнению $\alpha' = p(x) - \alpha^2$,

$$\alpha' < 0 \quad \text{при} \quad \alpha > D_1 = \sqrt{\max_{[0, X]} p(x)}$$

и

$$\alpha' > 0 \quad \text{при} \quad \alpha < D_2 = \sqrt{\min_{[0, X]} p(x)}.$$

Соответствующее поле интегральных кривых изображено на рис. 9.4.1. Поэтому решение $\alpha(x)$, удовлетворяющее условию $\alpha(x) \sim x^{-1}$ при $x \rightarrow 0$, монотонно убывает по крайней мере до тех пор, пока не попадет в область $D_2 < \alpha < D_1$, где оно и

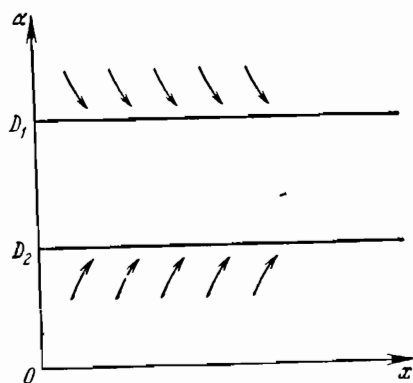


Рис. 9.4.1.

остается. Уравнение (14) линейно относительно $\beta(x)$. Так как коэффициент $\alpha(x)$ конечен при $x \neq 0$, то решение этого уравнения остается ограниченным при всех $x \neq 0$. Таким образом, были бы все основания признать замыкание алгоритма решения рассматриваемой задачи регулярным, если бы не неограниченность коэффициентов $\alpha(x)$ и $\beta(x)$ при $x \rightarrow 0$. В рассматриваемом случае значения α_0 и β_0 в точке $n = 0$ являются величинами порядка h^{-1} , далее с ростом n они

должны иметь тенденцию к убыванию вследствие аналогичной тенденции решений дифференциальных уравнений. Мы уже свыклись с тем, что во многих случаях влияние вычислительной погрешности порядка $2^{-t}h^{-p}$ является неизбежным и допустимым и возможно, что такая нерегулярность алгоритма не приведет к недопустимо большому влиянию вычислительной погрешности.

Обратимся к случаю, когда не всюду $p(x) \geq 0$. Тогда может оказаться, что в некоторой внутренней точке отрезка $W^1(y) = 0$ и, следовательно $\alpha(y) = \infty$. В этом случае замыкание алгоритма вычислений следует признать нерегулярным; однако и здесь не следует спешить окончательно отказываться от применения метода прогонки. Если оказалось, что во всех узлах сетки $W^1(nh)h^{-2} \gg 1$, то, согласно (16)–(18), имеем

$$\max_n |\alpha_n| = O(h^{-2}). \quad (20)$$

Мы уже говорили, что такого рода нерегулярность алгоритма не всегда следует считать катастрофической. Функция $W^1(x)$ гладкая с производной, отличной от нуля в точках y , где $W^1(y) = 0$. Поэтому часто значение $\min_{n > 0} W^1(nh)$ будет оказываться величиной порядка h и соотношение (20) будет выполняться. Таким образом, следует ожидать, что и в случае $p(x) \leq 0$ вычисления по методу прогонки, как правило, будут устойчивыми к различного рода возмущениям.

Оказалось, что замыкание алгоритма прогонки регулярно, если $W^1(x) \neq 0$ при $x \in (0, X]$. Условие $W^1(x_0) \neq 0$ равносильно условию, что краевая задача

$$y(0) = a, \quad y(x_0) = b, \quad y'' - p(x)y = f(x) \quad (21)$$

разрешима на $[0, x_0]$ при любых a, b и $f(x)$. Таким образом, условие регулярности замыкания алгоритма прогонки можно сформулировать в следующем виде:

Замыкание метода прогонки регулярно, если для любого отрезка $[0, x_0]$ при $0 < x_0 \leq X$ краевая задача (21) разрешима.

Получим этот вывод еще одним путем. Обратимся к формулам обратного хода прогонки. Коэффициенты C_j и Φ_j при $j < n$ зависят только от a и значений p_j и f_j при $j < n$, т. е. в точках $jh \in [0, nh]$. Следовательно, после отыскания значения y_n вычисления происходят в точности так же, как в случае краевой задачи для уравнения $y'' - p(x)y = f(x)$ на $[0, nh]$ при заданных $y(0) = a$ и $y(nh) = y_n$. Если эта краевая задача неразрешима при любых правых частях $f(x)$, то мало надежды, что соответствующая сеточная задача будет обладать хорошими свойствами при ее численном решении. Поэтому в случае, когда при каком-то $x_0 \in (0, X]$ краевая задача (21) разрешима не при любых правых частях, есть подозрения, что алгоритм вычислений будет сильно чувствителен к погрешности от округлений и требуется его более детальное исследование.

Отметим, что регулярность замыкания еще не обеспечивает малости суммарной вычислительной погрешности в силу следующих обстоятельств. Суммарная погрешность определяется погрешностями в ходе вычислений и множителями пропорциональности, с которыми эти погрешности входят в суммарную погрешность. Регулярность замыкания обеспечивает лишь малость погрешностей округления в ходе вычислений. Чтобы множители пропорциональности не были большими, нужна слабая чувствительность решения уравнений замыкания к возмущениям коэффициентов. Однако и этого, вообще говоря, недостаточно.

На примере уравнения $y' = My$ при $M < 0$ мы видели, что зачастую решение дифференциальной задачи слабо чувствительно к таким возмущениям, в то время как решение сеточной задачи сильно чувствительно.

§ 5. Обсуждение постановок краевых задач для линейных систем первого порядка

Рассмотрим краевую задачу

$$y' - A(x)y = f(x), \quad By(0) = b, \quad Dy(X) = d; \quad (1)$$

здесь y, f, b, d — векторы размерности, соответственно, $l, l, l - r, r$, а A, B, D — матрицы размерности $l \times l, (l - r) \times l, r \times l$. Всюду в дальнейшем предполагается, что ранг матрицы B равен $l - r$, а ранг матрицы D равен r .

Прежде чем отыскивать практически пригодные методы решения задачи (1), обсудим вопрос о чувствительности решений краевых задач к разного рода возмущениям. В качестве модели возьмем краевую задачу

$$y' - Ay = 0, \quad A = \text{const}, \quad By(0) = b, \quad Dy(X) = d. \quad (2)$$

Рассмотрим случай, когда все собственные значения матрицы A различны; в общем случае ход рассуждений изменяется не существенно. Пусть $\lambda_j = \alpha_j + i\beta_j$ — собственные значения матрицы A , упорядоченные в порядке возрастания α_j , а e_j — соответствующие собственные векторы, причем $\|e_j\| = 1$. Общее решение системы $y' - Ay = 0$ запишется в виде

$$y(x) = \sum_{j=1}^l c_j e_j \exp(\lambda_j x). \quad (3)$$

Собственные значения λ_j разделим на три группы, присваивая собственным значениям каждой группы соответствующий верхний индекс. Собственные значения λ_j , для которых величина $\exp(|\alpha_j|X)$ очень большая, обозначаем, как λ_j^+ , если $\alpha_j > 0$, и как λ_j^- , если $\alpha_j < 0$; остальные собственные значения, т. е. те, для которых величина $\exp(|\alpha_j|X)$ не очень велика, обозначаем, как λ_j^0 . Соответственно снабдим верхними индексами собственные векторы e_j ; суммирование по индексам j , соответствующим этим группам, будем обозначать $\sum^+$, $\sum^0$, $\sum^-$. Пусть l^+ , l^0 , l^- — число собственных значений в соответствующих группах. Форма записи решения (3) ставит в неординарное положение концы отрезка интегрирования: все функции $\exp(\lambda_j x)$ в точке $x = 0$ имеют порядок 1, в то время как в точке $x = X$ одни из них очень большие, другие очень малые. Удобнее другая форма записи общего решения:

$$y(x) = \sum^- c_j^- e_j^- \exp(\lambda_j^- x) + \sum^0 c_j^0 e_j^0 \exp(\lambda_j^0 x) + \\ + \sum^+ c_j^+ e_j^+ \exp(\lambda_j^+ (x - X)).$$

Выпишем систему уравнений $By(0) = b$, $Dy(X) = d$, из которой следует определить постоянные c_j , соответствующие отыскиваемому решению:

$$By(0) = \sum^- c_j^- B e_j^- + \sum^0 c_j^0 B e_j^0 + \sum^+ c_j^+ B e_j^+ \exp(-\lambda_j^+ X) = b, \\ Dy(X) = \sum^- c_j^- B e_j^- \exp(\lambda_j^- X) + \sum^0 c_j^0 B e_j^0 \exp(\lambda_j^0 X) + \\ + \sum^+ c_j^+ B e_j^+ = d. \quad (4)$$

Систему уравнений (4) можно записать в форме

$$Gc = g;$$

здесь

$$c = (c_1^-, \dots, c_l^+)^T, \quad g = (b_1, \dots, b_{l-r}, d_1, \dots, d_r)^T,$$

матрица G имеет клеточный вид

$$G = \begin{pmatrix} G_1^- & G_1^0 & G_1^+ \\ G_2^- & G_2^0 & G_2^+ \\ l^- & l^0 & l^+ \end{pmatrix} \begin{matrix} l-r \\ r \\ \end{matrix},$$

где

$$\begin{aligned} G_1^- &= [(Be_j^-)_q], \quad 1 \leq j \leq l^-, \quad 1 \leq q \leq l-r, \\ G_1^0 &= [(Be_j^0)_q], \quad l^- < j \leq l^- + l^0, \quad 1 \leq q \leq l-r, \\ G_1^+ &= [(Be_j^+)_q \exp(-\lambda_j^+ X)], \quad l^- + l^0 < j \leq l, \quad 1 \leq q \leq l-r, \\ G_2^- &= [(De_j^-)_q \exp(\lambda_j^- X)], \quad 1 \leq j \leq l^-, \quad 1 \leq q \leq r, \\ G_2^0 &= [(De_j^0)_q \exp(\lambda_j^0 X)], \quad l^- < j \leq l^- + l^0, \quad 1 \leq q \leq r, \\ G_2^+ &= [(De_j^+)_q], \quad l^- + l^0 < j \leq l, \quad 1 \leq q \leq r. \end{aligned}$$

Предположим, что $\Delta = \det G \neq 0$, и представим решение системы (4) в виде $c = G^{-1}g$; согласно известным формулам для элементов обратной матрицы:

$$G^{-1} = \left[\frac{(-1)^{i+j} G_{ji}}{\Delta} \right]; \quad (5)$$

G_{ji} — миноры матрицы G . При сформулированных ранее предположениях величины $\exp(-\lambda_j^+ X)$, $\exp(\lambda_j^- X)$ ничтожно малы; поэтому ничтожно малы элементы матриц G_1^+ и G_2^- и определитель

$$\Delta = \det \begin{pmatrix} G_1^- & G_1^0 & G_1^+ \\ G_2^- & G_2^0 & G_2^+ \end{pmatrix}$$

близок к определителю

$$\Delta_0 = \det \begin{pmatrix} G_1^- & G_1^0 & 0 \\ 0 & G_2^0 & G_2^+ \end{pmatrix}.$$

Рассмотрим отдельно случаи:

- а) определитель Δ_0 не мал,
- б) определитель Δ_0 мал, в частности равен нулю.

Поскольку среди элементов матрицы G нет больших, то, вследствие формулы (5), в случае а) элементы матрицы G^{-1} обычно не очень большие.

Предположим, что правые части граничных условий $\mathbf{b}$ и $\mathbf{d}$ содержат некоторые ошибки $\delta\mathbf{b}$ и $\delta\mathbf{d}$. Пусть $\delta\mathbf{g} = (\delta\mathbf{b}, \delta\mathbf{d})^T$, тогда ошибка вектора $\mathbf{c}$ равна $G^{-1}\delta\mathbf{g}$. Если элементы матрицы G^{-1} не очень велики, то влияние ошибки $\delta\mathbf{g}$ на коэффициенты c_j будет не очень большим. Решение задачи является линейной комбинацией с коэффициентами c_j слагаемых

$$e_j^- \exp(\lambda_j^- x), \quad e_j^0 \exp(\lambda_j^0 x), \quad e_j^+ \exp(\lambda_j^+ (x - X)).$$

Поэтому ошибка приближенного решения, являющаяся следствием ошибок $\delta\mathbf{b}$ и $\delta\mathbf{d}$, также будет приемлемой.

Заметим, что наши высказывания носят довольно неопределенный характер: «малый», «очень малый», «небольшой», «очень большой»; при анализе своей конкретной задачи исследователь должен сам решать, насколько приемлем для него тот или иной порядок рассматриваемых величин. В частности, если решается система «большого» порядка, то при «умеренных» значениях коэффициентов системы и «не очень малом» определителе Δ_0 возможно, что миноры, состоящие из сумм произведений большого числа элементов, окажутся «недопустимо большими».

Если определитель Δ_0 очень мал или равен нулю, то, вследствие равенства

$$\det G \det G^{-1} = 1,$$

среди элементов матрицы G^{-1} встретятся большие. Тогда малые возмущения правых частей граничных условий могут приводить к большим возмущениям коэффициентов c_j , а следовательно, и решения задачи.

Зная элементы матрицы G^{-1} и собственные значения матрицы A , из полученных выше соотношений можно получить довольно точную информацию о возмущении решения дифференциальной задачи. Однако получение этой информации само по себе требует большого объема вычислений; перенос этих построений на случай переменной матрицы $A(x)$ потребует еще большего объема вычислений. Попытаемся поэтому получить критерии устойчивости решения к возмущениям $\delta\mathbf{b}$ и $\delta\mathbf{d}$, требующие меньшей информации о задаче, хотя, может быть, и несколько менее надежные. Таким простым критерием могут служить соотношения между числами l^- , l^0 , l^+ , $l-r$ и r . Среди элементов первых $l-r$ строк матрицы G^0 ненулевые элементы могут находиться в первых $l^- + l^0$ столбцах, соответствующих матрицам G_1^- , G_1^0 . Если $l^+ > r$, то $l^- + l^0 < l-r$, и тогда все миноры порядка $(l-r) \times (l-r)$, лежащие в первых $l-r$ строках, обращаются в нуль. Раскрывая определитель Δ_0 по первым $l-r$ строкам, получаем $\Delta_0 = 0$. Точно так же, если $l^- > l-r$, то все миноры порядка $r \times r$, лежащие в последних r строках, обращаются в нуль, и поэтому $\Delta_0 = 0$. Если $l^+ \leq r$, а $l^- \leq l-r$, то определитель Δ_0 окажется линейной комбинацией произведе-

ний элементов матриц B и D и координат собственных векторов e_j , причем коэффициентами при этих произведениях будут произведения чисел $\exp(\lambda_j^0 X)$, не очень больших и не очень маленьких по модулю. Можно принять гипотезу, что этот определитель оказывается малым числом довольно редко. Тогда решение задачи (2) будет малочувствительно к возмущениям δb и δd правых частей граничных условий.

Мы можем сформулировать полученные выводы в качестве следующего утверждения. Если $l^+ > r$ или $l^- > l - r$, то решение дифференциальной задачи сильно чувствительно к возмущениям правых частей граничных условий; если $l^+ \leq r$ и $l^- \leq l - r$, то, как правило, решение задачи (2) будет малочувствительно к изменениям правых частей граничных условий.

Первую часть этого утверждения можно переформулировать еще в такой форме: для малой чувствительности задачи к возмущениям граничных условий необходимо, чтобы число независимых частных решений $e_j \exp(\lambda_j x)$, сильно растущих на $[0, X]$ с ростом x , не превосходило числа граничных условий на правом конце, а число частных решений $e_j \exp(\lambda_j x)$, сильно убывающих на $[0, X]$ с ростом x , не превосходило числа граничных условий на левом конце.

Эта формулировка при определенных уточнениях может быть перенесена и на случай задачи (1) с переменной матрицей $A(x)$. Строгая переформулировка этого утверждения будет довольно громоздкой; однако если матрица $A(x)$ относительно плавно меняется на $[0, X]$, то при первоначальном исследовании устойчивости задачи к возмущениям граничных условий зачастую можно ограничиться подсчетом числа собственных значений матриц $A(x)$ с большим положительным и большим отрицательным значением величины $X \operatorname{Re} \lambda_j(x)$.

Краевую задачу называют хорошо обусловленной (хорошо поставленной), если малые возмущения коэффициентов и правых частей уравнения и граничных условий приводят к столь же малым по порядку изменениям решения задачи. Более аккуратное определение хорошей обусловленности можно дать следующим образом. Наряду с краевой задачей (1) рассмотрим краевые задачи

$$\tilde{y}' - (A(x) + \delta A(x)) \tilde{y} = f(x) + \delta f(x),$$

$$(B + \delta B) \tilde{y}(0) = b + \delta b, \quad (D + \delta D) \tilde{y}(X) = d + \delta d$$

с не очень большой мерой возмущения

$$\varepsilon = \max_{0 \leq x \leq X} (\|\delta A(x)\| + \|\delta f(x)\|) + \|\delta B\| + \|\delta D\| + \|\delta b\| + \|\delta d\|.$$

Если для всех решений таких краевых задач выполняются не равенства

$$\max_{0 \leq x \leq X} \|\tilde{y}(x) - y(x)\| \leq M\varepsilon \quad (6)$$

с не очень большим значением постоянной M , то исходную задачу называют *хорошо обусловленной*, в противном случае задачу называют *плохо обусловленной*; минимальное значение $M(\varepsilon_0)$, при котором неравенство (6) выполняется при всех $\varepsilon \leq \varepsilon_0$ ($\varepsilon_0 > 0$ фиксировано), иногда называют *мерой обусловленности* данной задачи (относительно возмущений с нормой, не большей ε_0). Обусловленность задачи характеризует устойчивость решения к возмущениям исходных данных, например, к неточности задания параметров, входящих в коэффициенты уравнения. Поскольку ошибки от округления при вычислениях эквивалентны возмущениям коэффициентов исходного уравнения, то мера обусловленности характеризует и устойчивость численного решения к возможным округлениям при численном решении. Если известна ориентировочная оценка ε возмущения коэффициентов задачи и погрешность порядка $M(\varepsilon)\varepsilon$ допустима, то имеет смысл непосредственное численное решение задачи.

Рассмотрим в качестве примера задачу Коши для системы $y_1' = y_2$, $y_2' = y_1$ при $y_1(0) = 1$, $y_2(0) = 1$ на отрезке $[0, 30]$. Собственные значения матрицы $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ равны $\pm 1 \cdot \exp(-30)$ — очень малая, а $\exp(30)$ — очень большая величины, $l^- = 1$, $l^0 = 0$, $l^+ = 1$, $l^- - r = 2$, $r = 0$ и, следовательно, $l^- = 1 > r = 0$. Малые возмущения граничного условия должны приводить к очень большим изменениям решения. В данном случае проводившиеся нами в этом параграфе построения не имеют особого смысла, поскольку и без них ясно, что, вследствие сильного роста решений исходного линейного уравнения, погрешность решения растет очень быстро.

Пусть для той же системы рассматривается задача с краевыми условиями $b_1 y_1(0) + b_2 y_2(0) = b$, $d_1 y_1(30) + d_2 y_2(30) = d$; собственные векторы, соответствующие собственным значениям $\lambda_1^- = -1$ и $\lambda_2^+ = 1$, равны, соответственно, $(1, -1)^T$ и $(1, 1)^T$, $l^+ = r$, $l^- = l - r$. Вообще говоря, следует ожидать, что задача будет устойчива к возмущениям δb и δd . Решение отыскивается в виде

$$c_1^- \begin{pmatrix} 1 \\ -1 \end{pmatrix} \exp(-x) + c_2^+ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \exp(x - 30).$$

Уравнения (4) имеют вид

$$\begin{aligned}(b_1 - b_2) c_1^- + (b_1 + b_2) c_2^+ \exp(-30) &= b, \\ (d_1 - d_2) c_1^- \exp(-30) + (d_1 + d_2) c_2^+ &= d.\end{aligned}\quad (7)$$

Отсюда

$$\begin{aligned}c_1^- &= \frac{b(d_1 + d_2) - d(b_1 + b_2) \exp(-30)}{\Delta}, \\ c_2^+ &= \frac{d(b_1 - b_2) - b(d_1 - d_2) \exp(-30)}{\Delta},\end{aligned}$$

где

$$\Delta = (b_1 - b_2)(d_1 + d_2) - (d_1 - d_2)(b_1 + b_2) \exp(-60)$$

близко к $\Delta_0 = (b_1 - b_2)(d_1 + d_2)$. Если коэффициенты b_1 и d_1 небольшие, а Δ_0 не мало, то коэффициенты c_1^- и c_2^+ мало изменяются при малых изменениях δb и δd . Если $\Delta = 0$, то система (7) имеет ненулевое решение c_1^- , c_2^+ при $b = d = 0$, как говорят, «задача лежит на спектре». Это означает, что однородная задача имеет ненулевое решение и бессмысленно говорить об устойчивости решения к возмущениям граничных условий. При $\Delta_0 = 0$ задача плохо обусловлена.

В дальнейшем, в §§ 6—8, предполагается, что рассматриваемая краевая задача хорошо обусловлена.

З а д а ч а. Доказать, что решение хорошо обусловленной краевой задачи единственно.

§ 6. Алгоритмы решения краевых задач для систем уравнений первого порядка

Простейшим по форме методом решения краевой задачи (5.1) является *метод стрельбы*. Рассмотрим систему уравнений

$$By(0) = b. \quad (1)$$

Поскольку по предположению ранг матрицы B равен $l - r$, то общее решение системы (1) записывается в виде

$$y_0 + \sum_{i=1}^r c_i y_i;$$

здесь y_0 — произвольное решение неоднородной системы $By = b$, а $y_1, \dots, y_r$ — произвольная система r линейно независимых решений системы $By = 0$. Пусть $y_1, \dots, y_r, y_0$ — какой-то набор таких векторов. Численным интегрированием найдем частное решение неоднородной системы

$$y'_0 = A(x) y_0 + f(x) \quad (2)$$

при начальном условии $y_0(0) = y_0$ и решения однородной системы

$$y_j' = A(x) y_j, \quad j = 1, \dots, r, \quad (3)$$

при начальных условиях $y_j(0) = y_j$. Пусть $y(x)$ удовлетворяет (2) и левому граничному условию (1). Запишем вектор $y(0)$ в виде

$$y(0) = y_0 + \sum_{j=1}^r c_j y_j;$$

вектор-функция

$$z(x) = y_0(x) + \sum_{j=1}^r c_j y_j(x)$$

удовлетворяет уравнению (2) и совпадает с $y(x)$ при $x = 0$. Следовательно,

$$y(x) = y_0(x) + \sum_{j=1}^r c_j y_j(x). \quad (4)$$

Всякая функция вида (4) удовлетворяет соотношениям (1) и (2). Таким образом, многообразие всех решений (2), удовлетворяющих левому граничному условию (1), задается равенством (4). Чтобы найти искомое решение, надо выделить из этого многообразия решение, удовлетворяющее правому граничному условию. Определим коэффициенты c_j из системы r уравнений с r неизвестными

$$D \left(y_0(X) + \sum_{j=1}^r c_j y_j(X) \right) = d. \quad (5)$$

В предположении однозначной разрешимости задачи (5.1) определитель этой системы отличен от нуля; действительно, предположив противное, мы получили бы, что однородная краевая задача ($f \equiv 0$, $b = 0$, $d = 0$) имеет ненулевое решение

$$y(x) = \sum_{j=1}^r c_j y_j(x).$$

Если $c_1, \dots, c_r$ — решение системы (5), то вектор-функция

$$y(x) = y_0(x) + \sum_{j=1}^r c_j y_j(x)$$

удовлетворяет уравнению и всем граничным условиям и, следовательно, является решением искомой задачи.

При необходимости экономить память следует найти

$$y(0) = y_0(0) + \sum_{j=1}^r c_j y_j(0)$$

или

$$y(X) = y_0(X) + \sum_{j=1}^r c_j y_j(X)$$

и затем решить численно задачу Коши вперед или назад.

Если среди решений однородной системы $y' = A(x)y$ есть быстро растущие с ростом x , столь же быстро может возрасти вычислительная погрешность в решениях $y_0(x), \dots, y_r(x)$. В этом случае алгоритм метода стрельбы окажется непригодным для практического использования. По-другому практическую непригодность метода стрельбы в этом случае можно объяснить следующим образом. Пусть метод стрельбы применяется при $A = \text{const}$; обозначим через e_j и λ_j собственные векторы и соответствующие собственные значения матрицы A , причем пусть

$$\text{Re } \lambda_1 \leq \dots \leq \text{Re } \lambda_l;$$

предположим, что

$$y_l = \sum_{k=1}^l \alpha_{lk} e_k;$$

тогда

$$y_l(X) = \sum_{k=1}^l \alpha_{lk} e_k \exp(\lambda_k X). \quad (6)$$

Если $\alpha_{ll} \neq 0$ и $\exp(\text{Re } \lambda_l X) \gg \exp(\text{Re } \lambda_{l-1} X)$, то

$$y_l(X) \sim \alpha_{ll} e_l \exp(\lambda_l X).$$

Таким образом, все строки определителя системы (5) оказываются приблизительно пропорциональными вектору e_l , и решение этой системы $c_1, \dots, c_r$ будет найдено с большой вычислительной погрешностью. Эти рассуждения в ряде мест не совсем корректны, но тем не менее их можно принять за правдоподобное объяснение возможности большого влияния вычислительной погрешности в методе стрельбы.

Поставим перед собой пока частную задачу определения значения $y(X)$; ее можно трактовать как задачу выделения из многообразия

$$y_0(X) + \sum_{j=1}^r c_j y_j(X) \quad (7)$$

вектора, удовлетворяющего правому граничному условию $Dy(X) = d_0$. Мы видели, что непосредственное решение этой задачи может привести к большой погрешности. Векторы $y_j(X)$ окажутся близкими к пропорциональным, или, как иногда говорят, базис из векторов $y_j(x)$ при продвижении к точке X «сплющится».

Однако никто не обязывает в каждый момент задавать многообразие (7) именно при помощи вектора $y_0(x)$ и базиса $\{y_j(x)\}$. Время от времени по мере «сплющивания» базиса имеет смысл переходить к новому способу задания многообразия (7). На этом и основана идея первого метода ортогональной прогонки.

Разобьем отрезок интегрирования $[0, X]$ на части точками $0 = X_0 < X_1 < \dots < X_m = X$. Определим операцию M_s над матрицей $G = (g_1, \dots, g_r, g_0)$ размерности $l \times (r+1)$ как операцию над ее векторами-столбцами g_j следующим образом.

Пусть $g_j^s(x)$, $j = 1, \dots, r$, решение системы $y' = A(x)y$ при начальном условии $y(X_s) = g_j$, а $g_0^s(x)$ — системы $y' = A(x)y + f(x)$ при начальном условии $y(X_s) = g_0$; тогда $M_s g_j = g_j^s(X_{s+1})$.

Определим операцию U ортогонализации столбцов матрицы G . Столбцы матрицы $Q = U(G)$ обозначаем соответственно $q_1, \dots, q_r, q_0$. Векторы $q_1, \dots, q_r, q_0$ получаются из векторов $g_1, \dots, g_r, g_0$ последовательной ортогонализацией и нормировкой, причем вектор q_0 не нормируется. Таким образом, последовательно получаем (как и в § 15 гл. II) векторы

$$z_j = g_j - \sum_{i=1}^{j-1} (g_j, q_i) q_i, \quad (8)$$

$$q_j = \frac{z_j}{V(z_j, z_j)}, \quad j = 1, \dots, r,$$

$$q_0 = g_0 - \sum_{i=1}^r (g_0, q_i) q_i. \quad (9)$$

Выразим z_j в соотношениях (8), (9) через q_i и перенесем все q_i в левые части; получим

$$V(z_j, z_j) q_j + \sum_{i=1}^{j-1} (g_j, q_i) q_i = g_j, \quad j = 1, \dots, r,$$

$$\sum_{i=1}^r (g_0, q_i) q_i + q_0 = g_0.$$

Эту систему соотношений можно переписать в матричной форме $Q\Omega = G$, где

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{21} & \dots & \omega_{r1} & \omega_{r+1,1} \\ 0 & \omega_{22} & \dots & \omega_{r2} & \omega_{r+1,2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & \omega_{rr} & \omega_{r+1,r} \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix},$$

$$\omega_{ij} = V(z_j, z_j), \quad \omega_{ji} = (g_j, q_i) \quad \text{при} \quad i < j \leq r, \quad \omega_{r+1,i} = (g_0, q_i);$$

иначе $Q = U(G) = G \cdot \Omega^{-1}$. Выберем векторы $y_1, \dots, y_r, y_0$ так же, как в случае метода стрельбы, и осуществим над матрицей $Y_0 = (y_1, \dots, y_r, y_0)$ операцию ортогонализации. Пусть

$$G_0(X_0) = U(Y_0) = (g_1^0(X_0), \dots, g_r^0(X_0), g_0^0(X_0)).$$

Применив к $G_0(X_0)$ операцию M_0 , получим матрицу $G_0(X_1) = M_0(G_0(X_0))$; затем получим матрицу

$$G_1(X_1) = U(G_0(X_1)) = G_0(X_1) \Omega_1^{-1} \text{ и т. д.} \quad (10)$$

Через $G_s(x)$ будем обозначать матрицы $(g_1^s(x), \dots, g_r^s(x), g_0^s(x))$. В результате проводимых вычислений получаются матрицы

$$\begin{aligned} G_0(X_0), \quad G_0(X_1) &= M_0(G_0(X_0)), \\ G_1(X_1) &= U(G_0(X_1)) = G_0(X_1) \Omega_1^{-1}, \\ G_1(X_2) &= M_1(G_1(X_1)), \dots, \\ G_{m-1}(X_{m-1}) &= U(G_{m-2}(X_{m-1})) = G_{m-2}(X_{m-1}) \Omega_{m-1}^{-1}, \\ G_{m-1}(X_m) &= M_{m-1}(G_{m-1}(X_{m-1})), \\ G_m(X_m) &= U(G_{m-1}(X_m)) = G_{m-1}(X_m) \Omega_m^{-1}. \end{aligned} \quad (11)$$

Лемма. Многообразие, пробегаемое концами векторов

$$y = g_0^s(X_s) + \sum_{j=1}^r c_j g_j^s(X_s) = G_s(X_s) c \quad (12)$$

при всевозможных $c_1, \dots, c_r$, образует многообразие L_s всех значений $y(X_s)$ решений системы $y' = A(x)y + f(x)$, удовлетворяющих левому граничному условию $By(0) = b$; здесь

$$c = (c_1, \dots, c_r, 1)^T.$$

Доказательство. Действительно, при $s = 0$ это утверждение очевидно. Пусть оно доказано для некоторого s . Осуществляя операцию M_s , мы переходим от многообразия L_s к многообразию L_{s+1} значений таких решений в точке x_{s+1} :

$$y = G_s(X_{s+1}) c.$$

Поскольку

$$G_s(X_{s+1}) = G_{s+1}(X_{s+1}) \Omega_{s+1},$$

то элементы этого многообразия можно записать в виде

$$y = G_{s+1}(X_{s+1}) \Omega_{s+1} c.$$

Положим $\Omega_{s+1} c = \tilde{c}$; так как $\omega_{r+1}^{s+1} = 1$, то последняя компонента вектора $\tilde{c}$ равна 1, и, переобозначая $\tilde{c} = c$, получим такую же, как (12), запись элементов многообразия L_{s+1} :

$y = G_{s+1}(X_{s+1})c$. Таким образом, требуемое утверждение доказано.

Вектор $g_0^s(x_s)$ по построению ортогонален ко всем векторам $g_j^s(x_s)$, $j = 1, \dots, r$, образующим базис многообразия L_s . Следовательно, его норма меньше нормы любого вектора, принадлежащего L_s . Пусть на каждом из отрезков $[X_s, X_{s+1}]$ искомое решение задачи записывается в виде

$$y(x) = g_0^s(x) + \sum_{j=1}^r \beta_j^s g_j^s(x),$$

или иначе

$$y(x) = G_s(x) \beta^s, \quad (13)$$

где $\beta^s = (\beta_1^s, \dots, \beta_r^s, 1)$.

Полагая в (13) $s = q$, $x = X_{q+1}$, имеем

$$y(X_{q+1}) = G_q(X_{q+1}) \beta^q,$$

полагая $s = q + 1$, $x = X_{q+1}$, имеем

$$y(X_{q+1}) = G_{q+1}(X_{q+1}) \beta^{q+1}.$$

Кроме того

$$(G_{q+1}(X_{q+1}))^{-1} G_q(X_{q+1}) = \Omega_{q+1}.$$

Из последних трех равенств следует, что

$$\beta^{q+1} = \Omega_{q+1} \beta^q. \quad (14)$$

Значение β^m может быть найдено из правого граничного условия

$$Dy(X) = D(G_m(X_m) \beta^m) = d.$$

Далее, пользуясь равенством (14) при $q = m - 1, \dots, 0$, последовательно определяем $\beta^{m-1}, \dots, \beta^0$.

После этого, рассуждая формально, для нахождения значений $y(x)$ в точках отрезка $[X_s, X_{s+1}]$ можно было бы воспользоваться формулой (13). Однако такой способ требует большой загрузки памяти из-за необходимости запоминать значения матриц $G_s(x)$ в большом числе точек. С позиций экономии памяти выгоднее найти значения $y(X_s)$ и затем на каждом из отрезков $[X_s, X_{s+1}]$ решать задачи Коши.

Мы не будем проводить полное исследование влияния вычислительной погрешности для этого алгоритма и ограничимся разного рода соображениями, говорящими в пользу того, что этот алгоритм не должен быть сильно чувствителен к ошибкам округления, если отрезки разбиения $[X_s, X_{s+1}]$ не очень велики.

Поскольку норма вектора $g_0^s(X_s)$ меньше нормы любого вектора многообразия L_s , то, в частности,

$$\|g_0^s(X_s)\| \leq \|y(X_s)\| \leq Y = \max_{0 \leq x \leq X} \|y(x)\|,$$

где $y(x)$ — искомое решение задачи. Предположим теперь, что $X_{k+1} \rightarrow X_k$ при фиксированном X_k . Тогда матрица

$$G_k(X_{k+1}) = (g_1^k(X_{k+1}), \dots, g_r^k(X_{k+1}), g_0^k(X_{k+1}))$$

стремится к матрице

$$G_k(X_k) = (g_1^k(X_k), \dots, g_r^k(X_k), g_0^k(X_k)).$$

Для матрицы $G_k(X_k)$ преобразование $U(G_k(X_k))$, приводящее ее к виду, когда первые r столбцов ортонормированы, а последний ортогонален к первым, есть тождественное, поскольку эта матрица уже имеет такой вид. Поэтому $\Omega_{k+1} = E$, если $X_{k+1} = X_k$. Естественно, что $\Omega_{k+1} \rightarrow E$ при $X_{k+1} \rightarrow X_k$. Векторы в первых r столбцах матрицы $G_k(X_k)$ имеют норму, равную 1, а

$$\|g_0^k(X_k)\| \leq Y.$$

Из равномерной ограниченности этих векторов следует, что стремление $G_{k+1}(X_k)$ к $G_k(X_k)$, а следовательно, и Ω_{k+1} к E равномерно по множеству всевозможных точек

$$X_1, \dots, X_m \text{ при } \delta = \max_k |X_{k+1} - X_k| \rightarrow 0;$$

иначе $\|\Omega_{k+1} - E\| \leq \omega(\delta)$, где $\omega(\delta) \rightarrow 0$ при $\delta \rightarrow 0$. Поэтому, если величина δ не очень велика, то нормы матриц $\|\Omega_k\|$, а также нормы функций $\|g_j^k(x)\|$ на отрезках $[X_k, X_{k+1}]$ ограничены равномерно не очень большой постоянной.

Из полученных утверждений нельзя сделать никаких заключений о хороших свойствах рассматриваемого алгоритма. Однако уже неплохо, что все встречающиеся в вычислениях величины оказываются равномерно ограниченными, причем не очень большими числами. В противном случае округления при вычислениях были бы равносильны заведомо большим возмущениям коэффициентов системы, и были бы все основания сомневаться в разумности использования этого алгоритма.

§ 7. Методы дифференциальной ортогональной прогонки

Построим замыкание при $\delta \rightarrow 0$ описанного в предшествующем параграфе алгоритма. Поскольку операторы M_s и операторы ортогонализации по-разному действуют на векторы $g_j^s(X_s)$ при $j \neq 0$ и $j = 0$, то удобно разбить рассматриваемые матрицы на клетки, выделив клетки, соответствующие

решениям g_s^s неоднородного уравнения. Положим

$$G_s(X_k) = (Y_s(X_k), y_s(X_k)),$$

где

$$Y_s(X_k) = (g_1^s(X_k), \dots, g_r^s(X_k)), \quad y_s(X_k) = g_0^s(X_k),$$

и

$$\Omega_k = \begin{pmatrix} E + \delta_{k-1} R_k & \delta_{k-1} r_k \\ 0 & 1 \end{pmatrix},$$

где

$$\delta_k = X_{k+1} - X_k, \quad r_k = \left(\frac{\omega_{r+1,1}^k}{\delta_k}, \dots, \frac{\omega_{r+1,r}^k}{\delta_k} \right)^T,$$

матрица R_k — правая треугольная. Введение здесь новых функций и матриц аналогично подобным преобразованиям при построении замыкания алгоритма прогонки в § 4. Обозначим $\beta_s = (\beta_1^s, \dots, \beta_r^s)^T$.

Введем обозначения $Z_s = Y_s(X_s)$, $z_s = y_s(X_s)$.

Предположим, существуют гладкие матричные и векторные функции

$$Z(x), \quad z(x), \quad R(x), \quad r(x), \quad \beta(x)$$

такие, что

$$Z_s \Rightarrow Z(x), \quad z_s \Rightarrow z(x), \quad R_s \Rightarrow R(x), \quad r_s \Rightarrow r(x), \quad \beta_s \Rightarrow \beta(x);$$

$g_s \Rightarrow g(x)$ здесь означает, что

$$\max_{X_s} \left(\|g_s - g(X_s)\| + \left\| \frac{g_{s+1} - g_s}{\delta_s} - g'(X_s) \right\| \right) \rightarrow 0 \quad \text{при} \quad \delta \rightarrow 0.$$

Производя поэлементное умножение, можно записать равенство

$$G_s(X_{s+1}) = G_{s+1}(X_{s+1}) \Omega_{s+1},$$

или, что то же,

$$(Y_s(X_{s+1}), y_s(X_{s+1})) = (Y_{s+1}(X_{s+1}), y_{s+1}(X_{s+1})) \begin{pmatrix} E + \delta_s R_{s+1} & \delta_s r_{s+1} \\ 0 & 1 \end{pmatrix},$$

как совокупность равенств

$$Y_s(X_{s+1}) = Y_{s+1}(X_{s+1}) (E + \delta_s R_{s+1}), \quad (1)$$

$$y_s(X_{s+1}) = Y_{s+1}(X_{s+1}) \delta_s r_{s+1} + y_{s+1}(X_{s+1}). \quad (2)$$

Поскольку столбцы матриц $Y_s(x)$, а следовательно, и сами матрицы удовлетворяют системе $y' = A(x)y$, то

$$Y_s(X_{s+1}) = Y_s(X_s) + \delta_s A_s Y_s(X_s) + O(\delta_s^2).$$

Точно так же

$$y_s(X_{s+1}) = y_s(X_s) + \delta_s (A_s y_s(X_s) + f_s) + O(\delta_s^2);$$

здесь $A_s = A(X_s)$, $f_s = f(X_s)$. Вспоминая определения матрицы Z_s и вектора z_s , мы можем переписать теперь равенства (1), (2) в виде

$$Z_s + \delta_s A_s Z_s = Z_{s+1} (E + \delta_s R_{s+1}) + O(\delta_s^2),$$

$$z_s + \delta_s (A_s z_s + f_s) = Z_{s+1} \delta_s r_{s+1} + z_{s+1} + O(\delta_s^2),$$

или

$$\frac{Z_{s+1} - Z_s}{\delta_s} = A_s Z_s - Z_{s+1} R_{s+1} + O(\delta_s).$$

$$\frac{z_{s+1} - z_s}{\delta_s} = A_s z_s + f_s - Z_{s+1} r_{s+1} + O(\delta_s).$$

Переходя к пределу при $\delta \rightarrow 0$, получаем дифференциальные уравнения

$$Z' = AZ - ZR, \quad (3)$$

$$z' = Az + f - Zr. \quad (4)$$

Условие, что столбцы матрицы $Z_m = Y_m(X_m)$ нормированы и ортогональны, можно записать в виде равенства $Z_m^T Z_m = E$, которое в пределе дает $Z^T(x) Z(x) = E$. Отсюда следует, что

$$(Z^T)' Z + Z^T Z' = E' = 0;$$

выражение для $(Z^T)'$ получаем, транспонируя (3); в результате имеем

$$(Z^T A^T - R^T Z^T) Z + Z^T (AZ - ZR) = 0,$$

или

$$R + R^T = Z^T (A + A^T) Z. \quad (5)$$

По построению вектор $z_{s+1} = y_{s+1}(x_{s+1})$ ортогонален векторам $g_l^{s+1}(x_{s+1})$ — столбцам матрицы Z_{s+1} . Это условие записывается в виде $Z_{s+1}^T z_{s+1} = 0$ и в пределе переходит в равенство $Z^T z = 0$. Дифференцируя его, получим

$$(Z^T)' z + Z^T z' = 0,$$

или

$$(Z^T A^T - R^T Z^T) z + Z^T (Az + f - Zr) = 0.$$

С учетом равенств $Z^T z = 0$, $Z^T Z = E$ последнее соотношение переписывается в виде

$$Z^T (A + A^T) z + Z^T f - r = 0,$$

или

$$r = Z^T (A + A^T) z + Z^T f.$$

Подставим это выражение r в правую часть (4):

$$z' = Az + f - Z Z^T (Az + f + A^T z) = (E - Z Z^T) (Az + f) - Z Z^T A^T z. \quad (6)$$

Осталось получить предельное соотношение для (6.14); в клеточной форме это равенство записывается в виде

$$\begin{pmatrix} E + \delta_s R_{s+1} & \delta_s r_{s+1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta_s \\ 1 \end{pmatrix} = \begin{pmatrix} \beta_{s+1} \\ 1 \end{pmatrix},$$

что равносильно равенству

$$(E + \delta_s R_{s+1}) \beta_s + \delta_s r_{s+1} = \beta_{s+1},$$

или

$$\frac{\beta_{s+1} - \beta_s}{\delta_s} = R_{s+1} \beta_s + r_{s+1}.$$

Переходя к пределу при $\delta \rightarrow 0$, получаем

$$\beta' = R\beta + r. \quad (7)$$

Выпишем вместе все полученные соотношения:

$$\begin{aligned} Z' &= AZ - ZR, \\ z' &= (E - ZZ^T)(Az + f) - ZZ^T A^T z, \end{aligned} \quad (8)$$

верхняя треугольная матрица R однозначно определяется равенством

$$R + R^T = Z^T(A + A^T)Z; \quad (9)$$

$$\beta' = R\beta + r, \quad r = Z^T(A + A^T)z + Z^T f. \quad (10)$$

Само решение задачи вследствие (6.12) записывается в виде

$$y(x) = Z(x)\beta(x) + z(x). \quad (11)$$

Таким образом, построенный в § 6 алгоритм ортогональной прогонки замыкается в следующий алгоритм решения исходной задачи. При начальных условиях $Z(0) = Y_0(0)$, $z(0) = g_0^0(0)$ определяем функции $Z(x)$, $z(x)$ и $R(x)$ из соотношений (8), (9). Затем определяем $\beta(x)$ из условия

$$D(Z(x)\beta(x) + z(x)) = d, \quad (12)$$

далее интегрируем уравнение (10) в направлении убывания x и получаем значения $y(x)$ по формуле (11). Такой способ решения задачи естественно также называть *ортогональной прогонкой*. Поскольку при всех x мы имели $\|z_x\| \leq Y$, то и предельная функция $z(x)$ удовлетворяет неравенству $\|z(x)\| \leq Y$. Столбцы матрицы Z нормированы и ортогональны; поэтому из (9) можно получить оценку $\|R\| \leq \sqrt{2l} \|A\|$; итак, все коэффициенты в получившихся уравнениях ограничены не очень большими величинами. Следовательно, формальное замыкание алгоритма ортогональной прогонки является регулярным.

Способ дифференциальной прогонки, соответствующий соотношениям (8) — (10), не является единственно возможным. Построим еще один способ дифференциальной прогонки. Рассмотрим некоторое линейное подпространство L решений однородной системы $y' = A(x)y$, имеющее размерность r . Обозначим через $L(x)$ след значений элементов этого подпространства в точке x .

Лемма. Пусть векторы-столбцы $z_1(x_0), \dots, z_r(x_0)$ образуют базис подпространства $L(x_0)$, $Q(x)$ — произвольная матрица размерности $r \times r$, $Z(x)$, матрица размерности $l \times r$, — решение системы $Z' = AZ - ZQ$ при начальном условии $Z(x_0) = (z_1(x_0), \dots, z_r(x_0))$. Тогда столбцы $z_1(x), \dots, z_r(x)$ матрицы $Z(x)$ принадлежат подпространству $L(x)$.

Доказательство. Если z — решение однородной системы $z' = Az$, а v — решение сопряженной системы

$$v' = -A^T v, \quad (13)$$

то

$$\begin{aligned} (v, z)' &= (v', z) + (v, z') = \\ &= (-A^T v, z) + (v, Az) = -(v, Az) + (v, Az) = 0. \end{aligned} \quad (14)$$

В ортогональном дополнении $L^T(x_0)$ подпространства $L(x_0)$ выберем базис $v_1(x_0), \dots, v_{l-r}(x_0)$. Через $v_i(x)$ обозначаем решения сопряженной системы (13) при начальных условиях $v_i(x_0)$, а через $V(x)$ матрицу $(v_1(x), \dots, v_{l-r}(x))$. Если вектор $v_i(x_0)$ ортогонален всем векторам подпространства $L(x_0)$, то вследствие (14) вектор $v_i(x)$ ортогонален всем векторам подпространства $L(x)$. Отсюда следует, что столбцы матрицы $V(x)$

образуют базис $L^T(x)$. Из (13) следует, что $V' = -A^T V$, и поэтому $(V^T)' = -V^T A$. Подставив в равенство $(V^T Z)' = V^T Z' + (V^T)' Z$ значения $Z' = AZ - ZQ$ и $(V^T)' = -V^T A$, получим

$$(V^T Z)' = -(V^T Z) Q. \quad (15)$$

Матрица $U = V^T Z$ имеет размерность $(l-r) \times r$; очевидно, $u_{ij} = (v_i, z_j)$. Поскольку векторы $z_j(x_0)$ принадлежат $L(x_0)$, а векторы $v_i(x_0)$ — ортогональному дополнению $L^T(x_0)$, то $u_{ij}(x_0) = 0$ при всех i, j . Согласно (15) векторы $u_i(x)$ — строки матрицы U — удовлетворяют однородной системе линейных уравнений $u_i' = u_i Q$. При начальном условии $u_i(x_0) = 0$ эта система имеет единственное решение $u_i(x) \equiv 0$. Это означает, что $(v_i(x), z_j(x)) = 0$, при всех i, j ; векторы $z_j(x)$ ортогональны векторам $v_i(x)$, образующим базис $L^T(x)$, и, следовательно, принадлежат $L(x)$. Лемма доказана.

Задавая произвольной матрицей $Q(x)$ и решая систему (12), получим матрицу $Z(x)$, столбцы которой при каждом x принадлежат подпространству $L(x)$.

Обсуждая метод стрельбы, мы видели, что при $Q(x) \equiv 0$ и некоторых $A(x)$ способ задания подпространства $L(x)$ при помощи столбцов матрицы $Z(x)$ является не очень удачным. Поэтому нужно уделить специальное внимание выбору матрицы $Q(x)$. Один из возможных способов выбора матрицы $Q(x)$ возник у нас в результате построения замыкания алгоритма ортогональной прогонки. Построим другой такой способ.

Найдем матрицу $Q(x)$, при которой столбцы $Z(x)$ меняются по возможности медленно. Плавность изменения матрицы $Z(x)$, вообще говоря, должна давать возможность интегрирования системы (12) с более крупным шагом. С этой целью рассмотрим задачу минимизации норм $\|z_i'\| = \sqrt{(z_i', z_i')}$ за счет выбора элементов $q_{ij}(x)$ матрицы $Q(x)$. Столбец z_i удовлетворяет системе уравнений

$$z_i' = Az_i - \sum_{j=1}^r q_{ij} z_j.$$

Векторы $\sum_{j=1}^r q_{ij} z_j$ образуют линейное подпространство. Для минимальности $\|z_i'\|$ необходимо и достаточно, чтобы вектор z_i' был ортогонален этому подпространству, т. е. $(z_i', z_j) = 0$ при всех j . Совокупность условий $(z_i', z_j) = 0$ при всевозможных i, j можно переписать в матричной форме

$$Z^T Z' = 0. \quad (16)$$

Подставляя сюда значение Z' , получим

$$Z^T AZ - Z^T ZQ = 0.$$

Если матрица $Z^T Z$ обратима, то $Q = (Z^T Z)^{-1} Z^T AZ$; поэтому

$$Z' = (E - Z(Z^T Z)^{-1} Z^T) AZ. \quad (17)$$

Умножая (17) на Z^T , получаем равенство

$$Z^T Z' = Z^T AZ - (Z^T Z)(Z^T Z)^{-1} Z^T AZ = 0;$$

транспонируя равенство $Z^T Z' = 0$, имеем

$$(Z^T)' Z = 0,$$

откуда

$$(Z^T Z)' = 0,$$

т. е.

$$Z^T(x) Z(x) = Z^T(0) Z(0). \quad (18)$$

Таким образом, если $Z(x)$ — решение (17), то матрица $Z^T(x)Z(x)$ остается обратимой при всех x , если она была обратимой при $x = 0$. Из выписанных соотношений вытекает существование решения (17), удовлетворяющего условию (18), если матрица $Z^T(0)Z(0)$ обратима. Возьмем в качестве столбцов матрицы $Z(0)$ r линейно независимых решений системы $Bz = 0$ и найдем решение системы (17) при таком начальном условии. Матрица $Z^T(0)Z(0)$ является матрицей Грама для этих r линейно независимых векторов и поэтому обратима. Для задания подпространства W решений неоднородной системы, удовлетворяющих левому граничному условию, необходимо задать еще некоторый вектор этого подпространства. Если $Bu(0) = b$, то решение системы $u' = Au + f$ при начальном условии $u(0)$ принадлежит W . Любой вектор $Z(x)q(x)$, где $q(x)$ — вектор-столбец размерности r , параллелен W , и поэтому решение системы

$$u' = Au + f - Zq \quad (19)$$

при любом $q(x)$ также принадлежит W . Потребовав минимальность $\|u'\|$ по множеству всевозможных q , получим систему, записываемую в матричной форме в виде $Z^T u' = 0$ или

$$Z^T (Au + f - Zq) = 0.$$

Отсюда

$$q = (Z^T Z)^{-1} Z^T (Au + f),$$

и окончательный вид системы (19) будет

$$u' = (E - Z(Z^T Z)^{-1} Z^T)(Au + f). \quad (20)$$

Искомое решение задачи принадлежит многообразию $u + Zc$. Значение $c(x)$ определяется из правого граничного условия $Du(X) = d$. Подставляя $y = u + Zc$ в уравнение $y' = Ay + f$, имеем

$$(E - P)(Au + f) + Zc' + (E - P)AZc = Au + AZc + f,$$

здесь

$$P = Z(Z^T Z)^{-1} Z^T;$$

отсюда

$$Zc' = PAZc + PAu + Pf,$$

иначе

$$c' = Z^{-1}P(AZc + Au + f). \quad (21)$$

Решаем эту систему в направлении убывания x , а затем определяем $y(x) = u(x) + Z(x)c(x)$.

Возникает вопрос, какой из двух методов дифференциальной прогонки — (8)–(10) или (17), (20), (21) — является более предпочтительным. На этот вопрос трудно дать однозначный ответ, не учитывая конкретные свойства рассматриваемой системы. Второй метод предпочтительнее тем, что согласно построению соответствующие ему функции $Z(x)$, $u(x)$, вообще говоря, должны быть более гладкими; поэтому можно ожидать, что при равной точности окончательного результата интегрирование системы (17) будет происхо-

дить с более крупным шагом, чем интегрирование системы (8). С другой стороны, при интегрировании системы уравнений (17) объем вычислений на шаге может оказаться несколько большим из-за необходимости обращения матрицы $Z^T Z$. Может возникнуть желание отказываться от обращения матрицы $Z^T Z$ на каждом шаге интегрирования, поскольку эта матрица не зависит от x , особенно в случае, когда столбцы матрицы $Z(0)$ ортонормированы и матрица $Z^T(0)Z(0)$ соответственно единичная. При таких упрощениях исходной системы (17), (20), (21) решения системы не меняются; однако может измениться в нежелательную сторону характер устойчивости решений системы к округлениям, и поэтому при отсутствии детального анализа от подобного предложения стоит отказаться.

Для уравнения $y'' - p(x)y = f(x)$, рассматривавшегося в § 4, случайно оказалось, что замыкание алгоритма регулярно и решения уравнений замыкания одновременно устойчивы к округлениям. Для системы уравнений (5.1) это уже не так. Замыкание алгоритма ортогональной (8)–(10) прогонки регулярно в том смысле, что решения (8)–(10) остаются ограниченными не очень большими величинами; в то же время система уравнений в вариациях для уравнения (8) может иметь сильно растущие решения, и поэтому возникают сомнения в практической устойчивости этого алгоритма. На самом деле, вопреки этим сомнениям, рассматриваемый алгоритм обладает практической устойчивостью к округлениям. Дело заключается в некоторой неоднозначности понятия замыкания вычислительного алгоритма. При формальном построении этого замыкания мы можем получать различные дифференциальные уравнения относительно одной и той же функции; решения некоторых из них устойчивы к округлениям, других — неустойчивы. Правильный ответ на вопрос об устойчивости будет получен в случае, если замыкание правильно отражает все свойства рассматриваемого алгоритма. В случае метода ортогональной прогонки вместо уравнения (8) было бы правильнее получить уравнение

$$Z' = AZ - Z(Z^T Z)^{-1}R,$$

решения которого действительно устойчивы к возмущениям.

Последние высказывания могли окончательно подорвать доверие читателя к изучению замыканий вычислительных алгоритмов. Однако следует иметь в виду, что мы предлагаем использовать замыкания алгоритмов именно при предварительном изучении и конструировании новых алгоритмов в случае, когда проведение полного анализа погрешности может потребовать большого времени и неоправданных затрат. Свойства получаемых при таком конструировании алгоритмов должны проверяться затем путем вычислительного эксперимента.

Задачи.

1. Построить численный метод решения краевой задачи, замыканием которого был бы алгоритм (17), (20), (21).

2. Построить аналоги описанных алгоритмов ортогональной прогонки, приспособленные специально для решения краевых задач в случае системы $y'' - A(x)y = f(x)$.

§ 8. Нелинейные краевые задачи

Существует аналогия между методами решения этих задач и нелинейных алгебраических задач. Так же как в последнем случае, мы проведем обсуждение различных методов, не приходя в конце концов ни к какой рекомендации по решению произвольных нелинейных краевых задач. По существу всякий метод решения нелинейной краевой задачи сводит ее решение к решению

некоторой нелинейной системы уравнений. Различные методы решения нелинейных краевых задач отличаются выбором параметров этих вспомогательных задач и, естественно, методом решения этих задач.

Подобно рассматриваемой ниже задаче стрельбы по цели многие краевые задачи успешно решались задолго до появления ЭВМ и самого понятия дифференциального уравнения. Аналогия с многими такими домашними методами решения краевых задач часто помогает и теперь при разработке новых методов решения.

Рассмотрим случай положения орудия, цели и наблюдателя на одном уровне над горизонтом. Траектория снаряда описывается системой трех дифференциальных уравнений второго порядка, граничные условия — траектория снаряда начинается от орудия и кончается у цели. Мы не будем выписывать эту систему дифференциальных уравнений, поскольку ее конкретный вид никак не используется. Стреляющий имеет в своем распоряжении два параметра — азимут ствола орудия α и его угол над горизонтом β . При отсутствии заранее проведенной пристрелки и учета влияния атмосферных условий первый снаряд (что соответствует решению задачи Коши для нулевого приближения к искомым начальным условиям) разорвется где-то в окрестности цели. Рассмотрим сначала простейший случай, когда наблюдатель находится возле орудия. Наблюдая отклонение разрыва от направления на цель $\Delta\alpha$, он изменяет угол α на величину $\Delta\alpha$ с тем, чтобы при следующем выстреле направление на разрыв и цель совпало. Вследствие ряда факторов, например неоднородности атмосферных условий, вообще говоря, такого совпадения не произойдет, однако зачастую дальнейшего уточнения направления на цель уже не требуется. Затем наблюдатель подбирает такие углы β , чтобы иметь в распоряжении два разрыва на одной прямой с целью — один ближе, другой дальше него, как говорят, «зажимает цель в вилку». Если в его распоряжении нет никаких технических средств, то он может судить только о том, происходит разрыв ближе или дальше цели. Если $g(\beta)$ — дальность полета снаряда, а g_0 — расстояние до цели, то ему приходится решать уравнение $g(\beta) = g_0$, получая при каждом выстреле только информацию о знаке $g(\beta) - g_0$. Поэтому он производит операцию деления вилки надвое, каждый раз выбирая угол β посередине между ближайшими между собой значениями β , соответствующим перелету ($\text{sign}(g(\beta) - g_0) > 0$) и недолету ($\text{sign}(g(\beta) - g_0) < 0$).

Задача существенно усложняется, если наблюдательный пункт находится не там, где орудие. В этом случае пристрелка может производиться по более сложному правилу, аналогичному методу Зейделя. Наблюдатель получает информацию об угле γ отклонения разрыва от направления на цель с наблюдательного пункта, а при $\gamma = 0$ также о сравнительной дальности от него положения разрыва и цели. Сначала при фиксированном угле β наблюдатель корректирует угол α так, чтобы направление на цель и разрыв совпали. Угол γ между направлениями на цель и разрыв является какой-то функцией $\gamma(\alpha, \beta)$. Таким образом, на этом этапе наблюдатель итерационно решает уравнение $\gamma(\alpha, \beta) = 0$ при фиксированном β . Поскольку известно значение $\gamma(\alpha, \beta)$, то здесь может быть применен метод, более быстрый по сравнению с делением отрезка пополам, например метод секущих. Получившееся значение угла α зависит от β , и мы будем обозначать через $\alpha(\beta)$ решение уравнения $\gamma(\alpha, \beta) = 0$. При положении разрыва, цели и наблюдателя на одной прямой последний может указать, в каком направлении следует изменить угол β ; при этом новом значении β окажется, что $\gamma \neq 0$ и далее потребуются уточнение значения α . Имея в распоряжении вилку из разрывов, находящихся на прямой наблюдатель — цель, он может начать делить пополам интервал, где находится искомое значение β . Обозначим через g_0 — расстояние от наблюдателя до цели, а через $g(\alpha, \beta)$ — расстояние от него до разрыва при

заданных α и β . Таким образом, задача состоит в решении системы

$$\gamma(\alpha, \beta) = 0, \quad g(\alpha, \beta) = g_0.$$

Совокупность действий наблюдателя в целом можно характеризовать следующим образом. Он решает задачу, подбирая β на основании информации о знаке разности $g(\alpha(\beta), \beta) - g_0$; здесь $\alpha(\beta)$ — решение уравнения $\gamma(\alpha, \beta) = 0$, которое в свою очередь находится каким-то итерационным методом. Конечно, опытный наблюдатель практически снижает требуемое число итераций (выстрелов), интуитивно применяя какую-то комбинацию указанных выше классических численных методов.

Существо большинства методов решения нелинейных краевых задач можно описать следующим образом. Рассматриваемая задача сводится к решению нелинейной системы уравнений относительно не очень большого числа неизвестных; далее эта система решается при помощи какого-либо способа решения нелинейных систем. Выбор вспомогательных переменных, как говорят, *параметризация задачи*, часто имеет решающее значение для ее успешного решения.

Рассмотрим нелинейную краевую задачу

$$\begin{aligned} y' &= f(x, y), \quad B(y(0)) = 0, \quad D(y(X)) = 0, \\ y &= (y_1, \dots, y_l)^T, \quad B = (b_1, \dots, b_{l-r})^T, \quad D = (d_1, \dots, d_r)^T. \end{aligned} \quad (1)$$

Пусть функции $g_1(y(0)), \dots, g_r(y(0))$ таковы, что система уравнений

$$\begin{aligned} b_1(y(0)) &= 0, \dots, b_{l-r}(y(0)) = 0, \\ g_1(y(0)) &= g_1, \dots, g_r(y(0)) = g_r \end{aligned} \quad (2)$$

однозначно определяет вектор $y(0) = \omega_0(g)$, $g = (g_1, \dots, g_r)$. Тогда задача (1) может быть сведена к системе нелинейных уравнений относительно параметров $g_1, \dots, g_r$.

Довольно часто граничное условие в точке 0 имеет вид

$$y_1(0) = 0, \dots, y_{l-r}(0) = 0;$$

тогда в качестве $g_1(y(0)), \dots, g_r(y(0))$ целесообразно взять $y_{l-r+1}(0), \dots, y_l(0)$. Таким образом, здесь в качестве искомых параметров выступают неизвестные компоненты решения в точке $x = 0$. Как правило, требуемая система нелинейных уравнений не выписывается явно, но можно указать алгоритм вычисления ее правой части.

При каждом $g_1, \dots, g_r$, решая систему (2) можно определить вектор $y(0) = \omega_0(g)$; решая задачу Коши при начальном условии $y(0)$, определяем $y(X) = \omega_X(g)$ и затем находим $\psi(g) = D(\omega_X(g))$. Таким образом, решение задачи (1) сводится к решению нелинейной системы

$$\psi(g) = 0. \quad (3)$$

Эта система имеет следующие особенности, которые следует учесть при выборе метода ее решения:

1. Вычисление каждого значения ψ обходится слишком дорого — оно требует численного решения задачи Коши; поэтому надо ориентироваться на методы, требующие по возможности меньшего вычисления значений ψ .

2. Объем вычислений при нахождении вектора ψ или одной его компоненты ψ_i практически одинаков; поэтому применение методов типа Зейделя, где на каждом шаге используется информация о значении только одной компоненты ψ_i , вообще говоря, должно быть нерациональным.

Описанный выше способ пристрелки по цели аналогичен методу Зейделя, но его применение оправдывается спецификой задачи в связи с ограниченными возможностями наблюдателя.

Значительную группу методов решения нелинейных систем составляют методы, где наряду со значениями функций ψ_i используются значения их производных, отыскиваемые двояко.

1. Для определенности речь пойдет о вычислении производных $\frac{\partial \psi_i}{\partial g_1}$. Пусть мы задались значениями $g_1^0, \dots, g_r^0$. Из системы уравнений (2) найдем соответствующие $y_1(0), \dots, y_l(0)$. Дифференцируя уравнения системы (2) по g_1 , получаем систему уравнений для отыскания производных $\frac{\partial y_j(0)}{\partial g_1}$

$$\sum_{j=1}^l \frac{\partial b_k}{\partial y_j(0)} \frac{\partial y_j(0)}{\partial g_1} = 0, \quad k = 1, \dots, l - r,$$

$$\sum_{j=1}^l \frac{\partial g_m}{\partial y_j(0)} \frac{\partial y_j(0)}{\partial g_1} = \delta_{m-1}, \quad m = 1, \dots, r.$$

Решение системы $y' = f(x, y)$ при начальных условиях $y_1(0), \dots, y_l(0)$ является функцией от параметров g_m ; вектор производных

$$\frac{\partial y(x)}{\partial g_1} = \left(\frac{\partial y_1(x)}{\partial g_1}, \dots, \frac{\partial y_l(x)}{\partial g_1} \right)^T$$

удовлетворяет системе дифференциальных уравнений

$$\frac{d}{dx} \left(\frac{\partial y(x)}{\partial g_1} \right) = f_y(x, y) \frac{\partial y(x)}{\partial g_1}, \quad f_y = \left[\frac{\partial f_i}{\partial y_j} \right]. \quad (4)$$

После численного решения этой системы получаем вектор $\frac{\partial y(X)}{\partial g_1}$. Теперь можно найти производные

$$\frac{\partial \psi_i}{\partial g_1} = \sum_{j=1}^l \frac{\partial d_i}{\partial y_j} \frac{\partial y_j(X)}{\partial g_1} = \left(\text{grad } d_i, \frac{\partial y(X)}{\partial g_1} \right). \quad (5)$$

Этот путь может быть целесообразен, если, например, многократно решается краевая задача (1) при одной и той же правой части $f(x, y)$, но при различных граничных условиях.

В качестве стандартного метода определения производных $\frac{\partial \psi_i}{\partial g_j}$ этот метод неудобен тем, что наряду с написанием программы вычисления правых частей уравнения требует от пользователя также написания программы вычисления производных $\frac{\partial f_i}{\partial y_j}$.

2. Для большинства итерационных методов решения систем уравнений значения производных правых частей требуются лишь с умеренной точностью. Поэтому для нахождения этих производных можно просто воспользоваться какими-либо формулами численного дифференцирования, например, простейшей

$$\frac{\partial \psi_i(g_1, \dots, g_r)}{\partial g_1} \approx \frac{\psi_i(g_1 + \Delta, g_2, \dots, g_r) - \psi_i(g_1, \dots, g_r)}{\Delta}.$$

Вычисление производных $\frac{\partial \psi_i}{\partial g_1}$ по этой формуле требует дополнительно (по сравнению с нахождением значений $\psi_i(g_1, \dots, g_r)$) еще одного численного решения задачи Коши, соответствующей параметрам $g_1 + \Delta, g_2, \dots, g_r$. Поскольку при численном нахождении значений функций ψ_i может иметь место большая погрешность, которая еще более возрастет при делении на Δ , довольно часто вопрос рационального выбора Δ требует дополнительного изучения. Для описываемого метода типично одновременное нахождение производных всех функций ψ_i по фиксированной переменной g_j ; поэтому здесь может оказаться разумным применение итерационных процессов с поочередным уточнением параметров g_j . Можно представить себе такой итерационный процесс: параметры g_j уточняются в циклическом порядке; уточненное значение параметра $\bar{g}_j$ выбирается из условия минимума некоторой функции

$$\Phi_j(\psi_1(g_1, \dots, g_r) + \frac{\partial \psi_1(g_1, \dots, g_r)}{\partial g_j}(\bar{g}_j - g_j), \dots, \psi_r(g_1, \dots, g_r) + \frac{\partial \psi_r(g_1, \dots, g_r)}{\partial g_j}(\bar{g}_j - g_j)).$$

(см. подробнее § 4 гл. VII).

Так же как при решении линейных краевых задач, возникает вопрос о применимости метода в условиях реальных округлений. Если решения η системы уравнений в вариациях $\eta' = f_\eta \eta$ сильно растут с ростом x , то погрешности в значениях g_j и вычислительные погрешности при численном интегрировании приводят

к большим ошибкам в значениях функции ψ_i ; это в конечном счете приведет к большой погрешности получаемого решения. Если такая погрешность окажется недопустимой, то следует ввести иную параметризацию задачи. Опишем способы решения, где в качестве формальных параметров выступают значения решения на всем отрезке.

Один из предлагавшихся ранее способов решения нелинейной задачи $P(u) = 0$ заключался в следующем: выбирался линейный оператор Lu , действующий в том же пространстве, что и P . Последующие приближения выбирались из равенства

$$L(u_{n+1} - u_n) + P(u_n) = 0.$$

Например, в конкретном случае задачи (1) можно рассмотреть оператор L , определяемый равенствами

$$Ly = C(x)y' - A(x)y, \quad 0 < x < X,$$

$$Ly|_{x=0} = B_0 y(0),$$

$$Ly|_{x=X} = D_0 y(X);$$

$C(x)$, $A(x)$, B_0 , D_0 — некоторые матрицы. Тогда последующие приближения будут определяться как решения линейной системы

$$C(x)(y_{n+1} - y_n)' - A(x)(y_{n+1} - y_n) + y_n' - f(x, y_n) = 0,$$

$$B_0(y_{n+1}(0) - y_n(0)) + B(y_n(0)) = 0, \quad (6)$$

$$D_0(y_{n+1}(X) - y_n(X)) + D(y_n(X)) = 0.$$

В конкретном случае метода Ньютона (см. § 2 гл. VII) оператор L зависит от n и является оператором производной, соответствующим приближению $y_n(x)$:

$$Lu = u' - f_y(x, y_n(x))u \quad \text{при } x \in (0, X),$$

$$Lu|_{x=0} = B_y(y_n(0))u(0),$$

$$Lu|_{x=X} = D_y(y_n(X))u(X);$$

здесь

$$B_y = \left[\frac{\partial b_i}{\partial y_j(0)} \right], \quad D_y = \left[\frac{\partial d_i}{\partial y_j(X)} \right].$$

Тогда соотношения (6) приобретут вид

$$y_{n+1}' - f_y(x, y_n(x))(y_{n+1} - y_n) - f(x, y_n(x)) = 0, \quad 0 < x < X,$$

$$B_y(y_n(0))(y_{n+1}(0) - y_n(0)) + B(y_n(0)) = 0, \quad (7)$$

$$D_y(y_n(0))(y_{n+1}(0) - y_n(0)) + D(y_n(0)) = 0.$$

Согласно общим теоремам о сходимости метода Ньютона, итерационный процесс (5) сходится в случае достаточно гладких $f(x, y)$, b_i и d_i , если начальное приближение достаточно близко к искомому решению. Естественно, что методы решения вспомогательных линейных задач, например задачи (7), следует выбирать в достаточной мере устойчивыми к влиянию вычислительной погрешности.

В только что описанном способе в качестве отыскиваемых параметров задачи формально выступают значения решения на всем отрезке интегрирования. Поскольку решение линейной краевой задачи производится численно, то в реальных вычислениях будут участвовать приближения y_n^k к значениям решения $y(x_k)$ в узлах сетки $0 = x_0 < x_1 < \dots < x_N = X$. Например, если численное интегрирование производится при помощи метода Эйлера, то вместо (7) имеем соотношения

$$\frac{y_{n+1}^{j+1} - y_{n+1}^j}{x_{j+1} - x_j} - f_y(x_j, y_n^j)(y_{n+1}^j - y_n^j) - f(x_j, y_n^j) = 0, \quad 0 \leq j < N,$$

$$B_y(y_n^0)(y_{n+1}^0 - y_n^0) + B(y_n^0) = 0,$$

$$D_y(y_n^N)(y_{n+1}^N - y_n^N) + D(y_n^N) = 0.$$

Такая параметризация краевой задачи неудобна тем, что при большом числе узлов интегрирования и большом порядке системы требуется большая загрузка памяти ЭВМ. Может случиться, что приближения $y_n(x)$ имеют различный характер гладкости в том смысле, что гладкость одних приближений требует мелкого шага интегрирования на одних участках, а гладкость других приближений — на других. Тогда описываемый метод потребует модернизации, связанной с пересчетом значений этих функций с одной сетки на другую.

При решении систем нелинейных уравнений иногда рекомендовалось с целью уменьшения объема вычислений проводить итерации по методу Ньютона с неточными значениями производных правых частей. Для задачи (1) аналогичный прием помогает ниже уменьшить объем используемой оперативной памяти ЭВМ. Зададимся точками

$$0 = X_0 < X_1 < \dots < X_m = X$$

и после завершения каждого шага итерации запоминаем лишь приближения Y_j^n , $j = 0, \dots, m-1$, к значениям решения в точках $X_0, \dots, X_{m-1}$. Если решения системы уравнений в вариациях на искомом решении не сильно растут на отрезках $[X_{j-1}, X_j]$, то после нахождения всех $y(X_j)$ значения решения на всем отрезке могут быть найдены численным интегрированием исходной системы по отрезкам $[X_{j-1}, X_j]$.

Распишем метод более подробно. Пусть $Y = (y_0, \dots, y_{m-1})$, P — оператор, ставящий в соответствие всякой функции $z(x)$ матрицу $Pz = (z(X_0), \dots, z(X_{m-1}))$, а Q — оператор, ставящий в соответствие матрице Y функцию $u(x) = QY$, определенную на $[0, X]$, причем $QPy = y$, если y — решение исходной задачи. Предположим, имеется какой-то исходный итерационный процесс, где последующее приближение находится по предыдущему как решение системы

$$Y'_{n+1} = A(x, Y_n, Y_{n+1})$$

при граничных условиях $b(Y_n(0), Y_{n+1}(0)) = 0$, $d(Y_n(X), Y_{n+1}(X)) = 0$, причем искомое решение является стационарной точкой и для итерационного процесса:

$$y' = A(x, y, y),$$

$$b(y(0), y(0)) = 0, \quad d(y(X), y(X)) = 0.$$

Очевидно, тогда искомое решение является стационарной точкой и для итерационного процесса:

$$Y'_{n+1} = A(x, QPY_n, Y_{n+1}),$$

$$b(QPY_n|_{x=0}, Y_{n+1}(0)) = 0, \quad (8)$$

$$d(QPY_n|_{x=X}, Y_{n+1}(X)) = 0,$$

т. е.

$$y' = A(x, QPy, y), \quad b(QPy|_{x=0}, y) = 0, \quad d(QPy|_{x=X}, y) = 0.$$

Рассмотрим одну конкретную реализацию этого итерационного процесса. В качестве оператора Q возьмем оператор, ставящий в соответствие матрице $Y = (Y_0, \dots, Y_{m-1})$ функцию $u(x)$, которая на полуинтервалах $[X_0, X_1), \dots, [X_{m-2}, X_{m-1})$ и отрезке $[X_{m-1}, X_m]$ является решением задачи Коши для системы $u' = f(x, u)$ при начальных условиях $u(X_j) = Y_j$. Предположим, что

$$A(x, y_n, y_{n+1}) = g_0(x, y_n) + g_1(x, y_n) y_{n+1},$$

$$b(y_n(0), y_{n+1}(0)) = b_0(y_n(0)) + b_1(y_n(0)) y_{n+1}(0),$$

$$d(y_n(X), y_{n+1}(X)) = d_0(y_n(X)) + d_1(y_n(X)) y_{n+1}(X).$$

Тогда для решения линейной краевой задачи

$$y'_{n+1} = g_0(x, QPy_n) + g_1(x, QPy_n) y_{n+1}, \quad (9)$$

$$b_0((QPy_n)|_{x=0}) + b_1((QPy_n)|_{x=0}) y_{n+1}(0) = 0,$$

$$d_0((QPy_n)|_{x=X}) + d_1((QPy_n)|_{x=X}) y_{n+1}(X) = 0$$

может быть применен какой-либо из методов решения линейных краевых задач.

Применим метод ортогональной прогонки в первой форме при условии, что точки $X_1, \dots, X_{m-1}$ являются как раз теми точками, в которых производится ортогонализация решения. При этом потребуется хранить в памяти значения некоторых векторов и матриц, соответствующих точкам X_j , и на отрезках $[X_j, X_{j+1}]$ осуществлять численное интегрирование нескольких решений однородной системы

$$v' = g_1(x, QPy_n) v \quad (10)$$

и решения неоднородной системы (9). Поскольку QPy_n , согласно определению, является решением системы

$$(QPy_n)' = f(x, QPy_n), \quad (11)$$

то интегрирование систем (9) — (11) можно осуществлять одновременно, не прибегая к запоминанию значений этих функций в промежуточных узлах.

Произведем усовершенствование алгоритма, представляющее интерес и при решении других нелинейных и линейных задач.

Приближения к искомому решению получаются в виде линейных комбинаций частных решений систем (9) и (10). В случае сходимости к решению имеем $QPy_n(x) \rightarrow y(x)$ и системы (9), (10) переходят в системы

$$u' = g_0(x, y) + g_1(x, y) u,$$

$$v' = g_1(x, y) v.$$

Например, в случае метода Ньютона эти системы имеют вид

$$u' = f(x, y(x)) + f_y(x, y(x)) (u - y(x)),$$

$$v' = f_y(x, y(x)) v.$$

Иногда (например, при решении задач с пограничным слоем) отыскиваемое решение довольно гладкое, за исключением окрестностей концов отрезка интегрирования; в то же время на протяжении всего отрезка интегрирования решения системы (10) резко меняются. В результате этого даже там, где искомое решение гладкое, для его получения с приемлемой точностью требуется интегрировать вспомогательные системы уравнений (10), (11) с очень мелким шагом. Чтобы избежать этого, полезно перейти к новой неизвестной функции

$$r_{n+1} = y_{n+1} - QPy_n.$$

Для определенности будем проводить рассуждения на примере метода Ньютона. Тогда в точках непрерывной дифференцируемости $r_{n+1}(x)$ имеем

$$r'_{n+1} = f(x, QPy_n) - (QPy_n)' + f_y(x, QPy_n) r_{n+1},$$

иначе

$$r'_{n+1} = f_y(x, QPy_n) r_{n+1}. \quad (12)$$

Соответственно граничными условиями будут

$$\begin{aligned} B((QPy_n)|_{x=0}) + B_y((QPy_n)|_{x=0}) r_{n+1}(0) &= 0, \\ D((QPy_n)|_{x=X}) + D_y((QPy_n)|_{x=X}) r_{n+1}(X) &= 0. \end{aligned} \quad (13)$$

Функция y_{n+1} определяется как непрерывно дифференцируемое решение (9), а QPy_n , возможно, имеет разрывы в точках $X_1, \dots, X_{m-1}$. Поэтому

$$[r_{n+1}]_{X_j} = -[QPy_n]_{X_j}, \quad j = 1, \dots, m-1;$$

здесь

$$[g]_x = \lim_{\varepsilon \rightarrow +0} (g(x + \varepsilon) - g(x - \varepsilon)).$$

Систему уравнений для отыскания $r_{n+1}(x)$ можно записать в обобщенной форме в виде

$$r'_{n+1} = g_n(x) + f_y(x, QPy_n) r_{n+1}, \quad (14)$$

$$g_n(x) = - \sum_{j=1}^{m-1} [QPy_n]_{X_j} \delta(x - X_j);$$

здесь $\delta(t)$ — δ -функция. При решении краевой задачи (13) — (14) методом ортогональной прогонки с точками ортогонализации $X_1, \dots, X_{m-1}$ наличие δ -функций в правых частях этой системы не вносит никаких осложнений: после интегрирования однородного уравнения (10) на отрезках $[X_{j-1}, X_j]$ к получившемуся значению нужно добавлять величину $-[QPy_n]_{X_j}$.

Могло бы показаться, что такая замена неизвестной функции ничего не меняет, поскольку однородная система уравнений, соответствующая системам (9) и (14), является одной и той же. В действительности между этими двумя итерационными процессами имеется существенное различие. В первом случае приближения $y_{n+1}(x)$ получают на каждом участке $[X_{j-1}, X_j]$ как ли-

нейные комбинации $u_n(x) + \sum_{i=1}^r c_{in} v_{in}(x)$ решений уравнений (9) и (10),

причем, поскольку правая часть (9) не стремится к нулю, то $u_n(x) \not\rightarrow 0$ и, соответственно, не стремится к нулю вектор коэффициентов c_{in} . Во втором случае также имеем

$$r_{n+1} = u_n(x) + \sum_{i=1}^r c_{in} v_{in}(x).$$

Поскольку векторы $u_n(X_j)$, $v_{in}(X_j)$ ортогональны, а $r_{n+1}(X_j) \rightarrow 0$, то $u_n(x) \rightarrow 0$ и $c_{in} \rightarrow 0$ при $n \rightarrow \infty$. Функции $v_{in}(x)$ умножаются на малые множители, и поэтому ошибки при их численном отыскании существенно меньше сказываются на получаемых приближениях. Точно так же и для неоднородного уравнения относительно $u_n(x)$ есть основания ожидать сужения требования на шаг интегрирования вследствие стремления $u_n(x)$ к нулю.

При грубом интегрировании уравнений для функций u_n , v_{in} итерационный процесс несколько видоизменится, не исключено, что произойдет некоторое замедление скорости сходимости; однако есть основания надеяться, что в целом свойства алгоритма улучшатся за счет меньшего объема вычислений на каждом шаге итерации. Поскольку высокая точность в определении частных решений систем (12), (14) не требуется, то, кроме непосредственного численного интегрирования (12), можно прибегнуть и к более грубым способам интегрирования. Для определения в точках X_j значений решений однородной (12) и неоднородной системы (14) при заданных условиях в точках X_{j-1} требуется осуществить численное интегрирование уравнений в вариациях $\eta' = f_y(x, QPy_n)\eta$. Согласно определению QPy_n удовлетворяет на $[X_{j-1}, X_j]$ исходной системе

$$(QPy_n)' = f(x, QPy_n).$$

Пусть $y_\varepsilon(x)$ — решение этой же системы при начальном условии

$$y_\varepsilon(X_{j-1}) = QPy_n|_{X_{j-1}} + \varepsilon v_{j-1}.$$

Согласно выводу системы уравнений в вариациях вектор-функция $(y_\varepsilon(x) - QPy_n)\varepsilon^{-1}$ при малых ε приближает решение $R(x)$ уравнения (12) при на-

тельном условии $R(X_{j-1}) = v_{j-1}$. На каждом из отрезков $[X_{j-1}, X_j]$ нам требуется найти $r + 1$ решение системы $R' = f_y(x, QPy_n)R$. Отыскание всех этих решений на каждом отрезке $[X_{j-1}, X_j]$ по описанной выше процедуре потребует проинтегрировать систему $y' = f(x, y)$ $r + 2$ раза.

К итерационным методам решения краевых задач в равной мере относятся рассуждения из гл. VII о целесообразности повышения точности вычислений лишь по мере приближения к решению. Из проведенных нами рассуждений по поводу последнего из описанных методов видно, что эту точность целесообразно существенно повышать лишь при вычислении правых частей неоднородных уравнений, т. е. величин $[QPy_n](X_j)$; таким образом, достаточно повышать точность лишь при интегрировании уравнения

$$(QPy_n)' = f(x, QPy_n).$$

Так же как в случае нелинейных алгебраических задач, можно привести много различных методов, которые сходятся к решению при достаточно хорошем начальном приближении. И так же как там, нельзя предложить метод, пригодный для эффективного решения любых нелинейных краевых задач. Поэтому при практическом решении конкретных нелинейных задач обычно приходится заниматься «доводкой» метода: предлагается какой-то специальный метод получения начального приближения, метод модернизируется с целью расширения области начальных условий, при которых он сходится для данного конкретного класса задач, и т. д.

§ 9. Аппроксимации специального типа

Рассмотрим краевую задачу

$$(k(x)y'_x)' - p(x)y = f(x), \quad y(0) = a, \quad y(X) = b, \quad (1)$$

где $p(x) \geq 0$, $k(x)$ трижды, $p(x)$, $f(x)$ — дважды непрерывно дифференцируемы, за исключением конечного числа точек, где они или их производные k' , k'' , k''' , p' , p'' , f' , f'' могут иметь разрывы первого рода.

Пусть K , P , F — множества точек разрыва, соответственно, функций k , p , f или их производных; $\Omega = K \cup P \cup F$.

Решением задачи (1) будем называть функцию $y(x)$, удовлетворяющую следующим условиям:

- 1) $y(x)$ непрерывна на $[0, X]$;
- 2) $y(x)$ удовлетворяет уравнению всюду на $[0, X]$, за возможным исключением точек множества Ω ;
- 3) функция $w(x) = k(x)y'_x$, называемая «поток», непрерывна на $[0, X]$.

Из последнего условия следует, что функция $y'_x = \frac{w(x)}{k(x)}$ непрерывна всюду, за исключением точек разрыва функции $k(x)$; в этих точках y'_x будет иметь разрывы первого рода.

Можно показать, что производные $y'''(x)$ и $y^{(4)}(x)$ непрерывны и равномерно ограничены на множестве $[0, X] \setminus \Omega$.

Если при построении разностной схемы не учитывается факт разрывности y'_x , то может получиться разностная задача, решение которой не сходится к решению задачи (1).

Для построения конечно-разностных схем в случае уравнений с разрывами и другими особенностями коэффициентов и решений полезно ввести новые переменные (функции или аргументы) так, чтобы решение новой задачи обладало большей гладкостью. Часто такой переход осуществляется не совсем явным образом, подобно тому, как показано ниже.

Поскольку $w(x)$ непрерывна, а производная

$$w'_x = p(x)y + f(x) \quad (2)$$

непрерывна, за исключением точек разрыва $p(x)$ и $f(x)$, то функция $w(x)$ непрерывно-дифференцируема. Как при целых, так и не при целых s будем обозначать $x_s = sh$. Интегрируя (2) в пределах от $x_{q-1/2}$ до $x_{q+1/2}$, получим

$$w(x_{q+1/2}) - w(x_{q-1/2}) = \int_{x_{q-1/2}}^{x_{q+1/2}} (p(x)y(x) + f(x)) dx. \quad (3)$$

Так как функция $y(x)$ кусочно-дифференцируема, то

$$y(x) = y(x_q) + O(h) \quad \text{при} \quad x - x_q = O(h).$$

Поэтому (3) можно переписать в виде

$$\begin{aligned} w(x_{q+1/2}) - w(x_{q-1/2}) &= \int_{x_{q-1/2}}^{x_{q+1/2}} (p(x)y(x_q) + f(x) + O(h)) dx = \\ &= y(x_q) \bar{p}_q + \bar{f}_q + O(h^2); \end{aligned}$$

здесь

$$\bar{p}_q = \frac{1}{h} \int_{x_{q-1/2}}^{x_{q+1/2}} p(x) dx, \quad \bar{f}_q = \frac{1}{h} \int_{x_{q-1/2}}^{x_{q+1/2}} f(x) dx.$$

После деления на h получится соотношение

$$\frac{w(x_{q+1/2}) - w(x_{q-1/2})}{h} - \bar{p}_q y(x_q) = \bar{f}_q + O(h). \quad (4)$$

В случае, если интервал (x_{j-1}, x_j) содержит точки разрыва y'_x , погрешность от непосредственной замены $\frac{w(x_{j-1/2})}{h}$ на $k(x_{j-1/2}) \frac{y(x_j) - y(x_{j-1}))}{h^2}$ может оказаться величиной порядка h^{-1} .

Для получения лучшей аппроксимации введем в рассмотрение вспомогательную независимую переменную $t = \int_0^x \frac{dx}{k(x)}$. Из ограниченности производной по x и равенства $\frac{du}{dt} = k(x) \frac{du}{dx}$ следует ограниченность производной по t . Функция $w = k(x) \frac{dy}{dx} = \frac{dy}{dt}$ имеет ограниченную производную по x , а следовательно и по t . Таким образом, вторая производная y''_{tt} ограничена и можно написать равенство

$$k(x) \frac{dy}{dx} \Big|_{x_{j-1/2}} = \frac{dy}{dt} \Big|_{t_{j-1/2}} = \frac{y(t_j) - y(t_{j-1}))}{t_j - t_{j-1}} + O(t_j - t_{j-1});$$

здесь

$$t_n = \int_0^{x_n} \frac{dx}{k(x)}; \quad t_j - t_{j-1} = \int_{x_{j-1}}^{x_j} \frac{dx}{k(x)} = O(h).$$

Поэтому

$$w(x_{j-1/2}) = k(x) \frac{dy}{dx} \Big|_{x_{j-1/2}} = \frac{y(x_j) - y(x_{j-1}))}{t_j - t_{j-1}} + O(h). \quad (5)$$

После подстановки $k(x) \frac{dy}{dx} \Big|_{x_{q \pm 1/2}}$ в левую часть (4) получим

$$r_q = \frac{\frac{y(x_{q+1}) - y(x_q)}{h \bar{k}_{q+1/2}^{-1}} - \frac{y(x_q) - y(x_{q-1}))}{h \bar{k}_{q-1/2}^{-1}}}{h} - \bar{p}_q y(x_q) - \bar{f}_q = O(1), \quad (6)$$

где

$$\bar{k}_{j+1/2}^{-1} = \frac{1}{h} \int_{x_j}^{x_{j+1}} \frac{dx}{k(x)};$$

соответствующая конечно-разностная схема имеет вид

$$L(y_q) = \frac{\frac{y_{q+1} - y_q}{h \bar{k}_{q+1/2}^{-1}} - \frac{y_q - y_{q-1}}{h \bar{k}_{q-1/2}^{-1}}}{h} - \bar{p}_q y_q = \bar{f}_q. \quad (7)$$

Максимум погрешности аппроксимации у полученной схемы есть $O(1)$; поэтому для получения оценки погрешности потребуется принять во внимание ряд дополнительных соображений.

В предположении отсутствия округлений погрешность приближенного решения, согласно (2.10), запишется в виде

$$R_n = -h \sum_{k=1}^{N-1} G_n^k r_k.$$

Используя определение G_n^k (см. § 2) и полученные там оценки, можно получить соотношения

$$\begin{aligned} \max_{0 \leq kh, nh \leq X} |G_n^k| &\leq D < \infty, \\ \max_{0 \leq kh, nh \leq X} |G_n^{k+1} - G_n^k| &= O(h). \end{aligned} \quad (8)$$

Из теории дифференциальных уравнений известно, что производные решения до четвертого порядка включительно равномерно ограничены всюду, за возможным исключением точек Ω .

Поэтому, если интервал (x_{q-1}, x_{q+1}) не содержит точек Ω , оценку (6) можно уточнить: в левой части (6) разложим $y(x)$, $k(x)$, $p(x)$ и $f(x)$ по формуле Тейлора в точке x_q с остаточными членами порядка h^4 , h^3 , h^2 и h^2 соответственно. После проведения элементарных выкладок с учетом равенства (1) получится, что $r_q = O(h^2)$.

Воспользовавшись этой оценкой, получаем равенство

$$R_n = O(h^2) - \sum'_k h G_n^k r_k.$$

Знак $\sum'_k$ означает, что суммирование ведется только по значениям k таким, что интервал (x_{k-1}, x_{k+1}) содержит точки разрыва коэффициентов уравнения или их производных соответствующего невысокого порядка. Для значений $q = k$, соответствующих $\sum'$, представим r_q в виде

$$\begin{aligned} r_q &= \frac{\frac{y(x_{q+1}) - y(x_q)}{h\bar{k}_{q+1/2}^{-1}} - \frac{y(x_q) - y(x_{q-1})}{h\bar{k}_{q-1/2}^{-1}}}{h} - \bar{p}_q y(x_q) - \bar{f}_q = \\ &= \frac{\omega_{q+1/2} - \omega_{q-1/2}}{h} + \left\{ \frac{w(x_{q+1/2}) - w(x_{q-1/2})}{h} - \bar{p}_q y(x_q) - \bar{f}_q \right\}, \end{aligned}$$

где

$$\omega_{l+1/2} = \frac{y(x_{l+1}) - y(x_l)}{h\bar{k}_{l+1/2}^{-1}} - w(x_{l+1/2}).$$

Согласно (4) выражение в фигурной скобке есть $O(h)$, число точек разрыва конечно, и поэтому выражение погрешности уже можно записать в виде

$$R_n = O(h^2) - \sum'_k G_n^k (\omega_{k+1/2} - \omega_{k-1/2}).$$

Если интервал (x_k, x_{k+1}) содержит точку разрыва, то соответствующее значение $\omega_{k+1/2}$ войдет в Σ' с коэффициентом $G_n^k - G_n^{k+1}$; если этот интервал не содержит точки разрыва, то непосредственным разложением по формуле Тейлора убеждаемся, что $\omega_{k+1/2} = O(h^2)$. Поэтому возможно дальнейшее преобразование выражения для R_n :

$$R_n = O(h^2) + \sum_k'' (G_n^{k+1} - G_n^k) \omega_{k+1/2};$$

здесь Σ'' означает суммирование по k , соответствующим интервалам (x_k, x_{k+1}) , содержащим точки разрыва $k(x)$. Согласно (5) имеем $\omega_{k-1/2} = O(h)$ равномерно по всем k . Воспользовавшись также (8), получаем окончательную оценку погрешности

$$\max_{0 \leq nh \leq x} |R_n| = O(h^2).$$

При использовании схемы (7) вычислительный процесс не зависит от положения точек разрыва. Поэтому ее относят к классу *однородных схем*.

Схема (7) на первый взгляд обладает следующим неудобством. Ее коэффициенты $\bar{k}_{q+1/2}$, $\bar{p}_q$, $\bar{f}_q$ записываются как некоторые интегралы. На самом деле можно показать, что при погрешности в значениях этих коэффициентов $O(h^2)$ погрешность приближенного решения оказывается также $O(h^2)$. Поэтому, если интервал (x_{q-1}, x_{q+1}) не содержит точек разрывов, то без потери порядка точности можно заменить $\bar{p}_q$ на $p(x_q)$, $\bar{f}_q$ на $f(x_q)$ и $\bar{k}_q \pm 1/2$ на $k(x_q \pm 1/2)$.

Погрешность аппроксимации конечно-разностных методов численного интегрирования записывалась нами в виде $E_{my}^{(m+1)}(\xi)$ и поэтому равнялась нулю в случае, если решение было многочленом степени m . Может оказаться, что решение задачи лучше приближается не многочленами, а линейными комбинациями каких-то функций, отличных от одночленов x^i . Например, известно, что решения дифференциального уравнения Бесселя $x^2 y'' + xy' + (x^2 - \nu^2)y = 0$ при больших x представляются в виде

$$(a_0 x^{-1/2} + a_1 x^{-3/2} + \dots) \cos x + (b_0 x^{-1/2} + b_1 x^{-3/2} + \dots) \sin x.$$

Предположим, по ходу решения задачи необходимо иметь в памяти ЭВМ таблицы функций Бесселя. Иногда оказывается целесообразным составить их путем численного интегрирования уравнения Бесселя, используя при больших x разностные аппроксимации точные на функциях

$$x^{-k-1/2} \cos x, \quad x^{-k-1/2} \sin x, \quad k = 0, \dots, l.$$

Иногда бывает удобно в окрестности каждого расчетного узла приблизить рассматриваемое уравнение дифференциальным уравнением, интегрируемым в явном виде, и построить

разностную схему, точную для его решений. Один алгоритм такого рода уже предлагался на стр. 511—512.

Рассмотрим случай дифференциального уравнения

$$\mu^2 y'' + p(x)y = f(x),$$

где μ — малое число, $p > 0$.

В случае $p = \text{const}$ решения этого уравнения $\exp\left(\pm i \frac{x\sqrt{p}}{\mu}\right)$ колеблются с периодом $2\pi\mu/\sqrt{p}$, т. е. очень сильно. Характерный размер изменения решения (т. е. расстояние, на котором решение существенно изменяется) имеет порядок $\mu/\sqrt{p}$, и поэтому при отказе от использования специфики данного уравнения для получения высокой точности необходимо выполнение довольно обременительного условия $h \ll \mu/\sqrt{p}$.

В окрестности каждого узла x_n рассматриваемое уравнение близко к уравнению $\mu^2 y'' + p_n y = f_n$, $p_n = p(x_n)$, $f_n = f(x_n)$. Общее решение этого уравнения можно записать в виде

$$y(x) = D_1 \exp\left(i \frac{(x - x_n)\sqrt{p_n}}{\mu}\right) + D_2 \exp\left(-i \frac{(x - x_n)\sqrt{p_n}}{\mu}\right) + \frac{f_n}{p_n}; \quad (9)$$

найдем схему вида

$$a_n y_{n+1} + b_n y_n + c_n y_{n-1} - d_n = 0, \quad (10)$$

точную на всех таких решениях. Для этого подставим (9) в соотношение (10). Получим

$$\begin{aligned} a_n \left(D_1 \exp\left(i \frac{h\sqrt{p_n}}{\mu}\right) + D_2 \exp\left(-i \frac{h\sqrt{p_n}}{\mu}\right) \right) + b_n (D_1 + D_2) + \\ + c_n \left(D_1 \exp\left(-i \frac{h\sqrt{p_n}}{\mu}\right) + D_2 \exp\left(i \frac{h\sqrt{p_n}}{\mu}\right) \right) + \\ + (a_n + b_n + c_n) \frac{f_n}{p_n} - d_n = 0. \end{aligned}$$

Чтобы это равенство выполнялось при всех D_1 и D_2 , необходимо и достаточно равенства нулю коэффициентов при D_1 и D_2 и свободного члена. Приравниваем их к нулю:

$$\begin{aligned} a_n \exp\left(i \frac{h\sqrt{p_n}}{\mu}\right) + b_n + c_n \exp\left(-i \frac{h\sqrt{p_n}}{\mu}\right) &= 0, \\ a_n \exp\left(-i \frac{h\sqrt{p_n}}{\mu}\right) + b_n + c_n \exp\left(i \frac{h\sqrt{p_n}}{\mu}\right) &= 0, \\ (a_n + b_n + c_n) \frac{f_n}{p_n} - d_n &= 0. \end{aligned} \quad (11)$$

Полагая $a_n = 1$, получим

$$c_n = 1, \quad b_n = -2 \cos \frac{h \sqrt{\rho_n}}{\mu}, \quad d_n = \left(2 - 2 \cos \frac{h \sqrt{\rho_n}}{\mu} \right) \frac{f_n}{\rho_n}.$$

Общее решение системы (11) пропорционально полученному частному решению. Помножим все коэффициенты a_n , b_n , c_n , d_n на $\mu^2 h^{-2}$. Тогда получим схему

$$\mu^2 \frac{y_{n+1} - 2 \cos \frac{h \sqrt{\rho_n}}{\mu} y_n + y_{n-1}}{h^2} - \left(2 - 2 \cos \frac{h \sqrt{\rho_n}}{\mu} \right) \frac{\mu^2 f_n}{h^2 \rho_n} = 0. \quad (12)$$

Это соотношение мы примем за разностное уравнение, соответствующее узлу x_n ; такая нормировка схемы (12) является наиболее естественной: если вместо y_n подставить в (12) $y(x_n)$ и при фиксированном x_n устремить h к 0, то в пределе получится исходное дифференциальное уравнение

$$\mu^2 y'' - p(x_n) y - f(x_n) = 0.$$

При получении новых стандартных программ принято производить проверку их качества на типичных задачах. При отработке одной стандартной программы решения задачи Коши для систем дифференциальных уравнений второго порядка методом Штермера было решено проверить качество программы на примере составления таблицы солнечных затмений на несколько веков вперед в заданной точке поверхности Земли. Была выбрана неподвижная система координат, связанная с Землей, и относительно этой системы координат были выписаны дифференциальные уравнения движения Луны и Солнца. После осуществления численного интегрирования оказалось, что результаты вычислений расходятся с таблицами солнечных затмений, составленных много веков назад до появления всяких ЭВМ и даже дифференциального исчисления. В этом случае желательно повторить вычисления, взяв другой шаг по времени. Результаты вычислений с различными шагами существенно отличались, что указывало на большую ошибку вычислений. Делались попытки объяснить ошибку низким качеством стандартных программ, однако последние были виноваты лишь косвенно. Просто рассматриваемая задача не принадлежит к классу задач, для решения которых рассчитана данная стандартная программа. В выбранной системе координат за один день Луна и Солнце совершают примерно один оборот, и, следовательно, характерный размер измерения решения равенется одним суткам. При применении стандартных методов решения задач для малости погрешности нужно, чтобы шаг был существенно менее этого характерного размера (т. е. суток). Век содержит около 36 000 суток, и поэтому число шагов, необходимое для расчета затмений в течение века должно быть где-то порядка сотен тысяч. Как мы увидим далее, вычислительная погрешность метода Штермера может накапливаться как n^2 . Поэтому относительная вычислительная погрешность результата будет величиной где-то около $10^{10} t^{-1}$; здесь t — разрядность чисел в машине. Этим и объясняется большая погрешность полученного решения. Для удовлетворительного решения этой задачи нужно или построить разностные схемы специального вида или производить интегрирование в какой-то специально выбранной системе координат.

Для повышения точности интегрирования колеблющихся и резко меняющихся решений иногда используются следующие приемы.

Рассматривается уравнение, интегрируемое в явном виде, с таким же характером поведения решений, как исходное. Находят функции от решений этого уравнения, являющиеся гладкими функциями независимой переменной.

Затем эти функции принимают за новые неизвестные и интегрируют численно получающуюся систему уравнений.

Рассмотрим уравнение

$$\mu^2 x'' + p(t)x = 0. \quad (13)$$

При $p = \text{const}$ имеем

$$\mu^2 (x')^2 + px^2 = C_0, \quad \text{arctg}(x \sqrt{p}/\mu x') = C_1 t + C_2.$$

Поэтому можно ввести новые неизвестные

$$\rho = \sqrt{\mu^2 (x')^2 + p(t)x^2}, \quad \theta = \text{arctg}(x \sqrt{p(t)}/\mu x').$$

При μ малом для решений (13) справедливы асимптотические соотношения

$$x(t) \sim \frac{C}{\sqrt{p(t)}} \sin\left(\frac{1}{\mu} \int_c^t \sqrt{p(\tau)} d\tau\right),$$

$$x'(t) \sim -\frac{C \sqrt{p(t)}}{\mu} \cos\left(\frac{1}{\mu} \int_c^t \sqrt{p(\tau)} d\tau\right);$$

отсюда имеем

$$\frac{\mu^2 (x')^2}{p(t)} + \sqrt{p(t)} x^2 \sim C^2, \quad \theta = \text{arctg}\left(\frac{x \sqrt{p(t)}}{\mu x'}\right) \sim -\frac{1}{\mu} \int_c^t \sqrt{p(\tau)} d\tau.$$

В соответствии с этими соотношениями можно попробовать ввести новые неизвестные

$$\rho_1 = \sqrt{\frac{\mu^2 (x')^2}{p(t)} + \sqrt{p(t)} x^2} \quad \text{и} \quad \theta$$

и новую независимую переменную $\varphi = \int_c^t \sqrt{p(\tau)} d\tau$.

Для принятия решения, какая замена неизвестных в данном конкретном случае более целесообразна, необходимо проведение пробных расчетов.

Существо другого метода состоит в том, что на каждом шаге интегрирования рассчитывается возмущение решения упрощенной системы, проходящего через уже полученную расчетную точку.

Для примера рассмотрим систему уравнений

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} + k \frac{x}{r^3} + e \|y\| y = 0, \quad \|x\| = r, \quad e - \text{малое.}$$

Пусть приближенные значения x^n, y^n величин $x(t_n), y(t_n)$ уже известны. Пусть $x^n(t), y^n(t)$ — решение упрощенной системы

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} + k \frac{x}{r^3} = 0$$

при начальных условиях x^n, y^n . Эта система интегрируется в явном виде. Вводим новые переменные

$$\eta(t) = x_n(t) - x^n(t), \quad \gamma(t) = y_n(t) - y^n(t),$$

где $x_n(t), y_n(t)$ — точное решение исходной системы при начальных условиях x^n, y^n . Получаем некоторую систему дифференциальных уравнений относительно переменных $\eta(t), \gamma(t)$. Осуществив шаг интегрирования каким-либо

из стандартных методов при начальных условиях $\eta(t_n) = 0$, $\gamma(t_n) = 0$, получаем приближенные значения η^{n+1} , γ^{n+1} величины $\eta(t_{n+1})$, $\gamma(t_{n+1})$. Полагаем $x^{n+1} = x^n(t_{n+1}) + \eta^{n+1}$, $y^{n+1} = y^n(t_{n+1}) + \gamma^{n+1}$.

В § 4 мы рассматривали уравнение (4.13), решение которого могло обращаться в бесконечность, если $p(x) < 0$. Тем не менее при вычислении значений этого решения с помощью расчетных формул (4.9) оказывалось, что, согласно оценке (4.19), погрешность приближенного решения после прохождения таких точек мала. Расчетные формулы (4.9) представляют интересный пример схем, удовлетворяющихся с высокой точностью на особенностях решения типа $(x - a)^{-1}$.

Задача. Посмотреть, как работает схема (4.9) при приближении к особенности решения в случае $p(x) \equiv \text{const} > 0$, когда уравнение (4.13) интегрируется в явном виде. На каких решениях она точна?

§ 10. Конечно-разностные методы отыскания собственных значений

Рассмотрим простейшую задачу на собственные значения

$$y'' - p(x)y = \lambda p(x)y, \quad y(0) = 0, \quad y(X) = 0. \quad (1)$$

Зададимся шагом $h = XN^{-1}$ и напомним сеточную задачу

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - p_n y_n = \lambda p_n y_n, \quad (2)$$

$$n = 1, \dots, N-1, \quad y_0 = y_N = 0; \quad p_n = p(x_n), \quad \rho_n = \rho(x_n).$$

Значения λ , при которых система уравнений (2) имеет ненулевое решение $y_0, \dots, y_N$, естественно назвать собственными значениями сеточной задачи. Пусть собственные значения задач (1), (2) упорядочены в порядке убывания

$$\lambda_1 \geq \lambda_2 \geq \dots,$$

$$\lambda_1^h \geq \lambda_2^h \geq \dots$$

Рассмотрим модельный пример: $p(x) \equiv 0$, $\rho(x) \equiv 1$; тогда (1) приобретает вид

$$y'' = \lambda y, \quad y(0) = y(X) = 0.$$

Собственные функции этой задачи $w^m(x) = \sin\left(\frac{\pi m x}{X}\right)$ и соответствующие собственные значения $\lambda_m = -\left(\frac{\pi m}{X}\right)^2$.

В случае сеточной задачи (2), приобретающей вид

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - \lambda y_n = 0, \quad y_0 = y_N = 0,$$

рассуждаем следующим образом: общее решение разностного уравнения

$$y_{n+1} - (2 + \lambda h^2) y_n + y_{n-1} = 0$$

записывается в виде

$$y_n = C_1 \mu_1^n + C_2 \mu_2^n,$$

где μ_1, μ_2 — корни характеристического уравнения

$$\mu^2 - (2 + \lambda h^2) \mu + 1 = 0; \quad (3)$$

очевидно, $\mu_2 = \mu_1^{-1}$, и поэтому $y_n = C_1 \mu_1^n + C_1 \mu_1^{-n}$. Условия $y_0 = y_N = 0$ дают систему уравнений $C_1 + C_2 = 0$, $C_1 \mu_1^N + C_2 \mu_1^{-N} = 0$; она имеет ненулевое решение если ее определитель $\mu_1^{-N} - \mu_1^N = 0$; отсюда

$$\mu_1 = \exp\left(\frac{2\pi i m}{2N}\right), \quad m = \dots, -1, 0, 1, \dots$$

Из (3) можно выразить значения λ^h через значения μ_1

$$\begin{aligned} \lambda_m^h &= \frac{\mu_1 + \mu_1^{-1} - 2}{h^2} = \frac{\exp\left(\frac{\pi i m}{N}\right) + \exp\left(-\frac{\pi i m}{N}\right) - 2}{h^2} = \\ &= \frac{N^2}{X^2} \left(2 \cos \frac{\pi m}{N} - 2\right) = -4 \frac{N^2}{X^2} \sin^2 \frac{\pi m}{2N}. \end{aligned}$$

Для определенности возьмем $C_1 = (2i)^{-1}$; тогда соответствующие собственные функции имеют вид

$$W_n^m = \frac{1}{2i} \left(\exp\left(\frac{\pi i m n}{N}\right) - \exp\left(-\frac{\pi i m n}{N}\right) \right) = \sin \frac{\pi m n}{N}.$$

Собственные значения $\lambda_1^h, \dots, \lambda_{N-1}^h$ различны между собой; поэтому соответствующие им собственные функции $\sin \frac{\pi n}{N}, \dots, \sin \frac{\pi(N-1)n}{N}$ также различны. Так как задача (2) является задачей на собственные значения для матрицы размерности $N-1$, то мы получим полную систему собственных функций; каждая из функций W_n^m при $m \leq 0$ или при $m \geq N$ равна тождественно нулю или пропорциональна одной из перечисленных выше функций $W_n^1, \dots, W_n^{N-1}$. Мы случайно выбрали такой пример, где в узлах сетки $x_n = nh$ выполняется равенство

$$W_n^m = w^m(nh).$$

В общем случае это равенство не имеет места; однако характер близости собственных значений этих задач типичен и для

общего случая. Поскольку $\cos x = 1 - \frac{x^2}{2} + (\cos \theta x) \frac{x^4}{24}$, где $|\theta| \leq 1$, то

$$\lambda_m^h = -\frac{N^2}{X^2} \left(\frac{\pi^2 m^2}{N^2} - \frac{\cos \theta_m}{12} \left(\frac{\pi m}{N} \right)^4 \right),$$

или

$$\lambda_m^h = -\left(\frac{\pi m}{X} \right)^2 + \frac{\cos \theta_m}{12} \left(\frac{\pi m}{X} \right)^4 h^2. \quad (4)$$

Из этой формулы видно, что $\lambda_m^h - \lambda_m = O(h^2)$ при фиксированном m ; в то же время с ростом m как абсолютная, так и относительная погрешности монотонно возрастают и, например,

$$\frac{\lambda_{N-1}^h}{\lambda_{N-1}^h} = \frac{-\pi^2 (N-1)^2}{-4N^2 \sin^2 \frac{\pi(N-1)}{2N}} \sim \frac{\pi^2}{4}.$$

Равенство (4) можно записать в виде оценки $|\lambda_m^h - \lambda_m| \leq C \lambda_m^2 h^2$, где C не зависит от λ_m и h . В случае дважды дифференцируемых $p(x)$, $\rho(x)$ можно получить такую же оценку.

Укажем путь, на котором можно было бы получить оценку типа

$$|\lambda_m^h - \lambda_m| \leq C_0(\lambda_m) h^2. \quad (5)$$

Обозначим через $w^\lambda(x)$ решение задачи Коши

$$w'' - p(x)w = \lambda \rho(x)w, \quad w(0) = 0, \quad w'(0) = 1;$$

через w_n^λ обозначаем решение сеточной задачи Коши $w_0^\lambda = 0$, $w_1^\lambda = h$,

$$\frac{w_{n+1}^\lambda - 2w_n^\lambda + w_{n-1}^\lambda}{h^2} - p_n w_n^\lambda = \lambda \rho_n w_n^\lambda.$$

Очевидно, нули функции $g(\lambda) = w^\lambda(X)$ являются собственными значениями дифференциальной задачи, а нули функции $g_h(\lambda) = w_N^\lambda$ — собственными значениями сеточной задачи. Вследствие оценок погрешности решения такой задачи Коши, приведенных в § 2, справедливо неравенство

$$|g_h(\lambda) - g(\lambda)| \leq C_1(\lambda) h^2.$$

Стандартной техникой можно показать, что $g'(\lambda) \neq 0$, если $g(\lambda) = 0$; поэтому функция $g(\lambda)$ меняет знак в окрестности каждого своего нуля.

Из рис. 9.10.1 видно, что функция $g_h(\lambda)$ также меняет знак вблизи нулей функции $g(\lambda)$ и не может обращаться в нуль вдали от ее нулей; поэтому нули функций $g_h(\lambda)$ и $g(\lambda)$ близки между собой. Количественная оценка (5) близости нулей этих функций получается следующим образом. Пусть λ^h и λ — близкие корни функций $g_h(\lambda)$ и $g(\lambda)$. Согласно последнему неравенству $|g(\lambda^h)| \leq C_1(\lambda^h) h^2$; это неравенство можно представить в форме

$$|g(\lambda^h) - g(\lambda)| \leq C_1(\lambda^h) h^2$$

или

$$|g'(\bar{\lambda})(\lambda^h - \lambda)| \leq C_1(\lambda^h) h^2;$$

отсюда

$$|\lambda^h - \lambda| \leq \frac{C_1 (\lambda^h)}{|g'(\bar{\lambda})|} h^2.$$

Далее потребуется оценить снизу $|g'(\lambda)|$ в области, где $|g(\lambda)| \leq C_1 (\lambda^h) h^2$. Кроме этого, потребуется доказать, что вблизи каждого нуля $g(\lambda)$ имеется только один корень функции $g_h(\lambda)$.

Для решения задачи (1) с более высокой точностью можно воспользоваться любыми разностными аппроксимациями уравнения $y'' - q(x)y = 0$ более высокой точности. Рассмотрим пример.

В § 1 была построена разностная схема, аппроксимирующая последнее уравнение с погрешностью $O(h^4)$:

$$\frac{\delta^2 y_n}{h^2} - q_n y_n - \frac{1}{12} \delta^2 (q_n y_n) = 0 \quad (6)$$

(с несколько иными обозначениями). Уравнение (1) записывается в рассматриваемом виде при $q(x) = p(x) + \lambda \rho(x)$. Отсюда по-

лучаем сеточную задачу на собственные значения:

$$\frac{\delta^2 y_n}{h^2} - (p_n + \lambda \rho_n) y_n - \frac{1}{12} \delta^2 ((p_n + \lambda \rho_n) y_n) = 0, \quad (7)$$

$$n = 1, \dots, N-1, \quad y_0 = y_N = 0.$$

Можно показать, что для собственных значений этой задачи выполняется оценка

$$\lambda_m^h - \lambda_m = O(\lambda_m^3 h^4).$$

Система уравнений (7) имеет вид $Ay - \lambda By = 0$, формально несколько более сложный, чем (2), поскольку матрица B не диагональная. При повышении порядка точности могут возникать сеточные задачи, имеющие на первый взгляд еще более сложный вид, например такая: требуется найти λ , при котором система соотношений

$$\alpha_n(\lambda, h) y_{n-1} - \beta_n(\lambda, h) y_n + \gamma_n(\lambda, h) y_{n+1} = 0, \quad (8)$$

$$n = 1, \dots, N-1, \quad y_0 = y_N = 0,$$

имеет ненулевое решение y_n .

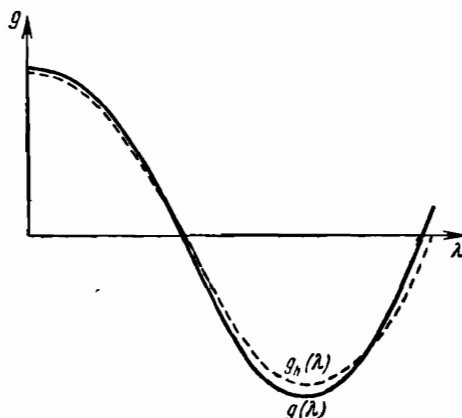


Рис. 9.10.1.

Рассмотрим случай $\gamma_n \neq 0$. Фиксируем некоторое λ , $w_1 \neq 0$, например $w_1 = h$, и из соотношения

$$\alpha_n(\lambda, h) w_{n-1}^\lambda - \beta_n(\lambda, h) w_n^\lambda + \gamma_n(\lambda, h) w_{n+1}^\lambda = 0 \quad (9)$$

последовательно определим $w_2^\lambda, \dots, w_N^\lambda$. Если $w_N^\lambda = 0$, то это λ окажется собственным значением и w_n^λ — собственной функцией; если $w_N^\lambda \neq 0$, то это λ не является собственным значением задачи (8).

Для отыскания собственных значений задачи (8), совпадающих с нулями w_N^λ , можно применить какой-либо итерационный метод отыскания нулей функции по ее значениям. Этот процесс облегчается следующим обстоятельством, имеющим место во многих случаях и позволяющим получить «вилку» для искомого корня: если функция w_n^λ имеет j перемен знака на $(0, X)$, то $\lambda_j < \lambda < \lambda_{j+1}$.

Для вычисления значений w_N^λ при различных λ , как правило, наиболее рационально воспользоваться непосредственно рекуррентными формулами (9).

Предлагаемый алгоритм вычисления значений w_n^λ совпадает с алгоритмом решения сеточной задачи Коши для (1). Если решения дифференциальной задачи Коши сильно растут, то, как и ранее, могут возникать большие вычислительные погрешности.

Для отыскания собственных значений может применяться также и метод прогонки. Не повторяя идеи метода, ограничимся написанием расчетных формул. Положим $\frac{w_n^\lambda}{w_{n+1}^\lambda} = C_n$, тогда (9) перепишется в виде $\alpha_n C_{n-1} C_n - \beta_n C_n + \gamma_n = 0$, откуда

$$C_n = \frac{\gamma_n}{\beta_n - \alpha_n C_{n-1}}. \quad (10)$$

Если w_n^λ и w_{n+1}^λ одного знака, то $C_n > 0$, если разного, то $C_n < 0$. Поэтому, наблюдая за переменами знака у C_n , можно определить число перемен знака у функции w_n^λ . Как мы видели в § 4, коэффициент C_n может оказываться очень большим, поэтому этот метод чаще применяется лишь для отыскания первого собственного значения.

Рассмотрим еще один способ сведения краевой задачи на собственные значения к нахождению нулей некоторой функции. Пусть рассматривается задача

$$y' - A(x)y - \lambda C(x)y = 0, \quad By(0) = 0, \quad Dy(X) = 0. \quad (11)$$

Задавая λ и осуществляя ортогональную прогонку, можно последовательно определить матрицы $G_s(X_s)$, $s = 1, \dots, m$. Если система уравнений (см. § 6)

$$D(G_m(X_m))\beta^m = 0$$

имеет ненулевое решение, то исходная задача (11) также имеет ненулевое решение и соответствующее значение λ будет собственным. Таким образом, задача отыскания собственного значения равносильна здесь задаче нахождения корня уравнения

$$\det(DG_m(X_m)) = 0.$$

§ 11. Оптимизация распределения узлов интегрирования

Решения дифференциальных уравнений и систем могут иметь различную гладкость на различных участках отрезка интегрирования. На примере оценки погрешности метода Рунге — Кутты было видно, что вклад от погрешности интегрирования на некотором шаге $[x_j, x_{j+1}]$ в суммарную погрешность пропор-

ционален в точке $x_N = x_0 + X$ множителю $\exp\left(\int_{x_j}^{x_N} f_y(x, y(x)) dx\right)$.

зависящему от j . Поэтому для некоторых классов дифференциальных уравнений становится актуальной задача оптимизации распределения узлов интегрирования. Для простоты предполагаем, что начальное условие задано точно и округления отсутствуют. Погрешность результата численного интегрирования по методу Рунге — Кутты в точке x_N в соответствии с (8.4.6) не превосходит

$$S_N = \sum_{j=1}^N |y_j - y_{j-1}(x_j)| \exp\left(\int_{x_j}^{x_N} f_y(x, \tilde{y}_j(x)) dx\right). \quad (1)$$

Предположим, что $x_j = \varphi\left(\frac{j}{N}\right)$, $\varphi(0) = x_0$, $\varphi(1) = x_0 + X$, $\varphi(t)$ — гладкая функция. Согласно формуле Лагранжа

$$x_j - x_{j-1} = \varphi\left(\frac{j}{N}\right) - \varphi\left(\frac{j-1}{N}\right) = \frac{1}{N} \varphi'(\tilde{t}_j),$$

где $\frac{j-1}{N} \leq \tilde{t}_j \leq \frac{j}{N}$, и поэтому

$$H = \max_{0 < j \leq N} (x_j - x_{j-1}) \leq \frac{1}{N} \max_{[0, 1]} |\varphi'(t)|, \quad H \rightarrow 0 \quad \text{при} \quad N \rightarrow \infty;$$

поэтому вследствие оценки из § 4 гл. VIII имеем

$$\max_{x_0 \leq x_j \leq x \leq x_0 + X} |y_j(x) - y(x)| \rightarrow 0$$

при $N \rightarrow \infty$. Следовательно, можно написать асимптотические равенства

$$\exp \left(\int_{x_j}^{x_N} f_y(x, \tilde{y}_j(x)) dx \right) \sim \exp \left(\int_{x_j}^{x_N} f_y(x, y(x)) dx \right),$$

$$\begin{aligned} y_j - y_{j-1}(x_j) &\sim \psi(x_j, y(x_j)) (x_j - x_{j-1})^{k+1} \sim \\ &\sim \psi \left(\frac{j}{N}, y \left(\frac{j}{N} \right) \right) \left(\frac{1}{N} \Phi' \left(\frac{j}{N} \right) \right)^{k+1}. \end{aligned}$$

Тогда выражение (1) для S_N представится в виде

$$S_N \sim \frac{1}{N^k} \sum_{j=1}^N \frac{1}{N} \Phi \left(\frac{j}{N} \right), \quad (2)$$

где

$$\Phi(t) = |\psi(\Phi(t), y(\Phi(t)))| (\Phi'(t))^{k+1} \exp \left(\int_{\Phi(t)}^{x_0+X} f_y(x, y(x)) dx \right). \quad (3)$$

При $\Phi(t)$ гладкой и $N \rightarrow \infty$ величина $\sum_{j=1}^N \frac{1}{N} \Phi \left(\frac{j}{N} \right)$ стремится к интегралу

$$I = \int_0^1 \Phi(t) dt \quad (4)$$

и, таким образом,

$$S_N \sim N^{-k} \int_0^1 \Phi(t) dt.$$

Примем φ за новую переменную в интеграле I . Тогда он запишется в виде

$$I = \int_{x_0}^{x_0+X} L(\varphi) d\varphi,$$

где

$$L(\varphi) = |\psi(\varphi, y(\varphi))| (t'(\varphi))^{-k} \exp \left(\int_{\varphi}^{x_0+X} f_y(x, y(x)) dx \right). \quad (5)$$

Задача минимизации этого интеграла в форме (4) за счет выбора $\varphi(t)$ и в форме (5) за счет выбора обратной функции $t(\varphi)$

эквивалентны. Вследствие равенства $\frac{\partial L}{\partial t} = 0$ уравнение Эйлера

$$\frac{d}{d\varphi} \left(\frac{\partial L}{\partial t'} \right) - \frac{\partial L}{\partial t} = 0$$

в данном случае приобретает вид $\frac{d}{d\varphi} \left(\frac{\partial L}{\partial t'} \right) = 0$; отсюда следует, что

$$\frac{\partial L}{\partial t'} = -k |\psi(\varphi, y(\varphi))| (t'(\varphi))^{-k-1} \exp \left(\int_{\varphi}^{x_0+X} f_y(x, y(x)) dx \right) = \text{const.}$$

Возвращаясь к переменной φ , получаем

$$(\varphi'(t))^{k+1} |\psi(\varphi, y(\varphi))| \exp \left(\int_{\varphi}^{x_0+X} f_y(x, y(x)) dx \right) = C_1. \quad (6)$$

Решение этого дифференциального уравнения зависит от C_1 и еще некоторой постоянной C_2 ; их значения следует определить из граничных условий $\varphi(0) = x_0$, $\varphi(1) = x_0 + X$. По поводу обоснования этих построений в случае, когда величины $\psi(\varphi, y(\varphi))$ или

$$\exp \left(\int_{\varphi}^{x_0+X} f_y(x, y(x)) dx \right)$$

не ограничены или негладкие, можно привести соображения, аналогичные случаю задачи численного интегрирования. В той же мере к этой задаче относятся и высказанные ранее соображения о практическом использовании соотношения (6).

Отметим одно неочевидное обстоятельство. Уравнение (6) может быть записано в виде

$$(\varphi'(t))^{k+1} |\psi(\varphi, y(\varphi))| \exp \left(- \int_0^{\varphi} f_y(x, y(x)) dx \right) = C, \quad (7)$$

куда не входят ни начальная, ни конечная точки интегрирования. Вместе с функцией $\varphi(t)$ этому уравнению удовлетворяет также функция $\varphi(at + b)$. Задавшись некоторым N , осуществим одновременное интегрирование исходного уравнения и уравнения (7), выбирая каждый раз шаги интегрирования из условия $x_i - x_{i-1} \sim \frac{1}{N} \varphi'(t_{i-1})$. Скорее всего мы придем в конечную точку $x_0 + X$ с числом узлов N_1 , отличным от N , и тогда это распределение узлов не будет являться оптимальным распределением узлов, соответствующим данному N . Это естественно, поскольку, начиная вычисления, мы не могли заранее предвидеть хода поведения решения и сказать, какую величину C следует взять в правой части (7). Однако можно показать, что вследствие указанных свойств решения уравнения (7) функция $\varphi(t)$ будет искомой и распределение узлов $\varphi(j/N_1)$ оптимальным (с точностью до погрешности численного интегрирования), соответствующим числу узлов N_1 .

Непосредственное интегрирование уравнений (1) и (7) встречает затруднение вследствие необходимости вычисления значений функции $\psi(x, y(x))$. Вместо непосредственного вычисления значений этой функции целесообразно использовать величину контрольного члена точности на шаге. При численном интегрировании уравнения (7) следует также иметь в виду, что эти соотношения носят асимптотический характер. В окрестности точек, где $\psi(x, y(x)) = 0$, в остаточном члене начинают играть существенную роль слабые порядки $(x_j - x_{j-1})^{k+2}$.

Дополнительное численное интегрирование уравнения (7) может сильно усложнить решение задачи. Поэтому к вопросу оптимизации распределения узлов часто подходят следующим образом. Пусть решаются задачи из некоторого определенного класса. Рассмотрим модельную для этого класса задачу, где можно в явном виде решить уравнение (7). Постараемся на ее примере установить зависимость шага (или меры погрешности на шаге (см. § 3 гл. VIII)) от поведения решения, при которой распределение узлов близко к оптимальному. Далее все задачи этого класса интегрируем с шагом, соответствующим этой зависимости.

В других случаях заранее задаются некоторой формой такой зависимости. Пусть система уравнений порядка m интегрируется с контролем точности на шаге, например, по формулам (8.3.2). В случае системы уравнений контрольный член r будет некоторым вектором $r_n = (r_n^1, \dots, r_n^m)$. В ряде программ шаг интегрирования выбирается из условия

$$\|r_n\| / (\max(M, \|y_n\|)) \approx \varepsilon = \text{const},$$

или из дающего большую гибкость условия

$$\max_{1 \leq k \leq m} \frac{|r_n^k|}{\max(|M_k|, |(y_n)_k|)} \approx \varepsilon = \text{const}.$$

Параметры M, M_k подбирают из соображений оптимальности распределения узлов в условиях конкретной задачи.

Рассмотрим примеры оптимизации распределения узлов интегрирования.

Пусть задача Коши $y' = My$ при начальном условии $y(0)$ решается методом Эйлера. Тогда

$$\psi(x, y(x)) = -\frac{1}{2} y''(x) = -\frac{M^2}{2} y(0) \exp(Mx)$$

и уравнение (6) имеет вид

$$(\varphi')^2 \frac{M^2}{2} |y(0)| \exp(M\varphi) \exp(M(X - \varphi)) = \text{const}.$$

Отсюда получаем $\varphi'_i = \text{const}$, т. е. распределение узлов следует взять равномерным.

Задача. Пусть мера погрешности на шаге ρ вычисляется, как и в § 3 гл. VIII, как разность просчета с двойным шагом и двух просчетов с одним шагом. Проверить, что в случае данного примера равномерное, т. е. оптимальное, распределение узлов, реализуется при условии

$$\left| \frac{r_n}{y_n} \right| \approx \varepsilon = \text{const}$$

Обратимся к краевой задаче $\mu^2 y'' - y = f(x)$ на отрезке $[0, 1]$ при заданных $y(0)$, $y(1)$ и малом μ . Представляет интерес изучение методов решения этой задачи как модели задач с так называемым *пограничным слоем*. Решением этой задачи при $f(x) \equiv 0$ является

$$y(x) = \frac{y(0) \left(\exp\left(-\frac{x}{\mu}\right) - \exp\left(\frac{x-2}{\mu}\right) \right) + y(1) \left(\exp\left(\frac{x-1}{\mu}\right) - \exp\left(-\frac{x+1}{\mu}\right) \right)}{1 - \exp\left(-\frac{2}{\mu}\right)}.$$

Непосредственным дифференцированием убеждаемся, что

$$|y^{(k)}(x)| = O\left(\frac{1}{\mu^k} y(0) \exp\left(-\frac{x}{\mu}\right)\right) + O\left(\frac{1}{\mu^k} y(1) \exp\left(\frac{x-1}{\mu}\right)\right).$$

При $f(x) \neq 0$ вместо этой оценки имеет место оценка

$$|y^{(k)}(x)| = O\left(\frac{1}{\mu^k} \exp\left(-\frac{x}{\mu}\right)\right) + O\left(\frac{1}{\mu^k} \exp\left(\frac{x-1}{\mu}\right)\right) + O(1). \quad (8)$$

Таким образом, дифференциальные свойства решения резко ухудшаются при малых μ вблизи точек $x=0$ и $x=1$, и там целесообразно применить более мелкий шаг интегрирования. Зададимся узлами интегрирования $x_j = \Phi\left(\frac{j}{N}\right)$, $\Phi(0)=0$, $\Phi(1)=1$ и заменим производную $y''(x_j)$ удвоенной разделенной разностью $y''(x_j) \approx 2y(x_{j-1}; x_j; x_{j+1})$. Получим систему разностных уравнений

$$l(y_j) = 2\mu^2 \frac{\frac{y_{j+1} - y_j}{x_{j+1} - x_j} - \frac{y_j - y_{j-1}}{x_j - x_{j-1}}}{x_{j+1} - x_{j-1}} - y_j = f_j, \quad (9)$$

$$j = 1, \dots, N-1, \quad y_0 = y(0), \quad y_N = y(1).$$

Согласно выражению разделенной разности через производную имеем

$$2y(x_{j-1}; x_j; x_{j+1}) = y''(\xi_j),$$

где $x_{j-1} \leq \xi_j \leq x_{j+1}$. Поэтому

$$r_n = l(y(x_n)) - f_n = \mu^2 (y''(\xi_n) - y''(x_n)) = \mu^2 (\xi_n - x_n) y^{(3)}(\eta_n). \quad (10)$$

Отсюда следует неравенство

$$|r_n| \leq \sigma_n = \mu^2 (\max(|x_{n+1} - x_n|, |x_n - x_{n-1}|)) \left(\max_{[x_{n-1}, x_{n+1}]} |y^{(3)}(x)| \right). \quad (11)$$

Из (9) и (10) получаем уравнение для ошибки приближенного решения $R_n = y_n - y(x_n)$

$$l(R_j) = -r_j, \quad j = 1, \dots, N-1, \quad R_0 = R_N = 0. \quad (12)$$

Далее следует получить асимптотическое представление погрешности и поставить и решить соответствующую асимптотическую вариационную задачу.

Если нам требуется минимизировать максимум погрешности на всем отрезке, а не в отдельных точках, то такой путь приведет к громоздким вычислениям. Поэтому мы заменим оптимизационную задачу более простой.

Предоставим читателю самому убедиться в справедливости неравенств

$$\max_n |R_n| \leq \max_n |r_n|.$$

Отсюда следует неравенство

$$\max_n |R_n| \leq \max_n |\sigma_n|.$$

Поставим задачу минимизации правой части этого неравенства. При $N \rightarrow \infty$ имеем

$$\sigma_n \sim \frac{\mu^2}{N} \varphi'(n/N) \left| y''' \left(\varphi \left(\frac{n}{N} \right) \right) \right|.$$

Поэтому асимптотически рассматриваемая задача сводится к минимизации выражения

$$z(\varphi) = \max_{[0, 1]} \varphi'(t) |y^{(3)}(\varphi(t))|$$

при условиях $\varphi(0) = 0$ и $\varphi(1) = 1$.

Покажем, что нижняя грань $z(\varphi)$, равная

$$z_0 = \left(\int_0^1 \frac{dt}{|y^{(3)}(\varphi(t))|} \right)^{-1}$$

достигается при $\varphi'(t) = (|y^{(3)}(\varphi(t))| z_0)^{-1}$. Действительно, предположим, что для некоторой функции $\varphi(t)$ при всех t справедливо $\varphi'(t) |y^{(3)}(\varphi(t))| < z_0$; из этого неравенства следует, что

$$\varphi'(t) < |y^{(3)}(\varphi(t))|^{-1} z_0.$$

Интегрируя это неравенство от 0 до 1, получаем

$$1 = \varphi(1) - \varphi(0) < \int_0^1 |y^{(3)}(\varphi(t))|^{-1} z_0 dt = 1.$$

Мы пришли к противоречию.

При таком $\varphi'(t)$ имеем

$$\mu \varphi'(t) |y^{(3)}(\varphi(t))| = \mu z_0^{-1} = \text{const.} \quad (13)$$

В рассматриваемом случае, согласно (8), справедливо

$$|y^{(3)}(t)| \leq c \left(\frac{1}{\mu^3} \left(\exp\left(-\frac{x}{\mu}\right) + \exp\left(\frac{x-1}{\mu}\right) \right) + 1 \right).$$

Поэтому вместо (13) рассмотрим уравнение

$$\varphi'(t) \left(\frac{1}{\mu^3} \left(\exp\left(-\frac{\varphi}{\mu}\right) + \exp\left(\frac{\varphi-1}{\mu}\right) \right) + 1 \right) = \text{const.}$$

Это уравнение и граничные условия $\varphi(0) = 0$, $\varphi(1) = 1$ однозначно определяют функцию $\varphi(t)$.

Отметим, что в действительности рассматриваемая схема имеет второй порядок аппроксимации и при реальных вычислениях распределение узлов следует выбирать с учетом этого обстоятельства.

§ 12. Влияние вычислительной погрешности в зависимости от формы записи конечно-разностного уравнения

Как было установлено ранее, вычислительная погрешность имеет различный характер роста для различных способов решения дифференциальных уравнений. Рассмотрим теперь, такой частный, но важный вопрос: как зависит вычислительная погрешность от формы записи конечно-разностных уравнений? Хотя все изложение ведется на примере задачи Коши, рассматриваемые соображения относятся в равной мере и к случаю решения краевых задач. Для примера обратимся к методу Эйлера

$$y_{n+1} = y_n + hf(x_n, y_n). \quad (1)$$

При реальных вычислениях будут получаться величины y_n^* , связанные соотношением

$$y_{n+1}^* = y_n^* + hf(x_n, y_n^*) + \delta_n; \quad (2)$$

наличие слагаемого δ_n , как уже упоминалось, является следствием ряда причин — ошибок при вычислении значений функции $f(x_n, y_n^*)$, ошибки при округлении произведения $hf(x_n, y_n^*)$ и ошибки при сложении чисел y_n^* и округленного значения $hf(x_n, y_n^*)$.

Введем обозначение $y_n^* - y_n = \Delta_n$; на основании формулы Лагранжа имеем

$$f(x_n, y_n^*) - f(x_n, y_n) = l_n \Delta_n,$$

где $l_n = f_y(x_n, \bar{y}_n)$. Предположим, что всегда $|f_y(x, y)| \leq L$. Вычитая (1) из (2), получаем

$$\Delta_{n+1} = (1 + l_n h) \Delta_n + \delta_n,$$

откуда

$$|\Delta_{n+1}| \leq (1 + Lh) |\Delta_n| + \delta, \quad \delta = \max_n \delta_n. \quad (3)$$

Рассмотрим разностное уравнение

$$z_{n+1} = (1 + Lh) z_n + \delta,$$

называемое *мажорирующим* для (3). Его решение при начальном условии $z_0 = |\Delta_0|$ есть

$$z_n^0 = |\Delta_0| (1 + Lh)^n + \frac{\delta (1 + Lh)^n - \delta}{Lh}.$$

Лемма. При всех $n \geq 0$ справедливо

$$|\Delta_n| \leq z_n^0. \quad (4)$$

Доказательство. При $n=0$ утверждение (4) очевидно. Пусть оно верно для некоторого n ; тогда имеем

$$|\Delta_{n+1}| \leq (1 + Lh) z_n^0 + \delta = z_{n+1}^0.$$

Лемма доказана.

Если интегрирование производится на отрезке $[x_0, x_0 + X]$, то

$$nh \leq X, \quad (1 + Lh)^n \leq (\exp(Lh))^n \leq \exp(LX).$$

Согласно формуле Лагранжа

$$x_2^N - x_1^N = N \bar{x}^{N-1} (x_2 - x_1),$$

и поэтому

$$(1 + Lh)^n - 1 = n(1 + \theta Lh)^{n-1} Lh, \quad \text{где } 0 \leq \theta \leq 1.$$

Отсюда получаем

$$(1 + Lh)^n - 1 \leq nLh (\exp(Lh))^{n-1} \leq LX \exp(LX).$$

В итоге имеем оценку

$$|\Delta_n| \leq |z_n^0| \leq |\Delta_0| \exp(LX) + \delta X h^{-1} \exp(LX).$$

Рассмотрим случай $\Delta_0 = 0$. Тогда погрешность $\Delta_n = y_n^* - y_n$ при X фиксированном оценивается сверху через $O(\delta h^{-1})$.

Эта оценка не улучшаема по порядку. Например, при $\Delta_0 = 0$, $f_y \equiv 0$, $\delta_n \equiv \delta$ имеем $\Delta_{n+1} = \Delta_n + \delta$ и, таким образом,

$$\Delta_N = N\delta = \delta X h^{-1}.$$

Вместо проведения этих аккуратных оценок можно привести правдоподобные рассуждения, из которых также следует, что погрешность округления может оказаться величиной порядка δh^{-1} . Соотношение

$$y_{n+1}^* = y_n^* + hf(x_n, y_n^*) + \delta_n$$

можно переписать в виде

$$y_{n+1}^* = y_n^* + hf^*(x_n, y_n^*),$$

где

$$f^*(x_n, y_n^*) = f(x_n, y_n^*) + \delta_n h^{-1}.$$

Таким образом, результат численного интегрирования уравнения $y' = f(x, y)$ при наличии округлений будет такой же, как если бы без округлений интегрировалось уравнение со значениями правой части в узлах сетки $f^*(x_n, y_n^*)$. Рассмотрим случай $\delta_n \equiv \delta$. Разность между решениями дифференциальных уравнений

$$y' = f(x, y) \quad \text{и} \quad y' = f(x, y) + \delta h^{-1}$$

имеет порядок разности между правыми частями этих уравнений, т. е. δh^{-1} . Нет оснований ожидать, что решения разностных уравнений

$$y_{n+1} = y_n + hf(x_n, y_n) \text{ и } \dot{y}_{n+1} = \dot{y}_n + h(\dot{f}(x_n, y_n) + \delta h^{-1})$$

будут отличаться менее чем на величину того же порядка, т. е. порядка δh^{-1} . Если δ — порядка 2^{-t} , t — разрядность чисел в ЭВМ, то решение разностного уравнения изменится на величину порядка $2^{-t}h^{-1}$.

При получении этого вывода сделано допущение $\delta_n \equiv \delta$. Рассмотрим задачу вычисления интеграла $\int_0^1 \tilde{f}(x) dx$, являющуюся частным случаем рассматриваемой задачи при $y(0) = 0$. Пусть $f(x) = 2/3$, $h = 2^{-k}$, $t - k$ — нечетно. При $x_n > 3/4$ y_n лежит в пределах $(1/2, 1)$. Тогда после округления при сложении

$$\begin{array}{r} + y_n = 0, 1 \alpha_2 \dots \alpha_t \\ hf_n = 0, 0 \dots 0 \underbrace{1010 \dots 1010101}_{k} \dots \end{array}$$

каждый раз происходит отбрасывание величины

$$2^{-t} \cdot 0, 0101 \dots = (1/3) 2^{-t}.$$

Таким образом, на этом участке действительно $\delta_n \equiv \delta$ порядка 2^{-t} .

Перейдем к случаю интегрирования уравнения $y'' = f(x, y)$ при помощи простейшего метода Штермера

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} = f(x_n, y_n)$$

по расчетной формуле

$$y_{n+1} = 2y_n - y_{n-1} + h^2 f(x_n, y_n);$$

реально получаемые значения y_n^* связаны некоторым соотношением

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + h^2 f(x_n, y_n^*) + \delta_n. \quad (5)$$

Значения y_n^* можно рассматривать как получаемые без округлений при вычислениях по формуле

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + h^2 f^*(x_n, y_n^*),$$

где

$$f^*(x_n, y_n^*) = f(x_n, y_n^*) + \delta_n h^{-2}.$$

Рассмотрим случай $\delta_n \equiv \delta$. Решения дифференциальных уравнений

$$y'' = f(x, y) \quad \text{и} \quad y'' = f(x, y) + \delta h^{-2}$$

отличаются между собой на величину порядка δh^{-2} . Можно показать, что в случае уравнения $y'' = \alpha$, $\alpha = \text{const}$, возможен случай $\delta_n \equiv \delta$, δ порядка 2^{-t} . Так же как для уравнения пер-

вого порядка, делаем вывод, что для рассматриваемого алгоритма суммарная вычислительная погрешность может оказаться величиной порядка $2^{-t}h^{-2}$.

Рассмотрим случай, когда такая величина вычислительной погрешности оказывается недопустимо большой. Можно было бы записать рассматриваемое уравнение в виде системы уравнений первого порядка

$$y' = v, \quad v' = f(x, y) \quad (6)$$

и применить какой-либо метод численного интегрирования этой системы. Как уже отмечалось, при этом произошла бы потеря эффективности вследствие того, что методы общего вида не учитывают специфики этой системы.

Попытаемся записать рассматриваемую расчетную формулу как некоторую расчетную формулу интегрирования системы (6). Введем новую дискретную переменную

$$\frac{y_n - y_{n-1}}{h} = z_n;$$

тогда уравнение (5) запишется в виде

$$\frac{z_{n+1} - z_n}{h} = f(x_n, y_n);$$

вычисления последовательных значений y_n, z_n будем производить при помощи пары расчетных формул

$$z_{n+1} = z_n + hf(x_n, y_n), \quad y_{n+1} = y_n + hz_{n+1}. \quad (7)$$

При наличии округлений, соответственно, имеем

$$z_{n+1}^* = z_n^* + hf(x_n, y_n^*) + \alpha_n, \quad y_{n+1}^* = y_n^* + hz_{n+1}^* + \beta_n,$$

где $\alpha_n, \beta_n = O(2^{-t})$. Эти соотношения можно представить в виде

$$\begin{aligned} z_{n+1}^* &= z_n^* + hf^*(x_n, y_n^*), \quad f^*(x_n, y_n^*) = f(x_n, y_n^*) + \alpha_n h^{-1}, \\ y_{n+1}^* &= y_n^* + h(z_{n+1}^* + g(x_{n+1})), \quad g(x_{n+1}) = \beta_n h^{-1}. \end{aligned} \quad (8)$$

Если формулы (7) можно трактовать как формулы численного интегрирования системы $z' = f(x, y), y' = z$, то формулы (8) соответствуют системе

$$z' = f^*(x, y), \quad y' = z + g(x).$$

Правые части этих систем отличаются на величины порядка $O(2^{-t}h^{-1})$; поэтому есть какие-то основания ожидать, что и решения разностных задач, т. е. решение разностной задачи с округлениями и разностной задачи без округлений будут отличаться на величину того же порядка. Конечно, к этому

заявлению следует отнестись с осторожностью; мы уже видели, что для некоторых конечно-разностных схем малые ошибки могут приводить к катастрофическому изменению результата.

Приведенные рассуждения о влиянии вычислительной погрешности в конкретных методах интегрирования уравнений первого и второго порядка опираются лишь на учет свойств конечно-разностной схемы, связанных с порядком дифференциального уравнения. Поэтому есть основания ожидать, что они переносятся на другие конечно-разностные методы. Например, при интегрировании уравнения $y^{(k)} = f(x, y)$ и прямом использовании схемы

$$\nabla^k y_n - h^k \sum_{i=0}^l a_i f(x_{n-i}, y_{n-i}) = 0$$

по-видимому, можно ожидать влияния вычислительной погрешности порядка $2^{-l} h^{-k}$. Поэтому в случае уравнений высокого порядка еще более актуальна задача преобразования схемы к форме, где влияние вычислительной погрешности будет меньше. Например, как и при $k = 2$, можно ввести вспомогательные переменные

$$z_n^i = \nabla^i y_n h^{-i}, \quad i = 1, \dots, k-1.$$

Попытаемся объяснить улучшение свойств разностной схемы (5) при переходе к расчетным формулам (7) с позиций оценки количества хранимой информации. При использовании расчетной формулы (5) при каждом n в памяти хранятся величины y_{n-1} и y_n и все дальнейшие значения y_j определяются по этим значениям. Пусть для определенности $1/2 \leq y_{n-1}$, $y_n \leq 1$ и $|y_n - y_{n-1}| \leq Mh$. Тогда в ячейке, содержащей значение $y_{n-1} = 0, 1\alpha_2 \dots \alpha_l$ имеется $l-1$ независимых двоичных знаков $\alpha_2, \dots, \alpha_l$; разряды числа $y_n = 0, 1\beta_2 \dots \beta_l$ уже не все несут новую информацию. Разность $y_n - y_{n-1} = \pm 0, 0\gamma_2 \dots \gamma_l$ по модулю меньше 2^{-l} , где l — наибольшее целое, такое, что $Mh < 2^{-l}$. Поэтому

$$y_n - y_{n-1} = \pm 0, 0 \dots 0\gamma_{l+1} \dots \gamma_l,$$

и для задания $y_n - y_{n-1}$, а следовательно, и y_n достаточен $l - l + 1$ двоичный знак (знак разности и $\gamma_{l+1}, \dots, \gamma_l$). Очевидно, $l \sim \log_2((Mh)^{-1})$ и, таким образом, общее количество независимой информации, которое имеется в распоряжении при каждом n , составляет $\sim 2l - \log_2((Mh)^{-1})$ двоичных разрядов.

В случае вычислений по формуле (7) все разряды чисел y_n и z_n независимы, и поэтому информация о решении задается независимыми двоичными разрядами.

Тот факт, что количество независимой информации для второго способа больше, конечно, сам по себе не означает, что этот способ лучше. Не исключено, что эта дополнительная информация не является содержательной и поэтому не позволяет точнее определить решение. Поясним, почему дополнительная информация при втором способе является содержательной.

Для задания решения дифференциального уравнения второго порядка с точностью $O(\epsilon)$ необходимо задание с такой же точностью значения решения и его производной в некоторой точке. Для разностного уравнения роль этих величин играют y_n и $\frac{y_n - y_{n-1}}{h}$. При первом способе задание значений y_n и y_{n-1} с t двоичными знаками позволяет определить значение $\frac{y_n - y_{n-1}}{h}$ с погрешностью порядка $O(2^{-t}h^{-1})$. Таким образом, здесь по известной нам информации мы располагаем возможностью найти дальнейшие значения решения сеточной задачи с погрешностью порядка $O(2^{-t}h^{-1})$ (если все последующие вычисления будут производиться абсолютно точно). В случае второго способа мы имеем значения y_n и $\frac{y_n - y_{n-1}}{h}$ с t двоичными знаками и поэтому обладаем возможностью найти решение сеточной задачи с погрешностью порядка $O(2^{-t})$. Таким образом, эта дополнительная информация действительно оказывается содержательной. На каждом шаге реального численного интегрирования ошибки округления вносят дополнительную неопределенность в значения вектора $(y_n, \frac{y_n - y_{n-1}}{h})$ во втором случае порядка $O(2^{-t})$, в первом — порядка $O(2^{-t}/h)$. Это и приводит к тому, что суммарная вычислительная погрешность решения во втором случае может оказаться величиной порядка $O(2^{-t}h^{-1})$, а в первом — порядка $O(2^{-t}h^{-2})$.

Обратим внимание на прием *практической оценки вычислительной погрешности путем изменения масштабов*. Пусть некоторым методом решается задача Коши

$$\frac{dy}{dx} = f(x, y), \quad y(0) = a.$$

Замена переменных $x = \mu t$, $y = \lambda z$ сводит эту задачу к задаче

$$\frac{dz}{dt} = \frac{\mu}{\lambda} f(\mu t, \lambda z), \quad z(0) = \frac{a}{\lambda}.$$

Предположим, что первая задача интегрировалась с шагами h_i ; осуществим численное интегрирование второй задачи тем же методом, но с шагами $h'_i = h_i/\mu$.

При отсутствии округлений будет иметь место равенство $y_i \equiv \lambda z_i$; если λ и μ оба не являются целыми степенями

двойки, то разность между реально получаемыми значениями величин y_i и λz_i , обычно дает определенное представление о величине вычислительной погрешности. Например, можно взять $\mu = \sqrt{3}$, $\lambda = \sqrt{2}$.

В случае интегрирования с автоматическим выбором шага надо обратить внимание на выбор нового значения параметра ε и начального шага h_0 с тем, чтобы равенство $h'_i = h_i/\mu$ имело место при всех i (см. подробнее § 17 гл. III).

§ 13. Оценка вычислительной погрешности при решении краевой задачи методом прогонки

В разделе численных методов алгебры было обещано привести пример оценки вычислительной погрешности при решении алгебраических систем методом Гаусса. Произведем такую оценку на примере решения системы алгебраических уравнений, возникающей при аппроксимации задачи (1.3), (1.4).

Рассматриваемая система уравнений имеет вид

$$y_{j-1} - (2 + p_j h^2) y_j + y_{j+1} = f_j h^2, \quad j = 1, \dots, N-1, \quad y_0 = a, \quad y_N = b, \quad (1)$$

При решении методом исключения Гаусса при последовательности исключения неизвестных $y_0, \dots, y_{N-1}$ вычисления производятся по формулам

$$C_{n+1} = (2 + p_n h^2 - C_n)^{-1}, \quad n = 0, \dots, N-2, \quad C_0 = 0, \quad (2)$$

$$\varphi_{n+1} = C_{n+1} (\varphi_n - f_{n+1} h^2), \quad n = 0, \dots, N-2, \quad \varphi_0 = a, \quad (3)$$

затем по формулам

$$y_n = C_n y_{n+1} + \varphi_n, \quad n = N-1, \dots, 1. \quad (4)$$

Всякое число z , получаемое в результате арифметической операции над двумя числами в ЭВМ, округляется затем до некоторого числа $z^* = z(1 + \varepsilon)$; при этом, как правило, выполняется неравенство $|\varepsilon| \leq 2^{-l}$, где l — разрядность чисел в машине. Будем предполагать, что последнее неравенство выполняется всегда; в настоящем параграфе греческими буквами обозначаем величины, по модулю не превосходящие некоторую абсолютную постоянную, умноженную на 2^{-l} ; полученное в результате реальных вычислений значение какой-либо величины z обозначаем z^* . Поскольку нас интересует вычислительная погрешность в процессе решения системы, будем для простоты предполагать, что коэффициенты системы $a, b, 2 + p_j h^2$ и $f_j h^2$ заданы без округлений. При наличии округлений соотношения (2), (3), (4) запишутся в виде

$$C_{n+1}^* = (1 : (((2 + p_{n+1} h^2) - C_n^*) (1 + \beta_n))) (1 + \alpha_n), \quad (5)$$

$$\varphi_{n+1}^* = (C_{n+1}^* ((\varphi_n^* - f_{n+1} h^2) (1 + \gamma_n))) (1 + \sigma_n), \quad (6)$$

$$y_n^* = ((C_n^* y_{n+1}^*) (1 + \omega_n) + \varphi_n^*) (1 + \delta_n). \quad (7)$$

Здесь после каждой скобки вида $(a \cdot b)$, где $\cdot$ — какая-либо арифметическая операция, стоит множитель вида $1 + \varepsilon$, соответствующий реализации этой операции при наличии округлений. Соотношение (5) можно переписать в

виде

$$C_{n+1}^* = (2 + p_{n+1}^* h^2 - C_n^*)^{-1},$$

где

$$p_{n+1}^* = p_{n+1} + g_{n+1}, \quad g_{n+1} = \frac{(2 + p_{n+1}^* h^2 - C_n^*) (\beta_n - \alpha_n)}{(1 + \alpha_n) h^2}. \quad (8)$$

Соотношение (7) можно переписать в виде

$$\frac{y_n^*}{1 + \delta_n} = C_n^* y_{n+1}^* + \Phi_n^* + \omega_n C_n^* y_{n+1}^*$$

или

$$y_n^* = C_n^* y_{n+1}^* + \Phi_n^{**};$$

здесь $\Phi_n^{**} = \Phi_n^* + d_n$, а в свою очередь

$$d_n = \omega_n C_n^* y_{n+1}^* + \frac{y_n^* \delta_n}{1 + \delta_n}.$$

Подставим $\Phi_n^* = \Phi_n^{**} - d_n$, $\Phi_{n+1}^* = \Phi_{n+1}^{**} - d_{n+1}$ в соотношение (6), представленное в виде

$$\frac{\Phi_{n+1}^*}{(1 + \gamma_n)(1 + \sigma_n)} = C_{n+1}^* (\Phi_n^* - f_{n+1} h^2);$$

получим

$$\frac{\Phi_{n+1}^{**} - d_{n+1}}{(1 + \gamma_n)(1 + \sigma_n)} = C_{n+1}^* (\Phi_n^{**} - d_n - f_{n+1} h^2).$$

Отсюда

$$\Phi_{n+1}^{**} = C_{n+1}^* (\Phi_n^{**} - f_{n+1} h^2),$$

$$f_{n+1} - f_{n+1} = q_{n+1} =$$

$$= \frac{1}{h^2} \left(d_n - (C_{n+1}^*)^{-1} \left(\frac{d_{n+1}}{(1 + \gamma_n)(1 + \sigma_n)} + \frac{\gamma_n + \sigma_n + \gamma_n \sigma_n}{(1 + \gamma_n)(1 + \sigma_n)} \Phi_{n+1}^{**} \right) \right). \quad (9)$$

Из приведенных выше равенств следует, что значения y_n^* , получившиеся при реальных вычислениях, совпадают со значениями, которые получаются при вычислениях без округлений в случае решения системы

$$y_{j-1}^* - (2 + p_j^* h^2) y_j^* + y_{j+1}^* = f_j^* h^2, \quad j = 1, \dots, N-1, \quad (10)$$

$$y_0^* = a, \quad y_N^* = b.$$

Оценка вычислительной погрешности свелась к оценке близости решений двух систем (1) и (10) с близкими коэффициентами. Удобнее переписать эти системы в виде

$$l_h(y_j) = \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} - p_j y_j = f_j, \quad (11)$$

$$l_h^*(y_j^*) = \frac{y_{j+1}^* - 2y_j^* + y_{j-1}^*}{h^2} - p_j^* y_j^* = f_j^*. \quad (12)$$

Мы ограничимся рассмотрением случая $p(x) \geq 0$, где оценка погрешности производится наиболее просто; будем предполагать также, что $|p_n h^2| \leq 2$ и $|f_n h^2| \leq \text{const}$. Вычитая (12) из (11), можно получить уравнение для ошибки $R_j = y_j - y_j^*$

$$l(R_j) = f_j - f_j^* + (p_j^* - p_j) y_j^* = v_j.$$

Вследствие леммы из § 1 имеем оценку погрешности

$$\max_j |R_j| \leq \frac{1}{4} \max_j |v_j| \leq \frac{1}{4} (\max_j |f_j - f_j^*| + (\max_j |p_j - p_j^*|) \max_j |y_j^*|). \quad (13)$$

Всякое число, не меньшее 2^s (s — целое), округляется до числа, не меньшего 2^s , а всякое число, не большее 2^s , округляется до числа, не большего 2^s . Очевидно,

$C_1 = \frac{1}{2 + p_1 h^2} \geq \frac{1}{4}$; если $\frac{1}{4} \leq C_n^* \leq 1$, то вследствие вышесказанного и рекуррентной формулы (2) также будет

$$C_{n+1}^* = \frac{1}{2 - C_n^*} \leq \frac{1}{2 - 1} = 1, \quad C_{n+1}^* \geq \frac{1}{4 - C_n^*} \geq \frac{1}{4}.$$

Таким образом,

$$\max_n |C_n^*| \leq 1, \quad \max_n |(C_n^*)^{-1}| \leq 4.$$

Поэтому на основании (8) имеем оценку

$$\max_n |g_n| \leq C' 2^{-t} h^{-2},$$

из (9) следует оценка

$$|q_n| \leq C'' 2^{-t} h^{-2} (|y_{n-1}^*| + |y_n^*| + |y_{n+1}^*|)$$

и, следовательно,

$$\max_n |q_n| \leq C''' 2^{-t} h^{-2} \max_n |y_n^*|.$$

Подставив эти оценки в правую часть (13), получим

$$\max_j |y_j - y_j^*| \leq C^0 2^{-t} h^{-2} \max_j |y_j^*|,$$

иначе

$$\|y - y^*\| \leq C^0 2^{-t} h^{-2} \|y^*\|. \quad (14)$$

Мы получили оценку приближенного решения сеточной задачи через его норму. Оценку через норму точного решения можно получить следующим образом. Из последнего неравенства следует, что

$$\|y^*\| \leq \|y\| + C^0 2^{-t} h^{-2} \|y^*\|,$$

откуда получаем

$$(1 - C^0 2^{-t} h^{-2}) \|y^*\| \leq \|y\|, \quad \|y^*\| \leq \frac{\|y\|}{1 - C^0 2^{-t} h^{-2}},$$

если $1 - C^0 2^{-t} h^{-2} > 0$; подставляя теперь эту оценку в правую часть (14), имеем

$$\|y - y^*\| \leq \frac{C^0 2^{-t} h^{-2}}{1 - C^0 2^{-t} h^{-2}} \|y\|.$$

При желании оценить вычислительную погрешность непосредственно через коэффициенты исходной системы следует воспользоваться неравенством

$$\max_j |y_j| \leq \max(|y_0|, |y_N|) + \frac{1}{4} \max_j |f_j|.$$

На первый взгляд могло бы показаться, что описанный метод решения системы (1) плох, поскольку велика погрешность результата вследствие округлений при вычислениях.

Однако погрешность такого порядка уже является следствием самой формы записи системы. Как правило, коэффициент $2 + p_j h^2$ запишется в машине с округлением ε_j порядка 2^{-1} , т. е. вместо него будет стоять число $s_j = 2 + p_j h^2 + \varepsilon_j$. Иначе можно написать

$$s_j = 2 + p_j^* h^2, \quad p_j^* = p_j + \varepsilon_j h^{-2}.$$

Таким образом, эти округления равносильны возмущениям коэффициента p_j исходной системы на величину порядка $2^{-1} h^{-2}$. Согласно рассуждениям из § 15 предшествующей главы это обстоятельство может привести к возмущению решения системы (1) на величину того же порядка.

Чтобы погрешность решения системы (1) была существенно меньше, необходимо по крайней мере задавать ее в форме, где округления коэффициентов системы равносильны существенно меньшим возмущениям коэффициентов исходной дифференциальной задачи. С этой целью можно, например, перейти к системе

$$\frac{y_j - y_{j-1}}{h} = z_j, \quad \frac{z_{j+1} - z_j}{h} - p_j y_j = f_j.$$

Соответственно при решении этой системы вместо рекуррентных соотношений (2), (3) относительно коэффициентов S_k , φ_k следует перейти к рекуррентным соотношениям (4.9), (4.10) относительно α_k , β_k .

- Дополнительная литература к главе IX

I [1, 4, 5, 13], II [2, 3, 28, 40, 44]; § 1: II [27], III [16]; § 2: II [40], III [8, 42]; § 3: II [10], III [4, 20]; § 4: III [43]; § 5: III [39]; §§ 6, 7: III [1, 2, 3, 4, 20, 21, 36, 52]; § 8: II [3, 41, 44]; § 9: II [40], III [14, 28, 37, 40, 42]; § 10: II [18, 44], III [38, 51]; § 11: III [12, 49, 50]; § 12: II [2, 3], III [8]; § 13: III [8].

СПИСОК ЛИТЕРАТУРЫ

(в списке указывается литература, содержащая более подробное изложение отдельных вопросов)

I. Учебная литература

1. Березин И. С., Жидков Н. П., Методы вычислений, т. 1, «Наука», 1966; т. 2, Физматгиз, 1962.
2. Воеводин В. В., Численные методы алгебры. Теория и алгоритмы, «Наука», 1966.
3. Гавурин М. К., Лекции по методам вычислений, «Наука», 1971.
4. Годунов С. К., Рябенский В. С., Разностные схемы (введение в теорию), «Наука», 1973.
5. Коллатц Л., Численные методы решения дифференциальных уравнений, ИЛ, 1953.
6. Коллатц Л., Функциональный анализ и вычислительная математика, «Мир», 1969.
7. Крылов А. Н., Лекции о приближенных вычислениях, Гостехиздат, М—Л., 1950.
8. Крылов В. И., Приближенное вычисление интегралов, «Наука», 1967.
9. Крылов В. И., Бобков В. В., Монастырский П. Н., Вычислительные методы высшей математики, 1, «Высшая школа», Минск, 1972.
10. Ланцош К., Практические методы прикладного анализа, Физматгиз, 1961.
11. Микеладзе Ш. Е., Численные методы математического анализа, Гостехиздат, 1953.
12. Милн В. Э., Численный анализ, ИЛ, 1951.
13. Милн В. Э., Численное решение дифференциальных уравнений, ИЛ, 1955.
14. Фаддеев Д. К., Фаддеева В. Н., Вычислительные методы линейной алгебры, Физматгиз, 1960.
15. Форсайт Дж., Молер К., Численное решение систем линейных алгебраических уравнений, «Мир», 1969.
16. Хаусхолдер А. С., Основы численного анализа, ИЛ, 1956.
17. Хемминг Р. В., Численные методы, «Наука», 1972.

II. Монографии и отдельные выпуски

1. Алберг Дж., Нильсон Э., Уолш Дж., Теория сплайнов и ее приложения, «Мир», 1972.
2. Бабушка И., Витасек Э., Прагер М., Численные процессы решения дифференциальных уравнений, «Мир», 1969.
3. Бахвалов Н. С., Вычислительные методы решения обыкновенных дифференциальных уравнений, Ин-т кибернетики АН УССР, Киев, 1970.

4. Бриллюэн Л., Наука и теория информации, Физматгиз, 1960.
5. Бусленко Н. П., Шрейдер Ю. А., Метод статистических испытаний (Монте-Карло) и его реализация на цифровых вычислительных машинах, Физматгиз, 1961.
6. Витушкин А. Г., Оценка сложности задач табулирования, Физматгиз, 1959.
7. Воробьев Ю. В., Метод моментов в прикладной математике, Физматгиз, 1958.
8. Вычислительные методы линейной алгебры, под ред. Марчука Г. И., Изд-во СО АН СССР, Новосибирск, 1972.
9. Гельфонд А. О., Исчисление конечных разностей, «Наука», 1967.
0. Годунов С. К., Рябенский В. С., Введение в теорию разностных схем, Физматгиз, 1962.
1. Гончаров В. Л., Теория интерполирования и приближения функций, Гостехиздат, 1954.
2. Дженкинс Г., Ваттс Д., Спектральный анализ и его приложения, т. 2, «Мир», 1972.
3. Ермаков С. М., Метод Монте-Карло и смежные вопросы, «Наука», 1971.
4. Жилейкин Я. М., Халус Е., Стандартная программа для вычисления многократных интегралов с помощью квадратур Гаусса, вып. 39, ВЦ МГУ, 1970.
5. Загускин В. Л., Справочник по численным методам решения алгебраических и трансцендентных уравнений, Физматгиз, 1960.
6. Зойтендейк Г., Методы возможных направлений, ИЛ, 1963.
7. Канторович Л. В. и Крылов В. И., Приближенные методы высшего анализа, Гостехиздат, М.—Л., 1949.
8. Коллатц Л., Задачи на собственные значения, «Наука», 1968.
9. Копченова Н. В., Марон И. А., Вычислительная математика в примерах и задачах, «Наука», 1972.
0. Коробов Н. М., Теоретико-числовые методы в приближенном анализе, Физматгиз, 1963.
1. Крылов В. И., Шульгина Л. Т., Справочная книга по численному интегрированию, «Наука», 1966.
2. Лаврентьев М. М., О некоторых некорректных задачах математической физики, Изд-во СО АН СССР, Новосибирск, 1962.
3. Латтес Р., Лионс Ж.-Л., Метод квазиобращения и его приложения, «Мир», 1970.
4. Линник Ю. В., Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений, Физматгиз, 1962.
5. Марчук Г. И., Лебедев В. И., Численные методы в теории переноса нейтронов, Атомиздат, 1971.
6. Метод статистических испытаний (метод Монте-Карло), СМБ, Физматгиз, 1962.
7. Микеладзе Ш. Е., Новые методы интегрирования дифференциальных уравнений и их приложения к задачам теории упругости, Гостехиздат, М.—Л., 1951.
8. Михлин С. Г., Численная реализация вариационных методов, «Наука», 1966.
9. Моисеев Н. Н., Численные методы в теории оптимальных систем, «Наука», 1971.
0. Морз Ф. М., Кембелл Д. Е., Методы исследования операций, «Советское радио», 1956.
1. Моррис У., Наука об управлении. Байесовский подход, «Мир», 1971.
2. Налимов В. В., Чернова Н. А., Статистические методы планирования экстремальных экспериментов, «Наука», 1965.
3. Никольский С. М., Квадратурные формулы, Физматгиз, 1958.

- 34 Островский А М, Решение уравнений и систем уравнений, ИЛ, 1963
- 35 Плэтт В, Информационная работа статистической разведки, ИЛ, 1958.
- 36 Процедуры численного интегрирования на языке Алгол, ВЦ МГУ, 1972.
- 37 Растринг Л А, Статистические методы поиска, «Наука», 1968
- 38 Ремез Е Я, Общие вычислительные методы чебышевского приближения, Изд-во АН УССР, Киев, 1957
- 39 Рябенский В С, Филиппов А Ф, Об устойчивости разностных уравнений, Гостехиздат, 1956
- 40 Самарский А. А, Введение в теорию разностных схем, «Наука», 1971
- 41 Современная математика для инженеров, под ред Беккенбаха Э Ф, ИЛ, 1958
- 42 Уилкинсон Дж Х, Алгебраическая проблема собственных значений, «Наука», 1970
- 43 Уолш Дж Л, Интерполяция и аппроксимация рациональными функциями в комплексной плоскости, ИЛ, 1961
- 44 Шаманский В Е, Методы численного решения краевых задач на ЭЦВМ, ч I, Изд во АН УССР, Киев, 1963, ч II, «Наукова думка», Киев, 1966
- 45 Яноши Л, Теория и практика обработки результатов измерений, «Мир», 1968
- 46 Dahlquist Y, Stability and error bounds in the numerical integration of ordinary differential equations, Uppsala, Almqvist & Wiksells boktr 130 (1959), 5—92
- 47 Stroud A H and Secrest D, Gaussian Quadrature Formulas, Prentice—Hall, Englewood Cliffs, N Y, 1966

III. Статьи

- 1 Абрамов А А, Вариант метода прогонки, Ж вычисл матем и матем. физ 1, № 2 (1961), 349—351
- 2 Абрамов А А, О переносе граничных условий для систем обыкновенных дифференциальных уравнений (вариант метода прогонки), Ж вычисл матем и матем физ 1, № 3 (1961), 542—543
- 3 Абрамов А А, О переносе условия ограниченности для некоторых систем обыкновенных линейных дифференциальных уравнений, Ж вычисл. матем и матем физ 1, № 4 (1961), 733—737
- 4 Абрамов А А, Андреев В Б, О применении метода прогонки к нахождению периодических решений дифференциальных и разностных уравнений, Ж вычисл матем и матем физ 3, № 2 (1963), 377—381
- 5 Бабури О В, Лебедев В И, О вычислении интегралов в смысле главного значения, весов и узлов квадратурных формул Гаусса, ЖВМ и МФ 5, № 3 (1965), 454—462
- 6 Бахвалов Н С, К оценке ошибки при численном интегрировании дифференциальных уравнений экстраполяционным методом Адамса, ДАН СССР 104, № 5 (1955), 683—686
- 7 Бахвалов Н С, Некоторые замечания к вопросу о численном интегрировании дифференциальных уравнений методом конечных разностей, ДАН СССР, 104, № 6 (1955), 805—808
- 8 Бахвалов Н С, О накоплении вычислительной погрешности при численном решении дифференциальных уравнений, Сб «Вычисл методы и программирование», вып I, Изд во МГУ, 1962, 47—68
- 9 Бахвалов Н С, Об определении начального шага и оценке главного члена погрешности при численном интегрировании с автоматическим выбором шага, Сб «Вычисл методы и программирование», вып I, Изд-во МГУ, 1962, 69—79

- 10 Бахвалов Н С, Об оптимальных оценках скорости сходимости квадратурных процессов и методов интегрирования типа Монте-Карло на классах функций, Сб «Числ методы решения диффер и интегр уравнений и квадратурные формулы», «Наука», 1964, 5—63
- 11 Бахвалов Н С, Об алгоритмах выбора шага интегрирования, Сб «Вычисл методы и программирование», вып V, Изд-во МГУ, 1966, 3—8
- 12 Бахвалов Н С, К оптимизации методов решения краевых задач при наличии пограничного слоя, Ж вычисл матем и матем физ 9, № 4 (1969), 841—859
- 13 Бахвалов Н С, Об оптимальности линейных методов приближения операторов на выпуклых классах функций, Ж вычисл матем и матем физ 11, № 4 (1971), 1014—1018
- 14 Бахвалов Н С, Об оптимизации методов решения обыкновенных дифференциальных уравнений с сильно осциллирующими решениями, Ж вычисл матем и матем физ 11, № 5 (1971), 1318—1322
- 15 Бродский М Л, Асимптотические оценки погрешности при численном интегрировании систем обыкновенных дифференциальных уравнений разностными методами, ДАН СССР 93, № 3 (1953), 599—602
- 16 Волков Е А, Эффективные оценки погрешности разностных решений краевых задач для обыкновенного дифференциального уравнения, Тр МИАН 112 (1971), стр 141—151
- 17 Валеев К Г, Ракитский Ю Р, Об отыскании разностными методами областей неустойчивости решений системы линейных дифференциальных уравнений, Сб «Числ методы решения диффер и интегр уравнений и квадратурные формулы», «Наука», 1964, 110—122
- 18 Гайсарян С С, О построении начальных условий при интегрировании систем обыкновенных дифференциальных уравнений методом Адамса, Сб «Вычисл методы и программирование», вып III, Изд во МГУ, 1965, 262—265
- 19 Гельфанд И М, Цейтлин М Л, Принципы нелокального поиска в системах автоматической оптимизации, ДАН СССР 137, № 2 (1961), 295—298
- 20 Годунов С К, Метод ортогональной прогонки для решения систем разностных уравнений, Ж вычисл матем и матем физ 2, № 6 (1962), 972—982
- 21 Годунов С К, О численном решении краевых задач для систем линейных обыкновенных дифференциальных уравнений, УМН 16, вып 3 (1961), 171—174
- 22 Горбунов А Д, Шахов Ю А, О приближенном решении задачи Коши для обыкновенных дифференциальных уравнений с наперед заданным числом верных знаков, I, Ж вычисл матем и матем физ 3, № 2 (1963), 239—253, II, там же, 6, № 3 (1964), 426—433
- 23 Гончар А А, О скорости рациональной аппроксимации непрерывных функций с характерными особенностями, Матем сб 73, вып 4 (1967), 630—638
- 24 Диткин В А, Люстерник Л А, Об одном способе практического гармонического анализа на сфере Сб «Вычисл матем и вычисл техника», № 1, Изд во АН СССР, 1953, 3—13
- 25 Долгополова Т Ф, Иванов В К, О численном дифференцировании, Ж вычисл матем и матем физ 6, № 3 (1966), 570—576
- 26 Дьяконов Е Г, О построении итерационных методов на основе использования операторов, эквивалентных по спектру, Ж вычисл матем и матем физ 6, № 1 (1966), 12—34
- 27 Жилейкин Я М, О квадратурных формулах на классах функции, Ж вычисл матем и матем физ 8, № 3 (1968), 507—516.

28. Ильин А. М., Разностная схема для дифференциального уравнения с малым параметром при старшей производной, Матем. заметки 6, № 2 (1969), 237—248.
29. Канторович Л. В., Функциональный анализ и прикладная математика, УМН 3, вып. 6 (28) (1948), 89—181.
30. Колмогоров А. Н., Тихомиров В. М., ϵ -энтропия и ϵ -емкость множеств в функциональных пространствах, УМН 14, вып. 2 (1959), 3—86.
31. Ким Г., О решении некоторых систем линейных алгебраических уравнений высокого порядка, Сб. «Вычисл. методы и программирование», вып. 3, Изд-во МГУ, 1965, 134—146.
32. Крейн С. Г., Шаблицкая И. О., Об устойчивости разностных схем для задачи Коши, Ж. вычисл. матем. и матем. физ. 6, № 4 (1966), 648—664.
33. Морозов В. А., О восстановлении функции методом регуляризации, Ж. вычисл. матем. и матем. физ. 7, № 4 (1967), 874—881.
34. Морозов В. А., О задаче дифференцирования и некоторых алгоритмах приближения экспериментальной информации, Сб. «Вычисл. методы и программирование», вып. 14, Изд-во МГУ, 1970, 46—62.
35. Мысовских И. П., О кубатурных формулах для вычисления интегралов по поверхности сферы, Сиб. матем. ж. 5, № 3 (1964), 721—723.
36. Огнева В. В., Метод «прогонки» для решения разностных уравнений, Ж. вычисл. матем. и матем. физ. 7, № 4 (1967), 803—812.
37. Повзнер А. Я., Павлов Б. В., Об одном методе численного интегрирования систем обыкновенных дифференциальных уравнений, Ж. вычисл. матем. и матем. физ. 3, № 4 (1973), 256—259.
38. Приказчиков В. Г., Однородные разностные схемы высокого порядка точности для задачи Штурма — Лиувилля, Ж. вычисл. матем. и матем. физ. 9, № 2 (1969), 315—336.
39. Рябенский В. С., Необходимые и достаточные условия хорошей обусловленности краевых задач для систем обыкновенных разностных уравнений, Ж. вычисл. матем. и матем. физ. 4, № 2 (1964), 242—255.
40. Ракитский Ю. В., Методика последовательного увеличения шага численного интегрирования систем обыкновенных дифференциальных уравнений, ДАН СССР 207, № 4 (1972), 793—795.
41. Салихов Н. П., О полярных методах решения задачи Коши для систем обыкновенных дифференциальных уравнений, Ж. вычисл. матем. и матем. физ. 2, № 4 (1962), 515—528.
42. Самарский А. А., Тихонов А. Н., Об однородных разностных схемах, Ж. вычисл. матем. и матем. физ. 1, № 1 (1961), 5—63.
43. Соболев С. Л., Некоторые замечания о численном решении интегральных уравнений, Изв. АН СССР, сер. матем. 20, № 4 (1956), 413—436.
44. Соболев С. Л., О формулах механических кубатур на поверхности сферы, Сиб. матем. ж. 3, № 5 (1962), 769—796.
45. Тихонов А. Н., О решении некорректно поставленных задач и методе регуляризации, ДАН СССР 151, № 3 (1963), 501—504.
46. Тихонов А. Н., Об устойчивости алгоритмов для решения вырожденных систем линейных алгебраических уравнений, Ж. вычисл. матем. и матем. физ. 5, № 4 (1965), 718—722.
47. Тихонов А. Н., О некорректных задачах оптимального планирования, Ж. вычисл. матем. и матем. физ. 6, № 1 (1966), 81—89.
48. Тихонов А. Н., Горбунов А. Д., Асимптотические разложения погрешности разностного метода решения задачи Коши для системы обыкновенных дифференциальных уравнений, Ж. вычисл. матем. и матем. физ. 2, № 4 (1962), 536—548.

49. Тихонов А. Н., Горбунов А. Д., Оценки погрешности метода типа Рунге — Кутта и выбор оптимальных сеток, Ж. вычисл. матем. и матем. физ. 4, № 2 (1964), 232—241.
50. Тихонов А. Н., Горбунов А. Д., Гайсарян С. С., Описание алгоритма построения оптимальных сеток при решении задачи Коши для обыкновенных дифференциальных уравнений методом Рунге — Кутта, Сб. «Вычисл. методы и программирование», вып. 6, Изд-во МГУ, 1966, 17—20.
51. Тихонов А. Н., Самарский А. А., Разностная задача Штурма — Лиувилля, Ж. вычисл. матем. и матем. физ. 1, № 5 (1961), 784—805.
52. Шапкин А. Ф., Замыкание двух вычислительных алгоритмов, основанных на идее ортогонализации, Ж. вычисл. матем. и матем. физ. 7, № 2 (1967), 411—416.
53. Bulirsch R., Bemerkungen zur Romberg — Integration, Numer. Math. 6, № 11 (1964), 6—16.
54. Butcher I. G., A Modified Multistep method for the numerical integration of ordinary differential equations, J. Assoc. Comput. mach., vol. 12, № 1 (1965), 124—135.
55. Dahlquist Y., Convergence and stability in the numerical integration of ordinary differential equations, Math. Skand, vol. 4 № 1 (1956), 33—53.
56. Distefano G. P., Causes of instabilities in numerical integration techniques, Internat. J. Comput. Math. vol. 2, № 2 (1968), 123—142.
57. Filippi S., Kraske F., Neue verallgemeinerte Multistep Methoden., ZAMM 48, № 8 (1968), Sonderh, 65—66.
58. Fyfe D. J., Economic evaluation of Runge — Kutta formulae, Math. Comp. 20, 95 (1966), 392—398.
59. Gragg W. B., Stetter H. I., Generalized Multistep Predictor Corrector Methods, J. Assoc. Comput. Mach., vol. 11, № 2 (1964), 188—209.
60. Grane R. L., Klopenstein. A Predictor — Corrector Algorithm with an Increased Range of Absolute Stability, J. Assoc. Comput. Mach., vol. 12, № 2 (1965), 227—242.
61. Haber S., Numerical evaluation of multiple integral, SIAM Rev., vol. 12, № 4 (1970), 481—526.
62. Liniger W., A criterion for A stability of linear multistep integration formylae, Computing 3, № 4 (1968), 280—285.
63. Reinsch R., Smoothing by Spline Functions, Numer. Mat. 10 (1967), 177—183.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

- Аппроксимация граничного условия 541
— дифференциального уравнения разностной схемой 472, 538
Асимптотически оптимальный метод 147
- Большое число 20
- Валле — Пуссена теорема 225
Верная цифра 20
Весовая функция 96
Выделение весовой функции 186
— корней 420
Вычислительная погрешность при решении задач алгебры 330, 343, 618
— — — задачи Коши 461, 483, 612
— — — краевых задач 549, 573, 617, 618
— — — численном дифференцировании 92
— — — интегрировании 614
— — — интерполировании 76
- Гарантированные оценки погрешности на классе функций 300
Главный член погрешности квадратур 159
— — численного интегрирования дифференциальных уравнений 466, 488, 539
- Двусторонние методы интерполирования 528
— — решения дифференциальных уравнений 532
— — численного интегрирования 529
($2m + 1$)-диагональная матрица 555
Дискретные коэффициенты Фурье 220
Диссипация 430
Дифференцирование численное 88
- Задача с пограничным слоем 610
Замыкание алгоритма 558
— метода прогонки 559, 577
— нерегулярное 558
— регулярное 558
Значащие цифры 20
- Изменение шага численного интегрирования 193, 459, 517
Интегрирование дифференциальных уравнений на большом отрезке 465, 493, 507, 510, 517
— осциллирующих функций 125
— системы уравнений 502
Интерполирование 32
— с кратными узлами 42
Интерполяционная формула Ньютона с разделенными разностями 40
Интерполяционные формулы 467
— — Бесселя 68
— — Гаусса 68
— — Лагранжа 36
— — Ньютона для интерполирования вперед 66
— — — — назад 66
— — — равных промежутков 66
— — Эверетта 72
Интерполяция 35
— квадратичная 70
— линейная 70
Итерационные методы построения многочлена наилучшего равномерного приближения 235
— — с использованием спектрально эквивалентных операторов 385
- Квадратуры Гаусса 107
— Грегори 166
— Лобатто 114
— Ньютона — Котеса 95
— обобщенные 131

Квадратуры прямоугольников 95, 99
 — Ромберга 175
 — Симпсона 100
 — со случайными узлами 311

— составные 131
 — трапещий 95, 99

— Филона 126

— Чебышева 102

— Эйлера 165

— Эрмита 117

Конечно разностные методы 467

— — схемы 467

Контроль погрешности на шаге 119, 459

Контрольный столбец системы 328

Линейная оценка погрешности 24

Линейно зависимая система элементов 30

— независимая система элементов 30, 79

Линейное множество 30

— нормированное пространство 31

— подпространство 30

— разностное уравнение (k -го порядка) 47

— — с постоянными коэффициентами 50, 56

Мажорирующая функция 542

Мажорирующее разностное уравнение 542, 612

Матрица ортогонализации 80

Машина с плавающей запятой 19

— — фиксированной запятой 18

Мера обусловленности задачи 570

— — матрицы 390

— — системы 389

Метод Адамса 468

— Адамса — Штермера 520, 522, 523

— Батчера 475

— вилки 419, 584

— вращений отыскания собственных значений 400

— Гаусса 324

— — с выбором главного элемента 329

— Гершгорина 542

— градиентного спуска 369, 422

— Зейделя 363, 408

— исключения Гаусса 324

— квадратного корня 330

— Ланцоша 377

— Милна 493

— Монте-Карло 301

Метод Монте-Карло решения системы линейных уравнений 378

— наименьших квадратов 272

— наискорейшего спуска 369, 422

— неопределенных коэффициентов 35, 109, 271, 471, 520

— Ньютона решения нелинейных уравнений 412

— оврагов 427

— оптимального по координатного спуска 422

— ортогонализации 333

— парабол 420

— по координатного спуска 366, 420, 422

— пристрелки 549, 584

> — прогонки 552, 574

— простой итерации 335, 407

— регуляризации 274, 393

— релаксации 368

— сверхрелаксации 368

— секущих 416

— сопряженных градиентов 372, 377

— спуска 420

— стрельбы 549, 584

— тяжелого шарика 431

— установления 429

— уточнения разностями высших порядков 540

> — Эйлера 450, 493

Методы дифференциальной ортогональной прогонки 577, 580, 582

— интерполяционные 467, 520

— Рунге — Кутты 450

— с забеганием вперед 467

— экстраполяционные 467, 520

Метрическое пространство 30

Многочлен наилучшего равномерного приближения 225

Многочлены Лагерра 86

> — Лежандра 85

—, наименее уклоняющиеся от нуля 58

— Чебышева 56

— — второго рода 85

— — первого рода 85

— Эрмита 86

— Якоби 85

Многошаговый способ численного интегрирования 461

Наилучшее равномерное приближение 225

Начальные условия для конечно разностной схемы 512

Недетерминированный метод 313

Некорректные задачи 391
 Неравенство Чебышева 302
 Неявная схема, метод 467, 514, 520
 Неявно заданное дифференциальное уравнение 515
 Норма 31
 Нормировка конечно-разностной схемы 473
 Нормы векторов и матриц 335

Область сходимости метода 406
 Обратная интерполяция 78
 Обратный ход метода Гаусса 328
 — — прогонки 552
 Обращение матриц 332
 Обусловленность матрицы системы 389
 — системы 389
 Однородные схемы 597
 Односторонние формулы численного дифференцирования 92
 Одношаговые методы численного интегрирования 461
 Оператор конечной разности 63
 Оптимальные квадратуры 134, 139, 140
 Оптимальный линейный итерационный процесс 360
 — — многошаговый итерационный процесс 356
 — — одношаговый итерационный процесс 356
 — метод 60, 134
 Оптимизация методов 60, 134
 — оценки погрешности интерполирования 59
 — распределения узлов интегрирования 147, 151, 606
 — скорости сходимости итерационного процесса 352
 Ортогональная прогонка 580
 — система 79
 Ортогональные многочлены 81, 85
 Ортонормированная система элементов 81
 Остаточный член формулы Лагранжа 37
 Оценка вычислительной погрешности путем изменения масштабов 330, 617
 Очень большое число 20

Погрешность абсолютная 19
 — аппроксимации дифференциального уравнения разностной схемой 472
 — вычислительная 15
 — квадратуры на классе функций 103, 139, 292
 — математической модели 18
 — метода 15
 — — на классе задач 60, 103, 139
 — — на шаге 452, 488
 — неустраняемая 15, 18
 — относительная 19
 — предельная 23
 Понижение размерности задач оптимизации 425
 — — нелинейных систем 428
 Порядок метода 410
 — погрешности аппроксимации 473
 — — метода Рунге — Кутты 452
 Правило Рунге 168, 492, 539
 Практическая оценка погрешности интерполяции 41, 67, 528
 — — — квадратур 119, 167
 — — — при решении дифференциальных уравнений 492, 539, 617
 — — — — линейных систем 330, 349
 Предельная точность интегрирования на классах 293, 296, 313
 Преобразование Фурье быстрое 222
 — — дискретное 218
 Проблема собственных значений полная 394
 — — — частичная 395
 Производная оператора 411
 δ^2 -процесс ускорения сходимости 351
 Прямое произведение формул интегрирования, интерполирования, дифференцирования 288, 289, 320
 Прямой ход метода Гаусса 328
 — — — прогонки 552

Разности вперед 62
 — высшего порядка 62
 — конечные 62
 — назад 62
 — первого порядка 62
 — разделенные 37
 — с кратными узлами 43
 — центральные 62
 Ряд Фурье дискретный, конечный 220

Параметризация задачи 438, 585
 Плохо обусловленные системы 391
 Повышение порядка точности разностной схемы 538

Сеточная функция Грина 542
 Сжатое отображение 407
 Сильный рост функции 19

- Симметризация системы уравнений 353, 362, 394
 Случайный покоординатный спуск 422
 Согласованные нормы 335
 Спектр семейства матриц 343
 Спектральный радиус матрицы 344
 — — семейства матриц 344
 > Сплайн 256
 Стандартные программы численного интегрирования 193, 201
 Строго нормированное пространство 31
 Схема Эйткена 40
 Схемы, точные на решениях специального вида 511, 597, 601

 Таблица разделенных разностей функции 39
 Телескопический метод 232
 Теорема Валле — Пуссена 225
 — о главном члене погрешности 489
 — об оценке погрешности 484
 — Чебышева 226
 Точка спектра семейства матриц 343
 Точки чебышевского альтернанса 226
 Тригонометрическая интерполяция 220

 Узлы интерполяции 35
 Уплотнение таблиц 72
 Уравнения в конечных разностях 47
 Ускорение сходимости метода Монте-Карло 307, 311
 Условие α 482
 Условия сходимости при решении задачи Коши конечно-разностными методами 482, 484, 524

 Формула Гаусса для интерполирования вперед 68
 Формула Гаусса для интерполирования назад 68
 — — прямоугольников 95, 99
 — — с кратными узлами 99
 — Ромберга 175
 — Симпсона 100
 — трапеций 95, 99
 — Филона 126
 Формулы Адамса 468
 — Грегори 166
 — Лобатто 114
 — Ньютона 66
 — с забеганием вперед 467
 — численного дифференцирования 88
 — Эрмита 117
 Фундаментальная система решений 54
 Функция сильно растущая 20

 Характеристическое уравнение 50
 Характеристическое уравнение разностной схемы 477
 Хорошо обусловленная (поставленная) краевая задача 569, 570
 — обусловленные системы 391

 Целевая функция 436

 Шаг таблицы 62

 Эквивалентные нормы 338
 Элемент наилучшего приближения 210
 Элементарные функции 249
 Энтропия 263
 ϵ -энтропия 265

 Явная схема, метод 467, 520

Николай Сергеевич Бахвалов

ЧИСЛЕННЫЕ МЕТОДЫ

М., 1975 г., 632 стр. с илл.

Редакторы *И. М. Овчинникова,*
Е. В. Шикин

Техн. редакторы *И. Ш. Аксельрод,*
А. П. Колесникова

Корректор *Л. Н. Боровина*

Печать с матриц.

Подписано к печати 28/1 1975 г.

Бумага 60×90¹/₁₆, тип. № 2. Физ. печ. л. 39,5.

Усл. печ. л. 39,5. Уч.-изд. л. 38,1. Тираж 40000 экз.

Цена книги 1 р. 47 к. Заказ № 456

Издательство «Наука»

Главная редакция

физико-математической литературы

117071, Москва, В-71, Ленинский проспект, 15

Ордена Трудового Красного Знамени

Ленинградская типография № 2

имени Евгении Соколовой Союзполиграфпрома

при Государственном комитете

Совета Министров СССР

по делам издательств, полиграфии

и книжной торговли

198052, Ленинград, Л-52,

Измайловский проспект, 29