УЧЕБНОЕ ПОСОБИЕ

И. Г. Черноруцкий

МЕТОДЫ ОПТИМИЗАЦИИ В ТЕОРИИ УПРАВЛЕНИЯ

Допущено Министерством образования Российской Федерации в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлениям подготовки бакалавров и магистров «Системный анализ и управление» и «Информатика и вычислительная техника»

300.piter.com

Издательская программа

300 лучших учебников для высшей школы в честь 300-летия Санкт-Петербурга

осуществляется при поддержке Министерства образования РФ

ПИТЕР®

Москва · Санкт-Петербург · Нижний Новгород · Воронеж
Ростов-на-Дону · Екатеринбург · Самара · Новосибирск
Киев · Харьков · Минск
2004

ББК 22.181я7
УДК 517.97(075)
Ч49

Федеральная целевая программа «Культура России»
Поддержка полиграфии и книгоиздания России

Рецензенты:

Заборовский В. С., доктор технических наук, профессор, заведующий кафедрой «Телематика» Санкт-Петербургского государственного политехнического университета

Фомин Б. Ф., доктор технических наук, академик Международной академии информатизации, профессор Санкт-Петербургского государственного электротехнического университета

Юревич Е. И., доктор технических наук, заслуженный деятель науки Российской Федерации

Ч49 **Методы оптимизации в теории управления: Учебное пособие** / И. Г. Черноруцкий. — СПб.: Питер, 2004. — 256 с.: ил.

ISBN 5-94723-514-5

В учебнике рассматриваются проблемы параметрической оптимизации и применение методов оптимизации при решении типовых задач управления. Излагаемый материал учитывает особенности постановок задач управления и описывающих их математических моделей. Особое внимание в учебнике уделяется наиболее сложным и часто встречающимся в практике управления особенностям параметрической оптимизации: многокритериальности и «плохой» обусловленности задачи. Рассматриваемый материал иллюстрируется представительными примерами.

Для студентов и преподавателей технических вузов.

ББК 22.181я7
УДК 517.97(075)

Все права защищены Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав

ISBN 5-94723-514-5

© ЗАО Издательский дом «Питер», 2004

Содержание

Предисловие 6

Основные обозначения и терминологические замечания 10

Глава 1. Задачи конечномерной оптимизации в теории управления 12
1.1. Основные понятия теории управления 12
1.2. Система управления сложным объектом 20
1.3. Примеры задач параметрической оптимизации в теории управления 26
1.3.1. Идентификация нелинейных детерминированных объектов 29
1.3.2. Идентификация стохастических объектов 32
1.3.3. Идентификация нестационарных объектов 36
1.3.4. Экстремальное регулирование 37
1.3.5. Синтез адаптивных систем автоматического управления 39
1.3.6. Синтез статистически оптимальных систем автоматического управления 40
1.3.7. Оптимальное проектирование систем 43
1.4. Выводы 45

Глава 2. Математические модели теории параметрической оптимизации 46
2.1. Канонические задачи 46
2.2. Многокритериальные задачи 47
2.3. Компромиссы Парето 49
2.4. Методы исключения ограничений 54
2.5. Влияние неопределенных факторов на процесс оптимизации 57
2.6. Методы декомпозиции 60
2.6.1. Метод агрегирования 60
2.6.2. Метод вспомогательных частных критериев 62
2.7. Особенности оптимизационных задач 62
2.8. Некоторые стандартные схемы параметрической оптимизации 64
2.8.1. Задачи аппроксимации 64
2.8.2. Системы неравенств 66
2.8.3. Решение систем неравенств в условиях неопределенности 69
2.8.4. Сигномиальная оптимизация 70
2.9. Основные результаты и выводы 70

Глава 3. Проблема плохой обусловленности 72
3.1. Явление овражности 72
3.2. Основные определения 75

3.3. Критерии овражности 81
3.4. Источники плохо обусловленных оптимизационных задач в теории управления 82
3.4.1. Естественная овражность 82
3.4.2. Внесенная овражность 88
3.5. Методы параметрической оптимизации 92
3.5.1. Ньютоновские методы 92
3.5.2. Методы доверительной окрестности 95
3.5.3. Квазиньютоновские методы 96
3.5.4. Задачи высокой размерности 97
3.5.5. Глобальная оптимизация 98
3.5.6. Анализ сложившейся ситуации 99
3.6. Основные результаты и выводы 101

Глава 4. Покоординатные стратегии параметрической оптимизации . . . 104
4.1. Метод циклического покоординатного спуска 104
4.2. Методы обобщенного покоординатного спуска 108
4.3. Реализация методов обобщенного покоординатного спуска 115
4.3.1. Нормализация основных переменных задачи 115
4.3.2. Методы диагонализации 118
4.4. Алгоритмы обобщенного покоординатного спуска 119
4.4.1. Реализации на основе конечно-разностных аппроксимаций производных 119
4.4.2. Реализации на основе рекуррентных алгоритмов оценивания 123
4.5. Специальные реализации методов обобщенного покоординатного спуска 128
4.5.1. Задачи аппроксимации 128
4.5.2. Идентификация нелинейных детерминированных объектов на основе функциональных рядов Вольтерра 130
4.5.3. Корреляционные методы идентификации стохастических объектов 133
4.5.4. Синтез статистически оптимальных систем автоматического управления 134
4.5.5. Идентификация нелинейных динамических систем 134
4.5.6. Оценивание состояний динамических систем: задача о наблюдении 136
4.5.7. Идентификация возмущающих воздействий 137
4.5.8. Решение систем неравенств 138
4.5.9. Управление технологическим процессом серийного выпуска изделий 139
4.5.10. Обеспечение максимального запаса работоспособности оптимизируемой системы 140
4.5.11. Оптимизация систем по сигномиальным целевым функционалам . . . 141
4.5.12. Оптимальное управление 142
4.6. Основные результаты и выводы 143

Глава 5. Градиентные стратегии параметрической оптимизации 146
5.1. Общая схема градиентных методов. Понятие функции релаксации 146

5.2. Классические градиентные схемы 150
5.2.1. Простой градиентный спуск (ПГС) 150
5.2.2. Метод Ньютона 153
5.2.3. Метод Левенберга 154
5.3. Методы с экспоненциальной релаксацией 156
5.3.1. Реализация методов с экспоненциальной релаксацией 161
5.3.2. Области применения и анализ влияния погрешностей 164
5.4. Методы многопараметрической оптимизации 167
5.4.1. Методы с чебышевскими функциями релаксации 169
5.4.2. Характеристики сходимости и сравнение с методами сопряженных градиентов 174
5.5. Применение процедур RELEX и RELCH в задачах теории управления . . . 179
5.6. Тактика решения общей задачи параметрической оптимизации 181
5.7. Основные результаты и выводы 182

Глава 6. Методы уменьшения размерности вектора оптимизируемых параметров 184
6.1. Методы теории жестких систем 184
6.1.1. Принцип квазистационарности производных для линейных систем с симметричными матрицами 185
6.1.2. Методы иерархической оптимизации: частный случай 189
6.1.3. Методы иерархической оптимизации: общий случай 192
6.1.4. Принцип повторных измерений 200
6.1.5. Алгоритмы иерархической оптимизации 208
6.2. Методы исключения переменных на основе спектрального разложения матрицы Гессе 210
6.2.1. Постановка задачи 210
6.2.2. Алгоритм исключения 214
6.2.3. Удаление переменных в задаче наименьших квадратов 219
6.3. Основные результаты и выводы 223

Глава 7. Примеры решения реальных задач параметрической оптимизации 225
7.1. Реализация оптимальной весовой функции линейной стационарной системы 226
7.2. Аппроксимация характеристик частотно-избирательных фильтров 229
7.3. Оптимизация параметров переключательных электронных схем 234
7.4. Управление химико-технологическими процессами производства высокомолекулярных соединений 238
7.4.1. Кинетическая модель процесса термоинициированной полимеризации стирола в массе 241
7.4.2. Методика воспроизведения моделей полимеризационных процессов 244
7.4.3. Параметрическая идентификация кинетических моделей полимеризационных процессов (полимеризация стирола) 247
7.5. Идентификация моделей теплообменников атомных реакторов 250

Литература 253

Предисловие

Предлагаемая читателю книга является учебником по методам конечномерной оптимизации для студентов различных специальностей высших технических учебных заведений, обучающихся по направлению «Информатика и вычислительная техника».

В основу учебника положен курс лекций по методам оптимизации, который читается на протяжении ряда лет на факультете технической кибернетики Санкт-Петербургского государственного политехнического университета.

Учитывая прикладную направленность подготовки специалистов в технических вузах, основное внимание уделено практическим вопросам оптимизации, тесно связанным с работой в конкретных предметных областях. Как показывает практика компьютерного моделирования и решения реальных оптимизационных задач, обычно представленные в учебной литературе универсальные «библиотечные» методы нелинейного программирования часто оказываются малоэффективными и требуют адаптации при практической реализации. Кроме того, важнейшее значение приобретают элементы сопутствующих вычислительных технологий, связанные с формализацией конкретной прикладной проблемы и разработкой сценария оптимизации. Для изучения соответствующих вопросов, обычно выпадающих из стандартных учебных курсов, целесообразно излагать теорию и методы оптимизации «на фоне» некоторой предметной области, а не в абстрактном виде. Это позволяет, с одной стороны, глубже понять возникающие проблемы, а с другой — получить наглядные и конкретные интерпретации применяемых технологий.

В качестве такой «модельной» предметной области в книге выбрана теория управления, изучаемая с различной степенью подробности по многим направлениям подготовки современных специалистов. Таким образом, эта предметная область знакома студентам соответствующих специальностей. С другой стороны, важность такого выбора объясняется широким практическим применением аппарата современной алгоритмической теории управления при создании прикладных программных систем, в том числе при разработке встроенных систем управления. Ведь когда мы говорим об управлении, то подразумеваем очень простую по своей сути задачу: необходимо сформировать такое целенаправленное воздействие на объект управления, чтобы перевести его в некоторое «желаемое» состояние. Понятно, что под эту простую схему подпадают многие ситуации, как

связанные с нашей повседневной жизнью, так и возникающие внутри технических, экономических, программных, эко- и других систем. Обычно речь идет о некотором наилучшем в определенном смысле, то есть «оптимальном», управлении, что и определяет необходимость создания теории управления, базирующейся на принципах оптимальности. Помимо обеспечения оптимальности самого управляющего воздействия, алгоритмизация и компьютерная реализация большинства этапов процесса управления сводятся к решению целой цепочки различных оптимизационных задач. В этом смысле можно утверждать, что теория конечномерной (параметрической) оптимизации является основным алгоритмическим фундаментом современной теории управления.

Специфика возникающих в теории управления оптимизационных задач отражает общую ситуацию, возникающую, как уже говорилось, при решении практически любых реальных задач. Поэтому и в этом смысле, изучая соответствующие оптимизационные проблемы на примере задач теории управления, можно получать общезначимые выводы и заключения.

Опыт компьютерного моделирования и проведения реальных вычислений показывает, что, как правило, невозможно при создании прикладных программных систем эффективно использовать универсальное алгоритмическое обеспечение нелинейного программирования «в чистом виде» из-за резкого понижения скорости сходимости предлагаемых поисковых процедур. Весьма вероятной оказывается хорошо знакомая специалистам ситуация полной остановки («заклинивания» или «залипания», что соответствует англоязычному термину jamming) алгоритма минимизации целевого функционала задолго до достижения искомого оптимума. Практика показывает, что один из основных факторов сложности решения реальных оптимизационных задач может быть связан со специальным случаем явления плохой обусловленности, когда целевой функционал имеет «жесткий», или «овражный», характер, то есть резко возрастает по одним направлениям и слабо изменяется вдоль других, что и вызывает указанные трудности.

Высокие степени жесткости возникают не в исключительных «патологических» случаях, а отражают обычную ситуацию при решении практически любой прикладной задачи параметрической оптимизации. Наиболее остро проблема жесткости стоит при решении задач параметрической идентификации объектов компьютерного моделирования и управления, а также в задачах оптимального параметрического синтеза, как реальных, то есть уже существующих, так и проектируемых систем при наличии ограничений и векторного критерия оптимальности.

В настоящее время уже трудно установить, когда впервые было указано на явление овражности как на типичную практическую ситуацию, затрудняющую работу классических оптимизирующих процедур. Во всяком случае уже в начале компьютерной эры, в 1959 г., в фундаментальном труде А. А. Фельдбаума «Вычислительные устройства в автоматических системах» были, по существу, рассмотрены все основные признаки овражной ситуации, включая явление «заклинивания» и понятие многомерного дна оврага. Широкую известность проблема овражности приобрела после опубликования работ И. М. Гельфанда и М. Л. Цетлина (1961–1962), посвященных методам управления техническими

системами и применению поисковых процедур в системах автоматической оптимизации. Эти исследования привели к созданию известного «метода оврагов». В 1967 г. блестящий компьютерный математик и инженер профессор Л. А. Растригин обобщил метод оврагов на общий случай овражной ситуации, когда размерность дна оврага может превышать единицу. Следующий шаг в изучении проблемы был предпринят в монографии Л. А. Растригина «Системы экстремального управления» (1974 г.), где в связи с общими вопросами применения поисковых методов параметрической оптимизации в системах экстремального управления рассмотрены различные аспекты овражной ситуации и, в частности, построены простейшие модели овражных экстремальных объектов. Значительный вклад в теорию плохо обусловленных экстремальных задач внесли работы Ю. В. Ракитского, возглавившего петербургскую школу по изучению проблемы жесткости и указавшего на принципиальную связь явления овражности с концепцией жесткости систем обыкновенных дифференциальных уравнений. Эти исследования нашли отражение в монографии Ю. В. Ракитского, С. М. Устинова и И. Г. Черноруцкого «Численные методы решения жестких систем» (М.: Наука, 1979 г.).

Однако несмотря на важность полученных результатов, проблема ими не исчерпывалась. Оставался не решенным целый ряд существенных вопросов, связанных с развитием теории жестких (овражных) оптимизационных задач, а также общих принципов построения методов и алгоритмов параметрической оптимизации по жестким целевым функционалам. Кроме того, представляла практический интерес разработка проблемно-ориентированного алгоритмического и программного обеспечения, учитывавшего специфику отдельных классов задач и структурные особенности применяемых на практике критериев качества. При построении таких методов и алгоритмов должны были учитываться такие естественные для приложений сопутствующие факторы сложности, как невыпуклость минимизируемых функционалов, их негладкость, а в ряде случаев и высокая размерность вектора варьируемых параметров. Такие иследования проводились в течение последних десятилетий, и были получены достаточно интересные для теории и практики результаты.

Сформулированные задачи составляют настолько важную алгоритмическую проблему при оптимизации современных технических, экономических, программных и других систем и методов управления ими, что назрела необходимость представить их в систематизированном виде и включить в соответствующие учебные курсы. Данный учебник, по мнению автора, позволяет восполнить этот пробел.

Есть хорошее замечание кого-то из известных людей: «Если вы, перечитывая в очередной раз „Фауста“ Гете, не находите для себя чего-то нового, то вы остановились в своем развитии». Автор далек от мысли сравнивать данную книгу с «Фаустом» и, тем более, себя с Гете, однако многим изучающим теорию оптимизации следует иметь в виду, что предлагаемая вашему вниманию книга также задумывалась как «многослойная». При первом чтении многие вопросы могут казаться неясными, хотя общая картина должна быть понятна. Когда вы лучше подготовитесь и изучите некоторые вопросы по указанной дополнительной литературе, можно приступать ко второму чтению и т. д. В итоге все зависит от вашей предварительной математической подготовки и опыта практической рабо-

ты. В книге даются ссылки на учебную и научную литературу. Как правило, предпочтение отдавалось ссылкам на учебники и учебные пособия. Многие из этих книг являются бестселлерами и многократно переиздавались и переиздаются. Однако, как правило, ссылки сделаны на наиболее ранние «непереработанные» издания, которые часто оказываются (по мнению автора) и наиболее интересными. При желании читатель без труда обнаружит в каталогах и более поздние версии соответствующих учебников.

Для полного усвоения представленных в книге материалов предполагается знакомство с курсами высшей математики (содержащими, в частности, достаточно полные сведения по линейной и матричной алгебре), численного анализа, теории вероятностей (в частности, корреляционной теории) и основ теории управления. Если вы владеете этим материалом в стандартном для технических вузов объеме, то никаких проблем возникнуть не должно. Для облегчения использования учебника и согласования применяемых терминов и обозначений в первой главе даны необходимые для данного курса общие сведения из теории управления и сформулированы характерные постановки оптимизационных задач.

А вообще, следуйте рецепту великого Даламбера: «Идите вперед, а понимание придет потом!»

Основные обозначения и терминологические замечания

R	Множество вещественных (действительных) чисел
R^n	n-мерное евклидово пространство
$\forall, \exists$	Кванторы всеобщности и существования
$f: X \to Y$	Отображение f множества X во множество Y
$k \in [1: N]$	Число k принимает последовательно все значения из множества натуральных чисел от 1 до N включительно
$\{x^n\}$	Последовательность элементов x^n
(a, b) $\langle a, b \rangle$	Скалярное произведение векторов a и b
$A > 0$	Матрица A положительно определена
$\triangleq$	Равно по определению
$D = \{x \in H \,/\, P(x)\}$	Подмножество элементов множества H, обладающих свойством $P(x)$
$\varnothing$	Пустое множество
$C^k(D)$	Множество k раз непрерывно дифференцируемых на множестве D функций
$(\cdot)^+$	Псевдообратная матрица
МНК	Метод наименьших квадратов
diag λ_i, diag (λ_i)	Диагональная матрица
cond $[A]$	Спектральное число обусловленности матрицы A
J', J''	Вектор градиента и матрица вторых производных функционала $J(x)$
$\arg\min J(x)$	Минимизатор функционала $J(x)$
$\operatorname{Arg}\min J(x)$	Множество всех минимизаторов функционала $J(x)$

Сделаем также некоторые терминологические замечания по тексту учебника. Согласно общепринятым математическим канонам (см., например, [38]), термин

функционал означает (однозначное) отображение произвольного множества во множество вещественных чисел. В силу этого замечания вещественная функция от вещественного переменного, а также вещественная функция от нескольких вещественных переменных также являются функционалами. Иногда, особенно в прикладных технических публикациях, предлагается называть функционалом более частный случай отображения некоторого *множества функций* во множество вещественных чисел. Пример такого функционала дает определенный интеграл от интегрируемой на заданном промежутке функции. Мы, однако, будем придерживаться стандартной математической терминологии.

Под задачами *математического программирования* будем понимать конечномерные задачи оптимизации, то есть задачи поиска максимума или минимума функционала, определенного на некотором подмножестве *конечномерного* евклидова пространства. В этом случае, например, задачи теории оптимального управления, формулируемые как задачи поиска оптимальных управляющих функций из некоторого подмножества бесконечномерного пространства, уже не будут относиться нами к задачам математического программирования.

Мы будем следовать сложившейся в математике традиции и называть определитель матрицы Гессе *гессианом*, определитель матрицы Якоби — *якобианом*, определитель матрицы Вронского — *вронскианом* и т. д. В некоторых книгах и учебниках по математическому программированию и оптимизации сама матрица Гессе называется гессианом, что является необоснованным и приводит к ненужной путанице.

Глава 1. Задачи конечномерной оптимизации в теории управления

1.1. Основные понятия теории управления

Обычно в учебных курсах по теории управления рассматриваются некоторые специальные разделы этой теории, такие как теория автоматического управления или математическая теория оптимальных процессов, основанная на принципе максимума Понтрягина или принципе оптимальности Беллмана. В данном же случае нас будет интересовать *управляемая система* как некая общая концепция *кибернетики* — науки об управлении.

Под управлением будем понимать процесс такого воздействия на некоторую систему или объект (объект управления), при котором состояние системы или объекта изменяется «в нужную сторону». Объектами управления, очевидно, могут быть: техническое устройство (например, автомобиль), экономическая ситуация на предприятии или фирме, экосистема региона, процесс разработки программного проекта, сам программный проект и его характеристики и т. п. Предполагается, что мы можем не только оказывать воздействие на объект, но и оценивать результаты этого воздействия по некоторым заданным критериям. Например, критериями качества процесса разработки программного проекта могут служить время завершения проекта и его бюджет (стоимость разработки). Влиять на эти характеристики (управлять ими) мы можем, например, с помощью перераспределения ресурсов между отдельными работами, составляющими данный программный проект.

Общая схема процесса управления приведена на рис. 1.1.

Объект управления рассматривается как сколь угодно сложная система, преобразующая входные управляющие воздействия $U(t)$ в выходные сигналы (траектории) $V(t)$, характеризующие *состояние объекта управления* в момент времени t. Очевидно, что реальный объект управления может иметь множество входов и выходов, определяющих его функциональное взаимодействие с внешней сре-

дой. Все эти каналы связи со средой на рис. 1.1 не показаны. Изображены лишь те входы и выходы, которые являются существенными для формулирования задачи управления.

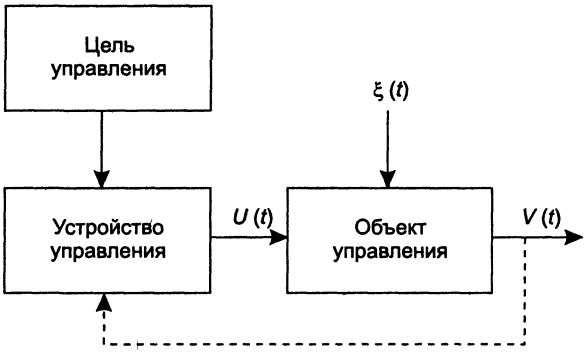

Рис. 1.1. Система управления

Объект управления и воздействующее на него устройство управления образуют *систему управления*. Предполагается, что на объект управления действуют также возмущения $\xi(t)$, изменяющие, как правило, в непредсказуемом направлении основные характеристики объекта управления.

Мы здесь для определенности будем предполагать, что все функции времени, изображенные на рис. 1.1, являются вектор-функциями с вещественными компонентами и заданными размерностями.

Сигнал управления вырабатывается в соответствии с некоторой заданной *целью управления*, определяемой теми задачами, которые поставлены перед системой управления. Довольно часто в системах управления для выработки управляющих воздействий оказывается необходимой информация о действительном состоянии объекта управления. Эта информация поступает по *цепи обратной связи*, показанной на рис. 1.1 пунктирной стрелкой.

В качестве примера объекта управления может выступать некоторый завод, перерабатывающий сырье в готовую продукцию. Здесь под управлением может пониматься вся система мероприятий и нововведений, направленная на достижение определенной цели или целей, связанных, например, с качеством и количеством выпускаемой продукции, а также, возможно, с определенными требованиями к срокам выпуска этой продукции.

Как справедливо указывается в книге [50], описание некоторого объекта как объекта управления и выявление каналов воздействия на него может производиться только исходя из заданной цели управления. Так, например, тот же завод с точки зрения служб санэпиднадзора или иных экологических служб является объектом управления, перерабатывающим сырье в промышленные отходы, загрязняющие окружающую среду. В этом случае управление со стороны этих служб будет направлено на снижение влияния вредных техногенных факторов, а вовсе не на интенсификацию производства, как в первом случае.

Основные задачи, решаемые большинством систем управления и отражающие главные цели управления, могут быть отнесены к одному из следующих типов: стабилизация, выполнение программы, слежение, экстремальное управление, оптимизация.

Задача стабилизации заключается в поддержании некоторых выходных (управляемых) характеристик объекта управления на заданных уровнях несмотря на постоянно действующие возмущения $\xi(t)$:

$$V(t) = \text{const}.$$

Задачи стабилизации возникают и решаются в живой природе (поддержание стабильной температуры тела у теплокровных животных), в технических системах (например, стабилизация напряжения и частоты в энергосистемах вне зависимости от нагрузки) и т. д.

Задача выполнения программы, или *задача программного управления*, возникает, когда необходимо обеспечить наперед заданные траектории $V(t)$. Иначе говоря, необходимо «заставить» объект управления изменять свои управляемые характеристики во времени по заданному закону, по определенной *программе* $V*(t)$. Например, процесс разработки некоторого технического изделия на заводе должен протекать в соответствии с заранее разработанным временным графиком. Метаморфоза насекомых в живой природе также происходит в соответствии с некоторыми биологическими программами развития. Запускаемый космический аппарат обычно выводится на траекторию в соответствии с заранее заданной расчетной траекторией. Легко видеть, что задача стабилизации является частным случаем задачи программного управления.

В *задачах слежения* основная проблема сводится к формированию такой выходной траектории $V(t)$ управляемого объекта, которая как можно более точно аппроксимировала бы другую, заранее не известную траекторию $V*(t)$. Например, при управлении пушечной турелью оператор (наводчик орудия) при наводке на цель вращает легкие маховички (задавая траекторию $V*(t)$), а тяжелая пушечная турель отслеживает движение маховичков, изменяя, ориентацию ствола по углу и по азимуту (реализуя соответствующую траекторию $V(t)$). Аналогичные задачи слежения возникают, когда антенна радиолокатора отслеживает непредвиденные движения какого-либо летящего объекта. Число подобных примеров легко может быть увеличено.

Задачи *экстремального управления*, или, как иногда говорят, *задачи настройки*, возникают довольно часто. Они заключаются в достижении некоторой экстремальной цели, которая к тому же может эволюционировать во времени. Предполагается, что на траекториях объекта управления задан некоторый функционал, отражающий эту цель (обычно его экстремум соответствует некоторому нормальному, благоприятному или наилучшему режиму работы) и зависящий как от управляемых, так и от неуправляемых параметров объекта. Требуется с помощью соответствующих управляющих воздействий добиваться того, чтобы значение *целевого функционала* в любой момент времени находилось в достаточно малой окрестности экстремума (максимума или минимума — в зависимости от смысловой интерпретации). Например, при настройке радиоприемника на какую-либо радиостанцию добиваются достижения максимальной громкости. Ав-

томатические системы экстремального управления часто называются *системами автоматической оптимизации*, или *самонастраивающимися* системами.

Задачи экстремального управления являются в определенном смысле более сложными, чем ранее перечисленные. Действительно, если, например, в задаче стабилизации достаточно одного измерения стабилизируемой величины для определения направления ее корректировки, то в данном случае для синтеза управляющего воздействия необходимо как минимум два *поисковых* измерения целевого функционала как основного выхода объекта. При настройке путем поиска — специальным образом организованных пробных движений — находится заранее не известное состояние объекта, соответствующее некоторому оптимальному режиму.

Важно отметить, что задачи экстремального управления являются не только более сложными, но и более общими. При необходимости практически любую задачу управления можно сформулировать на языке экстремального управления. Например, задачи стабилизации и программного управления могут быть сформулированы как задачи минимизации невязки (ошибки) между заданными и действительными выходными траекториями объекта. Другое дело — всегда ли оправданы такие формулировки? Это отдельный вопрос, и ответ на него однозначный — не всегда.

Когда мы говорим о *задачах оптимизации* в теории управления, то подразумеваем задачи реализации некоторых оптимальных (по заданному критерию) выходных траекторий управляемой системы. В частности, может ставиться задача перевода объекта управления из одной точки фазового пространства в другую, например, за минимальное время при соблюдении заданных ограничений, в том числе фазовых. Очевидно, такие задачи непосредственно не могут быть отнесены ни к одному из ранее рассмотренных типов задач (стабилизация, выполнение программы, слежение, настройка).

В качестве примера можно рассмотреть теперь уже ставший классическим пример задачи об оптимальном в смысле расхода горючего режиме набора высоты и скорости летательным аппаратом, например, самолетом [14]. На рис. 1.2 показано фазовое пространство самолета как объекта управления. Ось абсцисс является осью высот H, а ось ординат — осью скоростей v. Состояние самолета изображается точкой фазового пространства — плоскости $v0H$.

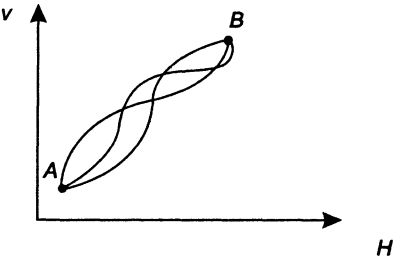

Рис. 1.2. Задача о наборе высоты и скорости самолетом

Существует множество траекторий в фазовом пространстве, соединяющих начальную (A) и конечную — *целевую* — точки (B). Требуется выбрать такое управление самолетом, чтобы набор высоты и скорости проходил в соответствии с некоторой оптимальной траекторией. Под критерием оптимальности в данном примере понимается суммарный расход горючего. Если такая оптимальная траектория рассчитана заранее, то она, в принципе, может выступать в качестве программы при последующем решении задачи программного управления. Однако обычно ситуация оказывается несколько сложнее. Может ставиться задача построения оптимального поведения объекта независимо от того, в какой точке фазового пространства он оказался в процессе реального движения в условиях возмущений. Собственно задача выполнения программы здесь уже не является определяющей.

Нужно ясно понимать, что все перечисленные задачи теории управления могут находиться в определенной иерархической взаимосвязи и присутствовать одновременно при проектировании той или иной системы управления. Например, на одном из иерархических уровней мы можем решать задачу стабилизации, строя соответствующую систему управления. В то же время (на более высоком уровне) может ставиться задача экстремального управления этой системой стабилизации в соответствии с некоторым критерием качества стабилизации и т. д. Число таких «вложений» теоретически не ограничено.

Еще раз отметим, что мы сейчас рассматривали основные задачи теории управления, а не методы их решения.

Методы решения сформулированных задач и их реализация в виде конкретных систем управления могут быть различными, и они связаны с основными *принципами управления.*

Жесткое управление. Принцип *жесткого (разомкнутого) управления* предполагает отсутствие обратной связи в общей схеме управления (см. рис. 1.1). Такие системы управления без обратной связи называются *разомкнутыми.* Чаще всего они применяются для целей программного управления.

Принцип жесткого управления проиллюстрирован на рис. 1.3, где представлена система жесткого управления, решающая задачу программного управления (задачу выполнения заданной программы).

Здесь цель управления $V*(t)$ задает желаемую программу изменения состояния объекта во времени. Эта программа передается управляющему устройству для построения необходимого управления $U(t)$. В результате такого управления состояние объекта должно изменяться по закону $V(t)$. Система управления стремится обеспечить равенство

$$V(t) \approx V*(t)$$

для любого момента времени. Рассмотрим данный процесс жесткого управления подробнее.

Рис. 1.3. Система жесткого программного управления

Предположим, что объект управления является безынерционным и описывается оператором F:

$$V(t) = F(U(t)).$$

Предполагаем, что данное равенство справедливо для любого момента времени t. Управляющее устройство вырабатывает управляющее воздействие по закону

$$U(t) = G(V*(t)),$$

где G — оператор устройства управления.

Очевидно, для достижения равенства

$$V(t) = F(U(t)) = F(G(V*(t))) = V*(t)$$

необходимо обеспечить закон управления G в виде

$$G = F^{-1}.$$

Таким образом, оператор, характеризующий закон управления, является обратным к оператору объекта.

Существенные моменты в процессе жесткого управления сводятся к следующему. Во-первых, для построения закона управления необходимо иметь полную информацию об операторе объекта (математической модели объекта). Во-вторых, предполагается стабильность характеристик объекта, то есть неизменность его оператора. При малейших изменениях в объекте точность жесткого управления может быть нарушена.

Примером системы жесткого управления является система радиосвязи, когда акустические колебания $V*(t)$, создаваемые, например, диктором в студии, поступают с микрофона в управляющее устройство, где преобразуются в радиоволны $U(t)$ (рис. 1.4).

После прохождения канала связи радиоволны попадают в приемник, детектируются, усиливаются и при помощи динамика преобразуются снова в акустические колебания. Здесь мы имеем факт управления на расстоянии колебаниями диффузора динамика таким образом, чтобы они копировали колебания в микрофонной цепи передатчика.

Таким образом, в системах жесткого управления управляющему устройству недоступна информация о действительном состоянии объекта управления — какая-либо обратная связь отсутствует. В гораздо большем числе случаев необходимо прибегать к более гибким принципам управления, оказывающимся более эффективными при наличии различных помех, возмущений и изменяющихся параметрах объекта управления.

Регулирование. В системах регулирования, в отличие от жесткого управления, управляющие воздействия строятся в зависимости от фактического текущего состояния объекта управления. Информация о состоянии объекта поступает по специально организованным каналам *обратной связи*. В этом случае говорят, что реализуется принцип *замкнутого управления*, а сама система управления называется *замкнутой*. Здесь так же, как и в предыдущем случае, для эффективного функционирования системы управления часто необходимо иметь модель (оператор) объекта управления. В противном случае построение управляющих воздей-

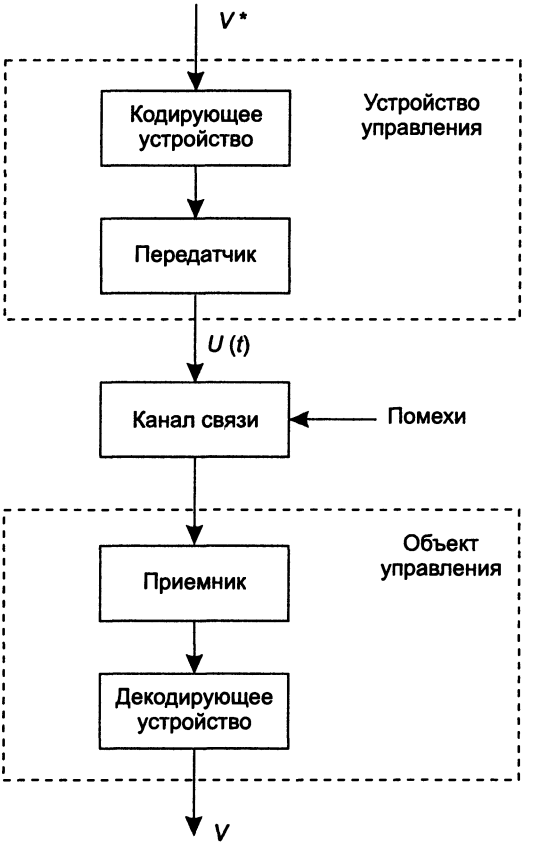

Рис. 1.4. Система радиосвязи

ствий оказывается невозможным или малоэффективным. Основное преимущество замкнутых систем перед разомкнутыми состоит в том, что они оказываются существенно менее зависимыми от неизмеримых возмущений и помех, особенно когда механизм влияния помех на объект управления неизвестен. Пример замкнутой системы программного управления представлен на рис. 1.5.

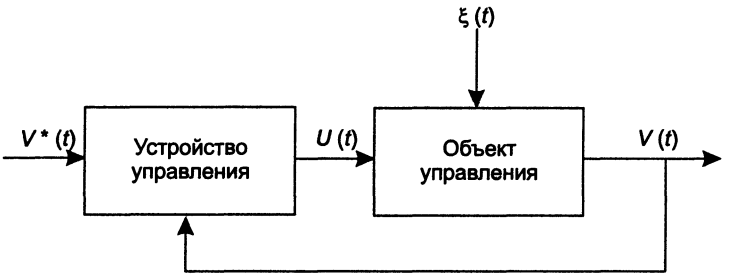

Рис. 1.5. Замкнутая система программного управления

В устройстве управления сравнивается желаемое $V*(t)$ и действительное значения функции $V(t)$. Это позволяет определить, насколько состояние объекта отличается от требуемого (задаваемого программой $V*(t)$). В результате управление строится как функция невязки:

$$U(t) = f\big(V(t) - V*(t)\big).$$

Наличие обратной связи обычно позволяет существенно расширить возможности управления и ослабить требования к знанию структуры и свойств объекта управления.

Отметим, что все сложности управления начинаются при попытке управления *сложным* объектом. Управление *простыми* объектами обычно не вызывает каких-либо проблем. Например, система управления подогревательными приборами в помещении (или в аквариуме) может быть основана на очень простом принципе: включить подогрев, если температура ниже заданной, и выключить его, если температура выше или равна заданной. В этом случае не нужна никакая модель объекта управления. Другое дело, что даже в этом простом примере нас могут интересовать возникающие при таком управлении колебания температуры. Для уменьшения этих колебаний мы можем начинать интересоваться мощностью нагревательного прибора, его инерционными свойствами, а также свойствами нагреваемой среды — в данном случае воды или воздуха. Например, существенным может оказаться учет нагреваемого объема воздуха или воды и т. д. Таким образом, нас уже интересует *модель объекта управления*.

Вообще, под *сложной системой* или объектом часто понимается система с *антиинтуитивным поведением*, то есть такая система, реакция которой на заданное входное воздействие или сигнал управления не может быть предсказана на основе «*здравого смысла*». Мы, следовательно, отличаем сложную систему от *большой системы*, характеризуемой, например, числовыми массивами или файлами высокой размерности. В этом смысле большая система может быть и простой по поведению.

Схема взаимодействия объекта управления с окружающей средой представлена на рис. 1.6.

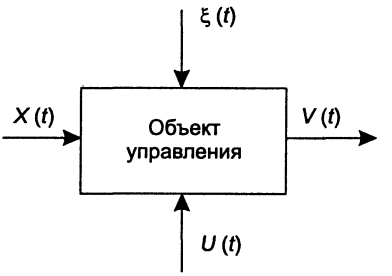

Рис. 1.6. Взаимодействие объекта со средой

Здесь появился новый канал связи объекта со средой — $X(t)$. Он означает неуправляемое, но наблюдаемое воздействие среды на объект. $\xi(t)$ по-прежнему

означает вектор неуправляемых и ненаблюдаемых возмущений (как внутренних, так и внешних). Под оператором объекта управления будем понимать оператор F^*, связывающий входы объекта с его выходом:

$$V = F^* (X, U, \xi).$$

Перечислим теперь характерные признаки сложности объекта управления. К ним в первую очередь относятся:

1. Отсутствие описания (в том числе алгоритмического) оператора F^*. В то же время, как уже говорилось, для целей управления сложным объектом оно совершенно необходимо.
2. Неожиданное, антиинтуитивное поведение объекта. Иногда это поведение моделируется и объясняется с помощью введения фактора *стохастичности*. Поэтому для построения систем управления сложными объектами привлекается аппарат статистической или стохастической теории управления.
3. Нестационарность (изменчивость во времени) оператора F^*.

1.2. Система управления сложным объектом

На рис. 1.7 представлена укрупненная алгоритмическая структура современной системы управления сложным объектом [53]. На вход объекта управления поступают вектор управляющих воздействий U, вектор возмущений ξ_1 и дополнительные «*идентифицирующие*» входные воздействия d. Измерительная система позволяет при наличии *измерительных шумов* ξ_2 измерять доступные (измеримые) характеристики состояния объекта управления (весьма часто сами переменные состояния оказываются непосредственно неизмеримыми). Сам процесс построения оценок V' переменных состояния по измеряемому выходу W реализуется с помощью алгоритма оценивания состояния. Информация о состоянии объекта далее используется для выработки управляющих воздействий, реализуя принцип замкнутого управления.

Алгоритм оценивания состояния настраивается с помощью соответствующего алгоритма оптимизации. Сам алгоритм оценивания состояния и алгоритм его настройки функционируют на основе оценок параметров P модели объекта управления, получаемых в соответствии с принятым алгоритмом идентификации (построения модели объекта управления). Алгоритм оптимизации устройства управления позволяет с помощью вектора параметров K организовать выбор оптимального алгоритма управления по полученным оценкам вектора состояний V' и по вектору параметров модели P. На алгоритм управления оказывает влияние алгоритм оптимизации самого объекта управления (оптимизации режимов функционирования объекта). Как правило, это влияние также носит параметрический характер — через вектор режимных параметров R. На вход алгоритма оптимизации объекта управления поступает информация о параметрах модели, полученной в результате идентификации.

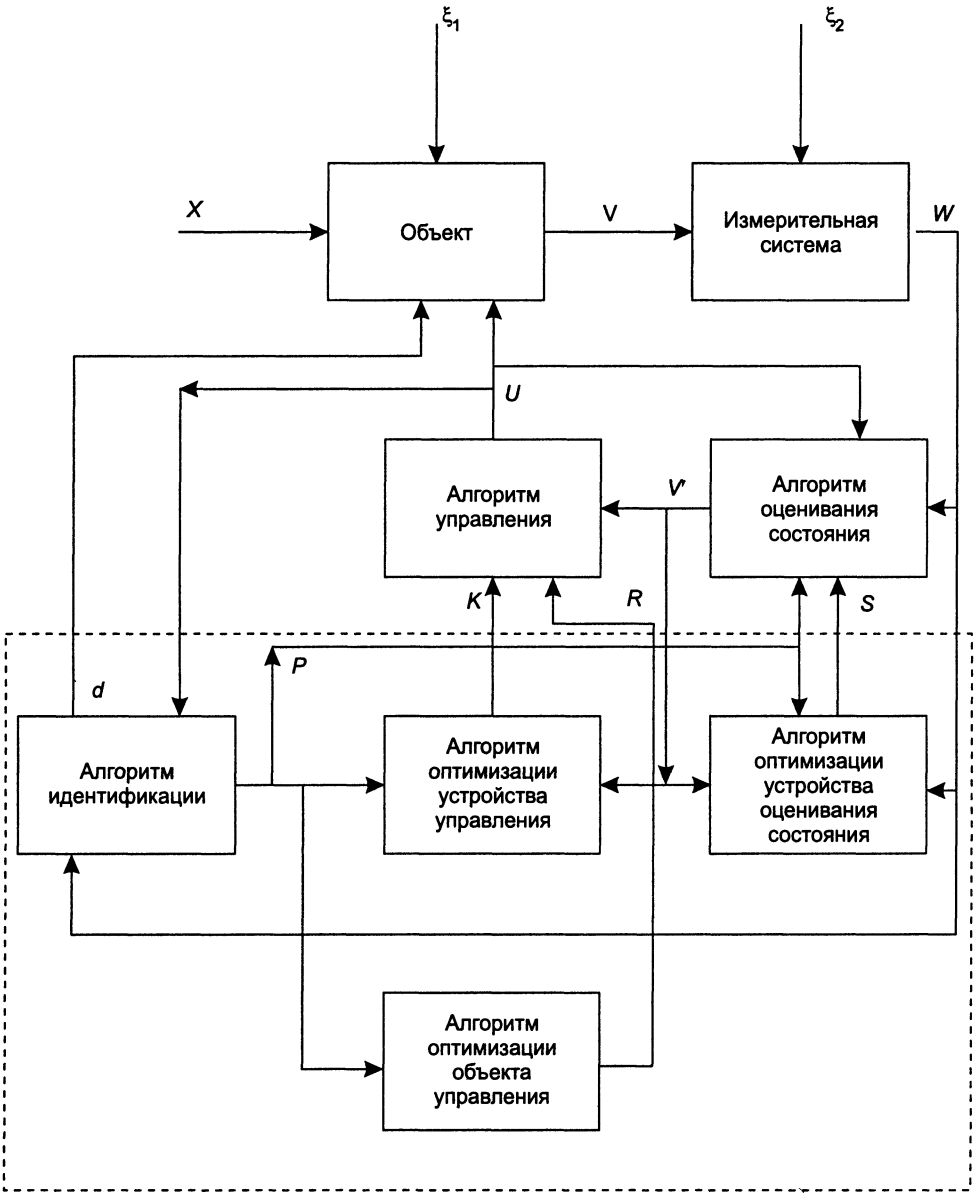

Рис. 1.7. Стратегия управления сложным объектом

Обведенная на рис. 1.7 пунктиром часть системы управления для некоторых постановок задач управления может функционировать вне контура управления с однократным или периодическим включением. Например, если параметры P объекта управления зависят от времени, то процедура идентификации должна периодически повторяться с соответствующей перенастройкой зависящих от этих параметров (от модели объекта) алгоритмов.

В большинстве современных систем управления, особенно автоматизированных, представленные на рис. 1.7 алгоритмы реализуются внутри соответствующих управляющих компьютеров и микропроцессоров со встроенным программным обеспечением. Однако совсем не исключен и другой, крайний вариант, когда приведенная схема управления сложным объектом полностью или частично будет реализовываться вообще «в ручном» режиме без использования вычислительной техники.

Совершенно ясно также, что в реальной схеме управления сложным объектом некоторые из представленных алгоритмов могут отсутствовать.

Опишем более подробно некоторые основные алгоритмы, реализующие систему управления сложным объектом.

Идентификация объектов управления. При решении задачи идентификации требуется определить наилучшую в некотором смысле модель объекта, описывающую соотношение между входными и выходными сигналами. Модель объекта необходима при реализации любого алгоритма управления сложным объектом, так как она позволяет предсказывать поведение объекта и определять наиболее эффективные управляющие воздействия с точки зрения целей управления.

Под моделью объекта управления понимается оператор F, связывающий состояние объекта V с его наблюдаемыми входами (см. рис. 1.6):

$$V = F(X, U).$$

Оператор модели, как правило, задается алгоритмически, то есть указывается правило, позволяющее по заданным входам определить выход без обращения к реальному объекту.

Следует отличать неизвестный оператор объекта F^* от оператора модели F. Ненаблюдаемые возмущения ξ при решении задачи идентификации оператора модели F рассматриваются как случайные помехи, затрудняющие процесс идентификации. Основная задача идентификации состоит в построении такого оператора модели F, который был бы в определенном смысле близок к оператору объекта F^*. При этом *близость* операторов оценивается исключительно по близости их реакций на одно и то же входное воздействие.

При построении оператора модели необходимо определить структуру S оператора F и вектор неизвестных параметров модели P:

$$F = \langle S, P \rangle.$$

Так, например, при построении модели динамического объекта может выбираться система обыкновенных дифференциальных уравнений в нормальной форме, правые части которой заданы с точностью до вектора неизвестных параметров P. Именно за счет подбора этих параметров производится «подгонка» этой модели под имеющиеся экспериментальные зависимости и тем самым обеспечивается близость реакций модели и реального объекта на идентичные входные воздействия.

Задача построения структуры S и параметров P оператора модели F называется *идентификацией в широком смысле*. Если структура оператора модели уже зада-

на и необходимо определить только вектор неизвестных параметров модели P, то имеем *задачу идентификации в узком смысле*, или задачу *параметрической идентификации*.

Весьма часто задача выбора структуры модели также может быть параметризована. Различные структуры могут кодироваться вектором структурных параметров D. Например, с помощью структурных параметров могут кодироваться порядок системы обыкновенных дифференциальных уравнений в приведенном ранее примере, а также вид и сложность правых частей уравнений.

Для выбора рациональных структур моделей объектов управления применяется также метод, основанный на использовании *избыточных топологических структур*. После задания избыточной структуры решается задача параметрической идентификации по имеющимся экспериментальным данным. В результате получаем сложную модель, достаточно точно описывающую поведение объекта. Затем решается *задача удаления переменных*, что позволяет приравнять к нулю некоторые из параметров и соответственно упростить структуру модели (например, из структуры модели удаляются элементы, отвечающие нулевым значениям параметров). Позже эти вопросы будут рассмотрены более подробно.

Таким образом, для задания оператора модели, вообще говоря, необходимо задавать две группы параметров:

$$F = \langle D, P \rangle.$$

Далее для простоты будем полагать, что структура оператора модели задана или уже определена, и решается задача параметрической идентификации. При рассмотрении конкретных задач теории управления мы укажем некоторые дополнительные подходы к решению задач структурной идентификации, основанные, в частности, на моделях Вольтерра.

Рассматривая задачи параметрической идентификации, будем предполагать, что оператор модели задан с точностью до вектора неизвестных параметров P:

$$V = F(X, U, P).$$

Различают два подхода к реализации процесса параметрической идентификации: пассивный и активный. *Пассивная идентификация* проводится в *режиме нормального функционирования* объекта управления — без оказания на него специальных идентифицирующих воздействий в виде специальным образом подобранных сигналов U и X (идентифицирующие, или, как иногда говорят, «раскачивающие», воздействия часто реализуются только по каналам управления). Идентифицирующие сигналы d показаны на рис. 1.7.

Предположим для простоты, что задача идентификации решается для построения будущей системы управления и объект пока не управляется. При этом модель объекта управления упрощается и принимает вид [51]

$$V = F'(X, P).$$

Управляемый вход U объекта здесь отсутствует.

Сама идентификация, то есть определение параметров P, осуществляется на основе информации о наблюдениях входов X и выходов V объекта в режиме нор-

мальной эксплуатации. После получения необходимой информации о поведении объекта формируется функция невязки ψ выходов модели и объекта. Например, в простейшем случае можно принять:

$$\psi(t,P)=\sum_{i=1}^{m}\left[V_i(t)-V_{iм}(t,P)\right]^2.$$

Здесь через $V_i(t)$ обозначена реакция реального объекта на заданное входное воздействие $X(t)$ по i-му выходу, а через $V_{iм}(t,P)$ — соответствующий (расчетный) выход модели (на вход модели поступает измеренный сигнал $X(t)$ — разумеется, его модельное представление). В данном выражении указаны i-е компоненты векторов V и $V_м$. Далее задача параметрической идентификации сводится к задаче поиска минимума некоторого целевого функционала, например, вида

$$J(P)=\sum_{k=1}^{N}\psi(t_k,P)\to\min.$$

Предполагается, что минимизируется сумма значений функции невязки на конечном множестве точек t_k. Минимизация целевого функционала осуществляется с помощью методов параметрической оптимизации, излагаемых в этой книге. В результате определяется искомый оптимальный вектор параметров P. Далее мы покажем, что сведение задачи идентификации к задаче минимизации функции многих переменных отнюдь не решает все проблемы идентификации. Хорошо известно, что возникают новые проблемы, например: как сформировать функции невязки и целевые функционалы в конкретном случае и как решить полученную задачу минимизации?

Здесь нам важно подчеркнуть, что в конечном итоге задача идентификации может быть сведена к задаче минимизации. Далее в основном тексте учебника мы более подробно рассмотрим важнейшие схемы параметрической идентификации и их реализации. Более подробно и полно проблема идентификации изложена в книге [52].

Обратимся теперь к *активной идентификации*. Активная идентификация предполагает подачу на вход объекта управления специальных идентифицирующих воздействий, вынуждающих объект управления *«проявить себя»* в максимальной степени. Такой подход по самой сути позволяет более эффективно по времени и точности получать оценки идентифицируемых параметров по сравнению с методом пассивной идентификации. При этом используются практически те же выражения для функций невязки и целевых функционалов. К сожалению, не все объекты допускают такое активное вмешательство в их работу, и это ограничивает область применимости метода активной идентификации.

Весьма часто как активная, так и пассивная идентификация осуществляется непрерывно, даже в процессе управления объектом («в контуре управления»). Это позволяет по мере поступления новой информации проводить последовательное уточнение параметров модели. В качестве таких уточняющих процедур в соответствующих разделах учебника будут описаны рекуррентный метод наименьших квадратов и алгоритм Качмажа.

Оценивание состояний объектов управления. Алгоритм оценивания состояния (см. рис. 1.7) по наблюдаемому (зашумленному) выходу объекта W и модели

объекта управления, построенной алгоритмом идентификации, строит наилучшую в некотором смысле оценку V' состояния V. В свою очередь, алгоритм оценивания состояния зависит от вектора параметров S, изменяя которые, мы можем влиять на эффективность алгоритма оценивания. Иногда задача оценивания состояния называется *задачей о наблюдении*. Объект называется *наблюдаемым*, если по измерениям выходного сигнала можно определить его состояние. Задача оценивания состояний является сложной самостоятельной проблемой, и ей посвящены многочисленные публикации.

Принято различать три типа оценок состояния динамического объекта: сглаживание, фильтрацию и прогноз.

При решении *задачи сглаживания* требуется построить оценку вектора состояния объекта в момент времени t по наблюдениям за выходом объекта вплоть до момента t', причем $t' > t$. Таким образом, состояние определяется с некоторым запаздыванием $(t' - t)$. В *задачах фильтрации* $t' = t$, а в *задачах прогноза* $t' < t$.

Рассмотрим в качестве примера задачу сглаживания.

Пусть движение некоторой динамической системы (объекта управления) определяется уравнением

$$\dot{z} = f(t,z),\ z(t_0) = x,\ t_0 \le t \le T,$$

где $z = (z_1, \ldots, z_y)$, — вектор переменных состояния, f — известная функция. Предполагается, что компоненты вектора состояния z непосредственно не измеряются. Измеряемым и непосредственно наблюдаемым является другой вектор $y = (y_1, \ldots, y_s)$, связанный с z линейной зависимостью

$$y(t) = H(t)z(t),$$

где $H(t)$ — известная прямоугольная матрица размерности $(s \times r)$. Матрица H характеризуется конструкцией измерительного устройства.

При сделанных предположениях траектория $z(t)$ системы однозначно определяется начальным вектором x, и решение представленной системы дифференциальных уравнений может быть записано в виде $z(t, x)$. Согласно методу наименьших квадратов вектор x может быть найден в результате минимизации следующего целевого функционала:

$$J(x) = 0{,}5 \sum_{j=1}^{N} \psi \left\| H(t_j)z(t_j,x) - y(t_j) \right\|^2 \to \min_x.$$

Очевидно, для однократного вычисления значения функционала при заданном x необходимо вначале проинтегрировать приведенную систему дифференциальных уравнений с начальным условием x. Далее полученная расчетная траектория в выбранных точках дискретности t_j сравнивается с соответствующими измеренными значениями $y(t_j)$. Полученная в соответствии с методом наименьших квадратов невязка и является значением целевого функционала в выбранной точке x. Минимизация целевого функционала может быть выполнена одним из представленных в этой книге методов параметрической оптимизации. Забегая вперед, уже здесь отметим, что поиск минимума полученного функционала с помощью традиционного «библиотечного» математического обеспечения может

быть сопряжен с весьма значительными вычислительными проблемами. Тем не менее сведение задачи оценивания состояния к задаче параметрической оптимизации является для нас принципиальным.

Алгоритмы оптимизации объектов управления. Часто объект управления и протекающие в нем процессы сами нуждаются в оптимизации. Например, в задачах программного управления возникает проблема построения наилучшей в каком-либо смысле программы поведения объекта. После построения такой оптимальной в смысле заданного критерия программы может ставиться задача о ее реализации, например, с помощью соответствующей системы регулирования или системы жесткого управления.

Сделаем некоторые выводы. Проведенное рассмотрение общей схемы управления сложным объектом показало, что при решении задач идентификации требуется определить *наилучшую* в некотором смысле модель объекта, описывающую соотношение между входными и выходными сигналами. Задача оценки состояния ставится как задача нахождения *наилучшей* в смысле заданного критерия оценки. Параметры устройства управления и параметры, определяющие режим поведения управляемого объекта, также обычно получаются в результате решения соответствующих оптимизационных задач. В результате устанавливается *оптимальный режим* протекания процессов в управляемом объекте и реализуется *оптимальная стратегия* поддержания заданного режима при наличии возмущающих воздействий.

В следующем разделе приведены более подробные схемы решения некоторых из указанных задач на основе аппарата параметрической (конечномерной) оптимизации.

1.3. Примеры задач параметрической оптимизации в теории управления

Под параметрической оптимизацией далее понимается процесс однократного достижения экстремальной цели в предположении стационарности экстремальной характеристики объекта оптимизации и конечномерности пространств входных и выходных параметров. При этом сам объект оптимизации может реально существовать либо представлять собой математическую модель. Как указано в 1.1 и 1.2 и как это будет продемонстрировано на примерах, приводимых ниже, методы и алгоритмы параметрической оптимизации играют важную роль в общем арсенале методов и средств расчета систем управления в традиционном понимании основных задач теории управления. С другой стороны, сам объект оптимизации можно рассматривать как статический объект оптимального управления с постоянными входными и выходными сигналами. В связи с этим задачи параметрической оптимизации, и в частности, задачи оптимального проектирования, часто изучаются в курсах теории управления в соответствующем контексте.

Рассматриваемые в книге объекты параметрической оптимизации далее будут характеризоваться совокупностью непрерывных параметров (предполагается, что ограничения типа дискретности отсутствуют), которые условно можно раз-

делить на три группы: входные (управляемые) x, внешние ξ и выходные y (рис. 1.8).

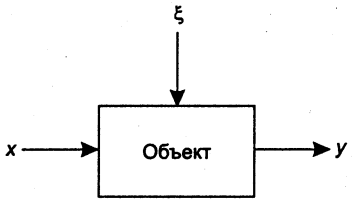

Рис. 1.8. Объект параметрической оптимизации

Входными параметрами $x = (x_1, x_2, \ldots, x_n)$ называются изменяемые в процессе оптимизации параметры, играющие роль управляющего воздействия при рассмотрении объекта оптимизации как объекта управления. *Внешние параметры* $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ характеризуют неопределенность обстановки. Отдельные компоненты вектора ξ могут иметь как случайный, так и неслучайный характер. Содержательный смысл введения внешних параметров состоит в описании следующих характерных явлений:

- влияния случайных отклонений при установке заданных (номинальных) значений x_i на реальном объекте;
- случайных воздействий внешней среды на объект, изменяющих реальные выходные характеристики по отношению к расчетным;
- влияний изменяющихся, как правило, не случайным, а заранее неизвестным образом, условий функционирования объекта, например таких, как температура, влажность, вибрации, уровень радиации и т. д.

Вектор *выходных параметров* $y = (y_1, \ldots, y_m)$ позволяет количественно оценить основные характеристики оптимизируемого объекта. С точки зрения общей теории управления, вектор y характеризует состояние объекта оптимизации. Далее предполагается, что существует функциональная связь

$$y = \varphi(x, \xi), \tag{1.1}$$

определяющая *оператор объекта оптимизации* и позволяющая по заданным x и ξ рассчитать соответствующий вектор y (задача *анализа*). Как правило, оператор объекта φ задан алгоритмически. Заметим, что мы здесь, вообще говоря, отличаем объект оптимизации от объекта управления в общей схеме управления, представленной в 1.2.

Задача параметрической оптимизации в общем случае ставится как многокритериальная задача с ограничениями:

$$y_i(x, \xi) \to \min_x, i \in [1{:}k], x \in D \subset R^n, \tag{1.2}$$

$$D = \{x \in R^n \mid g_i(x, \xi) \le 0, i \in [1{:}m], g_i(x, \xi) = 0, j \in [m+1{:}S]\}.$$

Множество $\{y_1, \ldots, y_k\}$ образует множество *критериальных* выходных параметров, имеющих смысл частных критериев оптимальности и характеризующих качество объекта оптимизации. Наличие нескольких частных критериев, по суще-

ству, отражает ту *неопределенность цели*, которая явно или неявно присутствует при оптимизации любого сколько-нибудь сложного объекта.

Интуитивный смысл задачи (1.2) состоит в выборе такого вектора x из допустимого множества D, чтобы каждый из критериальных выходных параметров принял по возможности меньшее значение. Математическое решение задачи (1.2) в виде конкретного вектора $x^* \in D$, вообще говоря, не существует, так как критериальные выходные параметры отражают противоречивые (конфликтные) требования к объекту оптимизации, и минимумы соответствующих функционалов достигаются в различных точках. Это предположение правомерно, так как в противном случае, если, например, две выходные функции y имеют минимумы в одной и той же точке, то одна из них может не рассматриваться. Как иногда шутят, «если два геолога говорят одно и то же, то один из них — лишний».

Допустимое множество D формируется на основе трех групп содержательно различных ограничений, имеющих вид равенств либо неравенств и задаваемых в (1.2) с помощью функций $g_i, i \in \{1:S\}$.

Прямые или *аргументные* ограничения накладываются непосредственно на компоненты вектора входных параметров:

$$a \le x_i \le b_i;\ a_i, b_i \in R^1. \tag{1.3}$$

В более общем случае границы интервалов могут быть функциями от других входных параметров:

$$a_i(x_j) \le x_i \le b_i(x_j);\ i \ne j. \tag{1.4}$$

Ограничения (1.3), (1.4) встречаются наиболее часто и вызываются причинами, связанными в основном с условиями физической реализуемости необходимых входных параметров.

Функциональные ограничения включают условия работоспособности, имеющие принципиальное значение при оценке правильности функционирования объекта оптимизации, исходя из его функционального назначения. Эти ограничения имеют вид

$$y_i \le t_i; t_i \in R^1; i \in [1:L]. \tag{1.5}$$

Например, к функциональным ограничениям относятся требования по устойчивости проектируемой системы автоматического управления, заключающиеся в необходимости расположения полюсов передаточной функции в левой полуплоскости комплексной плоскости.

Критериальные ограничения имеют вид

$$y_l \le t_l; t_l \in R^1; l \in [1:K]. \tag{1.6}$$

и отражают требования к характеристикам качества объекта оптимизации, подлежащие безоговорочному выполнению и приобретающие по существу характер функциональных ограничений.

Основное отличие функциональных ограничений от критериальных заключается в следующем. Как правило, выполнение неравенств (1.5) с большим запасом не

требуется; важно только гарантировать их выполнение. С другой стороны, в силу критериального характера выходных параметров y в (1.6) необходимо добиваться максимально возможных запасов при выполнении соответствующих неравенств.

Далее приводятся конкретные примеры, иллюстрирующие методологию применения аппарата параметрической оптимизации в некоторых традиционных задачах теории управления.

В основном рассмотрены задачи идентификации, а также задачи синтеза управляющих устройств при различных предположениях о целях управления и характеристиках входных воздействий.

При анализе и синтезе систем автоматического управления, как следует из общих курсов теории автоматического управления, часто необходимо учитывать стохастический (статистический) характер воздействий, приложенных к системе в различных точках. В некоторых подразделах этого раздела будем предполагать, что входные и выходные сигналы системы являются *случайными функциями*, изучаемыми в курсах теории вероятностей и математической статистики. Как известно, полной характеристикой случайной функции является ее закон распределения. Часто оказывается достаточным использование менее полных характеристик — математических ожиданий и корреляционных функций случайных воздействий.

В данном разделе при постановке некоторых задач мы ограничиваемся рассмотрением частных видов случайных функций — стационарных и стационарно связанных. Напомним, что случайная функция называется *стационарной*, если ее математическое ожидание постоянно (например, равно нулю), а корреляционные функции зависят только от разности своих двух аргументов (а не от их абсолютных значений). Две стационарные случайные функции называются *стационарно связанными*, если их взаимная корреляционная функция также зависит только от разности своих аргументов.

1.3.1. Идентификация нелинейных детерминированных объектов

В 1.3.1–1.3.3 рассматриваются некоторые варианты схем идентификации в *режиме нормальной работы* (то есть без пробных воздействий на объект). Все представленные далее стратегии идентификации имеют нерекуррентную форму, и поэтому собственно алгоритм идентификации включается в работу по истечении некоторого конечного интервала наблюдения за объектом. Такой выбор обусловлен двумя причинами. Во-первых, такие схемы широко распространены на практике и в целом ряде случаев оказываются численно более устойчивыми, чем рекуррентные процедуры. Во-вторых, указанный подход практически полностью базируется на аппарате конечномерной оптимизации.

Определение оптимальных параметров модели, имеющей заданную структуру

Пусть оператор модели F задан с точностью до вектора x неизвестных параметров

$$H_{\text{м}}(t) = F[G(t), x], \tag{1.7}$$

где $G(t)$ — входной сигнал модели и объекта; $H_м(t)$ — выходной сигнал модели; $H(t)$ — выходной сигнал объекта (рис. 1.9).

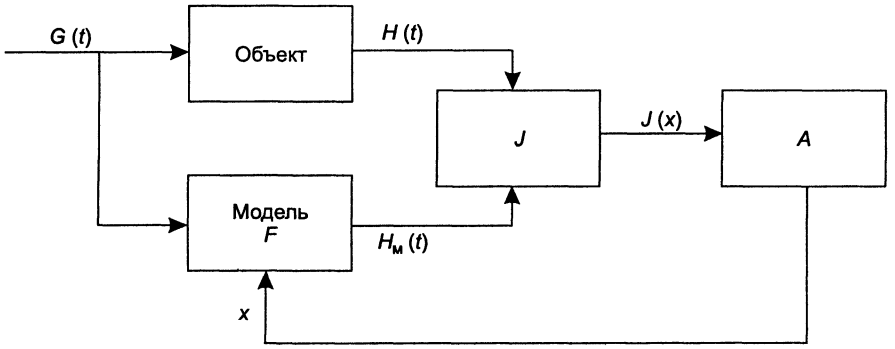

Рис. 1.9. Схема параметрической идентификации детерминированного объекта

Подстройка параметров модели осуществляется в дискретные моменты времени iT, $i \in [1:N]$, где $T = \tau_1 + \tau_2$; τ_1 — продолжительность одного цикла наблюдений за объектом; τ_2 — время обработки результатов наблюдений. В 1.3.1 и 1.3.2 предполагается, что в течение одного цикла настройки идентифицируемые параметры могут считаться постоянными. При этом сам объект может быть как стационарным, так и нестационарным (см. соответствующее пояснение в начале следующего пункта). Управляющее устройство A (реализованное аппаратно на базе соответствующего микропроцессора) по полученной за отрезок времени τ_1 информации вырабатывает (в течение отрезка времени τ_2) вектор x, исходя из условия минимума функционала невязки выходов модели и объекта:

$$J(x) = \int_0^{\tau_1} \|F[G(t),x] - H(t)\|^2 \, dt. \qquad (1.8)$$

В (1.8) отсчет времени t производится от начала i-го интервала наблюдений. Предполагается, что оператор модели задан алгоритмически. Алгоритм работы блока A основан на реализации принятой стратегии параметрической оптимизации.

Идентификация с использованием моделей Вольтерра

1. Пусть входной сигнал $G(t)$ нелинейного детерминированного объекта идентификации и выходной сигнал $H(t)$ являются одномерными, причем $G(t) = 0$ при $t < 0$ (см. рис. 1.9).

Рассмотрим *стационарный* случай, когда оператор объекта не зависит от времени. Для стационарных объектов реакция $H(t)$ не зависит от начала действия входного сигнала $G(t)$, а зависит только от интервала времени между началом действия $G(t)$ и данным моментом. При таких предположениях связь между входной и выходной переменными объекта может быть задана рядом Вольтерра [52]:

$$H(t) = \sum_{i=1}^{\infty} \int_0^{\infty} \dots \int_0^{\infty} \omega_i(\tau_1, \dots, \tau_i) \prod_{j=1}^{i} G(t - \tau_j) \, d\tau_1 \dots d\tau_i, \qquad (1.9)$$

где $\omega_i(\tau_1, \ldots, \tau_i)$ — обобщенные весовые функции i-го порядка ($i \in [1:l]$), подлежащие идентификации. На практике обычно ограничиваются отрезком ряда, содержащим члены первых двух порядков ($l = 2$). В этом случае при конечном интервале наблюдения имеем

$$H(t) = \int_0^T \omega_1(\tau) G(t-\tau) d\tau + \int_0^T \int_0^T \omega_2(\tau_1, \tau_2) G(t-\tau_1) G(t-\tau_2) d\tau_1 d\tau_2. \quad (1.10)$$

Параметризация задачи приводит к следующим разложениям:

$$\omega_1(t_1) = \sum_{v=1}^{N_1} c_{1v} \varphi_{v1}(t_1);$$
$$\omega_2(t_1, t_2) = \sum_{v=1}^{N_2} c_{2v} \varphi_{v2}(t_1, t_2), \quad (1.11)$$

где $\{\varphi_{v1}\}$, $\{\varphi_{v2}\}$ — наборы линейно-независимых опорных функций; $x = (c_{11}, \ldots, c_{1N_1}; c_{21}, \ldots, c_{2N_2})$ — вектор идентифицируемых параметров.

Интегральная невязка выходов модели и объекта имеет вид квадратичного функционала:

$$J(x) = \int_0^T \left[\sum_{v=1}^{N_1} c_{1v} y_{v1}(t) + \sum_{v=1}^{N_2} c_{2v} y_{v2}(t) - H(t) \right]^2 dt, \quad (1.12)$$

где коэффициенты

$$y_{v1}(t) = \int_0^T \varphi_{v1}(\tau) G(t-\tau) d\tau;$$
$$y_{v2}(t) = \int_0^T \int_0^T \varphi_{v2}(\tau_1, \tau_2) G(t-\tau_1) G(t-\tau_2) d\tau_1 d\tau_2$$

вычисляются вне цикла оптимизации. В данном случае схема процесса идентификации по существу совпадает с рис. 1.9, однако при этом структура модели непосредственно не задана, а формируется с помощью ряда Вольтерра.

При выборе опорных функций φ_{vi} необходимо учитывать, что ядра ω_i рядов Вольтерра для физически реализуемых операторов должны удовлетворять следующим требованиям:

1. $\omega_i(t_1, \ldots, t_i) = 0, \forall t_i < 0, i \in [1:S]$;
2. $\lim\limits_{t_i \to \infty} \omega_i(t_1, \ldots, t_i) = 0, \forall i \in [1:S]$. (1.13)

Дискретный вариант применения модели Вольтерра основан на представлении ряда Вольтерра в форме

$$H[n] = \sum_{i=0}^{\infty} \sum_{m_1=0}^{\infty} \ldots \sum_{m_i=0}^{\infty} \omega_i [m_1, \ldots, m_i] \prod_{j=1}^{i} G[n-m_j]. \quad (1.14)$$

Измерение значений входных и выходных сигналов осуществляется в дискретные моменты времени $t = n$ и приводит к построению решетчатых функций

$\{H[n]\}, \{G[n]\}, n \in [1{:}N_0]$. Ограничиваясь двумя членами ряда и используя параметризацию

$$\omega_1[m_1] = \sum_{v=1}^{N_1} c_{1v}\varphi_{v1}[m_1];$$
$$\omega_2[m_1, m_2] = \sum_{v=1}^{N_2} c_{2v}\varphi_{v2}[m_1, m_2], \tag{1.15}$$

где φ_{v1}, φ_{v2} — опорная система решетчатых функции, получим выражение, аналогичное (1.12):

$$J(x) = \sum_{n=1}^{N_0}\left[\sum_{v=1}^{N_1} c_{1v}y_{v1}[n] + \sum_{v=1}^{N_2} c_{2v}y_{v2}[n] - H[n]\right]^2, \tag{1.16}$$

где $x = (c_{11}, \ldots, c_{1N_1}; c_{21}, \ldots, c_{2N_2})$;

$$y_{v1}[n] = \sum_{m_1=0}^{N_0}\varphi_{v1}[m_1]G[n-m_1];$$

$$y_{v2}[n] = \sum_{m_1=0}^{N_0}\sum_{m_2=0}^{N_0}\varphi_{v2}[m_1, m_2]G[n-m_1]G[n-m_2].$$

Таким образом, при использовании моделей на основе ряда Вольтерра задача идентификации сводится к задаче параметрической оптимизации по квадратичным критериям типа (1.12) и (1.16). При этом ограничения на варьируемые параметры отсутствуют, и мы имеем задачу параметрической оптимизации без ограничений. Рассмотренные подходы естественным образом обобщаются на случай $l > 2$.

1.3.2. Идентификация стохастических объектов

Методы, основанные на процедурах сглаживания

Схема связи модели и объекта в рассматриваемом случае показана на рис. 1.10.

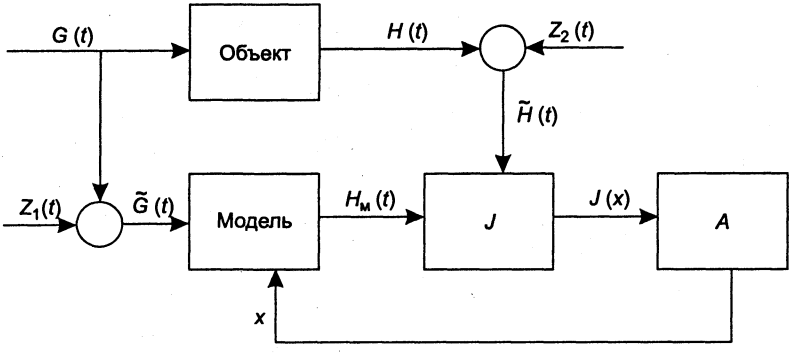

Рис. 1.10. Схема идентификации стохастического объекта

Предполагается, что стохастичность является следствием ошибок измерений входного $G(t)$ и выходного $H(t)$ сигналов детерминированного объекта. Пред-

полагается также, что помехи Z_1, Z_2 аддитивны, а рабочие сигналы $G(t)$, $H(t)$ регулярны и обладают определенными характеристиками гладкости (например, непрерывно дифференцируемы).

Исходная информация об идентифицируемом объекте О (вообще говоря, нелинейном) содержится в наблюдениях входа $\tilde{G}(t)$ и выхода $\tilde{H}(t)$, где

$$|\tilde{G}|(t) = G(t) + Z_1(t);$$
$$|\tilde{H}|(t) = H(t) + Z_2(t),\ t \in [0,T].$$

Рассматриваемая группа методов позволяет свести задачу идентификации стохастического объекта к детерминированному случаю [52]. Для этого входные и выходные сигналы $|\tilde{G}|(t)$, $|\tilde{H}|(t)$ подвергаются воздействию некоторого оператора усреднения. Классический подход связан с решением вспомогательных задач параметрической оптимизации вида

$$J_1[c^{(1)}] = \sum_{j=1}^{N_1}\left[\sum_{i=1}^{m_1} c_i^{(1)}\varphi_i^{(1)}(t_j) - |\tilde{G}|(t_j)\right]^2 \to \min_{c^{(1)}}; \tag{1.17}$$

$$J_2[c^{(2)}] = \sum_{k=1}^{N_2}\left[\sum_{i=1}^{m_2} c_i^{(2)}\varphi_i^{(2)}(t_k) - |\tilde{H}|(t_k)\right]^2 \to \min_{c^{(2)}}, \tag{1.18}$$

основанных на стандартном методе наименьших квадратов (МНК). Здесь $\{\varphi^{(1)}\}$, $\{\varphi^{(2)}\}$ — заданные системы координатных функций, обычно непрерывно дифференцируемых, что позволяет получить гладкие аппроксимации входных и выходных сигналов. Далее применяются процедуры идентификации детерминированных объектов на основе полученных «усредненных» зависимостей:

$$G(t) \cong \sum_{i=1}^{m_1} \hat{c}_i^{(1)}\varphi_i^{(1)}(t) = \left\langle \hat{c}^{(1)}, \varphi_i^{(1)}(t)\right\rangle; \tag{1.19}$$

$$H(t) \cong \sum_{i=1}^{m_2} \hat{c}_i^{(2)}\varphi_i^{(2)}(t) = \left\langle \hat{c}^{(2)}, \varphi_i^{(2)}(t)\right\rangle, \tag{1.20}$$

где $\hat{c}^{(1)}$, $\hat{c}^{(2)}$ — векторы, являющиеся решением задач (1.17), (1.18). Указанное приближение в среднем позволяет использовать избыточность информации для сглаживания случайных ошибок $Z_i(t)$. Как отмечается в работе [30], эффективность сглаживания критична к выбору координатных функций $\{\varphi^{(1)}\}$, $\{\varphi^{(2)}\}$, и эта критичность возрастает с ростом рабочего диапазона по t. Поэтому более эффективной может оказаться процедура разбиения области изменения независимой переменной на подобласти и построения для каждой из них своих аппроксимаций (1.19), (1.20). Если $\varphi^{(i)}(t)$ — полиномы, то это может соответствовать построению кусочно-полиномиальной аппроксимации. Другие стратегии усреднения, основанные на *методе локальной аппроксимации* (МЛА), представлены в [30]. В МЛА реализована идея использования «скользящего» интервала усреднения, определяемого как некоторая окрестность текущей точки t. Например, для входного сигнала $G(t)$, согласно МЛА, имеем:

$$G(t) \cong \langle c(t), \varphi(0) \rangle;$$
$$\hat{c}(t) = \arg\min_{c} J_N(t, c, \delta); \qquad (1.21)$$
$$J_N(t, c, \delta) = N^{-1} \sum_{i=1}^{N} \rho\left[\frac{t - t_i}{\delta}\right] F[|\tilde{G}|(t) - c^{\tau} \varphi(t - t_i)],$$

где $\{\varphi_i(t)\}$ — заданная система координатных функций; $F(\cdot)$ — неотрицательная функция потерь; $\rho(U)$ — функция локальности (например, $\exp[-|U|]$); δ — параметр усреднения, определяющий размеры локальной окрестности текущей точки t, в которой строится аппроксимирующая зависимость (1.21). Выбор параметра δ может быть основан на *методе перекрестного* (скользящего) *экзамена*, связанного с построением аппроксимаций, имеющих наилучшие интерполяционные свойства для узлов, не вошедших в базовый набор $\{t_1, t_2, \ldots, t_N\}$. МЛА по сравнению с классической и «кусочной» схемами МНК позволяет обеспечить большую точность аппроксимации при использовании небольшого количества «простых» координатных функций. Однако достигается это ценой увеличения сложности реализации и затрат необходимой памяти компьютера.

Корреляционные методы идентификации

Рассмотрим простейший случай линейного объекта со стационарным эргодическим по отношению к корреляционным функциям случайным входным сигналом $G(t)$ (рис. 1.11).

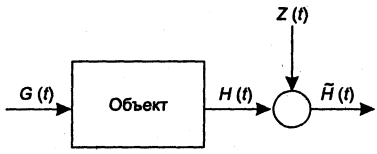

Рис. 1.11. Схема объекта идентификации

Помеха $Z(t)$ аддитивна:

$$|\tilde{H}|(t) = H(t) + Z(t), \qquad (1.22)$$

обладает указанным ранее свойством эргодичности и некоррелирована с входным сигналом. Оператор объекта предполагается детерминированным. Требуется построить оценку весовой функции объекта [52].

Предполагая, что все используемые реализации случайных процессов центрированы, получаем из (1.22):

$$R_{\tilde{H}G} = \lim_{T\to\infty} T^{-1} \int_0^T G(t-\tau) \left\{ \int_0^\infty \omega(\lambda) G(t-\lambda)\, d\lambda \right\} dt + R_{ZG}(\tau) =$$
$$= \int_0^\infty \omega(\lambda) \left\{ \lim_{T\to\infty} T^{-1} \int_0^T G(t-\tau) G(t-\lambda)\, dt \right\} d\lambda + R_{ZG}(\tau) = \qquad (1.23)$$
$$= \int_0^\infty \omega(\lambda) R_{GG}(\tau - \lambda)\, d\lambda + R_{ZG}(\tau).$$

Из некоррелированности случайных процессов $Z(t)$ и $G(t)$ имеем $R_{ZG}(\tau)=0$. Здесь с учетом эргодичности использованы следующие выражения для корреляционных функций:

$$R_{\tilde{H}G}(\tau)\triangleq\lim_{T\to\infty}T^{-1}\int_0^T\tilde{H}(t)G(t-\tau)\,dt;$$

$$R_{GG}(\tau)\triangleq\lim_{T\to\infty}T^{-1}\int_0^T G(t)G(t-\tau)\,dt; \qquad (1.24)$$

$$R_{ZG}(\tau)\triangleq\lim_{T\to\infty}T^{-1}\int_0^T Z(t)G(t-\tau)\,dt, \qquad (1.24)$$

а также предположение, что $G(t)=0,\forall t<0$.

Таким образом, для определения неизвестной весовой функции $\omega(t)$ линейного стационарного объекта построено интегральное уравнение вида:

$$\hat{R}_{\tilde{H}G}=\int_0^{T_\omega}\omega(\lambda)\hat{R}_{GG}(\tau-\lambda)\,d\lambda, \qquad (1.25)$$

известное как уравнение Винера—Хопфа. Здесь $\hat{R}_{\tilde{H}G}$, $\hat{R}_{GG}$ — оценки соответствующих корреляционных функций, полученные по ограниченному объему измерений; T_ω — рабочий промежуток времени, определяемый из условия $|\omega(t)|<\varepsilon\omega_{max}\ \forall t>T_\omega$, где, например, $\varepsilon=0{,}05$. Так как $\omega(t)$ первоначально неизвестна, то обоснованный выбор T_ω может быть произведен итеративно при решении нескольких задач (1.25) для различных T_ω. Аналогично выбираются рабочие промежутки интегрирования при получении оценок корреляционных функций.

Рассмотрим один из возможных способов решения интегрального уравнения (1.25). Перейдем в (1.25) к дискретному времени:

$$\hat{R}_{\tilde{H}G}[qj]=\sum_{i=1}^{N}\omega_i\hat{R}_{GG}[q(j-i)],\ 1\le j\le N, \qquad (1.26)$$

где q — шаг дискретности времени; N — число временных точек; ω_i — неизвестные значения весовой функции, $\omega_i\cong\omega[q_i]$. Функция невязки системы уравнений (1.26) имеет вид

$$J(x)=\sum_{j=1}^{N}\left[\sum_{i=1}^{N}\omega_i\hat{R}_{GG}[q(j-i)]-\hat{R}_{\tilde{H}G}[qj]\right]^2, \qquad (1.27)$$

где $x=(\omega_1,\omega_2,\ldots,\omega_N)$. Выбирая x из условия минимума (1.27), приходим к схеме идентификации, показанной на рис. 1.12, где БУ — блоки умножения; $\nearrow$ — блок переменной задержки; И — интеграторы.

Получили задачу параметрической оптимизации с квадратичным целевым функционалом без ограничений на вектор варьируемых параметров.

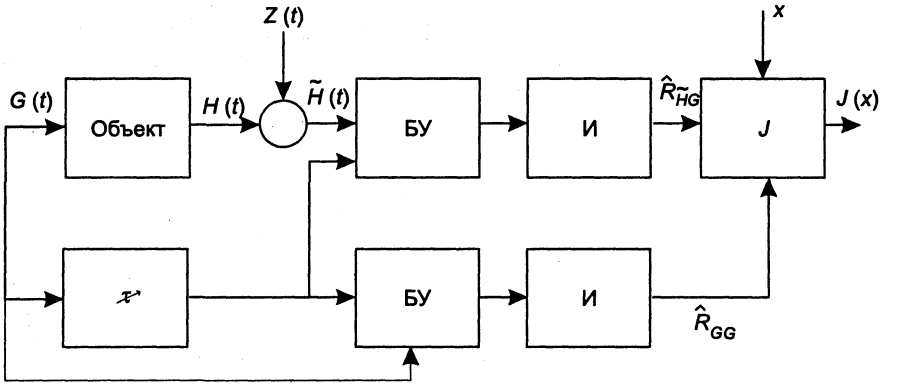

Рис. 1.12. Схема реализации корреляционного метода идентификации

1.3.3. Идентификация нестационарных объектов

Если объект нестационарен и его характеристики изменяются во времени достаточно интенсивно, то в течение одного цикла настройки параметры x уже нельзя считать постоянными. Однако в ряде случаев с помощью введения дополнительных переменных можно либо устранить, либо существенно ослабить указанный эффект.

Пусть оператор модели F задан с точностью до вектора $\alpha(t)$ неизвестных параметров:

$$H_{\text{м}}(t) = F[G(t), \alpha(t)]. \tag{1.28}$$

Задача идентификации сводится к задаче слежения за переменными параметрами объекта, отраженными в модели в виде вектор-функции

$$\alpha(t) = [\alpha_1(t), \alpha_2(t), \ldots, \alpha_m(t)]. \tag{1.29}$$

Параметризация $\alpha(t)$ приводит к следующему представлению

$$\alpha_i(t) = \sum_{j=1}^{N_i} c_{ij}\varphi_{ij}(t),\ i \in [1{:}m], \tag{1.30}$$

где $\varphi_{ij}(i \in [1{:}m]; j \in [1{:}k_m])$ — заданная система координатных функций. Подставляя (1.30) в (1.28), получим

$$H_{\text{м}}(t) = F_1[G(t), x],$$

где $x = (c_{11}, \ldots, c_{mk_m})$ — вектор неизвестных параметров. Существенно, что сам объект при этом не перестает быть нестационарным, как это иногда ошибочно отмечается.

1.3.4. Экстремальное регулирование

Постановка задачи экстремального регулирования предполагает нестационарность управляемого *экстремального* объекта (ЭО). Один из возможных общих подходов к организации процесса управления в указанных условиях связан с процедурой адаптации на основе *метода обучающейся модели* (рис. 1.13). Блок идентификации И в режиме нормальной работы подстраивает характеристики модели М под изменяющиеся во времени характеристики объекта. При этом каналы связи идентификатора и объекта, необходимые для собственно процесса идентификации, на рис. 1.13 не показаны. В результате строится модельная зависимость $I(x, t)$, аппроксимирующая на некотором отрезке времени реальную зависимость $J(x, t)$. В дискретные моменты времени τ_i (текущие или упреждающие) модель объекта фиксируется: $J_м(x) \triangleq I(x, \tau_i)$. Далее с помощью управляющего устройства УУ находится оптимальный режим из условия

$$x = \arg \min_{x_м \in D} J_м(x_м). \qquad (1.31)$$

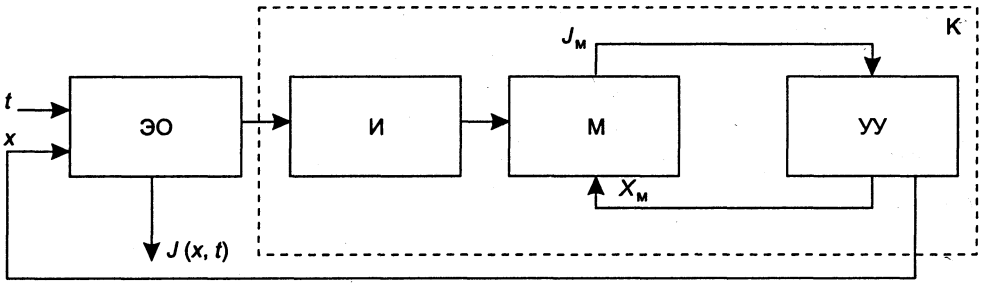

Рис. 1.13. Система экстремального регулирования (К — часть системы, реализованная в управляющем компьютере)

Процесс оптимизации протекает на модели в ускоренном масштабе времени. После завершения цикла оптимизации результаты переносятся на реальный объект.

Возможные временные характеристики указанной процедуры показаны на рис. 1.14, где $\Delta t_И$ означает время одного цикла идентификации, а $\Delta t_П$ — время поиска экстремума функционала $J_М$. В моменты времени τ_i происходит обновление модели за счет изменения вида функционала $J_М$. В точках t_i по результатам решения задачи (1.31) корректируется вектор x управляющих параметров. Предполагается, что в течение одного цикла $\Delta t_И$ идентифицируемые параметры с достаточной точностью можно считать постоянными (см. 1.3.3).

Возможны и другие формы проведения указанного процесса экстремального регулирования. Например, в ряде случаев новый цикл идентификации следует начинать непосредственно после предыдущего, не ожидая окончания процесса оптимизации (наложение интервалов $\Delta t_И$ и $\Delta t_П$). При достаточно медленном дрейфе экстремума функционала $J(x, t)$ процедуры подстройки модели и вектора управляющих параметров x могут включаться эпизодически по мере необходимости. Иногда целесообразно чаще повторять сеансы подстройки x по неизменной модели $I(x, t)$ и т. д. В любом случае для удовлетворительной работы

устройства в целом необходимо выполнение определенных соотношений между скоростью дрейфа экстремума (как по аргументу, так и по функционалу) и временными интервалами $\Delta t_{\text{И}}$, $\Delta t_{\text{П}}$. Эти соотношения определяются конкретной формой реализации изложенной процедуры и требованиями к максимально допустимой величине отклонения процесса от оптимального режима.

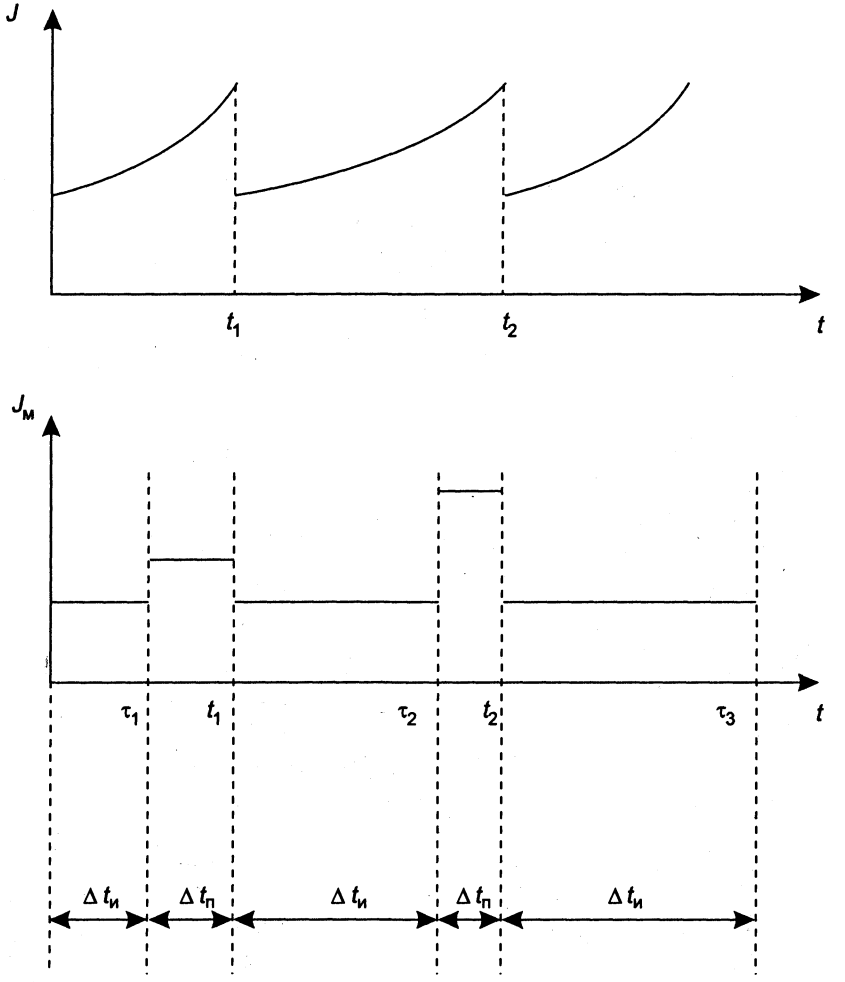

Рис. 1.14. Процесс экстремального регулирования с подстройкой управляющих параметров в дискретные моменты времени

В схеме управления (см. рис. 1.13) методы параметрической оптимизации используются дважды, определяя, с одной стороны, работу идентификатора И, а с другой — управляющего устройства УУ. В конечном счете, от эффективности соответствующих алгоритмов существенно зависят рабочие промежутки $\Delta t_{\text{И}}$, $\Delta t_{\text{П}}$, а следовательно, быстродействие и область применимости всей системы регулирования.

1.3.5. Синтез адаптивных систем автоматического управления

Изложенный в предыдущем разделе метод обучающейся модели может рассматриваться как некоторый общий *принцип адаптации*. Он допускает конкретные реализации, различающиеся способом формирования функционала $J_м(x)$. Далее рассмотрены возможные подходы к построению адаптивных систем автоматического управления, основанные на схемах с *эталонной моделью*.

Использование метрики в пространстве состояний

Данный подход можно пояснить схемой адаптации, изображенной на рис. 1.15.

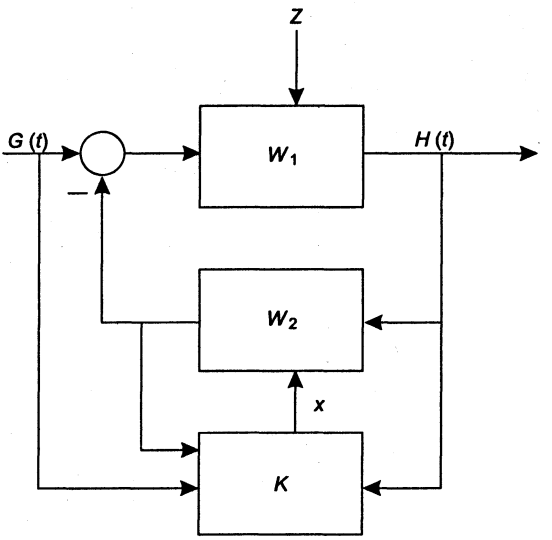

Рис. 1.15. Схема адаптации с корректирующим звеном в цепи обратной связи

Внешние условия воздействуют на передаточную функцию W_1. W_2 — передаточная функция корректирующего звена, которая подстраивается таким образом, чтобы скомпенсировать изменение W_1 под действием возмущения Z. Предполагается, что желаемые характеристики системы заданы в управляющем компьютере (К) в виде некоторой *эталонной модели* (в данном случае — цифровой). Идентификатор, реализованный в управляющем компьютере К, определяет неизвестные (дрейфующие во времени) параметры W_1. В результате получаем полную модель М системы управления, характеризуемую функциями W_1, W_2. Далее управляемые параметры x корректирующего звена W_2 определяются как решение задачи параметрической оптимизации вида

$$J_м(x) = \left\| H_э(t) - H_м(x,t) \right\|_t^2 \to \min_{x \in D}, \qquad (1.32)$$

где $H_э(t)$ — реакция эталона на заданный отрезок входного сигнала $G(t)$; $H_м(x,t)$ — реакция модели М на тот же отрезок входного сигнала; $\|\cdot\|_t$ — норма

в пространстве выходных сигналов $H(t)$. Полученные оптимальные значения x переносятся на реальный объект.

В данном случае близость эталона и реальной системы оценивается по *близости реакций* на одно и то же входное воздействие. При этом реальная структура системы автоматического управления может отличаться от структуры эталонной модели (последняя, например, может иметь более низкий порядок).

Использование метрики в пространстве параметров

В данном случае предполагается, что структура эталонной модели полностью соответствует структуре реальной системы управления с учетом корректирующего звена W_2. Процесс адаптации протекает аналогично описанию из предыдущего подраздела, однако вектор корректирующих параметров x выбирается из условия совпадения модельных коэффициентов $a_i(x)$ передаточной функции $W = \dfrac{W_1}{(1+W_1W_2)}$ реальной системы с соответствующими эталонными коэффициентами b_i:

$$J_{\text{м}}(x) = \sum_{i=0}^{N} [a_i(x) - b_i]^2 \to \min_{x \in D}. \qquad (1.33)$$

Множество D определяет множество физически реализуемых коэффициентов корректирующего звена W_2.

Пример. Пусть $W_1(p) = \dfrac{1}{p^2 + g_1 p + g_0}$, $W_2(p) = x_1 p + x_0$, причем помеха Z изменяет коэффициенты g_1, g_0 заранее неизвестным образом. Тогда на каждом шаге идентификации необходимо строить оценки $\hat{g}_1$, $\hat{g}_0$ и далее определять $x = (x_0, x_1)$ из условия (1.33), где $W = \dfrac{1}{p^2 + a_1 p + a_0}$, $a_0(x) = x_0 + g_0$, $a_1(x) = x_1 + g_1$. Если эталонная модель имеет достаточно простой вид, например, $W_{\text{Э}}(p) = \dfrac{1}{p^2 + b_1 p + b_0}$, то параметры корректирующего звена, доставляющие минимум функционалу $J_{\text{м}}(x)$, могут быть найдены непосредственно:

$$x_0 = b_0 - g_0, \quad x_1 = b_1 - g_1.$$

1.3.6. Синтез статистически оптимальных систем автоматического управления

Задача определения оптимальной весовой функции линейной стационарной системы автоматического управления

На рис. 1.16 представлена расчетная схема рассматриваемой задачи [2].

На вход оптимизируемой замкнутой линейной системы с весовой функцией $\omega(\tau)$ в установившемся режиме поступает сумма

$$\tilde{G}(t) = G(t) + Z(t)$$

полезного сигнала $G(t)$ и помехи $Z(t)$, являющихся стационарными и стационарно связанными случайными функциями. Перед системой ставится задача реализа-

ции заданного алгоритма преобразования входного сигнала с помощью идеальной весовой функции $\omega_{и}(\tau)$. Ошибка системы может быть представлена в виде

$$E(t) = H(t) - |\tilde{H}|(t).$$

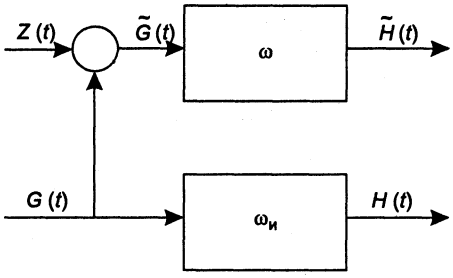

Рис. 1.16. Взаимодействие сигналов в задаче определения оптимальной весовой функции

Требуется найти весовую функцию $\omega_{*}(\tau)$ замкнутой системы таким образом, чтобы минимизировать *дисперсию ошибки*

$$D_E = M\{E^2(t)\}. \tag{1.34}$$

Используются следующие вероятностные характеристики входного сигнала и помехи: $M\{G(t)\} = M\{Z(t)\} = 0$; $R_{GG}(\tau)$, $R_{ZZ}(\tau)$, $R_{GZ}(\tau)$ — заданные корреляционные функции. Можно показать, что отсюда следуют равенства

$$R_{\tilde{G}\tilde{G}} = R_{GG} + 2R_{GZ} + R_{ZZ},\ M\{E(t)\} = 0.$$

Последнее соотношение определяет правомерность критерия (1.34). В курсах по теории автоматического управления показывается, что при выполнении некоторых дополнительных условий оптимальная в смысле критерия (1.34) весовая функция $\omega_{*}(\tau)$ удовлетворяет интегральному уравнению Винера—Хопфа

$$\int_0^{\infty} R_{\tilde{G}\tilde{G}}(\tau - \lambda)\,\omega_{*}(\lambda)\,d\lambda = R_{\tilde{G}H}(\tau),\quad \tau \geq 0. \tag{1.35}$$

Корреляционная функция $R_{\tilde{G}H}$ выражается через заданные корреляционные функции с учетом конкретной структуры идеальной системы. В частном случае, когда соответствующая $\omega_{и}$ передаточная функция $H_{и}(p) = 1$ (обычная следящая система при наличии помех), имеем

$$R_{\tilde{G}H}(\tau) = R_{\tilde{G}G}(\tau) = R_{GG}(\tau) + R_{GZ}(\tau).$$

Для решения уравнения Винера—Хопфа (1.35) можно воспользоваться методами из 1.3.2. В результате приходим к задаче минимизации квадратичного функционала типа (1.27).

Нахождение оптимальной весовой функции еще не означает, что реальная автоматическая система может быть выполнена оптимальной. Необходимо решить дополнительную, обычно оптимизационную, задачу *реализации* по определению реальной весовой функции, наименее уклоняющейся от найденной оптимальной.

Задача параметрической оптимизации стационарной линейной системы с заданной структурой

Пусть структурная схема замкнутой системы автоматического управления с передаточной функцией $W(p, x)$ задана с точностью до вектора $x = (x_1, x_2, \ldots, x_n)$ управляемых параметров. На систему, согласно рис. 1.17, воздействуют стационарные и стационарно-связанные полезный сигнал $G(t)$ и аддитивная помеха $Z(t)$. Через $W_И$ обозначена желаемая (идеальная) передаточная функция.

Требуется определить вектор $x \in D$ из условия минимума дисперсии стационарной ошибки [2]:

$$D_E = D_E(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\{|W(j\omega, x)|^2 S_{\tilde{G}\tilde{G}}(\omega) - \\ -2W(j\omega, x)W_И(-j\omega)S_{\tilde{G}G}(\omega) + |W_И(j\omega)|^2 S_{GG}(\omega)\}\,d\omega, \tag{1.36}$$

где $S_{\tilde{G}\tilde{G}}(\omega)$, $S_{GG}(\omega)$ — спектральные плотности сигналов $\tilde{G}$, G; $S_{\tilde{G}G}$ — взаимная спектральная плотность сигналов $|\tilde{G}|$, G.

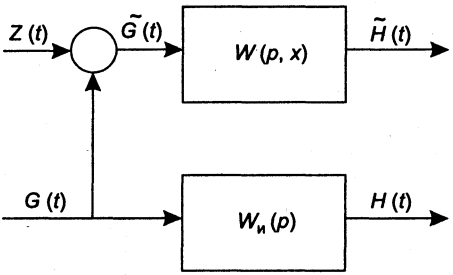

Рис. 1.17. Взаимодействие сигналов в задаче оптимизации системы с заданной структурой

Можно получить также выражение для D_E, используя вместо передаточных функций и спектральных плотностей весовые и корреляционные функции. Наиболее удобная форма представления D_E определяется условиями решения конкретной задачи.

Таким образом, задача построения статистически оптимальной системы с заданной структурой по критерию минимума дисперсии ошибки сводится к задаче минимизации функционала $D_E(x)$.

Задача синтеза оптимальной весовой функции линейной системы при нестационарных воздействиях

В предыдущих подразделах рассматривались стационарные воздействия на систему, что приводило к возникновению стационарных ошибок. Такие ситуации встречаются на практике. Однако более распространены случаи, когда входные воздействия нестационарны [2].

Пусть полезный сигнал и помеха являются нестационарными случайными функциями, причем полезный сигнал допускает следующую аппроксимацию:

$$G(t) = \sum_{i=0}^{m} G_i t^i + G'(t), \tag{1.37}$$

где $G'(t)$ — случайная функция времени с $M\{G'(t)\} = 0$; G_i — случайные величины с известными законами распределения.

В качестве критерия оптимальности принимается второй начальный момент ошибки

$$\Gamma_E(t_1, t_2) = M\{E(t_1)E(t_2)\}$$

при $t_1 = t_2 = t$. Таким образом, имеем

$$\Gamma_E(t) = M\{E(t)^2\},$$

где

$$E(t) = \int_0^T \left\{ \sum_{i=1}^{m} G_i \tau^i + G'(\tau) + Z(\tau) \right\} \omega(t, \tau)\, d\tau - H(t)$$

(согласно рис. 1.16). В данном случае в силу нестационарности искомая весовая функция ω зависит от двух аргументов. Кроме этого предполагается, что класс допустимых весовых функций ω ограничен функциями, равными нулю при $t < \tau$ (принцип причинности) и при $\tau > T$, $\tau < 0$, где T — время регулирования. Полагаем также, что $G(t) = 0$ при $t < 0$. Случайные величины G_i, $i \in [1:m]$ считаются не коррелированными между собой и со случайными функциями $G'(t)$, $Z(t)$.

Необходимое условие, которому удовлетворяет оптимальная весовая функция, имеет вид линейного интегрального уравнения Фредгольма первого рода

$$\int_0^T \left\{ \sum_{i=0}^{m} \alpha_i \tau^i \lambda^i + R_{X'X'}(\tau, \lambda) \right\} \omega(\tau, \lambda)\, d\lambda - R_{X'H}(\tau, t) = 0, \tag{1.38}$$

где $X'(t) \triangleq G(t) + Z(t)$; $\alpha_i \triangleq M\{G_i^2\}$; $R_{X'X'}$, $R_{X'H}$ — известные корреляционные функции.

Уравнение (1.38) при фиксированном t_j имеет вид (1.25) и его решение $\omega(t_j, \lambda)$, $0 \le \lambda \le T$, $j \in [1:N]$ может быть получено рассмотренными в 1.3.2 методами на основе алгоритмов параметрической оптимизации. Учет нестационарности приводит в данном случае к необходимости решения множества указанных задач параметрической оптимизации на сетке $\{t_j, j \in [1:N]\}$; $t_i \in [0, T]$, что выдвигает дополнительные требования к эффективности соответствующих оптимизирующих процедур.

1.3.7. Оптимальное проектирование систем

Под задачей проектирования понимается задача создания нового объекта или системы, обладающих заданными свойствами и характеристиками. Объектами проектирования могут быть технические системы, такие как компьютеры, компьютерные сети, программные комплексы, экономические системы, финансовые системы и т. д. Основные проблемы, возникающие при решении задачи проектирования, связаны с заданием структуры проектируемого объекта (*структурный*

синтез) и выбором параметров в рамках уже известной структуры (*параметрический синтез*). Предполагается, что проектируемый объект достаточно полно характеризуется некоторым вектором выходных параметров y, отражающих основные требования к создаваемому объекту.

Задача проектирования объекта или системы в целом ряде случаев может быть поставлена как задача решения системы неравенств — *спецификаций* вида

$$y_i(x, \xi) \le t_i, \quad t_i \in R^1, \quad i \in [1{:}m], \tag{1.39}$$

имеющих смысл *условий работоспособности* и составляющих основную часть *технического задания*. Здесь x означает неизвестный вектор *параметров проектирования*, подлежащих выбору. Вектор ξ характеризует наличие факторов неопределенности обстановки, например, влияние технологического разброса параметров при серийном выпуске изделий, а также влияние изменяющихся непредсказуемым образом условий функционирования объекта проектирования. Может быть поставлена задача *оптимального проектирования*, то есть задача построения вектора x из условия

$$x \in \arg\min_x J(x),$$

где, например, функционал $J(x)$ характеризует качество решения системы неравенств (1.39). Возможны и другие оптимизационные постановки задачи оптимального проектирования. Некоторые конкретные формы критериев оптимальности будут рассматриваться в 2.8.

В приведенной постановке объект проектирования может трактоваться как статический объект управления с управляемыми параметрами x, а сама проблема оптимального проектирования — изучаться в контексте основных задач теории управления. При этом оператор объекта задается алгоритмом вычисления выходных параметров y по входным параметрам x. Реализация такого алгоритма называется решением *задачи анализа* объекта проектирования. Обычно в процессе проектирования решается множество задач анализа для различных пробных значений входных параметров.

Упрощенная схема процесса оптимального проектирования показана на рис. 1.18

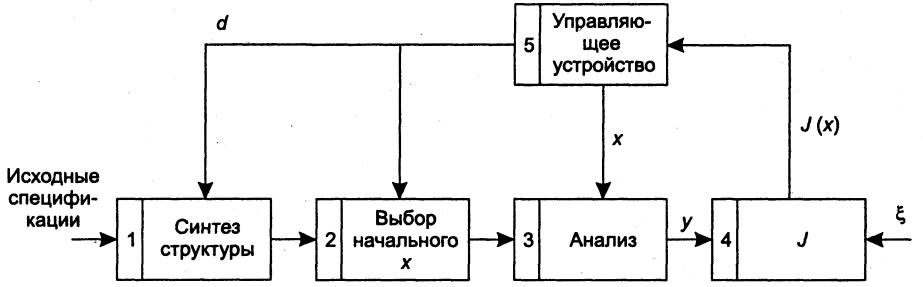

Рис. 1.18. Схема процесса оптимального проектирования

В блоке 1 синтезируется структура объекта с точностью до вектора *параметров проектирования* x.

Начальный вектор x выбирается в блоке 2 на основе имеющейся априорной информации. В блоке 3 реализуется *процедура анализа*, то есть вычисление y по заданному x. По вычисленному вектору y в критериальном блоке 4 формируется значение функционала $J(x)$, отражающего качество проекта, соответствующее текущему значению вектора x. В управляющем устройстве 5 реализован принятый алгоритм параметрической оптимизации. При этом предполагается, что влияние вектора ξ учтено в структуре функционала J, например, по методике, изложенной в 2.5.

Стрелка «*d*» означает возврат к оператору синтеза структуры при невозможности удовлетворить спецификациям с помощью текущей структуры проектируемого объекта. Кроме этого предполагается возможность корректировки спецификаций после исчерпания всего допустимого множества структур. На схеме не показан блок оценки получаемых результатов, позволяющий при выполнении соответствующих условий выйти из данного алгоритма. Как правило, выход означает получение решения задачи с приемлемой точностью, но может также означать невозможность решения задачи в рамках существующей системы оптимизации. Основные критерии остановки, применяемые в современных системах оптимизации, рассмотрены в книге [74].

Реальные процессы проектирования имеют сложную иерархическую структуру, состоящую из нескольких уровней проектирования [15]. На самом высшем уровне формируется *облик объекта*, то есть самые общие и основные требования-спецификации к проектируемому объекту в целом. Далее, на последующих уровнях происходит постепенная декомпозиция и детализация проекта. На каждом уровне реализуется представленная на рис. 1.18 схема или ей аналогичная (обычные усложнения связаны с многокритериальными постановками задач оптимального проектирования). При этом спецификации более низкого уровня формируются на основе результатов проектирования, полученных на соответствующем более высоком уровне.

1.4. Выводы

1. Современные методы расчета систем управления в значительной степени основываются на концепции *оптимальности*, что определяет широкое применение методов и алгоритмов теории оптимизации как при проектировании новых систем управления, так и при совершенствовании характеристик уже действующих объектов.
2. Большое число задач теории управления могут быть сформулированы как конечномерные оптимизационные задачи. К таким задачам, в частности, относятся:
   - задачи параметрической идентификации нелинейных детерминированных объектов;
   - задачи идентификации стохастических объектов;
   - задачи экстремального регулирования;
   - задачи синтеза адаптивных систем управления;
   - задачи синтеза статистически оптимальных систем управления;
   - задачи оптимального проектирования.

Важный раздел алгоритмического обеспечения современной теории управления объектами и системами различной физической природы составляют методы экстремизации (максимизации или минимизации) целевых функционалов, определенных в конечномерных векторных пространствах.

Глава 2. Математические модели теории параметрической оптимизации

2.1. Канонические задачи

Канонической задачей параметрической оптимизации далее называется задача однокритериальной оптимизации без возмущений и без ограничений на возможные значения управляемых (входных) параметров x:

$$J(x) \to \min_{x}, x \in R^n. \tag{2.1}$$

Такие задачи в курсах по теории оптимизации называются также задачами *безусловной оптимизации*, то есть задачами без дополнительных условий — без ограничений. В англоязычной научной и учебной литературе используется термин *unconstrained optimization*. Существуют книги, почти целиком посвященные решению таких задач, например учебник [24].

Задачи вида (2.1) довольно часто возникают в приложениях. Например, как было показано в 1.3.1, задача идентификации нелинейного детерминированного объекта на основе модели Вольтерра сводится к проблеме безусловной оптимизации. Однако обычно практическая ситуация оказывается более сложной, что определяется наличием многих частных критериев оптимальности, ограничений и неопределенных факторов. Кроме этого возникают существенные вычислительные проблемы, связанные с высокой размерностью n вектора x.

В данной главе обсуждаются некоторые методы сведения общей задачи параметрической оптимизации к конечному (или бесконечному) множеству задач (2.1). При этом мы следуем довольно традиционному в компьютерной математике подходу, связанному со сведением «сложных» задач к последовательности «простых». Например, существуют технологии, позволяющие сводить нелинейные задачи оптимзации к последовательности линейных, неквадратичные к последовательности квадратичных, задачи с ограничениями — к последовательности задач без ограничений, многокритериальные задачи — к некоторому набору одно-

критериальных задач, задачи с возмущениями — к последовательности задач без возмущений и т. д. На самом деле часто существуют иные, альтернативные методы решения. В частности, существует класс оптимизирующих процедур, специально ориентированных на работу в условиях ограничений. Однако если это возможно, везде в этой книге будет отдаваться предпочтение принципу сведения сложного к простому, сформулированному ранее.

Основной итог глав 1 и 2 должен состоять в обосновании важной роли канонических оптимизационных задач в алгоритмическом обеспечении современной теории управления.

2.2. Многокритериальные задачи

Когда мы говорим о многокритериальных задачах оптимизации, то имеем в виду довольно часто реализуемый случай *неопределенности цели*. В этом случае выбор вариантов осуществляется не по их оценкам с помощью единой целевой функции, а по целой группе оценок, находящихся в противоречии друг с другом. Например, задача покупки автомобиля явно или неявно ставится как задача выбора наилучшего, оптимального варианта из множества доступных. Если бы у нас были только один критерий и отражающий его целевой функционал (целевая функция), скажем, стоимость автомобиля, то в этом смысле проблемы бы не было. Имея конечное множество исходных вариантов, мы бы просто выбрали самый дешевый автомобиль. Но обычно ситуация осложняется тем, что при покупке нас начинают интересовать и другие показатели (оценки), такие как мощность двигателя, потребление топлива, цвет, год выпуска, пробег, модель автомобиля и т. д. При этом возникают и нечисловые характеристики, такие как цвет и модель автомобиля. При необходимости таким показателям тоже можно дать соответствующие числовые интерпретации, например, проранжировав все имеющиеся цвета с помощью принятых балльных оценок. Более подробный анализ нечисловых критериев выходит за рамки нашего рассмотрения.

Общая детерминированная многокритериальная задача параметрической оптимизации с ограничениями формулируется следующим образом:

$$f_i(x) \to \min_x,\ x \in D,\ i \in [1:k]; \tag{2.2}$$

$$D = \left\{ x \in R^n \,\middle|\, g_i(x) \le 0,\ i \in [1:m];\ g_i(x) = 0,\ i \in [m+1:s];\ a_i \le x_i \le b_i,\ j \in [1:q] \right\}.$$

Множество D называется множеством допустимых решений. В пределах этого множества выполняются прямые, функциональные и критериальные ограничения, представленные в виде общей системы неравенств и равенств. Прямые (*интервальные*) ограничения обычно указываются в явном виде, так как их учет может производиться отдельно от других типов ограничений, имеющих более сложную структуру.

Предполагается, что каждый из критериальных выходных параметров f_i необходимо минимизировать. Это не ограничивает общности, так как максимизация функции $\varphi(x)$ эквивалентна минимизации $f(x) = -\varphi(x)$. Кроме того мы учитываем здесь, что замена знака у левых частей неравенств $p(x) \ge 0$ меняет знаки не-

равенств на противоположные и приводит их к стандартному виду $g(x) \le 0$, где $g(x) = -p(x)$.

Задача (2.2) не является стандартной с точки зрения традиционных методов нелинейной оптимизации, главным образом, из-за наличия векторного критерия оптимальности. Поэтому приобретают важное значение различные приемы ее сведения к более удобным конструкциям, допускающим эффективное численное решение обычными средствами. Такое сведение не будет однозначным и обычно вызывает известные трудности.

Характерные практические ситуации, приводящие к многокритериальности в задачах управления, могут быть изучены по книгам [50], [25].

Существуют различные способы сведения исходной многоцелевой задачи (2.2) к задачам с единым критерием. Формулирование скалярного критерия оптимальности (целевой функции) должно производиться исходя из списка выходных параметров, имеющих смысл частных критериев оптимальности. Между тем, многие из критериев являются противоречивыми; улучшение одного из них при изменении вектора управляемых параметров приводит к ухудшению другого. Возникает проблема выбора разумного компромисса, то есть определения такого допустимого вектора управляемых параметров $x_1^*, \ldots, x_n^*$, при котором все критериальные параметры будут принимать приемлемые значения. Фактор противоречивости предъявляемых к объекту оптимизации требований, с одной стороны, значительно затрудняет формальный подход к формированию единой целевой функции и требует привлечения различных неформальных процедур, а с другой — приводит к плохо обусловленным оптимизационным задачам для построенных скалярных критериев качества.

Далее рассмотрены наиболее употребительные в практике компьютерного моделирования методы скаляризации векторного критерия оптимальности [74].

В *методе главного критерия* в качестве целевого функционала выбирается один из критериальных выходных параметров, наиболее полно с точки зрения исследователя отражающий цели оптимизации. Остальные частные критерии оптимальности учитываются с помощью введения необходимых критериальных ограничений, определяющих совместно с прямыми и функциональными ограничениями допустимое множество D. Основные трудности такого подхода связаны с проблемой назначения критериальных ограничений. Кроме того, в большом числе случаев всегда есть несколько главных критериев, находящихся в противоречии друг с другом.

Наиболее простой и часто применяемый метод формирования единой целевой функции основан на *линейной свертке* всех частных критериев в один:

$$J(x) = \sum_{i=1}^{k} \alpha_i f_i(x) \to \min_{x \in D}, \ \alpha_i > 0, \ \sum_{i=1}^{k} \alpha_i = 1. \qquad (2.3)$$

Весовые коэффициенты α_i могут при этом рассматриваться как показатели относительной значимости отдельных критериев f_i. Они характеризуют чувствительность целевого функционала $J(x)$ к изменению частных критериев: $\frac{\partial J}{\partial f_i} = \alpha_i, i \in [1:k]$. При наличии существенно разнохарактерных частных крите-

риев обычно бывает довольно сложно указать окончательный набор коэффициентов α_i, исходя из неформальных соображений, связанных, как правило, с результатами экспертного анализа.

Заметим, что иногда применяемые в предположении $f_i > 0$ мультипликативные критерии

$$J(x) = \prod_{i=1}^{k} f_i^{\alpha_i}(x) \to \min_{x \in D} \tag{2.4}$$

принципиально ничем не отличаются от конструкции (2.3), так как от функционалов $J, f_1, \ldots, f_k$ можно перейти к их логарифмам, и тогда (2.4) примет вид

$$\ln J(x) = \sum_{i=1}^{k} \alpha_i \ln f_i(x).$$

Использование *минимаксных* целевых функционалов обычно связано с введением *контрольных показателей* $t_1, \ldots, t_k$, фигурирующих в правых частях критериальных ограничений:

$$f_i(x) \le t_i. \tag{2.5}$$

В этом случае в качестве скалярного критерия можно использовать условия вида

$$J(x) = \min_i \alpha_i (t_i - f_i(x)) \to \max_{x \in D} \tag{2.6}$$

или

$$J(x) = \max_i \alpha_i (f_i(x) - t_i) \to \min_{x \in D}, \tag{2.7}$$

где D задается списком функциональных и прямых ограничений.

Для задания контрольных показателей применяется экспертный анализ либо производится их вычисление с помощью решения однокритериальных задач:

$$t_i = \min_{x \in D} f_i(x).$$

В последнем случае набор чисел t_i характеризует предельные, вообще говоря, недостижимые возможности по каждому из критериальных выходных параметров. В ряде случаев подбор t_i целесообразно определять в интерактивном режиме работы соответствующей программной системы. Весовые коэффициенты α_i выполняют в этом случае функции нормирования частных целевых функционалов по значению.

2.3. Компромиссы Парето

Пусть решается задача

$$f_i(x) \to \min_{x \in D}, \; i \in [1{:}k], \; D \subset R^n. \tag{2.8}$$

Рассмотрим две точки x', $x'' \in D$. Если выполняются неравенства

$$f_i(x') \le f_i(x'') \tag{2.9}$$

для всех $i \in [1:k]$, причем по крайней мере одно из неравенств строгое, то будем говорить, что точка x' *предпочтительнее*, чем x''. Если для некоторой точки $x^0 \in D$ не существует более предпочтительных точек, то будем называть x^0 *эффективным*, или *Парето-оптимальным*, *решением* многокритериальной задачи (2.8). Множество, включающее все эффективные решения, обозначается $P(D)$ и называется множеством Парето для векторного отображения $f: D \to R^n, f = (f_1, \ldots, f_k)$. Очевидно, $P(D) \subset D \subset R^n$. Образ множества $P(D)$ в пространстве критериев R^k обозначается $P(f)$. Множество $P(f) = f(P(D))$ называется *множеством эффективных оценок* [47].

Смысл введенного понятия эффективного решения состоит в том, что оптимальное решение многокритериальной задачи следует искать только среди множества $P(D)$ (*принцип Парето*). В противном случае всегда найдется точка x, оказывающаяся более предпочтительной, независимо от расстановки приоритетов и относительной важности отдельных частных критериев.

Точка $x' \in D$ называется *слабо эффективным* решением задачи (2.8), если не существует такой точки $x'' \in D$, для которой выполняются строгие неравенства $f_i(x'') < f_i(x'), i \in [1:k]$. Иначе говоря, решение называется слабо эффективным, если оно не может быть улучшено сразу по всем критериям. Множество слабо эффективных решений будет обозначаться через $S(D)$. Очевидно, $P(D) \subset S(D)$. Аналогично полагаем $S(f) \triangleq f(S(D))$.

Упражнение. Докажите включение $P(D) \subset S(D)$.

Введение понятия слабо эффективного решения вызвано тем, что в процессе оптимизации часто получаются именно эти решения, обычно представляющие с точки зрения практики меньший интерес, чем эффективные решения. С другой стороны, понятие слабо эффективного решения может играть важную роль при выборе набора «существенных» критериев [47]. Действительно, нетрудно доказать, что решение, которое слабо эффективно по сокращенному набору критериев, будет слабо эффективным и по расширенному набору.

Упражнение. Докажите, что этот вывод несправедлив для эффективных решений. Приведите «опровергающий» пример.

Поэтому после построения полного набора критериев следует выделить именно слабо эффективные решения, среди которых будут находиться и искомые решения, эффективные по окончательному (может быть, сокращенному) набору. На рис. 2.1 $P(f) = [b, c]$; $S(f) = [a, b] \cup [b, c] \cup [c, d]$. Данное замечание оказывается важным при первоначальной формализации реальной задачи и выборе рационального набора критериев оптимальности. Обычно вначале приходится «на всякий случай» учитывать большее число частных целевых функционалов, чем это в действительности окажется необходимым.

Схемы построения эффективных и слабо эффективных решений могут быть основаны на следующих теоремах.

Теорема 2.1. Пусть заданы произвольные $a_i > 0, i \in [1:k]$, тогда решение задачи

$$\min_i \alpha_i (t_i - f_i) \to \max_{x \in D} \tag{2.10}$$

при любых фиксированных t_i есть слабо эффективный вектор. Наоборот, любой слабо эффективный вектор может быть получен как решение задачи (2.10) при некоторых $\alpha_i > 0$ и $t_i > f_i(x)$, $i \in [1:k]$, $x \in D$.

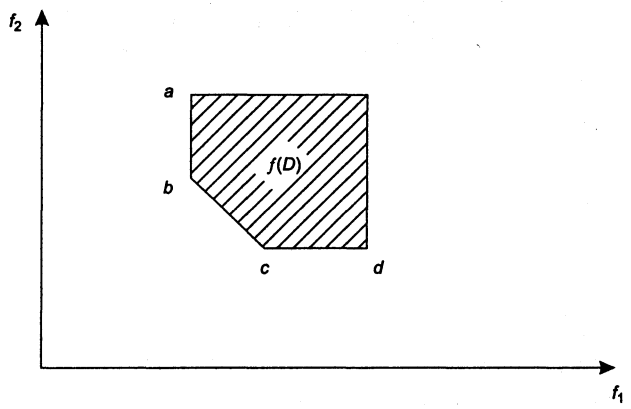

Рис. 2.1. Эффективные и слабо эффективные оценки задачи $f_i \to \min$, $i = 1, 2$

Доказательство. Пусть x^0 есть решение задачи (2.10) и существует $x' \in D$ такой, что $f_i(x') < f_i(x^0)$, $i \in [1:k]$. Тогда $\min\limits_i \alpha_i(t_i - f_i(x')) > \min\limits_i \alpha_i(t_i - f_i(x^0))$, что противоречит свойству x^0. Наоборот, пусть $x^0 \in S(D)$; тогда, согласно условию теоремы, $t_i - f_i(x^0) > 0$. Примем $\alpha_i' = 1/(t_i - f_i(x^0))$ и покажем, что

$$\max_{x \in D} \min_i \alpha_i'(t_i - f_i(x)) = \min_i \alpha_i'(t_i - f_i(x^0)) = 1. \tag{2.11}$$

Из $x^0 \in S(D)$ следует, что для любого $x' \in D$ найдется хотя бы один номер i^0, для которого $f_{i_0}(x') > f_{i_0}(x^0)$, а значит

$$\alpha_{i_0}'(t_{i_0} - f_{i_0}(x')) \le \alpha_{i_0}'(t_{i_0} - f_{i_0}(x^0)) = 1.$$

Но тогда и $\min\limits_i \alpha_i'(t_i - f_i(x')) \le 1$, что означает выполнение (2.11), так как x' — любой элемент D. Теорема доказана.

Замечание 1. Обратное утверждение теоремы 1 справедливо в следующей формулировке: любой слабо эффективный вектор x^0 может быть получен как решение задачи (2.10) при $t_i > f_i(x)$,

$$\alpha_i > 0, \sum_{i=1}^{k} \alpha_i = 1. \tag{2.12}$$

Действительно, если $x = \arg\max\limits_x \varphi(x)$, то $x = \arg\max \beta\varphi(x)$, если $\beta > 0$. Полагая $\varphi(x) = \min\limits_i \alpha_i^0(t_i - f_i)$, $x^0 = \arg\max\limits_{x \in D} \min\limits_i \alpha_i^0(t_i - f_i)$, $\beta = \dfrac{1}{\sum\limits_{i=1}^{k} \alpha_i^0}$, получим, что

$$x^0 = \arg\max_{x \in D} \min_i \alpha_i'(t_i - f_i),\ \alpha_i' > 0,\ \sum_{i=1}^{k} \alpha_i' = 1.$$

Замечание 2. Из доказательства второй части теоремы 1 следует, что для любого слабо эффективного вектора x^0 могут быть выбраны такие $\alpha_i > 0$, для которых он получается как решение задачи (2.10), причем

$$\alpha_1(t_1 - f_1(x^0)) = \alpha_2(t_2 - f_2(x^0)) = \ldots = \alpha_k(t_k - f_k(x^0)). \qquad (2.13)$$

Теорема 2.2. Пусть $\alpha_i > 0$, $i \in [1:k]$; D' — произвольное подмножество D, содержащее хотя бы один эффективный вектор, тогда решение задачи

$$\sum_{i=1}^{k} \alpha_i f_i(x) \to \min_{x \in D'} \qquad (2.14)$$

есть эффективный вектор.

Доказательство. Пусть x^0 есть решение задачи (2.14) и существует $x' \in D'$ такой, что $f_i(x') \le f_i(x^0)$, а для $i = i^0$ имеем $f_{i_0}(x') < f_{i_0}(x^0)$. Тогда

$$\sum_{i=1}^{k} \alpha_i f_i(x') < \sum_{i=1}^{k} \alpha_i f_i(x^0),$$

что противоречит свойству вектора x^0.

Замечание. Обратное утверждение без дополнительных предположений неверно. Существуют эффективные векторы, не являющиеся решением задачи (2.14).

Теорема 2.3. Пусть T_α — непустое множество решений задачи (2.10). Тогда существует вектор $x^* \in T_\alpha$, являющийся эффективным.

Доказательство. Допустим, что существует такой $x' \in D$, что $f_1(x') \le f_1(x^0)$, $i \in [1:k]$, где x^0 — произвольный элемент множества T_α. Тогда $\min\limits_i \alpha_i(t_i - f_i(x')) \ge$ $\ge \min\limits_i \alpha_i(t_i - f_i(x^0)) = \max\limits_{x \in D} \min\limits_i \alpha_i(t_i - f_i(x))$, то есть $x' \in T_\alpha$ и, следовательно, вне множества T_α не может существовать решений x', для которых $f_i(x') \le$ $\le f_i(x^0)$, $i \in [1:k]$. Теорема доказана.

Следствие. Если множество T_α содержит единственный вектор, то он эффективен.

По схеме теоремы 1 могут быть получены все слабо эффективные решения, если вектор $\alpha = (\alpha_1, \ldots, \alpha_k)$ пробегает все множество векторов (2.12). Напротив, вычислительная схема (2.14) позволяет получать только эффективные решения, но не все. А. Джоффриону принадлежит идея выделения эффективных точек из множества $S(D)$, построенного согласно (2.10) с помощью процедуры (2.14). Справедливо следующее утверждение [18].

Теорема 2.4. Пусть $\alpha_i > 0$, $i \in [1:k]$, тогда решение задачи

$$\sum_{i=1}^{k} f_i(x) \to \min_{x' \in T_\alpha}; \qquad (2.15)$$
$$T_\alpha = \{x' \in D \mid x' = \arg \max_{x \in D} \min_i \alpha_i(t_i - f_i)\} \ne \varnothing$$

есть эффективный вектор. Наоборот, любой эффективный вектор x^0 может быть получен как решение задачи (2.15) при некоторых $\alpha_i > 0$ и $t_i > f_i(x)$, $i \in [1:k]$.

Доказательство. Первое утверждение теоремы есть следствие теорем 2, 3. Действительно, согласно теореме 3 множество T_α будет содержать эффективный вектор, а согласно теореме 2 решение задачи (2.14) на множестве $D' = T_\alpha$ при любых $\alpha_i > 0$, в том числе и при $\alpha_i = 1$, является эффективным вектором.

Докажем обратное утверждение. Пусть x^0 — эффективный вектор. Он будет и слабо эффективным, поэтому, согласно теореме 1, при $\alpha = \alpha'$ он получается как решение задачи (2.10) в составе множества $T_{\alpha'}$ с выполнением соотношений (2.13). Пусть $x' \in T_{\alpha'}$ и $\sum_{i=1}^{k} f_i(x') < \sum_{i=1}^{k} f_i(x^0)$.

Тогда существует по крайней мере один $i = i_0$ такой, что $f_{i_0}(x') < f_{i_0}(x^0)$. Но тогда, согласно (2.13), мы получили бы

$$\alpha'_{i_0}(t_{i_0} - f_{i_0}(x')) > \alpha'_{i_0}(t_{i_0} - f_{i_0}(x^0)) = \max_{x \in D} \min_i \alpha'_i(t_i - f_i(x)),$$

что противоречит свойству $x^0 \in T_{\alpha'}$. Следовательно, x^0 есть решение задачи (2.15). Теорема доказана.

Доказанные теоремы позволяют понять сущность методов максиминной, минимаксной и линейной сверток, обсуждавшихся в 2.2. Становится ясной роль весовых коэффициентов, параметризующих множество слабо эффективных решений. С точки зрения приложений, представляет интерес соотношение (2.13), позволяющее в случае выпуклых границ *множества достижимости* $f(D)$ выбирать весовые коэффициенты, исходя из требуемых соотношений между окончательными значениями критериальных выходных параметров. Для более подробного и наглядного изложения методов и принципов многокритериальной оптимизации мы отсылаем читателя к учебнику [74].

Метод главного критерия также может быть проинтерпретирован помощью понятия эффективного решения. Справедливы следующие утверждения.

Теорема 2.5. Решение задачи

$$f_1(x) \to \min_{x \in D'}, \tag{2.16}$$

где множество

$$D' = \{x \in D \mid f_i(x) \le t_i,\ i \in [2:k]\} \tag{2.17}$$

не пусто, есть слабо эффективный вектор.

Доказательство. Пусть x^0 есть решение задачи (2.16) и существует $x' \in D$ такой, что

$$f_i(x') < f_i(x^0),\ i \in [1:k]. \tag{2.18}$$

Тогда $x' \notin D'$, так как в противном случае это противоречило бы свойству $f_1(x^0) \le f_1(x)$ для $x \in D'$. Следовательно, $x' \notin D'$, и поэтому существует такой $i = i_0$, для которого $f_{i_0}(x') > t_{i_0}$, что противоречит (2.18).

Теорема 2.6. Пусть T_t — непустое множество решений задачи (2.16). Тогда существует вектор $x^* \in T_t$, являющийся эффективным.

Доказательство. Пусть существует $x' \in D$, для которого $f_i(x') \le f_i(x^0)$, $i \in [1:k]$, где x^0 — произвольный элемент множества T_t. Тогда $x' \in D'$, причем, по предположению, $f_1(x') \le f_1(x^0)$ и, следовательно, $f_1(x') = f_1(x^0)$, так как $x^0 = \arg\min_{x \in D'} f_1(x)$. Таким образом, $x' \in T_t$, и вне множества T_t не может существовать решений x', для которых $f_i(x') \le f_i(x^0)$, $i \in [1:k]$. Теорема доказана.

Следствие. Если решение задачи (2.16) единственно, то оно эффективно.

Теорема 2.7. Любой эффективный вектор может быть получен как решение задачи (2.16) при некоторых t_i, $i \in [2:k]$.

Доказательство. Пусть $x^0 \in P(D)$; примем $t_i = f_i(x^0)$, $i \in [2:k]$ и покажем, что

$$f_1(x^0) = \min f_1(x), \; x \in D'. \tag{2.19}$$

Пусть $x' \in D'$; тогда $f_i(x') \le f_i(x^0)$, $i \in [2:k]$. Если предположить, что $f_1(x') < f_1(x^0)$, то это будет противоречить эффективности вектора x^0, следовательно, $f_1(x') > f_1(x^0)$, что эквивалентно (2.19).

Из доказанных теорем следует, что в качестве «главного» может быть выбран любой критерий. Независимо от этого выбора любое эффективное решение может быть получено как решение задачи (2.16) при соответствующем назначении t_i.

Ни один из методов, представленных ранее, не позволяет выделить единственное оптимальное решение. Решения, соответствующие различным наборам весовых коэффициентов, являются равноправными элементами множества (слабо) эффективных решений. Выбор окончательного результата осуществляется на основе дополнительной информации о предпочтениях лица, принимающего решения. Принцип Парето позволяет лишь сузить класс возможных претендентов на решение и исключить из рассмотрения заведомо неконкурентоспособные варианты.

Методы выбора единственного решения существуют и связаны с использованием моделей и процедур, предназначенных для структуризации и количественного описания субъективного мнения лица, принимающего решение. В результате мы приходим к оптимизационным задачам максимизации *функций ценности* или (если компоненты векторного критерия оптимальности являются случайными переменными) *функций полезности.*

2.4. Методы исключения ограничений

Рассмотрим методы учета ограничений в однокритериальной задаче

$$J(x) \to \min_x, \; x \in D, \tag{2.20}$$

где D задано в (2.2).

Формально наиболее просто снимаются прямые ограничения, Для этого достаточно выполнить замену переменных по одной из формул табл. 2.1, где z_i означают новые независимые переменные. Однако, как указано, например, в книге [19],

подобная замена переменных в ряде случаев приводит к различным вычислительным осложнениям. В частности, может возрастать обусловленность матриц Гессе целевого функционала по новым переменным вплоть до получения вырожденных матриц.

Таблица 2.1

№	Ограничение	Преобразование
1	$x_i > a_i$	$x_i = a_i + \exp(z_i)$
2	$x_i > a_i$	$x_i = a_i + z_i^2$
3	$x_i > x_j, i \neq j$	$x_j = z_j, x_i = z_j + z_i^2$
4	$a_i \le x_i \le b_i$	$x_i = b_i + (a_i - b_i)\sin^2 z_i$ $x_i = 0{,}5(a_i + b_i) + 0{,}5(b_i - a_i)\sin z_i$
5	$a_i < x_i < b_i$	$x_i = b_i + (a_i - b_i)\frac{1}{\pi}\operatorname{arcctg} z_i$ $x_i = b_i + (a_i - b_i)\exp(z_i) / (1 + \exp(z_i))$
6	$a \le x_i \le b$ $a \le x_j \le b$ $x_i > x_j, i \neq j$	$x_j = b + (a - b)\sin^2 z_j$ $x_i = b + (a - b)\sin^2 z_j \sin^2 z_i$
7	$a < x_i < b$	$x_j = b + (a - b)\frac{1}{\pi}\operatorname{arcctg} z_j$
	$a < x_j < b$ $x_i > x_j, i \neq j$	$x_i = b + (a - b)\frac{1}{\pi^2}\operatorname{arcctg} z_i \operatorname{arcctg} z_i$
8	$a_i(x_j) \le x_i \le b_i(x_j)$ $i \neq j$	$x_j = z_j$ $x_i = b_i(z_j) + (a_i(z_j) - b_i(z_j))\sin^2 z_i$
9	$a_i(x_k) \le x_i \le b_i(x_j)$ $i \neq j, i \neq k$	$x_j = z_j$ $x_k = z_k$ $x_i = b_i(z_i) + (a_i(z_k) - b_i(z_j))\sin^2 z_i$

В задачах параметрической оптимизации, возникающих при алгоритмизации процессов управления, наиболее часто встречаются сложные нелинейные ограничения. Это не позволяет использовать специальную технику параметрической оптимизации, направленную, например, на решение задач с линейными или прямыми ограничениями [19].

Существует два класса методов оптимизации, ориентированных на решение однокритериальных задач с ограничениями общего вида. Первый класс составляют алгоритмы, реализующие методы проекции градиента, отсечения, а также различные методы возможных направлений. Эти алгоритмы дают возможность на каждой итерации свести исходную задачу к формально более простой задаче с ограничениями, например к задаче линейного программирования.

Во вторую группу входят методы штрафных функций и модифицированных функций Лагранжа, основанные на учете ограничений непосредственно в конструкции критерия оптимальности с последующим использованием алгоритмов безусловной оптимизации [21]. Далее рассматривается второй, в определенном смысле более перспективный, подход.

Основная идея *метода штрафных функций* состоит в следующем. Рассмотрим задачу нелинейной оптимизации вида

$$J(x) \to \min_{x \in D},\ D = \left\{ x \in R^n \middle| h_i(x) = 0,\ i \in [1:q] \right\}, \tag{2.21}$$

не содержащую ограничений в виде неравенств; тогда вместо (2.21) решается последовательность задач безусловной минимизации однопараметрического семейства функционалов $\{J_k\}$, где

$$J_k(x) = J(x) + \sigma_k \sum_{i=1}^{q} h_i^2(x),\ \ \sigma_k \to \infty,\ \ k \to \infty. \tag{2.22}$$

Второе слагаемое в (2.22) имеет смысл «штрафа» за нарушение ограничений, что и определяет название метода. Справедлива следующая теорема.

Теорема 2.8. Пусть задача (2.21) имеет единственное решение x^*; функции J, h_i непрерывны в R^n; для любого $k = 1, 2\ldots$ существует $x_k = \arg\min J_k(x) \in D \subset R^n$, где D — ограниченное замкнутое множество; $\sigma_k \to \infty$, $k \to \infty$. Тогда $\lim x_k \to x^*$, $k \to \infty$.

Доказательства различных утверждений, аналогичных сформулированной теореме, содержатся в учебнике [29].

Наиболее распространенный вариант метода штрафных функций для решения общей задачи

$$J(x) \to \min_{x},\ x \in D; \tag{2.23}$$

$$D = \left\{ x \in R^n \middle| g_i(x) \le 0;\ \ i \in [1:m];\ \ g_i(x) = 0;\ \ i \in [m+1:s] \right\}$$

состоит в применении вспомогательных функционалов

$$J_k(x) = J(x) + \sigma_k \sum_{i=1}^{s} [g_i^+(x)]^p, \tag{2.24}$$

где

$$g_i^+ = \begin{cases} \max\{g_i(x);\ 0\}, & i \in [1:m]; \\ |g_i(x)|, & i \in [m+1:s]. \end{cases}$$

Если функционалы J, g_i являются r раз непрерывно дифференцируемыми на множестве D, то при любом $p > r$ этим же свойством будут обладать функционалы $J_k(x)$ [13].

Основной недостаток метода штрафных функций заключается в ухудшении обусловленности вспомогательных задач при больших σ_k. Соответствующие вопросы рассмотрены в третьей главе.

Наиболее перспективным общим методом учета ограничений считается *метод модифицированных функций Лагранжа* [19]. Применительно к задаче (2.21) он формулируется следующим образом.

Введем в качестве обобщенного критерия оптимальности функционал

$$M(x,\lambda,\sigma) = J(x) + \langle \lambda, h(x)\rangle + \frac{\sigma}{2}\|h(x)\|^2, \tag{2.25}$$

где $\sigma > 0$ — параметр метода. Тогда алгоритм оптимизации сводится к итерационному процессу

$$\begin{aligned} x^{k+1} &= \arg\min_x M(x,\lambda^k,\sigma), \\ \lambda^{k+1} &= \lambda^k + \sigma h(x^{k+1}), \quad h(x) = [h_1(x), \ldots, h_q(x)], \end{aligned} \tag{2.26}$$

обобщающему методы штрафных функций и множителей Лагранжа. Основная особенность сформулированного алгоритма по сравнению с методом штрафных функций заключается в отсутствии неограниченно растущего штрафного коэффициента σ. При правильной организации процесса величина σ влияет лишь на скорость сходимости, но не на сам факт сходимости последовательности $\{x^k\}$ к оптимуму. При решении практических задач значение σ целесообразно подбирать в интерактивном режиме, так как надежные методы априорного задания σ в настоящее время отсутствуют.

2.5. Влияние неопределенных факторов на процесс оптимизации

Рассмотренные в 1.3 примеры классических задач теории управления показывают, что существуют различные специальные методы учета внешних возмущений, основанные на корреляционном анализе сигналов, методах фильтрации и сглаживания экспериментальных зависимостей и т. д. В результате часто удается свести проблему к стандартной задаче параметрической оптимизации без неопределенных факторов.

Однако для целого ряда практических ситуаций таких специальных методов не существует. Поэтому представляют интерес некоторые общие принципы параметрической оптимизации в условиях неопределенности обстановки.

Рассмотрим математическую модель объекта оптимизации

$$y = F(x, \xi), \tag{2.27}$$

включающую неопределенный вектор ξ. В этих условиях с методологической точки зрения целесообразно различать три основные ситуации: 1) ξ — случайный вектор с известным законом распределения; вектор выходных параметров y реализуется многократно для различных значений вектора ξ; 2) вектор y реализуется однократно при заранее не известном векторе ξ; 3) выходные параметры y реализуются многократно; вектор ξ изменяется неизвестным образом, но не носит случайный характер либо статистические характеристи-

ки ξ оказываются неизвестными. В качестве примера рассмотрим задачу оптимального проектирования. Тогда первый случай характерен для анализа технологического разброса параметров при серийном производстве прдукции. Проектирование уникальных изделий приводит ко второму типу неопределенности. Третий вариант возникает при моделировании влияния неконтролируемых параметров внешней среды, определяющих условия эксплуатации объекта.

На практике возникают и более сложные ситуации. Например, часть компонент вектора ξ может носить случайный характер с известными законами распределения, а некоторые компоненты могут меняться непредсказуемым образом, но не обладать свойством статистической устойчивости.

Обратимся к методам параметрической оптимизации при наличии неопределенности первого типа. Наиболее простой путь заключается в переходе от выражения (2.27) к зависимости

$$y = F\left(x, \bar{\xi}\right) \tag{2.28}$$

где $\bar{\xi} = M\{\xi\}$ — математическое ожидание случайного вектора ξ. В модели (2.28) неопределенность формально отсутствует и могут применяться методы исследования детерминированных моделей. Однако получаемые при этом результаты должны интерпретироваться с учетом вероятностной природы вектора ξ. В случае решения задачи (2.28) гарантируется лишь оптимальность «в среднем» для достаточно большого числа реализаций ξ.

Второй возможный подход основан на использовании представления

$$y = \overline{F(x,\xi)} = M\{F(x,\xi)\}. \tag{2.29}$$

Переход от (2.27) к (2.28) или (2.29) является неформальным актом, и построение окончательной математической модели должно опираться на дополнительную информацию о задаче и на эвристические представления исследователя о действительных целях оптимизации.

При использовании подхода (2.29) однокритериальная задача параметрической оптимизации может быть сформулирована следующим образом:

$$\begin{aligned} &F_0(x) = M\{J(x,\xi)\} \to \min_{x \in D}; \\ &D = \left\{x \in R^n \middle| F_i(x) = M\{g_i(x,\xi)\} \le 0,\ i \in [1\colon l]\right\}. \end{aligned} \tag{2.30}$$

Предполагается, что исходная многокритериальная задача предварительно сведена к одной или нескольким задачам со скалярным критерием качества (функционалом).

Задача (2.30) является задачей *стохастического программирования* [40]. В отличие от детерминированной постановки, функционалы задачи не заданы в явном виде, и для их вычисления необходимо проводить усреднение по ξ, что, вообще говоря, связано с вычислением многомерных интегралов. Последнее приводит к нереальным вычислительным затратам в случае прямого применения детерминированных методов нелинейного программирования. Более эффективными оказываются процедуры стохастического программирования, основанные на ин-

формации о конкретных реализациях функционалов $J(x, \xi)$, $g_i(x, \xi)$, отвечающих различным значениям вектора ξ.

Один из вариантов подобных методов сводит задачу (2.30) к последовательности детерминированных задач

$$J(x, \xi^k) \to \min_{x \in D_k};$$
$$D = \{x \in R^n | g_i(x, \xi^k) \le 0, i \in [1: l]\} \tag{2.31}$$

при фиксированных значениях случайного вектора $\xi = \xi^k$, $k = 1, 2, \ldots$ Эти значения должны вырабатываться датчиком случайных чисел в соответствии с заданной плотностью распределения ξ. Решение задачи (2.31), отвечающее вектору ξ^k, обозначим через x^k. Тогда последовательность векторов $\{\tilde{x}^k\}$, сходящаяся к решению исходной задачи (2.30), строится следующим образом: $\tilde{x}^{k+1} = \tilde{x}^k + \alpha_k(x^{k+1} - \tilde{x}^k)$, $\alpha_k > 0$, где $\tilde{x}^2 = x^1 + \alpha_1(x^2 - x^1)$.

Для сходимости процесса необходимо выполнение условий

$$\alpha_k \underset{k \to \infty}{\to} 0, \sum_{k=1}^{\infty} \alpha_k = \infty, \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

Этим требованиям удовлетворяет, например, последовательность $\left\{\alpha_k = \frac{1}{k}\right\}$. На практике, однако, возникают проблемы, связанные с эффективным выбором α_k для повышения скорости сходимости метода.

Рассмотрим второй и третий типы неопределенности. Информация о статистических свойствах вектора ξ, даже если она и имеется, не может быть эффективно использована. В указанных условиях целесообразно производить расчет «на наихудший случай», используя *принцип гарантированного результата* и дополнительную информацию вида $\xi \in G_\xi$, где G_ξ — некоторое ограниченное множество. Соответствующая задача оптимизации формулируется следующим образом:

$$F_0(x) = \max_{\xi \in G_\xi} J(x, \xi) \to \min_{x \in D};$$
$$D = \left\{x \in R^n \middle| F_i(x) = \max_{\xi \in G_\xi} g_i(x, \xi) \le 0, i \in [1: l]\right\}. \tag{2.32}$$

Из (2.32) видно, что вычисление функционалов, задающих критерий и ограничения, сопряжено с решением вспомогательных задач оптимизации. В результате трудоемкость процедуры в целом оказывается довольно высокой.

Как видно из изложенного, регулярный учет случайных факторов в общей задаче параметрической оптимизации довольно сложен и в настоящее время ограничен лишь относительно простыми ситуациями. Поэтому чаще вначале решается задача детерминированной оптимизации при фиксированных, например средних, значениях ξ. Далее в отношении наилучшего детерминированного варианта проводится тот или иной вид статистического анализа с целью расчета реальных характеристик объекта оптимизации.

2.6. Методы декомпозиции

Под методами *разделения*, или *декомпозиции*, понимаются методы сведения исходной задачи большой размерности к решению нескольких более простых задач. Будем предполагать, что внешние условия, влияющие на объект, фиксированы (фактор неопределенности обстановки отсутствует) и задача сведена к однокритериальной с целевым функционалом $J(x)$. Цель оптимизации заключается в построении вектора

$$x^* \in \operatorname{Arg}\min_{x \in D} J(x);$$
$$\operatorname{Arg}\min_{x \in D} J(x) = \left\{ x \in D \middle| J(x) = \min_{x' \in D} J(x') \right\}. \tag{2.33}$$

Если размерность n пространства поиска велика, то это, с одной стороны, приводит к резкому возрастанию необходимого для целей оптимизации количества вычислений значений $J(x)$, а с другой — к большому времени, необходимому для однократного вычисления значения $J(x)$. В результате прямое решение задачи (2.33) стандартными методами нелинейной оптимизации оказывается невозможным из-за нереальных вычислительных затрат.

2.6.1. Метод агрегирования

Метод агрегирования допускает следующее формальное представление.

Задача параметрической оптимизации формулируется в виде (2.33). Далее вводятся *агрегированные характеристики*

$$z_i = \varphi_i(x_1, x_2, \ldots, x_n), \tag{2.34}$$

где вектор $z = (z_1, z_2, \ldots, z_s)$ должен иметь существенно меньшую по сравнению с x размерность. Функции φ_i, по предположению, могут быть построены таким образом, что: 1) целевой функционал $J(x)$ может быть представлен в виде суперпозиции отображений

$$J(x) = F(\varphi(x)) = F(z),\ z \in D_z = \varphi(D); \tag{2.35}$$

2) существуют обратные отображения $\varphi_i^{-1}(\cdot)$, позволяющие для $\forall\, z \in D_z$ достаточно эффективно вычислять $x = \varphi^{-1}(z) \in D$.

Если структура функционала $J(x)$ допускает введение агрегированных переменных (2.34) при выполнении сформулированных требований, то процесс минимизации $J(x)$ может быть реализован следующим образом. На первом этапе решается задача построения вектора

$$z^* \in \operatorname{Arg}\min_{z \in D_z} F(z) \tag{2.36}$$

в пространстве пониженной размерности s, $s << n$. На втором этапе по найденному z^* восстанавливается вектор

$$x^* = \varphi^{-1}(z^*). \tag{2.37}$$

Легко видеть, что вектор (2.37) есть решение исходной задачи (2.33). Экономия вычислительных затрат достигается за счет уменьшения размерности пространства поиска при решении задачи (2.36), а также в силу предположения о достаточно простой алгоритмической структуре оператора $\varphi^{-1}(\cdot)$.

Приведем одну из возможных интерпретаций изложенной формальной конструкции, известную в теории автоматического управления и теории синтеза электронных устройств как метод *«аппроксимации и реализации»* [33].

Пусть требуется спроектировать звено, имеющее заданную амплитудно-частотную характеристику (АЧХ). Тогда на этапе аппроксимации строится дробно-рациональная передаточная функция $W(p, z)$ комплексной переменной p и вектора $z = (z_1, z_2, \ldots, z_s)$ коэффициентов полиномов в числителе и знаменателе. Вектор z вычисляется как решение задачи

$$F(z) \to \min_{z \in D_z}, \tag{2.38}$$

где $F(z)$ характеризует близость расчетной и желаемой АЧХ, а множество D_z задается условиями физической реализуемости функции $W(p, z)$. На этапе реализации по найденным z_i^* выбирается структурная схема устройства, конкретизирующая вид функций φ_i в (2.34). А далее определяется одно из решений системы так называемых *компонентных* уравнений

$$\varphi_i(x_1, x_2, \ldots, x_n) = z_i^*, \; i \in [1{:}s]. \tag{2.39}$$

Разрешимость (2.39) должна следовать из выполненных на этапе аппроксимации условий физической реализуемости.

Такое разбиение процесса оптимизации на два этапа обычно мотивируется соображениями удобства, позволяющими на этапе аппроксимации не рассматривать конкретные объекты и получать некоторые общие результаты. Однако не менее важная особенность данного подхода связана с идеей декомпозиции. Пусть трудоемкость решения задачи минимизации функционала линейно зависит от размерности n и приближенно оценивается величиной kn. В книге [54] предлагается считать $k \cong 50$. Допустим также, что основная вычислительная работа выполняется при вычислении «расстояния» между аппроксимируемой и аппроксимирующей АЧХ. Иначе говоря, коэффициенты z_i по заданным x_i рассчитываются относительно просто, и напротив, реализация зависимостей $F(z)$, $J(x) \triangleq F(\varphi(x))$ как функций z и x оказывается довольно трудоемкой (например, в связи с необходимостью сравнения заданной и расчетной АЧХ в достаточно большом числе точек).

В этом случае трудоемкость прямого решения задачи $J(x) \to \min$ без разбиения ее на этапы аппроксимации и реализации оценивается числом kn, а трудоемкость этапа аппроксимации — числом ks. Пренебрегая вычислительными затратами на этапе реализации, связанными с получением значений z_i по формулам (2.34), получим выигрыш за счет декомпозиции приблизительно в $\frac{n}{s}$ раз. Дополнительные примеры, иллюстрирующие целесообразность изложенного подхода, содержатся в книге [40].

2.6.2. Метод вспомогательных частных критериев

Данный метод связан с сокращением множества D конкурирующих решений [40]. Это может быть выполнено на основе идеи введения вектора *вспомогательных частных критериев* $u(x) = [u_1(x), u_2(x), \ldots, u_m(x)]$.

Предполагается, что критерий $J(x)$ удовлетворяет следующему *условию монотонности*: для любых двух точек x', x'' из системы неравенств $u_i(x') \geq u_i(x'')$, $i \in [1: m]$, следует $J(x') > J(x'')$. Таким образом, если решение x'' оказывается более предпочтительным, чем x', по векторному критерию $u(x)$, то оно будет более предпочтительным и с позиций скалярного критерия $J(x)$.

При выполнении условия монотонности исходная задача (2.33) может быть заменена следующей:

$$J(x) \to \min_{x \in P_u(D)}, \tag{2.40}$$

где $P_u(D)$ — множество решений, эффективных (Парето-оптимальных) по векторному критерию u на множестве D. Очевидно, $P_u(D) \subset D$. Если достигнутое сужение множества D значительно, то задача (2.40) оказывается проще исходной и цель декомпозиции считается достигнутой.

Алгоритмическая сторона изложенного подхода заключается в следующем. Предполагается, что критерии $u_i(x)$, в отличие от $J(x)$, оказываются эффективно вычислимыми с относительно малыми затратами компьютерного времени. Решая последовательность оптимизационных задач вида

$$\max_i \alpha_i u_i(x) \to \min_{x \in D}, \ \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m) \in \alpha^\varepsilon,$$

где α^ε — некоторая дискретная сетка в множестве

$$\overline{\alpha} = \left\{ \alpha = (\alpha_1, \ldots, \alpha_n) \,\middle|\, \alpha_i > 0, \sum_{i=1}^{m} \alpha_i = 1 \right\},$$

мы фактически реализуем некоторую функцию $x(\alpha)$, определяющую точки x, в которых необходимо вычислять исходный глобальный критерий $J(x(\alpha))$. Число таких точек определяется количеством узлов сетки α^ε. Существенно, что размерность пространства векторов α может оказаться значительно меньше (на несколько порядков) размерности исходного пространства векторов x.

Для построения α^ε-сетки может быть использован метод зондирования пространства векторов α, основанный на построении $ЛП_\tau$-последовательностей, обладающих свойством равномерного заполнения заданной многомерной области [55].

2.7. Особенности оптимизационных задач

Рассмотренные методы позволяют свести общую задачу параметрической оптимизации с ограничениями, возмущениями и векторным критерием оптимальности к последовательности канонических оптимизационных задач.

Возникающие при этом канонические экстремальные задачи обладают рядом особенностей, позволяющих выделить их среди всех задач безусловной нелинейной оптимизации. Основные, с точки зрения приложений в области теории управления, особенности заключаются в следующем.

1. Алгоритмическое задание функционалов, задающих критерии и ограничения, существенно увеличивает трудоемкость их вычисления и вынуждает ограничиваться методами оптимизации, не использующими в явном виде выражения для производных.
2. Критерии оптимальности, применяемые в задачах управления, часто имеют характерную структуру, позволяющую строить специальные методы оптимизации второго порядка, использующие упрощенные выражения для вторых производных.
3. Однократное вычисление функционалов задачи связано с довольно сложным и трудоемким решением соответствующей задачи анализа. Наиболее эффективными целесообразно считать алгоритмы, которые в процессе оптимизации наименьшее число раз обращаются к вычислению значений минимизируемых функционалов и функционалов ограничений для получения решений с требуемой точностью.
4. Если оптимизируется сложная многопараметрическая система, то ее обычно можно представить как некоторую совокупность связанных подсистем меньшей размерности. Учет подобной структуры системы позволяет строить более рациональные методы по сравнению с традиционными универсальными алгоритмами нелинейного программирования.
5. Невыпуклая структура минимизируемых функционалов существенно понижает эффективность обычных методов нелинейной оптимизации, особенно если такой структуре сопутствует описываемая далее овражная ситуация.
6. Как свидетельствует практика решения реальных задач параметрической оптимизации, возникающие канонические оптимизационные задачи являются, как правило, плохо обусловленными. Это определяет характерную овражную структуру поверхностей уровня минимизируемых функционалов и вызывает резкое замедление сходимости стандартных методов оптимизации.
7. Тейлоровское разложение гладкого целевого функционала указывает на возможность его аппроксимации в окрестности любой точки пространства поиска с помощью квадратичного функционала. Как показывает практика решения реальных задач из области управления, участки «квадратичности» минимизируемых функционалов оказываются достаточно обширными. По этой причине соответствующие алгоритмы параметрической оптимизации могут строиться, исходя из квадратичных приближений целевых функционалов.

Указанные характерные черты канонических оптимизационных задач определяют конкретные требования к методам параметрической оптимизации. С позиций традиционных методов нелинейной оптимизации, наиболее существенными оказываются особенности, отмеченные в пп. 5 и 6. Основная трудность состоит в том, что с математической точки зрения проблема невыпуклости оказывается неразрешимой в силу сложности класса невыпуклых оптимизационных задач.

Согласно результатам, представленным в книге [41], основной вывод, сформулированный в интуитивных терминах, заключается в том, что даже для гладких одноэкстремальных функционалов в задачах с не очень малой размерностью пространства управляемых параметров скорость сходимости любого метода (равномерно по всем задачам) безнадежно мала, и попытка построить общий метод, эффективный для всех гладких невыпуклых задач, заранее обречена на неудачу.

Однако с точки зрения практики представляют интерес методы оптимизации, вырабатывающие эффективные направления поиска в точках пространства, где стандартные процедуры оказываются неработоспособными либо неэффективными. В последующих разделах описаны методы, которые как в выпуклой, так и в невыпуклой и одновременно овражной ситуации локально (с точки зрения квадратичной модели) дают существенно более удовлетворительные по скорости убывания функционала результаты по сравнению с традиционными методами.

О правомерности использования соответствующих алгоритмов для задач параметрической оптимизации в системах управления техническими объектами можно судить по результатам решения реальных задач. Применение нестандартных методов будет оправдано, если они окажутся эффективными для заведомо непустого множества практических ситуаций, вызывающих трудности для традиционных поисковых процедур. В данном случае такое множество можно указать заранее — это множество задач с целевыми функционалами, близкими к кусочно-квадратичным, не обязательно выпуклым зависимостям. Типичность подобных функций подтверждается практикой реальных компьютерных вычислений в области теории управления. Речь, следовательно, идет о некотором и, может быть, существенном расширении обычно применяемого арсенала методов и алгоритмов конечномерной оптимизации.

Подтверждаемая экспериментально довольно высокая степень работоспособности рассматриваемых методов никак не противоречит результатам уже упоминавшейся работы [41]. Дело заключается в том, что на практике довольно редко реализуются те специальные структуры невыпуклых задач, которые приводят к пессимистическим теоретическим оценкам.

2.8. Некоторые стандартные схемы параметрической оптимизации

В зависимости от реальных ситуаций, примеры которых были рассмотрены в 1.3, могут формироваться различные алгоритмические схемы параметрической оптимизации. Далее представлены некоторые варианты таких схем, охватывающие значительное количество практических задач, решаемых при проектировании, а также в процессе функционирования систем управления.

2.8.1. Задачи аппроксимации

Задачи аппроксимации наиболее часто возникают в практике параметрической оптимизации. К таким задачам, в частности, приводят:

- алгоритмы идентификации нелинейных детерминированных объектов по схеме с настраиваемой моделью;
- алгоритмы идентификации стохастических объектов, основанные на процедурах сглаживания;
- корреляционные алгоритмы идентификации линейных стохастических объектов;
- алгоритмы адаптации систем автоматического управления, основанные на методе обучающейся модели;
- алгоритмы оптимального проектирования устройств с заданными, например, амплитудно-частотными характеристиками;
- корреляционные алгоритмы синтеза статистически оптимальных систем автоматического управления.

Критерии оптимальности в задачах аппроксимации чаще всего формируются одним из следующих способов:

$$J_1(x) = \sum_{i=1}^{N} \alpha_i^2 [W(x, s_i) - W^*(s_i)]^2 \to \min_{x \in D}; \tag{2.41}$$

$$J_2(x) = \max_{i \in [1:N]} \alpha_i \left| W(x, s_i) - W^*(s_i) \right| \to \min_{x \in D}, \tag{2.42}$$

где D — множество допустимых значений x; α_i — весовые коэффициенты, определяющие необходимую точность аппроксимации в отдельных точках диапазона изменения независимой переменной s; s_i, $i \in [1:N]$ — дискретная сетка значений s, при которых происходит сравнение заданной $W^*(s)$ и расчетной $W(x, s)$ характеристик.

Каждый из приведенных критериев имеет свои достоинства и недостатки. Функционал J_1 довольно прост и обладает свойством «гладкости». А именно, если функция $W(x, s)$ является дважды непрерывно дифференцируемой функцией x, то этим же свойством будет обладать зависимость $J_1(x)$, что существенно облегчает последующую процедуру оптимизации. Недостаток J_1 заключается в возможности выбросов по точности для отдельных слагаемых. Иначе говоря, плохая точность аппроксимации в некоторых точках s_i при больших значениях N может компенсироваться хорошей точностью в других точках. Этот недостаток устранен в критерии (2.42), однако он не сохраняет характеристики гладкости функции $W(x, s)$, что требует привлечения специальных методов оптимизации.

Как показывает практика, довольно простой и надежный способ решения задач аппроксимации, возникающих в теории управления, заключается в использовании гладких среднестепенных аппроксимаций минимаксного критерия J_2. Согласно этому подходу, вместо решения задачи (2.42) ищется минимум функционала со *среднестепенной структурой* [52]:

$$J_3(x) = \sum_{i=1}^{N} \varphi_i^{\nu}(x) \to \min_{x \in D}, \quad \nu = 2, 3, \ldots \tag{2.43}$$

где $\varphi_i(x) \triangleq \alpha_i \left| W(x, s_i) - W^*(s_i) \right|$.

При достаточно больших значениях ν решения задач (2.42), (2.43) будут практически совпадать. Действительно, справедливо предельное соотношение

$$\left[\sum_{i=1}^{N}\varphi_i^{\nu}\right]^{\frac{1}{\nu}} \to \max_i \varphi_i,\ \nu \to \infty,$$

где $\varphi_i > 0$, $i \in [1:N]$ — произвольные числа. Кроме того, можно показать, что операция извлечения корня ν-й степени не влияет на локализацию точки минимума.

Функционал J_3 совмещает в себе достоинства функционалов J_1 и J_2. Являясь гладким, подобно J_1, он не допускает значительных отличий по точности аппроксимации в отдельных точках.

При решении практических задач на основе критерия J_3 целесообразно пошаговое увеличение параметра ν, начиная с ν = 2. Таким способом обычно удается избежать переполнения разрядной сетки компьютера при возведении первоначально больших значений локальных ошибок аппроксимации φ_i в высокую степень ν. Кроме того, проводя в интерактивном режиме оценку получаемых в процессе увеличения ν решений, мы можем вовремя прервать процесс, если получены удовлетворительные результаты. Априорное задание оптимального значения ν обычно оказывается затруднительным. Как правило, при практических расчетах значение ν не превышает 10–15. Учет ограничений $x \in D$ производится методами, изложенными в 2.4.

Рассмотренный подход очевидным образом обобщается на случай вектор-функций W. При этом в качестве функций $\varphi_i(x)$ могут использоваться зависимости $\varphi_i(x) = \alpha_i \| W(x, s_i) - W^*(s_i) \|$.

2.8.2. Системы неравенств

Наиболее часто задачи с неравенствами возникают при формализации основных спецификаций к оптимизируемому объекту (см. 1.3.7). В ряде случаев аппроксимационные по смыслу задачи также целесообразно формулировать и решать как системы неравенств. Например, если ставится задача аппроксимации заданной характеристики $W(s)$ с помощью расчетной зависимости $W(x, s)$ в точках $s_i, i \in [1:N]$, и если при этом заданы допустимые значения $\delta_i > 0$ точности аппроксимации в различных точках диапазона, то есть $W(x, s_i) \in [W(s_i) - \delta_i, W(s_i) + \delta_i]$, $i \in [1:N]$, то требование $W(x, s_i) \cong W(s_i)$ очевидно эквивалентно двум неравенствам:

$$W(x, s_i) > W(s_i) - \delta_i;$$

$$W(x, s_i) \le W(s_i) + \delta_i.$$

В общем случае имеем

$$y_i(x) \le t_i, i \in [1:m]. \qquad (2.44)$$

С помощью задания системы (2.44) могут учитываться все виды ограничений на объект оптимизации: прямые, функциональные и критериальные.

Особенность рассматриваемого класса задач заключается в том, что обычно нельзя считать, что «чем меньше $y_i(x)$, тем лучше». Напротив, на одну и ту же выходную характеристику, как правило, накладываются двусторонние ограничения:

$$\varphi_{\text{КН}} \le \varphi_k(x) \le \varphi_{\text{КВ}},$$

что приводит к двум неравенствам вида (2.44)

$$y_i(x) \triangleq \varphi_k(x) \le t_i \triangleq \varphi_{\text{КВ}};$$

$$y_j(x) \triangleq -\varphi_k(x) \le t_j \triangleq -\varphi_{\text{КН}}.$$

Поэтому трактовка задачи (2.44) как многокритериальной вида $y_i(x) \to \min$ с применением соответствующих процедур свертки обычно нецелесообразна. Как правило, необходимо обеспечить *безусловное выполнение всех неравенств* (2.44), не допуская больших значений $y_i(x)$, (превышающих соответствующие t_i) за счет очень малых значений $y_j(x)$ других выходных параметров.

Рациональный подход связан с применением критериев вида

$$J(x) = \min_i \gamma_i^{-1}(t_i - y_i(x)) \to \max_{x \in R^n} \tag{2.45}$$

или

$$J(x) = \max_i \gamma_i^{-1}(y_i(x) - t_i) \to \min_{x \in R^n}.$$

При этом параметры γ_i должны задавать «единицы измерения» разностей $t_i - y_i$ для различных i. Эти разности характеризуют *запас*, с которым выполняются неравенства (2.44). Согласно (2.45) максимизируется минимальный из запасов.

Функционал (2.45), так же как и (2.42), не является гладким, что существенно усложняет ситуацию и требует применения специальных оптимизирующих процедур. Далее излагается альтернативный подход, основанный на процедуре сглаживания исходного функционала с последующим обращением к методам гладкой оптимизации.

Пусть $z_i(x) \triangleq t_i - y_i(x)$. Тогда, очевидно,

$$\min_i z_i(x) = \max_i [\exp(-z_i(x))].$$

Поэтому задача (2.45) эквивалентна задаче

$$\max_i [\exp(-z_i(x))] \to \min_x. \tag{2.46}$$

Для задачи (2.46) применима среднестепенная свертка (2.43), если принять $\varphi_i(x) \triangleq \exp[-z_i(x)]$, $i \in [1:m]$. В результате приходим к следующему критерию оптимальности:

$$J(x) = \sum_{i=1}^{m} \exp[-\nu z_i(x)] \to \min_x, \ \nu = 1, 2, \ldots. \tag{2.47}$$

Точно так же могут решаться и многокритериальные задачи вида

$$f_i(x) \to \min_x, \; i \in [1:l]; \; x \in D \subset R^n;$$
$$D = \left\{ x \in R^n \middle| \, g_j(x) \le t_j, \; j \in [l+1:s] \right\},$$

где функции $g_i(x)$ задают функциональные ограничения. Сформулированная задача может быть представлена в виде системы неравенств

$$f_i(x) \le \hat{f}_i, \; g_j(x) \le t_j; \; i \in [1:l]; \; j \in [l+1:s],$$

где $\hat{f}_i$ — некоторые предельные значения для критериальных выходных параметров f_i.

Полагая

$$z_i(x) = \alpha_i [\hat{f}_i - f_i(x)] / \gamma_i; \; i \in [1:l];$$
$$z_j(x) = \alpha_j [t_j - g_j(x)] / \gamma_j; \; j \in [l+1:s],$$

приходим к следующему целевому функционалу, аналогичному (2.47):

$$J(x) = \sum_{i=1}^{s} \exp[-\nu z_i(x)] \to \min_x.$$

Весовые коэффициенты α_i, α_j должны выбираться из условия $\alpha_i << \alpha_j$ ($i \in [1:l]$; $j \in [l+1:s]$). В этом случае (при достаточно больших ν) критерий $J_j(x)$ эквивалентен минимаксному критерию (2.46), что, в свою очередь, эквивалентно задаче

$$\min_i z_i \to \max_x.$$

Очевидно, в силу больших значений α_j, $j \in [l+1:s]$ вес соответствующих слагаемых в выражении для $J(x)$ резко возрастает при нарушении хотя бы одного из условий $g_j(x) \le t_j$ (так как тогда $z_j(x) < 0$). Если же указанные условия выполнены, то, по существу, производится выбор x исходя из максимальности минимального запаса по критериальным неравенствам. В данном случае параметры α_j, $j \in [l+1:s]$ играют роль штрафных коэффициентов при учете функциональных ограничений. Параметры α_i, $i \in [1:l]$ могут использоваться для построения на основе максиминной свертки аппроксимации множества Парето в допустимом множестве D либо отражать интуитивные представления об относительной значимости отдельных критериальных выходных параметров.

Рассмотренный подход в силу своего единообразия и простоты часто оказывается наиболее предпочтительным при решении реальных задач параметрической оптимизации.

Как показывает вычислительная практика, трудности оптимального синтеза параметров по критериям типа максимума минимального запаса (2.45) часто вызываются исключительно негладкостью критериев, приводящей к преждевременной остановке поисковой процедуры. Целесообразно поэтому сразу обращаться к модифицированным критериям (2.47) с применением на первом этапе простейших алгоритмов оптимизации типа метода простого покоординатного спуска.

Использование среднестепенных критериев оптимальности в задачах параметрической оптимизации, где, по существу, необходим минимаксный подход, оправдано также с позиций рассматриваемого в главе 3 явления плохой обусловленности. Развитая в настоящее время техника решения негладких оптимизационных задач достаточно сложна, и в невыпуклой овражной ситуации многие алгоритмы теряют эффективность. В то же время излагаемые в главах 4 и 5 методы позволяют получить удовлетворительные результаты для невыпуклых овражных функционалов при условии их гладкости. При этом удается использовать структурные особенности функционалов (2.43), (2.47) для увеличения эффективности соответствующих вычислительных процедур.

2.8.3. Решение систем неравенств в условиях неопределенности

В данном разделе рассматривается специальный случай неопределенности обстановки. Предполагается, что вектор ξ, входящий в левые части решаемых неравенств

$$y(x,\xi) \le t,\ y = (y_1,\ldots,y_m),\ t = (t_1,\ \ldots,\ t_m), \tag{2.48}$$

имеет случайный характер. Подобные постановки задач возникают, например, при алгоритмизации процедур *управления технологическими процессами* в условиях серийного производства продукции.

В качестве критерия, достаточно глубоко отражающего конечную цель решения указанной задачи, как правило, можно выбрать вероятность P выполнения условий (2.48):

$$J(x) = P\left[y(x,\xi) \le t\right] \to \max_x. \tag{2.49}$$

В данном случае все возможные ограничения на компоненты вектора x считаются учтенными за счет расширения системы неравенств (2.48). Процедура однократного вычисления функционала $J(x)$ основывается на проведении статистических испытаний по методу Монте-Карло в соответствии с заданной плотностью распределения $\Psi(x,\xi)$ случайного вектора ξ. Можно показать, что одновременно с расчетом значения целевого функционала можно практически без дополнительных вычислительных затрат рассчитать составляющие вектора градиента $J'(x)$ и матрицы Гессе $J''(x)$. Данное обстоятельство положено в основу построения некоторых реализаций рассматриваемых в главах 4, 5 методов параметрической оптимизации.

Альтернативный подход к решению задачи (2.49) рассмотрен в учебнике [42]. Он заключается в следующем. Потребуем, чтобы каждое неравенство (2.48) выполнялось с некоторым запасом:

$$y_i(x,\bar{\xi}) + \delta_i \le t_i,\ \delta_i > 0, \tag{2.50}$$

где δ_i характеризует величину рассеяния i-го выходного параметра за счет статистических вариаций компонент вектора ξ относительно своих средних (как правило, нулевых) значений $\bar{\xi}_i$. Требования (2.50) эквивалентны неравенствам

$$z_i(x) \triangleq \left[\frac{(t_i - y_i)}{\delta_i} - 1\right] \geq 0. \qquad (2.51)$$

Величина z_i имеет смысл *запаса работоспособности* по i-му выходному параметру.

На практике получила распространение максиминная форма целевого функционала

$$J(x) = \min_i z_i(x) \to \max_{x \in R^n}, \qquad (2.52)$$

аналогичная (2.45). От представления (2.52) можно по методике, изложенной в 2.8.2, перейти к выражению (2.47), более удобному для практических расчетов.

Иногда в выражение (2.52) вводятся дополнительные весовые коэффициенты α_i, позволяющие регулировать степень «перевыполнения» требований (2.50) по отдельным выходным параметрам.

Для определения δ_i проводится статистический анализ в окрестности текущей точки x. Значения δ_i обычно имеют смысл трехсигмовых допусков, которые периодически уточняются в процессе оптимизации. Весьма часто величины δ_i задаются как исходные данные на основе априорной информации, что значительно сокращает трудоемкость процедуры оптимизации, особенно при решении идентичных задач.

2.8.4. Сигномиальная оптимизация

В некоторых практических ситуациях, связанных, например, с задачами оптимального проектирования, целевые функционалы или их аппроксимации имеют стандартную *сигномиальную* структуру:

$$J(x) = \sum_{i=1}^{k} s_i(x), \qquad (2.53)$$

где $s_i(x) = c_i x_1^{\alpha_{i1}} x_2^{\alpha_{i2}} \ldots x_n^{\alpha_{in}}$; $\alpha_{ij} \in R^1$; $x_j > 0$; $j \in [1:n]$.

В частном случае $c_i \geq 0$ выражение (2.53) называется *позиномом*, а соответствующая задача параметрической оптимизации — *задачей геометрического программирования* [60].

В 4.5 развиты общие стандартные процедуры сигномиальной оптимизации, основанные на простоте вычисления первых и вторых производных функционалов (2.53). При этом предполагается, что возможные дополнительные ограничения на параметры x_i также имеют вид сигномов и учитываются методами штрафных функций или модифицированных функций Лагранжа без нарушения сигномиальной структуры расширенного целевого функционала.

2.9. Основные результаты и выводы

1. Общая постановка задачи параметрической оптимизации характеризуется такими факторами сложности, как многокритериальность, наличие ограничений, необходимость учета случайных и неопределенных воздействий.

Кроме того, пространство управляющих параметров в достаточно сложной реальной задаче может иметь высокую размерность, что исключает прямое применение стандартных алгоритмических средств поисковой оптимизации.

2. Методы устранения указанных факторов сложности могут быть основаны на преобразовании исходной задачи к последовательности относительно более простых канонических оптимизационных задач.
3. Характерные особенности канонических оптимизационных задач теории управления определяют следующие требования к соответствующим методам и алгоритмам конечномерной оптимизации:
   - реализации применяемых методов решения канонических задач должны иметь нулевой порядок (не использовать производные в своей схеме вычислений);
   - наиболее эффективными целесообразно считать алгоритмы, которые в процессе оптимизации наименьшее число раз обращаются к вычислению значений целевых функционалов для получения решения с требуемой точностью;
   - специальные структуры применяемых в теории управления критериев качества необходимо учитывать при построении более эффективных методов и алгоритмов решения соответствующих классов задач по сравнению с традиционными универсальными алгоритмами нелинейного программирования;
   - применяемые в оптимизационных задачах теории управления методы и алгоритмы должны вырабатывать эффективные направления поиска в предположении невыпуклости целевого функционала и при наличии сложного овражного рельефа поверхностей уровня на значительных отрезках траектории спуска.
4. В зависимости от вида решаемой задачи могут быть использованы конкретные стандартные схемы параметрической оптимизации. Большое число задач теории управления допускает формальное представление в виде:
   - задач аппроксимации на основе метода наименьших квадратов, а также минимаксных критериев;
   - задач решения систем детерминированных неравенств, определяющих требования-спецификации (как функциональные, так и критериальные) к оптимизируемой системе по заданному списку выходных параметров;
   - задач решения систем неравенств в условиях неопределенности.
5. В главе 2 рассмотрена единая стратегия решения сформулированных в п. 4 задач на основе среднестепенных целевых функционалов, традиционно применяемых только при решении задач аппроксимации непрерывных зависимостей.

Глава 3. Проблема плохой обусловленности

В данной главе анализируется часто возникающая на практике ситуация неудовлетворительного поведения стандартных методов поисковой оптимизации при решении канонических задач. Как правило, это выражается в резком замедлении сходимости применяемых поисковых процедур, а в ряде случаев — в полной остановке алгоритма задолго до достижения оптимальной точки (ситуация «ложного» локального экстремума).

Возникновение подобных трудностей связывается далее со специальной формой плохой обусловленности матрицы вторых производных минимизируемых функционалов, приводящей к характерной овражной структуре поверхностей уровня. Указываются причины частого появления плохо обусловленных экстремальных задач в теории управления.

3.1. Явление овражности

Рассмотрим следующий пример критерия оптимальности, зависящего от двух управляемых параметров x_1, x_2 [49]:

$$J(x_1, x_2) = g_0^2(x_1, x_2) + \sigma g_1^2(x_1, x_2) \to \min_x, \tag{3.1}$$

где σ — достаточно большое положительное число. Рассмотрим также уравнение

$$g_1(x_1, x_2) = 0, \tag{3.2}$$

определяющее в простейшем случае некоторую зависимость $x_2 = \varphi(x_1)$. Тогда при стремлении параметра σ к бесконечности значение функционала J в каждой точке, где $g_1(x_1, x_2) \neq 0$, будет неограниченно возрастать по абсолютной величине, оставаясь ограниченным и равным $g_0^2(x_1, x_2)$ во всех точках на кривой $x_2 = \varphi(x_1)$. То же самое будет происходить с нормой вектора градиента $J'(x) = \left[\frac{\partial J}{\partial x_1}, \frac{\partial J}{\partial x_2}\right]$, где

$$\frac{\partial J}{\partial x_1} = 2g_0(x_1, x_2)\frac{\partial g_0}{\partial x_1} + 2\sigma g_1(x_1, x_2)\frac{\partial g_1}{\partial x_1};$$
$$\frac{\partial J}{\partial x_2} = 2g_0(x_1, x_2)\frac{\partial g_0}{\partial x_2} + 2\sigma g_1(x_1, x_2)\frac{\partial g_1}{\partial x_2}.$$

Линии уровня $J(x) = \text{const}$ для достаточно большого σ представлены на рис. 3.1. Там же стрелками показано векторное поле антиградиентов, определяющее локальные направления наискорейшего убывания $J(x)$.

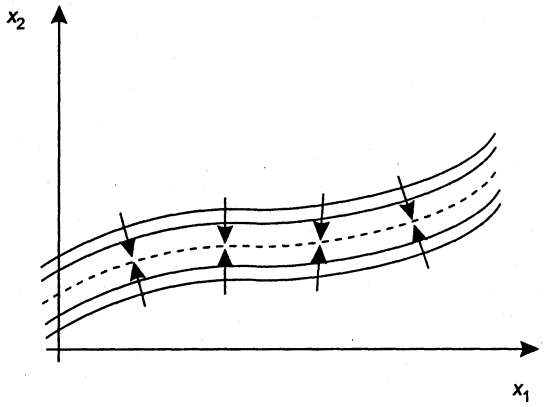

Рис. 3.1. Явление овражности

Ясно, что при достаточно больших σ минимальные значения $J(x)$ следует искать вдоль зависимости $x_2 = \varphi(x_1)$, определяющей так называемое *дно оврага*. Из (3.1) следует, что изменение $J(x)$ вдоль дна задается выражением $g_0^2(x_1, x_2)$ и не зависит от величины параметра σ. Таким образом, задача минимизации $J(x)$ сводится к минимизации функционала $g_0^2(x_1, \varphi(x_1))$ от одной переменной x_1. В общем случае уравнение $x_2 = \varphi(x_1)$ обычно неизвестно.

Приведенный пример *овражной ситуации* является достаточно простым, хотя и из него уже видны принципиальные трудности, связанные, например, с применением широко распространенных методов спуска по антиградиенту. Действительно, из рис. 3.1 следует, что направления поиска, задаваемые антиградиентами, оказываются неэффективными. Приводя довольно быстро процесс поиска на дно оврага, они в окрестности дна начинают осциллировать, оставаясь почти перпендикулярными направлению в точку минимума.

Возможны различные усложнения и обобщения рассматриваемой ситуации. Например, уравнение (3.2) может, вообще говоря, определять не одно, а несколько решений $x_2 = \varphi_i(x_1)$, каждое из которых означает свой овраг. С точки зрения приложений более существенным обобщением является предположение о наличии многомерного дна оврага. Чтобы проиллюстрировать это явление, обратимся к следующему примеру:

$$J(x) = g_0^2(x) + \sigma\sum_{i=1}^{m} g_i^2(x),\ x \in R^n,\ n > 2,\ m < n. \qquad (3.3)$$

В этом случае дно оврага задается системой уравнений

$$g_i(x) = 0,\ i \in [1:m], \tag{3.4}$$

что, в принципе, позволяет выразить m параметров x_i через оставшиеся $n-m$ переменных. Предположим, не ограничивая общности, что уравнения (3.4) определяют следующие зависимости:

$$x_1 = \varphi_1(x_{m+1}, \ldots, x_n),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$x_m = \varphi_m(x_{m+1}, \ldots, x_n).$$

Аналогично предыдущему случаю устанавливаем, что задача минимизации (3.3) при достаточно больших σ эквивалентна минимизации функции $g_0^2(\varphi_1, \ldots, \varphi_m, x_{m+1}, \ldots, x_n)$ от $r \triangleq n-m$ переменных $x_{m+1}, \ldots, x_n$. Число r носит название размерности дна оврага. Легко представить, что, например, для $n = 3$ поверхности уровня одномерного оврага имеют характерный «сигарообразный» вид, а для двумерного оврага они будут близки к деформированным дискам.

В любом случае для овражной ситуации определяющим фактором является специальная структура поверхностей уровня $J(x)$, весьма сильно отличающаяся от сферической. Характерно наличие некоторой области притяжения (по сути — дна оврага) $Q \subset R^n$, содержащей оптимальную точку $x^* = \arg\min J(x)$. При этом норма вектора градиента $J'(x)$ для $x \in Q$, как правило, существенно меньше, чем в остальной части пространства.

Овражную структуру могут иметь не только функционалы вида (3.1), (3.3), явно содержащие большой параметр σ. Можно привести следующий пример квадратичного функционала:

$$f(x_1, x_2) = 0{,}250025x_1^2 + 0{,}49995x_1x_2 + 0{,}250025x_2^2 - x_1 - x_2. \tag{3.5}$$

Линии уровня $f(x) = \text{const}$ функционала (3.5) представляют семейство подобных эллипсоидов с центром в точке (1:1). Длины полуосей эллипсоидов относятся при этом как $1 : 10^2$. Указать большой параметр σ в выражении (3.5) нельзя, хотя овражная ситуация налицо, и так же, как и в предыдущих случаях, явно выделяется дно оврага (прямая ab на рис. 3.2), имеющее уравнение $x_2 = 2 - x_1$. Подставляя выражение для x_2 в (3.5), снова приходим к эквивалентной задаче меньшей размерности:

$$f_1(x_1) = 10^{-4}(x_1 - 1)^2 - 1 \to \min_{x_1}.$$

Необходимость выделения овражных оптимизационных задач в отдельный класс обусловлена, с одной стороны, значительными вычислительными трудностями при их решении стандартными для компьютерного моделирования методами, а с другой стороны, бесспорным фактом важности данного класса задач для большинства практических ситуаций. Специалисты по моделированию сталкивались с овражной ситуацией уже на заре современной компьютерной эры, когда компьютеры стали регулярно использоваться при решении реальных задач. Приведем лишь некоторые свидетельства специалистов, подтверждающие тезис о типичности овражной ситуации:

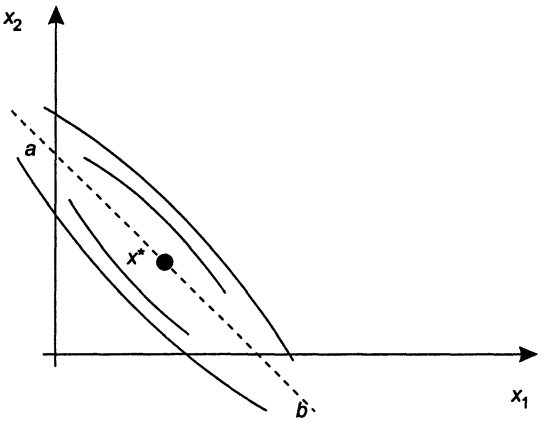

Рис. 3.2. Овражный квадратичный функционал

- [40], с. 226: «с увеличением размерности задачи возрастает вероятность появления оврагов»;
- [50], с. 353: «большинство практических задач многопараметрической оптимизации, особенно из области оптимального проектирования, страдает обилием такого рода ловушек (то есть оврагов)»;
- [42], с. 272: «в экстремальных задачах проектирования необходимо использовать методы оптимизации, приспособленные для поиска экстремума, в овражных ситуациях»;
- [43], с. 161: «особенностью целевых функций при решении задач схемотехнического проектирования является их гребневый (овражный при поиске минимума) характер, что приводит к большим вычислительным трудностям;
- [17], с. 35: «Наибольшие трудности при поиске локального оптимума доставляют так называемые "овражные" ситуации».

Аналогичные утверждения делаются также во многих других работах, связанных с компьютерным моделированием. Сейчас известны различные методы, ориентированные на решение рассматриваемого круга оптимизационных задач, однако и в настоящее время проблема минимизации овражных функционалов является актуальной. Особенно остро стоит вопрос минимизации овражных и одновременно невыпуклых функционалов, так как именно в этой ситуации отказывает большинство методов поисковой оптимизации.

3.2. Основные определения

Пусть решается задача $J(x) \to \min$, $J \in C^2(D)$, $x \in D \subset R^n$. Будем предполагать далее, что функционал $J(x)$ ограничен снизу на D.

Траектория наискорейшего спуска (ТНС) $x(\tau)$ функционала $J(x)$ задается известным из курсов численного анализа векторным дифференциальным уравнением

$$\frac{dx}{d\tau} = -J'(x),\ J'(x) \triangleq \left[\frac{\partial J}{\partial x_1}, \ldots, \frac{\partial J}{\partial x_n}\right]. \tag{3.6}$$

Эти траектории в случае овражных функционалов обладают рядом специфических черт. Например, для функционала (3.5) имеем

$$x(\tau) = \sum_{i=1}^{2} \left\{ \alpha_i^* + (\alpha_i^0 - \alpha_i^*) \exp(-\lambda_i \tau) \right\} u_i, \tag{3.7}$$

где

$$x(0) = \sum_{i=1}^{2} \alpha_i^0 u_i,\ x^* = (1,\ 1) = \sum_{i=1}^{2} \alpha_i^* u_i;$$

$$\lambda_1 = 1,\ \lambda_2 = 10^{-4},\ u_1 = \frac{1}{\sqrt{2}}(1,1),\ u_2 = \frac{1}{\sqrt{2}}(-1,1).$$

Из (3.7) видно, что из-за наличия быстро затухающей и медленно затухающей экспонент отчетливо выделяются два участка с существенно различным поведением решения. Первый, сравнительно непродолжительный, характеризуется большими значениями производных $\frac{dx_i(\tau)}{d\tau}$ и означает спуск на дно оврага. На дне выполняются условия типа (3.2) и норма вектора градиента, а с ней и производные $\frac{dx_i(\tau)}{d\tau}$ становятся относительно малыми. Поэтому для второго участка характерно относительно плавное изменение переменных x_1, x_2. Таким образом, прослеживается полная аналогия с поведением решений так называемых *жестких систем обыкновенных дифференциальных уравнений*. В связи с этим в [49] предложено следующее общее определение.

Определение 1. Функционал $J(x)$ называется *овражным*, если отвечающая ему система дифференциальных уравнений (3.6) — жесткая.

Однако с точки зрения приложений к области оптимизации более конструктивным оказывается приведенное далее определение овражного функционала. Это определение не содержит, в частности, таких неестественных для задач оптимизации требований, как необходимость задания промежутка интегрирования уравнения (3.6) [72].

Определение 2. Функционал $J(x) \in C^2(D)$, $D \subset R^n$ называется *овражным* в множестве $Q \in D$, если найдутся такие числа $\delta > 0$, $\sigma >> 1$ и множество $Q_\delta \subset D$, что

$$\begin{aligned} &1)\ \forall x \in Q_\delta,\ \lambda_1[J''(x)] \geq \sigma \left|\lambda_n[J''(x)]\right|; \\ &2)\ \forall x \in Q,\ \mathrm{Arg}\min_{x' \in X_\delta(x)} J(x') \subset Q; \\ &3)\ \forall x \in Q,\ L[X_\delta(x) \cap Q] \leq \sigma^{-1} L[X_\delta(x)], \end{aligned} \tag{3.8}$$

где $X_\delta(x) = \{x' \in R^n \mid \|x' - x\| \le \delta\}$; $Q_\delta = \bigcup_{x \in Q} X_\delta(x)$, $\lambda_i[A]$ — собственные числа матрицы $A = J''(x)$, упорядоченные по убыванию:

$$\lambda_1[A] \ge \lambda_2[A] \ge \ldots \ge \lambda_n[A];$$

$L[S]$ — константа Липшица в соотношении

$$\|J'(x') - J'(x)\| \le L[S]\|x' - x\|, \ \forall x', x \in S \subset R^n.$$

Множество Q называется *дном оврага*.

Основным в определении 2 является условие 1, констатирующее резко несимметричное расположение спектра матрицы вторых производных $J''(x)$ относительно начала координат: $\lambda_i \in [-m, M]$, $M >> m > 0$. Условия 2, 3 необходимы для описания свойства «устойчивости» множества Q: можно показать, что все ТНС, начинавшиеся в любой точке $x \in Q_\delta$, быстро попадают в достаточно малую окрестность Q_ε ($\varepsilon << \delta$) множества Q и остаются там до выхода из множества Q_δ.

Как правило, для приложений оказывается достаточной более грубая модель явления овражности, когда предполагается, что собственные числа матрицы вторых производных можно отчетливо разделить на две группы, в одну из которых входят собственные числа, по модулю намного превосходящие элементы второй группы. Будет использоваться следующее определение [72].

Пусть в $D \in R^n$ задана r-мерная поверхность (конфигурационное пространство) $Q = \{x \in D \mid g_i(x) = 0, i \in [1:n-r]\}$, $g_i \in C^2(D)$.

Определение 3. Функционал $J(x) \in C^2(D)$, $D \subset R^n$ называется овражным на множестве Q, если найдутся такие числа $\delta > 0$, $\sigma >> 1$ и множество $Q_\delta \subset D$, что

$$\begin{aligned} &1)\ \forall x \in Q_\delta,\ \lambda_1[J''(x)] \ge \ldots \ge \lambda_{n-r}[J''(x)] \ge \\ &\quad \ge \sigma\left|\lambda_{n-r+1}[J''(x)]\right| \ge \ldots \ge \sigma\left|\lambda_n[J''(x)]\right|; \\ &2)\ \forall x \in Q,\ \operatorname{Arg\,min} J(x') \subset Q; \\ &3)\ \forall x \in Q,\ L[X_\delta(x) \cap Q] \le \sigma^{-1} L[X_\delta(x)]. \end{aligned} \tag{3.9}$$

Число r называется размерностью оврага (дна оврага) Q.

Пример. Рассмотрим квадратичный функционал

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c, c = \text{const}. \tag{3.10}$$

Пусть собственные числа $\lambda_i(A)$, $i \in [1:n]$, удовлетворяют неравенствам (3.9) (при $\forall x \in R^n$). Построим множество Q, являющееся в силу определения 3 дном оврага. Предположим без существенного ограничения общности, что $\det A \ne 0$, и обозначим через x^* решение уравнения $Ax = b$. Тогда

$$f'(x) = Ax - b = A(x - x^*) = \sum_{i=1}^{n} \beta_i \lambda_i u_i, \tag{3.11}$$

где $x - x^* = \sum_{i=1}^{n} \beta_i u_i$.

Примем

$$Q = \left\{ x \in R^n \middle| \beta_i = \langle x - x^*, u_i \rangle = 0, \quad i \in [1:n-r] \right\}. \tag{3.12}$$

Тогда, очевидно, $Q_\delta = R^n$, $\sigma \cong \dfrac{\lambda_1}{|\lambda_{n-r+1}|}$, и условие 1 выполнено.

Обратимся к проверке выполнения условия 2. Рассмотрим сферическую окрестность

$$X_\delta = \left\{ x \in R^n \middle| \| x - \bar{x} \| \le \delta \right\}$$

некоторой точки $\bar{x} = \sum_{i=n-r+1}^{n} \alpha_i u_i + x^* \in Q$. Для произвольного $x = \sum_{i=1}^{n} \beta_i u_i + x^* \in R^n$ условие $x \in X_\delta$ эквивалентно требованию

$$\left\| \sum_{i=1}^{n-r} \beta_i u_i + \sum_{i=n-r+1}^{n} (\beta_i - \alpha_i) u_i \right\| \le \delta$$

или

$$\beta = (\beta_1, \ldots, \beta_n) \in S_\beta = \left\{ \beta \in R^n \middle| \sum_{i=1}^{n-r} \beta_i^2 + \sum_{i=n-r+1}^{n} (\beta_i - \alpha_i)^2 \le \delta^2 \right\}.$$

Имеем

$$f(x) = \frac{1}{2} \sum_{i=1}^{n} \beta_i^2 \lambda_i + c_1 \triangleq \varphi(\beta),$$

где c_1 не зависит от β. Задача поиска $x \in X_\delta$, минимизирующего $f(x)$, сводится к следующей задаче условной оптимизации:

$$\varphi(\beta) \to \min_{\beta}, \; g(\beta) = \sum_{i=1}^{n-r} \beta_i^2 + \sum_{i=n-r+1}^{n} (\beta_i - \alpha_i)^2 - \delta^2 \le 0.$$

Условие стационарности для функции Лагранжа

$$L(\mu_1, \mu_2, \beta) = \mu_1 \varphi(\beta) + \mu_2 g(\beta)$$

имеет вид

$$L'_\beta = \mu_1 \varphi'(\beta) + \mu_2 g'(\beta) = 0$$

или

$$\beta_i (\mu_1 \lambda_i + 2\mu_2) = 0, \; i \in [1:n-r];$$

$$\mu_1 \beta_j \lambda_i + 2\mu_2 (\beta_j - \alpha_j) = 0, \; j \in [n-r+1:n],$$

причем $\mu_i \ge 0$ и не равны нулю одновременно.

Отсюда в силу положительности «больших» собственных чисел λ_i ($i \in [1{:}n-r]$) с необходимостью следует $\beta_i = 0$ ($i \in [1{:}n-r]$). Поэтому

$$\min_{\beta \in S_\beta} \varphi(\beta) = \min_{\beta \in \bar{S}_\beta} \bar{\varphi}(\bar{\beta}), \tag{3.13}$$

где

$$\bar{\beta} = (\beta_{n-r+1}, \ldots, \beta_n); \quad \bar{\varphi}(\bar{\beta}) = \frac{1}{2} \sum_{i=n-r+1}^{n} \beta_i^2 \lambda_i + c_1;$$

$$\bar{S}_\beta = \left\{ \gamma = (\gamma_1, \ldots, \gamma_r) \in R^r \,\middle|\, \sum_{i=1}^{r} (\gamma_i - \alpha_{n-r+i})^2 \le \delta^2 \right\}.$$

Из (3.13) следует 2. Для

$$x' = \sum_{i=1}^{n} \beta_i' u_i + x^*, x'' = \sum_{i=1}^{n} \beta_i'' u_i + x^*, x', x'' \in X_\delta$$

имеем

$$\| J'(x') - J'(x'') \| = \| A(x'-x'') \| = \left\| \sum_{i=1}^{n} (\beta_i' - \beta_i'') \lambda_i u_i \right\| \le$$

$$\le \max_i \lambda_i \left\| \sum_{i=1}^{n} (\beta_i' - \beta_i'') u_i \right\| = \lambda_1 \| x'-x'' \|.$$

Аналогично, для $\forall \bar{x}', \bar{x}'' \in X_\delta \cap Q$ получим

$$\| J'(\bar{x}') - J'(\bar{x}'') \| \le |\lambda_{n-r+1}| \cdot \| \bar{x}' - \bar{x}'' \|.$$

Поэтому $L(X_\delta) = \lambda_1 >> L(X_\delta \cap Q) = |\lambda_{n-r+1}|$, и условие 3 выполнено при $\sigma = \dfrac{\lambda_1}{|\lambda_{n-r+1}|}$.

Таким образом, для квадратичных функционалов при сдвинутом в точку x^* начале координат дно оврага Q совпадает с линейной оболочкой (3.12) собственных векторов, отвечающих «малым» собственным числам. Это согласуется с интуитивными представлениями, развитыми в 3.1.

На этом же примере можно проиллюстрировать роль требований 1–3 в определениях 2 и 3. Действительно, для квадратичного функционала (3.10) условия 1, 2 могут выполняться для всего пространства $Q = R^n$ и любого $\delta > 0$. Необходимая линейная оболочка собственных векторов может быть выделена только при дополнительном требовании, эквивалентном 3. В то же время требования 1 и 3 без 2 также оказываются недостаточными. В этом случае сдвиг линейной оболочки Q, являющейся дном оврага, вдоль любого из не вошедших в оболочку собственных векторов не приведет к нарушению 1 и 3, а условие 2 при этом нарушится.

Рассмотренные ранее модели явления овражности не являются исчерпывающими. Однако они описывают наиболее существенные стороны большинства практических ситуаций, связанных с решением задач параметрической оптимизации в системах управления.

Определение 4. Пусть $\forall x \in Q$, $\det J''(x) \neq 0$. Наименьшее из чисел σ, удовлетворяющих определению 2, называется *степенью овражности* $J(x)$ в Q и обозначается $\eta(Q)$. Отношение $\eta(x) = \lambda_1(x)\left|\min_i \lambda_i(x)\right|^{-1}$, $x \in Q$, называется локальной степенью овражности в точке x. Для вырожденных матриц $J''(x)$ величина $\eta(x)$ полагается равной ∞ [72].

Если $J''(x) > 0$, то $\eta(x) = \operatorname{cond}[J''(x)] \triangleq \dfrac{\max_i \lambda_i(x)}{\min_i \lambda_i(x)}$. В общем случае справедливо неравенство $1 \le \eta \le \operatorname{cond}\,[J''(x)]$. При наличии больших по модулю отрицательных собственных чисел $\lambda_i(x)$, то есть при отсутствии овражной ситуации, возможно неравенство $\eta(x) << \operatorname{cond}\,[J''(x)]$. Следовательно, из высокой степени овражности $J(x)$ в точке x следует плохая обусловленность матрицы $J''(x)$, а обратное неверно. Действительно, пусть спектр матрицы $J''(x)$ расположен в множестве $[-M, -m] \cup [m, M]$, $M >> m > 0$, включая граничные значения. Тогда $\eta(x) = 1$, а $\operatorname{cond}\,[J''(x)] = \dfrac{M}{m} >> 1$. Данный функционал не будет относиться к классу овражных, что естественно, так как трудностей при его минимизации, например, методом наискорейшего спуска, не возникает.

Различие между двумя характеристиками $\eta(x)$ и $\operatorname{cond}[J''(x)]$ функционала $J(x)$ часто игнорируется, и овражными называются функционалы с большим числом $\operatorname{cond}\,[J''(x)]$. Подобная точка зрения не оправдывается с позиций основных вычислительных трудностей, возникающих при решении экстремальных задач. Однако, учитывая отмеченную ранее связь между $\eta(x)$ и $\operatorname{cond}\,[J''(x)]$, овражные задачи далее будут называться *плохо обусловленными экстремальными задачами.*

В каждом конкретном случае различные значения $\eta(x)$ следует считать большими. Здесь существует аналогия с понятием плохой обусловленности матрицы. В большинстве случаев все определяется точностью вычислений и типом применяемого алгоритма оптимизации. Традиционно принято классифицировать задачу как плохо обусловленную, если

$$\log_2 \eta > t, \tag{3.14}$$

где t — длина разрядной сетки вычислительной машины. Однако и при меньших значениях η для целого ряда алгоритмов могут возникать значительные вычислительные трудности, особенно если овражность сопровождается отсутствием выпуклости $J(x)$.

Дополнительным фактором, характеризующим степень сложности экстремальной задачи и затрудняющим применение традиционных алгоритмов минимизации, является наличие многомерных оврагов с $r > 1$. В указанной ситуации целый ряд методов, специально ориентированных на решение плохо обусловленных задач, становятся неэффективными.

3.3. Критерии овражности

Рассмотрим практические методы распознавания овражной ситуации, играющие роль *критериев овражности*. Наиболее существенной характеристикой оказывается значение показателя η в допустимой области изменения управляемых параметров.

Своеобразным индикатором может служить *метод простого градиентного спуска* (ПГС), реализуемый по схеме

$$x^{k+1} = x^k - hJ'(x^k) \tag{3.15}$$

с постоянным шагом $h \in R^1$.

Принадлежность $J(x)$ к классу овражных в этом случае проявляется в необходимости применения относительно малых значений h. Попытки увеличения h вызывают потерю свойства релаксационности (монотонного убывания) последовательности $\{J(x^k)\}$, и значения $J(x^k)$ начинают резко возрастать. Если для некоторого фиксированного h (наибольшего из возможных) удалось заставить процесс (3.15) протекать без полной остановки, то по результатам работы метода можно количественно оценить величину η.

Для этого процесс (3.15) продолжается до тех пор, пока отношение $\dfrac{\|J'(x^{k+1})\|}{\|J'(x^k)\|}$ не стабилизируется около некоторого значения μ. Тогда справедливо следующее равенство:

$$\eta \cong \frac{2}{|1-\mu|}. \tag{3.16}$$

Соотношение (3.16) справедливо независимо от выпуклости функционала $J(x)$ и является основным для грубой практической оценки степени овражности решаемой задачи в окрестности текущей точки. Доказательство соотношения (3.16) дано в 5.2.

В силу изложенного можно рекомендовать процесс оптимизации начинать с помощью метода ПГС. Если задача простая и степень овражности невелика, то уже этот метод довольно быстро приведет в малую окрестность оптимума. В противном случае будет получена оценка η, что позволит правильно оценить ситуацию и выбрать наиболее рациональный алгоритм.

Другой, прямой метод оценки η сводится к вычислению матрицы Гессе функционала и решению для нее полной проблемы собственных значений. Тогда на основе непосредственной проверки выполнения неравенств (3.9) для вычисленных собственных чисел делается вывод о значении η. При этом определяется также размерность r дна оврага. Главный недостаток такого подхода заключается в существенных вычислительных трудностях принципиального характера, возникающих при определении малых собственных значений. Можно показать, что абсолютная погрешность $|d\lambda_i|$ представления любого собственного значения λ_i матрицы A за счет относительного искажения δ ее элементов удовлетворяет

неравенству $|d\lambda_i| \le n\delta|\lambda_1|$, где $|\lambda_1| \triangleq \max_i |\lambda_i|$. Полагая $\delta = \varepsilon_m = 2^{-t}$, где ε_m — относительная машинная точность, а t — длина разрядной сетки мантиссы числа, получим оценку для абсолютных искажений собственных чисел за счет ошибок округления:

$$|d\lambda_i| \le n\varepsilon_m|\lambda_1|. \tag{3.17}$$

Параметр ε_m известен для каждой вычислительной системы.

Из последнего неравенства можно сделать следующее заключение. Если все вычисленные собственные числа матрицы $A = J''(x)$ достаточно велики, то есть $|\lambda_i| \ge n\varepsilon_m|\lambda_1|$, то параметр η может быть вычислен непосредственно. Если же некоторые из вычисленных собственных чисел удовлетворяют неравенству $|\lambda_i| \le n\varepsilon_m|\lambda_1|$, то все они должны быть отнесены к блоку «малых» собственных чисел, а для η имеем границу снизу: $\eta \ge \dfrac{1}{n\varepsilon_m}$.

Качественным признаком плохой обусловленности может служить существенное различие в результатах оптимизации, например, методом ПГС, при спуске из различных начальных точек. Получаемые результирующие точки обычно расположены довольно далеко друг от друга и не могут интерпретироваться как приближения к единственному решению или конечной совокупности решений (при наличии локальных минимумов). Описанная ситуация, как правило, означает наличие оврага, а точки остановки применяемой поисковой процедуры трактуются как элементы дна оврага Q.

3.4. Источники плохо обусловленных оптимизационных задач в теории управления

Несмотря на то что типичность овражной ситуации может считаться установленным экспериментальным фактом, представляет определенный интерес выяснение на качественном уровне основных причин появления оврагов в задачах управления. Рассмотрим два типа овражной ситуации — так называемую *естественную овражность* и *внесенную овражность*. В первом случае подразумевается, что задачи параметрической оптимизации могут оказываться плохо обусловленными в силу естественных причин для каждого из выходных параметров объекта оптимизации, имеющих смысл частных критериев оптимальности. Напротив, внесенная овражность возникает в связи с применением специальных методов учета ограничений типа штрафных функций и модифицированных функций Лагранжа, а также в связи с образованием обобщенных критериев оптимальности в виде некоторых сверток частных критериев.

3.4.1. Естественная овражность

Данный тип овражной ситуации может быть проиллюстрирован на примерах решения задач идентификации линейных динамических объектов, а также задач

синтеза статистически оптимальных систем автоматического управления, приводящих к интегральному *уравнению Винера—Хопфа* (см. 1.2.2, 1.2.6)

$$\int_0^T \varphi(\tau-\lambda)\omega(\lambda)d\lambda = R(\tau), \tag{3.18}$$

играющего важную роль в общей теории управления.

Как известно [52], задача решения уравнения Винера—Хопфа является плохо обусловленной, что проявляется в сильной чувствительности решения к малым вариациям исходных функций φ, R, имеющих смысл корреляционных функций и получаемых на практике с ограниченной точностью. Задача оказывается некорректной по Тихонову из-за потери свойства единственности решения [61]. Для построения решения уравнения (3.18) применяются алгебраические методы, основанные на минимизации невязки.

Пример соответствующего целевого функционала был дан в 1.3.2:

$$J(x) = \sum_{j=1}^{N}\left\{\sum_{i=1}^{N}\omega_i\varphi[q(j-i)] - R[qj]\right\}^2 \to \min_x, \tag{3.19}$$

где $\omega_i = \omega(qi)$, $i \in [1{:}N]$; $x = (\omega_1, \omega_2, \ldots, \omega_N)$.

Свойство некорректности исходной задачи приводит к плохой обусловленности матрицы $J''(x)$ с резко выраженным овражным характером соответствующих поверхностей уровня. Характер решения задачи типа (3.19) показан на рис. 3.3. Как отмечается в [52], получаемые таким образом функции $\omega(t)$, как правило, имеют среднюю квадратичную погрешность, близкую к минимальной, однако из-за резко колебательного характера они сильно отличаются от точных решений. При применении стандартных методов параметрической оптимизации, не учитывающих особенностей оптимизационных задач, отмеченных в 2.7, амплитуда паразитных колебаний может быть весьма значительной.

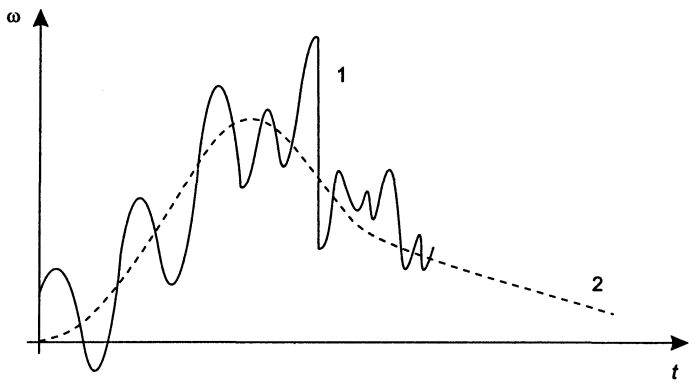

Рис. 3.3. *1* — решение задачи (3.19); *2* — точное решение уравнения Винера—Хопфа

Для ослабления отмеченного эффекта на практике применяются регуляризованные формы целевых функционалов. Простейший критерий с регуляризацией имеет вид

$$J_1(x) = J(x) + \alpha \sum_{j=2}^{N} [\omega_i - \omega_{i-1}]^2 \to \min_x, \tag{3.20}$$

где $\alpha > 0$ — параметр регуляризации, осуществляющий сглаживание получаемого решения.

Существуют и другие формы регуляризации, связанные с расширением исходных целевых функционалов типа (1.34) за счет сглаживающих (регуляризующих) функционалов [22]:

$$I[\omega(t)] = M\{E^2(t)\} + \alpha\Omega[\omega(t)],$$

где

$$\Omega[\omega(t)] = \int_0^T \left\{ k(t)\left[\frac{d\omega(t)}{dt}\right]^2 + p(t)\omega^2(t) \right\} dt. \tag{3.21}$$

Условие минимума $I[\omega(t)]$ при $\alpha = 0$ совпадает с уравнением Винера—Хопфа. При $\alpha \neq 0$ получаются уравнения, для решения которых могут строиться функционалы невязки и решаться экстремальные задачи, аналогичные (3.19) с учетом параметра регуляризации α. Если $\alpha \neq 0$, задача решения «расширенного» уравнения Винера—Хопфа является корректно поставленной и имеет единственное, непрерывное и дифференцируемое решение. Однако решение задач минимизации регуляризованных функционалов, например (3.20), также сопряжено с известными трудностями, вызванными некорректностью исходной задачи. Действительно, регуляризующие добавки помимо эффекта сглаживания дают искажение решения, приводя к увеличению невязки $J(x)$ [52]. Поэтому с целью получения достаточно малых невязок необходимо работать с минимально возможными α, что приводит к достаточно высокой степени «остаточной» обусловленности матрицы $J_1''(x)$. Данное обстоятельство, а также отсутствие надежных априорных методов задания α, приводят к необходимости применять специальные овражно-ориентированные алгоритмы параметрической оптимизации, позволяющие надежно получать достаточно гладкие и точные решения. Дополнительная причина, затрудняющая применение стандартного алгоритмического обеспечения, заключается в факте ухудшения обусловленности задачи с увеличением числа точек дискретности для повышения точности дискретизации при применении алгебраических методов.

В качестве другой, по-видимому, более распространенной причины возникновения овражной ситуации можно указать фактор *агрегированности* управляемых (входных) параметров x. В достаточно большом числе случаев оказывается, что вектор выходных параметров y в действительности зависит не от n независимых переменных $x_1, x_2, \ldots, x_n$, а от s агрегатов

$$y = \Phi(z_1, z_2, \ldots, z_s), \tag{3.22}$$

где

$$z_i = \varphi_i(x),\ i \in [1{:}s];\ s < n. \tag{3.23}$$

Как показывается, например, в [72], прямое решение задачи минимизации целевого функционала $J(x)$ в пространстве параметров x связано с наличием овражной ситуации. Действительно, значение $J(x)$ мало меняется на множестве, определяемом равенствами

$$\varphi_i(x) = z_i^*,\ i \in [1:s], \tag{3.24}$$

где z_i^* — оптимальные значения агрегатов, доставляющих минимум целевому функционалу $I(z) = I[\varphi(x)] \triangleq J(x)$. Поэтому уравнения (3.24) фактически являются уравнениями дна оврага.

Образование агрегатов можно пояснить на примере описания динамических свойств оптимизируемого объекта. Распространенные методы синтеза линейных оптимальных систем управления (см., например, 1.3.5, 1.3.6) основаны на построении передаточных функций замкнутых систем, обладающих необходимыми свойствами (так называемый *этап аппроксимации*). Далее на этапе *реализации* строятся модели реальных систем с конкретной структурой, имеющих заданные (полученные на этапе аппроксимации) передаточные характеристики.

Таким образом, процедура синтеза распадается на две стадии: *синтез в пространстве коэффициентов* передаточных функций, имеющих смысл агрегированных переменных, и *синтез в пространстве «физических» параметров* x_i, подбираемых из условия (3.24) [32].

Пример. Рассмотрим в качестве иллюстрации задачу параметрического синтеза частотного фильтра, схема которого представлена на рис. 3.4 [72].

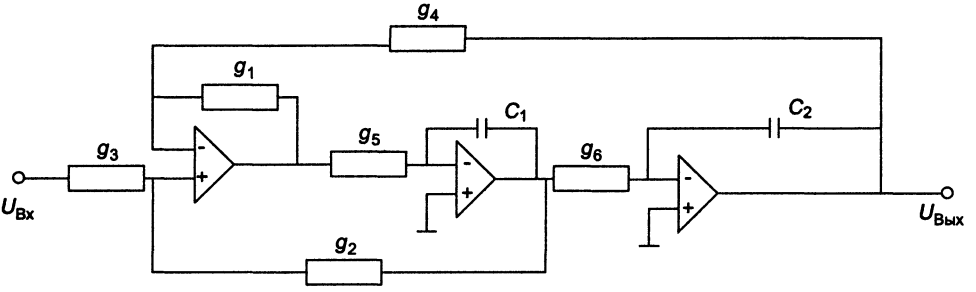

Рис. 3.4. Фильтр нижних частот

Передаточная функция фильтра имеет вид

$$W(j\omega) = \frac{\dot{U}_{\text{вых}}}{\dot{U}_{\text{вх}}} = \frac{1}{z_3 p^2 + z_2 p + z_1},\ p = j\omega, \tag{3.25}$$

где

$$z_0 = g_3 g_5 g_6 (g_1 + g_4);\quad z_1 = g_4 g_5 g_6 (g_2 + g_3)/z_0;$$
$$z_2 = g_2 g_5 (g_1 + g_4) c_2 / z_0;\quad z_3 = g_1 c_1 c_2 (g_2 + g_3)/z_0.$$

Требуется так подобрать компоненты вектора $x = (c_1, c_2, \ldots, g_6)$ неизвестных параметров элементов схемы, чтобы оптимизировать отклонение характеристики затухания $a(\omega) = 20 \lg \dfrac{1}{|W(j\omega)|}$ от заданного значения a_n в полосе пропускания $0 < \omega < \omega_n$ при условии, что в полосе задерживания $\omega_н < \omega < \omega_в$ затухание не меньше заданной величины $a_з$. Выберем N_n равноотстоящих точек на промежутке $[0, \omega_n]$ и $N_з$ равноотстоящих точек в полосе задерживания $[\omega_н, \omega_в]$. Тогда задача аппроксимации может быть сформулирована в виде следующей задачи параметрической оптимизации:

$$J(z) = \max_i |a(z, \omega_i) - a_n| \to \min_z, \; z \in Z \subset R^3, \; \omega_i \in [0, \omega_n], \; i \in [1{:}N_n];$$

$$Z = \left\{ z \in R^3 \middle| z_i \geq 0; \; a(z, \omega_j) \geq a_з, \; \omega_j \in [\omega_н, \omega_в], \; j \in [1{:}N_з] \right\}.$$

Найденные в результате решения этой задачи значения z_i^* являются коэффициентами искомой дробно-рациональной функции, аппроксимирующей желаемую функцию цепи. Для решения задачи реализации необходимо по найденным коэффициентам z_i^* восстановить составляющие вектора x согласно соотношениям (3.25), конкретизирующим (3.24). В результате приходим к следующей задаче параметрической оптимизации:

$$\left\| z^* - \varphi(x^*) \right\|^2 \to \min_x, \; x \in D \subset R^8;$$

$$D = \{ x \in R^8 \mid x_i \geq 0, \; i \in [1{:}8] \}.$$

В данном случае предполагалось, что $Z \subset \varphi(D)$.

Из приведенного примера видно, что агрегаты обычно фигурируют как коэффициенты уравнений, описывающих определенные динамические свойства объекта. При этом количество исходных переменных x_i, составляющих действительный вектор x варьируемых параметров, обычно оказывается значительно большим.

Весьма важным обстоятельством в данном случае является тот факт, что в подавляющем числе случаев при оптимизации *нелинейных* систем, а также при отсутствии явных представлений для соответствующих математических моделей мы вынуждены проводить оптимизацию непосредственно в пространстве параметров элементов оптимизируемой системы. Последнее, согласно изложенному ранее, подразумевает работу в условиях овражной ситуации. Заметим, что иногда и при синтезе линейных систем необходимо осуществлять поиск в пространстве x из-за трудностей в задании множества достижимости Z при сложных функциональных ограничениях на компоненты вектора x.

В действительности обычно реализуется следующая зависимость:

$$J(x) = I(z_1, z_2, \ldots, z_s) + \varepsilon g(x)$$

обобщенного показателя качества $J(x)$ от агрегатов z_i и вектора исходных переменных x. Наличие «объединяющего» функционала $g(x)$, с одной стороны, позволяет говорить лишь о приближенном агрегировании с точностью до малого

параметра ε, а с другой, при соответствующих конструкциях $g(x)$, решает проблему единственности решения в пространстве переменных x.

Высока вероятность появления овражной ситуации из-за образования так называемых асимптотических агрегатов при оптимизации динамических объектов, описываемых в пространстве состояний *жесткими системами* обыкновенных дифференциальных уравнений с большим разбросом значений основных постоянных времени.

Возникновение жестких дифференциальных уравнений при моделировании реальных систем в теории управления определяется разбросом временных характеристик, заложенным в самой их физической природе. Примерами могут служить задачи электротехники, химической кинетики, моделирования сложных систем путем формального объединения описаний различных по инерционности подсистем. Например, в задачах управления электроэнергетическими системами жесткие дифференциальные уравнения возникают из-за резко различающихся по скорости протекания быстрых электромагнитных процессов и медленных электромеханических [49].

Высокие значения чисел обусловленности матрицы Якоби дифференциальных уравнений динамики характерны для моделей гироскопических устройств, а также различных типов электронных схем.

Пример. Пусть функционирование объекта описывается вектор-функцией

$$u(t) = [u_1(t),\ u_2(t)]; \tag{3.26}$$

$$u_1(t) = t[x_1 \exp(-At) + (x_1 + x_2)\exp(-at)];$$

$$u_2(t) = t[x_1 \exp(-Bt) + (x_1 + x_2)\exp(-bt)],$$

являющейся при $A >> a > 0$, $B >> b > 0$ решением жесткой системы уравнений.

Требуется получить оптимальные значения параметров x_1, x_2 из условия наилучшего совпадения $u(t)$ с заданной вектор-функцией $\bar{u}(t)$ для $t \in [t_0, T]$. Если предположить, что $t_0 > \tau_{\text{п.с}}$, где $\tau_{\text{п.с}}$ — длина пограничного слоя, определяющего практически полное затухание экспонент $\exp(-At)$, $\exp(-Bt)$, то из (3.26) видно, что поведение решения $u(t)$ для $t \in [t_0, T]$ будет определяться агрегатом $z = x_1 + x_2$. Если продолжать считать x_1 и x_2 независимыми параметрами, то степень овражности, например, следующего квадратичного целевого функционала:

$$J(x) = [u_1(t_1) - \bar{u}_1(t_1)]^2 + [u_2(t_1) - \bar{u}_2(t_1)]^2 \to \min_x$$

будет достаточно высока. При $t_1 = 1$; $a = 1$; $b = 1{,}5$; $A = 20$; $B = 15$ получим

$$\eta \cong \frac{\exp[-\min(a,b)]}{\exp[-\max(A,B)]} \cong 1{,}8 \cdot 10^8.$$

Таким образом, по крайней мере до тех пор, пока не разработаны достаточно универсальные регулярные методы выделения агрегатов для последующей декомпозиции исходной задачи параметрической оптимизации, мы вынуждены считаться с необходимостью работы в пространстве переменных x_i в условиях овражной ситуации.

3.4.2. Внесенная овражность

Учет ограничений

Овражная ситуация может быть внесена в задачу оптимизации в силу используемой конструкции обобщенного критерия оптимальности. Рассмотрим эффект овражности, возникающий при использовании методов штрафных функций и модифицированных функций Лагранжа. На наличие овражной ситуации в подобных случаях указывается в большом числе работ.

Рассмотрим в качестве примера ограничения в виде равенств $g_j(x) = 0$, $j \in [1:p]$. Тогда, согласно методу штрафных функций, задача сводится к минимизации вспомогательных функционалов

$$J_0(x, \sigma) = J(x) + \sigma \sum_{j=1}^{p} g_j^2(x) \to \min_x$$

с достаточно большим положительным коэффициентом σ. При этом структура расширенного критерия J_0, содержащего большой параметр σ, как правило, оказывается овражной, даже если исходный функционал $J(x)$ этим свойством не обладает.

Пример. Пусть требуется найти x_1, x_2, минимизирующие квадратичный функционал $f(x) = x_1^2 + x_2^2$ при условии $x_1 = 2$. Задача имеет очевидное решение $x_1^* = 2$, $x_2^* = 0$. Поступим формально и составим вспомогательный функционал согласно общей рецептуре метода штрафных функций:

$$J_0(x) = x_1^2 + x_2^2 + \sigma(x_1 - 2)^2 = \frac{(x_1 - b_1)^2}{a_1^2} + \frac{(x_2 - b_2)^2}{a_2^2} + d,$$

где $a_1 = \dfrac{1}{\sqrt{1+\sigma}}$; $a_2 = 1$; $b_1 = \dfrac{2\sigma}{\sigma+1}$; $b_2 = 0$; $d = \dfrac{4\sigma}{\sigma+1}$.

Уравнение линии уровня $J_0(x) = \text{const}$ является уравнением эллипса с центром в точке (b_1, b_2) и длинами полуосей, относящимися как $\dfrac{a_2}{a_1} = \sqrt{1+\sigma}$. При больших значениях σ, обеспечивающих относительно точное выполнение ограничения $x_1 = 2$, линии уровня оказываются сильно вытянутыми (рис. 3.5). Чем точнее выполняются ограничения, тем ярче выражен эффект овражности. В данном случае степень овражности равна $\eta = 1 + \sigma$ и стремится к бесконечности при $\sigma \to \infty$. Заметим, что линии уровня исходного функционала являются сферами и явление овражности отсутствует.

На практике метод штрафных функций широко используется на начальных этапах оптимизации с применением таких, возможно больших, значений σ, для которых удается достигнуть относительно быстрого убывания $J_0(x)$ при достаточно точном выполнении ограничений. Для последующего улучшения решения привлекаются более тонкие стратегии, которые, как правило, оказываются и существенно более трудоемкими. Кроме этого метод штрафных функций до сих пор не имеет разумных альтернатив в ряде критических ситуаций, характерных для прикладных задач параметрической оптимизации.

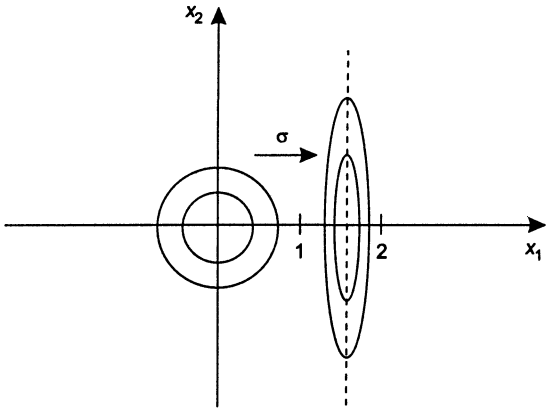

Рис. 3.5. Метод штрафных функций

Например, при наличии вырожденного минимума, когда не выполняется условие линейной независимости градиентов ограничений $g_i'(x)$, могут потерять работоспособность все методы учета ограничений, основанные на обычной и модифицированной функции Лагранжа, а также на линеаризации ограничений. Метод же штрафных функций в указанной ситуации применим. Он оказывается наименее чуствительным ко всем формам вырождения.

Второй критической ситуацией, возникающей на практике, является несогласованность требований-спецификаций к объекту оптимизации, которая приводит к пустому множеству D допустимых значений управляемых параметров. Предположим, что решается задача с ограничениями-равенствами. В этом случае снова наиболее целесообразным образом ведет себя метод штрафных функций, позволяя получить такое решение задачи $\|g(x)\| \to \min$, для которого величина $J(x)$ минимальна. Другие методы либо теряют смысл, либо заведомо не будут сходящимися.

Более подробно эти довольно тонкие вопросы могут быть изучены по учебнику [48].

В силу изложенного наличие алгоритмов оптимизации, сохраняющих работоспособность при достаточно высокой степени овражности минимизируемых функционалов, оказывается чрезвычайно желательным. Этот вывод подтверждается также тем фактом, что и при использовании модифицированных функций Лагранжа мы сталкиваемся с овражной ситуацией, хотя и в несколько ослабленной форме. Как показано в учебнике [19], выбор параметра σ в модифицированных функциях Лагранжа весьма сильно влияет на обусловленность соответствующей канонической задачи. При этом как слишком малые, так и слишком большие σ приводят к овражной ситуации. Если учесть, что надежные рекомендации по априорному заданию σ в настоящее время отсутствуют, то становится совершенно ясна необходимость применения специальных овражно-ориентированных процедур для решения вспомогательных канонических задач.

Пример. Решается задача [19]

$$J(x) = x_1 x_2^2 \to \min,\ x \in D,$$
$$D = \left\{ x \in R^n \,\middle|\, g(x) = 2 - x_1^2 - x_2^2 = 0 \right\}.$$

На рис. 3.6. показаны линии уровня вспомогательного функционала

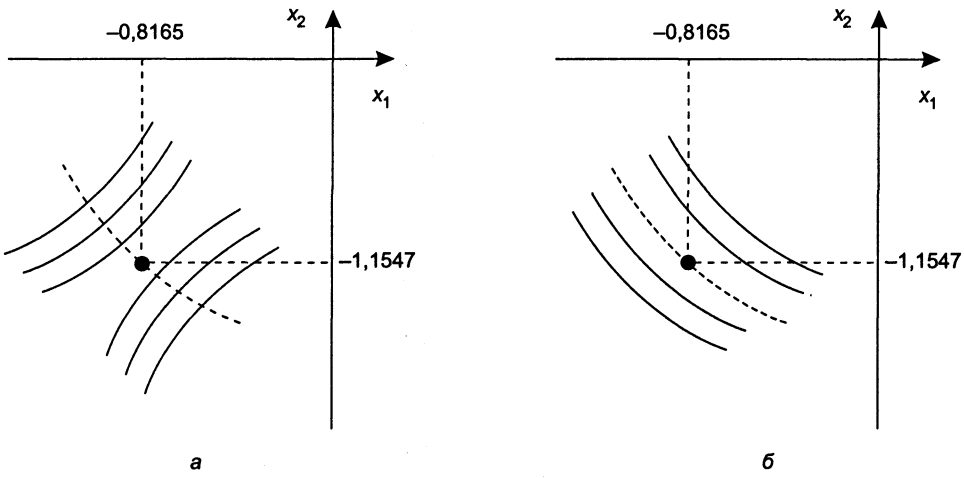

Рис. 3.6. Линии уровня модифицированной функции Лагранжа при *а* — σ = 0,075; *б* — σ = 100

$$M(x^*, \sigma) = J(x) + \lambda g(x) + \frac{\sigma}{2} g^2(x)$$

при $\lambda = \lambda^* \cong 0{,}82$; $\sigma^* \cong 0{,}2$.

В обоих случаях линии уровня практически параллельны, что и определяет наличие овражной ситуации.

Объединение конфликтных выходных параметров

Учет многокритериальных требований к объекту оптимизации с помощью единого критерия оптимальности также является важнейшим фактором, обусловливающим возникновение овражной ситуации. Изложенные в 2.3 принципы построения Парето-оптимальных решений приводят к двум основным видам сверток: линейной и минимаксной (максиминной). Остановимся на минимаксной свертке двух частных критериев:

$$J(x) = \max\{\alpha_1 J_1(x),\ \alpha_2 J_2(x)\} \to \min_x,\ \alpha_i > 0,\ \alpha_1 + \alpha_2 = 1.$$

Очень часто отдельные критериальные выходные параметры как функции от входных параметров имеют монотонный, существенно неовражный характер. Однако и в этом случае их объединение почти неизбежно приводит к овражной ситуации. При этом крутые «склоны» оврага характеризуют доминирующее влияние на обобщенный критерий какого-то одного из «частных» критериев. Как следует из рис. 3.7, объединение критериев J_1, J_2 приводит к образованию «клювообразной» зависимости, порождающей в многомерном случае овраг с крутыми склонами. Характерно, что движение по любой поверхности $\alpha_i J_i$ в отдельности с помощью практически любых методов оптимизации может не вызывать никаких затруднений.

В качестве второго примера рассмотрим критерии оптимальности, построенные на основе МНК-функционалов вида

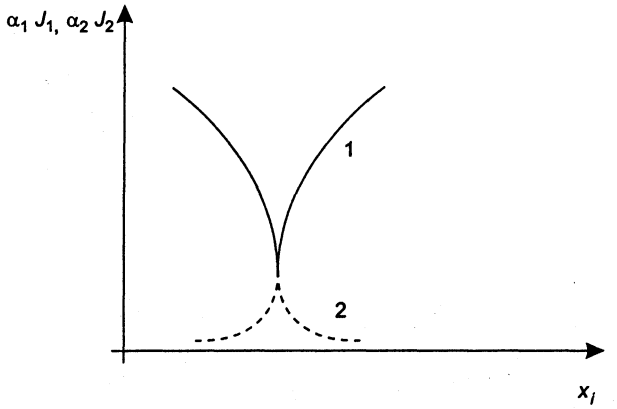

Рис. 3.7. Объединение частных критериев: *1* – $\alpha_1 J_1$; *2* – $\alpha_2 J_2$

$$J(x)=\sum_{i=1}^{N}\alpha_i f_i(x),\ f_i(x)\triangleq\left[\sum_{j=1}^{n}x_j\varphi_j(t_i)-y_i\right]^2, \tag{3.27}$$

где $y_i=y(t_i)$; $\{\varphi_j,\ j\in[1:n]\}$ — заданная система функций. Задача $J(x)\to\min$ ставится при решении задач аппроксимации

$$y(t)\cong\sum_{j=1}^{n}x_j\varphi_j(t),\ \ 0\le t\le T, \tag{3.28}$$

характерных для многих разделов теории управления. Например, необходимость в применении параметризации (3.28) возникает в задачах сглаживания экспериментальных зависимостей при идентификации стохастических объектов, а также в задачах идентификации нелинейных объектов на основе функциональных рядов Вольтерра (см. 1.3). Уже обсуждавшиеся ранее алгоритмические методы решения уравнения Винера—Хопфа также приводят к задачам минимизации МНК-функционалов.

Функционал $J(x)$ можно рассматривать как линейную свертку функционалов $f_i(x)$, характеризующих частные критерии оптимальности

$$f_i(x)\to\min_x,\ i\in[1:N].$$

Поэтому трудности минимизации (3.27) в какой-то степени отражают общую ситуацию, связанную с применением линейных сверток при большом числе частных критериев. В наиболее употребительном на практике случае используется полиномиальная аппроксимация на основе системы базисных функций вида

$$\varphi_j(t)=t^{j-1},\ 1\le j\le n.$$

Полагая N достаточно большим и заменяя сумму интегралом

$$J(x)=\sum_{i=1}^{N}\alpha_i f_i(x)\cong\int_0^T\alpha(t)\left[\sum_{j=1}^{N}x_j\varphi_j(t)-y(t)\right]^2dt,$$

получим при $\alpha(t) \equiv 1$:

$$J''(x) = H_n,\ H_n = \{h_{ij}\},\ \ h_{ij} = \frac{1}{i+j-1},\ \ i,j \in [1:n],$$

где H_n – $(n \times n)$-матрица Гильберта, являющаяся также матрицей линейной системы так называемых *нормальных уравнений* $J'(x) = 0$. Спектральное число обусловленности положительно определенной матрицы H_n, совпадающее со степенью овражности квадратичного сильно выпуклого функционала $J(x)$, оказывается достаточно высоким: матрицы Гильберта являются стандартным примером очень плохо обусловленных матриц. При этом cond $[H_n]$ быстро растет с увеличением n:

n	2	3	4	5
cond $[H_n]$	$1{,}93 \cdot 10^1$	$5{,}24 \cdot 10^2$	$1{,}55 \cdot 10^4$	$4{,}77 \cdot 10^5$
n	6	7	8	9
cond $[H_n]$	$1{,}50 \cdot 10^7$	$4{,}75 \cdot 10^8$	$1{,}53 \cdot 10^{10}$	$4{,}93 \cdot 10^{11}$

В результате уже для небольших n минимизация (3.27) происходит в условиях сильно выраженной овражной ситуации.

3.5. Методы параметрической оптимизации

В 2.7 сформулированы особенности канонических оптимизационных задач, возникающих в теории управления. В связи с этим далее обсуждаются возможности некоторых современных методов параметрической оптимизации. Основное внимание уделяется анализу работоспособности алгоритмов в характерных для реальных компьютерных вычислений условиях невыпуклой и одновременно овражной ситуации. Набор рассматриваемых методов по необходимости ограничен и включает в основном ньютоновские и квазиньютоновские процедуры, считающиеся в настоящее время наиболее эффективными при решении задач с гладкими функционалами. Мы дадим не очень подробный обзор методов, так как этот материал довольно полно представлен в учебной литературе.

3.5.1. Ньютоновские методы

Ньютоновские методы (Н-методы) генерируют последовательность точек $\{x^k\}$ по правилу

$$x^{k+1} = x^k + h_k p_N^k,\ h_k \in R^1, \tag{3.29}$$

где $p_N^k = -[G_k]^{-1} g^k$ определяет так называемое *ньютоновское* направление спуска; h_k – длина шага; $G_k \triangleq J''(x^k)$; $g \triangleq J'(x^k)$. При использовании алгоритма (3.29) предполагается, что все матрицы G_k положительно определены $(G_k > 0)$. Последнее условие гарантирует разрешимость задачи вычисления p_N^k, исходя из системы уравнений

$$G_k p_N^k = -g^k. \tag{3.30}$$

Вектор x^{k+1}, построенный согласно (3.29), (3.30) при $h_k = 1$ является минимизатором аппроксимирующего квадратичного функционала (*параболоида*), являющегося отрезком соответствующего ряда Тейлора:

$$F_k(x) = J(x^k) + \langle g^k, x - x^k \rangle + \frac{1}{2}\langle G_k(x - x^k), x - x^k \rangle. \tag{3.31}$$

Поэтому метод (3.29) называется также методом параболоидов [44].

Н-методы (3.29) оказываются весьма эффективными при решении задач безусловной минимизации выпуклых функционалов. Некоторые авторы утверждают, что это наиболее эффективные из всех применяемых ими для решения реальных задач методов при условии выпуклости целевого функционала.

Если матрица G_k не является положительно определенной, параболоид (3.31) может не иметь конечного минимизатора и, как правило, будет неограничен снизу. Поэтому для невыпуклых целевых функционалов процедура (3.30) непосредственно неприменима. Ньютоновское направление p_N^k при наличии отрицательных спектральных составляющих матрицы G_k может указывать в сторону возрастания функционала $J(x)$, что соответствует расходимости процесса.

Как отмечается в книге [19], «в настоящее время нет общепринятого определения "метода Ньютона" для расчета направления спуска при знаконеопределенной матрице G_k, поскольку среди специалистов нет согласия относительно того, как использовать локальную квадратичную аппроксимацию F_k в этом случае». Такая ситуация сохраняется и поныне.

Наиболее употребительны модификации Н-метода, в которых направление спуска p^k находится из решения линейной системы

$$\overline{G}_k p^k = -g^k, \tag{3.32}$$

где $\overline{G}_k$ — некоторая положительно определенная матрица, совпадающая с исходной матрицей Гессе G_k, если последняя положительно определена. Указанный метод выбора направления p^k гарантирует глобальную сходимость процедуры независимо от характера выпуклости функционала $J(x)$ в окрестности текущей точки x^k.

Далее рассмотрены наиболее известные процедуры построения $\overline{G}_k$ на основе различных матричных разложений, позволяющих учитывать знаки собственных чисел G_k.

Методы, основанные на спектральном разложении

Спектральное разложение имеет вид

$$G_k = U\Lambda U^{\mathrm{T}} = \{g_{ij}\},\ g_{ij} = \sum_{m=1}^{n} \lambda_m u_i^m u_j^m, \tag{3.33}$$

где $U = [u^1, u^2, \ldots, u^n]$; $\Lambda = \mathrm{diag}\,(\lambda_i)$; $\{u^m, m \in [1{:}n]\}$ — система ортонормальных собственных векторов матрицы G_k; $\{\lambda_m, m \in [1{:}n]\}$ — спектр матрицы G_k. Базовая схема, реализующая рассматриваемый подход, основана на выборе

$$\overline{G}_k = U\overline{\Lambda}U^{\mathrm{T}},\ \overline{\Lambda} = \mathrm{diag}\left[\{|\lambda_i|, \delta\}\right], \tag{3.34}$$

где $\delta > 0$ — параметр метода, определяющий границу *«существенной положительности»* любого из собственных чисел. Отмечаемый в литературе недостаток метода (3.32) с матрицей (3.34) связывается с трудоемкостью процедуры построения спектрального разложения (3.33), требующей до $4n^3$ арифметических операций. Кроме этого, возникают известные из вычислительной линейной алгебры трудности в определении малых собственных чисел плохо обусловленной матрицы. Выбор параметра δ также до конца не алгоритмизирован.

Методы, основанные на модифицированной факторизации Холесского

Разложение Холесского для симметричной положительно определенной матрицы B имеет вид

$$B = LDL^{\mathrm{T}},\ D = \mathrm{diag}\,(d_i), \tag{3.35}$$

где L — нижняя треугольная матрица с единицами на диагонали; D — положительная диагональная матрица. Факторизация (3.35) непосредственно неприменима к знаконеопределенной симметричной матрице $B = G_k$. В работе [19] предложена модифицированная процедура, позволяющая равномерно ограничить рост элементов треугольного фактора L на уровне $|r_{ik}| \le \beta \left(i > k,\ r_{ik} \triangleq l_{ik}\sqrt{d_k}\right)$ и гарантирующая «существенную» положительность $d_i > \delta > 0$ диагональных элементов матрицы D. В результате получаем

$$\overline{G}_k = LDL^{\mathrm{T}} = G_k + C,\ \overline{G}_k > 0, \tag{3.36}$$

где C — неотрицательная диагональная матрица, зависящая от выбранного параметра β. Для сохранения численной устойчивости процедуры построения $\overline{G}_k$, а также для совпадения $\overline{G}_k$ и G_k в случае положительно определенной G_k целесообразно вычислять β из условия

$$\beta^2 = \max\left\{\gamma, \frac{\xi}{\sqrt{n^2 - 1}}, \varepsilon_{\text{м}}\right\},$$

где $\varepsilon_{\text{м}}$ — машинное эпсилон; ξ — максимальный модуль недиагонального элемента G_k; γ — значение максимального из диагональных элементов G_k.

Основная цель построения разложения (3.36) заключается в сокращении вычислительных затрат по сравнению с (3.34) приблизительно до $\frac{1}{6}n^3$ арифметических операций. Существо дела при этом не затрагивается.

Аналогичные модификации метода Ньютона, основанные на других численно устойчивых процедурах факторизации знаконеопределенных симметричных матриц, рассматриваются также во многих работах других исследователей. Большое число публикаций, посвященных рассматриваемым вопросам, указывает, с одной стороны, на актуальность проблемы, а с другой — на отсутствие метода, полностью отвечающего предъявляемым практикой требованиям.

3.5.2. Методы доверительной окрестности

Основная идея *методов доверительной окрестности* (МДО) сводится к следующему [19]. Мы рассматриваем *квадратичные* методы, то есть методы, основанные на последовательной квадратичной аппроксимации минимизируемой функции. Надежность прогноза в таких методах определяется областью справедливости (достаточной точности) локальной квадратичной модели. В методах Ньютона мы определяем направление в сторону минимума аппроксимирующей квадратичной зависимости, а затем регулируем норму вектора продвижения, чтобы не уйти «слишком далеко». Здесь же предлагается поступать по иному, а именно, решать задачу условной минимизации аппроксимирующей квадратичной функции при условии, что норма вектора продвижения ограничена сверху некоторым заданным параметром.

Таким образом, минимизирующая последовательность строится по правилу $x^{k+1} = x^k + p^k$, где вектор p^k на каждом шаге определяется как решение вспомогательной задачи вида

$$\frac{1}{2}\langle G_k p, p\rangle + \langle g^k, p\rangle \to \min,\ p \in D \subset R^n;$$
$$D = \left\{ p \in R^n \middle| \langle p, p\rangle^2 \le \Delta \right\}. \tag{3.37}$$

Величина Δ характеризует область (окрестность) «квадратичности» исходного функционала. Сформулированная задача может быть решена известным из курсов высшей математики и численного анализа *методом множителей Лагранжа.*

Условие стационарности функции Лагранжа для задачи (3.37) приводит к методу определения p из системы линейных уравнений

$$[G_k + \beta E]p = -g^k,\ \beta > 0, \tag{3.38}$$

где β играет роль множителя Лагранжа. Алгоритм выбора β зависит от конкретной реализации МДО. Возможен непосредственный подбор оптимального β на основе многократного решения линейной системы (3.38). В ряде случаев [19] вначале полагают $\beta = 0$. Если в процессе решения (3.38) при $\beta = 0$ выясняется, что G_k знаконеопределена или $\|p\| > \Delta$, где Δ — установленное пороговое значение, то β определяется как решение нелинейного уравнения вида

$$\left\| [G_k + \beta E]^{-1} g^k \right\| = \Delta. \tag{3.39}$$

Далее по найденному β согласно (3.38) определяется искомый вектор p.

Основной недостаток всех методов рассматриваемого класса состоит в необходимости решения на каждом шаге итерационного процесса нелинейных алгебраических уравнений типа (3.39), что требует привлечения различных процедур регуляризации. Кроме этого, найденному оптимальному значению может отвечать очень плохо обусловленная матрица $G_k + \beta E$, что приводит к значительным вычислительным трудностям при решении системы (3.38). Аналогичные недостатки присущи и ньютоновским методам, рассмотренным в 3.5.1.

Указанные проблемы становятся практически неразрешимыми, если степень обусловленности матрицы G_k настолько высока, что информация о малых спектральных составляющих оказывается полностью утерянной на фоне больших собственных чисел за счет погрешностей задания G_k. Эти погрешности могут вызываться приближенным характером соотношений, применяемых для расчета производных, а также ограниченностью разрядной сетки компьютера. В указанных условиях, даже при использовании регуляризированных форм матричных разложений, гарантирующих, в частности, положительную определенность матрицы $\overline{G}_k$, генерируемые векторы p^k, оставаясь направлениями спуска, оказываются практически случайными, что резко замедляет сходимость соответствующей поисковой процедуры.

3.5.3. Квазиньютоновские методы

Квазиньютоновские методы (КН-методы), так же как и ньютоновские, основаны на использовании квадратичных моделей минимизируемых функционалов. Однако аппроксимация матрицы Гессе (либо обратной к ней) осуществляется последовательно на основе наблюдений за изменением градиентов целевых функционалов в последовательных точках минимизирующей последовательности.

Часто указывается, что КН-методы существенно более эффективны, чем Н-методы, по причине более низкого порядка производных, определяющих схему метода. Легко, однако, видеть, что реальная ситуация значительно сложнее. Пусть, например, минимизируется сильно выпуклый квадратичный функционал $J(x)$. Тогда для построения оптимизатора x^* КН-методу потребуется в общем случае вычислить n градиентов, где n — размерность пространства поиска. Последнее эквивалентно $2n^2$ вычислениям значений функционала $J(x)$ если используются двусторонние конечно-разностные аппроксимации первых производных. Указанного количества значений функционала $J(x)$, очевидно, достаточно для построения аппроксимации $J''(x)$, необходимой для реализации любого из вариантов Н-метода. Следовательно, трудоемкости рассматриваемых процедур в указанных условиях приблизительно равны, если не учитывать дополнительные вычислительные затраты на процедуры одномерного поиска в КН-методах. С другой стороны, доказано, что большинство вариантов КН-методов (например, одна из наиболее эффективных схем Бройдена—Флетчера—Гольдфарба—Шенно) при минимизации сильно выпуклых квадратичных функционалов приводят к одной и той же траектории спуска, вырождаясь в хорошо изученные методы сопряженных градиентов (СГ). В то же время известна особенность методов СГ, существенно ограничивающая область их эффективного применения. Она заключается в понижении скорости сходимости для плохо обусловленных задач оптимизации. Оценки, приведенные, например, в учебнике [48], показывают, что стандартные методы СГ сходятся по закону геометрической прогрессии со знаменателем q, близким к единице: $q \cong 1 - \frac{2}{\sqrt{\eta}}$, где η — степень овражности минимизируемого функционала. Там же имеются указания на достаточно высокую скорость сходимости метода СГ «по функционалу», независимо от величины η:

$$J(x^k) - J(x^*) \le \frac{L\left\| x^0 - x^* \right\|^2}{2(2k+1)^2}, \; L = \text{const}. \qquad (3.40)$$

Однако оценки типа (3.40) получены в предположении $G(x) > 0$. Кроме этого, согласно (3.40) достаточно эффективно получаются значения функционала порядка $J(x^*)$, где x^* — минимизатор аппроксимирущего параболоида $f(x)$, что в общем случае не решает задачи. Сказанное иллюстрируется на рис. 3.8. Отрезок $[x', x'']$ в условиях высоких значений η будет преодолеваться методом СГ, а вместе с ним и КН-методом, с малым шагом по аргументу, хотя $f(x') \cong f(x'') \cong f(x^*)$. Предположение о невыпуклости вносит дополнительные трудности. Можно показать, что в этих условиях метод СГ по характеристикам сходимости эквивалентен градиентному методу наискорейшего спуска со всеми вытекающими отсюда последствиями. Кроме отмеченных дефектов, общих для СГ- и КН-методов, последние имеют дополнительные недостатки, связанные с проблемой потери положительной определенности квазиньютоновских матриц за счет накопления вычислительных погрешностей в рекуррентных процедурах аппроксимации матриц Гессе [19].

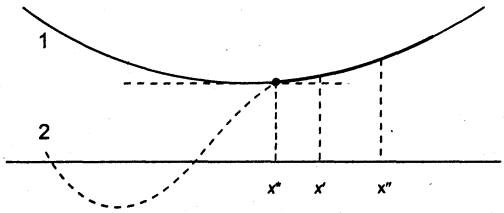

Рис. 3.8. Метод СГ:
1 — аппроксимирующий на участке $[x', x'']$ параболоид;
2 — целевой функционал

В настоящее время на основе экспериментальных данных принята точка зрения, согласно которой Н-методы с конечно-разностной аппроксимацией матрицы Гессе на основе аналитических выражений для градиентов надежнее, чем КН-методы, и сходятся в более широком классе задач, в частности, в задачах с очень плохой обусловленностью матрицы Гессе. С другой стороны, согласно тем же источникам, если аналитические выражения для градиентов отсутствуют и применяются версии нулевого порядка соответствующих алгоритмов, в большинстве случаев более эффективными оказываются КН-методы с конечно-разностной аппроксимацией градиентов. Данные замечания не снимают отмеченные выше недостатки указанных процедур, а отражают лишь некоторые сравнительные оценки, основанные на опыте проведения реальных вычислений.

3.5.4. Задачи высокой размерности

При решении канонических оптимизационных задач с высокой размерностью n аргумента минимизируемого функционала возникают значительные трудности, вызванные существенным возрастанием времени вычислений, с помощью традиционных алгоритмов нулевого и первого порядков. Эти трудности могут стать непреодолимыми, если функционал характеризуется высокой степенью овраж-

ности в широкой области изменения аргументов. С другой стороны, стандартные быстросходящиеся КН-схемы минимизации типа методов Давидона—Флетчера—Пауэлла, Бройдена, Пирсона, Мак-Кормика, Бройдена—Флетчера— Гольдфарба— Шенно, Пауэлла—Бройдена и др. не могут непосредственно применяться для оптимизации больших систем из-за потери свойства разреженности матриц и нехватки оперативной памяти компьютера. Применение Н-методов для минимизации функционалов с разреженными и структурированными матрицами Гессе алгоритмически возможно, так как при этом удается использовать специальную структуру обычно сильно разреженных матриц Гессе. Соответствующие Н-методы, основанные на различных предположениях о структуре матриц Гессе, представлены, например, в [19]. Однако при отсутствии выпуклости функционала в окрестности текущей точки генерируемые Н-методами направления спуска оказываются локально неэффективными. Возникающие трудности связаны с указанными в 3.5.2 недостатками модифицированных Н-методов и методов доверительной окрестности. Кроме этого возникают специфические для задач высокой размерности дополнительные трудности, обусловленные необходимостью решения больших систем линейных уравнений вида

$$\overline{G}_k p^k = -g^k \tag{3.41}$$

с разреженной матрицей $\overline{G}_k$. Дело заключается в том, что решение (3.41) обычно проводится на основе некоторой процедуры факторизации положительно определенной матрицы $\overline{G}_k$. Наиболее часто применяется симметричный вариант гауссова исключения, основанный на факторизации Холесского. При этом матрица L в представлении $\overline{G}_k = LL^T$ может оказаться заполненной несмотря на разреженность исходной матрицы $\overline{G}_k$. Развитые в работе [23] методы оптимального упорядочения строк и столбцов матрицы линейной системы позволяют достичь определенной экономии памяти и сократить число вновь появляющихся ненулевых элементов. Однако проблема в целом не решается. Кроме того, становятся, по существу, неприемлемыми методы, основанные на модифицированном разложении Холесского (см. 3.5.1), так как в данном случае непосредственно не удается воспользоваться результатами упомянутой работы [23].

Исследования по построению аналогов КН-методов, ориентированных на решение задач высокой размерности, пока находятся в ранней стадии, и, кроме того, этим методам присущ общий недостаток КН-методов, связанный с их низкой эффективностью в невыпуклой и одновременно овражной ситуации, характерной для базовых задач теории управления.

В настоящее время общепринятым средством решения канонических задач высокой размерности являются методы сопряженных градиентов (СГ). Однако их заведомо низкая скорость сходимости в условиях невыпуклости минимизируемых функционалов не позволяет считать вопрос решенным.

3.5.5. Глобальная оптимизация

Все рассмотренные ранее методы являются «локальными», так как с их помощью может быть найден один из локальных минимумов функционала $J(x)$. С точки зрения практики представляет интерес поиск глобального минимума,

то есть такого локального минимума, где значение критерия оптимальности оказывается наименьшим.

Трудность вопроса заключается в том, что для произвольного функционала $J(x)$ задача глобальной оптимизации неразрешима с помощью вычислений $J(x)$ в любом сколь угодно большом, но конечном числе точек. Поэтому алгоритмы глобальной оптимизации должны развиваться для достаточно узких классов задач на основе имеющейся априорной информации.

В настоящее время известен один довольно общий класс критериев оптимальности, для которых обеспечивается возможность локализации глобального минимума за обозримое машинное время. Речь идет о классе липшицевых функционалов, удовлетворяющих условиям

$$|J(x') - J(x'')| \le L\|x' - x''\|, \; L = \text{const} > 0.$$

Соответствующие методы глобальной оптимизации изложены в учебниках [13], [48]. Недостатком этих методов является требование знания константы Липшица для всей области изменения x. Неправильное назначение L может резко замедлить метод либо привести к потере глобального минимума. Вопросы глобальной оптимизации, включая и различные эвристические процедуры, рассмотрены также в [57].

Реальная ситуация в области глобальной оптимизации в настоящее время расценивается как неблагоприятная. Существующие методы поиска глобального экстремума, особенно в овражной ситуации, не могут рассматриваться как исчерпывающие при решении задач достаточно высокой размерности.

Наиболее распространенный и эффективный эвристический метод заключается в задании некоторой грубой сетки начальных точек в допустимом множестве с последующим применением методов локальной оптимизации. Для построения таких сеток целесообразно применять $ЛП_\tau$-последовательности, обладающие свойством равномерного заполнения многомерной области [55].

В качестве начальных точек для локальных процедур спуска могут использоваться только некоторые узлы сетки, которым отвечают наименьшие значения функционала. Таким образом, в настоящее время основным инструментом параметрической оптимизации продолжают оставаться локальные методы.

В заключение отметим, что в ряде случаев проблема многоэкстремальности возникает в результате определенного непонимания реальной ситуации. Регистрируемые на практике многочисленные «локальные экстремумы» в действительности оказываются точками остановки применяемых поисковых процедур. Можно утверждать, что наличие многих локальных минимумов в практических задачах управления встречается значительно реже, чем об этом принято говорить (за исключением специально сконструированных тестовых многоэкстремальных задач). Данная точка зрения подтверждается в известных работах [48], [66].

3.5.6. Анализ сложившейся ситуации

Указанные в 2.7 основные особенности канонических оптимизационных задач, возникающих в теории управления, определяют требования к соответствующим

алгоритмам. Вычислительная практика показывает, что основные трудности при решении собственно задач оптимизации связаны с одновременным присутствием двух факторов сложности, обусловленных невыпуклостью минимизируемых функционалов и их плохой обусловленностью. В ряде случаев присоединяется третий фактор, определяемый высокой размерностью вектора варьируемых параметров.

Кроме задач поиска минимизаторов целевых функционалов часто возникают сопутствующие задачи, решение которых также затруднено в условиях овражной ситуации. В частности, важное практическое значение имеют методы построения некоторого вектора $\hat{x}$, отличающегося от минимизатора x^* функционала $J(x)$ рядом компонент, имеющих предписанные (например, нулевые) значения при условии $J(\hat{x}) \le J(x^*) + \delta J$, где δJ — заданное допустимое отклонение целевого функционала от оптимального (минимального) значения.

Подобные формальные модели могут иметь различную интерпретацию. Например, зануление ряда компонент вектора x может трактоваться как упрощение исходной структуры оптимизируемой системы, приводя к некоторому алгоритму структурного синтеза. Кроме того, указанные алгоритмы построения вектора $\hat{x}$ имеют большое значение при поиске «минимальных» параметрических представлений непрерывных функций в следующих задачах управления:

- задачах идентификации нелинейных детерминированных объектов с применением рядов Вольтерра;
- задачах идентификации стохастических объектов на основе методов сглаживания;
- задачах параметризации при идентификаций существенно нестационарных объектов,

а также в целом ряде других задач, основанных на параметрическом представлении искомых непрерывных зависимостей. Регулярные методы решения указанного класса задач в условиях плохой обусловленности $J(x)$ будут изложены в последующих главах.

Учитывая изложенное, а также основные характеристики стандартных методов параметрической оптимизации, рассмотренных в 3.5.1–3.5.3, можно поставить следующие важные в прикладном аспекте задачи по совершенствованию имеющегося алгоритмического и программного обеспечения:

- разработка общих процедур параметрической оптимизации, сохраняющих эффективность в условиях невыпуклости целевых функционалов в предположении их гладкости и «кусочной квадратичности»;
- разработка проблемно-ориентированных реализаций алгоритмов, рассчитанных на конкретные схемы параметрической оптимизации, характерные для отдельных стандартных классов задач теории управления;
- разработка методов решения задач параметрической оптимизации большой размерности;
- разработка элементов структурного синтеза и принципов построения «минимальных» параметрических представлений искомых непрерывных функций.

Решению указанных задач в предположении высокой степени обусловленности целевых функционалов, а также рассмотрению вопросов, связанных с созданием необходимого программного обеспечения, посвящено дальнейшее изложение. Предполагается, что все решаемые задачи параметрической оптимизации разделены на два класса по степени обусловленности матриц Гессе минимизируемых функционалов. В первый класс входят оптимизационные задачи с «умеренной» степенью овражности, для которых характерна «информативность» матриц Гессе целевых функционалов по малым спектральным составляющим. Иначе говоря, предполагается, что при условии точного решения полной проблемы собственных значений для матрицы $J''(x)$, записанной в явном виде в памяти компьютера, получается достаточно полная информация о ее спектральном составе. Именно в этих условиях, которые обычно специально не оговариваются, сохраняют работоспособность все известные варианты Н-методов.

Во второй класс включаются задачи со «сверхвысокой» степенью обусловленности. При этом предполагается, что при явном формировании матрицы $J''(x)$ в памяти компьютера в широкой области изменения аргумента x информация о малых спектральных составляющих полностью теряется.

Представленные в данной книге и рекомендуемые для практического использования матричные градиентные методы решения задач с «умеренной» степенью овражности являются квадратичными и имеют ньютоновский характер, так как основаны на построении квадратичной модели целевого функционала в окрестности каждой текущей точки минимизирующей последовательности. Однако в отличие от классического подхода их вычислительные схемы не используют конечные результаты решения плохо обусловленных систем линейных уравнений. Вместо этого применяются различные рекуррентные процедуры решения таких систем, в которых все промежуточные результаты имеют «физический смысл», обладают свойством релаксационности, и поэтому возможен непрерывный контроль точности, исключающий накопление вычислительных погрешностей до неприемлемого уровня. В качестве основы для построения методов и алгоритмов оптимизации в условиях «сверхвысокой» степени овражности выбран класс методов обобщенного покоординатного спуска, реализующих известную идею приведения квадратичного функционала к главным осям. Указанные методы, по существу, осуществляют локальную декомпозицию исходной задачи на несколько подзадач с малой степенью овражности, что приводит к необходимому вычислительному эффекту.

3.6. Основные результаты и выводы

В данной главе представлены следующие основные результаты:

1. Обоснована необходимость выделения плохо обусловленных (овражных) оптимизационных задач в отдельный класс, обусловленная, с одной стороны, значительными вычислительными трудностями при их решении стандартными средствами, а с другой стороны, экспериментально установленным фактом типичности овражной ситуации для большинства задач теории управления.

2. Выявлены и в явном виде сформулированы следующие основные причины появления плохо обусловленных экстремальных задач в теории управления:
   - частные структурные особенности решаемой задачи, например, некорректность задачи решения уравнения Винера—Хопфа, играющего важную роль во многих разделах теории управления;
   - наличие фактора агрегированности управляющих параметров и отсутствие регулярных методов выделения агрегатов;
   - жесткость систем обыкновенных дифференциальных уравнений, описывающих динамику объекта оптимизации;
   - специальные структуры расширенных целевых функционалов, получаемых при учете ограничений методами штрафных функций и модифицированных функций Лагранжа;
   - специальные структуры обобщенных целевых функционалов при решении многокритериальных оптимизационных задач на основе линейных и минимаксных (максиминных) сверток.
3. Дано формальное описание явления плохой обусловленности (определения 2, 3), оказывающееся более конструктивным для теории оптимизации по сравнению с общим определением 1, опирающимся на теорию жестких систем. Введено понятие степени овражности, отличное от обычно используемого спектрального числа обусловленности матрицы Гессе и более точно отражающее возникающие в процессе оптимизации вычислительные трудности. В частности, показано, что из высокой степени овражности всегда следует плохая обусловленность матрицы Гессе, а обратное, вообще говоря, неверно. Последнее обстоятельство обычно игнорировалось не только в учебной, но и научной литературе, что не позволяло адекватно оценивать трудность решаемой задачи и эффективность применяемых алгоритмов при минимизации невыпуклых целевых функционалов. Показана *регулярность* введенных формальных моделей овражных функционалов: для случая квадратичных функционалов доказано совпадение «дна оврага» с линейной оболочкой собственных векторов матрицы Гессе, отвечающих «малым» собственным значениям.
4. Описаны алгоритмические методы распознавания овражной ситуации, играющие роль критериев овражности. Построена оценка степени овражности, позволяющая по результатам работы метода ПГС указать нижнюю границу для степени овражности, что имеет первостепенное значение при оценке трудности решаемой задачи и при автоматизации процесса выбора алгоритма параметрической оптимизации. Кроме того, предложена альтернативная методика оценки степени овражности на основе прямого спектрального разложения матрицы Гессе с учетом погрешностей вычислений. Указаны качественные признаки плохой обусловленности задачи, выражающиеся в регистрации ложных локальных минимумов.
5. Дан анализ стандартного алгоритмического обеспечения (включающего ньютоновские методы, методы доверительной окрестности, квазиньютоновские методы и методы сопряженных градиентов), показавший актуальность задачи его совершенствования по следующим основным направлениям:

- построение общих процедур решения плохо обусловленных задач, сохраняющих эффективность в условиях невыпуклости целевых функционалов;
- построение проблемно-ориентированных реализаций алгоритмов, рассчитанных на конкретные схемы параметрической оптимизации, характерные для отдельных классов задач теории управления;
- построение методов решения плохо обусловленных задач параметрической оптимизации большой размерности;
- разработка элементов структурного синтеза и принципов построения минимальных параметрических представлений искомых непрерывных зависимостей на основе методов удаления переменных.

Глава 4. Покоординатные стратегии параметрической оптимизации

4.1. Метод циклического покоординатного спуска

Решается каноническая задача построения минимизирующей последовательности $\{x^k\}$ для функционала $J(x)$.

Переход от вектора x^i к вектору x^{i+1} по методу циклического покоординатного спуска (ЦПС) происходит следующим образом: для $l \in [1:n]$ компонента x_l^{i+1} определяется как

$$x_l^{i+1} \in \operatorname{Arg}\min_{x \in R^1} J(x_1^{i+1}, x_2^{i+1}, \ldots, x_{l-1}^{i+1}, x, x_{l+1}^{i}, \ldots, x_n^{i}).$$

Для плохо обусловленных экстремальных задач этот метод применим в специальных случаях ориентации оврагов вдоль координатных осей. Трудности применения процедур, построенных на основе метода ЦПС, проиллюстрированы на рис. 4.1. Продвижение к точке минимума становится замедленным при наличии вытянутых поверхностей уровня. Как правило, имеет место ситуация «заклинивания», вызываемая дискретным характером представления информации в памяти компьютера.

Множество чисел F в форме с плавающей запятой, которые могут быть представлены в компьютере, конечно и содержит $2(\beta - 1)\beta^{t-1}(M - m + 1) + 1$ чисел вида

$$x = \pm\left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \ldots + \frac{d_t}{\beta^t}\right)\beta^e,$$

где $0 \le d_i \le \beta - 1$; d_i — целые числа; β — основание системы исчисления (обычно $\beta = 2$), t — длина мантиссы числа; показатель степени e лежит в заданном интервале $[m, M]$. Следовательно, плоскость (x_1, x_2) аппроксимируется также конечным множеством точек, лежащих в узлах сетки, показанной на рис. 4.2.

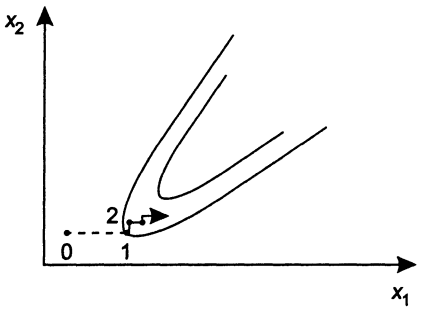

Рис. 4.1. Траектория спуска метода ЦПС при наличии овражной ситуации

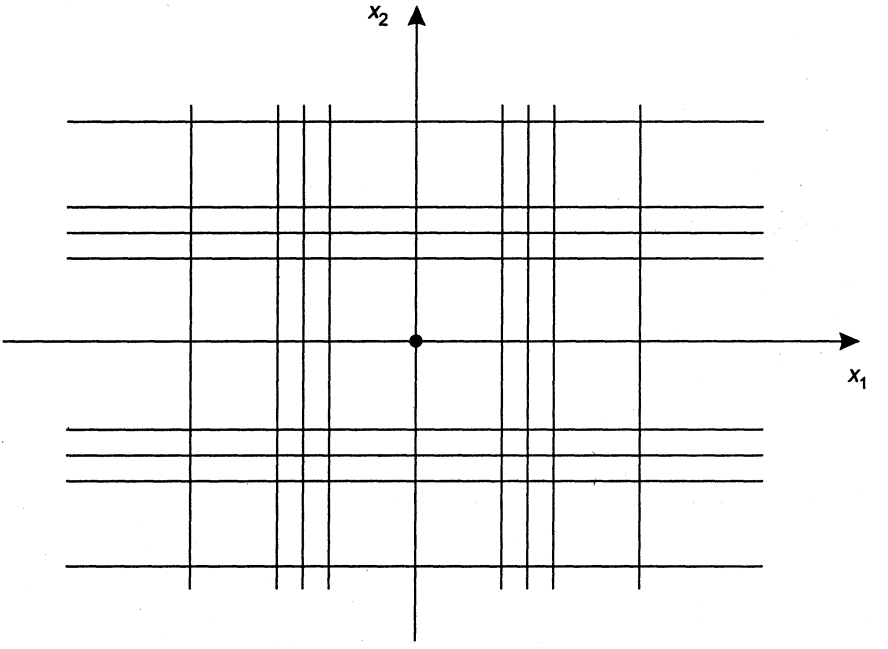

Рис. 4.2. Представление плоскости R^2 в памяти компьютера

Ситуация заклинивания представлена на рис. 4.3. Если процесс попадает в точку A, то ближайшие доступные точки B, B' и C, C' соответствуют большему значению функционала. Заклинивание (в англоязычной литературе — jamming) может происходить на значительных расстояниях от точки минимума, приводя к регистрации «ложного» локального минимума.

Вероятность заклинивания резко возрастает при использовании негладких функционалов, имеющих минимаксную или максиминную структуру. Типичный случай показан на рис. 4.4.

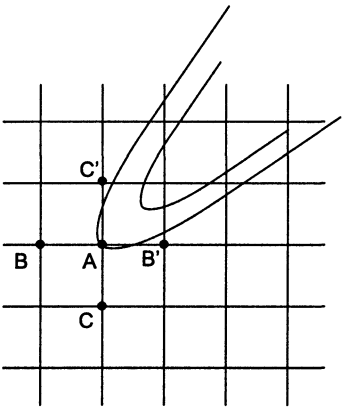

Рис. 4.3. Заклинивание метода ЦПС

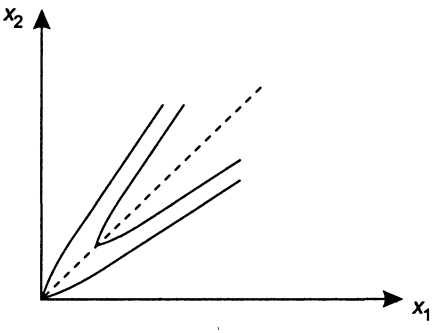

Рис. 4.4. Заклинивание в точке «излома» линии уровня

Такого типа ситуации, по существу, не являются следствием высокой степени овражности, и могут быть устранены переходом к гладким аппроксимациям, построенным по методике 2.8. В дальнейшем будет приведен пример практической ситуации, иллюстрирующий данное замечание.

Несмотря на отмеченную низкую эффективность метода ЦПС в овражной ситуации, его включение в библиотеку алгоритмов целесообразно при решении практически любого класса оптимизационных задач, по крайней мере, как стартового алгоритма с целью получения разумного начального приближения для последующих процедур. Причина этого заключается в высокой надежности метода по отношению к различным сбойным ситуациям, а также в простоте процесса подготовки задачи к моделированию на компьютере. Метод имеет нулевой порядок, то есть не требует включения в вычислительную схему информации о производных от минимизируемого функционала.

При реализации метода могут быть использованы известные из курса численного анализа способы одномерного поиска минимума типа золотого сечения, квадратичной аппроксимации и др. Однако это приводит к заметному и часто неоправданному усложнению алгоритма. Можно привести примеры, когда точный поиск минимума вдоль координатных направлений не только не обязателен, но

даже вреден. Поэтому на практике применяются более простые стратегии выбора шагов. Рассмотрим в качестве примера вариант, изложенный в работе [54].

Задается вектор начальных шагов $h = (h_1, \ldots, h_n)$ продвижений из точки x в направлении координатных ортов $e^1, e^2, \ldots, e^n$. Далее шаги h_i модифицируются от итерации к итерации. Если выполняется неравенство $J(x + h_i e^i) \le J(x)$, то текущая точка x заменяется на $x + h_i e^i$, а величина h_i утраивается: $h_i := 3h_i$. После этого осуществляется переход к следующему номеру i. Если $J(x + h_i e^i) > J(x)$, то производится умножение h_i на $-0{,}5$ и также осуществляется переход к следующему координатному орту. Таким образом, алгоритм адаптируется к конкретным условиям оптимизации за счет изменения величин и знаков шагов. Если начальные значения шагов были выбраны неудачно, то они быстро скорректируются до необходимых значений.

Указанный метод выбора координатных шагов достаточно эффективен и реализован в алгоритме GZ1. Это, по-видимому, простейшая из возможных реализаций метода ЦПС.

Алгоритм GZ1 [74].

Шаг 1. Ввести начальную точку $x = (x_1, \ldots, x_n)$ и шаг s; принять $F := J(x)$.

Шаг 2. Принять $h_i := s$, $i \in [1: n]$.

Шаг 3. Принять $m := 1$.

Шаг 4. Принять $x_m := x_m + h_m$; вычислить $F_1 = J(x)$.

Шаг 5. Если $F_1 \le F$, принять $h_m := 3h_m$, $F := F_1$ и перейти к шагу 7; иначе — перейти к шагу 6.

Шаг 6. Принять $x_m := x_m - h_m$, $h_m := -0{,}5h_m$.

Шаг 7. Принять $m := m + 1$. Если $m \le n$, перейти к шагу 4; иначе — к шагу 3.

Выход из алгоритма осуществляется по достижении заданного числа NFM вычислений J. Обычно программа составляется таким образом, чтобы обеспечить возможность продолжения работы с прерванного места после повторных входов в GZ1. В данном случае целесообразно отказаться от применения каких-либо внутренних критериев сходимости процесса оптимизации и обрывать его после заранее обусловленного числа шагов. При необходимости такой анализ сходимости может выполняться во внешней программе путем сравнения результатов, полученных при двух последовательных обращениях к GZ1.

Предполагается, что при каждом новом входе в алгоритм GZ1 счетчик числа вычислений функционала зануляется и, следовательно, разрешается еще NFM обращений к подпрограмме вычисления $J(x)$. Задавая различные значения NFM, можно регулировать частоту выходов из GZ1 во внешнюю программу для оценки и вывода получаемых результатов.

Из-за рассмотренного ранее явления «заклинивания» Розенброк в 1960 г. был вынужден модифицировать процедуру ЦПС. Подробное описание полученного алгоритма содержится в работе [54]. Основная идея заключается в организации процесса покоординатного спуска не вдоль фиксированных координатных ортов, а вдоль осей специальным образом выбираемой системы координат. При этом одна из осей должна составлять достаточно малый угол с образующей дна

одномерного оврага. В результате смещения по этой оси позволяют продвигаться вдоль дна оврага в направлении точки минимума. Схема метода Розенброка сводится к трем основным этапам. Пусть x^{m-1} и x^m — две соседние точки в последовательности $\{x^k\}$, построенной рассматриваемым методом. Тогда переход к x^{m+1} осуществляется следующим образом.

Шаг 1. Выбрать новую систему координат, первая ось которой направлена вдоль вектора $x^m - x^{m-1}$, а остальные дополняют ее до ортонормированного базиса («поворот осей»).

Шаг 2. В новой системе координат для поиска x^{m+1} осуществить алгоритм GZ1 до выполнения условия поворота осей.

Шаг 3. Возвратиться к старой системе координат и перейти к шагу 1.

Точка x^0 задается, а x^1 получается из x^0 с помощью алгоритма GZ1.

Различные модификации метода отличаются друг от друга способом организации одномерного поиска вдоль координатных ортов, способом построения ортогонального дополнения к оси $x^m - x^{m-1}$ (методом ортогонализации), а также выбором условия окончания процесса спуска для перехода к очередному повороту осей.

К недостаткам всех вариантов метода Розенброка следует отнести невозможность продолжения процесса оптимизации, если в качестве начальной точки выбрана «точка заклинивания» метода покоординатного спуска. Другим, более существенным с практической точки зрения, недостатком является то обстоятельство, что метод применим лишь к задачам оптимизации с одномерными оврагами. При наличии многомерных оврагов метод теряет эффективность, так как в нем не принимается специальных мер для погружения необходимого количества координатных осей (равного размерности дна оврага) в подпространство, образующее дно оврага. В результате целенаправленное изменение ориентации лишь одной из координатных осей не позволяет эффективно продвигаться по многомерному дну.

4.2. Методы обобщенного покоординатного спуска

Пусть решается задача $J(x) \to \min, x \in R^n, J \in C^2(R^n)$ с овражным в смысле определения, данного в 3.2.3, функционалом. Основной процедурой при реализации рассматриваемого далее класса методов обобщенного покоординатного спуска (ОПС) является процедура диагонализации матрицы $J''(x)$ с последующим циклическим покоординатным спуском вдоль собственных векторов. Целесообразность такого подхода вытекает из геометрически очевидного факта, заключающегося в том, что оси наиболее рациональной системы координат при минимизации квадратичных функционалов (независимо от их выпуклости) методом покоординатного спуска совпадают с собственными векторами матрицы вторых производных, являющихся осями симметрии соответствующих поверхностей уровня. Эта идея неоднократно высказывалась в литературе, и даже строились соответствующие алгоритмы. Однако численные эксперименты показали низкую эффективность такого подхода. Существует и объяснение этих неудовлетворительных результатов. Оно основано на том, что при определении собственных

векторов, соответствующих близким или кратным собственным значениям, возникают принципиальные вычислительные трудности. Аналогичные трудности, связанные с ограниченной точностью задания исходной информации, а также последующих вычислений, наблюдаются и при диагонализации плохо обусловленных матриц, имеющих относительно малые по модулю спектральные составляющие. Указанные обстоятельства, по-видимому, явились основным сдерживающим фактором, не позволившим внедрить изучаемые далее методы в вычислительную практику. Однако из дальнейшего изложения следует, что неудачи при численном экспериментировании были вызваны, по-видимому, особенностями реализации метода, рассчитанной на случай выпуклых функционалов. Как показано далее, для целей оптимизации, как в выпуклой ситуации, так и в невыпуклой, достаточно вычислить произвольный ортонормированный базис в инвариантном подпространстве, отвечающем каждой изолированной группе собственных значений. При этом собственные векторы могут быть вычислены со значительными погрешностями.

Отвечающие этим базисам линейные оболочки с высокой точностью совпадают с истинными подпространствами, определяемыми невозмущенной диагонализируемой матрицей.

Этот вывод в известной степени подтверждаются следующей теоремой [74].

Теорема 4.1. Пусть A — симметричная матрица $n \times n$; $\{u^i\}$ — ортонормированные собственные векторы; λ_i — собственные значения. Тогда при $\lambda_i \neq \lambda_j$ с точностью до величин второго порядка малости имеем

$$\langle u^i + du^i, u^j\rangle = (\lambda_i - \lambda_j)^{-1}\langle (dA)u^i, u^j\rangle, \tag{4.1}$$

где dA — возмущение матрицы A; du^i — соответствующее возмущение вектора u^i.

Доказательство. Отбрасывая величины второго порядка малости, из равенства $Au^i = \lambda_i u^i$ получим

$$(dA)u^i + Adu^i = \lambda_i du^i + d\lambda_i u^i.$$

Отсюда, умножая обе части равенства скалярно на u^j, имеем

$$\langle (dA)u^i, u^j\rangle + \langle Adu^i, u^j\rangle = \langle \lambda_i du^i, u^j\rangle + \langle d\lambda_i u^i, u^j\rangle.$$

В силу равенств $\langle u^i, u^j\rangle = 0$, $\langle Adu^i, u^j\rangle = \lambda_j\langle du^i, u^j\rangle$ имеем

$$(\lambda_i - \lambda_j)\langle du^i, u^j\rangle = \langle (dA)u^i, u^j\rangle,$$

откуда следует (4.1). Теорема доказана.

Пусть теперь

$$M_1 = \sum_{i=1}^{n-r} \alpha_i u^i,\ M_2 = \sum_{j=n-r+1}^{n} \alpha_j u^j$$

есть два линейных многообразия, порожденных непересекающимися системами собственных векторов $\{u^i,\ i \in [1:\ n - r]\}$, $\{u^j, j \in [n - r + 1:\ n]\}$ матрицы A. Если соответствующие множества собственных значений $\{\lambda_i,\ i \in [1:\ n - r]\}$, $\{\lambda_j,$

$j \in [n - r + 1: n]\}$ строго разделены в смысле $|\lambda_i| >> |\lambda_j|$, то из (4.1) следует $\langle u^i + du^i, u^j \rangle \cong 0$ при достаточно малой величине $\frac{\|dA\|}{|\lambda_i|}$. Это означает, что все собственные векторы под действием возмущения dA изменяются только в пределах своих линейных многообразий, сохраняя с высокой точностью свойство ортогональности к векторам из дополнительных многообразий. При этом сами вариации векторов при близких $\lambda_i \cong \lambda_j$ собственных значениях в пределах фиксированного линейного многообразия, как это следует из (4.1), могут быть весьма значительными.

Изложенное позволяет в качестве модели программ, реализующих различные методы диагонализации матрицы A, использовать оператор $\Lambda(A)$, ставящий в соответствие произвольной симметричной матрице A ортогональную матрицу V, отличную, вообще говоря, от истинной матрицы U, состоящей из собственных векторов матрицы A. Оператор Λ характеризуется тем, что если спектр матрицы A разделяется на p групп

$$\lambda_i^t(A),\ i \in [1: k_t],\ \sum_{t=1}^{p} k_t = n$$

«близких» между собой собственных чисел, то каждой группе соответствует набор столбцов $\{v^{it}\}$ матрицы V, задающий точное линейное многообразие, порожденное соответствующими столбцами $\{u^i\}$ точной матрицы U.

Рассмотрим квадратичную аппроксимацию

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c \tag{4.2}$$

исходного функционала $J(x)$ в окрестности точки $x \in Q$. Допустим, что известны матрица A и ортогональная матрица U, приводящая ее к диагональному виду $U^T A U = \operatorname{diag} \lambda_i$. Тогда замена переменных $x = Uy$ приводит квадратичный функционал к сепарабельному виду

$$f(x) = f(Uy) = \sum_{i=1}^{n} f_i(y_i), \tag{4.3}$$

где f_i — квадратичные функции одной переменной (параболы). Таким образом достигается *полная локальная декомпозиция* исходной задачи, и последняя сводится к n независимым экстремальным задачам. В результате поиск оптимального вектора y^* может осуществляться покомпонентно, так как связь между аргументами y_i фактически исчезает. В указанной идеализированной ситуации явление заклинивания метода покоординатного спуска невозможно, и все вычислительные проблемы при применении покоординатных стратегий поиска оптимума, связанные с большими значениями η, формально снимаются.

В действительности бывает задана не матрица A, а возмущенная матрица $A + dA$, где dA отражает как неопределенность задания исходной матрицы A, так и последующие ошибки округления при проведении собственно процесса диагонали-

зации. В связи с этим вместо точной матрицы U оказывается доступной некоторая матрица $V = \Lambda(A)$. Свойства оператора Λ были рассмотрены ранее. Замена переменных $x = Vy$ уже не приводит к представлению (4.3). Для изучения создавшейся ситуации важное значение имеет следующая теорема.

Теорема 4.2. Пусть собственные значения $\{\lambda_i\}$ и отвечающие им ортонормированные собственные векторы $\{u_i\}$, $i \in [1{:}\ n]$ некоторой симметричной матрицы A разделены произвольным образом на p групп λ_i^t, u^{it}, $i \in [1{:}\ k_t]$, $t \in [1{:}\ p]$, $\sum_{t=1}^{p} k_t = n$ так, что $u^{it'} \neq u^{jt''}$, $t' \neq t''$, $i \in [1{:}\ k_{t'}]$, $j \in [1{:}\ k_{t''}]$, где λ_j^t, u^{jt} означают j-е собственное число и соответствующий собственный вектор группы t. Тогда, если в каждом линейном многообразии M_t размерности k_t с базисом $\{u^{it}\}$, $i \in [1{:}\ k_t]$, задать иной ортонормированный базис $\{w^{it}\}$, $i \in [1{:}\ k_t]$, связанный с исходным базисом линейным соотношением

$$w^{it} = \sum_{m=1}^{k_t} \alpha_{mi}^t u^{mt},\ i \in [1{:}\ k_t],\ \alpha_{mi}^t \in R^1,$$

то существует такая матрица P перестановок столбцов, что:

1) преобразование подобия $\overline{W}^{\mathrm{T}} A \overline{W}$, $W = \{w^{11}, \ldots, w^{k_1 1}, \ldots, w^{k_p p}\}$, $\overline{W} = WP$ приводит матрицу A к блочно-диагональному виду $\overline{W}^{\mathrm{T}} A \overline{W} = \operatorname{diag}(A_1, A_2, \ldots, A_p)$, $\overline{W}^{\mathrm{T}} = \overline{W}^{-1}$ с квадратными $(k_t \times k_t)$ матрицами A_t на главной диагонали;
2) собственные значения матрицы A_t есть λ_i^t, $i \in [1{:}\ k_t]$, $t \in [1{:}\ p]$.

Доказательство. Первое утверждение проверяется непосредственно с учетом ортонормированности векторов базиса $\{u^i\}$. Для доказательства второго утверждения достаточно заметить, что вид и расположение матрицы A_m при фиксированном многообразии M_m не зависят от способа задания остальных многообразий M_t, $t \neq m$. Поэтому, предположив, что все $k_t = 1$ при $t \neq m$, получим, что все оставшиеся многообразия будут содержать по одному собственному вектору и собственному числу. Тогда будем иметь:

$$\overline{W}^{\mathrm{T}} A \overline{W} = \operatorname{diag}(\lambda_1^1, \lambda_1^2, \ldots, A_m, \ldots, \lambda_1^p).$$

Учитывая, что преобразование подобия не изменяет спектр матрицы, приходим к требуемому заключению. Теорема доказана.

Теорема 4.3. Пусть $V = \Lambda(A)$, тогда

1) замена переменных $x = Vy$ с точностью до нумерации компонентов вектора y приводит функционал $f(x)$ (4.2) к блочно-сепарабельному виду

$$f(x) = f(Vy) = f_s(y) = \sum_{t=1}^{p} f_t(y^t), \tag{4.4}$$

где $y = (y_1, \ldots, y_n) = (y^1, \ldots, y^p)$, $y^t = (y_1^t, \ldots, y_{k_t}^t)$,

$$f_t(y^t) \triangleq \frac{1}{2}\langle A_t y^t, y^t \rangle - \langle b^t, y^t \rangle + c_t,\ c_t \in R^1;$$

2) собственные значения матрицы f_t'' равны $\lambda_i^t(A)$, $i \in [1{:}k_t]$, $t \in [1{:}p]$.

Доказательство. Имеем $V = \overline{W}P$, где P — некоторая матрица перестановок столбцов. Поэтому

$$f(x) = \frac{1}{2}\langle V^{\mathrm{T}}AVy, y\rangle - \langle V^{\mathrm{T}}b, y\rangle + c = \frac{1}{2}\langle P^{\mathrm{T}}\overline{W}^{\mathrm{T}}A\overline{W}Py, y\rangle -$$

$$-\langle P^{\mathrm{T}}\overline{W}^{\mathrm{T}}b, y\rangle + c = \frac{1}{2}\langle \overline{W}^{\mathrm{T}}A\overline{W}z, z\rangle - \langle \overline{W}^{\mathrm{T}}b, z\rangle + c,\ z \triangleq Py.$$

Согласно теореме 4.2, матрица $\overline{W}^{\mathrm{T}}A\overline{W}$ имеет блочно-диагональную структуру, что и доказывает первое утверждение. Второе утверждение есть прямое следствие второго утверждения теоремы 4.2.

Следствие. Пусть собственные числа матрицы A удовлетворяют неравенствам $\lambda_1 \geq \ldots \geq \lambda_{n-r} >> |\lambda_{n-r+1}| \geq \ldots \geq |\lambda_n|$, тогда

1) замена переменных $x = Vy$, $V = \Lambda(A)$, где $V = (v^{11}, \ldots, v^{1n-r}, v^{21}, \ldots, v^{2r})^{\mathrm{T}}$;

$$v^{1i} = \sum_{m=1}^{n-r} \alpha_{mi}^1 u^m;\quad v^{2i} = \sum_{m=1}^{r} \alpha_{mi}^2 u^{n-r+m}$$

с точностью до нумерации компонентов вектора y приводит $f(x)$ к виду

$$f_s(y) = f_1(y^1) + f_2(y^2),\ y = (y^1, y^2), \tag{4.5}$$

где $y^1 = (y_1, \ldots, y_{n-r})$, $y^2 = (y_{n-r+1}, \ldots, y_n)$;

2) $\eta_1 << \eta$, $\eta_2 << \eta$, где η_i — показатели овражности задач $f_i \to \min$, $i = 1, 2$.

Таким образом, исходная оптимизационная задача локально может быть сведена к двум эквивалентным задачам с существенно меньшими числами η_i. Представление (4.5) реализует принцип *частичной локальной декомпозиции* и является аналогом идеализированного соотношения (4.3).

Если собственные числа матрицы квадратичного функционала разделяются более чем на две группы, то будет справедливо представление (4.5), содержащее соответствующее число слагаемых.

Согласно (4.5), появляется возможность независимого решения не связанных между собой оптимизационных задач для функционалов f_i с невысокими показателями овражности.

Полученные результаты носят локальный характер и справедливы в рамках квадратичной аппроксимации исходного функционала $J(x)$. Для неквадратичных функционалов приближенное выполнение соотношений типа (4.5) позволяет говорить о существенном ослаблении связей между различными группами переменных, что определяет достаточно высокую эффективность покоординатного спуска и в общем случае.

Исследование сходимости методов ОПС в предположении точной линейной оптимизации вдоль направляющих ортов может быть основано на общем подходе к исследованию алгоритмов нелинейного программирования [5].

Пусть решается задача $J(x) \to \min$, $x \in R^n$, $J(x) \in C^1(R^n)$. Рассмотрим произвольный алгоритм A, строящий последовательность точек $\{x^k\}$, причем каждая точка x^{k+1} получается из x^k последовательной минимизацией функционала $J(x)$ вдоль направлений $d^1, \ldots, d^n$, начиная из точки x^k. Предполагается, что матрица $D = (d^1, \ldots, d^n)$ может зависеть от номера k, являясь при любом k ортогональной. Легко видеть, что метод ЦПС, метод Розенброка, а также методы ОПС описываются приведенной общей схемой. Далее будет показано, что все эти методы при некоторых естественных ограничениях сходятся, так как сходится базовый алгоритм A.

Лемма 4.1. Пусть $J(x) \in C^1(R^n)$ и пусть задана последовательность точек $\{x^k, k = 1, 2, \ldots\}$ из R^n такая, что:

1) $\forall k\ J(x^{k+1}) \le J(x^k)$;

2) $J(x^k) \to J$, $k \in K$,

где K — некоторое бесконечное подмножество индексов.

Тогда

$$\lim_{k\to\infty} J(x^k) = \lim_{k\in K} J(x^k) = J.$$

Доказательство содержится в книге [28]. В лемме утверждается, что если последовательность значений функционала $\{J_k\}$ является монотонно невозрастающей и ее подпоследовательность $\{J_k, k \in K\}$ сходится к некоторому значению J, то вся последовательность $\{J_k\}$ сходится к этому же значению.

Теорема 4.4. Пусть

1) $J(x) \in C^1(R^n)$;

2) множество решений, определяемое как

$$X^* = \{x \in R^n \mid J'(x) = 0\},$$

непусто;

3) множество $\Theta = \{x \in R^n \mid J(x) \le J(x^0)\}$, где x^0 — заданная начальная точка, ограничено и замкнуто (компактно) в R^n;

4) минимум функционала J вдоль любой прямой в R^n единственен;

5) если $J'(x^k) = 0$, то алгоритм останавливается в x^k.

Тогда каждая предельная точка последовательности $\{x^k\}$, построенной алгоритмом A, принадлежит множеству X^*.

Доказательство. Если последовательность $\{x^k\}$ конечна, то это значит, что алгоритм остановился в точке $x^k \in X^*$, и утверждение теоремы очевидно. Допустим поэтому, что алгоритм порождает бесконечную последовательность точек $\{x^k\}$. Согласно третьему предположению, все точки расположены в компактном множестве Θ и, следовательно, должна существовать подпоследовательность $\{x^k, k \in K\}$, сходящаяся при $k \to \infty$ к некоторой точке x. Покажем, что $x \in X^*$, то есть $J'(x) = 0$.

Рассмотрим подпоследовательность $\{x^{k+1}, k \in K\}$. Она также содержится в Θ, и поэтому существует $K' \subset K$ такое, что $x^{k+1} \to x'$ при $k \in K'$, где x' — некоторая точка

из Θ. Из сходимости последовательностей $\{x^k, k \in K\}$, $\{x^{k+1}, k \in K'\}$ следует в силу непрерывности J сходимость последовательностей $\{J(x^k), k \in K\}$, $\{J(x^{k+1}), k \in K'\}$, соответственно, к значениям $J(x), J(x')$. В силу леммы 4.1, вся последовательность $\{J(x^k)\}$ также оказывается сходящейся к некоторой величине $\tilde{J}$, и при этом $J(x) = J(x') = \tilde{J}$.

Далее будет показано, что точка x' получается из x последовательной минимизацией вдоль некоторых ортогональных направлений $q^1, \ldots, q^n$. Поэтому в силу четвертого предположения из равенства $J(x) = J(x')$ следует $x = x'$. Это значит, что убывания J нет ни в одном из направлений q^i, то есть проекции вектора градиента $J'(x)$ на q^i, $i \in [1: n]$ равны нулю. Отсюда следует, что $J'(x) = 0$.

Для завершения доказательства покажем, что существуют такие ортогональные направления $q^1, \ldots, q^n$, которые позволяют получить точку x' из x с помощью последовательной минимизации по этим направлениям.

Пусть D_k — матрица порядка $n \times n$, столбцами которой являются построенные на k-й итерации направления спуска $\{d^{ik}\}$, $i \in [1: n]$; то есть $x^{k+1} = x^k + D_k\lambda^k$, где $\lambda^k = (\lambda_{k1}, \ldots, \lambda_{kn})^T$ — вектор значений шагов вдоль направлений $d^{1k}, \ldots, d^{nk}$ соответственно.

Обозначая $y^{1k} = x^k$, $y^{j+1,k} = y^{jk} + \lambda_{kj}d^{jk}$, $j \in [1: n]$, получим, что $x^{k+1} = y^{n+1,k}$ и

$$J(y^{j+1,k}) \le J(y^{jk} + \tilde{\lambda}d^{jk}), \tilde{\lambda} \in R^1, j \in [1: n]. \quad (4.6)$$

Рассмотрим последовательность матриц $\{D_k\}$, $k \in K'$. Из этой последовательности можно выделить подпоследовательность, для которой первый столбец d^{1k} будет сходиться к некоторому вектору q^1. Это следует из того, что множество векторов, норма которых равна единице, компактно. Из этой подпоследовательности можно снова извлечь подпоследовательность, для которой не только $d^{1k} \to q^1$, но и $d^{2k} \to q^2$ и т. д. В итоге существует такое $K'' \subset K'$, что $D_k \to Q$. Причем $\det Q = 1 \ne 0$, так как для всех k $\det D_k = 1$. Имеем, таким образом,

$$x^k \to x, x^{k+1} \to x', D_k \to Q, k \in K''. \quad (4.7)$$

Поэтому из соотношения $x^{k+1} = x^k + D_k\lambda^k$ имеем

$$\lambda^k \to \lambda = Q^{-1}(x' - x). \quad (4.8)$$

Следовательно, $x' = x + Q\lambda$. Пусть $y_1 = x$ и $y^{j+1} = y^j + \lambda_j d^j$. Чтобы доказать, что вектор x' может быть получен из x последовательной минимизацией J по направлениям q^i, $i \in [1: n]$, достаточно показать, что

$$J(y^{j+1}) \le J(y^j + \tilde{\lambda}q^j), \tilde{\lambda} \in R^1, j \in [1: n]. \quad (4.9)$$

Из (4.7), (4.8) следует, что $y^{jk} \to y^j, j \in [1: n+1]$ при $k \in K''$. Поэтому в силу непрерывности J неравенство (4.9) следует из (4.6). Теорема доказана полностью.

Сделаем несколько замечаний общего характера.

Вопросам выяснения условий сходимости алгоритмов оптимизации уделяется большое внимание. При этом исследуется сходимость последовательности $\{x^k\}$, генерируемой алгоритмом, к некоторому фиксированному множеству X^*, которое называется множеством решений. Наиболее естественный подход

к введению понятия множества решений заключается в определении: $X^* = \{x^* \in D \subset R^n \mid J(x^*) \le J(x), x \in D\}$, где D — множество допустимых точек из R^n. Иначе говоря, x^* — решение, если это точка глобального минимума. Однако в действительности такой подход не всегда возможен, и мы вынуждены останавливать процесс при выполнении более слабых предположений относительно полученных алгоритмом точек. Чаще всего для описания множества X^* применяются необходимые условия экстремума, и тогда полагают: $X^* = \{x^* \in D \subset R^n \mid J'(x^*) = 0\}$. Возможны и другие способы задания X^*, которые столь же легко позволяют проверить принадлежность точки x^k множеству решений. Типичное для приложений множество решений может быть задано следующим образом:

$$X^* = \{x^* \in D \subset R^n \mid J(x^*) \le \tilde{J}\},$$

где $\tilde{J}$ — некоторое приемлемое значение минимизируемого функционала. Очевидно, что сходящийся для фиксированного X^* алгоритм может оказаться несходящимся, если определить множество решений другим способом.

Сходимость алгоритма при заданном X^* является чрезвычайно желательным свойством. Однако полезность этого свойства часто переоценивается. Дело в том, что исследуя бесконечные последовательности генерируемых алгоритмом точек, мы фактически обращаемся к некоторым математическим фикциям. В действительности всегда имеют дело с конечной последовательностью, и, как указано в [48], свойство сходимости алгоритма в общем случае не является ни необходимым, ни достаточным для окончательной оценки алгоритма.

Другим фактором, несколько снижающим ценность понятия сходимости, являются различные погрешности, которые всегда присутствуют в реальных вычислениях и почти никогда не фигурируют в доказательствах теорем о сходимости. Вместе с тем известны многочисленные примеры, когда влияние погрешностей оказывается решающим.

4.3. Реализация методов обобщенного покоординатного спуска

При реализации методов ОПС в первую очередь необходимо учитывать три фактора, связанные с нормализацией основных переменных задачи, построением аппроксимации матрицы $G(x)$, а также с выбором метода диагонализации матрицы $G(x)$.

4.3.1. Нормализация основных переменных задачи

Цель нормализации состоит в улучшении обусловленности задачи, а также в согласовании масштабов основных переменных, часто имеющих различный физический смысл и измеряемых в различных единицах.

Масштабирование управляемых параметров

Масштабирование может выполняться на двух уровнях: на уровне пользователя и на уровне стандартных программных модулей, реализующих конкретный алгоритм параметрической оптимизации. Как показывает опыт решения реальных задач, эффективность процедуры масштабирования существенно зависит от конкретной структуры решаемой задачи. Поэтому, как правило, целесообразно выполнять масштабирование на уровне пользователя, несмотря на известное усложнение процесса подготовки задачи к ее компьютерной реализации. Наиболее часто переход от исходных переменных $x = (x_1, \ldots, x_n)$ к новым $y = (y_1, \ldots, y_n)$ осуществляется с помощью замены вида

$$x = Dy,\ D = \mathrm{diag}\,(d_i). \tag{4.10}$$

Диагональная матрица масштабов D хранится в соответствующей «общей» области подпрограммы, вычисляющей значения $J(x)$, и модифицируется в начале каждого нового цикла оптимизации, исходя из равенства

$$d_i = \min\{\delta_2,\ \max\{|\tilde{x}_i|, \delta_1\}\},\ \delta_i > 0, \tag{4.11}$$

где $\tilde{x}_i$ — лучшее значение i-й переменной, полученное к началу следующего цикла; $\delta_1 \cong \varepsilon_м$ — заданное пороговое значение, исключающее появление нулевых d_i; $\delta_2 \cong \varepsilon_M^{-1}$. В методах второго порядка, рассматриваемых в данной книге, считается, что новый цикл начинается с процедуры вычисления аппроксимации матрицы вторых производных $G(x)$. Начальное масштабирование проводится на основе заданных начальных значений x_i^0. В ряде случаев лучшие результаты дает комбинированный метод, использующий дополнительное масштабирование на основе *принципа «равного влияния»*, согласно которому диагональная матрица масштабов D выбирается из условия балансирования производных

$$d_i = k\gamma_i,\ \gamma_i = \left(\left|\frac{\partial J}{\partial x_i}\right| + \varepsilon_м\right)^{-1} + \varepsilon_м, \tag{4.12}$$

где, например, $k = 1$. В этом случае (при отсутствии влияния $\varepsilon_м$) будем иметь

$$\frac{\partial J(Dy)}{\partial y} = \frac{D\partial J}{\partial x} = (1, \ldots, 1)^T.$$

При реализации масштабов (4.12) в начале каждого цикла оптимизации необходимо проводить анализ чувствительности для получения грубой аппроксимации вектора $\partial J/\partial x$ в окрестности точки $\tilde{x}$.

Указанных двух методов нормализации управляемых параметров, обычно применяемых совместно, бывает достаточно для решения большинства практических задач рассматриваемыми в книге методами параметрической оптимизации.

Нормализация значений минимизируемого функционала.

Теоретически масштаб значений целевого функционала $J(x)$ не оказывает влияния на процесс поиска. Однако при численной реализации алгоритма выбор масштаба $J(x)$ является существенным фактором. Как показывает практика, целесо-

образно поддерживать значения $J(x)$ на уровне $J(x) \cong 1$ с целью уменьшения вычислительных погрешностей и предотвращения влияния концов диапазона чисел, представимых в памяти компьютера. Для этого в подпрограмме пользователя значения $J(x)$ умножаются на постоянный для текущего цикла оптимизации множитель. Как правило, необходимо вначале нормализовать управляемые параметры согласно (4.11), затем промасштабировать значения $J(x)$, а далее перейти к уравниванию производных (4.12).

Специальные приемы нормализации

В ряде случаев на уровне аналитического исследования проблемы оптимизации и подготовки ее для компьютерного моделирования удается существенно улучшить структуру задачи с точки зрения ее обусловленности и эффективности последующей процедуры минимизации. Рассмотрим два характерных случая.

Первый случай связан с нормализацией некоторой независимой переменной t в задачах аппроксимации по среднестепенным критериям (см. 2.8) и, в частности, по методу наименьших квадратов (МНК). Как правило, переменная t, содержательно трактуемая, например, как частота или время, изменяется в широком диапазоне $[t', t'']$, образуя комбинации с управляемыми параметрами вида $x_i t_i^{s_i}$, $t \in [t', t'']$, где s_i может быть достаточно большим. В результате величина $t_i^{s_i}$ оказывается большой, и ее прямое вычисление может привести к переполнению разрядной сетки машины. Соответственно, оптимальное значение x_i оказывается сравнимым с нулем (по сути, имеем численный вариант неопределенности типа $0 \cdot \infty$). При таких вычислениях, даже при отсутствии аварийных остановов компьютера, катастрофически теряется точность. Проблема снимается, если нормализовать переменную t (а тем самым и все x_i) с помощью соотношения $t = T\tau$, $T >> 1$, где τ — новая независимая переменная, $\tau \in [\tau', \tau'']$, $\tau' = \frac{t'}{T}$, $\tau'' = \frac{t''}{T}$. Новый управляемый параметр при этом равен $y_i = T^{s_i} x_i$ и имеет увеличенное в T^{s_i} раз значение.

Второй случай относится к устранению аддитивных фоновых добавок к целевому функционалу $J(x)$. Если возможно представление $J(x) = I(x) + c$, $c = \text{const}$, то при достаточно больших абсолютных значениях c необходимо решать задачу минимизации $I(x)$ вместо исходной задачи минимизации $J(x)$. В противном случае могут возникать серьезные вычислительные трудности при построении конечно-разностных аппроксимаций первых и вторых производных $J(x)$. Потеря точности происходит при условии $\Delta J = \Delta I = I(x + \Delta x) - I(x) << c$. Существенно, что общая нормализация значений $J(x)$, согласно п. 2, не изменяет ситуацию к лучшему.

Нормализация ограничений

Рассмотрим общий вид ограничений — неравенств

$$g(x) \le 0, g(x) = (g_1(x), g_2(x), \ldots, g_s(x)).$$

Основная цель нормализации ограничений состоит в достижении их сбалансированности. Предполагается, что нормализация управляемых параметров x про-

изведена. Переход к нормализованному варианту ограничений $\alpha^T g(x) \le 0$, $\alpha_i > 0$ ($i \in [1: s]$) может быть выполнен согласно условию аналогичному (4.12):

$$\alpha_i = \left[\left\| \frac{\partial g_i}{\partial x} \right\| + \varepsilon_м \right]^{-1} + \varepsilon_м .$$

Оценка соответствующих градиентов может производиться на основе грубых односторонних приращений независимых переменных x_i.

Более тонкие проблемы, связанные с масштабированием ограничений, рассмотрены в [19].

4.3.2. Методы диагонализации

В качестве основной процедуры приведения симметричной матрицы к главным осям в методах ОПС может быть использован метод Якоби [63], несмотря на наличие конкурирующих, вообще говоря, более эффективных вычислительных схем [45]. Такой выбор обусловлен следующими обстоятельствами. Во-первых, вычисленные методом Якоби собственные векторы всегда строго ортонормальны с точностью, определяемой точностью компьютера, даже при кратных собственных числах. Последнее весьма существенно при использовании этих векторов в качестве базиса, так как предотвращается возможность вырождения базиса, существующая, например, в методе Пауэлла [71]. Во-вторых, многие вычислительные схемы имеют преимущество перед методом Якоби лишь при решении частичной проблемы собственных значений. В нашем же случае всегда решается полная проблема, и поэтому выигрыш во времени оказывается несущественным при существенно более сложных вычислительных схемах. В-третьих, алгоритмы, основанные на методах Якоби, часто оказываются наиболее доступными, так как соответствующие программы имеются в большинстве вычислительных систем. И наконец, определенное влияние на выбор алгоритма оказала простота логики метода Якоби, что приводит к компактности реализующих его программ.

В задачах большой размерности по сравнению с методом Якоби более предпочтительным по объему вычислительных затрат оказывается метод, использующий преобразование Хаусхолдера для приведения матрицы к трехдиагональной форме с последующим обращением к QR-алгоритму определения собственных векторов симметричной трехдиагональной матрицы [45].

В методе Якоби исходная симметричная матрица A приводится к диагональному виду с помощью цепочки ортогональных преобразований вида

$$A_{k+1} = U_k^T A_k U_k, \; A_0 = A, \; k = 1, 2, \ldots, \qquad (4.13)$$

являющихся преобразованиями вращения. В результате надлежащего выбора последовательности $\{U_k\}$ получаем $\lim A_k = D = U^T A U$, $k \to \infty$, где $D = \operatorname{diag}(\lambda_i)$ — диагональная матрица; $U = U_0 U_1 U_2 \ldots$ — ортогональная матрица. Так как (4.13) есть преобразования подобия, то на диагонали матрицы D расположены собственные числа матрицы A; столбцы матрицы U есть собственные векторы матрицы A. В действительности, как указано в 4.2, вместо матрицы U получается некоторая матрица $V = \Lambda(A)$, отличная, вообще говоря, от истинной матрицы U. Основные характеристики оператора Λ были рассмотрены ранее.

Элементарный шаг (4.13) процесса Якоби заключается в преобразовании посредством матрицы $U_k = \{u_{ij}\}$, отличающейся от единичной элементами $u_{pp} = u_{qq} = \cos\varphi$, $u_{pq} = -u_{qp} = \sin\varphi$. Угол вращения φ выбирается таким образом, чтобы сделать элемент a_{pq} матрицы A нулем. Вопросы сходимости различных численных схем, реализующих метод Якоби, рассмотрены в книге [65].

За основу может быть взят алгоритм *jacobi* из сборника алгоритмов линейной алгебры [64], реализующий так называемый *частный циклический метод Якоби*. В этом методе аннулированию подвергаются все элементы верхней треугольной части матрицы A с применением построчного выбора. При таком выборе индексы элементов a_{pq} пробегают последовательность значений (1, 2), (1, 3), ..., (1, n); (2, 3), (2, 4), ..., (2, n); ...; (n – 1, n). Затем начинается новый цикл перебора элементов в том же порядке. Эмпирическая оценка трудоемкости процесса построения матрицы $\Lambda(A)$ позволяет выразить необходимое время T работы процессора через размерность n решаемой задачи. Известно, что для матриц до 50-го порядка и длин машинных слов от 32 до 48 двоичных разрядов общее число циклов в процессе вращений Якоби в среднем не превышает 6–10 (под циклом понимается любая последовательность из $\frac{(n^2-n)}{2}$ вращений). При этом $T \cong kn^3$, где коэффициент k определяется быстродействием применяемой вычислительной системы и приблизительно равен $40t_y$, где t_y — время выполнения операции умножения.

Полученная оценка, а также опыт практической работы показывают, что при умеренных n время реализации оператора Λ для многих практических случаев невелико и сравнимо с временем однократного вычисления значения минимизируемого функционала. Упоминавшаяся ранее комбинация метода Хаусхолдера и QR-алгоритма оказывается приблизительно в 1,5–2 раза быстрее, что может иметь значение при достаточно больших n.

Все рассматриваемые в книге методы и алгоритмы ОПС основаны на общих результатах материала, изложенного в 4.3. Различные версии алгоритмов, учитывающие специфику решаемых классов прикладных задач, отличаются, в основном, методами построения аппроксимаций матриц $G(x)$. Соответствующие вопросы излагаются в следующем разделе.

4.4. Алгоритмы обобщенного покоординатного спуска

4.4.1. Реализации на основе конечно-разностных аппроксимаций производных

Достоинством подхода, основанного на применении формул численного дифференцирования, кроме его универсальности, является низкая стоимость подготовки задачи к компьютерному моделированию. От пользователя требуется лишь

написание программы для вычисления значения $J(x)$ при заданном x. Реализованные на основе численных производных методы оптимизации оказываются, по существу, *прямыми* методами, то есть методами, не использующими в своей схеме производные от J. Действительно, заменяя $\frac{\partial I}{\partial x_i}$, например, конечно-разностным отношением $\frac{[J(x+se^i)-J(x)]}{s}$, где $e^i=(0,\ldots,1,\ldots,0)$, $s\in R^1$, мы фактически используем лишь значения J, вычисленные при определенных значениях аргумента.

Все рассматриваемые в данной книге методы оптимизации строятся на основе использования локальной квадратичной модели минимизируемого функционала, получаемой из общего разложения в ряд Тейлора. Естественно поэтому при выборе формул численного дифференцирования также руководствоваться идеей локальной квадратичной аппроксимации. Исходя из этого, целесообразно вместо формул с односторонними приращениями применять двусторонние конечно-разностные аппроксимации производных, оказывающиеся точными для квадратичных функционалов. Действительно, легко проверить, что если

$$f(x)=\frac{1}{2}\langle Ax,x\rangle-\langle b,x\rangle,$$

то равенство

$$\frac{\partial f}{\partial x_i}=\frac{f(x+se^i)-f(x-se^i)}{2s} \tag{4.14}$$

оказывается точным при любом $s\neq 0$.

Упражнение. Докажите последнее утверждение.

Точно так же точным оказывается следующее представление для вторых производных квадратичного функционала

$$\frac{\partial^2 f}{\partial x_i\partial x_j}=$$
$$=\frac{f(x+se^i+se^j)-f(x-se^i+se^j)-f(x+se^i-se^j)+f(x-se^i-se^j)}{4s^2}. \tag{4.15}$$

При использовании соотношений (4.14), (4.15) вычислительные затраты характеризуются числом обращений к вычислению значений $J(x)$: для вычисления градиента — $2n$, для вычисления матрицы Гессе — $2n^2+1$, где n — размерность вектора x.

Для излагаемых далее методов оптимизации достаточно определять $J'(x)$ и $J''(x)$ с точностью до множителя, поэтому при реализации формул (4.14), (4.15) деление, соответственно, на $2s$ и $4s^2$ не производится. Это позволяет избежать известных вычислительных трудностей, если значение s оказывается относительно малым.

В данном случае мы отказываемся от работы с различными шагами s_i по отдельным компонентам x_i вектора x, предполагая, что последние надлежащим образом нормализованы.

В обсуждаемых реализациях метода ОПС шаг дискретности s определяется автоматически в зависимости от величины нормы вектора продвижения $\|x^i - x^{i-1}\|$ в пространстве переменных x за один цикл работы алгоритма. Чем большая величина s используется, тем шире предполагаемая область справедливости локальной квадратичной модели исходного функционала. Наибольшая точность вычислений по формуле (4.15) применительно к квадратичным зависимостям достигается при работе с максимально возможным s, так как в этом случае вклад погрешностей задания значений J в окончательный результат становится наименьшим. Поэтому чем дальше удалось продвинуться на основе построенной квадратичной аппроксимации функционала, то есть чем шире зона ее «полезного действия», тем большие значения s целесообразно выбирать для вычисления производных на следующем этапе поиска. Возможны и другие стратегии регулировки параметра s. Например, сама подпрограмма, реализующая метод оптимизации, может быть настроена на работу с постоянным шагом дискретности. Изменения s в этом случае осуществляются во внешней программе в зависимости от получаемых результатов.

Полученные в результате диагонализации матрицы Гессе новые координатные орты используются далее для реализации базового алгоритма ЦПС с процедурой выбора шагов продвижения по осям, применяемой в алгоритме GZ1. Переход к новым осям координат целесообразно осуществлять после того, как текущие оси «исчерпали себя» и дальнейшего существенного улучшения ситуации не ожидается. В изучаемых алгоритмах обновление осей координат происходит после того, как по каждому из координатных направлений вслед за успешным продвижением последовала неудача в смысле возрастания значения $J(x)$.

Укрупненное описание базового алгоритма, реализующего метод ОПС, сводится к следующей последовательности шагов.

Алгоритм SPAC1.

Шаг 1. Ввести исходные данные: x, s.

Шаг 2. Вычислить матрицу $B = \{b_{ij}\}$ по формулам

$$b_{ij} = J(x + se^i + se^j) - J(x - se^i + se^j) - J(x + se^i - se^j) + J(x - se^i - se^j),$$

$$i, j \in [1: n]; \; e^i = (0, \ldots, 1, \ldots, 0). \qquad (4.16)$$

Шаг 3. С помощью процедуры *jacobi* построить ортогональную матрицу U, приводящую матрицу B к диагональному виду $U^{\mathrm{T}}BU$.

Шаг 4. В осях $\{u^i\}$, совпадающих со столбцами матрицы U, реализовать процесс ЦПС (алгоритм GZ1) из точки x до выполнения условия поворота осей; присвоить x полученное лучшее значение, модифицировать s и перейти к шагу 2.

Процесс заканчивается по исчерпании заданного числа обращений к процедуре вычисления $J(x)$. Так же как и в алгоритме GZ1, предусмотрена возможность повторных входов в алгоритм и продолжения вычислений с прерванного места. Приводимые далее результаты работы SPAC1 соответствуют автоматическому выбору шагов дискретности для численного дифференцирования, исходя из равенства

$$s_{i+1} = 0{,}1\| x^{i+1} - x^i \|.$$

Построенный алгоритм имеет весьма простую структуру, однако его эффективность может быть достаточно высокой, несмотря на необходимость построения матрицы B. В ряде случаев более эффективной оказывается модификация метода ОПС, реализованная в алгоритме SPAC2.

В алгоритме SPAC2 вычисление матрицы вторых производных осуществляется в текущих осях $\{u^i\}$ без возврата к исходному единичному базису $\{e^i\}$. В результате информация о последнем используемом базисе не теряется, что позволяет сократить трудоемкость решения задачи.

Алгоритм SPAC2 в упрощенном варианте состоит из следующих шагов.

Алгоритм SPAC2.

Шаг 1. Ввести исходные данные: x, s.

Шаг 2. Принять $U = E$, где E — единичная матрица; в качестве координатных векторов взять столбцы $\{u^i\}$ матрицы U.

Шаг 3. Построить матрицу $B = \{b_{ij}\}$ по формулам

$$b_{ij} = J(x + su^i + su^j) - J(x - su^i + su^j) - J(x + su^i - su^j) + J(x - su^i - su^j),$$
$$i, j \in [1: n].$$

Шаг 4. Принять $U := UT$, где T — ортогональная матрица, приводящая матрицу B к диагональному виду $T^{\mathrm{T}}BT$.

Шаг 5. В осях $\{u^i\}$ реализовать процесс ЦПС из точки x до выполнения условия поворота осей; присвоить x полученное лучшее значение. Модифицировать s и перейти к шагу 3.

Окончание процесса и выбор шагов дискретности такие же, как и в алгоритме SPAC1.

Дадим необходимые пояснения к алгоритму, касающиеся построения матрицы U на шаге 4.

Выбор в качестве координатных направлений столбцов $\{u^i\}$ некоторой ортогональной матрицы U, очевидно, эквивалентен замене переменных $x = Uy$. В этом случае изменения компонент y_i вектора y приводят в исходном пространстве к смещениям вдоль одноименных векторов u^i.

Функционал $J(Uy) = I(y)$, как функция от y, имеет матрицу Гессе вида $I''(y) = U^{\mathrm{T}}J''U$. Таким образом, если необходимо работать с функционалом, имеющим гессову матрицу вида $U^{\mathrm{T}}J''U$, то для этого достаточно в качестве базиса взять столбцы матрицы U. Если требуется изменить матрицу Гессе и привести ее к виду $T^{\mathrm{T}}(U^{\mathrm{T}}J''U)T = U_1^{\mathrm{T}}J''U_1$, где T — новая ортогональная матрица, $U_1 = UT$, то в качестве базисных векторов достаточно выбрать столбцы матрицы $U_1 = UT$. Указанная процедура и реализована на шаге 4 сформулированного алгоритма.

Остановимся на некоторых принципиальных отличиях алгоритма SPAC2 от SPAC1. Во-первых, если минимизируемый функционал близок к квадратичному и матрица Гессе меняется относительно мало, то повторное ее построение в осях $\{u^i\}$ опять приведет практически к диагональной матрице, и поэтому число якобиевых циклов вращений при последующей диагонализации вновь полученной матрицы J'' будет сведено к минимуму. В результате матричная поправка T

к матрице U, вычисляемая на шаге 4, оказывается близкой к единичной матрице. В алгоритме SPAC1 в указанных условиях весь процесс диагонализации должен каждый раз целиком повторяться; то, что оси $\{u^i\}$ фактически меняются мало при переходе к следующей точке x^i, никак не используется.

Во-вторых, если на каком-либо этапе поиска минимума шаг дискретности s оказывается меньше, чем, скажем, $\min_i |\varepsilon_м x_i|$, то это приведет к получению нулевой матрицы на этапе построения матрицы B. Подпрограмма, реализующая алгоритм *jacobi*, выполнена таким образом, что при этом на выходе в качестве диагонализирующей матрицы T выдается единичная матрица. Если указанная ситуация возникает в процессе работы SPAC2, то на шаге 4 не произойдет изменения матрицы U и процесс ЦПС будет продолжен в текущих осях координат. Таким образом, будет сохраняться возможность медленного продвижения, скорость которого затем может вновь возрасти. В случае же применения SPAC1 нулевая матрица B автоматически приведет к получению матрицы $U = E$, что эквивалентно возврату к исходному единичному базису. В результате вероятность заклинивания на участках медленного продвижения для SPAC1 оказывается существенно большей, чем для SPAC2.

Отмеченные особенности SPAC2 не позволяют, однако, полностью отказаться от применения SPAC1. Это объясняется, во-первых, тем, что трудоемкость вычисления матрицы B в SPAC2 оказывается заметно выше, чем в SPAC1, так как вектор x варьируется в направлениях u^i, отличных, вообще говоря, от единичных векторов e^i. Кроме этого важное отличие заключается в том, что алгоритм SPAC2 предполагает практически единственно возможный способ построения матриц Гессе, основанный на формулах (4.15). При применении же базового алгоритма SPAC1 пригоден любой из описанных далее методов. Последнее обстоятельство обычно оказывается решающим при выборе метода.

4.4.2. Реализации на основе рекуррентных алгоритмов оценивания

В процессе работы алгоритма ОПС получается последовательность векторов $\{x^k\}$ и отвечающая им последовательность значений минимизируемого функционала $\{J_k\}$. В указанных алгоритмах используются только те точки x^k, которые приводят к монотонному убыванию $J(x)$, а «неудачные» точки отбрасываются и далее никак не участвуют в процессе поиска. Далее показано, что соответствующая информация может быть использована для построения квадратичной модели функционала $J(x)$ с целью последующего вычисления собственных векторов матрицы $J''(x)$ в качестве новых направлений поиска.

Рассмотрим последовательности $\{x^k\}$, $\{J_k\}$, генерируемые методом ОПС в текущей системе координат. В данном случае имеются в виду «полные» последовательности, включающие как удачные, так и неудачные шаги. Опишем процедуру, позволяющую по этой информации вычислить матрицу Гессе аппроксимирующего квадратичного функционала. Задачу квадратичной аппроксимации будем решать на основе метода наименьших квадратов

$$F_N(c) \triangleq \sum_{i=1}^{N} \left[J(x^i) - f(x^i, c) \right]^2 \to \min_{c \in R^m}; \tag{4.17}$$

$$x^i = (x_1^i, x_2^i, \ldots, x_n^i),\ c = (c_1, c_2, \ldots, c_m),\ m = \frac{n^2}{2} + \frac{3n}{2} + 1,\ N \geq m.$$

Здесь $f(x, c)$ означает аппроксимирующий функционал с неизвестными коэффициентами c_i:

$$\begin{aligned} f(x,c) = {} & c_1 x_1^2 + c_2 x_1 x_2 + c_3 x_1 x_3 + \ldots + c_n x_1 x_n + \\ & + c_{n+1} x_2^2 + c_{n+2} x_2 x_3 + \ldots + c_{2n-1} x_2 x_n + \\ & \ldots \\ & + c_{\frac{n^2+n}{2}} x_n^2 + c_{\frac{n^2+n}{2}+1} x_1 + \ldots + c_{m-1} x_n + c_m. \end{aligned} \tag{4.18}$$

Полагая $y(x) \triangleq (x_1^2, x_1 x_2, \ldots, x_1 x_n, x_2^2, \ldots, x_n^2, x_1, \ldots, x_n, 1)$, представим (4.18) в виде

$$f(x, c) = \langle c, y \rangle = y^{\mathrm{T}} c,$$

что позволяет говорить о линейной регрессионной задаче оценки параметров $c_1, \ldots, c_m$. Система нормальных уравнений, отвечающая МНК-функционалу (4.17), имеет вид

$$Y_N Y_N^{\mathrm{T}} c \triangleq \sum_{k=1}^{N} y^k (y^k)^{\mathrm{T}} c = Y_N J^N \triangleq \sum_{k=1}^{N} J_k y^k, \tag{4.19}$$

где $Y_N = (y^1, y^2, \ldots, y^N)$ — матрица размерности $m \times N$; $y^k \triangleq y(x^k)$; $J^N \triangleq \triangleq (J_1, J_2, \ldots, J_N)^{\mathrm{T}}$. Заметим, что так как квадратичный функционал $F(c) \geq 0$ и $F''(c) = YY^{\mathrm{T}}$, то матрица YY^{T} неотрицательно определена. Можно рассчитать коэффициенты квадратичной модели непосредственно из системы линейных алгебраических уравнений (4.19). Однако с вычислительной точки зрения более рациональным может оказаться другой подход, позволяющий избегать решения заведомо плохо обусловленных линейных систем.

Известно [1], что для псевдообратной матрицы $(\cdot)^+$ выполняется соотношение

$$(Y^{\mathrm{T}})^+ = \lim (\delta E + YY^{\mathrm{T}})^{-1} Y,\ \delta \to 0,\ \delta > 0.$$

Отсюда следует, что вместо системы уравнений

$$Y_r Y_r^{\mathrm{T}} c = Y_r J^r \tag{4.20}$$

можно рассматривать систему

$$(\delta E + Y_r Y_r^{\mathrm{T}}) c = Y_r J^r, \tag{4.21}$$

решение которой при $\delta \to 0$ сходится к решению (4.20) с минимальной нормой среди всех векторов, минимизирующих значение

$$\left\| Y_r^{\mathrm{T}} c - J^r \right\|^2 \to \min_c,$$

что при $r = N$ совпадает с выражением (4.17). Далее на основе известных рекуррентных алгоритмов оценивания, соответствующих случаю $\delta = 0$, будут построены методы решения регуляризованных систем (4.21) при конечных малых значениях параметра δ. В данном случае δ играет роль параметра регуляризации, обеспечивая устойчивость получаемых решений к ошибкам округления. При $\delta = 0$ решение системы (4.20) может наталкиваться на существенные вычислительные трудности, так как при $N < m$ матрицы $Y_N Y_N^{\mathrm{T}}$ будут вырождены, а при $N > m$ — плохо обусловлены. Введем обозначение

$$P_r^{-1} \triangleq \sum_{k=1}^{r} y^k (y^k)^{\mathrm{T}} + \delta E = P_{r-1}^{-1} + y^r (y^r)^{\mathrm{T}}. \tag{4.22}$$

Используя так называемую вторую лемму об обращении матриц (эквивалентную формуле Шермана—Моррисона—Вудбери [26])

$$[K^{-1} + B^{\mathrm{T}} R^{-1} B]^{-1} = K - KB^{\mathrm{T}} [BKB^{\mathrm{T}} + R]^{-1} BK,$$

получим рекуррентное соотношение [3], [20]

$$P_r = [P_{r-1}^{-1} + y^r (y^r)^{\mathrm{T}}]^{-1} = P_{r-1} - P_{r-1} y^r [(y^r)^{\mathrm{T}} P_{r-1} y^r + 1]^{-1} (y^r)^{\mathrm{T}} P_{r-1}. \tag{4.23}$$

Так как, очевидно, $P_1^{-1} = y^1 (y^1)^{\mathrm{T}} + \delta E$, то из (4.22) следует, что необходимо принять $P_0^{-1} = \delta E$ или $P_0 = \delta^{-1} E$.

Матрица P_0^{-1} симметрична и положительно определена. Предположим, что P_{r-1}^{-1} симметрична и положительно определена, и покажем, что P_r^{-1} обладает этими же свойствами. Последнее, согласно теории симметричных возмущений [63], немедленно следует из представления (9), так как матрица $y^r (y^r)^{\mathrm{T}}$ симметрична и неотрицательно определена. Поэтому, по принципу индукции, все матрицы P_r^{-1}, а вместе с ними и P_r будут симметричны и положительно определены. Таким образом, все обратные матрицы в (4.23) существуют.

Построим рекуррентное соотношение для определения оценок c_i^r. Уравнение (4.21) имеет вид

$$P_r^{-1} c^r = \sum_{k=1}^{r} J_k y^k, \tag{4.24}$$

откуда

$$P_r^{-1} c^r = \sum_{k=1}^{r} J_k y^k + y^r J_r = P_{r-1}^{-1} c^{r-1} + y^r J_r. \tag{4.25}$$

Здесь c^r означает оценку вектора c по r вычислениям функционала $J(x)$. Прибавляя и вычитая $y^r (y^r)^{\mathrm{T}} c^{r-1}$ в правой части (4.25), получим

$$P_r^{-1} c^r = P_{r-1}^{-1} c^{r-1} + y^r [J_r - (y^r)^{\mathrm{T}} c^{r-1}]. \tag{4.26}$$

Из (4.26) имеем окончательное выражение

$$c^r = c^{r-1} + P_r y^r [J_r - (y^r)^{\mathrm{T}} c^{r-1}]. \qquad (4.27)$$

По формуле (4.27) может быть вычислена новая, оценка c^r вектора параметров при условии, что известны предыдущая оценка c^{r-1}, матрица P_r и вновь полученные численные данные J_r, $y^r = y(x^r)$.

Последовательный метод оценки параметров квадратичной модели функционала J позволяет заменить процедуру обращения матрицы полной нормальной системы уравнений (4.21) операцией вычисления скаляра, обратного к заданному $(y^r)^{\mathrm{T}} P_{r-1} y^r + 1$, выполняемой на каждом шаге итерационного процесса (4.23).

Непосредственно из построения уравнения видно, что результат для $r = N$, полученный согласно (4.23), (4.27), приводит к оценке, которая получается из решения полной системы (4.21). При этом необходимо принять $c^0 = 0$. Последнее следует из соотношений (4.24), (4.27), записанных для $r = 1$.

Действительно, согласно (4.24) имеем оценку $c^1 = P_r J_1 y^1$, полученную в результате решения системы (4.21). Из (4.27) следует $c^1 = c^0 + P_1 y^1 [J_1 - (y^1)^{\mathrm{T}} c^0]$. Поэтому для совпадения обеих оценок c^1, а значит и последующих оценок, необходимо и достаточно принять $c^0 = 0$.

На основе вычисленных оценок c_i^N, $i \in [1{:}n^2/2 + n]$, задающих аппроксимацию матрицы $J''(x)$, может быть реализована процедура ОПС. При этом возможны различные стратегии применения изложенного общего подхода, конкретизирующие способ выбора числа «измерений» y^r, J_r, участвующих в коррекции текущей оценки, а также самих точек y^r. Целесообразно после каждого поворота осей обновлять процесс и вновь начинать процедуру построения аппроксимации. Такая тактика позволяет не учитывать «устаревшие» значения J, расположенные достаточно далеко от текущей точки.

Изложенная процедура обладает определенными свойствами адаптации по локализации окрестности текущей точки, в которой строится аппроксимирующая квадратичная модель. Действительно, если норма результирующего вектора продвижения в текущих осях достаточно велика, то исходный функционал заменяется квадратичным в достаточно широкой области пространства поиска. Если же оси выбраны неудачно и продвижение мало, то автоматически на формирование квадратичной модели оказывают влияние только близкие точки, и тем самым область предполагаемой «квадратичности» функционала сжимается.

Опыт применения такого типа алгоритмов для целей оптимизации в настоящее время недостаточен. Однако можно ожидать, что в ряде случаев будут возникать трудности, связанные с рациональным выбором δ, определяющим, в частности, погрешности промежуточных вычислений и их влияние на результат. В этом смысле подбор δ необходимо начинать с относительно больших значений, позволяющих с достаточной точностью получать «малые разности больших величин» при реализации соотношений (4.23). Кроме того, необходимо учитывать, что указанная реализация методов ОПС приводит к увеличению объема используемой памяти компьютера.

Другой подход к применению рекуррентных методов оценивания параметров линейных моделей для целей оптимизации может быть основан на *модифицированном алгоритме Качмажа* [11].

Используя представление (4.18), записанное в виде

$$f = y^{\mathrm{T}} c, \tag{4.28}$$

где c — вектор оцениваемых параметров, получим следующую рекуррентную процедуру уточнения оценок c_i параметров c_i^*:

$$c^k = d^k + \frac{f_k - \langle d^k, y^k \rangle}{\langle y^k, y^k \rangle} y^k; \tag{4.29}$$

$$d^k = \begin{cases} d^{k-1}, \text{ если выполнено неравенство (4.30)}, \\ c^{k-1}, \text{ в противном случае;} \end{cases}$$

$$\frac{\left(f_k - \langle y^k, d^{k-1} \rangle\right)^2}{\langle y^k, y^k \rangle} > \frac{\left(f_{k-1} - \langle y^{k-1}, d^{k-1} \rangle\right)^2}{\langle y^{k-1}, y^{k-1} \rangle} + \frac{\left(f_k - \langle y^k, c^{k-1} \rangle\right)^2}{\langle y^k, y^k \rangle}. \tag{4.30}$$

Геометрически алгоритм (4.29) реализует операцию проецирования вектора d^k на k-ю гиперплоскость (4.28), что приводит к монотонной (в евклидовой норме) сходимости последовательности оценок к точным значениям. В классическом варианте алгоритма Качмажа $d^k \equiv c^{k-1}$, что, однако, вызывает более медленную сходимость.

Основные достоинства алгоритма (4.29) заключаются в небольшом количестве вычислений для реализации соотношений (4.29), (4.30), а также в существенно меньших объемах используемой памяти компьютера по сравнению с рекуррентным методом наименьших квадратов. Кроме того, алгоритм (4.29) сохраняет эффективность при наличии малых помех измерений и медленном дрейфе вектора параметров c^*, приводя к достаточно точным оценкам.

Реализация методов ОПС с применением модифицированного алгоритма Качмажа сводится к следующей процедуре.

Алгоритм KACZM.

Шаг 1. Ввести исходные данные: $x \in R^n$, $c \in R^N \left(N = \frac{n^2}{2} + \frac{3n}{2} + 1\right)$, $s \in R^1$; принять $F := J(x)$, $d := c$; $y^0 := y(x)$, $F_0 := F$.

Шаг 2. Принять $h_i := s$, $i \in [1: n]$.

Шаг 3. Принять $U = E$, где E — единичная матрица; в качестве координатных векторов взять столбцы $\{u^i\}$ матрицы U.

Шаг 4. Принять $m := 1$.

Шаг 5. Принять $x := x + h_m u^m$; $y^1 := y(x)$; вычислить $F_1 = J(x)$. Если

$$\frac{\left(F_1 - \langle y^1, d\rangle\right)^2}{\langle y^1, y^1\rangle} \le \frac{\left(F_0 - \langle y^0, d\rangle\right)^2}{\langle y^0, y^0\rangle} + \frac{\left(F_1 - \langle y^1, c\rangle\right)^2}{\langle y^1, y^1\rangle},$$

принять $d = c$.

Шаг 6. Принять

$$c := d + \frac{F_1 - \langle d, y^1\rangle}{\langle y^1, y^1\rangle} y^1;\ F_0 := F_1;\ y^0 := y^1.$$

Шаг 7. Если $F_1 \le F$, принять $h_m := 3h_m$, $F := F_1$ и перейти к шагу 9, иначе — перейти к шагу 8.

Шаг 8. Принять $x := x - h_m u^m$, $h_m := -0{,}5h_m$.

Шаг 9. Принять $m := m + 1$. Если m ≤ n, перейти к шагу 5, иначе — к шагу 10.

Шаг 10. Проверить условия поворота осей. Если они выполнены, перейти к шагу 11; иначе — к шагу 4.

Шаг 11. На основе вычисленных оценок c_i, $i \in [1: (n^2 + n)/2]$, построить аппроксимацию G матрицы $J''(x)$.

Шаг 12. С помощью процедуры *jacobi* построить ортогональную матрицу $U = \{u^i\}$, приводящую матрицу G к диагональному виду $U^T G U$; перейти к шагу 4.

Процесс заканчивается по исчерпании заданного числа вычислений $J(x)$. Условия поворота осей совпадают с таковыми в алгоритмах GZ1, SPAC1, SPAC2. При выполнении шагов 5, 6 производится проверка на корректность соответствующих операций деления. Если деление невозможно, очередное измерение F_1 игнорируется.

4.5. Специальные реализации методов обобщенного покоординатного спуска

Специальные реализации методов ОПС позволяют использовать структурные особенности отдельных классов задач теории управления для повышения эффективности соответствующих оптимизирующих процедур.

4.5.1. Задачи аппроксимации

Характерные для теории управления целевые функционалы, отражающие меру «близости» расчетных и желаемых зависимостей, могут быть представлены в виде

$$J(x) = \nu^{-1} \sum_{k=1}^{m} \varphi_k^{\nu}(x),\ \nu = 2, 3, \ldots, \qquad (4.31)$$

где φ_k — алгоритмически заданные функции ($R^n \to R$) вектора управляемых параметров. Как отмечалось в 2.8.1, класс (4.31) включает в себя наиболее часто

используемые на практике МНК-критерии, а также минимаксные целевые функционалы.

Из (4.31) имеем следующие выражения для составляющих вектора-градиента:

$$\frac{\partial J}{\partial x_i} = \sum_{k=1}^{m} \varphi_k^{\nu-1}(x) \frac{\partial \varphi_k(x)}{\partial x_i}, \; i \in [1{:}\, n]. \tag{4.32}$$

Естественный способ аппроксимации вторых производных, принятый, в частности, в процедурах оптимизации типа Гаусса—Ньютона [44], а также в теории чувствительности систем автоматического управления, состоит в линеаризации функций φ_k вблизи текущей точки x':

$$\varphi_k(x) \cong \varphi_k(x') + \left\langle \frac{\partial \varphi_k(x')}{\partial x}, \; x - x' \right\rangle. \tag{4.33}$$

Используя (4.33), получим

$$\frac{\partial^2 J}{\partial x_i \partial x_j} \cong (\nu - 1) \sum_{k=1}^{m} \varphi_k^{\nu-2}(x) \frac{\partial \varphi_k(x)}{\partial x_i} \frac{\partial \varphi_k(x)}{\partial x_j}; \; \nu = 2, 3, \ldots; \; j \geq i; \; i \in [1{:}\, n]. \tag{4.34}$$

В векторно-матричных обозначениях использование аппроксимаций (4.34) эквивалентно отбрасыванию второго слагаемого в представлении

$$G(x) = (\nu - 1) F^{\mathrm{T}}(x) \operatorname{diag}[\varphi_i^{\nu-2}(x)] F(x) + Q(x), \tag{4.35}$$

где $F(x)$ — $m \times n$-матрица Якоби вектор-функции $\varphi(x) = [\varphi_1(x), \ldots, \varphi_m(x)]$; $G_i(x)$ — матрица Гессе функции $\varphi_i(x)$; $Q(x) = \sum_{i=1}^{m} \varphi_i^{\nu-1} G_i(x)$.

Дополнительным доводом в пользу правомерности указанного подхода является предположение о «малости» функций φ_i, по крайней мере, в некоторой окрестности минимизатора x^* функционала (4.31) [19].

Метод ОПС с вычислением аппроксимации матрицы Гессе $G(x)$ на основе формул (4.34) реализован с помощью следующего алгоритма.

Алгоритм SPAC5.

Шаг 1. Ввести исходные данные: x, s, ν.

Шаг 2. Вычислить матрицу $F = \{f_{ij}\}$, $f_{ij} = \varphi_i(x + se^j) - \varphi_i(x - se^j)$; $i \in [1{:}\, m]$; $j \in [1{:}\, n]$.

Шаг 3. Принять $B = F^T \operatorname{diag}[\varphi_i^{\nu-2}] F$.

Шаг 4. С помощью процедуры *jacobi* построить ортогональную матрицу U, приводящую матрицу B к диагональному виду $U^{\mathrm{T}} B U$.

Шаг 5. В осях $\{u^i\}$, совпадающих со столбцами матрицы U, реализовать процесс ЦПС (алгоритм GZ1) из точки x до выполнения условия поворота осей (совпадает с условием поворота в алгоритме SPAC1); присвоить x полученное лучшее значение, модифицировать s и перейти к шагу 2.

Окончание процесса и процедура пересчета s — такие же, как и в случае алгоритма SPAC1. Предполагается, что пользователь должен иметь подпрограмму, вычисляющую вектор $\varphi(x) = [\varphi_1(x), \ldots, \varphi_m(x)]$.

Далее рассматриваются вопросы применения алгоритмов типа SPAC1, SPAC5 для решения конкретных классов задач теории управления.

4.5.2. Идентификация нелинейных детерминированных объектов на основе функциональных рядов Вольтерра

Согласно выражению (1.16), имеем следующее представление для целевого функционала, отражающего ошибку идентификации:

$$J(x) = \frac{1}{2}\sum_{k=1}^{N_0}\left[\sum_{\nu=1}^{N_1} x_\nu y_{\nu 1}[k] + \sum_{\nu=1}^{N_2} x_{N_1+\nu} y_{\nu 2}[k] - H[k]\right]^2, \tag{4.36}$$

где $y_{\nu 1}[k]$, $y_{\nu 2}[k]$ — заданные функции дискретного переменного k; $H[k]$ — заданная дискретная аппроксимация выходного сигнала объекта. Множитель 0,5 здесь добавлен для удобства последующих записей. На процесс оптимизации он, очевидно, не влияет.

Функционал (4.36) имеет вид (4.31) с параметром $\nu = 2$, где

$$\varphi_k(x) = \langle x, y[k]\rangle - H[k]; \tag{4.37}$$

$$x = (x_1, x_2, \ldots, x_n);\ n = N_1 + N_2;\ y[k] = (y_{11}[k], \ldots, y_{N_1 1}[k], y_{12}[k], \ldots, y_{N_2 2}[k]).$$

Из (4.37) следует, что в данном случае приближенные равенства (4.33), (4.34) выполняются точно.

Элементы гессовой матрицы $\{g_{ij}\}$ равны

$$g_{ij} = \sum_{k=1}^{N_0} y_i[k] y_j[k];\ i, j \in [1{:}n]. \tag{4.38}$$

В силу линейной зависимости φ_k от x функционал (4.36) является параболоидом, и его минимизатор x^* может быть найден из решения линейной системы нормальных уравнений вида $J'(x) = 0$ с матрицей $G = \{g_{ij}\}$, имеющей элементы (4.38). Однако с вычислительной точки зрения более эффективным обычно оказывается подход, основанный на непосредственной минимизации (4.36). Дадим необходимые разъяснения.

Как отмечалось, задачи минимизации функционалов типа (4.36) оказываются очень плохо обусловленными. Это приводит к известным вычислительным трудностям при решении линейных систем нормальных уравнений

$$Gx = b, \tag{4.39}$$

отражающих необходимые условия экстремума. Здесь $G = Y^{\mathrm{T}}Y$, где Y — $(N_0 \times n)$-матрица вида

$$Y = \begin{bmatrix} y_1[1] & \dots & y_n[1] \\ \dots & \dots & \dots \\ y_1[N_0] & \dots & y_n[N_0] \end{bmatrix}. \tag{4.40}$$

Задача решения (4.39) оказывается некорректной по Тихонову [61], а привлечение известных методов регуляризации наталкивается на принципиальные трудности, связанные с отсутствием необходимой априорной информации, определяемой истинными целями решения исходной задачи идентификации.

Механизм влияния вычислительных погрешностей на решение системы (4.39) можно проследить из представления решения в виде

$$x^* = [u^1, u^2, \dots, u^n] \begin{bmatrix} \lambda_1^{-1}\langle b, u^1\rangle \\ \lambda_2^{-1}\langle b, u^2\rangle \\ \dots \\ \lambda_n^{-1}\langle b, u^n\rangle \end{bmatrix}, \tag{4.41}$$

где $J(x) = \frac{1}{2}\langle Gx, x\rangle - \langle b, x\rangle + c$, $c \in R^1$, u^i, λ_i — соответственно, собственные векторы (ортонормированные) и собственные числа матрицы G. Данное соотношение следует из очевидного матричного представления

$$x^* = G^{-1}b = U\,\mathrm{diag}\,(\lambda_i^{-1})U^{\mathrm{T}}b,$$

где $U = [u^1, u^2, \dots, u^n]$.

Из (4.41) видно, что компоненты x^* в основном определяются малыми собственными числами, и уже небольшая погрешность в их представлении приводит к большой ошибке в компонентах вектора x^*. В силу этого использование для определения x^* линейной системы (4.39), содержащей в явном виде матрицу G, недопустимо, так как из-за ограниченной точности представления элементов g_{ij}

$$g_{ij} = \sum_{k=1}^{n} u_i^k u_j^k \lambda_k \tag{4.42}$$

матрицы G информация о «малых» собственных числах $\lambda_{n-r+1}, \dots, \lambda_n$ теряется на фоне «больших» $\lambda_1, \dots, \lambda_{n-r}$. Указанное обстоятельство и приводит к некорректности задачи (4.39). Приведем конкретный численный пример.

Пример [74]. Рассмотрим квадратичный функционал с высокой степенью овражности $\eta = 10^{12}$:

$$J = 0{,}5\sum_{i=1}^{4} \lambda_i \langle x, u^i\rangle^2 - \langle b, x\rangle;\ b = (1; 1; 1; 1).$$

Собственные числа матрицы Гессе J'' равны: $\lambda_1 = 10^8, \lambda_2 = \lambda_4 = 10^{-4}, \lambda_3 = 10^6$. Собственные векторы u^i есть

$$u^1 = \frac{1}{\sqrt{3}}(1;-1;1;0);\ u^2 = \frac{1}{\sqrt{6}}(1;2;1;0);$$

$$u^3 = \frac{1}{\sqrt{3}}(1;0;-1;1);$$

$$u^4 = \frac{1}{\sqrt{6}}(1;0;-1;-2);\ x^0 = (0;0;0;0);$$

$$x^* \cong (3333{,}3;\ 13333;\ 10000;\ 6666{,}7);\ J(x^*) \cong -16667.$$

Точные значения компонентов вектора x^* задаются выражением (4.41):

$$x^* = (u^1; u^2; u^3; u^4)\left(\lambda_1^{-1}\langle b, u^1\rangle;\ \lambda_2^{-1}\langle b, u^2\rangle;\ \lambda_3^{-1}\langle b, u^3\rangle;\ \lambda_4^{-1}\langle b, u^4\rangle\right) \cong$$

$$\cong (u^1; u^2; u^3; u^4)\left(0;\ 10^4\langle b, u^2\rangle;\ 0;\ 10^4\langle b, u^4\rangle\right).$$

Отсюда видно, что компоненты x^* в основном определяются малыми собственными числами λ_2, λ_4, и уже небольшая погрешность в их представлении приводит к большой ошибке в результате. Необходимо отметить, что, например, при написании тестовой программы, осуществляющей вычисление значений J, следует использовать вышеприведенное представление функционала в виде суммы. Применение для этой цели обычного выражения квадратичного функционала $J = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$, содержащего в явном виде матрицу $A = J''$, недопустимо, так как из-за ограниченной точности представления элементов a_{ij} матрицы в памяти компьютера информация о малых собственных числах λ_2, λ_4 теряется на фоне больших λ_1, λ_3. Указанное обстоятельство приводит к резкой потере эффективности методов ньютоновского типа, основанных на существенном использовании информации о малых собственных числах при явном представлении аппроксимаций матриц Гессе минимизируемого функционала. Все сказанное, очевидно, относится к попыткам искать решение с помощью линейных систем (4.39).

Привлечение методов регуляризации для решения (4.39) позволяет определить «квазирешение» $\tilde{x}^*$, отражающее некоторый компромисс между величиной нормы $\|\tilde{x}^*\|$ и невязкой $\|G\tilde{x}^* - b\|$. При этом относительно малым невязкам могут соответствовать относительно большие ошибки как по аргументу $\|\tilde{x}^* - x^*\|$, так и по функционалу $J(\tilde{x}^*) - J(x^*)$. Действительно, рассмотрим одномерный случай параболы $y(x) = \frac{1}{2}ax^2$. Тогда невязка будет определяться величиной градиента (в данном случае — производной) $y'(x) = ax$. Здесь коэффициент a моделирует влияние некоторого «малого» собственного числа. Полагая $a = 2 \cdot 10^{-6}$, $\tilde{x}^* = 10^4$, получим $y'(\tilde{x}^*) = 2 \cdot 10^{-2}$, $y(\tilde{x}^*) = 10^2$ при $y(x^*) = 0$.

Этот пример полностью отражает общую ситуацию. Важно при этом понимать, что в данном случае только для простоты выкладок начало координат совмещено с точкой x^*. При проведении регуляризации в общем случае из условия «ма-

В силу изложенного на практике довольно редко обращаются к непосредственному решению нормальных уравнений (4.39). Стандартный подход описан в учебнике [31] и заключается в построении сингулярного разложения прямоугольной матрицы (4.40)

$$Y = U\Sigma V^{\mathrm{T}},$$

где U — ортогональная $N_0 \times N_0$-матрица; V — ортогональная $n \times n$-матрица; Σ — диагональная $N_0 \times n$-матрица, у которой $\sigma_{ij} = 0$ при $i \neq j$ и $\sigma_{ii} \triangleq \sigma_i \geq 0$. Вектор $\hat{x}^*$, аппроксимирующий точный минимизатор функционала (4.36), выражается соотношениями $\hat{x}^* = V\hat{z}^*$, где $\hat{z}_i^* = \dfrac{d_i}{\sigma_i}$, если $\sigma_i \geq \delta$; в противном случае $\hat{z}_i^*$ — произвольно. Обычно при $\sigma_i < \delta$ полагают $\hat{z}_i^* = 0$, снова минимизируя длину $\left\| \hat{x}^* \right\|$ и добиваясь за счет этого единственности решения. В данном случае $d = U^{\mathrm{T}}h$, $h = (H[1], \ldots, H[N_0])^{\mathrm{T}}$. Величину δ, отражающую уровень «малости» соответствующего сингулярного числа, целесообразно полагать равной $n\lambda_1\varepsilon_{\text{м}}$, где $\lambda_1 = \max\limits_i |\lambda_i|$. Введение δ, по существу, реализует некоторый алгоритм регуляризации и оказывает заметное влияние на окончательный результат. Как и в предыдущем случае, необходимая априорная информация для обоснованного выбора δ и значений $\hat{z}_i^*$ при $\sigma_i < \delta$ (с позиций исходной задачи идентификации) здесь также отсутствует.

Иная ситуация складывается при решении задачи идентификации с помощью прямой минимизации целевого функционала (4.31) методами типа ОПС. Предполагая, что полная информация о минимизаторе x^* не теряется при реализации алгоритма вычисления $J(x)$, мы используем матрицу G только для выбора наиболее рациональной системы координат. Далее в процессе минимизации производятся многократные дополнительные вычисления $J(x)$ с поступлением новой полезной информации об истинных значениях компонентов вектора x^*. Это приводит к определенному эффекту усиления «полезного сигнала». Таким образом, прямая минимизация функционала (4.31) является наиболее предпочтительным подходом.

4.5.3. Корреляционные методы идентификации стохастических объектов

В соответствии с результатами, полученными в 1.3.2, задача идентификации линейного объекта со стационарным случайным входным сигналом сводится к интегральному уравнению Винера—Хопфа. Один из возможных подходов к его решению заключается в минимизации регуляризованных функционалов вида (3.20)

$$J(x) = 0{,}5\sum_{j=1}^{N}\left[\sum_{i=1}^{N}\omega_i\hat{R}_{GG}\left[q(j-i)\right] - \hat{R}_{\hat{H}G}[qj]\right]^2 + 0{,}5\alpha\sum_{j=2}^{N}\left[\omega_j - \omega_{j-1}\right]^2, \quad (4.43)$$

где $x = (\omega_1, \omega_2, \ldots, \omega_N) \in R^N$; $\hat{R}_{GG}$, $\hat{R}_{\hat{H}G}$ — заданные оценки соответствующих корреляционных функций (см. рис. 1.11); α — параметр регуляризации.

При отсутствии необходимости проводить регуляризацию имеем $\alpha = 0$, и минимизация функционала (4.43), осуществляется по методике, изложенной в 4.5.2, алгоритмами типа SPAC5.

При $\alpha \neq 0$ целесообразно использовать алгоритмы со специальной реализацией шага 2 вычисления матрицы Гессе. Элементы $g(x)$ и $G(x)$ в данном случае вычисляются по точным формулам:

$$\frac{\partial J}{\partial x_l} = \sum_{j=1}^{N}\left[\sum_{i=1}^{N}\omega_i \hat{R}_{GG}[q(j-i)] - \hat{R}_{\hat{H}G}[qj]\right]\hat{R}_{GG}[q(j-i)] + \alpha\psi_l, l \in [1:N], \quad (4.44)$$

$$\Psi_l = \begin{cases} (\omega_l - \omega_{l-1}) - (\omega_{l+1} - \omega_l), 1 < l < N; \\ -(\omega_2 - \omega_1), l = 1; \\ (\omega_N - \omega_{N-1}), l = N. \end{cases}$$

$$\frac{\partial^2 J}{\partial x_l \partial x_m} = \sum_{j=1}^{N}\hat{R}_{GG}[q(j-m)]\hat{R}_{GG}[q(j-l)] + \alpha\varphi_{lm}, l, m \in [1:N], \quad (4.45)$$

где матрица $\varphi = \{\varphi_{lm}\}$ имеет вид

$$\varphi = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & \dots & 1 & 1 \end{bmatrix}.$$

4.5.4. Синтез статистически оптимальных систем автоматического управления

Как известно, многие задачи синтеза линейных оптимальных систем по статистическим критериям как при стационарных, так и нестационарных воздействиях сводятся к решению интегральных уравнений Винера—Хопфа. Учет нестационарности входных воздействий приводит к необходимости решения последовательности стационарных задач.

Решение последних с помощью методов ОПС может быть проведено по методике, изложенной в 4.5.3, и не требует дополнительных разъяснений.

4.5.5. Идентификация нелинейных динамических систем

Пусть траектория $z(t, x)$ динамической системы описывается уравнением

$$\dot{z}(t,x) = f(t,z,x), z(t_0,x) = z^0, t_0 \le t \le T, \quad (4.46)$$

где z — r-мерный вектор фазовых координат; z^0 — известный начальный вектор; x — n-мерный вектор постоянных неизвестных параметров; f — непрерывно дифференцируемая функция своих аргументов, удовлетворяющая в некоторой замкнутой области при $x \in X$ условиям теоремы существования и единственности решения. Требуется по результатам измерения вектор-функции

$$y(t) = H(t)\, z(t, x),\ t \in [t_0, T] \tag{4.47}$$

определить вектор параметров x из условия минимума функционала

$$J(x) = 0{,}5\sum_{j=1}^{N} \left\| H(t_j) z(t_j, x) - y(t_j) \right\|^2, \tag{4.48}$$

где $H(t)$ — непрерывная матрица размерности $l \times r$, которая связывает вектор измерения с вектором состояния $(l < r)$; t_j $(j \in [1{:}\ N], N \geq n)$ — дискретные моменты съема измерительной информации.

Функционал (4.48) с точностью до обозначений имеет, очевидно, вид (4.31)

$$J(x) = 0{,}5\sum_{k=1}^{N}\sum_{s=1}^{l} \varphi_{sk}^2(x), \tag{4.49}$$

где $\varphi_{sk}(x) = \langle h_s(t_k), z(t_k, x)\rangle - y_s(t_k)$; h_s — s-я строка матрицы H.

Из (4.49) получаем представление, аналогичное (4.34):

$$\frac{\partial^2 J}{\partial x_i \partial x_j} \cong \sum_{k=1}^{N}\sum_{s=1}^{l} \frac{\partial \varphi_{sk}}{\partial x_i}\frac{\partial \varphi_{sk}}{\partial x_j} = \sum_{k=1}^{N}\sum_{s=1}^{l} \left\langle h_s(t_k), \frac{\partial z(t_k, x)}{\partial x_i} \right\rangle \left\langle h_s(t_k), \frac{\partial z(t_k, x)}{\partial x_j} \right\rangle, \tag{4.50}$$

$$i, j \in [1{:}\ n].$$

Таким образом, для данного класса задач построение аппроксимации матрицы $G(x)$ сводится к построению матрицы Якоби $F(t, x) = \dfrac{\partial z(t, x)}{\partial x}$ вектор-функции $z(t, x)$ в дискретные моменты времени t_k, $k \in [1{:}\ N]$. Для вычисления матрицы $F(t, x)$, имеющей в качестве элементов соответствующие функции чувствительности системы (4.46), при фиксированном $x = \tilde{x}$ решается линейное матричное уравнение

$$\frac{d}{dt} F(t, \tilde{x}) = \frac{\partial f(t, z, \tilde{x})}{\partial z} F(t, \tilde{x}) + \frac{\partial f(t, z, \tilde{x})}{\partial x} \tag{4.51}$$

с $F(t_0, \tilde{x}) = 0$, $t \in [t_0, T]$. При этом зависимость $z(t, x)$ при фиксированном x получается непосредственным интегрированием (4.46). Уравнение (4.51) выводится в любом стандартном учебнике по дифференциальным уравнениям.

Важно понимать, что в данном случае квадратичная модель целевого функционала (4.48) вводится лишь для выбора наиболее рациональной системы координат, а минимизации подвергается исходный функционал (4.48).

Пример. Пусть фазовые траектории некоторой динамической системы имеют вид

$$z(t, x) = \varphi(t)[\exp(-x_1 t) + \exp[-x_2 t]],\ t_0 = 0,\ z^0 = 2\varphi(0),$$

где $\varphi(t)$ — заданная функция, принимающая значения $\varphi(1) = \exp(0, 1)$, $\varphi(2) = 0{,}5\exp(0, 1)$. Требуется определить вектор $x = (x_1, x_2)$ из условия минимума функционала

$$J(x)=\sum_{j=1}^{2}\left[z(t_j,x)-\bar{z}_j\right]^2,$$

где $\bar{z}_j=z(t_j,x^*)$; $t_1=1$, $t_2=2$; $x_1^*=1$, $x_2^*=1{,}5$.

Для начальной точки $x^0=(0{,}1;\ 0{,}1)$, $J(x^0)=2{,}6006$ имеем согласно (4.50) (элемент матрицы g_{12} не показан, так как матрица симметрична)

$$G(x)\cong\left[\begin{array}{cc|c} \dfrac{\partial z(t_1,x)^2}{\partial x_1}+\dfrac{\partial z(t_2,x)^2}{\partial x_1} & & \\ \hline \dfrac{\partial z(t_1,x)}{\partial x_1}\dfrac{\partial z(t_1,x)}{\partial x_2}+\dfrac{\partial z(t_2,x)}{\partial x_1}\dfrac{\partial z(t_2,x)}{\partial x_2} & & \dfrac{\partial z(t_1,x)^2}{\partial x_2}+\dfrac{\partial z(t_2,x)^2}{\partial x_2} \end{array}\right]=\begin{bmatrix}2 & 2\\ 2 & 2\end{bmatrix}.$$

Прямая минимизация аппроксимирующего параболоида в данном случае затруднительна из-за вырожденности аппроксимации матрицы $G(x)$ в начальной точке $x=x^0$ и ее плохой обусловленности в окрестности x^0. При использовании же алгоритма SPAC5 никаких вычислительных проблем не возникает. В результате вычислений, эквивалентных по трудоемкости ста обращениям к подпрограмме, по которой осуществляется вычисление значений $J(x)$, была получена точка $\tilde{x}=(1{,}0001;\ 1{,}5000)$ с $J(\tilde{x})=1{,}5164\cdot10^{-11}$.

Начальное значение шага дискретности s полагалось равным 0,01.

4.5.6. Оценивание состояний динамических систем: задача о наблюдении

Уже говорилось, что для управления некоторой сложной системой или объектом необходимо иметь информацию о текущем состоянии системы. В то же время измеряемыми и непосредственно наблюдаемыми являются лишь некоторые физические выходные переменные $y=(y_1, y_2, \ldots, y_l)$, функционально связанные с вектором $z=(z_1, z_2, \ldots, z_r)$ состояний. Возникает задача определения $z(t_0)$ в некоторый заданный момент времени $t=t_0$ по данным измерения $y(t)$ (а иногда и управляющего воздействия $u(t)$) на конечном интервале времени $t_0\le t\le T$, $T>t_0$. Согласно введенной в 1.2 терминологии, сформулированная задача является задачей сглаживания.

Пусть движение системы определяется уравнением

$$\dot{z}(t)=f(t,z),\ z(t_0)=x,\ t_0\le t\le T, \tag{4.52}$$

где $f(t, z)$ — известная нелинейная непрерывно дифференцируемая функция своих аргументов, удовлетворяющая в некоторой замкнутой области условиям теоремы существования и единственности решения. Предположим, что векторы z и y связаны линейным соотношением

$$y(t)=H(t)z(t), \tag{4.53}$$

где $H(t)$ — известная $l\times r$-матрица, определяемая конструкцией измерительного устройства. При сделанных предположениях траектория $z(t)$ системы (4.52)

однозначно определяется начальным вектором $x = (x_1, ..., x_n)$. Оценки x_i, $i \in [1: n]$ могут быть определены из условия минимума целевого функционала

$$J(x) = 0{,}5\sum_{j=1}^{N} \left\| H(t_j)z(t_j) - y(t_j) \right\|^2. \tag{4.54}$$

Аналогично предыдущему подразделу имеем следующие формулы для элементов аппроксимации матрицы Гессе:

$$\frac{\partial^2 J}{\partial x_i \partial x_j} \cong \sum_{k=1}^{N}\sum_{s=1}^{l} \left\langle h_s(t_k), \frac{\partial z(t_k)}{\partial x_i} \right\rangle \left\langle h_s(t_k), \frac{\partial z(t_k)}{\partial x_j} \right\rangle, \tag{4.55}$$

$i, j \in [1: n]$.

В силу известных результатов теории дифференциальных уравнений матрица $F(t,x) \triangleq \dfrac{\partial z}{\partial x}$ производных по начальным условиям удовлетворяет в точке $x = \tilde{x}$ матричному уравнению

$$\frac{d}{dt}F(t,\tilde{x}) = \frac{\partial f(t,z)}{\partial z}F(t,\tilde{x}),\ F(t_0,\tilde{x}) = E,$$

где E — единичная матрица, а $z = z(t,\tilde{x})$ — результат предварительного интегрирования (4.52).

4.5.7. Идентификация возмущающих воздействий

Задача идентификации *возмущающих сил* формулируется следующим образом.

Пусть движение некоторой динамической системы описывается уравнением

$$\dot{z}(t) = f(t,z,v),\ z(t_0) = z_0, \tag{4.56}$$

где v — вектор возмущающих сил. Предполагая, что известна параметризация

$$v = v(t, x),\ x = (x_1, x_2, \ldots, x_n),\ t \in [t_0, T], \tag{4.57}$$

требуется определить вектор неизвестных параметров x. Не останавливаясь на обсуждении очевидных необходимых свойств функций f, v, отметим, что так поставленная задача идентификации возмущающих сил с точностью до обозначений совпадает с задачей, поставленной в 4.5.5. Более того, обе эти задачи могут решаться совместно по единой методике.

Неизвестный вектор x по измерениям y (4.53) может быть найден из условия минимума функционала

$$J(x) = 0{,}5\sum_{j=1}^{N} \left\| H(t_j)z(t_j,x) - y(t_j) \right\|^2, \tag{4.58}$$

где $z(t_j, x)$ — результат интегрирования (4.56) при фиксированном x. Приближенные выражения для элементов матрицы $G(x)$ определяются выражениями (4.55), где матрица $F(t,x) \triangleq \dfrac{\partial z(t,x)}{\partial x}$ является при $x = \tilde{x}$ решением матричного уравнения

$$\frac{d}{dt}F(t,\tilde{x}) = \frac{\partial f(t,z,v)}{\partial z}F(t,\tilde{x}) + \frac{\partial f(t,z,\tilde{v})}{\partial v}\frac{\partial v(t,\tilde{x})}{\partial x}, \tag{4.59}$$

где $\tilde{v} = v(t,\tilde{x})$; $F(t_0,\tilde{x}) = 0$.

4.5.8. Решение систем неравенств

Задачи с неравенствами весьма часто возникают в теории управления. При этом задача решения системы вида

$$y_j \le t_j, j \in [1:m] \tag{4.60}$$

одним из рассмотренных в 2.8.2 способов формулируется как оптимизационная. Соответствующие достаточно общему подходу целевые функционалы представимы в виде

$$J(x) = \nu^{-1}\sum_{k=1}^{m}\exp[-\nu z_k(x)], \nu = 1,2,3, \ldots, \tag{4.61}$$

где $z_k(x)$ отражает k-й «запас» в выполнении неравенств (4.60).

Из (4.61) получаем:

$$\frac{\partial J}{\partial x_i} = -\sum_{k=1}^{m}\exp[-\nu z_k(x)]\frac{\partial z_k(x)}{\partial x_i}; \tag{4.62}$$

$$\frac{\partial^2 J}{\partial x_i \partial x_j} = \sum_{k=1}^{m}\exp[-\nu z_k(x)]\left[\frac{\partial z_k}{\partial x_i}\frac{\partial z_k}{\partial x_j}\nu - \frac{\partial^2 z_k}{\partial x_i \partial x_j}\right]. \tag{4.63}$$

Выполняя линеаризацию z_k в окрестности текущей точки x'

$$z_k(x) \cong z_k(x') + \left\langle \frac{\partial z_k(x')}{\partial x}, x - x' \right\rangle,$$

получим из (4.63) используемую на практике аппроксимацию матрицы Гессе

$$\frac{\partial^2 J}{\partial x_i \partial x_j} \cong \nu\sum_{k=1}^{m}\exp[-\nu z_k(x)]\frac{\partial z_k}{\partial x_i}\frac{\partial z_k}{\partial x_j}, \nu = 1,2,3, \ldots; j \ge i, i \in [1:n]. \tag{4.64}$$

Из (4.63) следует, что при увеличении ν точность аппроксимации (4.64), вообще говоря, возрастает.

Для конкретных классов прикладных задач в ряде случаев возможно использование точных формул для вторых производных критериев оптимальности без существенного увеличения трудоемкости решения задачи.

4.5.9. Управление технологическим процессом серийного выпуска изделий

При выпуске изделий массового производства (например, интегральных микросхем) одним из важнейших производственных показателей является вероятность выхода годных изделий (см. формулу (2.49)):

$$J(x) \triangleq P\{y(x,\xi) \le t\} \to \max_x \tag{4.65}$$

Ставится задача выбора такого вектора управляемых параметров $x = (x_1, x_2, \ldots, x_n)$, чтобы обеспечить максимум вероятности выхода годных изделий. Будем далее предполагать, что помимо случайного вектора ξ, отражающего фактор неопределенности обстановки, величина P зависит также от некоторого контролируемого вектора параметров $\zeta(t)$, характеризующих, например, измеряемые параметры очередной партии сырья. В дискретные моменты времени t_i, отвечающие моментам скачкообразного изменения вектора ζ либо соответствующие существенному (в некотором определенном смысле) ухудшению показателя качества $J(x)$ за счет накопившихся непрерывных изменений $\zeta(t)$, целесообразно осуществлять перенастройку производственного процесса с помощью повторного решения задачи (4.65) и выбора нового оптимального вектора управляемых параметров x^*. Следовательно, в данном случае можно говорить об оптимальном *управлении процессом производства* по критерию вероятности выхода годных изделий. Вероятность P при фиксированном векторе параметров x вычисляется на основе статистических испытаний по методу Монте-Карло в соответствии с заданной плотностью $\Psi(x, \xi)$ распределения случайного вектора ξ. Здесь предполагается, что составляющие вектора ξ имеют смысл некоторых внутренних либо внешних параметров процесса и, что существенно, включают в себя все компоненты вектора x. При этом под x_i будут пониматься средние значения $\overline{\xi}_i$, за счет выбора которых и происходит управление процессом.

Критерий (4.65) может быть представлен в виде [8]

$$P(x) = \int_D \ldots \int \varphi(\xi)\Psi(x,\xi)\,d\xi, \tag{4.66}$$

где $\varphi(\xi)$ — калибровочная функция, равная 1, если при данном ξ система неравенств $y(x, \xi) \le t$, имеющих смысл условий работоспособности, выполнена; в противном случае $\varphi(\xi)$ полагается равной 0. В данном случае существенно, что $\varphi(\xi)$ не зависит от x. На основе предположения о возможности операции дифференцирования по x под знаком интеграла (4.66) и перехода к статистическим аналогам соответствующих соотношений могут быть получены выражения для составляющих градиента и матрицы Гессе критерия (4.65). Для случая, когда $\Psi(x, \xi)$ задает многомерное гауссовское распределение вида

$$\Psi(x,\xi) = \left[(2\pi)^{\frac{n}{2}} \prod_{i=1}^{n} \sigma_{\xi_i} \sqrt{R}\right]^{-1} \exp\left[-\frac{1}{2R}\sum_{i=1}^{n}\sum_{j=1}^{n} R_{ij}\frac{(\xi_i - x_i)(\xi_j - x_j)}{\sigma_{\xi_i}\sigma_{\xi_j}}\right], \tag{4.67}$$

справедливы следующие представления для производных:

$$\frac{\partial P}{\partial x_s} = \frac{1}{NR\sigma_s}\sum_{i=1}^{N}\varphi(\xi^l)\sum_{i=1}^{N}R_{si}\frac{\xi_i^l - x_i}{\sigma_i}; \tag{4.68}$$

$$\frac{\partial^2 P}{\partial x_s \partial x_t} = \frac{1}{NR\sigma_s\sigma_t}\left[-kR_{si} + R^{-1}\sum_{i=1}^{N}\varphi(\xi^l)\sum_{i=1}^{n}R_{si}\frac{\xi_i^l - x_i}{\sigma_i}\sum_{j=1}^{n}R_{ij}\frac{\xi_i^l - x_j}{\sigma_j}\right], \tag{4.69}$$

где N — число статистических испытаний по методу Монте-Карло; R, R_{ij} — соответственно, определитель и алгебраические дополнения, составленные из элементов матрицы коэффициентов парных корреляций; $\sigma_i = \sigma_{\xi_i}$ — среднее квадратичное отклонение параметра ξ_i; k — число «успешных» испытаний, при которых условия работоспособности оказываются выполненными.

Из (4.68), (4.69) следует, что элементы вектора градиента и матрицы Гессе критерия (4.65) при сделанных предположениях могут рассчитываться *без дополнительных затрат* по вычислению значений выходных параметров y одновременно с расчетом целевого функционала $P(x)$.

Рассмотренная техника дифференцирования при соответствующем обосновании выполняемых операций дает эффективный способ реализации методов ОПС для решения довольно широкого класса прикладных задач теории управления.

4.5.10. Обеспечение максимального запаса работоспособности оптимизируемой системы

Как отмечалось в 2.8, основные требования-спецификации к оптимизируемой системе могут быть выражены в виде системы неравенств (4.60). При этом основная задача формулируется следующим образом: обеспечить такой набор управляемых параметров x, при которых наилучшим образом выполняются все спецификации (4.60) во всем диапазоне изменения внешних параметров.

Количественная оценка степени выполнения j-го неравенства имеет смысл *запаса работоспособности* z_j. Наиболее объективно цели оптимизации отражаются при использовании конструкций вида (2.51):

$$z_j(x) \triangleq \left[\frac{(t_j - y_j(x))}{\delta_j} - 1\right]. \tag{4.70}$$

В случае использования представлений (4.70), где δ_j, по существу, играет роль весовых коэффициентов, общие соотношения (4.62), (4.63), приближенно реализующие принцип максимума минимального из запасов, принимают следующий вид:

$$\frac{\partial J}{\partial x_i} = \sum_{k=1}^{m}\exp\left[-\nu z_k(x)\right]\delta_k^{-1}\frac{\partial y_k(x)}{\partial x_i}; \tag{4.71}$$

$$\frac{\partial^2 J}{\partial x_i \partial x_j} = \sum_{k=1}^{m}\exp\left[-\nu z_k(x)\right]\left[\nu\delta_k^{-2}\frac{\partial y_k(x)}{\partial x_i}\frac{\partial y_k(x)}{\partial x_j} + \delta_k^{-1}\frac{\partial^2 y_k(x)}{\partial x_i \partial x_j}\right]. \tag{4.72}$$

Выполняя линеаризацию k-го выходного параметра y_k в окрестности текущей точки x'

$$y_k(x) \cong y_k(x') + \left\langle \frac{\partial y_k(x')}{\partial x},\ x - x' \right\rangle, \tag{4.73}$$

приходим к следующему представлению, аналогичному (4.64):

$$\frac{\partial^2 J}{\partial x_i \partial x_j} \cong \nu \sum_{k=1}^{m} \exp[-\nu z_k(x)] \left[\delta_k^{-2} \frac{\partial y_k(x)}{\partial x_i} \frac{\partial y_k(x)}{\partial x_j} \right], \tag{4.74}$$

$$\nu = 1, 2, 3, \ldots,\ j \geq i,\ i \in [1{:}n].$$

При реализации методов ОПС множитель ν в (4.74) может быть опущен. В ряде случаев соотношение (4.73), а вместе с ним и (4.74) выполняются точно. В случае достаточно простой аналитической структуры зависимостей $y_k(x)$ могут быть получены явные выражения для производных $\dfrac{\partial^2 y_k}{\partial x_i\, \partial x_j}$, что позволяет непосредственно воспользоваться соотношениями (4.72).

4.5.11. Оптимизация систем по сигномиальным целевым функционалам

Так называемые «простые» задачи оптимального параметрического синтеза систем характеризуются наличием известных аналитически заданных связей между управляемыми параметрами (параметрами оптимизации) и соответствующими значениями целевых функционалов. При этом, как показывает практика [62], весьма характерна *сигномиальная* структура указанных связей:

$$J(x) = \sum_{i=1}^{N} \sigma_i t_i(x), \tag{4.75}$$

$$t_i(x) = c_i \prod_{j=1}^{n} x_j^{\alpha_{ij}};\ c_i > 0;\ \alpha_{ij} \in R^1;\ \sigma_i = \pm 1;\ x_j > 0,\ j \in [1{:}n].$$

Кроме того, функционалы с такой структурой используются (по аналогии с квадратичными функционалами) как хорошие локальные модели функционалов общего вида. Поэтому наряду с *квадратичными методами* конечномерной оптимизации могут строиться *сигномиальные методы*. Функционалы вида (4.75) изучаются в специальном разделе теории нелинейного программирования — *геометрическом программировании*.

Существуют регулярные методы минимизации функционалов (4.75), основанные на теории двойственности в геометрическом программировании в предположении $\sigma_i = 1$, $i \in [1{:}\ N]$ [60]. Однако если число *степеней трудности* $\gamma > 0$, где $\gamma = N - (n + 1)$, то целесообразен прямой поиск минимума $J(x)$ с использованием ньютоновских процедур второго порядка [62]. Указанный подход становится практически наиболее оправданным в общем случае произвольных σ_i, когда

основные предпосылки метода двойственного геометрического программирования нарушаются.

Как показано в книге [62], применение Н-методов для решения задачи минимизации (4.75) сопряжено с известными трудностями из-за неустойчивости и расходимости численных процедур вследствие знаконеопределенности матриц Гессе. Поэтому по причинам, изложенным ранее, целесообразно обращение к методам ОПС. Применение процедур ОПС (так же как и Н-методов) в данном случае облегчается из-за наличия явных выражений для первых и вторых производных функционала (4.75). Действительно, выполняя, например, замену переменных $x_j = \exp(y_j)$, получим:

$$t_i(x) = c_i \prod_{j=1}^{n} \left[\exp(y_j)\right]^{\alpha_{ij}} = c_i \exp\left[\sum_{j=1}^{n} \alpha_{ij} y_j\right] = \tilde{t}_i(y);$$

$$\frac{\partial \tilde{t}_i(y)}{\partial y_k} = c_i \alpha_{ik} \exp\left[\sum_{j=1}^{n} \alpha_{ij} y_j\right] = \alpha_{ik} \tilde{t}_i(y); \qquad (4.76)$$

$$\frac{\partial^2 \tilde{t}_i(y)}{\partial y_k \partial y_l} = \alpha_{ik} \alpha_{il} \tilde{t}_i(y).$$

Из (4.76) имеем

$$\frac{\partial \tilde{J}}{\partial y_k} = \sum_{i=1}^{N} \sigma_i \alpha_{ik} c_i \exp\left[\sum_{j=1}^{n} \alpha_{ij} y_j\right]; \qquad (4.77)$$

$$\frac{\partial^2 J}{\partial y_k \partial y_l} = \sum_{i=1}^{N} \sigma_i \alpha_{ik} \alpha_{il} c_i \exp\left[\sum_{j=1}^{n} \alpha_{ij} y_j\right]. \qquad (4.78)$$

Перспективность указанной реализации методов ОПС определяется достаточно широкой сферой приложений математических моделей типа (4.75), не ограниченной задачами оптимального проектирования систем с функционалами качества, непосредственно сводимых к сигномиальному виду. Как уже указывалось, функциональная зависимость (4.75) часто дает хорошее приближение к эмпирическим данным в широком диапазоне изменения переменных x_i. Кроме того, функционалы (4.75) используются непосредственно в задачах дискретного оптимального управления.

Более детальное обсуждение соответствующих вопросов выходит за рамки данной книги и может составить предмет отдельного изложения.

4.5.12. Оптимальное управление

Основные постановки задач теории оптимального управления изучаются в базовых курсах по теории управления. Несмотря на интенсивные исследования в области создания эффективных численных методов решения задач оптимального управления, при их практической реализации возникают значительные трудности. Они обусловлены трудностями решения существенно нелинейных краевых задач, получаемых из принципа максимума Л. С. Понтрягина, а также чрезмер-

ной громоздкостью соответствующих методу динамического программирования Р. Беллмана численных процедур. В силу указанных причин интенсивно развиваются методы решения задач оптимального управления, базирующиеся на идеях конечномерной оптимизации. Несмотря на свои трудности указанный подход оказался чрезвычайно эффективным, позволяя использовать всесторонне развитый арсенал методов решения канонических задач безусловной минимизации. Для реализации соответствующих методов разработаны рекуррентные процедуры вычисления как первых, так и вторых производных от характерных для теории оптимального управления целевых функционалов вида

$$R(x,\omega) = b(z_q) + \sum_{i=1}^{q-1} h_i B(z_i), \tag{4.79}$$

где $\omega = [u^1, u^2, j, u^q] \in R^{rq}$ — полный вектор управлений; $x = x(\omega) = [x^1, x^2, \ldots, x^q] \in R^q$ — полный фазовый вектор $z_i \triangleq [x_i, u_i, t_i]$; $i \in [1:q-1]$. Здесь предполагается, что управляемый процесс описывается неавтономной системой дифференциальных уравнений вида

$$\frac{dx}{dt} = f[x(t), u(t), t],\ t \in [0, T],\ x(0) = x_1;$$

$$x(t) \in R^n,\ u(t) \in R^r;\ x_i = x(t_i),\ u_i = u(t_i);\ 0 = t_1 < t_2 < \ldots < t_q = T. \tag{4.80}$$

К виду (4.79), в частности, приводят такие методы учета ограничений, как методы штрафных функций и модифицированных функций Лагранжа.

Эффективные вычислительные процедуры для построения первых и вторых производных функционалов (4.79) описаны в работе [26]. Это позволяет непосредственно обращаться к эффективным методам оптимизации второго порядка с целью решения задач теории оптимального управления методами конечномерной оптимизации. При этом представляет несомненный практический интерес решение следующих вопросов.

1. Обобщение формул численного дифференцирования на случаи применения различных схем численного интегрирования.
2. Уточнение структуры матрицы $R''_{\omega\omega}$ с целью организации упакованной формы ее хранения в памяти компьютера при решении задач высокой размерности методами, рассматриваемыми в главе 5.
3. Организация программных интерфейсов между соответствующими модулями, реализующими конкретные выражения для производных, и библиотекой методов решения канонических оптимизационных задач.

К сожалению, рассмотрение этих вопросов выходит за рамки настоящей книги.

4.6. Основные результаты и выводы

1. Анализ явления заклинивания метода циклического покоординатного спуска (ЦПС) в условиях овражной ситуации показывает, что заклинивание вызывается не столько наличием изломов в поверхностях уровня негладких

целевых функционалов (что обычно отмечается в литературе), сколько дискретным представлением информации в компьютере.

2. Рассмотрена базовая версия метода ЦПС (алгоритм GZ1) с адаптивной настройкой шагов. Алгоритм GZ1 имеет повышенную вычислительную надежность (в частности, не содержит операций деления) и используется непосредственно в качестве стартовой поисковой процедуры, а также в составе более сложных процедур обобщенного покоординатного спуска (ОПС).
3. Теорема об устойчивости линейных оболочек изолированных групп собственных векторов матрицы Гессе целевого функционала позволяет сформулировать и доказать теоремы, выражающие принцип *частичной локальной декомпозиции* задачи оптимизации при реализации методов ОПС. В результате показано, что применение методов ОПС для решения овражных оптимизационных задач, по существу, реализует принцип «разделения движений». Исходная плохо обусловленная задача локально аппроксимируется несколькими хорошо обусловленными задачами, что резко повышает эффективность процедуры оптимизации в целом.
   Теорема о сходимости методов ОПС для достаточно широкого класса невыпуклых функционалов подтверждает тезис о высокой степени универсальности соответствующих алгоритмов.
4. Описаны методы масштабирования и адаптивной нормализации основных переменных оптимизационной задачи с учетом конечного значения машинного эпсилон. Применение указанных методов позволяет в среднем в 1,5–2 раза сократить вычислительные затраты и предотвратить появление возможных сбойных ситуаций.
5. Изложены общие реализации алгоритмов ОПС (алгоритмы SPAC1, SPAC2) на основе конечноразностных двусторонних аппроксимаций производных с адаптивной настройкой шагов дискретности.
6. Изучены общие реализации алгоритмов ОПС на основе рекуррентных алгоритмов оценивания параметров линейных регрессионных моделей. В качестве базовых методов используются рекуррентный метод наименьших квадратов и модифицированный алгоритм Качмажа. Применение указанных процедур исключает необходимость прямого вычисления производных при реализации методов ОПС и представляется весьма перспективным для снижения общей трудоемкости решения задачи. Рассмотренные версии алгоритмов позволяют по сравнению с общими процедурами типа SPAC1–2 в среднем сократить время минимизации приблизительно в $0{,}8n$ раз, где n — размерность пространства поиска. Однако надежность и универсальность алгоритмов SPAC1–2 остаются более высокими.
7. Изучены специальные схемы реализация методой ОПС для следующих классов задач теории управления:
   - идентификации нелинейных детерминированных объектов на основе функциональных рядов Вольтерра;
   - идентификации стохастических объектов на основе корреляционных методов;
   - синтеза статистически оптимальных систем автоматического управления;

- идентификации нелинейных динамических систем;
- оценки состояний динамических систем (задача о наблюдении);
- идентификации возмущающих сил;
- управления технологическим процессом серийного выпуска изделий по критерию вероятности выхода годных;
- обеспечения максимального запаса работоспособности системы по заданному списку выходных параметров;
- задач оптимального управления.

Указанные специальные реализации методов ОПС позвóляют за счет учета конкретных структурных особенностей некоторых из вариантов критериев качества, применяемых в перечисленных задачах, существенно (приблизительно в n раз) сократить трудоемкость построения матриц Гессе целевых функционалов.

Приведенные результаты по применению процедур ОПС позволяют утверждать, что по сравнению с традиционными методами нелинейной оптимизации методы ОПС дают существенное расширение класса эффективно решаемых прикладных задач. При этом наблюдается как качественный эффект (решение задач, не решаемых известными методами), так и количественный — существенно меньшие временные затраты на решение задачи по сравнению с показателями работы стандартных оптимизирующих процедур.

Глава 5. Градиентные стратегии параметрической оптимизации

В данной главе строятся оптимизирующие процедуры второго порядка ньютоновского типа с учетом основных особенностей оптимизационных задач теории управления и требований к соответствующим методам и алгоритмам (см. 2.7, 3.5.6). Представленные в книге матричные градиентные процедуры, в отличие от классического подхода, не используют в своих вычислительных схемах результаты решения плохо обусловленных систем линейных алгебраических уравнений. Вместо этого применяются различные рекуррентные процедуры с непрерывным контролем точности, исключающим накопление вычислительных погрешностей до неприемлемого уровня. По сравнению с предыдущей главой, далее вводится дополнительное предположение об *информативности* матриц Гессе минимизируемых функционалов в смысле, указанном в 3.5.6.

С учетом представленных в главе 3 моделей явления овражности эффективность алгоритмов оценивается по свойствам соответствующих *функций релаксации*, полностью определяющих локальные характеристики рассматриваемых методов.

5.1. Общая схема градиентных методов. Понятие функции релаксации

Пусть решается задача

$$J(x) \to \min_x, \; x \in R^n, \; J \in C^2(R^n). \tag{5.1}$$

Рассмотрим класс матричных градиентных методов вида

$$x^{k+1} = x^k - H_k(G_k, h_k)\, g(x^k), \; h_k \in R^1, \tag{5.2}$$

где $G_k \triangleq J''(x^k)$; $g(x^k) \triangleq J'(x^k)$, H_k — матричная функция от G_k. Предполагается, что в некоторой ζ_k-окрестности $\{x \in R^n \mid \|x - x^k\| \le \zeta_k\}$ точки x^k функционал $J(x)$ достаточно точно аппроксимируется параболоидом

$$f(x)=\frac{1}{2}\langle G_k x,x\rangle-\langle b_k,x\rangle+c_k, \tag{5.3}$$

где G_k — симметричная, не обязательно положительно определенная матрица. Без существенного ограничения общности можно считать, что $b_k = 0$, $c_k = 0$. Действительно, полагая $\det G_k \neq 0$, $x = x^* + z$, где $x^* = G_k^{-1} b_k$, получим представление

$$f_1(z)=f(x^*+z)=\frac{1}{2}\langle G_k z,z\rangle+\bar{c}_k. \tag{5.4}$$

При этом константа $\bar{c}_k = c_k - 0{,}5\langle G_k x^*, x^*\rangle$ может не учитываться как не влияющая на процесс оптимизации.

Формула (5.2) обладает свойством инвариантности относительно смещения начала координат: будучи записанной для $f(x)$, она преобразуется в аналогичное соотношение для $f_1(z)$. А именно, имеем для $f(x)$

$$x^{k+1} = x^k - H_k(G_k x^k - b_k). \tag{5.5}$$

Полагая $z^k = x^k - x^*$, получим из (5.5): $z^{k+1} = z^k - H_k G_k z^k$. А это есть запись метода (5.2) для функционала f_1.

Ставится задача построения таких матричных функций H_k, чтобы выполнялись условия релаксационности процесса $f(x^{k+1}) < f(x^k)$ и чтобы при этом величина нормы $\| x^{k+1} - x^k \|$ ограничивалась сверху только параметром ζ_k, характеризующим область справедливости локальной квадратичной модели (5.3). При высокой степени овражности $\eta(x^k)$ для большинства классических схем поиска имеем $\| x^{k+1} - x^k \| << \zeta_k$, что в результате приводит к медленной сходимости.

Определение 5.1. Скалярная функция $R_h(\lambda) = 1 - H(\lambda, h)\lambda$; $\lambda, h \in R^1$ называется *функцией релаксации* метода (5.2), а ее значения $R_h(\lambda_i)$ на спектре матрицы G_k — *множителями релаксации* для точки x^k [72].

В ряде случаев для сокращения записи индекс «h» в выражении для функции релаксации будет опускаться.

Здесь $H(\lambda, h)$ означает скалярную зависимость, отвечающую матричной функции $H(G, h)$ в представлении (5.2).

Если G — симметричная матрица и

$$G = T\,\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)\,T^{\mathrm{T}},$$

где T — ортогональная матрица, столбцы которой есть собственные векторы матрицы G, то любая матричная функция $F(G)$ представима в виде

$$F(G) = T\,\mathrm{diag}(F(\lambda_1), F(\lambda_2), \ldots, F(\lambda_n))T^{\mathrm{T}}.$$

Матричная функция $F(G)$ имеет смысл, если скалярная функция $F(\lambda)$ определена в точках $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Теорема 5.1. Для выполнения условия

$$f(x^{k+1}) \le f(x^k) \tag{5.6}$$

при $\forall x^k \in R^n$ необходимо и достаточно, чтобы

$$|R(\lambda_i)| \geq 1,\ \lambda_i < 0;\ |R(\lambda_i)| \leq 1,\ \lambda_i > 0 \qquad (5.7)$$

для всех собственных чисел λ_i, $i \in [1: n]$, матрицы G_k.

Доказательство. Пусть $\{u^i\}$ — ортонормированный базис, составленный из собственных векторов матрицы G_k. Тогда, разлагая x^k по векторам базиса, получим

$$x^k = \sum_{i=1}^{n} \xi_{i,k} u^i,\ x^{k+1} = x^k - H_k(G_k, h_k) f'(x^k) = (E - H_k G_k) x^k =$$

$$= \sum_{i=1}^{n} \xi_{i,k} [1 - H_k(\lambda_i, h_k)\lambda_i] u^i = \sum_{i=1}^{n} \xi_{i,k} R(\lambda_i) u^i.$$

С другой стороны, имеем $x^{k+1} = \sum_{i=1}^{n} \xi_{i,k+1} u^i$. Поэтому $\xi_{i,k+1} = \xi_{i,k} R(\lambda_i)$. Из сравнения выражений

$$f(x^k) = 0{,}5 \sum_{i=1}^{n} \xi_{i,k}^2 \lambda_i;\ f(x^{k+1}) = 0{,}5 \sum_{i=1}^{n} \xi_{i,k+1}^2 \lambda_i = 0{,}5 \sum_{i=1}^{n} \xi_{i,k}^2 \lambda_i R^2(\lambda_i) \qquad (5.8)$$

следует, что при выполнении (5.7) каждое слагаемое суммы в представлении $f(x^k)$ не возрастает. Достаточность доказана. Докажем необходимость. Рассмотрим первое условие (5.7). Пусть существует такой $i = i_0$, для которого $|R(\lambda_{i_0})| < 1, \lambda_{i_0} < 0$. Выберем $x^k = u^{i_0}$. Тогда $f(x^k) = 0{,}5\lambda_{i_0} < f(x^{k+1}) = 0{,}5\lambda_{i_0} R^2$, что противоречит условию релаксационности (5.6). Аналогично рассматривается второе неравенство (5.7). Теорема доказана.

Замечание 1. Для строгого выполнения неравенства (5.6) необходимо и достаточно кроме (5.7) потребовать, чтобы существовал такой $i = i_0$, для которого $\xi_{i_0,k} \neq 0$ и соответствующее неравенство (5.7) было строгим.

Замечание 2. Выражения (5.8) позволяют оценить скорость убывания функционала f в зависимости от «запаса», с которым выполняются неравенства (5.7). Действительно, обозначим через λ_i^+, λ_i^- положительные и отрицательные собственные числа матрицы G_k. Этими же индексами снабдим соответствующие собственные векторы. Суммирование по соответствующим i будем обозначать Σ^+, Σ^-. Тогда

$$2|f(x^k) - f(x^{k+1})| = \Sigma^+ \xi_{i,k}^2 \lambda_i^+ [1 - R^2(\lambda_i^+)] + \Sigma^- \xi_{i,k}^2 |\lambda_i^-| [R^2(\lambda_i^-) - 1].$$

Из полученного выражения следует, что наибольшее подавление будут испытывать слагаемые, для которых значение множителя релаксации наиболее существенно отличается от единицы (при выполнении условий (5.7)).

Далее будут рассматриваться в основном зависимости $R_h(\lambda)$, обладающие свойством

$$R_h(\lambda) \to 1,\ h \to 0. \qquad (5.9)$$

В этом случае из равенства

$$\left\| x^{k+1} - x^k \right\| = \sum_{i=1}^{n} \xi_{i,k}^2 \left[R_{h_k}(\lambda_i) - 1 \right]^2 \qquad (5.10)$$

следует, что для $\forall \zeta_k \in R^1$ всегда можно выбрать такой h_k, что $\| x^{k+1} - x^k \| \le \zeta_k$. Таким образом, с помощью параметра h_k можно регулировать норму вектора продвижения в пространстве управляемых параметров с целью предотвращения выхода из области справедливости локальной квадратичной модели (5.3).

Иногда для ограничения нормы (5.10) параметр h может вводиться в схему оптимизации как множитель в правой части (5.2):

$$x^{k+1}(h) = x^k - hH_k g(x^k),\ h \in [0, 1]. \qquad (5.11)$$

При этом $\| x^{k+1}(h) - x^k \| = h \| x^{k+1} - x^k \|$, а второе равенство (5.8) трансформируется к виду

$$f(x^{k+1}) = 0{,}5 \sum_{i=1}^{n} \xi_{i,k}^2 \lambda_i \overline{R}^2(\lambda_i),$$

где $\overline{R}(\lambda_i) = (1-h) + hR(\lambda_i)$. Таким образом, новые множители релаксации $\overline{R}(\lambda_i)$ принимают промежуточные значения между 1 и $R(\lambda_i)$, что и требуется для обеспечения свойства релаксационности, определяемого требованиями (5.9).

Введенное понятие функции релаксации позволяет с единых позиций оценить локальные свойства различных градиентных схем поиска. Удобство такого подхода заключается также в возможности использования наглядных геометрических представлений. Подобно областям устойчивости методов численного интегрирования обыкновенных дифференциальных уравнений, построенных на основе «тестового» (линейного, скалярного) уравнения, мы можем для любого метода (5.2) построить функцию релаксации, характеризующую область его релаксационности в множестве собственных чисел матрицы Гессе аппроксимирующего параболоида. При этом роль тестового функционала играет квадратичная зависимость (5.3). Требуемый характер функции релаксации представлен на рис. 5.1; заштрихована «запрещенная» область, где условия релаксационности (5.7) не выполняются.

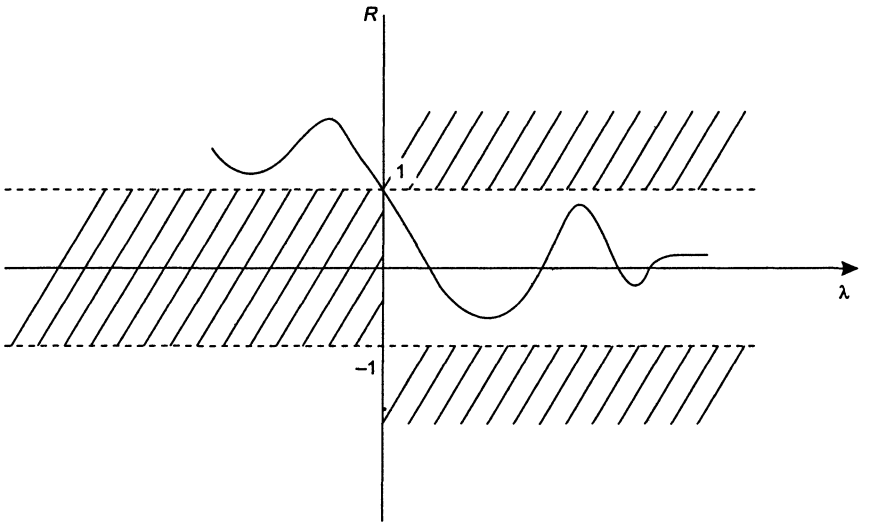

Рис. 5.1. Общий вид функции релаксации

Важное свойство функций релаксации заключается в возможности использования соответствующих представлений для синтеза новых процедур (5.2), обладающих некоторыми желательными свойствами при решении конкретных классов задач параметрической оптимизации.

5.2. Классические градиентные схемы

Рассмотрим некоторые конкретные методы (5.2) и отвечающие им функции релаксации.

5.2.1. Простой градиентный спуск (ПГС)

Формула метода ПГС имеет вид

$$x^{k+1} = x^k - hg(x^k),\ h \in R^1. \tag{5.12}$$

Соответствующая функция релаксации линейна (рис. 5.2):

$$R(\lambda) = 1 - \lambda h. \tag{5.13}$$

Пусть собственные значения матрицы G_k расположены в замкнутом интервале $[m, M]$, причем $M >> m > 0$, так что $\eta = \dfrac{M}{m} >> 1$. В этом случае условие (5.9), очевидно, выполняется, а неравенства (5.7) сводятся к требованию $|R(\lambda_i)| \le 1$, $i \in [1: n]$, или

$$|1 - h\lambda_i| \le 1,\ i \in [1: n]. \tag{5.14}$$

Эти оценки иллюстрируются рис. 5.2. Точка пересечения прямой $R(\lambda)$ с осью абсцисс есть точка $\lambda = \dfrac{1}{h}$, и чтобы при $\lambda \in [m, M]$ зависимость $R(\lambda)$ находилась в разрешенной области, необходимо выполнение неравенства $\dfrac{1}{h} \ge \dfrac{M}{2}$ или $h \le \dfrac{2}{M}$. При этом ординаты функции релаксации характеризуют величину соответствующих множителей релаксации, которые в окрестности $\lambda = m$ будут тем ближе к единице, чем больше величина $\dfrac{M}{m}$. Будем считать, что для собственных чисел матрицы G_k выполняются неравенства

$$\lambda_1 \ge \ldots \ge \lambda_{n-r} \ge \sigma|\lambda_{n-r+1}| \ge \ldots \ge \sigma|\lambda_n|,\ \sigma >> 1,$$

характерные для овражной ситуации. Тогда для точки $x^k \in Q$, где Q — дно оврага, имеем, согласно (5.10),

$$\left\| x^{k+1} - x^k \right\| \cong 4 \sum_{i=n-r+1}^{n} \xi_{i,k}^2 \left(\frac{\lambda_i}{\lambda_1} \right)^2 \le 4\sigma^{-2} \sum_{i=n-r+1}^{n} \xi_{i,k}^2,$$

что может быть существенно меньше ζ_k.

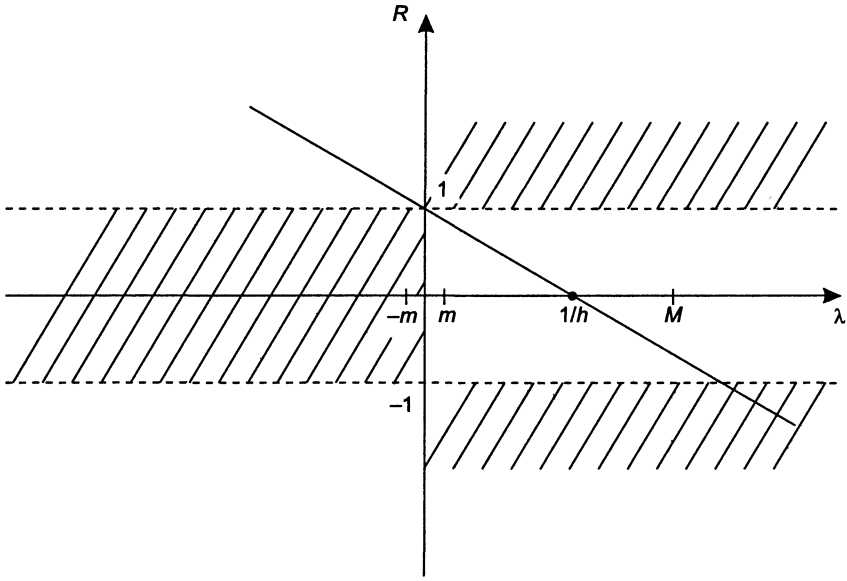

Рис. 5.2. Функция релаксации метода простого градиентного спуска

В результате соответствующие малым собственным значениям из окрестности $\lambda = 0$ слагаемые в (5.8) практически не будут убывать, а продвижение будет сильно замедленным. Это и определяет низкую эффективность метода (5.12).

В области $\lambda < 0$ функция (5.13) удовлетворяет условиям релаксации при любом значении параметра h. Практически параметр h в методе ПГС выбирается из условия монотонного убывания функционала на каждом шаге итерационного процесса. При отсутствии убывания величина h уменьшается до восстановления релаксационности процесса. Существуют различные стратегии выбора h, однако при больших η все эти методы, включая и метод наискорейшего спуска

$$x^{k+1} \in \operatorname{Arg}\min_{h\geq 0} J[x^k - hg(x^k)],$$

ведут себя приблизительно одинаково плохо даже при минимизации сильно выпуклых функционалов. Так же, как и в методе покоординатного спуска, здесь возможна ситуация заклинивания, представленная на рис. 5.3.

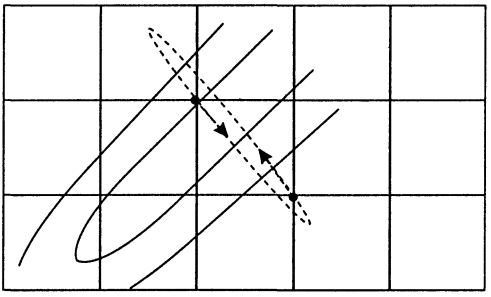

Рис. 5.3. Остановка метода ПГС в овражной ситуации

Как указывалось в 3.3, метод ПГС представляет определенный интерес как средство оценки локальной степени овражности в окрестности точки замедления алгоритма. Выведем соответствующие соотношения.

Пусть замедление метода ПГС при минимизации некоторого функционала $J(x)$ произошло в окрестности некоторой точки x^0. Тогда можно предположить, что достаточно длинный отрезок последовательности $\{x^k\}$, построенной из точки x^0, будет оставаться в области $\|x - x^0\| \le \zeta_0$, и для $J(x)$ справедлива квадратичная аппроксимация

$$f(x) = 0{,}5\langle Gx, x\rangle - \langle b, x\rangle,\ \lambda(G) \in [-m, M]. \tag{5.15}$$

Метод (5.12) для (5.15) примет вид

$$x^{k+1} = x^k - h(Gx^k - b) = Bx^k + d, \tag{5.16}$$

где $B \triangleq E - hG$, $d \triangleq hb$.

Записывая (5.16) для двух последовательных номеров k и вычитая полученные равенства, приходим к соотношению

$$y^k \triangleq x^{k+1} - x^k = B(x^k - x^{k-1}) = B^k(x^1 - x^0) = B^k y^0. \tag{5.17}$$

Согласно хорошо известному из курса численного анализа степенному методу определения максимального собственного числа симметричной матрицы, в результате проведения процесса (5.17) может быть получена оценка максимального собственного числа матрицы B:

$$\frac{\|y^{k+1}\|}{\|y^k\|} \to \max_i |\lambda_i(B)|,\ k \to \infty. \tag{5.18}$$

Полагая, что шаг h в итерационном процессе (5.16) выбирается из условия релаксационности (5.7), можно заключить, что $h \le \frac{2}{M}$. Построив график функции релаксации $1 - h\lambda$ как функции от λ при фиксированном h, легко установить, что при $h \le \frac{2}{M}$ будем иметь

$$|1 - hM| \le 1;\ 1 + mh > 1.$$

Таким образом,

$$\max_i |\lambda_i(B)| = \max_i |1 - h\lambda_i(G)| = 1 + mh.$$

Следовательно, для достаточно больших k

$$\frac{\|y^{k+1}\|}{\|y^k\|} = \frac{\|g(x^{k+1})\|}{\|g(x^k)\|} = \mu_k \cong 1 + mh > 1. \tag{5.19}$$

В результате приходим к требуемой оценке степени овражности функционала $J(x)$ в окрестности точки x^0:

$$\eta(x^0) = \frac{M}{m} \cong \frac{2}{(\mu_k - 1)}.$$

Практически поступают следующим образом. Проводят релаксационный процесс (5.16) до тех пор, пока он не замедлится и отношение $\frac{\|g^{k+1}\|}{\|g^k\|}$ не стабилизируется около некоторого значения μ_k. Далее используют полученную оценку степени овражности для выбора последующей стратегии оптимизации.

Рассуждая аналогично, для случая $\lambda \in [m, M]$ из того же графика получим вместо (5.19) равенство $\mu_k = 1 - mh < 1$ и соответствующую оценку $\eta(x^0) = \frac{2}{(1-\mu_k)}$.

Легко геометрически видеть, что в этом случае для более устойчивого выделения в качестве максимального по модулю собственного числа матрицы B значения $1 - mh$ достаточно проводить процесс (5.16) с уменьшенным шагом h. Например, на практике удобно в любом случае осуществлять «лишнее» деление шага h на два после получения устойчивой релаксации.

Общая оценка может быть записана в виде

$$\eta(x^0) \cong \frac{2}{|1-\mu_k|},$$

причем, сравнивая μ_k с единицей, можно установить характер выпуклости $J(x)$ в окрестности точки x^0, что дает дополнительную полезную информацию.

5.2.2. Метод Ньютона

Схема Н-метода основана на формуле (3.29):

$$x^{k+1} = x^k - h_k[G_k]^{-1}g^k, \tag{5.20}$$

где $0 \le h_k \le 1$.

Соответствующая методу (5.20) функция релаксации имеет вид

$$R(\lambda) = 1 - H(\lambda, h_k)\lambda = 1 - h_k, \lambda \ne 0. \tag{5.21}$$

При $h_k = 1$, что соответствует классическому варианту метода Ньютона без регулировки шага, имеем $R(\lambda) \equiv 0$ при $\forall\lambda \ne 0$. И аналогично, при любых значениях h_k прямая релаксация $1 - h_k$ будет параллельна оси абсцисс и захватывает запрещенную область при $\lambda < 0$. Положение ее при $h_k = 0$ соответствует остановке процесса. В указанных условиях эффективный выбор h_k оказывается затруднительным. Таким образом, неприменимость Н-метода в невыпуклой ситуации получает здесь наглядное геометрическое подтверждение.

Традиционное возражение против Н-методов, связанное с необходимостью вычисления вторых производных, с учетом материала, изложенного в 4.4, 4.5, оказывается менее существенным.

5.2.3. Метод Левенберга

Если известно, что собственные значения матрицы G_k расположены в интервале $[-m, M]$, где $M >> m$, то можно построить метод, имеющий нелинейную функцию релаксации (рис. 5.4)

$$R(\lambda) = \frac{h}{(h+\lambda)}, \; h > 0, \tag{5.22}$$

удовлетворяющую требованиям (5.7) при $\forall\lambda \in [-m, M]$, если $h > m$.

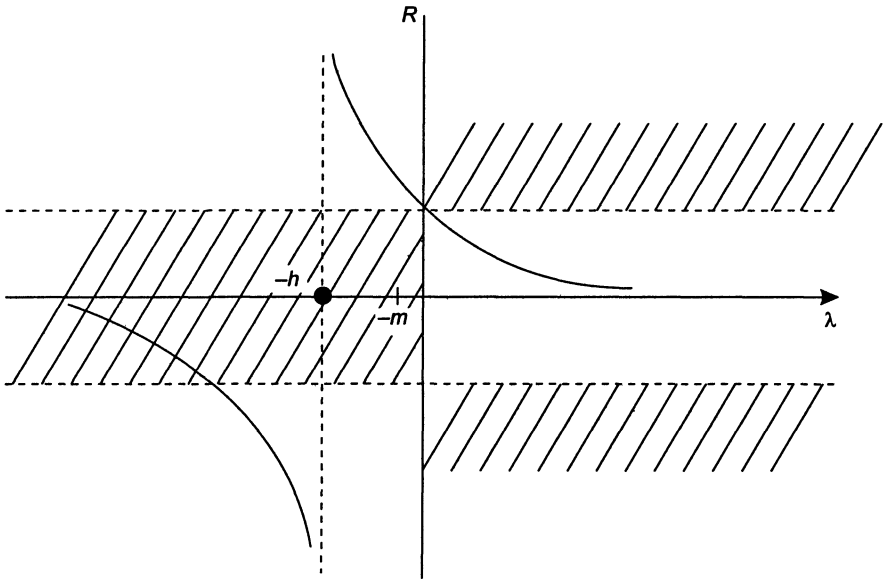

Рис. 5.4. Функция релаксации метода Левенберга

Соответствующий метод предложен Левенбергом для специальных классов оптимизационных задач метода наименьших квадратов и имеет функцию

$$H(\lambda, h) = \frac{(1-R)}{\lambda} = (h+\lambda)^{-1}. \tag{5.23}$$

Соответствующий метод называется *методом Левенберга.*

Схема метода (5.2) с функцией (5.23) имеет вид

$$x^{k+1} = x^k - [hE + G_k]^{-1} g^k \tag{5.24}$$

и реализует некоторый метод доверительной окрестности (см. 2.5.3). Данный метод часто называется также *регуляризованным методом Ньютона* или *методом*

Маркуардта—Левенберга. Скаляр h на каждом шаге итерационного процесса подбирается так, чтобы матрица $hE + G(x^k)$ была положительно определена и чтобы $\| x^{k+1} + x^k \| \le \Delta$, где величина Δ может меняться от итерации к итерации.

Реализация метода (5.23) сводится к решению на каждом шаге по k линейной алгебраической системы

$$[hE + G_k]\Delta x^k = -g^k, \ \Delta x^k \stackrel{\Delta}{=} x^{k+1} - x^k. \tag{5.25}$$

Главный недостаток метода заключается в необходимости достаточно точного подбора параметра h, что сопряжено с известными вычислительными трудностями. Значение m, как правило, неизвестно и не может быть вычислено с приемлемой точностью. Лучшее, что обычно можно сделать на практике, — это принять

$$h \ge \max\left\{\varepsilon_м n \| G_k \|, \left| \min_i \lambda_i(G_k) \right| \right\}, \tag{5.26}$$

где $\varepsilon_м$ — машинное эпсилон. При этом оценка для m существенно ухудшается при возрастании размерности n. Правая часть неравенства (5.26) обусловлена тем, что абсолютная погрешность представления любого собственного числа матрицы G_k за счет ограниченности разрядной сетки равна [72]

$$| d\lambda_i | \le n\lambda_1 \varepsilon_м \cong n \| G_k \| \varepsilon_м.$$

При невыполнении условия $h > m$ система (5.25) может оказаться вырожденной. Кроме этого слева от точки $\lambda = -m$ функция релаксации быстро входит в запрещенную область, и метод может стать расходящимся. Попытки использования алгоритмического способа более точной локализации h приводят к необходимости многократного решения плохо обусловленной линейной системы (5.25) с различными пробными значениями h.

Легко видеть, что число обусловленности матрицы $hE + G_k$ может превышать $\mathrm{cond}\,(G_k)$. Действительно, потребуем, например, чтобы $R(-m) = 10$ для обеспечения заданной скорости убывания $J(x)$. Определим необходимую величину параметра h. Имеем $R = \dfrac{h}{(h+\lambda)} = \dfrac{h}{(h-m)} = 10$ или $h = \dfrac{m}{0{,}9}$. В этом случае $\lambda_{\min}(hE + G_k) = -m + \dfrac{m}{0{,}9} = \dfrac{m}{9} > 0$. Полагая $\lambda_{\max}(hE + G_k) \cong \lambda_{\max}(G_k)$, получим, что $\mathrm{cond}(hE + G_k) \cong \dfrac{9\lambda_{\max}(G_k)}{|\lambda_{\min}(G_k)|}$.

При выборе заведомо больших значений h, что реализуется, например, когда определяющим в (5.26) является первое выражение в скобках, мы имеем $m << h$, и величина $|R(-m)| \cong 1$, что приводит к медленной сходимости. Ограничение снизу на величину h не позволяет также уменьшить до желаемой величины множители релаксации для $\lambda > 0$.

Эти трудности усугубляются при аппроксимации производных конечными разностями, так как при малых значениях g^k для точек x^k, расположенных на дне оврага, мы приходим к необходимости получать компоненты вектора градиента

как малые разности относительно больших величин порядка $J(x^k)$. В результате компоненты вектора Δx^k будут находиться с большими относительными погрешностями порядка $\dfrac{\eta(x^k)\left|\varepsilon_м J(x^k)\right|}{\left\|g^k\right\|}$. Коэффициент овражности η в данном случае играет роль коэффициента усиления погрешности. Для метода Ньютона справедливо аналогичное замечание. В то же время для метода ПГС точность задания g^k может оказаться достаточной для правильного определения направления убывания $J(x)$.

Еще раз укажем, что методы, рассмотренные в последующих разделах этой главы, также используют в своей схеме производные, которые вычисляются с теми же погрешностями. Однако их вычислительные схемы таковы, что они не используют окончательных результатов решений плохо обусловленных линейных алгебраических систем. Например, в методах с экспоненциальной релаксацией на участках выпуклости $J(x)$ решение эквивалентной (5.25) линейной системы выполняется итеративно, причем каждая итерация имеет «физический» смысл, что дает возможность непрерывно контролировать точность вычислений и прерывать процесс, когда накопленная ошибка начинает превышать допустимый уровень.

Несмотря на отмеченные недостатки, методы типа (5.25) часто оказываются достаточно эффективными, и их присутствие в библиотеке методов оптимизации следует признать весьма желательным.

Дополнительное положительное свойство соответствующих алгоритмов связано с возможностью их обобщения на решение «больших» задач параметрической оптимизации, рассмотренных в 3.5.4.

5.3. Методы с экспоненциальной релаксацией

Начало систематического изучения методов описываемого класса было положено в книге [49]. Далее рассматривается другой подход к построению и анализу алгоритмов, основанный на понятии функции релаксации.

Исходя из вышеизложенных требований к функциям релаксации, естественно рассмотреть экспоненциальную зависимость вида (рис. 5.5)

$$R(\lambda) = \exp(-\lambda h),\ h > 0, \tag{5.27}$$

для которой условие (5.7) выполняется при любых значениях параметра h. Кроме того реализуется предельное соотношение (5.9), что позволяет эффективно регулировать норму вектора продвижения независимо от расположения спектральных составляющих матрицы G_k на вещественной оси λ.

Функция (5.27) обобщает ранее рассмотренные функции релаксации и является в определенном смысле оптимальной. Действительно, разлагая экспоненту (5.27) в ряд Тейлора и ограничиваясь двумя первыми членами разложениями, получим

$$\exp(-\lambda h) = \frac{1}{\exp(\lambda h)} \cong \frac{1}{1+\lambda h} = \frac{h'}{(h'+\lambda)}, \quad h' = \frac{1}{h},$$

что совпадает с (5.22) при $h = h'$. И аналогично, полагая $\exp(-\lambda h) = 1 - \lambda h$, приходим к зависимости (5.13). Для достаточно больших значений параметра h имеем $\exp(-\lambda h) \cong 0$ при любых $\lambda \geq m > 0$, что позволяет говорить о вырождении метода в классический метод Ньютона без регулировки шага.

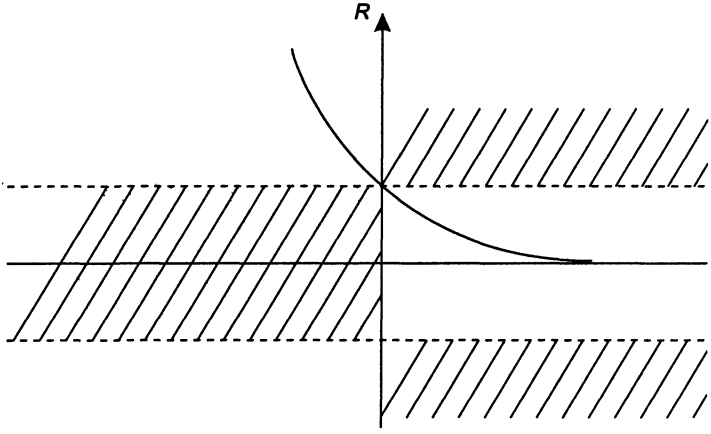

Рис. 5.5. Экспоненциальная функция релаксации

Для представления схемы метода в виде (5.2) необходимо определить соответствующую функцию $H(\lambda, h)$. Имеем

$$\lambda H(\lambda, h) = 1 - R(\lambda) = 1 - \exp(-\lambda h).$$

Полагая $\lambda \neq 0$, получим

$$H(\lambda, h) = \lambda^{-1}\left[1 - \exp(-\lambda h)\right] = \int_0^h \exp(-\lambda\tau)\, d\tau. \tag{5.28}$$

Доопределяя $H(0, h)$ из условия непрерывности, получим $H(0, h) = h$. В результате схема *метода с экспоненциальной релаксацией* (*ЭР-метода*) примет вид

$$x^{k+1} = x^k - H(G_k, h_k)\, g(x^k); \tag{5.29}$$

$$H(G, h) = \int_0^h \exp(-G\tau)\, d\tau. \tag{5.30}$$

Параметр h_k определяется равенством

$$h_k \in \operatorname{Arg}\min_{h \geq 0} J\left[x^k - H(G_k, h) g\left(x^k\right)\right], \tag{5.31}$$

однако возможны и другие способы выбора h_k.

Принципиальная схема ЭР-метода была получена исходя из анализа локальной квадратичной модели минимизируемого функционала. Представляет интерес

выяснение возможностей метода в глобальном смысле, без учета предположений о квадратичной структуре $J(x)$.

Можно доказать, что алгоритм (5.29), (5.30) сходится практически при тех же ограничениях на минимизируемый функционал, что и метод наискорейшего спуска [29], имея в определенных условиях существенно более высокую скорость сходимости.

Следующая теорема устанавливает факт сходимости ЭР-метода для достаточно широкого класса невыпуклых функционалов в предположении достижимости точки минимума (условие 2) и отсутствия точек локальных минимумов (условие 3) [49].

Теорема 5.2. Пусть

1) $J(x) \in C^2(R^n)$;
2) множество $X_* = \{x^* | J(x^*) = \min J(x)\}$ непусто;
3) для любого $\varepsilon > 0$ найдется такое $\delta > 0$, что $\|g(x)\| \geq \delta$, если $x \notin S(X_*)$, где

$$S(X_*) = \left\{x \middle| d(x, X_*) \leq \varepsilon\right\},\ d(x, X_*) = \min_{x^* \in X_*} \|x - x^*\|;$$

4) для любых $x, y \in R^n$

$$\|g(x+y) - g(x)\| \leq l\|y\|,\ l > 0;$$

5) собственные числа матрицы $G(x)$ заключены в интервале $[-M, M]$, где $M > 0$ не зависит от x.

Тогда независимо от выбора начальной точки x^0 для последовательности $\{x^k\}$, построенной согласно (5.29), (5.30), выполняются предельные соотношения

$$\lim d(x^k, X_*) = 0; \tag{5.32}$$

$$\lim J(x^k) = J(x^*),\ k \to \infty. \tag{5.33}$$

Доказательство. Используя соотношения

$$J(x+y) = J(x) + \int_0^1 \langle g(x+\theta y), y\rangle d\theta,$$

$$\left|\int_0^1 \langle x(\theta), y(\theta)\rangle d\theta\right| \leq \int_0^1 \|x(\theta)\| \cdot \|y(\theta)\| d\theta$$

и обозначая $J_k = J(x^k)$, $g_k = J'(x^k)$, $G_k = J''(x^k)$, получим

$$\begin{aligned} J_k - J\left[x^k - H(G_k, h)g^k\right] &= \int_0^1 \left\langle g\left[x^k - \theta H(G_k, h)g^k\right], H(G_k, h)g^k\right\rangle d\theta = \\ &= \left\langle H(G_k, h)g^k, g^k\right\rangle - \int_0^1 \left\langle g^k - J\left[x^k - \theta H(G_k, h)g^k\right], H(G_k, h)g^k\right\rangle d\theta \geq \\ &\geq \left\langle H(G_k, h)g^k, g^k\right\rangle - l\left\|H(G_k, h)g^k\right\|^2 \int_0^1 \theta d\theta = \\ &= \left\langle H(G_k, h)g^k, g^k\right\rangle - \frac{l}{2}\left\|H(G_k, h)g^k\right\|^2 \geq \rho\left\|g^k\right\|^2 - \frac{l}{2}\left\|g^k\right\|^2 R^2 = \alpha\left\|g^k\right\|^2; \end{aligned} \tag{5.34}$$

$$\alpha \triangleq \rho - \left(\frac{l}{2}\right)R^2.$$

При этом использованы неравенства

$$\rho \| y \|^2 \le \langle H(G_k, h)y, y\rangle \le R \| y \|^2, \tag{5.35}$$

где ρ, R — соответственно, минимальное и максимальное собственные числа положительно определенной матрицы $H(G_k, h)$.

Левое неравенство (5.35) следует из представления минимального собственного числа λ любой симметричной матрицы B в виде $\lambda = \min\limits_{x\ne 0}\dfrac{\langle Bx, x\rangle}{\langle x, x\rangle}$, а правое — из условия согласования $\| Bx \| \le \| B \| \cdot \| x \|$ сферической нормы вектора $\| x \| = \sqrt{\langle x, x\rangle}$ и спектральной нормы симметричной матрицы $\| B \| = \max\limits_i |\lambda_i(B)|$, где $\lambda_i(B)$, $i \in [1: n]$ — собственные числа матрицы B. Для значений ρ и R получим:

$$\rho = \min_i \lambda_i\left[H(G,h)\right] = \min_i \int_0^h \exp\left[-\lambda_i(G)\tau\right] d\tau,\ R = \max_i \int_0^h \exp\left[-\lambda_i(G)\tau\right] d\tau.$$

Согласно пятому предположению, имеем

$$\int_0^h \exp(-M\tau)\, d\tau \le \int_0^h \exp\left[-\lambda_i(G)\tau\right] d\tau \le \int_0^h \exp(M\tau)\, d\tau,$$

поэтому

$$\rho \ge \int_0^h \exp(-M\tau)\, d\tau = M^{-1}\left[1 - \exp(-Mh)\right];$$

$$R \le \int_0^h \exp(M\tau)\, d\tau = M^{-1}\left[\exp(Mh) - 1\right];$$

$$\alpha = \rho - \frac{l}{2}R^2 \ge M^{-1}\left[1 - \exp(-Mh)\right] - \frac{l}{2M^2}\left[\exp(Mh) - 1\right]^2.$$

Полагая $h = \dfrac{1}{M}$ и считая без ограничения общности, что $M > \dfrac{le(e-1)}{2}$, где e — основание натуральных логарифмов, получим

$$\alpha \ge \frac{(e-1)[2M - le(e-1)]}{2M^2 e} > 0. \tag{5.36}$$

Из (5.31), (5.34), (5.36) следует

$$J_k - J_{k+1} \ge J_k - J\left[x^k - H(G_k, M^{-1})\, g^k\right] \ge \alpha \| g^k \|^2. \tag{5.37}$$

Следовательно, последовательность $\{J_k\}$ является монотонно невозрастающей и ограниченной снизу величиной $J(x^*)$, поэтому она имеет предел и $J_{k+1} - J_k \to 0$ при $k \to \infty$. Из (5.37) следует $\| g^k \|^2 \le \alpha^{-1}(J_k - J_{k+1})$, поэтому $\| g^k \| \to 0$ при $k \to \infty$.

А так как по условию $\|g^k\| \geq \delta$ при $x^k \in S(X_*)$, то найдется такой номер N, что $x^k \in S(X_*)$ при $k \geq N$, и, следовательно, справедливо утверждение (5.32).

Обозначим через $\bar{x}^k$ проекцию x^k на множество X_*. Тогда по теореме о среднем $J_k - J(\bar{x}^k) = \langle g(x^{kc}), x^k - \bar{x}^k \rangle$, где $x^{kc} = \bar{x}^k + \lambda_k(x^k - \bar{x}^k)$, $\lambda_k \in [0,1]$ Учитывая, что $g(\bar{x}^k) = 0$, получим

$$J_k - J(\bar{x}^k) = \langle g(x^{kc}) - g(\bar{x}^k), x^k - \bar{x}^k \rangle \leq$$
$$\leq \|g(x^{kc}) - g(\bar{x}^k)\| \cdot \|x^k - \bar{x}^k\| \leq ld^2(x^k, X_*).$$

И в силу (5.32) получаем (5.33). Теорема доказана.

Замечание 1. Утверждения теоремы, очевидно, выполняются, если h_k выбирать не из условия (5.31), а из условия $J[x^k - H(G_k, h_k)g^k] = \min\limits_{h \in [0, \bar{h}]} J[x^k - H(G_k, h)g^k]$, где $\bar{h} > 0$ — произвольное число. Действительно, легко видеть, что равенство (5.37) только усилится, если брать любое другое значение h_1 (может быть, даже большее, чем $\dfrac{2}{le(e-1)}$ с меньшим значением функционала, чем при $h = \dfrac{1}{M}$, и в то же время, если при $h = \dfrac{1}{M}$ сходимость имеет место, то она сохраняется и при меньших значениях h. Последнее следует из возможности выбора сколь угодно больших значений M при установлении сходимости.

Замечание 2. Утверждения (5.32), (5.33) сохраняются также при замене условия (5.31) на следующее:

$$J_{k+1} = J[x^k - H(G_k, h_k)g^k] \leq (1 - \gamma_k)g^k + \gamma_k \min_{h>0} J[x^k - H(G_k, h)g^k], \quad (5.38)$$
$$0 < \gamma < \gamma_k \leq 1.$$

Действительно, из (5.38) будем иметь

$$J_k - J_{k+1} \geq \gamma_k(J_k - \min_{h>0} J[x^k - H(G_k, h)g^k]) \geq \gamma_k(J_k - J[x^k - H(G_k, h)g^k])$$

и, согласно (5.34), $J_k - J_{k+1} \geq \gamma_k \alpha \|g^k\|^2 = \bar{\alpha}\|g^k\|^2$, $\bar{\alpha} > 0$.

Получено неравенство, аналогичное (5.37), и далее доказательство проводится по той же схеме с заменой α на $\bar{\alpha}$.

В случае сильной выпуклости функционала $J(x)$ удается получить оценку скорости сходимости.

Теорема 5.3. Пусть:

1) $J(x) \in C^2(R^n)$;
2) для любых $x, y \in R^n$ выполняются условия $\lambda\|y\|^2 \leq \langle G(x)y, y \rangle \leq \Lambda\|y\|^2$, $\|G(x + y) - G(y)\| \leq L\|x\|$, $\lambda > 0$, $L \geq 0$.

Тогда независимо от выбора начальной точки x^0 для метода (5.29) справедливы соотношения (5.32), (5.33), и оценка скорости сходимости

$$\left\| x^{k+1} - x^* \right\| \le \frac{\left(\frac{\Lambda}{\lambda}\right)^{\frac{1}{2}} L \left\| x^k - x^* \right\|^2}{2\lambda}.$$

Доказательство содержится в книге [49].

Таким образом, установлена квадратичная скорость сходимости, характерная для Н-методов.

5.3.1. Реализация методов с экспоненциальной релаксацией

Алгоритм вычисления матричных функций (5.30) может быть основан на использовании известного рекуррентного соотношения

$$H(G, 2h) = H(G, h)[2E - GH(G, h)]. \tag{5.39}$$

Так как все рассматриваемые матричные функции симметричны и, следовательно, обладают простой структурой, то для доказательства (5.39) достаточно проверить его для соответствующих скалярных зависимостей, что тривиально.

Формула (5.39) используется также для получения обратной матрицы G^{-1}, так как выполняется предельное соотношение:

$$\lim H(G, h) = G^{-1}, \; h \to \infty.$$

Этот факт еще раз указывает на связь ЭР-метода с методом Ньютона, который является предельным вариантом рассматриваемого алгоритма при условии положительной определенности матрицы G. Практический выбор параметра h при известной матрице G или ее аппроксимации может осуществляться различными способами. В каждом из них приближенно реализуется соотношение (5.31). Наиболее простой прием, приводящий к так называемым *системным* алгоритмам оптимизации, заключается в следующем.

Задаются некоторой малой величиной h_0, такой, что матрица $H(G_k, h_0)$ может быть заменена отрезком соответствующего степенного ряда:

$$H(G_k, h_0) \cong h_0 \sum_{i=1}^{l} \frac{(-G_k h_0)^{i-1}}{i!}. \tag{5.40}$$

Далее последовательно наращивают h с помощью соотношения (5.39), вычисляя каждый раз значение $J[x^k - H(G_k, 2^q h_0) J'_k]$, $q = 0, 1, \ldots$. Процесс продолжается до тех пор, пока функция убывает либо достаточно быстро убывает. Точка с минимальным значением J принимается за x^{k+1}. При этом вместо точной реализации соотношения (5.31) оптимальный шаг выбирается на дискретной сетке значений $h_q = 2^q h_0$, $q = 0, 1, \ldots$. Как правило, предельное значение q не превышает 30–40. В самом деле, если функционал $J(x)$ квадратичный и $G(x) > 0$, то оптимальное значение параметра $h = +\infty$, $q = +\infty$, а $H(G_k, h) = G_k^{-1}$, и метод вырождается в классический вариант метода Ньютона без регулировки шага. Однако в действительности при использовании (5.39) для построения матрицы $H(G_k, h)$

необходимое число итераций q оказывается конечной величиной, ибо все вычисления проводятся с ограниченной точностью, и процесс автоматически останавливается при попадании результата в достаточно малую окрестность решения. Количество обращений к рекуррентному соотношению (5.39) при этом оказывается сравнительно небольшим, что подтверждается опытом практического применения (5.39) в качестве алгоритма построения обратной матрицы.

Сказанное подтверждается следующими рассуждениями для случая $G > 0$.

Соотношение (5.39) может быть преобразовано к виду

$$E - GH(G, 2h_0) = [E - GH(G, h_0)]^2.$$

Из равенств $\|E - GH(G, h_0)\| = \|\exp(-h_0 G)\| = \exp(-mh_0)$, при $h_0 = \frac{0{,}1}{M}$, где $m = \min_i \lambda_i(G)$, $M = \max_i \lambda_i(G)$, следует, что $\|E - GH(G, h_0)\| = \exp\left(\frac{-0{,}1m}{M}\right)$; рассмотренный выбор h_0 аргументируется позже. Поэтому

$$E - GH(G, 2^q h_0) = (E - GH(G, h_0))^{2^q}$$

и

$$\|E - GH(G, 2^q h_0)\| \le \|E - GH(G, h_0)\|^{2^q} = \exp\left(-0{,}1 \cdot 2^q \frac{m}{M}\right).$$

Необходимое число итераций q можно определять из условия выполнения с машинной точностью равенства $\|E - GH(G, 2^q h_0)\| = 2^{-t}$, где t — длина разрядной сетки мантиссы в представлении числа в форме с плавающей точкой, или, что то же самое, из условия $\exp\left(-0{,}1 \cdot 2^q \frac{m}{M}\right) = 2^{-t}$. Полагая, например, $t = 28$, $\eta = \frac{M}{m} = 10^8$, получим, что $q \cong \frac{\ln\left(10 \frac{t \ln 2M}{m}\right)}{\ln 2} \cong 34$. Таким образом, показано, что максимальное число итераций при реализации соотношения (5.39) зависит от степени овражности η и обычно не превышает указанных ранее значений.

В целом ряде случаев более эффективной оказывается реализация метода с элементами адаптации, в которой значение J не вычисляется для всех промежуточных значений q. Функционал вычисляется для трех значений q: $q^* - 1$, q^*, $q^* + 1$ — с последующим выбором лучшего значения. Здесь q^* — оптимальное значение q, полученное на предыдущей итерации по k. На первой итерации для определения q^* необходимо вычислить весь ряд значений J.

С целью более точной локализации минимума на каждом шаге по k могут использоваться процедуры одномерного поиска по h, например, известный из курса численного анализа *метод золотого сечения*. Для этого вначале изложенным ранее грубым способом определяется промежуток $[h_{\min}, h_{\max}]$, содержащий оптимальное в смысле (5.31) значение h_*. Далее полагаем $\varphi(h) \triangleq J[x^k - H(G_k, h)g^k]$. Тогда $h_{\min} = 2^q h_0$, $h_{\max} = 2^{q'+2} h_0$, $h_* = 2^{q'+1} h_0$, причем предполагается, что $\varphi(h_{\min}) > \varphi(h_*)$, $\varphi(h_{\max}) > \varphi(h_*)$. Фиксируя число пересчетов q', получим, что

выбирая новый параметр h_0' из промежутка $h_0' \in [h_0, 4h_0]$, где h_0 — первоначально выбранное значение, мы получим $h = 2^{q'} h_0' \in [h_{\min}, h_{\max}]$. Далее можно определить $\varphi(h) = \varphi(2^{q'} h_0') = \Psi(h_0')$, и задача сводится к стандартней задаче минимизации функции одной переменной $\Psi(h_0')$ на заданном промежутке.

Для приближенного вычисления матрицы G_k вторых производных в оптимизационных задачах теории управления могут применяться методы, изложенные в 4.4, 4.5. Рассмотрим наиболее универсальный алгоритм, основанный на конечноразностных соотношениях. В результате вычислений по формулам (4.14), (4.15) приходим к матрице $G_k = \dfrac{D_k}{\beta_k^2}$ и к вектору $g^k = \dfrac{d^k}{\beta}$, где $\beta_k = 2s_k$, s_k — шаг дискретности. Как уже говорилось, производить деление матрицы D_k на β_k^2 или вектора d^k на β_k с целью получения G_k и g^k нецелесообразно с вычислительной точки зрения. Поэтому далее принципиальная схема ЭР-метода будет преобразована к виду, удобному для непосредственного применения D_k и d^k вместо G_k и g^k.

Имеем

$$H(D_k, h) = \int_0^h \exp(-D_k \tau)\, d\tau = \beta_k^{-2} \int_0^h \exp\left(-G_k \beta_k^2 \tau\right) d\beta_k^2 \tau = \beta_k^{-2} \int_0^{\beta_k^2 h} \exp(-G_k t)\, dt$$

или

$$\beta_k^2 H\left(\beta_k^2 G_k, h\right) = H\left(G_k, h_k\right),\ h_k = \beta_k^2 h. \tag{5.41}$$

С учетом (5.41) основное соотношение (5.29) приводится к виду

$$\begin{aligned} x^{k+1} &= x^k - H(G_k, h_k) g^k = x^k - \beta_k^2 H\left(D_k, \frac{h_k}{\beta_k^2}\right) \frac{d^k}{\beta_k} = \\ &= x^k - 2s_k H\left(D_k, h\right) d^k,\ h = \frac{h_k}{4s_k^2}. \end{aligned} \tag{5.42}$$

Имеем также

$$H(D_k, 2h) = H(D_k, h)[2E - D_k H(D_k, h)].$$

Оптимальное значение h находится непосредственно из соотношения

$$J(x^{k+1}) = \min_{h>0} J[x^{k+1} - 2s_k H\left(D_k, h\right) d^k].$$

При использовании разностного уравнения (5.42) укрупненная вычислительная схема метода с экспоненциальной релаксацией может быть сведена к следующей последовательности действий.

Алгоритм RELEX.

Шаг 1. Ввести исходные данные x^0, s.

Шаг 2. Принять $x := x^0$; $J := J(x)$; $x^l := x$; $J_l := J$.

Шаг 3. Вычислить матрицу $D = \{D_{ij}\}$ и вектор $d = \{d_{ij}\}$ в точке x по формулам

$$D_{ij} = J(x + se_i + se_j) - J(x - se_i + se_j) - J(x + se_i - se_j) + J(x - se_i - se_j), \tag{5.43}$$

$$i, j \in [1: n];$$

$$d_i = J(x + se_i) - J(x - se_i),\ i \in [1: n],\ e_i = (0, \ldots, 1, \ldots, 0); \qquad (5.44)$$

принять $h_0 := \dfrac{0{,}1}{\|D\|}$.

Шаг 4. Принять $k = 0$. Вычислить матрицу $H = H(D, h_0)$

$$H = \sum_{i=1}^{7} \frac{(-D)^{i-1} h_0^i}{i!}. \qquad (5.45)$$

Шаг 5. Принять $x^t := x - 2sHd; J_t := J(x^t)$; $k := k + 1$.

Шаг 6. Если $J_t < J_l$, принять $x^l := x^t, J_l := J_t$.

Шаг 7. Если $k > 20$, перейти к шагу 8; в противном случае принять $H := H(2E - DH)$ и перейти к шагу 5;

Шаг 8. Проверить условия окончания процесса оптимизации в целом; если они выполняются, остановить работу алгоритма; в противном случае принять $x := x^l$, $J := J_l$ и перейти к шагу 3.

Заметим, что выбор на шаге 3 параметра $h_0 := \dfrac{0{,}1}{\|D\|}$ эквивалентен при $G = J''(x)$ равенству $h_0 := \dfrac{0{,}1}{\|G\|}$ в исходной схеме алгоритма. А последнее равенство, как это следует из результатов, полученных при доказательстве теоремы 5.2, гарантирует убывание функционала:

$$J[x^k - H(G_k, h_0)g^k] < J(x^k). \qquad (5.46)$$

Действительно, как было сказано ранее, для выполнения неравенства (5.46) достаточно принять $h_0 \le \dfrac{1}{M}$, где $M > \dfrac{le(e-1)}{2} \cong 2{,}3l$. Для параболоида, аппроксимирующего $J(x)$ в окрестности точки x, имеем $l = \|G\|$, где $G = J''(x)$. Поэтому можно выбрать h_0 из условия $h_0 \le \dfrac{1}{2{,}3\|G\|} \cong \dfrac{0{,}4}{\|G\|}$. Замена коэффициента 0,4 на 0,1 позволяет более точно реализовать шаг 4 алгоритма, одновременно гарантируя выполнение (5.46).

Параметр s_k может меняться в зависимости, например, от величины $\|x^k - x^{k-1}\|$ так, как это было описано в 4.4. Возможны и другие способы регулировки шага.

5.3.2. Области применения и анализ влияния погрешностей

Обратимся к анализу влияния погрешностей вычислений при реализации ЭР-методов.

Рассмотрим итерационный процесс, определяемый рекуррентным соотношением

$$x^{k+1} = x^k - H_k(G_k, h_k)\, G_k x^k = q(G_k)x^k, \tag{5.47}$$

где $q(G) = E - H(G, h)\, G$. Данный процесс является упрощенной моделью ЭР-метода, характеризуя его локальные свойства. Здесь предполагается, что ищется минимум квадратичной формы

$$f(x) = \frac{1}{2}\langle G_k x, x\rangle.$$

Оценим влияние погрешностей в представлении матрицы G_k на характеристики релаксационности последовательности $\{f(x^k)\}$.

Кроме предположения о квадратичном характере $J(x)$ в окрестности точки x^k, мы неявно ввели еще одно допущение. А именно, заменяя в (5.47) матрицу G_k на возмущенную матрицу $G + dG$ (индекс «k» у матрицы далее будем опускать), мы предполагаем, что ошибки в вычислении G и g определенным образом согласованы. В действительности эквивалентное возмущение dG у матрицы, определяющей величину градиента Gx^k, может не совпадать с возмущением матрицы G, так как g и G вычисляются раздельно. Однако с позиций последующего анализа данное отличие не является принципиальным.

Предположим, что собственные числа матрицы G разделены на две группы

$$\lambda_1 \geq \ldots \geq \lambda_{n-r} >> |\lambda_{n-r+1}| \geq \ldots \geq |\lambda_n|. \tag{5.48}$$

Возмущение dG матрицы G приводит к появлению возмущений $d\lambda_i$ для собственных чисел и возмущений du^i для отвечающих им собственных векторов. Согласно результатам, полученным в 4.2, будем считать, что вариации собственных векторов происходят в пределах линейных оболочек

$$M_1 = \sum_{i=1}^{n-r} \alpha_i u^i,\ M_2 = \sum_{j=n-r+1}^{n} \alpha_j u^j,$$

порожденных собственными векторами $\{u^i, i \in [1{:}\, n-r]\}$, $\{u^j, j \in [n-r+1, n]\}$ исходной невозмущенной матрицы G. В данном случае матрицы G и $G + dG$ одновременно не приводятся к главным осям, что вносит дополнительный элемент сложности в анализ влияния погрешностей.

Пусть

$$U^{\mathrm{T}} G U = \operatorname{diag}(\lambda_i),\ U = (u^1, u^2, \ldots, u^n); \tag{5.49}$$

$$W^{\mathrm{T}}(G + dG)W = \operatorname{diag}(\lambda_i + d\lambda_i),\ W = (w^1, w^2, \ldots, w^n).$$

Имеем теперь

$$q(G) = E - WD_1 W^{\mathrm{T}} W D_2 W^{\mathrm{T}} = E - WD_3 W^{\mathrm{T}},$$

где

$$D_1 = \operatorname{diag}\left[\int_0^h \exp\left((-\lambda_i - d\lambda_i)\,\tau\right)\,d\tau\right];$$

$$D_2 = \operatorname{diag}\,(\lambda_i + d\lambda_i);\ D_3 = D_1 D_2.$$

Таким образом, матрица $q\,(G)$ имеет собственные векторы w^i и соответствующие им собственные числа $\lambda_i\,(q) = 1 - \lambda_i\,(D_3)$.

Полагая

$$x^k = \sum_{i=1}^{n} \xi_{i,k} w^i,\ w^i = \sum_{j=1}^{n} \alpha_{ji} u^i,$$

получаем

$$f(x^k) = 0{,}5\left\langle Gx^k, x^k\right\rangle = 0{,}5\sum_{i=1}^{n} \xi_{i,k}^2 \left(\sum_{j=1}^{n} \alpha_{ji}\right)^2 \lambda_i.$$

Аналогично имеем

$$x^{k+1} = q(G)x^k = \sum_{i=1}^{n} \xi_{i,k}\lambda_i(q) w^i = \sum_{i=1}^{n} \xi_{i,k+1} w^i;$$

$$f\left(x^{k+1}\right) = 0{,}5\sum_{i=1}^{n} \xi_{i,k+1}^2 \left(\sum_{j=1}^{n} \alpha_{ji}\right)^2 \lambda_i = 0{,}5\sum_{i=1}^{n} \xi_{i,k}^2 \left(\sum_{j=1}^{n} \alpha_{ji}\right)^2 \lambda_i \lambda_i^2(q),$$

где

$$\lambda_i(q) = 1 - (\lambda_i + d\lambda_i)\int_0^{h_k} \exp((-\lambda_i - d\lambda_i)\tau)\,d\tau = \exp\left((-\lambda_i - d\lambda_i)\,h_k\right). \qquad (5.50)$$

Для выполнения неравенства $f(x^{k+1}) \le f(x^k)$, согласно теореме 5.1, должны выполняться условия релаксационности

$$|\lambda_i\,(q)| \le 1,\ \lambda_i > 0;\ |\lambda_i\,(q)| \ge 1,\ \lambda_i < 0. \qquad (5.51)$$

Теперь легко видеть, что если возмущение $d\lambda_i$ таково, что собственное число меняет знак:

$$\operatorname{sign}\,(\lambda_i) \ne \operatorname{sign}\,(\lambda_i + d\lambda_i), \qquad (5.52)$$

то условия (5.51), вообще говоря, нарушаются. Это приводит к резкому замедлению сходимости процесса оптимизации.

Пусть вариация dG матрицы G вызывается только погрешностями округления. Тогда неравенство (5.52) невозможно, если все малые собственные числа ограничены снизу величиной $n\lambda_1\varepsilon_м$. Действительно, в этом случае $\operatorname{sign}\,(\lambda_i) = \operatorname{sign}\,(\lambda_i + d\lambda_i)$, так как $|d\lambda_i| \le n\lambda_1\varepsilon_м \le |\lambda_i|$. Отсюда имеем следующее ограничение на степени овражности функционалов, эффективно минимизируемых ЭР-методами:

$$\eta(x^k) \le \frac{1}{n\varepsilon_м}. \tag{5.53}$$

Проведенный анализ показывает, что вычислительные погрешности при достаточно больших значениях η могут приводить к практически случайному характеру множителей релаксации для малых собственных чисел, что определяет резкое снижение эффективности метода. Из (5.53) следует, что трудности возрастают при увеличении размерности n решаемой задачи и уменьшении длины разрядной сетки компьютера. Вычисления с двойной точностью приводят к оценке $\eta(x^k) \le \frac{1}{n\varepsilon_м^2}$ и позволяют решать существенно более широкий класс задач.

Как показывают результаты численных экспериментов, в отличие от методов ОПС матричные градиентные схемы типа RELEX оказываются менее универсальными. Однако там, где они применимы, может быть получен заметный вычислительный эффект. Кроме того, как показано далее, на базе градиентных методов могут быть построены алгоритмы оптимального параметрического синтеза систем с большим числом управляемых параметров. Следовательно, рассматриваемые классы методов взаимно дополняют друг друга, не позволяя выделить какой-то один «наилучший» подход.

5.4. Методы многопараметрической оптимизации

Под большими системами будем понимать системы, описываемые моделями с большим числом управляемых параметров. Если степень овражности соответствующих критериев оптимальности достаточно высока, то стандартные вычислительные средства оказываются неэффективными в силу изложенных в 3.5.4 причин. Методы ОПС, а также ЭР-методы неприменимы, так как их вычислительные схемы содержат заполненные матрицы размерности $n \times n$, что при больших (порядка 1000) n определяет чрезмерные требования к объему необходимой памяти компьютера.

Наиболее часто в указанной ситуации рекомендуется применять различные нематричные формы метода сопряженных градиентов (СГ). Однако далее будет показано, что в классе матричных градиентных схем (5.2) существуют более эффективные для рассматриваемых задач алгоритмы, чем методы СГ.

Пусть оптимизируемая система может быть представлена как совокупность взаимосвязанных подсистем меньшей размерности, в этом случае требования к выходным параметрам системы могут быть сформулированы в виде следующих неравенств:

$$y_j(x^j, x^q) \le t_j, j \in [1: q-1]; y_q(x^q) \le t_q, \tag{5.54}$$

где x^j есть n_j-мерный частный вектор управляемых параметров; x^q — n_q-мерный вектор управляемых параметров, влияющий на все q выходных параметров и

осуществляющий связь отдельных подсистем оптимизируемой системы. Размерность полного вектора управляемых параметров $x = [x^1, x^2, \ldots, x^q]$ равна

$$n = \sum_{i=1}^{q} n_i. \tag{5.55}$$

Используя технику оптимизации, представленную в 2.8.2, можно привести задачу решения системы неравенств (5.54) к виду

$$J(x) = \sum_{j=1}^{q} \Psi_j (x^j, x^q) \to \min, \; x \in R^n, \tag{5.56}$$

где критерий (5.56) является сглаженным вариантом критерия минимального запаса работоспособности.

Функционалы (5.56) возникают и при других постановках задач оптимального параметрического синтеза, не основанных непосредственно на критериях минимального запаса работоспособности. Поэтому задача (5.56) имеет достаточно общий характер.

Далее будут рассмотрены методы решения задачи (5.56) при следующих предположениях.

1. Критерий (5.56) обладает относительно высокой степенью овражности, а его выпуклость гарантируется только в окрестности точки минимума.
2. Размерность (5.55) полного вектора управляемых параметров x велика, что, с одной стороны, затрудняет применение стандартных методов оптимизации из-за нехватки доступной памяти компьютера, а с другой — не позволяет реализовать предельно возможные характеристики сходимости алгоритмов.
3. Решение задачи анализа оптимизируемой системы требует значительных вычислительных затрат. Поэтому в процессе оптимизации требуется минимизировать количество обращений к вычислению значений $J(x)$.
4. Коэффициент заполнения γ матрицы $G(x) = J''(x)$ достаточно мал. Обычно можно полагать $\gamma \sim 1/q$.

Структура матрицы $G(x)$ не зависит от точки x:

$$G(x) = \begin{bmatrix} G_{11} & & 0 & & C_{1q} \\ & G_{22} & & & G_{2q} \\ 0 & & G_{33} & & G_{3q} \\ & & & \ddots & \vdots \\ G_{q1} & G_{q2} & G_{q3} & \cdots & G_{qq} \end{bmatrix}.$$

Подматрица G_{ij} имеет размеры $n_i \times n_j$, а общее число ненулевых элементов равно

$$\sum_{i=1}^{q} n_i^2 + 2n_q \sum_{i=1}^{q-1} n_i.$$

Таким образом, учитывая симметричность матрицы $G(x)$, в памяти компьютера необходимо хранить

$$\sum_{i=1}^{q}\frac{(n_i^2+n_i)}{2}+n_q\sum_{i=1}^{q-1}n_i$$

ненулевых элементов. Необходимые сведения о схемах хранения разреженных матриц содержатся, например, в [12], [23].

5.4.1. Методы с чебышевскими функциями релаксации

Обратимся снова к классу матричных градиентных схем.

Пусть $\lambda_i(G_k) \in [-m, M]$, $M >> m > 0$. В силу приведенных ранее предположений и сформулированных в 5.1 требований к функциям релаксации, наиболее рациональный метод должен иметь функцию релаксации, значения которой резко снижаются от $R = 1$ при $\lambda = 0$, оставаясь малыми во всем диапазоне $[0, M]$. И напротив, при $\lambda < 0$ функция $R(\lambda)$ должна интенсивно возрастать. Кроме того, отвечающая $R(\lambda)$ матричная функция H должна строиться без матричных умножений для сохранения свойства разреженности матрицы $G_k = J''(x^k)$.

Покажем, что в качестве такой $R(\lambda)$ с точностью до множителя могут быть использованы смещенные полиномы Чебышева второго рода $P_s(\lambda)$, удовлетворяющие следующим соотношениям [65]:

$$P_1(\lambda) = 1,\ P_2(\lambda) = 2(1-2\lambda);\ P_{s+1}(\lambda) = 2(1-2\lambda)P_s(\lambda) - P_{s-1}(\lambda). \quad (5.57)$$

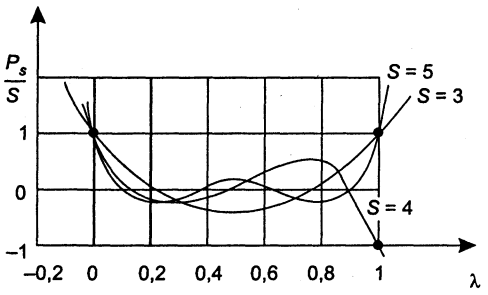

Рис. 5.6. Чебышевские функции релаксации

Графики зависимостей $\frac{P_s(\lambda)}{s}$ для $s = 3, 4, 5$ представлены на рис 5.6. Действительно, полагая $R(\lambda) = \frac{P_L(\lambda)}{L}$ при достаточно большом значении L, получим сколь угодно быструю релаксацию любого слагаемого в представлении

$$f(x^{k+1}) = \frac{1}{2}\sum_{i=1}^{n}\xi_{i,k}^2\lambda_i R^2(\lambda_i), \quad (5.58)$$

где

$$x^k = \sum_{i=1}^{n}\xi_{i,k}u^i.$$

Данное утверждение вытекает из известного факта равномерной сходимости последовательности $\left\{\frac{P_s(\lambda)}{s}\right\}$ к нулю при $s \to \infty$ на открытом промежутке (0, 1). Далее будем предполагать, что собственные числа матрицы G_k нормированы к промежутку (0, 1). Для этого достаточно вместо матрицы G рассматривать матрицу $\frac{G}{\|G\|}$, а вместо вектора g — вектор $\frac{g}{\|G\|}$.

Отвечающая принятой $R(\lambda)$ зависимость $H(\lambda)$ имеет вид

$$H(\lambda) = [1 - R(\lambda)] / \lambda = [1 - P_L(\lambda) / L] / \lambda. \qquad (5.59)$$

Построение методов (5.2) непосредственно с функцией (5.59) возможно, но приводит к необходимости решения на каждом шаге по k больших линейных систем уравнений с разреженной матрицей. Далее показано, что существуют более эффективные приемы реализации.

Из (5.59) следует, что $H(\lambda)$ является полиномом степени $L - 2$, в то время как $R(\lambda)$ имеет степень $L - 1$. Поэтому для реализации матричного градиентного метода с указанной функцией $H(\lambda)$, вообще говоря, нет необходимости решать линейные системы. Метод будет выглядеть следующим образом:

$$x^{k+1} = x^k - \left(\alpha_1 E + \alpha_2 G_k + \ldots + \alpha_{L-1} G_k^{L-2}\right) g^k = x^k - H(G_k) g^k. \qquad (5.60)$$

Реализация метода (5.60) может быть основана на методах вычисления коэффициентов α_i для различных степеней L. При этом число L должно выбираться из условия наиболее быстрого убывания $J(x)$. Далее обсуждается альтернативный подход, основанный на других соображениях.

Для функции

$$H_s(\lambda) \triangleq \alpha_1 + \alpha_2 \lambda + \ldots + \alpha_{s-1} \lambda^{s-2}, \; s = 2, 3, \ldots$$

из (5.57) можно получить рекуррентное соотношение

$$\begin{aligned} &(s+1) H_{s+1} = 2s(1 - 2\lambda) H_s - (s-1) H_{s-1} + 4s; \\ &H_1 = 0, \quad H_2 = 2, \quad s \in [2: L-1]. \end{aligned} \qquad (5.61)$$

Из (5.61) имеем

$$x^{k+1}[s+1] \triangleq x^k - H_{s+1} g^k = x^k - \frac{2s}{s+1}(E - 2G_k) H_s g^k + \frac{s-1}{s+1} H_{s-1} g^k - \frac{4s}{s+1} g^k,$$
$$s \in [2: L-1]$$

или

$$\begin{aligned} &\theta_{s+1} \triangleq x^{k+1}[s+1] - x^k = \frac{2s}{s+1}(E - 2G_k)\theta_s - \frac{s-1}{s+1}\theta_{s-1} - \frac{4s}{s+1} g^k; \\ &\theta_1 = 0, \; \theta_2 = -2g^k, \; s \in [2: L-1]. \end{aligned} \qquad (5.62)$$

Здесь $x^{k+1}[s]$ есть s-е приближение к вектору $x^{k+1} = x^{k+1}[L]$.

Таким образом, при фиксированной квадратичной аппроксимации $f(x)$ функционала $J(x)$ в окрестности $x = x^k$ мы имеем возможность переходить от P_s к P_{s+1} за счет одного умножения матрицы $E - 2G_k$ на вектор θ_s, в полной мере используя свойство разреженности матрицы G_k и не прибегая к дополнительным вычислениям градиента. Эффективность алгоритма (5.62) при больших значениях η определяется множителями релаксации для малых собственных значений матрицы G_k. Рассмотрим положительную часть спектра ($\lambda > 0$), что особенно важно в окрестности оптимума, где матрица $G(x)$ положительно определена. Основное достоинство метода с $R_s(\lambda) = \dfrac{P_s(\lambda)}{s}$ состоит в том, что уже при малых s происходит заметное подавление слагаемых из (5.58) в широком диапазоне значений λ. Далее представлены значения R_s для внутреннего максимума $R_s(\lambda)$ и границы диапазонов $\alpha_s \le \lambda \le \beta_s$, где $|R_s(\lambda)| \le R_s$:

s	...	3	4	5	6	7	8
R_s	...	0,333	0,272	0,250	0,239	0,233	0,230
α_s	...	0,147	0,092	0,061	0,044	0,033	0,025
β_s	...	0,853	0,908	0,939	0,956	0,967	0,975
$-R_s'(0)$	...	5,30	10,0	16,0	23,3	32,0	42,0

В левой части спектра ($\lambda < 0$) имеем

$$R_s(\lambda) > 1 + R_s'(0)\lambda,$$

поэтому значения производных $R_s'(0)$ в последней строке таблицы характеризуют множители релаксации для отрицательных слагаемых в (5.58). Вычисление производных $R_s'(0)$ может быть выполнено, исходя из следующих рекуррентных соотношений:

$$P_1' = 0,\ P_2' = -4;\ P_{s+1}' = 2P_s' - 4s - P_{s-1}';\ R_L'(0) = \frac{P_L'}{L}.$$

Значения α_s, β_s для $s > 8$ (при $\lambda > 0$) могут быть вычислены по асимптотической формуле

$$\alpha_s = \frac{1{,}63}{s^2},\quad \beta_s = 1 - \alpha_s; \tag{5.63}$$

при этом $R_s < 0{,}22$.

Соотношение (5.63) получается из следующего представления полиномов Чебышева:

$$P_L(\lambda) = \frac{\sin L\zeta}{L\sin\zeta},\quad \lambda = \sin^2\frac{\zeta}{2},\ \lambda,\ \zeta \in [0,1].$$

Действительно, при достаточно малых ζ имеем [34]

$$P_L(\lambda) \cong \Phi(\xi) = \frac{\sin\sqrt{\xi}}{\sqrt{\xi}},\quad \xi \triangleq 4L^2\lambda.$$

График функции $\Phi(\xi)$ представлен на рис. 5.7. Полагая $x = \sqrt{\xi}$, получим

$$\Phi(\xi) = \varphi(x) = \frac{\sin x}{x}.$$

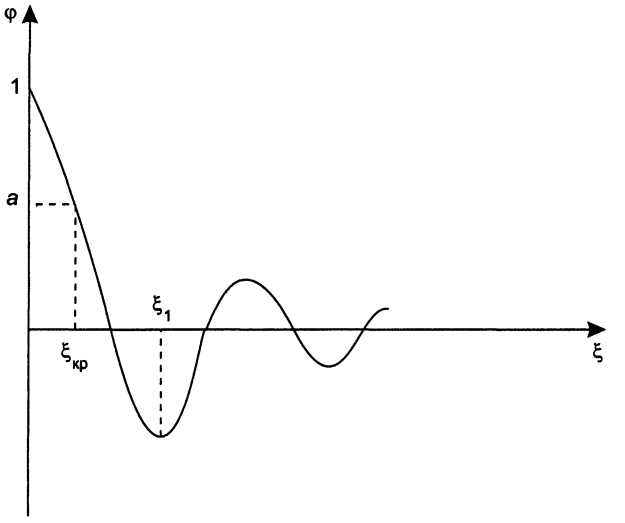

Рис. 5.7. График зависимости $\Phi(\xi) = \dfrac{\sin\sqrt{\xi}}{\sqrt{\xi}}$

Последняя функция и ее числовые характеристики показаны на рис. 5.8.

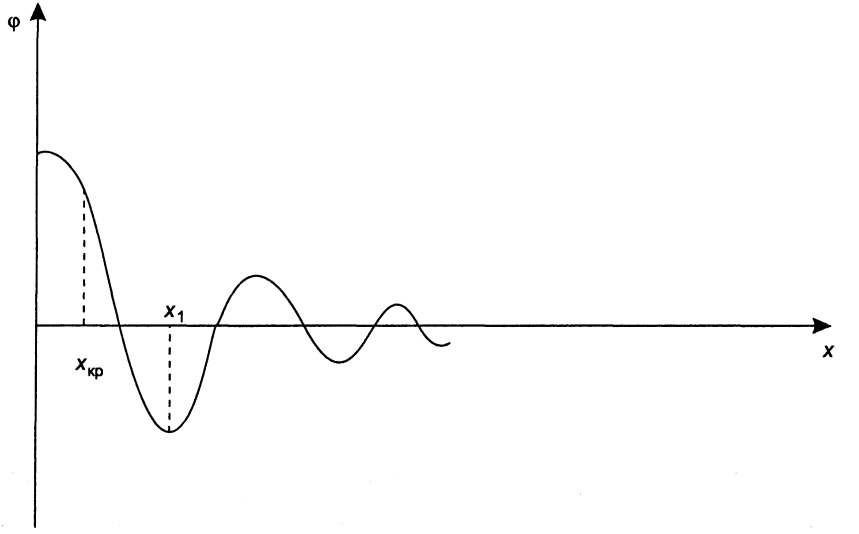

Рис. 5.8. График функции $\varphi(x) = \dfrac{\sin x}{x}$; $x_1 = 4{,}4934$, $x_{кр} = 2{,}5536$, $\varphi(x_1) = -0{,}2172 \cong -0{,}22$, $\varphi(x_{кр}) = |\varphi(x_1)|$

Из рис 5.7 и 5.8 имеем

$$\Phi(\xi) \le \Phi(\xi_{кр}) \text{ при } \xi \ge \xi_{кр},$$

где $\xi_{кр} = x_{кр}^2 = 6{,}523$; $\Phi(\xi_{кр}) = \varphi(x_{кр}) \cong 0{,}22$.

Таким образом, полагая

$$\xi_{кр} = 4L^2\lambda_{кр},$$

получим следующее утверждение: если для наименьшего (положительного) собственного числа m выполняется неравенство

$$\xi_{min} = 4L^2 m \ge \xi_{кр} = 6{,}523,$$

то есть если

$$m \ge \frac{6{,}523}{4L^2} = \frac{1{,}63}{L^2}, \tag{5.64}$$

то для всех $\lambda > m$ будем иметь

$$|R_L(\lambda)| \le 0{,}22.$$

Из (5.64) следует (5.63).

Укрупненная схема алгоритма, построенного на основе соотношения (5.62), может быть реализована с помощью следующей последовательности шагов. При этом предполагается, что все переменные задачи надлежащим образом нормализованы (см. 4.3.1). Предполагается также, что переменные пронумерованы некоторым оптимальным способом, гарантирующим эффективное хранение разреженной матрицы $E - G_k$ в памяти компьютера.

Алгоритм RELCH.

Шаг 1. Задать начальную точку x; вычислить $J := J(x)$; задать L, определяющее количество пересчетов по формуле (5.62) (об априорном выборе L см. далее).

Шаг 2. Вычислить $g := J'(x)$, $G := J''(x)$; принять $g := g/\|G\|$, $G := G/\|G\|$; $\alpha := 1$.

Шаг 3. По формуле (5.62) построить θ_L; принять $x^t := x + \theta_L$.

Шаг 4. Вычислить $J_t := J(x^t)$. Если $J_t > J$, перейти к шагу 5, иначе — к шагу 6.

Шаг 5. Принять $\alpha := \frac{\alpha}{2}$, $x^t := x + \alpha\theta_L$ и перейти к шагу 4.

Шаг 6. Принять $x := x^t$, $J := J_t$ и перейти к шагу 2.

Критерий окончания процесса здесь не указан. Как правило, вычисления заканчиваются по исчерпании заданного количества вычислений функционала либо при явной остановке алгоритма. Число пересчетов L по формуле (5.62) является параметром, задаваемым пользователем. Согласно (5.63) первоначально целесообразно полагать $L \cong \sqrt{\frac{1{,}63}{\alpha_L}} \cong 1{,}3\sqrt{\eta}$, где η — оценка степени овражности минимизируемого функционала. При таком выборе L множители релаксации в положительной части спектра будут гарантированно меньше 0,23. При конструировании алгоритмических способов задания L необходимо учитывать, что последовательность $\{J_s\}$, где $J_s \triangleq J(x^k + \theta_s)$ не будет при $s \to \infty$ убывать монотонно. На шаге 5

алгоритма применена регулировка нормы вектора продвижения с целью предотвращения выхода из области справедливости локальной квадратичной модели функционала.

5.4.2. Характеристики сходимости и сравнение с методами сопряженных градиентов

Дадим оценку эффективности метода (5.62) по сравнению с методами сопряженных градиентов (СГ-методами). Для задач большой размерности (когда число итераций меньше размерности) можно гарантировать сходимость СГ-методов только со скоростью геометрической прогрессии даже для сильно выпуклых квадратичных функционалов [48].

Действительно, рассмотрим случай

$$f(x) = \frac{1}{2}\langle Gx, x\rangle,\ G > 0$$

и оценим скорость сходимости метода СГ к экстремальной точке $x = 0$. Итерация x^k, полученная методом СГ, может быть представлена в виде [65]

$$x^k = (E + c_1 G + c_2 G^2 + \ldots + c_k G^k)\, x^0 = P_k(G)\, x^0,$$

где $P_k(G)$ — матричный полином k-й степени. При этом из свойств метода СГ следует, что коэффициенты $c_1, \ldots, c_k$ полинома $P_k(G)$ на каждой итерации принимают такие значения, чтобы минимизировать величину $f(x^k)$, только множителем отличающуюся от функции ошибки. Иначе говоря, k-е приближение минимизирует $f(x^k)$ среди векторов $x^0 + V$, где вектор V является элементом подпространства, натянутого на векторы $Gx^0, G^2x^0, \ldots, G^kx^0$. Полагая

$$x^0 = \sum_{i=1}^{n} \xi_{i,0} u^i,$$

где $\{u^i\}$ — ортонормальный базис собственных векторов матрицы G, получим

$$x^k = P_k(G)\sum_{i=1}^{n} \xi_{i,0} u^i = \sum_{i=1}^{n} \xi_{i,0} P_k(\lambda) u^i,\ P_k(0) = 1;$$

$$f(x^k) = \frac{1}{2}\left\langle Gx^k, x^k\right\rangle = \frac{1}{2}\sum_{i=1}^{n} \xi_{i,0}^2 P_k^2(\lambda_i)\lambda_i. \qquad (5.65)$$

Отсюда имеем

$$\left\|x^0\right\|^2 = \sum_{i=1}^{n} \xi_{i,0}^2;$$

$$\left\|x^k\right\|^2 = \sum_{i=1}^{n} \xi_{i,0}^2 P_k^2(\lambda_i) \le \max_i P_k^2(\lambda_i)\left\|x^0\right\|^2. \qquad (5.66)$$

Выберем в качестве полинома $P_k(\lambda)$ близкий к оптимальному полином, наименее уклоняющийся от нуля на промежутке $[m, M]$, содержащем все собственные

значения положительно определенной матрицы G и нормированный так, что $P_k(0) = 1$.

Линейной заменой переменных

$$\lambda = \frac{M+m}{2} - \frac{M-m}{2}t$$

задача сводится к построению полинома наименее отклоняющегося от нуля на промежутке $t \in [-1, 1]$ и принимающего в точке $t_0 = \dfrac{M+m}{M-m}$, (соответствующей $\lambda = 0$) значение 1. Решение последней задачи дается полиномом [7]

$$\tilde{T}_k(t) = \frac{T_k(t)}{\cos(k \arccos t_0)} = \frac{T_k(t)}{T_k(t_0)},$$

где $T_k(t) = \cos(k \arccos t)$ — полином Чебышева. При этом

$$\max_{-1 \le t \le 1} \left| \tilde{T}_k(t) \right| = \frac{1}{\left| T_k(t_0) \right|} \max_{-1 \le t \le 1} \left| T_k(t) \right|.$$

Очевидно,

$$\max_{-1 \le t \le 1} \left| T_k(t) \right| = 1,$$

$$\text{поэтому } L_k = \max_{\lambda} \left| P_k(\lambda) \right| = \max_{t} \left| \tilde{T}_k(t) \right| = \frac{1}{T_k(t_0)}, \quad \lambda \in [m, M], \; t \in [-1, 1].$$

Так как справедливо представление

$$T_k(t) = 0{,}5\left[\left(t + \sqrt{t^2 - 1} \right)^k + \left(t - \sqrt{t^2 - 1} \right)^k \right],$$

то

$$L_k = \frac{2}{\left(\dfrac{M+m}{M-m} + \sqrt{\left(\dfrac{M+m}{M-m} \right)^2 - 1} \right)^k + \left(\dfrac{M+m}{M-m} - \sqrt{\left(\dfrac{M+m}{M-m} \right)^2 - 1} \right)^k} =$$

$$= \frac{2}{\left(\dfrac{\sqrt{M} + \sqrt{m}}{\sqrt{M} - \sqrt{m}} \right)^k + \left(\dfrac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \right)^k}.$$

При достаточно больших k ($k \ge k_0$) имеем

$$L_k \cong \frac{2}{\left(\dfrac{\sqrt{M} + \sqrt{m}}{\sqrt{M} - \sqrt{m}} \right)^k} = 2\left(\frac{\sqrt{\eta} - 1}{\sqrt{\eta} + 1} \right)^k \cong 2\left(1 - \frac{2}{\sqrt{\eta}} \right)^k, \quad \eta \triangleq \frac{M}{m}. \tag{5.67}$$

Из (5.66) и (5.67) получаем

$$\| x^k \| \le L_k \| x^0 \|$$

или

$$\| x^k \| \le 2q^k \| x^0 \|, \; k \ge k_0, \tag{5.68}$$

где $q \cong 1 - \frac{2}{\sqrt{\eta}}$. Таким образом, сходимость метода СГ со скоростью геометрической прогрессии доказана. Точное значение L_k, справедливое для любых k, будет при этом равно

$$L_k = \frac{2}{\left[\left(1 + \frac{2}{\sqrt{\eta}}\right)^k + \left(1 - \frac{2}{\sqrt{\eta}}\right)^k\right]}, \; L_0 = 1.$$

Из (5.68) следует, что при $\eta >> 1$ сходимость может быть очень медленной.

«Конечность» метода СГ, то есть точное решение задачи минимизации квадратичной функции за n шагов, где n — размерность пространства поиска, проявляется только при достаточно большом количестве итераций. При этом степень полинома $P_k(\lambda)$ в (5.65) будет равна n, и оптимальный выбор этого полинома сводится к локализации его n корней в точках $\lambda_1, \lambda_2, \ldots, \lambda_n$, что приведет к точному решению задачи ($f(x^n) = 0$).

Легко видеть, что оценка (5.68) для квадратичной функции общего вида

$$f(x) = \frac{1}{2}\langle Gx, x\rangle - \langle G, x\rangle + c$$

преобразуется к виду

$$\| x^k - x^* \| \le 2q^k \| x^0 - x^* \|,$$

где x^* — оптимальная точка, не совпадающая в общем случае с началом координат.

Важная особенность алгоритмов типа RELCH заключается в том, что соответствующие множители релаксации будут определяться только числом итераций L и степенью η обусловленности задачи независимо от размерности n. В то же время в схемах методов СГ для завершения каждого цикла спуска требуется порядка n итераций; в противном случае, согласно последней формуле, скорость сходимости может быть очень малой. Кроме того, каждая итерация метода СГ даже для квадратичного случая требует нового вычисления градиента, то есть дополнительных вычислительных затрат по анализам функционирования оптимизируемой системы.

Будем далее полагать, что алгоритм RELCH реализован с постоянным $L = 1{,}3\sqrt{\eta}$, имея в области $\lambda > 0$ множители релаксации, не превышающие значения 0,23.

Рассмотрим задачу минимизации квадратичного функционала $f(x) = \frac{1}{2}\langle Gx, x\rangle$ с положительно определенной матрицей G. Оценим количество вычислений $f(x)$, требуемое для достижения контрольного вектора x' с нормой $\|x'\| \le 0{,}23$ методом СГ и алгоритмом RELCH из начальной точки x^0 с $\|x^0\| = 1$. По достижении точки x' вся ситуация повторяется, поэтому полученные далее сравнительные оценки эффективности имеют достаточно общий характер.

Будем предполагать, что для вычисления производных применяются двусторонние конечноразностные соотношения, что в нижеследующем анализе дает дополнительные преимущества методу СГ.

Для достижения вектора x' алгоритму RELCH требуется вычислить в точке x^0 слабо заполненную матрицу Гессе и вектор градиента $f'(x^0)$. При коэффициенте заполнения γ это потребует около $2\gamma n^2$ вычислений f. Далее выполняется $L = 1{,}3\sqrt{\eta}$ итераций по формуле (5.62), не требующих дополнительных вычислений целевого функционала f.

Чтобы получить вектор x', методу СГ потребуется N итераций, где число N определяется из условия:

$$\|x^N\| = 2q^N = 0{,}23,$$

то есть $N \cong \frac{-2{,}2}{\ln q}$. Для выполнения каждой итерации необходимо обновление вектора градиента, что связано с $2n$ вычислениями $f(x)$. Общее число вычислений f равно $\frac{-4{,}4n}{\ln q}$. Относительный выигрыш в количестве вычислений f методом RELCH по сравнению с методом СГ задается функцией $\Psi(\eta) \cong -2{,}2/(\gamma n \ln q)$. Очевидно, при $\eta \to \infty$ имеем $q(\eta) \to 1$ и $\Psi(\eta) \to \infty$. Характерные значения Ψ для $\gamma = 0{,}01$ и $n = 1000$ даны ниже:

η	...	100	1000	1500	10^4	10^5
Ψ	...	1,0	3,4	4,0	11,0	35,0

Таким образом, для получения сравнимых результатов при $\eta = 10^4$ алгоритму RELCH потребуется приблизительно в 11 раз меньше вычислений f, чем методу СГ. Следует, однако, учитывать, что при увеличении η возрастает количество L пересчетов по формуле (5.62). Это может приводить к возрастанию влияния вычислительных погрешностей при вычислении θ_s с большими номерами s.

Пример. Рассмотрим модельную задачу минимизации квадратичного функционала $f(x)$ с $n = 200$, $\eta = 1500$, $\gamma = 0{,}025$. Для определенности примем, что время однократного вычисления $f(x)$ эквивалентно выполнению $10^2 n$ операций умножения с плавающей точкой. Время выполнения одной операции умножения для определенности и чисто условно положим равным $t_y = 3\cdot 10^{-5}$ с. Вычисление значения $f(x)$ занимает при этом $t_f = 0{,}6$ с процессорного времени. Для вычисления f' и f'' с помощью общих конечноразностных формул потребуется, соответствен-

но, $t' = 2nt_f = 4$ мин, $t'' = 2\gamma n^2 t = 20$ мин. Число пересчетов по формуле (5.62) равно $L = 1{,}3\sqrt{\eta} = 50$. При каждом пересчете производится умножение слабо заполненной матрицы $E - 2G_k$ на вектор θ_s, что требует $\gamma n^2 t_y \cong 3 \cdot 10^{-2}$ с машинного времени. Время построения вектора θ_{50} без учета вычисления f', f'' составит около $50 \cdot 3 \cdot 10^{-2} = 1{,}5$ с и может в расчет не приниматься.

В результате получается, что для построения контрольного вектора x' с $\|x'\| < 0{,}23$ методом RELCH потребуется около $t'' = 20$ мин машинного времени. Метод СГ затратит, соответственно, $\Psi(1500) \cdot 20 \cong 1{,}3$ ч.

При повторном применении алгоритма RELCH к построенному вектору x' мы получим вектор x'' с $\|x''\| \le 0{,}23\|x'\|$, и т. д. Следовательно, если обозначить соответствующую последовательность векторов через $\{x^m\}$, то норма вектора x будет убывать по закону геометрической прогрессии $\|x^m\| \le d^m \|x^0\|$, где $d < 0{,}23$ независимо от величины η и n.

Важным дополнительным преимуществом алгоритма RELCH по сравнению с методом СГ является его достаточно высокая эффективность в невыпуклом случае, так как функция релаксации метода в левой полуплоскости целиком расположена в разрешенной области и множители релаксации для $\lambda < 0$ быстро растут по абсолютной величине при переходе от θ_s к θ_{s+1}. Характеристики роста были приведены ранее.

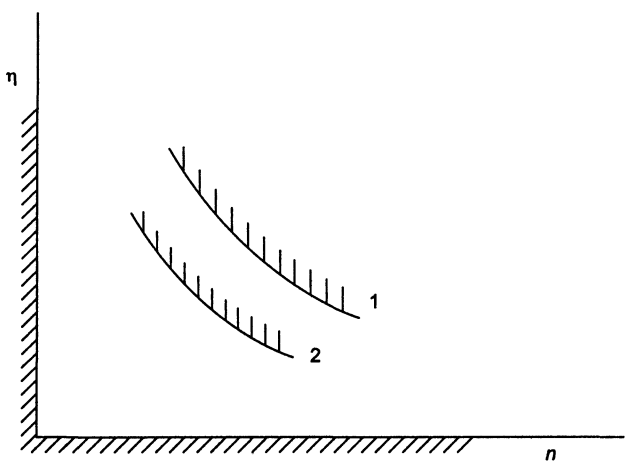

Рис. 5.9. Области работоспособности: *1* — $\varepsilon_м = \varepsilon'_м$; *2* — $\varepsilon_м = \varepsilon''_м > \varepsilon'_м$

Так же как и в случае ЭР-методов, можно показать, что эффективность рассматриваемого подхода сохраняется при степенях овражности, удовлетворяющих неравенству $\eta < \dfrac{1}{n\varepsilon_м}$. Области работоспособности алгоритмов RELEX, RELCH в плоскости (n, η) представлены на рис. 5.9. Ясно, однако, что при умеренных размерностях n более эффективными, вообще говоря, оказываются алгоритмы

типа RELEX. Они позволяют за меньшее число N_y операций умножения матрицы на вектор получить заданные значения множителей релаксации. При больших η это приводит к существенному уменьшению накопленной вычислительной погрешности. Для подтверждения данного замечания достаточно проанализировать характер изменения множителей релаксации при применении формул пересчета (5.39) и (5.62). Характерные зависимости для рассмотренных случаев (для фиксированного $\lambda_i < 0$) представлены на рис. 5.10.

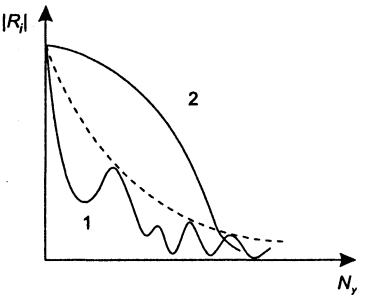

Рис. 5.10. Характер изменения множителей релаксации: *1* — RELCH; *2* — RELEX

Видно, что если область локальной квадратичности функционала $J(x)$ невелика (ζ_k мало), то необходимые значения $|R_i| \cong 1$ и более эффективными могут оказаться методы типа RELCH.

5.5. Применение процедур RELEX и RELCH в задачах теории управления

Рассмотренные процедуры RELEX и RELCH целесообразно реализовывать согласно общим рекомендациям, данным в 4.3.1, 5.3. При этом необходимо использовать специальные приемы вычисления матриц Гессе, основанные на конкретных структурных особенностях отдельных классов задач теории управления (см. 4.5). Априорно затруднительно гарантировать преимущество какого-либо из методов при решении конкретной задачи. Однако имеющийся вычислительный опыт позволяет сделать следующие общие выводы:

1. Для задач с квадратичным критерием качества следует ожидать наибольшей эффективности от процедур класса RELEX при условии, что степень овражности целевого функционала позволяет считать матрицу Гессе информативной. К таким задачам, в частности, относятся (см. 1.3):
   - задачи идентификации нелинейных детерминированных объектов при использовании рядов Вольтерра с параметрическим представлением обобщенных весовых функций;
   - задачи усреднения входных и выходных сигналов при идентификации стохастических объектов;

- задачи идентификации линейных стохастических объектов с аддитивной помехой и применением корреляционных методов идентификации;
- задачи определения статистически оптимальных весовых функций линейных стационарных и нестационарных систем автоматического управления при использовании алгебраических методов решения соответствующих уравнений Винера—Хопфа и их регуляризованных вариантов.

При потере информативности матриц Гессе минимизируемых функционалов, что выражается в резком замедлении сходимости процедур типа RELEX, оправдан переход к более универсальных методам ОПС.

2. В условиях высокой размерности вектора управляемых параметров x целесообразно непосредственно обращаться к процедурам типа RELCH, основанным на компактном представлении матриц вторых производных в памяти компьютера.

Из рассмотренных в 1.3 примеров задач теории управления наиболее часто высокие размерности вектора x возникают при использовании алгебраических методов решения интегральных уравнений Винера—Хопфа. При этом, как показала практика, за счет искусственного увеличения размерности в этих задачах удается добиться регуляризующего эффекта, предотвращающего появление паразитных высокочастотных составляющих в решении.

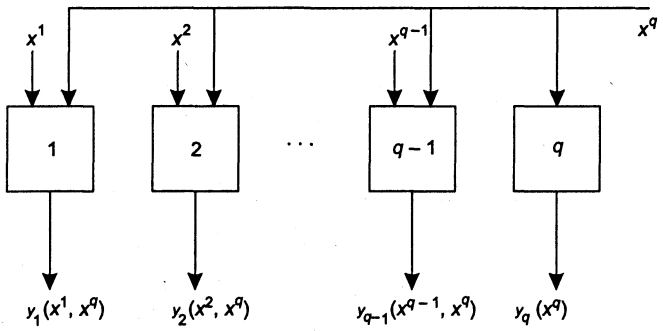

Рис. 5.11. Совокупность взаимосвязанных подсистем

Типичным примером, иллюстрирующим область рационального использования алгоритмов RELCH, является задача параметрической оптимизации (или параметрической настройки — при работе в режиме реального времени) сложной системы, состоящей из нескольких взаимосвязанных подсистем. Структурная схема такой системы показана на рис. 5.11. Один из возможных подходов к формированию соответствующего критерия оптимальности с указанием структуры матрицы Гессе был изложен ранее (см. соотношения (5.54), (5.56)). В качестве конкретного реального примера можно рассмотреть задачу параметрического синтеза *большой интегральной схемы* (БИС), расположенной на единой подложке. В этом случае блоки 1, ..., $q-1$ характеризуют функциональные подблоки БИС, а параметры x^q связующего блока q отражают общие для всех подсистем параметры подложки, источников питания, а также внешней среды, определяющей условия функционирования системы.

5.6. Тактика решения общей задачи параметрической оптимизации

На основе представленных в книге методов можно сформулировать общий подход к решению «незнакомой» задачи параметрической оптимизации. Мы здесь рассмотрим ограниченный класс однокритериальных задач безусловной оптимизации с гладкими целевыми функционалами, к которым, как было показано, сводятся многие более сложные практические ситуации. Основные факторы, которые учитываются при выборе метода, — это машинное эпсилон $\varepsilon_м$, характеризующее точность вычислений, размерность пространства поиска n и степень овражности η минимизируемого функционала (рис. 5.12).

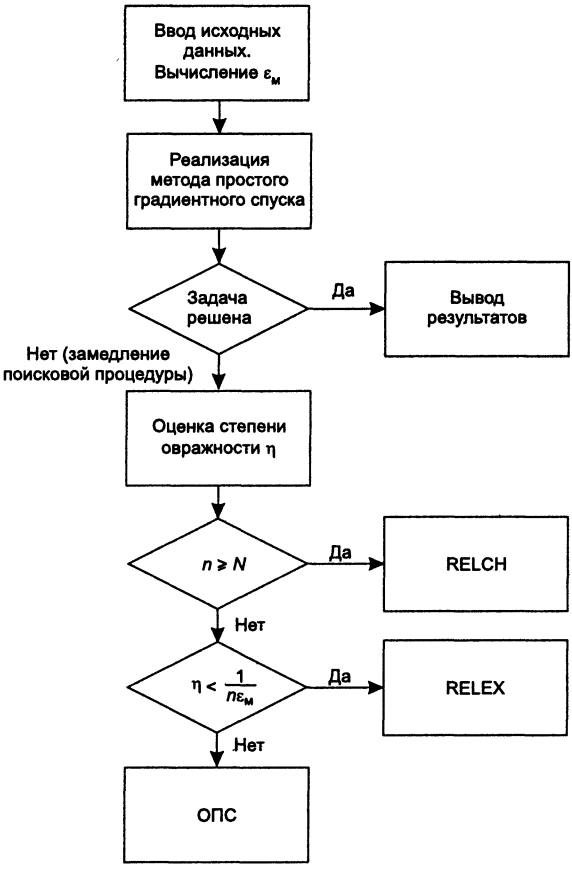

Рис. 5.12. Тактика решения общей задачи параметрической оптимизации

После ввода исходных данных определяется машинное эпсилон на основе известных из курса численного анализа методов. Далее применяются стартовые алгоритмы на основе метода циклического покоординатного спуска и метода простого градиентного спуска. Если задача алгоритмически проста, то уже на

этом этапе может быть получено разумное приближение к искомому результату. К сожалению, обычно ситуация оказывается более сложной и происходит резкое замедление сходимости задолго до достижения приемлемых точек в пространстве поиска. В этом случае, как это было показано ранее, на основе результатов работы метода простого градиентного спуска может быть получена оценка степени овражности минимизируемого функционала. После этого необходимо осуществить переход к более мощным вычислительным процедурам. Если размерность решаемой задачи превышает некоторый заданный порог N, определяемый объемом доступной памяти компьютера, то производим переход к алгоритмам типа RELCH. В противном случае определяется «информативность» матрицы Гессе на основе проверки известного неравенства, содержащего степень овражности и машинное эпсилон. Если матрица Гессе информативна, обращаемся к методам с экспоненциальной релаксацией, в противном случае — к различным вариантам методов обобщенного покоординатного спуска. Можно сразу же переходить к методам обобщенного покоординатного спуска как к наиболее мощным и универсальным процедурам.

5.7. Основные результаты и выводы

В пятой главе представлены следующие основные результаты.

1. Для класса матричных градиентных методов введено понятие функции релаксации, позволяющее с единых позиций оценивать эффективность градиентных оптимизирующих процедур, а также синтезировать новые методы, ориентированные на специальные классы плохо обусловленных оптимизационных задач. Получены соотношения, позволяющие по произвольно заданной функции релаксации строить соответствующие поисковые процедуры.
2. Доказана теорема об условиях релаксационности произвольного матричного градиентного метода. На основе понятия функции релаксации рассмотрена геометрическая интерпретация релаксационных свойств градиентных методов, позволяющая строить области релаксационности, а также оценивать скорость убывания значений минимизируемых функционалов.
3. Дан анализ классических матричных градиентных схем: простого градиентного метода, метода Ньютона, метода Левенберга. Показано, что функция релаксации дает практически исчерпывающую информации о свойствах и возможностях соответствующей поисковой процедуры.
4. Рассмотрен класс методов с экспоненциальной функцией релаксации, естественным образом обобщающих классические методы спуска по антиградиенту, Н-методы, а также методы доверительной окрестности. В отличие от указанных методов, ЭР-методы имеют функции релаксации, целиком расположенные в разрешенной области, что существенно облегчает процедуру настройки параметра h, характеризующего норму вектора продвижения в пространстве поиска.

   Доказаны теоремы, устанавливающие сходимость ЭР-методов для достаточно широкого класса невыпуклых функционалов и гарантирующие квад-

ратичную скорость сходимости в предположении сильной выпуклости целевого функционала.

5. Изучены реализации ЭР-методов (алгоритм RELEX). Дан анализ влияния погрешностей на релаксационные свойства ЭР-методов, что позволило указать области эффективного применения соответствующих алгоритмов в пространстве параметров: n (размерность решаемой задачи), η (степень обусловленности), $\varepsilon_м$ (машинное эпсилон). Модульная структура алгоритма RELEX позволяет воспользоваться развитой в главе 4 методикой аппроксимации матриц вторых производных с целью повышения эффективности решения конкретных классов задач теории управления.
6. Изучены методы параметрической оптимизации с большим числом управляемых параметров по функционалам со слабозаполненными и структурированными матрицами Гессе. В качестве основы для создания соответствующих оптимизирующих процедур рассмотрены новые матричные градиентные схемы с чебышевскими функциями релаксации. Построено рекуррентное соотношение, позволяющее сохранять свойство разреженности матриц и упакованные формы их хранения в памяти компьютера на протяжении каждого цикла процесса оптимизации. Исследованы релаксационные свойства методов и особенности их реализации. Описан базовый алгоритм RELCH. Даны характеристики сходимости алгоритма RELCH и проведено его сравнение с методами сопряженных градиентов (СГ). Получена оценка скорости сходимости метода СГ, а также выражение для «коэффициента выигрыша» Ψ алгоритмов класса RELCH по сравнению с методами СГ в зависимости от размерности решаемой задачи, ее обусловленности и степени разреженности соответствующей матрицы Гессе. Показано, что уже для относительно небольших (порядка 10^4) степеней обусловленности при достаточно больших (порядка 10^3) n выигрыш, получаемый алгоритмом RELCH, может быть весьма значительным. Указаны области работоспособности алгоритмов RELCH.
7. Даны общие рекомендации по применению процедур RELEX, RELCH и ОПС в задачах теории управления. Рассмотрена тактика решения общей конечномерной оптимизационной задачи с помощью изученного алгоритмического обеспечения.

Глава 6. Методы уменьшения размерности вектора оптимизируемых параметров

В этой главе рассматриваются две причины, приводящие к необходимости сокращения размерности пространства управляемых параметров оптимизируемой системы. Первая причина (см. 6.1) связана с разработкой методов понижения локальной степени овражности минимизируемых функционалов при решении плохо обусловленных (овражных) оптимизационных задач.

Вторая причина (см. 6.2) обусловлена необходимостью исключения максимально возможного числа компонентов управляемого вектора, что эквивалентно фиксированию нулевых значений соответствующих переменных. Интерпретация последнего преобразования заключается в следующем. Во-первых, может ставиться задача структурного синтеза с помощью упрощения исходной, заведомо избыточной структуры посредством удаления «несущественных» элементов. Во-вторых, часто возникает необходимость в построении «минимальных» параметрических представлений искомых непрерывных зависимостей, например, в таких задачах теории управления, как задачи идентификации нелинейных объектов на основе функциональных рядов Вольтерра (дополнительные примеры содержатся в главе 1).

6.1. Методы теории жестких систем

Как следует из общего определения овражного функционала, дифференциальные уравнения, описывающие траектории наискорейшего спуска таких функционалов, относятся к классу жестких. При этом уравнения дна оврага можно трактовать как уравнения алгебраических связей, устанавливающихся вне пограничного слоя между фазовыми переменными жестких дифференциальных систем. Эти связи могут быть найдены на основе *принципа квазистационарности производных* (ПКП), примененного к уравнению

$$\frac{dx}{dt} = -g(x),\ g(x) \triangleq J'(x) \triangleq \frac{\partial J}{\partial x} \tag{6.1}$$

и обобщающего известный метод квазистационарных концентраций Семенова—Боденштейна. ПКП для общего нелинейного случая был предложен Ю. В. Ракитским [49]. Анализ, приведенный далее и относящиеся к линейным системам теоремы и вычислительные технологии были получены автором настоящей книги.

6.1.1. Принцип квазистационарности производных для линейных систем с симметричными матрицами

Решение уравнения спуска

$$\frac{dx}{dt} = -Dx + b \tag{6.2}$$

для параболоида

$$f(x) = 0{,}5\langle Dx, x\rangle - \langle b, x\rangle + c,\ x \in R^n,\ c \in R^1, \tag{6.3}$$

являющегося локальной моделью исходного функционала $J(x)$, может быть представлено в виде

$$x(t) = \sum_{k=1}^{n}\left[\alpha_k \exp\left(-\lambda_k(D)t\right) + \beta_k \int_0^t \exp(-\lambda_k(D)\tau)d\tau\right]u^k, \tag{6.4}$$

где постоянные α_k, β_k определяются соотношениями

$$x^0 = \sum_{k=1}^{n}\alpha_k u^k;\ b = \sum_{k=1}^{n}\beta_k u^k,$$

а $\{u^k\}$, $k \in [1: n]$, есть ортонормированный базис, составленный из собственных векторов матрицы D.

Пусть собственные числа матрицы D удовлетворяют неравенствам

$$\lambda_1(D) \geq \ldots \geq \lambda_{n-r}(D) \geq \sigma\left|\lambda_{n-r+1}(D)\right| \geq \ldots \geq \sigma\left|\lambda_n(D)\right|,\ \sigma >> 1, \tag{6.5}$$

что соответствует овражной структуре функционала (6.3) в смысле определений, приводимых в 3.2.

Умножая обе части равенства (6.4) скалярно на u^i, $i \in [1: n - r]$, и используя свойство ортонормированности базиса $\{u^i\}$:

$$\left\langle u^i, u^j\right\rangle = \begin{cases} 1, & i = j; \\ 0, & i \neq j, \end{cases}$$

получаем

$$\langle x, u^i \rangle = \alpha_i \exp[-\lambda_i(D)t] - \frac{\beta_i}{\lambda_i(D)}[\exp(-\lambda_i(D)t) - 1] \underset{t>\tau_{\text{п.с}}}{\cong} \frac{\beta_i}{\lambda_i(D)},$$

где $\tau_{\text{п.с}}$ — длина *пограничного слоя*, $\tau_{\text{п.с}} \sim 1/\sigma$. Понятие пограничного слоя вводится в курсе численного анализа при изучении жестких систем дифференциальных уравнений.

Так как, очевидно, $\beta_i = \langle b, u^i \rangle$, то алгебраические связи между фазовыми траекториями $x_i(t)$ имеют вид

$$\langle x, u^i \rangle = \lambda_i^{-1}(D)\langle b, u^i \rangle, \quad i \in [1:n-r]. \tag{6.6}$$

Полученные соотношения, согласно развитым в 3.2 представлениям, определяют дно оврага.

На основе равенств (6.6) для рассматриваемого частного случая (6.2) уравнений спуска ПКП может быть доказан непосредственно.

Теорема 6.1 (ПКП для линейных систем с симметричными матрицами).

Пусть:

1) $x(t)$ — решение уравнения (6.2) с постоянной симметричной матрицей D;

2) собственные числа матрицы D удовлетворяют неравенствам (6.5).

Тогда существует такой индекс $j \in [1: n]$, что уравнение

$$\frac{d^s x_j}{dt^s} = 0 \tag{6.7}$$

асимптотически ($t \to \infty$, $s \to \infty$) эквивалентно соотношению

$$\langle x, u^1 \rangle = \lambda_1^{-1}(D)\langle b, u^1 \rangle,$$

где u^1 — собственный вектор матрицы D, соответствующий собственному числу $\lambda_1(D)$.

Доказательство. Дифференцируя уравнение спуска (6.2) $s - 1$ раз и обозначая j-й вектор-столбец матрицы A^s, где $A \triangleq -D$, через c_j^s, имеем

$$\frac{d^s x_j}{dt^s} = \langle c_j^s, x \rangle + \langle c_j^{s-1}, b \rangle,$$

где

$$c_j^s = A^s e^j, e^j \triangleq (0, \ldots, 1, \ldots, 0) = \sum_{k=1}^{n} \langle e^j, u^k \rangle u^k \tag{6.8}$$

j-й единичный орт. Последнее представление для e^j получается следующим образом. Разложим вектор e^j по векторам ортонормального базиса $\{u^k\}$:

$$e^j = \sum_{k=1}^{n} \alpha_k u^k$$

и определим коэффициенты α_k, умножая это равенство скалярно на некоторый вектор базиса u^m. Имеем

$$\left\langle e^j, u^m \right\rangle = \sum_{k=1}^{n} \alpha_k \left\langle u^k, u^m \right\rangle = \alpha_m,$$

так как $\left\langle u^k, u^m \right\rangle = 1$ при $k = m$ и $\left\langle u^k, u^m \right\rangle = 0$ — в противном случае. В результате имеем (6.8).

Отсюда

$$\frac{d^s x_j}{dt^s} = \lambda_1^{s-1}(A) \sum_{k=1}^{n} u_j^k \left[\frac{\lambda_k(A)}{\lambda_1(A)}\right]^{s-1} [\lambda_k(A)\left\langle u^k, x \right\rangle + \left\langle u^k, b \right\rangle].$$

Пренебрегая слагаемыми с номерами $k = l + 1, \ldots, n$, где l — кратность максимального собственного числа $\lambda_1(D)$, получим

$$\frac{d^s x_j}{dt^s} = \lambda_1^s(A)[\left\langle v^j, x \right\rangle + \lambda_1^{-1}(A)\left\langle v^j, b \right\rangle], \tag{6.9}$$

где $v^j = \sum_{k=1}^{l} u_j^k u^k$ — собственный вектор матрицы D, соответствующий $\lambda_1(D)$. Действительно,

$$Dv^j = \sum_{k=1}^{l} u_j^k \lambda_k u^k = \sum_{k=1}^{l} u_j^k \lambda_1 u^k = \lambda_1 \sum_{k=1}^{l} u_j^k u^k = \lambda_1 v^j.$$

Из (6.9) видно, что условие квазистационарности (6.7) $(s - 1)$-й производной от j-го компонента вектора $x(t)$, соответствующего ненулевому вектору v^j, эквивалентно искомому уравнению связи (6.6) для $i = 1$.

Если выбранный номер j определяет нулевые компоненты u_j^k, $k = 1, \ldots, l$, и, соответственно, нулевой вектор v^j (что в общем случае маловероятно), то необходимо перейти к другому номеру j. Покажем, что такой номер j, для которого $v^j \neq 0$, всегда найдется.

Действительно, пусть $v^j = 0$ при $\forall j \in [1: n]$.

Тогда

$$\forall j \in [1:n], \forall i \in [1:l] : u_j^i = \left\langle v^j, u^i \right\rangle = 0,$$

то есть $u^i = 0$, $i \in [1: l]$, а это противоречит тому, что u^i — собственные векторы и поэтому не равны нулю. Теорема доказана.

Остановимся несколько подробнее на практическом определении номера j уравнения (6.7) для получения соответствующей линейной связи.

Предположим для простоты, что кратность максимального собственного числа $l = 1$. Тогда

$$v^j = \sum_{k=1}^{l} u_j^k u^k = u_j^1 u^1,$$

и из (6.9) получим

$$\frac{d^s x_j}{dt^s} = \lambda_1^s(A) u_j^1 [\langle u^1, x\rangle + \lambda_1^{-1}\langle u^1, b\rangle].$$

Ясно, что если для некоторых j имеем $u_j^1 \neq 0$, то соответствующее уравнение (6.7) доставляет искомую линейную связь. Если матрица D не является блочно-диагональной матрицей вида

$$D = \begin{bmatrix} L & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & N \end{bmatrix}, \tag{6.10}$$

то пригодно любое уравнение (6.7).

Действительно, пусть, например, выделено первое уравнение

$$\frac{d^s x_1}{dt^s} = 0$$

и при этом случайно оказалось $u_1^1 = 0$. Из-за ошибок округления при каком-то s получим

$$c_1^s = (A^s + \Pi_s) e^1$$

вместо точного соотношения (6.8).

Здесь Π_s — матрица погрешности в задании матрицы A_s. Имеем

$$c_1^s = \sum_{j=2}^{n} u_1^j \lambda_j^s(A) u^j + \xi_{s1},\ \xi_{s1} = \Pi_s e^1 = \sum_{j=1}^{n} \xi_{s1}^j u^j.$$

И так как

$$c_1^{s+m} = A^m c_1^s = \sum_{j=2}^{n} u_1^j \lambda_j^{m+s}(A) u^j + \sum_{j=1}^{n} \xi_{s1}^j \lambda_j^m(A) u^j,$$

то

$$\frac{d^{s+m} x_1}{dt^{s+m}} = \langle c_1^{s+m}, x\rangle + \langle c_1^{s+m-1}, b\rangle = \sum_{j=2}^{n} u_1^j \lambda_j^{s+m-1}(A)[\lambda_j(A)\langle u^j, x\rangle + \langle u^j, b\rangle] +$$

$$+ \sum_{j=1}^{n} \xi_{s1}^j \lambda_j^{m-1}(A)[\lambda_j(A)\langle u^j, x\rangle + \langle u^j, b\rangle].$$

При дальнейшем увеличении m слагаемое

$$\xi_{s1}^1 \lambda_1^m(A)[\langle u^1, x\rangle + \lambda_1^{-1}(A)\langle u^1, b\rangle]$$

начнет быстро возрастать по сравнению с остальными и в конце концов станет доминирующим, определяя искомое уравнение связи.

Если матрица D имеет блочную структуру (6.10), то исходная оптимизационная задача распадается на несколько независимых подзадач для разных параболоидов.

На практике целесообразно выбирать уравнение (6.7) со сравнительно большими коэффициентами при компонентах вектора x — это свидетельствует о наличии полезной информации о больших собственных числах. Порядок дифференцирования s может быть выбран достаточно малым; обычно $s = 2{,}3$ (в условиях применимости метода Семенова—Боденштейна имеем $s = 1$). Необходимо лишь, чтобы вектор, аналогичный v^j из выражения (6.9), оказался в подпространстве, натянутом на собственные векторы, соответствующие доминирующим собственным числам λ_i, $i \in [1: n - r]$. Последнее условие выполняется уже при малых s, так как предполагается, что

$$\lambda_{n-r} \geq \sigma |\lambda_{n-r+1}|, \ \sigma >> 1.$$

Замечание 1. Изложенный метод построения линейных зависимостей (6.6) никак не связан с проблемой разделения близких собственных чисел. Если, например, λ_1 существенно больше λ_2, то такое разделение при надлежащем выборе номера уравнения (6.8) действительно произойдет. Если же $\lambda_1 \cong \lambda_2$ — почти кратное собственное число, то нет необходимости увеличивать s до разделения λ_1 и λ_2. В качестве коэффициентов при компонентах вектора x в (6.5) в этом случае будут использоваться не компоненты вектора u^1, а компоненты вектора вида $\alpha_1 u^1 + \alpha_2 u^2$.

Замечание 2. В общем случае задача определения u^1 по приближенной матрице D может оказаться плохо обусловленной, то есть малые вариации элементов матрицы D при наличии близких к λ_1 собственных чисел могут приводить к значительным вариациям составляющих вектора u^1. Однако для излагаемого подхода указанное обстоятельство не имеет существенного значения, ибо, как показано в 4.2, вариации u^1 будут необходимо происходить в пределах линейной оболочки собственных векторов, соответствующих близким к λ_1 собственным числам, и поэтому любой из построенных векторов u^1 может быть использован в уравнении (6.6).

Аналогично изложенному ранее, можно показать, что существуют такие номера $j_1, j_2, \ldots, j_{n-r} \in [1: n]$, которые определяют асимптотически эквивалентную соотношениям (6.6) подсистему уравнений (6.7). Однако вышеизложенный подход, связанный с последовательным выделением одиночных доминирующих связей, алгоритмически наиболее рационален.

6.1.2. Методы иерархической оптимизации: частный случай

Рассмотрим частный случай задачи минимизации параболоида

$$f(x) = 0{,}5 \langle Dx, x \rangle - \langle b, x \rangle + c \qquad (6.11)$$

в предположении

$$\lambda_1(D) \geq \sigma\lambda_i(D) > 0,\ \sigma >> 1,\ i \in [2{:}\ n]. \tag{6.12}$$

Поставим задачу построения нового параболоида, в некотором смысле эквивалентного заданному, но матрица Гессе которого уже не имеет значительного различия в собственных числах.

С помощью линейного соотношения (6.6) для $i = 1$ выразим компонент x_k вектора x, которому соответствует максимальный по модулю компонент u_k^1 собственного вектора u^1:

$$x_k = L(z) =$$
$$= -\frac{1}{u_k^1}\left[\sum_{\substack{j=1\\ j\neq k}}^{n} u_j^1 x_j - \lambda_1^{-1}(D)\langle b, u^1\rangle\right],\ z \triangleq (x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n). \tag{6.13}$$

Определим функционал $f_1(z)$:

$$f_1(z) \triangleq f(x_1, x_2, \ldots, x_{k-1}, L(z), x_{k+1}, \ldots, x_n) =$$
$$= 0{,}5\langle f_1'' z, z\rangle - \langle d, z\rangle + c_1. \tag{6.14}$$

Теорема 6.2. Пусть заданы функционалы (6.11), (6.14) и выполняются неравенства (6.12). Тогда

1) Матрица f_1'' положительно определена ($f_1'' > 0$) и минимизаторы x^*, z^* параболоидов f, f_1 связаны соотношениями

$$x^* = (z_1^*, \ldots, z_{k-1}^*, L(z^*), z_k^*, \ldots, z_{n-1}^*) \triangleq [L_k(z^*), z^*],$$
$$z^* = (x_1^*, \ldots, x_{k-1}^*, x_{k+1}^*, \ldots, x_n^*); \tag{6.15}$$

2) спектральные числа обусловленности матриц D, f_1'' удовлетворяют неравенству

$$k(f_1'') \leq (n-1)(n+1)\sigma^{-1}k(D). \tag{6.16}$$

Доказательство. Матрицы f'', f_1'', очевидно, не изменятся, если в выражении (6.11) принять $b = 0$, $c = 0$. При этом для любого $z \neq 0$ по определению $f_1(z)$ будем иметь

$$0{,}5\langle f_1'' z, z\rangle = 0{,}5\langle Dx, x\rangle, \tag{6.17}$$

где $x = [L_k(z), z]$.

Выражение справа в равенстве квадратичных форм (6.17) больше нуля из-за условия $D > 0$. Отсюда следует неравенство $f_1'' > 0$.

Минимизатор x^* параболоида $f(x)$, очевидно, удовлетворяет уравнению $x_k^* = L(z^*)$, так как имеем

$$\langle u^1, x^*\rangle = \langle u^1, D^{-1}b\rangle = \langle D^{-1}u^1, b\rangle = \lambda_1^{-1}(D)\langle b, u^1\rangle.$$

Поэтому $x^* = [L_k(\tilde{z}), \tilde{z}]$, где $\tilde{z} \triangleq (x_1^*, \ldots, x_{k-1}^*, x_{k+1}^*, \ldots, x_n^*)$.

Первое соотношение (6.15) доказано. Покажем теперь, что, $\tilde{z} = z^*$. Имеем

$$\forall x^0 \in R^n : f_1(\tilde{z}) = f(x^*) \le f(x^0).$$

Пусть существует $z' : f_1(z') < f_1(\tilde{z})$. Тогда

$$f([L_k(z'), z']) = f_1(z') < f_1(\tilde{z}) = f(x^*),$$

что противоречит свойству минимизатора x^*. Следовательно, $\tilde{z} = z^*$, и первое утверждение теоремы доказано.

Предположим, не ограничивая общности, что $|u_1^1| \ge |u_i^1|$, $i \in [2:n]$. Непосредственным дифференцированием получаем выражение элементов g_{ij} матрицы f_1'' через элементы d_{ij} матрицы D:

$$g_{ij} = d_{ij} = \frac{u_i^1}{u_1^1} d_{1j} - \frac{u_j^1}{u_1^1} d_{1i} - \frac{u_i^1 u_j^1}{(u_1^1)^2} d_{11}; \quad i, j \in [2:n]. \tag{6.18}$$

Учитывая представление

$$d_{ij} = \sum_{k=1}^{n} u_i^k u_j^k \lambda_k(D),$$

получим для диагональных элементов матрицы f_1'':

$$g_{ii} = \sum_{k=1}^{n} \left(u_i^k - \frac{u_i^1}{u_1^1} u_1^k \right)^2 \lambda_k(D) = \sum_{k=2}^{n} \left(u_i^k - \frac{u_i^1}{u_1^1} u_1^k \right)^2 \lambda_k(D), \quad i \in [2:n].$$

Суммируя по i, приходим к выражению для следа Sp матрицы f_1'':

$$Sp(f_1'') = \sum_{i=2}^{n} g_{ii} = \sum_{k=2}^{n} \lambda_k(D) \sum_{i=1}^{n} \left(u_i^k - \frac{u_i^1}{u_1^1} u_1^k \right)^2 = \sum_{k=2}^{n} \lambda_k(D) \left[1 + \left(\frac{u_1^k}{u_1^1} \right)^2 \right], \quad \left(\frac{u_1^k}{u_1^1} \right)^2 \le n.$$

Следовательно,

$$\lambda_1(f_1'') \le Sp(f_1'') \le (n+1) \sum_{k=2}^{n} \lambda_k(D) \le (n+1)(n-1) \sigma^{-1} \lambda_1(D). \tag{6.19}$$

Пусть $\min_i \lambda_i(f_1'') \triangleq \lambda_{n-1}(f_1'')$.

Для соответствующего этому числу собственного вектора z имеем

$$\lambda_{n-1}(f_1'') = \frac{\langle f_1'' z, z \rangle}{\langle z, z \rangle}.$$

Тогда для вектора $x = [(L(z), z)]$ получаем, согласно равенству (6.17)

$$\frac{\langle Dx,x\rangle}{\langle x,x\rangle}=\frac{\langle f_1''z,z\rangle}{\langle z,z\rangle+L^2(z)}\le\frac{\langle f_1''z,z\rangle}{\langle z,z\rangle}=\lambda_{n-1}(f_1'').$$

А так как

$$\lambda_n(D)=\min_i\lambda_i(D)=\min_{x\neq0}\frac{\langle Dx,x\rangle}{\langle x,x\rangle},$$

то, очевидно,

$$\lambda_n(D)\le\lambda_{n-1}(f_1''). \tag{6.20}$$

Из (6.19), (6.20) имеем

$$k(f_1'')=\frac{\lambda_1(f_1'')}{\lambda_{n-1}(f_1'')}\le\frac{(n+1)(n-1)\sigma^{-1}\lambda_1(D)}{\lambda_n(D)}=(n+1)(n-1)\sigma^{-1}k(D), \tag{6.21}$$

что доказывает неравенство (6.16).

Из доказанной теоремы следует, что в результате исключения переменной x задача минимизации овражного параболоида f сведена к аналогичной задаче для параболоида f_1 с существенно меньшей степенью овражности. Необходимая линейная связь (6.13). согласно теореме 6.1, может быть найдена с помощью ПКП.

6.1.3. Методы иерархической оптимизации: общий случай

Перейдем к исследованию общего случая r-овражного ($1 \le r \le n-1$) необязательно выпуклого параболоида.

Пусть собственные числа постоянной симметричной ($n\times n$)-матрицы D удовлетворяют неравенствам

$$\lambda_1(D)\ge\ldots\ge\lambda_{n-r}(D)\ge\sigma\,|\lambda_{n-r+1}(D)|\ge\ldots\ge\sigma\,|\lambda_n(D)|,$$

так что параболоид

$$f(x)=0{,}5\,\langle Dx,x\rangle-\langle b,x\rangle+\text{const} \tag{6.22}$$

является r-овражным (где r — размерность дна оврага) в смысле определений главы 3.

Пусть для некоторого $i\in[1{:}\,n-r]$ задано уравнение

$$\langle x,u^i\rangle=\lambda_i^{-1}\langle b,u^i\rangle. \tag{6.23}$$

Для определенности примем $i=1$. Пусть

$$k=\arg\max_j|u_j^1|. \tag{6.24}$$

Тогда, согласно (6.23) можно выразить k-й компонент вектора x через остальные и аналогично (6.14) построить новый функционал

$$f_1(z) = 0{,}5\left\langle G_k^D z, z\right\rangle - \langle d, z\rangle + \text{const}, \; G_k^D = f_1''(z), \; z = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n). \quad (6.25)$$

Таким образом, определен оператор $(D,k) \to G_k^D$, ставящий в соответствие матрице D размерности $n \times n$ и номеру $k \in [1: n]$ исключаемой переменной матрицу G_k^D размерности $(n-1) \times (n-1)$.

Лемма 6.1. Пусть задано уравнение относительно переменной $\lambda \in R^1$:

$$\lambda + \sum_{i=1}^{k} a_i b_i^2 (b_i - \lambda)^{-1} = \sum_{i=1}^{k} a_i b_i, \; a_i > 0, \; b_i \neq 0, \; b_j > b_{j+1}, \; j \in [1, k-1]. \quad (6.26)$$

Тогда:

1. Корни λ_i уравнения (26) различны, их число равно $k+1$, и они строго разделяются числами b_i:

$$b_1 < \lambda_1; b_i \leq \lambda_i \leq b_{i-1}; \; i \in [2: k]; \; \lambda_{k+1} < b_k, \quad (6.27)$$

при этом один из корней $\lambda_j = 0, j \in [1: k+1]$.

2. Если $b_1 > 0$, $b_k < 0$, то

$$b_k(1+s) \leq \lambda_i \leq b_1(1+s), \; i \in [1: k+1]; \quad (6.28)$$

если $b_1 > 0$, $b_k > 0$, то

$$b_k < \lambda_i \leq b_1(1+s), \; i \in [1: k], \; \lambda_{k+1} = 0; \quad (6.29)$$

если $b_1 < 0$, $b_k < 0$, то

$$b_k(1+s) \leq \lambda_i < b_1, \; i \in [2: k+1], \; \lambda_1 = 0, \quad (6.30)$$

где $s \overset{\Delta}{=} \sum_{i=1}^{k} a_i$.

3. Если $b_1 > |b_i|$, $i \in [2: k]$, то корни μ_i уравнения

$$\mu + \sum_{i=2}^{k} a_i b_i^2 (b_i - \mu)^{-1} = \sum_{i=2}^{k} a_i b_i \quad (6.31)$$

удовлетворяют неравенствам

$$|\mu_j| \geq |\lambda_{j+1}|, \; j \in [1: k]. \quad (6.32)$$

Доказательство. 1. Числа b_i являются полюсами первого порядка функции

$$\varphi(\lambda) = \lambda + \sum_{i=1}^{k} a_i b_i^2 (b_i - \lambda)^{-1},$$

при переходе через которые (при увеличении λ) она меняет знак с плюса на минус. Поэтому из монотонности $\varphi(\lambda)$ на интервалах (b_i, b_{i+1}), $i \in [1: k-1]$, а также $(b_1, +\infty)$, $(-\infty, b_k)$ следует утверждение (6.27).

2. Для доказательства (6.28)–(6.30) достаточно доказать, что для $\lambda_i \neq 0$

$$\underline{\lambda} \le \lambda_i \le \overline{\lambda},$$

где

$$\overline{\lambda} = \begin{cases} b_1, b_1 < 0, & (6.33) \\ b_1(1+s), b_1 > 0; & (6.34) \end{cases}$$

$$\underline{\lambda} = \begin{cases} b_k, b_k > 0, & (6.35) \\ b_k(1+s), b_k < 0. & (6.36) \end{cases}$$

Пусть

$$\varphi_1(\lambda) = \sum_{i=1}^{k} a_i b_i - \lambda - \sum_{i=1}^{k} a_i b_i^2 (b_i - \lambda)^{-1}; \quad (6.37)$$

$$\varphi_2(\lambda) = b_1 \sum_{i=1}^{k} a_i - \lambda - b_1^2 \sum_{i=1}^{k} a_i (b_1 - \lambda)^{-1}; \quad (6.38)$$

$$\varphi_3(\lambda) = b_k \sum_{i=1}^{k} a_i - \lambda - b_k^2 \sum_{i=1}^{k} a_i (b_k - \lambda)^{-1}; \quad (6.39)$$

$$\varphi_4(\lambda) = \sum_{i=2}^{k} a_i b_i - \lambda - \sum_{i=2}^{k} a_i b_i^2 (b_i - \lambda)^{-1}. \quad (6.40)$$

Тогда

$$\varphi_2 - \varphi_1 = \sum_{i=1}^{k} a_i (b_1 - b_i) + \sum_{i=1}^{k} a_i \left(\frac{b_1^2}{\lambda - b_1} - \frac{b_i^2}{\lambda - b_i} \right).$$

Для $\lambda > b_1 > 0$ имеем $\varphi_2' < 0$, $\varphi_1' < 0$. Поэтому φ_1, φ_2 строго монотонны на интервале $(b_1, +\infty)$ и, согласно выражениям (6.37), (6.38), меняются от $-\infty$ до $+\infty$. Следовательно, в указанном интервале существуют корни λ_{φ_1}, λ_{φ_2} функций φ_1, φ_2, причем из неравенства $\varphi_2 - \varphi_1 > 0$ получим

$$\lambda_{\varphi_1} < \lambda_{\varphi_2} = b_1(1+s).$$

Соотношение (6.34) доказано. (6.33) следует из (6.27), так как в этом случае $\lambda_1 = 0$.

При $b_k < 0$ разность

$$\varphi_1 - \varphi_3 = \sum_{i=1}^{k} a_i \left(b_i + \frac{b_i^2}{\lambda - b_i} - b_k - \frac{b_k^2}{\lambda - b_k} \right) = \sum_{i=1}^{k} \frac{\lambda^2 a_i}{b_k - \lambda} \frac{b_i - b_k}{b_i - \lambda}$$

строго положительна на интервале $\lambda \in (-\infty, b_k)$ и $\lambda_{\varphi_3} = b_k(1+s) < \lambda_{\varphi_1} = \lambda_{k+1} < \lambda_i$, $i \in [1:k]$. Доказана справедливость утверждения (6.36). Соотношение (6.35) следует из (6.27), так как $\lambda_{k+1} = 0$.

3. Докажем третье утверждение леммы. Вычитая (6.37) из (6.40), имеем

$$\varphi_4 - \varphi_1 = a_1 b_1 + \frac{a_1 b_1^2}{b_1 - \lambda} = \frac{a_1 b_1 \lambda}{b_1 - \lambda}. \tag{6.41}$$

На промежутках (b_i, b_{i+1}) для $i \in [1: k-1]$, $(-\infty, b_k)$, $(b_1, +\infty)$ имеем $\varphi_1' < 0$, $\varphi_4' < 0$, и, следовательно, φ_1, φ_4 кусочно-монотонны.

При $\lambda \in (0, b_1)$ разность $\varphi_4 - \varphi_1 > 0$, и для всех корней $\lambda_i > 0$ выполняются неравенства

$$\mu_i > \lambda_{i+1}.$$

При $\lambda < 0$ разность (6.41) отрицательна и

$$|\mu_i| > |\lambda_{i+1}|.$$

Поэтому в общем случае

$$|\mu_i| > |\lambda_{i+1}|,$$

и неравенство (6.32) доказано.

Теорема 6.3. Пусть заданы параболоиды (6.22), (6.25). Тогда:

1. Для собственных чисел матрицы G_k^D выполняются неравенства

$$\lambda_1(G_k^D) \geq \ldots \geq \lambda_{n-r-1}(G_k^D) \geq n^{-1}\sigma\left|\lambda_{n-r}(G_k^D)\right| \geq \ldots \geq n^{-1}\sigma\left|\lambda_{n-1}(G_k^D)\right|; \tag{6.42}$$

$$\lambda_1(G_k^D) \leq n\lambda_2(D); \tag{6.43}$$

$$\lambda_{n-r-1}(G_k^D) \geq \lambda_{n-r}(D); \tag{6.44}$$

2. Для «малых» по модулю собственных чисел матриц D и G_k^D выполняются неравенства

$$\left|\lambda_i(D)\right| \leq \left|\lambda_{i-1}(G_k^D)\right| \leq n\left|\lambda_{n-r+1}(D)\right|,\ i \in [n-r+1: n]; \tag{6.45}$$

3. Если $D > 0$, то $G_k^D > 0$ и минимизаторы x^*, z^* функционалов f, f_1 связаны соотношениями (6.15).

Доказательство. Принимая без ограничения общности $k = n$, получим представление элементов g_{ij} матрицы G_k^D через элементы d_{ij} матрицы D:

$$g_{ij} = d_{ij} - d_{nj}\frac{u_i^1}{u_n^1} - d_{ni}\frac{u_i^1}{u_n^1} + d_{nn}\frac{u_i^1}{(u_n^1)^2},\ i, j \in [1: n-1]. \tag{6.46}$$

При этом, очевидно, $\forall i\colon g_{in} = 0$, $\forall j\colon g_{nj} = 0$.

Матрица G_n^D является блоком $(n \times n)$-матрицы $G = \{g_{ij}\}$:

$$G = \begin{bmatrix} G_n^D & 0 \\ 0 & 0 \end{bmatrix}. \tag{6.47}$$

Введем с помощью преобразования подобия матрицу $F = T^{\mathrm{T}}GT$, где $T = [u^1, u^2, \ldots, u^n]$ — ортогональная матрица, составленная из ортонормированных собственных векторов матрицы D.

Непосредственно проверяется, что F является окаймленной диагональной матрицей вида

$$F = \begin{bmatrix} \alpha & a_2 & a_3 & \ldots & a_n \\ a_2 & \lambda_2(D) & 0 & \ldots & 0 \\ a_3 & 0 & \lambda_3(D) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_n & 0 & 0 & \ldots & \lambda_n(D) \end{bmatrix}; \alpha, a_i \in R^1, i \in [2: n].$$

Действительно, используя представление (6.46), получаем для элементов f_{ij} матрицы F:

$$f_{ij} = \lambda_i(D)\delta_{ij} - \frac{1}{u_n^1}\langle u^1, u^i\rangle \sum_{m=1}^{n} u_m^1 d_{nm} - \frac{1}{u_n^1}\langle u^1, u^j\rangle \sum_{k=1}^{n} u_k^1 \alpha_{nk} + d_{nn}(u_n^1)^{-2}\langle u^1, u^i\rangle\langle u^1, u^j\rangle,$$

где

$$\delta_{ij} = \begin{cases} 1, i = j, \\ 0, i \neq j, i, j \in [1:n]. \end{cases}$$

Отсюда имеем: $f_{ij} = 0$ при $i > 1, j > 1$ и при $i \neq j$; $f_{ij} = \lambda_i(D)$ при $i = j > 1$.

Для $i \in [2: n]$ получаем:

$$a_i = f_{i1} = -\frac{1}{u_n^1}\sum_{k=1}^{n} u_k^i d_{nk} = -\frac{1}{u_n^1} u_n^i \lambda_i(D). \qquad (6.48)$$

Из (6.47) следует, что матрица G, а следовательно, и F имеют те же собственные числа, что и G_n^D и, кроме того, нулевое собственное число. Таким образом, исследование спектра матрицы G_n^D может быть сведено к изучению относительно более простой матрицы F.

Пусть лишь s компонентов вектора $a = (a_2, \ldots, a_n)$ не равны нулю. Тогда, выполняя последовательность преобразований подобия с помощью матриц перестановок [16], затрагивающих последние $n - 1$ строк и столбцов матрицы F, получим матрицу

$$\Phi = \begin{bmatrix} \alpha & h^{\mathrm{T}} & 0 \\ h & \mathrm{diag}(\beta_i) & 0 \\ 0 & 0 & \mathrm{diag}(\gamma_i) \end{bmatrix},$$

где компоненты вектора $h = (h_1, \ldots, h_s)$ являются перестановкой ненулевых величин (6.48); $\mathrm{diag}(\beta_i)$ — порядка s; $\mathrm{diag}(\gamma_i)$ — порядка $n - s - 1$; диагональ $\mathrm{diag}(\beta_i, \gamma_i)$ матрицы Φ является перестановкой множества собственных чисел $\lambda_i(D)$, $i \in [2: n]$.

Следовательно, $k = n - s - 1$ собственных чисел $\lambda_i(F)$ совпадают с некоторыми входящими в $\operatorname{diag}(\gamma_i)$ собственными числами $\lambda_i(D)$, $i \in [2{:}\,n]$: $\lambda_{i_1}(D)$, ..., $\lambda_{i_k}(D)$; $i_m \in [2{:}\,n]$, $m \in [1{:}\,k]$, а остальные собственные числа $\lambda_i(F)$ являются собственными числами матрицы

$$Z = \begin{bmatrix} \alpha & h^{\mathrm{T}} \\ h & \operatorname{diag}(\beta_i) \end{bmatrix}, \; i \in [1{:}\,s]. \tag{6.49}$$

Характеристический полином матрицы Z равен [63]

$$\varphi(\lambda) = (\alpha - h)\prod_{i=1}^{s}(\beta_i - \lambda) - \sum_{j=1}^{s} h_j^2 \prod_{i \neq j}(\beta_i - \lambda). \tag{6.50}$$

Пусть t значений β_i различны ($t \le s$). Обозначим их как $\lambda_{m_1}(D)$, ..., $\lambda_{m_t}(D)$, а их кратности — как q_1, ..., q_t:

$$q_1 + q_2 + \ldots + q_t = s.$$

Тогда характеристическое уравнение

$$\varphi(\lambda) = 0 \tag{6.51}$$

будет иметь множитель

$$\prod_{i=1}^{t}\left[\lambda_{m_i}(D) - \lambda\right]^{q_i - 1}. \tag{6.52}$$

Следовательно, $\lambda_{m_i}(D)$ являются собственными числами матрицы Z, а значит, и матрицы F с кратностями $q_i - 1$. Поделив (6.51) на

$$\prod_{i=1}^{t}\left[\lambda_{m_i}(D) - \lambda\right]^{q_i}, \tag{6.53}$$

приходим к уравнению для оставшихся собственных чисел матрицы Z

$$\alpha - \lambda - \sum_{i=1}^{t} \eta_i^2 \left[\lambda_{m_i}(D) - \lambda\right]^{-1} = 0, \tag{6.54}$$

где η_i^2 есть сумма q_i величин h_j^2, связанных с β_i.

Согласно (6.48) имеем

$$\eta_i^2 = \frac{1}{(u_n^1)^2}\lambda_{m_i}^2(D)\sum_{j=2}^{n}{}'(u_n^j)^2;$$

штрих у знака суммы означает, что индекс j пробегает значения, соответствующие номерам собственных чисел $\lambda_i(D)$, кратных $\lambda_{m_i}(D)$, включая номер m_i.

Таким образом, показано, что $n - t - 1 = (n - 1 - s) + (s - t)$ собственных чисел $\lambda_i(F)$ совпадают с некоторыми из чисел

$$\lambda_{i_1}(D), \ldots, \lambda_{i_{n-i-1}}(D); \quad i_k \in [2{:}n],\ k \in [1{:}n-t-1];$$

остальные являются корнями уравнения

$$\alpha - \lambda - \sum_{i=1}^{t} c_{m_i}^2 \lambda_{m_i}^2(D)\left[\lambda_{m_i}(D) - \lambda\right]^{-1} = 0, \tag{6.55}$$

где

$$c_{m_i}^2 = \frac{1}{(u_n^1)^2} \sum_{j=2}^{n}{}' (u_n^j)^2. \tag{6.56}$$

Известно, что матрица F имеет нулевое собственное число, не вошедшее в группу $\{\gamma_i\}$ и поэтому сохранившееся у матрицы Z (деление на множитель (6.53) убирает лишь кратные собственные числа, и хотя бы одно нулевое собственное число останется). Подставляя $\lambda = 0$ в уравнение (6.55), получим

$$\alpha = \sum_{i=1}^{t} c_{m_i}^2 \lambda_{m_i}(D),$$

и уравнение (6.55) примет вид

$$\lambda + \sum_{i=1}^{t} c_{m_i}^2 \lambda_{m_i}^2(D)\left[\lambda_{m_i}(D) - \lambda\right]^{-1} = \sum_{i=1}^{t} c_{m_i}^2 \lambda_{m_i}(D), \tag{6.57}$$

где $m_i \neq i_k$ при $\forall i \in [1{:}\,t]$, $\forall k \in [1{:}\,n-t-1]$. Кроме того, $\lambda_{m_i}(D) \neq \lambda_{m_j}(D)$ при $i \neq j$.

Теперь для доказательства пунктов 1, 2 достаточно показать, что число больших корней уравнения (6.57) равно числу больших собственных чисел в наборе $\{\lambda_{m_i}(D)\}$, $i \in [1{:}t]$, а число малых корней равно числу малых собственных чисел в том же наборе, причем справедливы неравенства (6.45).

Пусть среди чисел $\lambda_{m_i}(D)$, $i \in [1{:}t]$, присутствует p чисел $\lambda_i(D)$, где $i \in [1{:}\,n-r]$; $0 \le p \le n-r+1$. Тогда p корней λ_i уравнения (6.57), согласно (6.27), будут заведомо большими, причем максимальный корень, согласно второму утверждению леммы, будет не больше, чем

$$n \max_{i \in [1:t]} \lambda_{m_i}(D).$$

Покажем, что оставшиеся корни малы.

Запишем уравнение (6.57) для p больших (положительных) собственных чисел $\lambda_{m_i}(D)$:

$$\mu + \sum_{i=1}^{t-p} \frac{c_{\eta_i}^2 \lambda_{\eta_i}^2(D)}{\lambda_{\eta_i}(D) - \mu} = \sum_{i=1}^{t-p} c_{\eta_i}^2 \lambda_{\eta_i}(D), \tag{6.58}$$

где $\left|\lambda_{\eta_i}(D)\right| \le \left|\lambda_{n-r+1}(D)\right|$. При этом $t-p$ ненулевых корней уравнения (6.58), согласно третьему утверждению леммы, будет, соответственно, не меньше, чем $t-p$ ненулевых малых корней уравнения (6.57). Следовательно, левая часть неравенств (6.45) доказана. Однако в силу второго утверждения леммы эти $t-p$ корней не могут быть большими.

Действительно, согласно неравенствам (6.28)–(6.30), имеем для корней уравнения (6.58)

$$\left|\mu_j\right| \le \max_{i\in[1:t-p]}\left|\lambda_{\eta_i}(D)\right|\left(1+\sum_{i=1}^{t-p} c_{\eta_i}^2\right) \le n \max_{i\in[1:t-p]}\left|\lambda_{\eta_i}(D)\right|, \tag{6.59}$$

что доказывает правую часть неравенств (6.45).

Поясним последнее неравенство в (6.59). Согласно (6.56) получим

$$\sum_{i=1}^{t-p} c_{\eta_i} \le (u_n^1)^{-2}\sum_{i=2}^{n}(u_n^i)^2, \tag{6.60}$$

так как если $m_i \neq m_j$, то $c_{m_i}^2$ и $c_{m_j}^2$ не содержат одинаковых слагаемых. Следовательно,

$$1+\sum_{i=1}^{t-p} c_{\eta_i}^2 \le (u_n^1)^{-2}\sum_{i=1}^{n}(u_n^i)^2 = (u_n^1)^{-2} \le n, \tag{6.61}$$

так как для максимального по модулю компонента нормированного к единице вектора u^1, очевидно, выполняется неравенство

$$\left|u_n^1\right| > \frac{1}{\sqrt{n}}.$$

Из (6.61) следует (6.59).

Из рассмотрения уравнения (6.57) получили, что p корней λ_i будут заведомо большими, что эквивалентно (6.44):

$$\lambda_{n-r-1}(G_k^D) > \lambda_{n-r}(D).$$

А так как имеем по условию, что

$$\lambda_{n-r}(D) \ge \sigma\left|\lambda_{n-r+1}(D)\right|,$$

то соотношение (6.42) доказано полностью. Неравенство (6.43) очевидным образом следует из соотношений, аналогичных (6.59), записанных для уравнения (6.57).

Для доказательства третьего утверждения теоремы заметим, что x^* удовлетворяет уравнению

$$\langle x, u^1\rangle = \lambda_1(D)^{-1}\langle b, u^1\rangle,$$

поэтому x^* имеет вид

$$x^* = [L_k(y^*), y^*],$$

где y^* есть подвектор вектора x^* с исключенным k-м компонентом. Доказательство равенства $y^* = z^*$ проводится по схеме доказательства соотношения (6.15) в теореме 6.2. Теорема доказана.

6.1.4. Принцип повторных измерений

Рассмотрим *принцип повторных измерений* (ППИ) в предположении, что применяются методы оптимизации второго порядка, содержащие в своих схемах информацию о первых и вторых производных минимизируемых функционалов [49]. Сам ППИ был предложен Ю. В. Ракитским. Им же была предложена идея его реализации в задачах оптимизации.

Будем предполагать далее, что необходимые аналитические зависимости отсутствуют и объекты (матрицы, векторы), участвующие в вычислениях, получаются на основе значений $J(x^i)$ в некотором множестве точек $\{x^i\}$, определяемом конкретным методом оптимизации. Кроме того, будем считать, что результатам вычисления (или измерения) функционала $J(x)$ в точке x^i является число

$$\tilde{J}_i = J(x^i) + \delta_J J(x^i) = (1 + \delta_J) J(x^i),$$

где δ_J — относительная погрешность вычисления (измерения), которая может иметь как случайный, так и детерминированный характер.

Применение методов оптимизации второго порядка основано на часто оправдываемом предположении, что по крайней мере в окрестности оптимума минимизируемый функционал с достаточной точностью может быть аппроксимирован параболоидом

$$f(x) = 0{,}5\,\langle Dx, x\rangle - \langle b, x\rangle + \text{const}, \tag{6.62}$$

где D — симметричная положительно определенная $(n \times n)$-матрица, b — n-мерный вектор.

Применение методов ньютоновского типа, по существу, сводится к определению элементов матрицы D и вектора b и вычислению x^* из решения системы линейных алгебраических уравнений

$$Dx = b. \tag{6.63}$$

Именно эта модель применяемых методов оптимизации будет далее использована.

Если функционал (6.62) овражный, то спектральное число обусловленности $k(D) >> 1$ и уже малые деформации коэффициентов системы (6.63) могут приводить к сильным вариациям компонентов искомого вектора x^*. При $k(D) > \frac{1}{\delta}$, где δ — относительная погрешность представления указанных коэффициентов ($\delta \cong \delta_J$), задачу решения системы уравнений (6.62) следует считать некорректной из-за возможного вырождения матрицы D и нарушения требования единственности решения.

Смысл отмеченного явления некорректности можно пояснить следующим образом. Разложим векторы x^*, b по полной ортонормальной системе $\{u^i\}$ собственных векторов матрицы D:

$$x^* = \sum_{i=1}^{n} \alpha_i u^i;\ b = \sum_{i=1}^{n} \beta_i u^i.$$

Подстановка в (6.63) дает

$$\sum_{i=1}^{n} \alpha_i \lambda_i(D) u^i = \sum_{i=1}^{n} \beta_i u^i,$$

где $\lambda_i(D)$ — собственные числа матрицы D. Отсюда последовательно получаем

$$\alpha_i = \frac{\beta_i}{\lambda_i(D)}; \; x^* = \sum_{i=1}^{n} \frac{\beta_i}{\lambda_i(D)} u^i; \quad \frac{\partial x^*}{\partial \lambda_i(D)} = -\lambda_i^{-2}(D)\beta_i u^i;$$

$$\left| \frac{\langle dx^*, u^i \rangle}{\langle x^*, u^i \rangle} \right| = \Delta_i,$$

где $\Delta_i = |d\lambda_i/\lambda_i|$ — относительная ошибка в представлении i-го собственного числа матрицы D.

Из последних двух равенств следует, что решение системы (6.63) может быть сильно чувствительным к изменению минимальных по модулю собственных чисел, и даже небольшое их искажение может существенно повлиять на величину x^*.

Оказывается, чем больше величина $k(D)$ спектрального числа обусловленности, тем большие относительные искажения минимальных собственных чисел возможны за счет погрешностей Δd_{ij}. Поясним последнее утверждение.

Для норм

$$\|D\|_1 = \max_i \sum_{j=1}^{n} |d_{ij}|,$$

$$\|D\| = \max_i |\lambda_i(D)| \stackrel{\Delta}{=} \Lambda_D$$

выполняется неравенство [65]:

$$\|D\|_1 \le \sqrt{n}\|D\| = \sqrt{n}\Lambda_D.$$

Так как, очевидно, $\Lambda_D \le \|D\|_1$, то

$$\|D\| = \Lambda_D \le \max_i \sum_{j=1}^{n} |d_{ij}| \le \sqrt{n}\Lambda_D. \tag{6.64}$$

Эти неравенства показывают следующее:

1. Наличие больших по модулю собственных чисел приводит к наличию больших по модулю элементов матрицы.
2. Наличие больших элементов при заданной относительной ошибке δ их представления приводит к большим по модулю элементам матрицы ошибок $\Pi \triangleq \hat{D} - D$, где $\hat{D}$ — приближенно заданная матрица D.
3. Наличие больших элементов в Π влечет наличие больших по модулю собственных чисел у Π (матрица Π считается симметричной, так как при работе

с симметричными матрицами вычисляется лишь верхняя треугольная часть матрицы, а элементы под диагональю полагаются равными соответствующим элементам над диагональю).

Согласно известному результату теории симметричных возмущений [63], если матрица Π добавляется к D, то все собственные числа матрицы D могут изменяться на величину, лежащую между наименьшим и наибольшим собственными числами Π.

Легко привести пример, когда минимальное собственное число матрицы D изменяется при этом весьма сильно. Рассмотрим матрицу

$$D = \begin{bmatrix} L & L-\varepsilon \\ L-\varepsilon & L \end{bmatrix}, \quad |\varepsilon| << L,\ L > 0$$

и диагональную матрицу ошибок

$$\Pi = \begin{bmatrix} h & 0 \\ 0 & h \end{bmatrix}.$$

Матрица D имеет спектр:

$$\lambda_1(D) = 2L - \varepsilon \cong 2L;$$
$$\lambda_2(\Delta) = \varepsilon.$$

Собственные числа матрицы Π одинаковы и равны h, поэтому спектр $D + \Pi$ равен

$$\lambda_1(D+\Pi) = \lambda_1(D) + h \cong 2L + h;$$
$$\lambda_2(D+\Pi) = \lambda_2(D) + h \cong \varepsilon + h.$$

В результате погрешность h незначительно изменяет максимальное число, то есть $\lambda_1(D + \Pi) \cong \lambda_1(D)$, если $h << L$, и в то же время число $\lambda_2(D + \Pi)$ может не иметь ничего общего с истинным значением $\lambda_2(D)$. Например, если $h = -\varepsilon$, то матрица из невырожденной превращается в вырожденную и, в частности, об использовании метода Ньютона не может быть и речи.

После этих предварительных замечаний перейдем к более точному анализу.

Имеем

$$\max_i \sum_{j=1}^{n} |\alpha_{ij}| \le n \max_{i,j} |d_{ij}|.$$

Отсюда, учитывая (6.64), получим (см. теорему 3.5.1 из [39])

$$n^{-1}\Lambda_D \le \max_{i,j} |d_{ij}| \le \Lambda_D \tag{6.65}$$

или

$$\max_{i,j} |d_{ij}| \le \Lambda_D \le n \max_{i,j} |d_{ij}| \tag{6.66}$$

Допустим, что элемент матрицы D с максимальным модулем вычислен с относительной ошибкой δ, и его абсолютная ошибка равна максимальному по модулю элементу матрицы $\Pi = (p_{ij})$. Умножая (6.65) на δ, получим

$$\frac{\delta\Lambda_D}{n} \le \max_{i,j}\left|p_{ij}\right| \le \delta\Lambda_D.$$

Записывая неравенства (6.66) применительно к матрице Π:

$$\max_{i,j}\left|p_{ij}\right| \le \Lambda_\Pi \le n\max_{i,j}\left|p_{ij}\right|,$$

приходим к соотношению

$$\frac{\delta\Lambda_D}{n} \le \Lambda_\Pi \le n\delta\Lambda_D.$$

Как уже указывалось, на величину Λ_Π может измениться любое собственное число матрицы D, при этом

$$\frac{\delta}{n}\frac{\Lambda_D}{\left|\lambda_i(D)\right|} \le \Delta_i \le n\delta\frac{\Lambda_D}{\left|\lambda_i(D)\right|}. \qquad (6.67)$$

Согласно правому неравенству (6.67), чем меньше величина $\dfrac{\Lambda_D}{\left|\lambda_i(D)\right|}$, тем меньше возможная величина Δ_i относительного искажения числа $\lambda_i(D)$ за счет погрешностей Δd_{ij}. Согласно левому неравенству, погрешность может быть особенно большой для собственных чисел с минимальными модулями в случае

$$\Lambda_D >> \left|\lambda_i(D)\right|,$$

что и наблюдается при минимизации овражных функционалов.

Таким образом, искажения элементов матрицы D в первую очередь приводят к большим относительным ошибкам Δ_i в представлении минимальных по модулю собственных чисел, которые, в свою очередь, оказывают наиболее существенное влияние на величину компонентов вектора x^*.

Указанное противоречие и лежит в основе рассматриваемого явления некорректности.

В работе [49] предложен новый подход к решению проблемы, в результате которого удается существенно уменьшить влияние погрешности δ_J вычисления значений функционала $J(x)$ на результат, получаемый с помощью квадратичных методов оптимизации. Суть излагаемого подхода заключается в следующем.

На основе вычислений значений минимизируемого функционала в некоторых точках $\{x^j\}$ вычисляются неизвестные коэффициенты линейной системы (6.63). Как уже указывалось, решать систему (6.63) при больших $k(D)$ бессмысленно, так как x^* существенно зависит от малых собственных чисел, информация о которых при приближенном вычислении элементов матрицы D утеряна. Однако оказывается, что приближенно найденные D и b несут вполне определенную информацию об x^*, заключающуюся в выполнении определенных связей между компонентами x^*, а не в точных значениях самих компонентов. Такая точка зрения позволяет по-новому использовать ньютоновскую технику второго порядка, сводя ее к некоторой многоэтапной процедуре принятия решений. На первых этапах определяются линейные связи между компонентами x_i^*, коэффициенты

которых мало зависят от погрешностей задания D и b. Затем с помощью найденных связей исключаются некоторые компоненты вектора x, и задача минимизации исходного функционала сводится к аналогичной задаче для нового функционала, заданного в некотором подпространстве исходного пространства R^n. В результате от системы (6.63) переходим к решению линейной системы

$$B\xi = d \tag{6.68}$$

относительно некоторого подвектора ξ вектора x. Система (6.68) по построению оказывается эквивалентной заданной с точностью до использованных уравнений связи.

Основным в таком подходе является, во-первых, то, что матрица B при надлежащем построении алгоритма исключения оказывается уже хорошо обусловленной и вектор ξ^* может быть вычислен относительно точно. Во-вторых, элементы матрицы B и вектора d допускают независимое от D и b вычисление по значениям $J(x^i)$, вычисленным (измеренным) в соответствующих найденным ранее связям точках, например, на основе метода наименьших квадратов или метода конечно-разностных аппроксимаций производных вдоль составляющих вектора ξ.

Принципиально важно независимое от D и b определение B и d. Вычисление B и d непосредственно по D и b формально возможно (соответствующие связи между x и ξ найдены), но не приводит к результату, так как в D и b информация о точных значениях компонентов x^* утеряна, и восстановить ее можно лишь при повторных вычислениях $J(x)$ в специально выбранных с учетом найденных связей точках $\{x^i\}$.

Указанная идея и носит название принципа повторных измерений.

Пример. Пусть задан выпуклый квадратичный функционал вида

$$J(\beta, x) = \beta_0 + \beta_1 x_1^2 + \beta_2 x_1 x_2 + \beta_3 x_2^2 - \beta_4 x_1 - \beta_5 x_2, \tag{6.69}$$

где $\beta = (\beta_0, \beta_1, \ldots, \beta_5)$ — вектор неизвестных коэффициентов.

Требуется получить оценки $\hat{\beta}_i$ на основе измерений значений J в некоторых точках и затем определить экстремальную точку $\hat{x}^*$ из условия

$$J(\hat{\beta}, \hat{x}^*) = \min_x J(\hat{\beta}, x). \tag{6.70}$$

При этом предполагается, что измеряемые значения J искажены шумом

$$\tilde{J}(x^i) = (1 + \delta_J) J(x^i),$$

где δ_j — заданная относительная погрешность.

Приведенная постановка задачи, в частности, весьма характерна для теории планирования экстремальных экспериментов, имеющей дело с алгоритмически заданными минимизируемыми функционалами, измерения значений которых искажаются сравнительно высоким уровнем шума.

Как показывается далее, уже достаточно малая степень овражности $J(x)$ не позволяет применять традиционные методы и вынуждает обращаться к специальным методам.

Предположим, что истинные значения параметров $\bar{\beta}_i$ равны:

$$\bar{\beta}_0 = 1111,\ \bar{\beta}_1 = 989,\ \bar{\beta}_2 = 984,\ \bar{\beta}_3 = 251,\ \bar{\beta}_4 = 1978,\ \bar{\beta}_5 = 984. \qquad (6.71)$$

При этом легко проверить, что

$$x^* = (1,0),\ J(\bar{\beta}, x^*) = 122{,}0. \qquad (6.72)$$

Используемые при получении оценок $\hat{\beta}_i$ экспериментальные данные имитируются при заданных значениях параметров (6.71). Шум измерений моделируется округлением результата измерений до двух цифр мантиссы в представлении числа в форме с плавающей точкой (запятой).

В табл. 6.1 приведен близкий к D-оптимальному насыщенный план эксперимента для определения $\hat{\beta}_i$ методом наименьших квадратов; $\bar{J}_i$ — истинные значения, $\tilde{J}_i$ — измеренные значения, искаженные шумом.

Таблица 6.1

i	x_1^i	x_2^i	$\tilde{J}_i$	$\bar{J}_i$
1	−1	−1	6300	6297
2	−1	+1	2400	2361
3	+1	−1	370	373
4	+1	+1	370	373
5	0	−1	2300	2346
6	0	0	1100	1111

Функционал

$$I(\beta) = \sum_{i=1}^{6} \left[J(\beta, x^i) - \tilde{J}_i \right]^2,$$

построенный по методу наименьших квадратов, принимает минимальное значение в точке

$$\bar{\beta}_0 = 1100,\ \bar{\beta}_1 = 1035,\ \bar{\beta}_2 = 975,\ \bar{\beta}_3 = 225,\ \bar{\beta}_4 = 1990,\ \bar{\beta}_5 = 975. \qquad (6.73)$$

Сравнивая с (6.71), видим, что погрешность в определении $\hat{\beta}_i$ лежит во 2-м–3-м знаке мантиссы, что согласуется с погрешностями экспериментальных данных.

Таким образом, можно считать, что задача определения коэффициентов β_i решена вполне успешно. Однако попытка определить x^* по полученным значениям (6.73) из системы

$$\frac{\partial J(\hat{\beta}, x)}{\partial x_i} = 0,\ i = 1,2,$$

приводит к неверному результату: $\hat{x}^* = (2{,}9; -4{,}0)$, $J(\bar{\beta}, \hat{x}^*) = 231{,}0$ вместо точного результата (6.72). Причина неудачи заключается в овражном характере $J(x)$. Если представить выражение (6.69) с параметрами (6.71) в виде

$$J = 0{,}5\langle Dx, x\rangle - \langle b, x\rangle + c, \tag{6.74}$$

получим

$$D = \begin{bmatrix} 1978 & 984 \\ 984 & 502 \end{bmatrix}, \; b = (1978, 984), \; c = 1111.$$

Собственные числа матрицы D равны

$$\lambda_1 = 2470, \; \lambda_2 = 10.$$

Соответствующие собственные векторы имеют вид

$$u^1 = \frac{2}{\sqrt{5}}(1; 0{,}5), \; u^2 = \frac{2}{\sqrt{5}}(0{,}5; -1).$$

Линии уровня функционала (6.74) изображены на рис. 6.1. Овраг направлен параллельно вектору u^2. Кружками обозначены точки плана эксперимента, использованные при получении оценок (6.73).

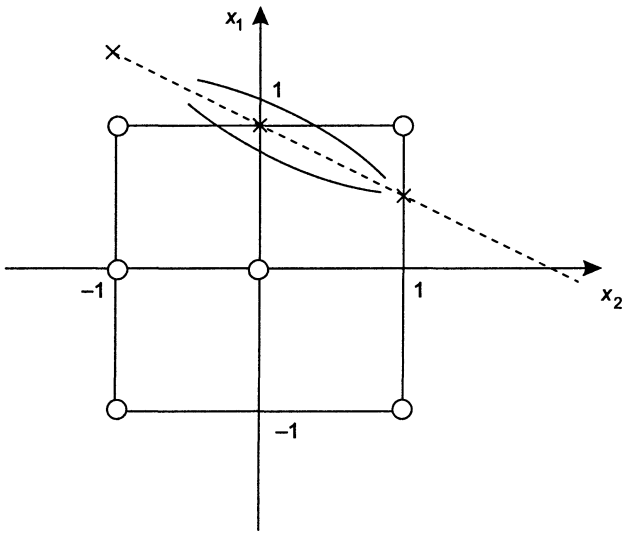

Рис. 6.1. План эксперимента

Полученная точка $\hat{x}^*$ существенно смещена вдоль образующей дна оврага по сравнению с истинным положением x^*.

Применим для решения этой же задачи метод иерархической оптимизации. С помощью оценок (6.73) не удалось получить приемлемое приближение к вектору x^*, однако построенное на их основе уравнение связи

$$x_1 = 0{,}968 - 0{,}469x_2 \cong 1 - 0{,}5x_2 \tag{6.75}$$

достаточно точно локализует дно оврага. При получении (6.75) было использовано первое уравнение системы вида (6.7).

Согласно общей методике, исключаем переменную x_1 и переходим к новому функционалу вида

$$J_1(c, x_2) = c_0 + c_1 x_2 + c_2 x_2^2. \tag{6.76}$$

В результате задача свелась к задаче одномерной минимизации по переменной x_2, а соответствующие значения x_1 вычисляются по уравнению (6.75).

На следующем этапе производим независимую от $\hat{\beta}_i$ оценку значений параметров c_i. Используется план эксперимента

$$x_2^1 = 0,\ x_2^2 = 1,\ x_2^3 = -1. \tag{6.77}$$

Для вычисления значений функционала (6.76) в точках плана (6.77) вычисляются соответствующие значения x_1^i по уравнению (6.75). Получается план для исходного функционала, представленный в табл. 6.2, точки которого расположены вдоль оврага (обозначены крестиками на рис. 6.1).

Таблица 6.2

i	x_1^i	x_2^i	$\tilde{J}_i$	$\bar{J}_i$
1	0,968	0	120	123
2	0,499	1	130	128,25
3	1,437	−1	130	131,51

Вычисление значений функционала J в точках плана табл. 6.2 как раз и реализует принцип повторных измерений. На основе этой дополнительной серии измерений поступает дополнительная информация, позволяющая более точно локализовать точку минимума.

Метод наименьших квадратов приводит к оценкам

$$\hat{c}_0 = 120,\ \hat{c}_1 = 0,\ \hat{c}_2 = 10$$

и к точке минимума $x_2^* = 0$ функционала $J(\hat{c}, x_2)$.

Согласно (6.75)

$$x_1^* = 0{,}968 - 0{,}469x_2^* = 0{,}968.$$

Таким образом, получен вектор

$$\hat{x}^* = (0{,}968;\ 0),\quad J(\bar{\beta}, \hat{x}^*) = 123{,}64,$$

близкий к истинному решению.

Погрешность в определении оптимальных значений компонентов x^*, как и у экспериментальных значений, лежит во втором знаке мантиссы. Задача решена.

6.1.5. Алгоритмы иерархической оптимизации

Пусть $J(x) \in C^2(R^n)$ — минимизируемый функционал. В качестве независимой переменной используем вектор приращения к начальной точке $\bar{x}$, так что исходная задача минимизации $J(x)$ сводится к эквивалентной для функционала $I(y)$, где

$$I(y) \triangleq J(\bar{x} + y). \tag{6.78}$$

Пусть $f(y)$ — параболоид, аппроксимирующий функционал $I(y)$ в окрестности точки $y = 0$:

$$f(y) = 0{,}5\langle Ay, y\rangle - \langle d, y\rangle + c, \tag{6.79}$$

где

$$A = f''(0) = J''(\bar{x}) = I''(0);$$
$$d = -f'(0) = -J'(\bar{x}) = -I'(0).$$

Дифференциальное уравнение спуска (6.2) для параболоида (6.79) и уравнение асимптотической связи между компонентами вектора y примут соответственно вид

$$\frac{dy}{dt} = -Ay + d;$$
$$\left\langle y, u^1 \right\rangle = \lambda_1^{-1}(A)\left\langle d, u^1 \right\rangle. \tag{6.80}$$

Коэффициенты уравнения (6.80) могут вычисляться непосредственно по A и d на основе любой численной процедуры, позволяющей построить максимальное по модулю собственное число и отвечающий ему собственный вектор. Это эквивалентно применению ПКП:

$$\frac{d^k y}{dt^k} = 0,$$

или

$$A^k y = A^{k-1} d. \tag{6.81}$$

при надлежащем задании параметра k.

Укрупненная вычислительная схема метода иерархической оптимизации сводится к следующей последовательности шагов.

Алгоритм МИО.

Шаг 0. Принять:

$$y^0 := y;\ I_0(y^0) \triangleq I(y);\ n_0 := n;\ y^0 = (y_1^0, \ldots, y_{n_0}^0);\ l := 1.$$

Шаг 1. С помощью конечно-разностных аппроксимаций производных в пространстве переменных y^{l-1} вычислить:

$$g^{l-1} = I'_{l-1}(0),\ G_{l-1} = I''_{l-1}(0).$$

Шаг 2. Вычислить максимальное по модулю собственное число λ_1^l матрицы G_{l-1} и отвечающий ему собственный вектор $u^{l,1}$; принять

$$k_l := \arg\max_j \left| u_j^{l,1} \right|.$$

Шаг 3. Из соотношения (6.80) выразить k_l-й компонент вектора $y^{l-1} = (y_1^{l-1}, \ldots, y_{n_{l-1}}^{l-1})$:

$$y_{k_l}^{l-1} = L^l(y^l) \triangleq -(u_{k_l}^{l,1})^{-1}\left[(\lambda_1^l)^{-1}\left\langle g^{l-1}, u^{l,1}\right\rangle - \sum_{i=1}^{n_{l-1}} y_i^{l-1} u_i^{l,1}\right],\ i \neq k_l, \qquad (6.82)$$

где

$$y^l \triangleq (y_1^{l-1}, \ldots, y_{k_l-1}^{l-1}, y_{k_l+1}^{l-1}, \ldots, y_{n_{l-1}}^{l-1}) = (y_1^l, \ldots, y_{n_l}^l);\ n_l := n_l - 1.$$

Шаг 4. Принять

$$I_l(y^l) \triangleq I_{l-1}([L_{k_l}(y^l), y^l]),$$

где $[L_{k_l}(y^l), y^l] \triangleq (y_1^l, \ldots, y_{k_l-1}^l, L^l(y^l), y_{k_l}^l, \ldots, y_{n_l}^l)$.

Шаг 5. Принять $l := l + 1$. Если $l \le n - r$, перейти к шагу 1, иначе — к шагу 6.

Шаг 6. Реализовать процедуру минимизации функционала

$$I_{n-r}(y^{n-r}), y^{n-r} = (y_1^{n-r}, \ldots, y_{n_{n-r}}^{n-r}),\ n_{n-r} = r,$$

где

$$I_{n-r}(y^{n-r}) = I_0(\tilde{y}) = I(\tilde{y});$$
$$\tilde{y} = [L_k(y^1), y^1];$$
$$y^1 = [L_{k_2}(y^2), y^2],$$

.....

$$y^{n-r+1} = [L_{k_{n-r}}(y^{n-r}), y^{n-r}].$$

(Согласно теореме 6.3, минимизация функционала I_{n-r} будет протекать в условиях существенно меньшей степени овражности, чем минимизация исходного функционала I_0).

Шаг 7. По найденному вектору y^{n-r} размерности r восстановить полный вектор y размерности n; принять $\bar{x} := \bar{x} + y$ и снова перейти к шагу 0.

Как обычно, условие окончания процесса оптимизации здесь не рассматривается.

Опыт практической работы позволяет сделать следующие замечания.

Замечание 1. Алгоритм MIO оказывается неэффективным, если исходный функционал является существенно неквадратичным и область справедливости локальной квадратичной модели невелика. В этом случае целесообразно непосредственно обращаться к процедурам ОПС, не использующим в своей схеме в явном виде предположений о квадратичном характере минимизируемого функционала и сохраняющим возможность достижения любой точки пространства поиска.

Замечание 2. Алгоритм MIO целесообразно включать в работу в точках замедления маршевой поисковой процедуры (на дне оврага) как средство ускорения сходимости и преодоления участков сложного рельефа с критическими значениями степеней овражности минимизируемых функционалов.

Замечание 3. Основная особенность процедуры MIO по сравнению с методами ОПС, в принципе, также реализующими идею локальной декомпозиции задачи минимизации, заключается в следующем. В отличие от методов ОПС, в алгоритме MIO для минимизации функционала $I_{n-r}(y^{n-r})$ может применяться любая достаточно эффективная в рассматриваемых условиях процедура (например, RELEX), имеющая приемлемую скорость сходимости при наличии высокой остаточной степени овражности. В то же время в методах ОПС минимизация в пределах линейного многообразия, определяющего дно оврага, может производиться только методами покоординатного спуска, имеющими при больших ошибках в определении «малых» собственных векторов (при сохранении точности задания самого подпространства) относительно малую скорость сходимости.

Сделанные замечания позволяют утверждать, что наиболее рациональная стратегия параметрической оптимизации должна основываться не на каком-то одном подходе, а на целом спектре процедур (например, типа RELEX, SPAC1, SPAC5, MIO), вызываемых некоторой мониторной системой автоматически либо в интерактивном режиме в зависимости от получаемых результатов в процессе решения конкретной задачи.

Дополнительная особенность процедуры MIO заключается в ее инвариантности относительно характера выпуклости минимизируемого функционала. Более точное в результате реализации ППИ измерение малых спектральных составляющих матрицы Гессе позволяет в этом случае более эффективно строить направления наискорейшего убывания функционала.

6.2. Методы исключения переменных на основе спектрального разложения матрицы Гессе

6.2.1. Постановка задачи

Пусть решается задача

$$J(x) \to \min_{x \in D},\ D \subset R^n = \{x \mid x = (x_1, x_2, \ldots, x_n)\},\ J: R^n \to R,\ J \in C^2(R^n). \quad (6.83)$$

Будем далее предполагать, что исходная условная задача оптимизации одним из известных методов сведена к безусловной, так что $D = R^n$. В частности, наиболее распространенные прямые ограничения вида

$$a_i \le x_i \le b_i$$

или

$$x_i \ge 0$$

могут быть устранены переходом к новым переменным y_i, где

$$x_i = b_i + (a_i - b_i)\sin^2 y_i \text{ или } x_i = y_i^2.$$

Пусть μ — минимизатор функционала $J(x)$. Основная задача формулируется следующим образом. Требуется построить такой вектор z, чтобы вектор $\mu' = \mu + z$ имел максимальное число компонентов, совпадающих с заданным вектором d (как правило, $d = 0$), и чтобы при этом

$$\Delta J \triangleq J(\mu') - J(\mu) \le \delta J, \tag{6.84}$$

где δJ характеризует заданное допустимое абсолютное отклонение значения функционала от $J(\mu)$.

Обычно применяемый подход к решению сформулированной задачи (в предположении $d = 0$) заключается в исключении достаточно малых компонентов минимизатора μ.

Если поверхности уровня функционала $J(x)$ не обладают овражной структурой и близки к сферам, то подобный метод оказывается правомерным. Однако в более реальных условиях овражной ситуации он может приводить к неоправданному ухудшению качества системы; кроме того, такой способ неприемлем, если, например, компоненты минимизатора приблизительно равны по величине. Легко видеть, что с помощью простого масштабирования можно добиться любых соотношений между компонентами вектора μ, поэтому уже с этих позиций ясно, что полагаться на «малость» некоторых параметров μ_i не всегда целесообразно. Это замечание иллюстрируется рис. 6.2, *а*.

Компонент μ_1 больше μ_2, однако из рисунка видно, что исключаться должна не переменная x_2, а переменная x_1, так как эллипсоидная зона допустимых значений функционала, являясь сильно вытянутой, пересекает ось x_2, не имея общих точек с осью x_1. Исключение же x_2 неизбежно приведет к превышению допустимого отклонения δJ даже при последующей «доводке» оставшегося компонента x_1. На рис. 6.2, *б* изображена ситуация, когда $\mu_1 \cong \mu_2$ и с точки зрения обычного подхода вообще неясно, какая переменная должна исключаться.

Далее предлагается регулярный метод решения проблемы, инвариантный относительно выбора масштабов оптимизируемых параметров и степени овражности функционала (6.83). В представленных на рис. 6.2 ситуациях этот метод приводит к выбору одной из точек отрезка $[c, d]$, то есть к исключению, соответствен-

но, переменных x_1, x_2 и необходимому пересчету оставшихся компонентов. Превышение допустимого уровня δJ при этом, очевидно, не происходит.

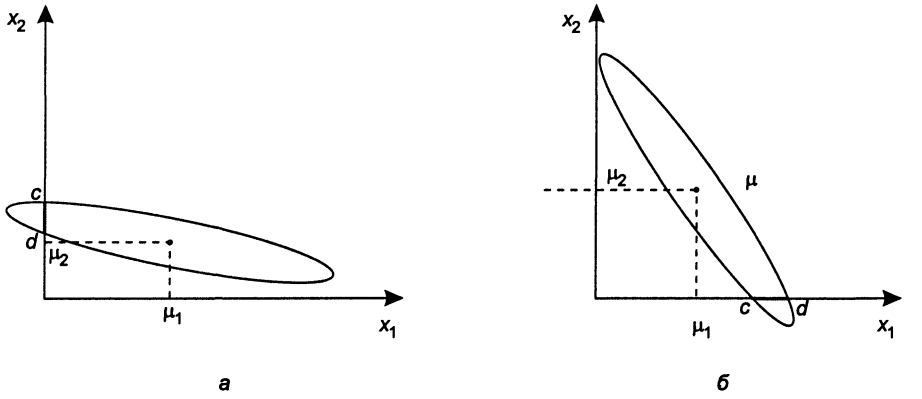

Рис. 6.2. Исключение переменных:

а — исключается переменная x_1, *б* — исключается переменная x_2

В принципе, решение рассматриваемой проблемы может быть получено последовательным решением задач типа (6.83) с различными наборами $\{x_{i_1}, ..., x_{i_k}\}$ управляемых параметров. При этом должны варьировать как число параметров, так и состав управляемого вектора. Однако такой подход приводит к нереальным вычислительным затратам для большинства практических ситуаций. Основная задача заключается в разработке метода, позволяющего указать необходимый состав управляемого вектора без многократного решения задач типа (6.83) и при минимальном числе обращений к вычислению значений $J(x)$. Последнее оказывается особенно важным, например, при оптимизации сложных систем, требующих для однократного вычисления значения критерия качества реализации достаточно трудоемкого процесса анализа функционирования.

При правильном выборе показателя качества $J(x)$ можно предположить, что в окрестности оптимума μ он достаточно точно аппроксимируется своей квадратичной моделью

$$J(x) \cong 0{,}5\langle Ax, x\rangle - \langle b, x\rangle + \text{const}, \qquad (6.85)$$

где $A = J''(\mu)$ — симметричная положительно определенная $(n \times n)$-матрица вторых производных, $b \in R^n$.

Овражный характер $J(x)$ выражается в выполнении следующих неравенств для собственных значений матрицы A, упорядоченных по величине модулей

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{n-r} >> \lambda_{n-r+1} \geq \ldots \geq \lambda_n.$$

Предлагаемый метод основан на решении полной проблемы собственных значений для матрицы A. Известно, что «малые» собственные значения $\lambda_{n-r+1}, \ldots, \lambda_n$ будут при этом получены с катастрофическими погрешностями и не будут содержать полезной информации. Известно также [63], что линейные оболоч-

ки M_1, M_2, натянутые, соответственно, на вычисленные собственные векторы $\{u^1, \ldots, u^{n-r}\}$, $\{u^{n-r+1}, \ldots, u^n\}$ будут практически совпадать с истинными. Последнее замечание оказывается решающим при реализации излагаемого ниже метода.

Основная идея метода состоит в следующем.

Предлагается искать вектор z в виде линейной комбинации векторов $u^{n-r+1}, \ldots, u^n$, задающих «дно» оврага M_2, где норма вектора градиента $\| J'(x) \|$ определяется малыми собственными значениями $\lambda_{n-r+1}, \ldots, \lambda_n$ и обычно оказывается существенно меньше, чем в остальной части пространства. Таким образом, $z = vc$, где $v = [u^{n-r+1}, \ldots, u^n]$ — матрица полного ранга размерности $n \times r$; $c = (c_1, \ldots, c_r)$ — вектор неизвестных коэффициентов, удовлетворяющий соотношению $\| c^2 \| \le \delta$. Значение δ вычисляется из условия (6.84).

Далее излагается построение алгоритма, реализующего эту идею. Мы здесь нигде не утверждаем ничего большего. Более того, легко привести пример, когда такой подход будет заведомо неэффективен. Однако, как показывает вычислительная практика, излагаемый метод часто дает желаемые результаты. Логика здесь крайне проста. Действительно, любое удаление переменных связано со сдвигом точки оптимума в пространстве аргументов до пересечения с соответствующими координатными осями и плоскостями. Так вот, все, что здесь обсуждается, — это попытка осуществлять эти сдвиги вдоль дна оврага, так как при этом целевой функционал изменяется наименее интенсивно и можно ожидать не слишком значительных ухудшений его значений по сравнению с точным оптимумом в полном пространстве.

Согласно (6.85) имеем

$$\Delta J = \langle J'(\mu), vc \rangle + 0{,}5 \langle Avc, vc \rangle = 0{,}5 \langle Avc, vc \rangle = 0{,}5 \sum_{i=n-r+1}^{n} c_{i-n+r}^2 \lambda_i, \; J'(\mu) = 0. \quad (6.86)$$

Полагая $|\lambda_i| \le \alpha$, $i \in [n - r + 1: n]$, получим $\Delta J \le 0{,}5\alpha \| c^2 \| \le \delta J$, откуда

$$\| c \|^2 = \sum_{i=1}^{r} c_i^2 \le \delta, \; \delta \overset{\Delta}{=} \frac{2\delta J}{\alpha}. \quad (6.87)$$

Введенная величина α формализует понятие малого собственного значения и в разных случаях может выбираться из условий, определяемых типом решаемой задачи.

Согласно (6.67) для погрешностей задания собственных значений при ограниченной разрядности применяемых компьютеров выполняются неравенства $|d\lambda_i| \le n\lambda_1 \varepsilon_м$, где $\varepsilon_м$ — относительная точность представления чисел в системе с плавающей точкой; $d\lambda_i$ — абсолютная погрешность представления i-го собственного значения. Поэтому в любом случае необходимо выбирать $\alpha \ge n\lambda_1 \varepsilon_м$, и если в результате решения задачи на собственные значения для некоторых из вычисленных значений выполняются неравенства

$$|\lambda_i| \le n\lambda_1 \varepsilon_м, \quad (6.88)$$

то соответствующие собственные значения целесообразно классифицировать как «малые», а отвечающие им собственные векторы включать в группу $\{u^{n-r+1}, \ldots, u^n\}$.

Можно назначать величину α из соображений, не связанных непосредственно с длиной разрядной сетки компьютера. При этом следует иметь в виду, что увеличение α с целью увеличения размерности оболочки M_2, а тем самым и количества исключаемых переменных будет приводить к убыванию значения $\delta = 2\delta J/\alpha$, что ограничивает наши возможности по увеличению нормы $\| c \|$, характеризующей длину вектора продвижения $\mu' - \mu$. Таким образом, при заданной величине δJ за счет выбора α мы сможем продвинуться далеко (и исключить компоненты вектора μ, существенно отличающиеся от компонентов вектора d), но вдоль ограниченного числа направлений, либо продвинуться по многим направлениям, но на ограниченное расстояние. В каждой конкретной задаче необходим выбор разумного компромисса. Целесообразно проводить пошаговое уменьшение величины α от некоторого большого значения порядка $\lambda_1/2$ до значения $n\lambda_1\varepsilon_м$ с реализацией на каждом шаге излагаемого далее базового алгоритма. Возможны и другие стратегии выбора α, например основанные на интерактивном режиме работы в соответствующей вычислительной системе.

Прямой путь решения сформулированной задачи чрезвычайно прост и заключается в решении всех возможных подсистем размерности $r \times r$ переопределенной, обычно несовместной, системы линейных уравнений вида

$$vc = f, \; f \triangleq d - \mu \tag{6.89}$$

и выбора решения с минимальной нормой $\| c \|$. Однако этот способ практически нереализуем из-за своей трудоемкости. Так, например, уже для $n = 50$, $r = 25$ потребуется решить около 10^{14} систем линейных уравнений, содержащих 25 переменных.

Не приводят в данном случае к цели и стандартные методы решения переопределенных систем, использующие идею псевдообращения. Они основаны на выборе вектора c, минимизирующего среднюю ошибку для всех n уравнений, тогда как нам необходимо отыскать вектор c из части системы, игнорируя остальные $n - r$ уравнений.

6.2.2. Алгоритм исключения

Для приближенного определения оптимальной в смысле величины нормы $\| c \|$ подсистемы (6.89) предлагается воспользоваться геометрически очевидной идеей проецирования точки $f \in R^n$ вдоль линейной оболочки M_2 на ближайшие подпространства R^l, $l = n - 1, \ldots, n - r$. Данный процесс по смыслу и по трудоемкости аналогичен гауссовскому методу исключения переменных со специальным выбором ведущего элемента.

Пусть $V = \{v_{ij}\}$, $f = \{f_i\}$, тогда укрупненная схема соответствующего алгоритма может быть представлена с помощью следующей последовательности действий.

Каждое из уравнений системы (6.89) имеет вид

$$\langle w, c\rangle = t, \tag{6.90}$$

где w — вектор коэффициентов, $c = (c_1, \ldots, c_r)$.

На первом шаге выберем уравнение, для которого норма вектора c, определяющего нормальное решение уравнения (6.90) с компонентами $c_i = tw_i / \|w\|^2$, $i \in [1:r]$, имеет минимальное значение. Разрешим данное уравнение относительно переменной c_i с наибольшим по модулю коэффициентом, полагая, без ограничения общности, что выбраны первое уравнение и первая переменная c_1. Подставляя выражение для c_1 в остальные уравнения, приходим к системе

$$\begin{aligned} &c_1 = L_1(c_2, \ldots, c_r),\\ &v'_{21}f_1 + v'_{22}c_2 + \ldots + v'_{2r}c_r = f_2;\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &v'_{n1}f_1 + v'_{n2}c_2 + \ldots + v'_{nr}c_r = f_n, \end{aligned} \tag{6.91}$$

где $v'_{i1} = \dfrac{v_{i1}}{v_{11}}$, $v'_{ij} = v_{ij} - \dfrac{v_{i1}v_{1j}}{v_{11}}$, $i \in [2:n]$, $j \in [2:r]$.

Каждое из уравнений системы (6.91), начиная со второго, также имеет вид (6.90) $\langle \bar{w}, \bar{c}\rangle = \bar{t}$, где $\bar{c} = (c_2, \ldots, c_r)$. На втором шаге выбирается уравнение, для которого норма вектора c с компонентами $c_1 = L_1(c_2, \ldots, c_r)$, $c_j = \dfrac{\bar{t}\bar{w}_j}{\|\bar{w}\|^2}$ имеет минимальное значение. Аналогично предыдущему получаем линейную связь $c_2 = L_2(c_3, \ldots, c_r)$, что позволяет исключить c_2 из оставшихся уравнений. Далее процесс повторяется до получения явного выражения для c_r. Реализуя обратные подстановки на основе иерархии линейных соотношений L_i, восстанавливаем полный вектор c, доставляющий искомое решение задачи. Номера исключенных компонентов управляемого вектора x при этом определяются номерами выбираемых в описанном процессе уравнений.

Для выполнимости сформулированного алгоритма существенно, что если уже исключены переменные $c_1, \ldots, c_k$, $k < r$, то всегда найдется уравнение вида $\beta_{k+1}c_{k+1} + \ldots + \beta_r c_r = t_k$, где не все $\beta_i = 0$, и, следовательно, процесс исключения может быть продолжен. Справедливость данного замечания следует из теории гауссовского исключения для прямоугольных матриц [56]. А именно, если прямоугольная матрица v размерности $n \times r$ приведена с помощью элементарных преобразований и перестановок строк и столбцов к матрице ступенчатого вида, то ранг матрицы v равен числу ненулевых ведущих элементов. Если предположить, что утверждение неверно, то отсюда следует, что (ранг v) $\le k < r$, а это противоречит условиям решения задачи.

Легко установить, что номера уравнений с нулевыми коэффициентами при всех c_i являются номерами, отвечающими уже выбранным уравнениям.

Если предположение о квадратичном характере зависимости $J(x)$ в окрестности оптимума выполняется с недостаточной точностью, то значения оставшихся после исключения переменных должны быть уточнены с помощью повторного од-

нократного решения задачи (6.83) с вычисленными на этапе исключения начальными условиями.

Далее приводятся два примера, иллюстрирующие работу алгоритма. Первый из них является модельным и позволяет детально проследить основные этапы проводимых вычислений. Второй связан с решением реальной задачи структурного синтеза.

Пример 1. Решалась задача минимизации квадратичного функционала $J(x)$ с $D = R^4$, в результате был получен вектор $\mu = -(10, 10, 10, 10), J(\mu) = 1{,}621 \cdot 10^{-1}$. При заданном допустимом абсолютном отклонении $\delta J = 10^{-2}$ требовалось изучить возможность исключения максимального числа параметров. Задача (6.83) решалась методом ОПС на компьютере с $\varepsilon_м \cong 10^{-8}$, что позволило получить необходимую для проведения процесса исключения информацию непосредственно в процессе оптимизации. Вычисленные собственные значения матрицы $J''(\mu)$ равны: $\lambda_1 = 143{,}12$; $\lambda_2 = 87{,}633$; $\lambda_3 = 4{,}141 \cdot 10^{-7}$; $\lambda_4 = -3{,}713 \cdot 10^{-6}$.

Согласно (6.88), положим $\alpha \cong 10^{-5}$, и поэтому значения λ_3, λ_4 являются «малыми» (λ_4 вычислено с неверным знаком). Отвечающие λ_3, λ_4 собственные векторы v^1, v^2 равны:

$$v^1 = 10^{-1}(1{,}400280;\ 5{,}601120;\ 7{,}001400;\ 4{,}200840)^{\mathrm{T}};$$

$$v^2 = 10^{-1}(-\,8{,}980265;\ 1{,}796053;\ -1{,}796053;\ 3{,}592106)^{\mathrm{T}}.$$

Полагая $r = 2$, решали задачу исключения двух переменных — по числу малых собственных значений. Система двух уравнений (6.89) с двумя неизвестными c_1, c_2 при $d_i = 0$ имеет вид

$$c_1 v^1 + c_2 v^2 = 10\,(1;\ 1;\ 1;\ 1)^{\mathrm{T}}. \tag{6.92}$$

На первом шаге, согласно базовому алгоритму, вычислялись нормы $\| tw/\| w \|^2 \| = |t| / \| w \|$ для всех четырех уравнений (6.92). Получены значения $\| \cdot \|_1 = 1{,}100257$; $\| \cdot \|_2 = 1{,}700091$; $\| \cdot \|_3 = 1{,}383489$; $\| \cdot \|_4 = 1{,}809224$, что позволило выбрать первое уравнение

$$1{,}400280 \cdot 10^{-1} c_1 - 8{,}980265 \cdot 10^{-1} c_2 = 10.$$

Отсюда имеем линейную связь L_1:

$$c_2 = 1{,}559285 \cdot 10^{-1} c_1 - 11{,}13552. \tag{6.93}$$

Подставляя (6.93) в уравнения 2–4 системы (6.92), получим на втором шаге

$$\begin{aligned} 5{,}881175 c_1 &= 120{,}0; \\ 6{,}721344 c_1 &= 80{,}0; \\ 4{,}760951 c_1 &= 140{,}0. \end{aligned} \tag{6.94}$$

С учетом (6.93) имеем для трех случаев (6.94)

$$c_1 = 20{,}40408,\ c_2 = -7{,}953943;$$

$$c_1 = 11{,}90238,\ c_2 = -9{,}279600; \qquad (6.95)$$

$$c_1 = 29{,}40588,\ c_2 = -6{,}550306.$$

Наименьшее значение нормы $\|c\|$ достигается во втором уравнении (6.95), что соответствует третьему уравнению исходной системы (6.92). В результате установлено, что исключению подлежат компоненты μ_1, μ_3 вектора μ. Новый вектор μ' находится из равенства

$$\mu' = \mu + c_1 v^1 + c_2 v^2 = -10(1;\ 1;\ 1;\ 1)^{\mathrm{T}} + 11{,}90238\cdot 10^{-1}(1{,}400280;\ 5.601120;\ 7{,}001400;\ 4{,}200840)^{\mathrm{T}} - 9{,}279600\cdot 10^{-1}(-8{,}980265;\ 1{,}796053;\ -1{,}79053;\ 3{,}592106)^{\mathrm{T}} \cong$$
$$\cong (0{,}0;\ -5{,}00;\ 0{,}0;\ -8{,}33)^{\mathrm{T}}.$$

При этом, согласно (6.87), имеем оценку

$$\Delta J = J(\mu') - J(\mu) \le \frac{1}{2}\|c\|^2 \alpha \le 2\cdot 10^{-3},$$

что удовлетворяет условию (6.84) допустимости отклонения. Задача решена. В данном случае простое отбрасывание любой из переменных μ_i приводило к многократному превышению заданного значения δJ.

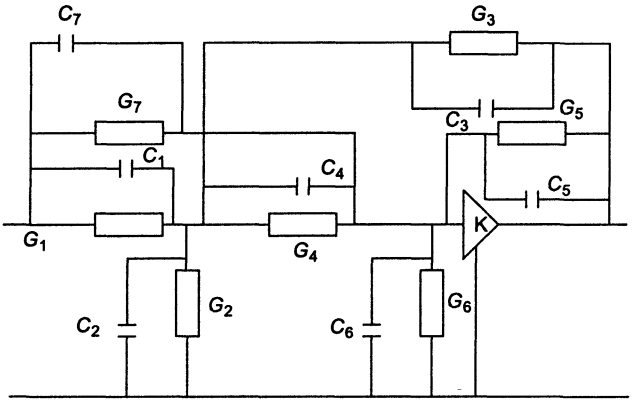

Рис. 6.3. Исходная схема с избыточной структурой

Пример 2. На рис. 6.3 представлена ARC-цепь с пятиузловой *RC*-схемой, имеющей избыточную структуру [33]. Решалась задача реализации передаточной функции звена фильтра высокой частоты $T(p) = \dfrac{p^2}{1 + 0{,}01p + p^2}$ с использованием минимального числа пассивных компонентов. Минимизируемый функционал в данном случае имеет вид

$$J(x) = f_1^2 + f_2^2 + (f_3 - 1)^2 + (f_4 - 1)^2 + (100 f_5 - 1)^2 + (f_6 - 1)^2, \qquad (6.96)$$

где $x = (x_1, \ldots, x_{15})$; $y = (G_1, \ldots, G_7, c_1, \ldots, c_7, k)$; $y_i = x_i^2$, $i \in [1{:}15]$; $f_i(x)$ — известные функции, определяющие коэффициенты передаточной функции

$$T(p) = \frac{f_1 + f_2 p + f_3 p^2}{f_4 + f_5 p + f_6 p^2}.$$

Минимизация (6.96) методом ОПС привела к следующей точке μ:

$$\mu_1 = 0{,}7240029 \cdot 10^{-5};\ \mu_2 = 0{,}2704296 \cdot 10^{-2};\ \mu_3 = 0{,}2324280 \cdot 10;$$

$$\mu_4 = 0{,}2430573 \cdot 10^{-7};\ \mu_5 = 0{,}7020994 \cdot 10^{-5};\ \mu_6 = 0{,}4297312;$$

$$\mu_7 = 0{,}1366784 \cdot 10^{-4};\ \mu_8 = 0{,}8707104;\ \mu_9 = 0{,}4076001;$$

$$\mu_{10} = 0{,}4429756 \cdot 10^{-4};\ \mu_{11} = 0{,}7618611;\ \mu_{12} = 0{,}1217431 \cdot 10^{-4};$$

$$\mu_{13} = 0{,}1281024 \cdot 10^{-1};\ \mu_{14} = 0{,}3692824 \cdot 10^{-6};\ \mu_{15} = 0{,}1507476 \cdot 10$$

и соответствующей передаточной функции

$$T(p) = \frac{0{,}4754518 \cdot 10^{-4} + 0{,}5167886 \cdot 10^{-4} p + 1{,}000001 p^2}{1{,}000004 + 0{,}01000007 p + 1{,}000002 p^2}.$$

Вычисленные собственные значения матрицы $J''(\mu)$ равны

$$\lambda_1 = 0{,}4238 \cdot 10^{7};\ \lambda_2 = 0{,}5342 \cdot 10^{2};\ \lambda_3 = 0{,}230 \cdot 10^{2};$$

$$\lambda_4 = 0{,}570 \cdot 10;\ \lambda_5 = 0{,}167;\ \lambda_6 = -0{,}795 \cdot 10^{-4};$$

$$\lambda_7 = 0{,}192 \cdot 10^{-5};\ \lambda_8 = 0{,}238 \cdot 10^{-6};\ \lambda_9 = 0{,}220 \cdot 10^{-7};$$

$$\lambda_{10} = 0{,}452 \cdot 10^{-9};\ \lambda_{11} = -0{,}189 \cdot 10^{-11};\ \lambda_{12} = 0{,}585 \cdot 10^{-12};$$

$$\lambda_{13} = 0{,}616 \cdot 10^{-13};\ \lambda_{14} = \lambda_{15} = 0.$$

Собственные числа $\lambda_6 - \lambda_{15}$ отнесены к блоку «малых», что соответствует исключению 10 компонентов вектора μ. В результате работы алгоритма исключения были оставлены компоненты μ_3, μ_6, μ_8, μ_{11}, μ_{15}. После уточнения параметров с помощью повторного решения задачи оптимизации (6.96) в 5-мерном пространстве были получены следующие окончательные результаты (рис. 6.4): $G_3 = 25{,}301$; $G_6 = 0{,}039525$; $C_1 = 0{,}63571$; $C_4 = 1{,}5715$; $k = 1{,}0019$;

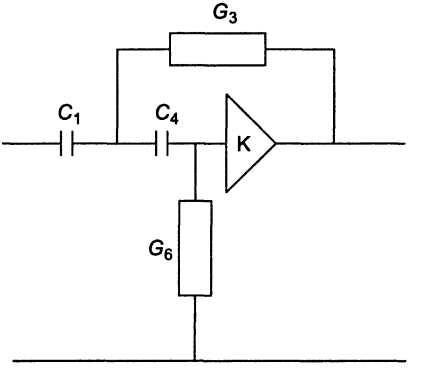

Рис. 6.4. Схема, полученная после упрощения

$$T(p) = \frac{1{,}00096p^2}{1{,}000003 + 0{,}9999997 \cdot 10^{-2} p + 0{,}9990286p^2}.$$

Последующее упрощение цепи на основе включения λ_5 в блок малых собственных значений оказалось невозможным из-за пропадания необходимых составляющих в выражении для $T(p)$.

6.2.3. Удаление переменных в задаче наименьших квадратов

Под линейной задачей наименьших квадратов (НК) будем понимать следующую задачу [35].

Пусть заданы действительная $(m \times n)$-матрица A ранга $k \le \min(m, n)$ и действительный m-вектор b. Требуется найти действительный n-вектор x^*, минимизирующий евклидову длину вектора $Ax - b$. Для обозначения задачи НК используется символика $Ax \cong b$.

Удаление переменной из задачи НК соответствует фиксированию значения этой переменной в нуле. Если удаляется одна переменная, например, x_n, то задача $Ax \cong b$ преобразуется в задачу $\tilde{A}\tilde{x} = b$, где $\tilde{A}$ — $m \times (n-1)$-матрица, состоящая из первых $n-1$ столбцов исходной матрицы A, а $\tilde{x}$ — $(n-1)$-мерный вектор. Наиболее часто к удалению переменных прибегают для определения наименьшего числа параметров x_i с сохранением приемлемо малой нормы невязки $\| Ax - b \|$.

Примеры задач НК из области теории управления приведены в главе 1.

Решение задачи об удалении переменных сводится к решению следующей задачи о выборе подмножеств. Для каждого значения $k \in [1: n]$ необходимо найти подмножество I_k из k индексов такое, что норма ρ_k невязки, полученной при решении задача НК относительно k указанных переменных x_i, $i \in I_k$, не превосходит значения нормы, получаемой при выборе любого другого набора из k переменных. Сформулированная задача полностью эквивалентна задаче, рассматриваемой в 6.2.2, в предположении

$$J(x) \triangleq 0{,}5 \| Ax - b \|^2. \tag{6.97}$$

Так же, как и в случае задачи (6.89), прямой перебор всех возможных комбинаций из k индексов для $\forall k \in [1: n]$ приводит к нереальным вычислительным затратам при достаточно больших значениях n.

Существует альтернативный подход. Соответствующий алгоритм называется *пошаговой регрессией* и заключается в следующем. При $k = 1$ решается сформированная ранее задача о выборе подмножеств. Пусть $\tilde{I}$ — ее решение. Далее, предположим, что построено подмножество $\tilde{I}_k$. Тогда в качестве $\tilde{I}_{k+1}$ выбирается такое множество $\tilde{I}_k \cup \{j\}$ $(j \notin \tilde{I}_k, j \in [1: n])$, для которого реализуется наименьшая невязка $\tilde{\rho}_{k+1}$.

Проведем сравнительный анализ алгоритма пошаговой регрессии и алгоритма из 6.2.2 на примере решения конкретной задачи НК.

Пример. Рассмотрим задачу НК [35] $Ax \cong b$, (15×5)-матрица A и вектор b которой приведены в табл. 6.3. Предполагается, что имеется неопределенность порядка $0{,}5 \cdot 10^{-8}$ в элементах A и порядка $0{,}5 \cdot 10^{-4}$ в элементах b.

Сингулярное разложение $(m \times n)$-матрицы A ранга k имеет вид

$$A = USV^{\mathrm{T}},$$

где U — ортогональная $(m \times m)$-матрица; V — ортогональная $(n \times n)$-матрица; $S = \{s_{ij}\}$ — диагональная $(m \times n)$-матрица, у которой из $s_{ij} \neq 0$ следует $i = j$; для $i \in [1{:}\ n]$ $s_{ii} \geq 0$, для $i \in [1{:}\ k]$ $s_{ii} > 0$. Диагональные элементы s называются *сингулярными числами* A.

Таблица 6.3. Матрица исходных данных [A:b]

−0,13405547	−0,20162827	−0,16930778	−0,18971990	−0,17387234	−0,4361
−0,10379475	−0,15766336	−0,13346256	−0,14848550	−0,13597690	−0,3437
−0,08779597	−0,12883867	−0,10683007	−0,12011796	−0,10932972	−0,2657
0,02058554	0,00335331	−0,01641270	0,00078606	0,00271659	−0,0392
−0,03248093	−0,01876799	0,00410639	−0,01405894	−0,01384391	0,0193
0,05967662	0,06667714	0,04352153	0,05740438	0,05024962	0,0747
0,06712457	0,07352437	0,04489770	0,06471862	0,05876455	0,0935
0,08687186	0,09368296	0,05672327	0,08141043	0,07302320	0,1079
0,02149662	0,06222662	0,07213486	0,06200069	0,05570931	0,1930
0,06687407	0,10344506	0,09153849	0,09508223	0,08393667	0,2058
0,15879069	0,18088339	0,11540692	0,16160727	0,14796479	0,2606
0,17642887	0,20361830	0,13057860	0,18385729	0,17005549	0,3142
0,11414080	0,17259611	0,14816471	0,16007466	0,14374096	0,3529
0,07846038	0,14669563	0,14365800	0,14003842	0,12571177	0,3615
0,10803175	0,16994623	0,14971519	0,15885312	0,14301547	0,3647

Известно, что положительно определенная симметричная $(n \times n)$-матрица $A^{\mathrm{T}}A$ допускает спектральное разложение

$$A^{\mathrm{T}}A = VDV^{\mathrm{T}}, \tag{6.98}$$

где $D = \{d_{ij}\}$ — диагональная $(n \times n)$-матрица, состоящая из собственных чисел матрицы $A^{\mathrm{T}}A$, причем $d_{ii} = s_{ii}^2$, $i \in [1{:}n]$; столбцы матрицы V суть ортонормированные собственные векторы матрицы $A^{\mathrm{T}}A$.

Далее представлены результаты сингулярного разложения матрицы A из табл. 6.3:

$$10^4 V = \begin{bmatrix} 3742 & -7526 & 3382 & -1981 & -3741 \\ 5196 & -636 & 2301 & 6349 & 5195 \\ 4123 & 6510 & 4741 & -1067 & -4123 \\ 4796 & 689 & -2493 & -6877 & 4797 \\ 4359 & 302 & -7388 & 2707 & -4359 \end{bmatrix},$$

$s_{11} = 1{,}00000$; $s_{22} = 0{,}1000$; $s_{33} = 0{,}0100$; $s_{44} = 0{,}9997 - 05$; $s_{55} = 0{,}9904 - 07$.

Собственные числа матрицы $J''(x) = A^{\mathrm{T}}A$, где $J(x)$ определен равенством (6.97), равны

$$\lambda_i(A^{\mathrm{T}}A) = s_{ii}^2,\ i \in [1\colon 5]$$

и, следовательно, cond $(A^{\mathrm{T}}A) \cong 10^{14}$.

Решим задачу удаления трех переменных в сформулированной задаче НК. Точное решение задачи, полученное простым перебором всех возможных комбинаций, согласно табл. 6.4, приводит к выбору первого и пятого опорных столбцов, которым соответствует минимальная норма невязки, равная 0,001. Таким образом, решение задачи известно, и далее она будет использована как тестовая при сравнении алгоритма из 6.2.2 и алгоритма пошаговой регрессии.

Таблица 6.4

Опорные столбцы	$\|\omega\|$	$\|b - A\omega\|$	Опорные столбцы	$\|\omega\|$	$\|b - A\omega\|$
1	2,46	0,40	1,5	5,07	0,001
2	1,92	0,22	2,3	3,03	0,053
3	2,42	0,07	2,4	20,27	0,030
4	2,09	0,19	2,5	17,06	0,128
5	2,30	0,19	3,4	3,07	0,056
1,2	5,09	0,039	3,5	2,97	0,058
1,3	2,72	0,052	4,5	17,05	0,175
1,4	4,53	0,023			

Вначале обратимся к методу пошаговой регрессии. Из табл. 6.4 видно, что если обозначить через $\omega^{(i,j,\ldots)}$ решение, опирающееся на столбцы $i, j, \ldots$, то сначала будет найден вектор $\omega^{(3)}$, которому соответствует наименьшее значение нормы невязки, равное 0,07. Далее будет выбрано наилучшее решение из множества

$$\omega^{(1,3)},\ \omega^{(2,3)},\ \omega^{(3,4)},\ \omega^{(3,5)}. \tag{6.99}$$

В результате получим вектор $\omega^{(1,3)}$ как наилучший среди векторов (6.99) по значению невязки $\|A\omega - b\|$. Соответствующая вектору $\omega^{(1,3)}$ норма невязки равна 0,052, что в 52 раза больше, чем невязка 0,001, доставляемая оптимальным вектором $\omega^{(1,5)}$.

Теперь покажем, что алгоритм, приведенный в 6.2.2, позволяет получить оптимальное решение этой задачи на основе информации о спектральном разложении (6.98), получаемом, например, из сингулярного анализа матрицы A.

Выберем собственные векторы, отвечающие трем наименьшим собственным числам $\lambda_i(A^{\mathrm{T}}A)$, $i = 3, 4, 5$. Это будут соответствующие столбцы матрицы V. Обозначим их через V^1, V^2, V^3. Тогда

$$V^1 = 10^{-4}\cdot(3382;\ 2301;\ 4741;\ -2493;\ -7388);$$

$$V^2 = 10^{-4}\cdot(-1981;\ 6349;\ -1067;\ -6877;\ 2707);$$

$$V^3 = 10^{-4}\cdot(-3741;\ 5195;\ -4123;\ 4797;\ -4359),$$

и уравнение (6.89) примет вид

$$c_1V^1 + c_2V^2 + c_3V^3 = f, \tag{6.100}$$

где f = (2,4857328; 0,52913252; 0,18414114; −1,6156794; −3,4547871), $f = -\mu$, а μ есть построенное в [35] решение задачи НК (опирающееся на все 5 столбцов матрицы A).

Согласно алгоритму, приведенному в 6.2.2, последовательно получаем $\|\cdot\|_1$ = = 4,587709; $\|\cdot\|_2$ = 0,6210388; $\|\cdot\|_3$ = 0,2889410; $\|\cdot\|_4$ = 1,847010; $\|\cdot\|_5$ = 3,840756.

Следовательно, выбирается третье уравнение

$$0{,}4741c_1 - 0{,}1067c_2 - 0{,}4123c_3 = 0{,}18414114,$$

откуда

$$c_1 = 0{,}225058c_2 + 0{,}8696477c_3 + 0{,}18414114 \triangleq L_1(c_2, c_3).$$

Исключая c_1 из оставшихся уравнений системы (6.100), получаем новую систему уравнений относительно двух переменных:

$$\begin{aligned}
&(1):\ -0{,}1219853c_2 - 7{,}998515\ 10^{-2}c_3 = 2{,}423456;\\
&(2):\ 0{,}686858c_2 + 0{,}7196059c_3 = 0{,}4867616;\\
&(4):\ -0{,}7438069c_2 + 0{,}2628968c_3 = -1{,}569773;\\
&(5):\ 0{,}1044271c_2 - 1{,}078395c_3 = -3{,}318743.
\end{aligned} \tag{6.101}$$

Имеем

(1): $c_2 = -13{,}89348$; $c_3 = -9{,}109885$; $c_1 = L_1(c_2, c_3) = -10{,}86508$; $\|\cdot\|_1 = 19{,}85116$

(приведенные значения c_2, c_3 доставляют нормальное решение уравнению (6.83) в (6.101));

(2): $c_2 = 0{,}3378435$; $c_3 = 0{,}3540399$; $c_1 = 0{,}5680655$; $\|\cdot\|_2 = 0{,}7497872$;

(4): $c_2 = 1{,}876087$; $c_3 = -0{,}6630985$; $c_1 = 2{,}970736\ 10^{-2}$; $\|\cdot\|_4 = 1{,}990046$;

(5): $c_2 = -0{,}2952416$; $c_3 = 3{,}048893$; $c_1 = 2{,}769158$; $\|\cdot\|_5 = 4{,}129304$.

Таким образом, выбирается уравнение (2) системы (6.101):

$$0{,}6866858c_2 + 0{,}7196059c_3 = 0{,}4867616.$$

Отсюда имеем вторую линейную связь:

$$c_3 = -0{,}9542525c_2 + 0{,}676428 \triangleq L_2(c_2).$$

Исключая c_3 из уравнений (1), (4), (5) системы (6.101), получаем

(1): $c_2 = -54{,}26191$; $c_3 = L_2(c_2) = 52{,}45599$; $c_1 = L_1(c_2, c_3) = 33{,}59029$; $\| c \| = 82{,}60927$;

(4): $c_2 = 1{,}756956$; $c_3 = -1{,}000151$; $c_1 = -0{,}2902209$; $\| c \| = 2{,}042406$;

(5): $c_2 = -2{,}284352$; $c_3 = 2{,}856276$; $c_1 = 2{,}153983$; $\| c \| = 4{,}244551$.

Наименьшее значение $\| c \|$ соответствует уравнению (4); следовательно, остаются 1-е и 5-е уравнения исходной системы (6.100), что соответствует сохранению компонентов μ_1, μ_5 вектора μ. Таким образом, номера оптимальных опорных столбцов, соответствующие номерам ненулевых компонентов μ_1, μ_5 вектора μ, получены почти без дополнительных вычислительных затрат. Задача решена.

6.3. Основные результаты и выводы

Доказан принцип квазистационарности производных (ПКП) для линейных систем с симметричными матрицами, что позволило указать условия применимости ПКП и обосновать методы практического определения линейных связей, устанавливающихся вне пограничного слоя между фазовыми переменными жестких дифференциальных систем.

На основе структуры асимптотических связей сформулирована и доказана теорема о понижении размерности пространства поиска решаемой оптимизационной задачи. Получена оценка для степени овражности результирующего целевого функционала, определенного в $(n-1)$-мерном пространстве. Доказана теорема об асимптотическом понижении размерности пространства поиска оптимизационной задачи при произвольной размерности дна оврага.

Доказанные утверждения позволили обосновать процедуру иерархической оптимизации в последовательности подпространств пониженной размерности. Получены выражения для коэффициентов чувствительности решения оптимизационной задачи к изменению собственных чисел соответствующей матрицы Гессе, получены соотношения, устанавливающие внутренний механизм переноса погрешностей в задании исходных данных на окончательный результат. Рассмотрен пример, иллюстрирующий на конкретном числовом материале смысл и эффективность методов иерархической оптимизации.

Построен общий алгоритм иерархической оптимизации МИО. Указаны области рационального применения алгоритма МИО.

Рассмотрены методы исключения переменных на основе спектрального разложения матрицы Гессе в точке оптимума целевого функционала. Показано, что эффективность обычно применяемого метода, основанного на удалении «малых» составляющих оптимального вектора аргументов, существенно зависит от степени овражности (обусловленности) целевого функционала. В результате применения такого подхода, например для упрощения избыточных структур оптимизируемых систем, могут получаться структуры, неоптимальные как по количеству составляющих их элементов, так и по качеству функционирования.

Изучен алгоритм исключения избыточных переменных, аналогичный гауссовскому методу исключения со специальным выбором ведущего элемента. В результате вариации управляемого вектора x производятся только в пределах

«дна» оврага, где общий показатель качества (критерий оптимальности) меняется относительно слабо. Реализация такого подхода позволяет на основе квадратичной модели критерия оптимальности системы указать максимальное число исключаемых параметров и оценить ожидаемое при этом ухудшение качества. Оставшиеся после исключения переменные соответствующим образом корректируются. Изложенный метод, в отличие от классического подхода, инвариантен относительно выбора масштабов управляемых переменных и степени овражности критерия оптимальности.

Изучена методология применения рассматриваемого подхода для удаления переменных в задаче наименьших квадратов. На классическом тестовом примере продемонстрировано преимущество спектрального метода исключения по сравнению с классическим методом пошаговой регрессии, не позволяющим решить указанную задачу.

Обоснована целесообразность применения изученного подхода в задачах структурного синтеза, а также при построении минимальных параметрических представлений искомых непрерывных зависимостей, например в таких задачах теории управления, как задачи идентификации нелинейных детерминированных объектов с использованием моделей Вольтерра, а также задачи идентификации и синтеза, приводящие к интегральным уравнениям Фредгольма 1-го рода (уравнения Винера—Хопфа), решаемым на основе алгебраических методов.

Глава 7. Примеры решения реальных задач параметрической оптимизации

Приводимые далее численные примеры имеют целью проиллюстрировать на небольших и легко обозримых задачах методику применения изученных в книге методов и алгоритмов параметрической оптимизации. Показывается, что при решении реальных задач оптимизации возникает много дополнительных, «неоптимизационных» проблем, связанных с различными «тонкостями» компьютерного моделирования. Иногда возникает даже проблема разработки специальных проблемно-ориентированных численных методов и технологий, не относящихся непосредственно к проблемам оптимизации. Например, в 7.4.2 представлен новый метод численного интегрирования жесткой системы обыкновенных дифференциальных уравнений, оказывающийся совершенно необходимым в соответствующей задаче идентификации. На примере рассмотренных задач демонстрируются также элементы вычислительных технологий, связанные с решением следующих важных практических вопросов:

- нормализация основных переменных задачи;
- выбор вида критерия оптимальности;
- выбор весовых коэффициентов в критериях метода наименьших квадратов;
- методика перехода с одного критерия оптимальности на другой в процессе решения одной и той же задачи;
- методика реализации среднестепенной аппроксимации минимаксных критериев с целью предотвращения появления «машинных» нулей;
- учет ограничений методом замены переменных и с помощью специальных реализаций метода штрафных функций;
- работа со спецификациями оптимизируемой системы в виде системы неравенств и т. д.

Нужно иметь в виду, что большинство приведенных в данной главе задач являются все-таки модельными, и не следует слишком серьезно относиться к пред-

ставленным численным результатам с точки зрения профессионала в соответствующей предметной области. Здесь в первую очередь преследовалась цель демонстрации применения сопутствующих оптимизации вычислительных технологий в условиях, приближенных к реальным.

Основной вывод сводится к утверждению, что для решения каждой практической задачи или класса задач оптимизации требуется длительный период предварительной подготовки перед собственно обращением к соответствующим «библиотечным» методам параметрической оптимизации.

7.1. Реализация оптимальной весовой функции линейной стационарной системы

На этапе синтеза статистически оптимальной системы (см. 1.3.6) строится желаемая весовая функция $\omega_*(t)$. Далее по найденной зависимости $\omega_*(t)$ необходимо решить задачу реализации по определению реальной весовой функции, максимально приближенной к оптимальной. Таким образом, решается задача аппроксимации функции $\omega_*(t)$ с помощью некоторой функции $\omega(x,t)$, для которой выполнены условия физической реализуемости, а также известны соответствующие методы решения задачи реализации. Вектор управляемых параметров $x = (x_1, x_2, \ldots, x_n)$ должен выбираться из условия максимальной близости заданной $\omega_*(t)$ и аппроксимирующей $\omega(x,t)$ зависимостей на промежутке $t \in [0, \infty]$. Однако для устойчивых систем справедливо предельное соотношение $\lim \omega(t) = 0$, $t \to \infty$, что означает возврат системы в исходное состояние после импульсного входного воздействия [52]. Поэтому нет необходимости рассматривать бесконечный интервал времени, и достаточно ограничиться значениями $t \le T$, где T определяется из условия

$$\forall t > T: |\omega(t)| \le \alpha \max_{t \in [0,T]} |\omega(t)|, \ \alpha \in (0{,}05; 0{,}1).$$

Форма представления функции $\omega(x,t)$, гарантирующая выполнение условий физической реализуемости, имеет вид [32]

$$\omega(x,t) = \sum_{i=1}^{r} \exp(\sigma_i t)(A_{i1} \cos \omega_i t + A_{i2} \sin \omega_i t),$$

$$x = (A_{11}, \ldots, A_{r1}, A_{12}, \ldots, A_{r2}, \sigma_1, \ldots, \sigma_r, \omega_1, \ldots, \omega_r); \ \sigma_i \le 0, \ i \in [1:r].$$

Наиболее часто в качестве критериев близости $\omega_*(x)$ и $\omega(x,t)$ используются функционалы, построенные с помощью принципов минимакса, наименьших квадратов и обобщающего эти подходы метода среднестепенной аппроксимации (см. 2.8.1).

Далее приводится пример численного решения сформулированной задачи [72].

Пусть требуется определить весовую функцию $\omega(x,t)$ при условии $r = 1$ по заданной зависимости $\omega_*(t)$:

t_i	0	0,5	1	1,5	2
ω_*	−0,998134	−0,747134	−0,406134	−0,0691941	0,197866
t_i	2,5	3	3,5	4	4,5
ω_*	0,357866	0,403866	0,355866	0,247866	0,115866
t_i	5	5,5	6		
ω_*	−0,006134	−0,096134	−0,144134		

График функции $\omega_*(t)$ представлен на рис. 7.1. Начальный вектор равен $x^0 = (A_{11}, A_{12}, \sigma_1, \omega_1) = (-1{,}005; 0{,}06105; -0{,}296; 0{,}9355)$.

На рис. 7.2 показан график зависимости

$$\Delta_0(t) \triangleq \omega_*(t) - \omega(x^0, t); \max_t |\Delta_0(t)| = 0{,}6866 \cdot 10^{-2}.$$

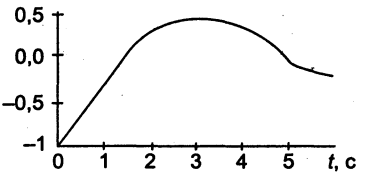

Рис. 7.1. Заданная весовая функция

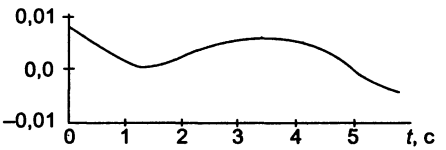

Рис. 7.2. График зависимости $\omega_*(t) - \omega(x^0, t)$

Критерий близости ω_* и ω построен на основе метода наименьших квадратов:

$$J(x) = \sum_{i=1}^{13} \left[\omega_*(t_i) - \omega(x, t_i)\right]^2, \tag{7.1}$$

где $\omega(x,t) = \exp(x_3 t)(x_1 \cos x_4 t + x_2 \sin x_4 t)$.

Ограничение $x_3 \le 0$ учитывалось с помощью замены переменных $x_i = y_i$, $i = 1, 2, 4$; $x_3 = -y_3^2$. Далее поиск минимума J осуществлялся в пространстве переменных y_i, на которые уже не накладывалось никаких ограничений. Начальное условие для y_3 было соответствующим образом пересчитано.

В результате решения задачи алгоритмом SPAC1 получены значения

$$y^* = (-0{,}99819; 0{,}061221; 0{,}53821; 0{,}93544),$$

которым соответствует вектор $x^* = (-0{,}99819;\ 0{,}061221;\ -0{,}28967;\ 0{,}93544)$. На рис. 7.3 приведен график $\Delta_*(t) = \omega_*(t) - \omega(x^*, t)$. Максимальная погрешность $\max_t |\Delta_*(t)| \cong 0{,}5 \cdot 10^{-3}$ уменьшилась приблизительно в 14 раз. Процесс поиска потребовал около 90 вычислений $J(x)$.

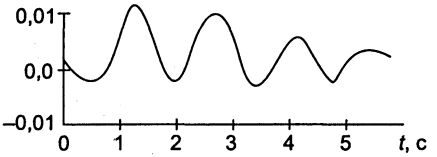

Рис. 7.3. График зависимости $\omega_*(t) - \omega(x^*, t)$

Незначительно уступающие по качеству и трудоемкости результаты в данном случае удалось также получить с помощью метода циклического покоординатного спуска, реализованного в алгоритме GZ1.

Для уменьшения значения максимальной погрешности вместо МНК-критерия (7.1) была использована среднестепенная аппроксимация минимаксного критерия

$$J_1(x) = \sum_{i=1}^{13} \left\{ \left[\omega_*(t_i) - \omega(x, t_i) \right] 10^3 \right\}^8. \qquad (7.2)$$

Множители 10^3 введены для предотвращения возникновения машинного нуля при возведении малых разностей в восьмую степень; существенно, что эти множители не могут быть вынесены за фигурные скобки.

Минимизация критерия (7.2) методом SPAC1 привела к результату $x^* = (-0{,}99796;\ 0{,}061193;\ -0{,}28950;\ 0{,}93546)$. При этом максимальная погрешность оказалась равна $0{,}38 \cdot 10^{-3}$, что хорошо согласуется с результатами работы [32], где минимизировался непосредственно минимаксный критерий методом возможных направлений. В данном случае за счет перехода от критерия (7.1) к (7.2) удалось уменьшить значение погрешности примерно в 18 раз по сравнению с исходной.

На процесс минимизации в целом было затрачено около 300 Горнеров (300 вычислений значений целевого функционала).

В результате можно сделать вывод, что для решения подобных задач нет необходимости прибегать к значительно более сложным, а подчас и менее эффективным методам условной минимизации (то есть специальным методам, работающим в условиях ограничений), а также к специальной технике минимаксной оптимизации.

За счет некоторого изменения в подготовке задачи к решению и обращения к алгоритму SPAC5 удалось сократить трудоемкость получения приведенных ранее результатов приблизительно в три раза. Для «больших» задач указанный выигрыш может оказаться принципиальным.

7.2. Аппроксимация характеристик частотно-избирательных фильтров

Обычный (классический) подход к проблеме синтеза электрических фильтров с заданными характеристиками заключается в последовательном выполнении этапов аппроксимации и реализации. Такой двухэтапный подход имеет ряд известных преимуществ, которые определили его широкое практическое применение (см. 2.6.1).

На этапе аппроксимации заданные условия обработки сигнала выражаются некоторой специальной физически реализуемой функцией, принадлежащей к одному из известных классов: Баттерворта, Чебышева и т. д. Для функций этих классов существует техника построения соответствующих передаточных функций в виде дробно-рациональных выражений, а также аппарат схемной реализации с соответствующими таблицами параметров элементов схемы.

Однако указанные методы разработаны лишь для некоторых основных типов фильтров и не позволяют строить оптимальные в каком-либо смысле передаточные функции, обладающие к тому же дополнительными свойствами, такими, например, как заданные добротность, стабильность и т. д. Поэтому более гибкий подход связан с компьютерным проектированием фильтров, когда на стадии аппроксимации непосредственно ищется дробно-рациональное выражение, удовлетворяющее всему комплексу требований, сформулированных в виде равенств, неравенств и заданных критериев оптимальности. Такой подход, в частности, является наиболее естественным при проектировании фильтров, идеальные амплитудно-частотные характеристики которых не являются кусочно-постоянными и не сводятся поэтому ни к одному из стандартных видов полосового фильтра. Соответствующий пример из области дальней телефонной связи приведен в [37] и заключается в построении схемы с монотонно нарастающей по заданному закону амплитудно-частотной характеристикой.

В качестве примера рассмотрим задачу построения дробно-рационального выражения вида [36]

$$W(p, f) = \frac{1}{(f_1 + f_2 p + f_3 p^2 + f_4 p^3)(f_5 + f_6 p + f_7 p^2 + f_8 p^3)},$$

обладающего следующими свойствами. Модуль коэффициента передачи $|W(j\omega, f)|$ в полосе пропускания $\frac{\omega}{2\pi} \in [0, \ldots, 600]$ Гц удовлетворяет неравенству $0{,}7 \le |W(i\omega, f)| \le 1$.

При $\frac{\omega}{2\pi} = 600$ Гц требуется, чтобы $|W| = 0{,}707 \pm 0{,}007$, а при $\frac{\omega}{2\pi} = 1200$ Гц необходимо выполнение неравенства $|W| \le 0{,}005$. Сформулированные требования соответствуют НЧ-фильтру (фильтру нижних частот), и геометрически представлены на рис. 7.4.

При выборе коэффициентов f_i кроме условия их положительности необходимо соблюсти требование устойчивости отдельных звеньев фильтра $\Delta_1 = f_2 f_3 - f_1 f_4 > 0$; $\Delta_2 = f_6 f_7 - f_5 f_8 > 0$, а также ограничения на добротности:

$$Q_1 = \sqrt{\frac{f_1}{f_3}}\,\frac{f_3^2 + f_2 f_4}{\Delta_1} < 10;\; Q_2 = \sqrt{\frac{f_5}{f_7}}\,\frac{f_7^2 + f_6 f_8}{\Delta_2} < 10.$$

Эти соотношения эквивалентны неравенствам

$$\sqrt{f_1}(f_3^2 + f_2 f_4) < 10\sqrt{f_3}\Delta_1;\; \sqrt{f_5}(f_7^2 + f_6 f_8) < 10\sqrt{f_7}\Delta_2. \qquad (7.3)$$

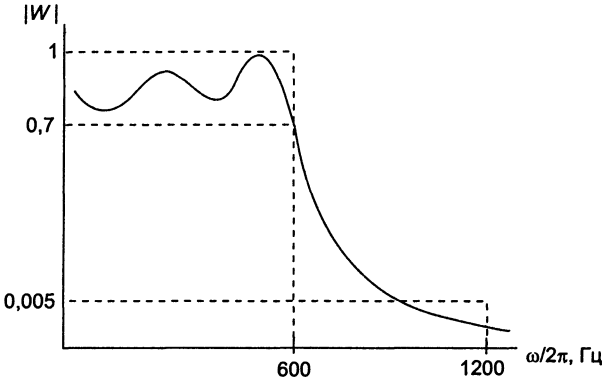

Рис. 7.4. Характеристика НЧ-фильтра

Представления (7.3) обычно являются более предпочтительными, так как не содержат операций деления, что существенно при возможных малых значениях Δ_i в случае выхода в процессе вычислений на границу области устойчивости. Ясно, что так как $f_i \geq 0$, то при $\Delta_1 \leq 0$ либо при $\Delta_2 \leq 0$ неравенства (7.3) не будут выполнены. Поэтому справедливость (7.3) гарантирует и выполнение требований по устойчивости.

Имеем $F(f,\omega) \triangleq |W(j\omega, f)| = \dfrac{1}{\sqrt{c_1 c_2}}$;

где

$$c_1 = (f_1 - f_3\omega^2)^2 + \omega^2 (f_2 - f_4\omega^2)^2;$$
$$c_2 = (f_5 - f_7\omega^2)^2 + \omega^2 (f_6 - f_8\omega^2)^2;\; f = (f_1, \ldots, f_8). \qquad (7.4)$$

Из полученного выражения видно, что максимальные степени частоты ω могут быть порядка ω^6. Чтобы исключить возможные вычислительные трудности, связанные с переполнением разрядной сетки при вычислении величины F для высоких частот ω, была проведена нормализация частоты. А именно, приняли значение ω_0 равным $\omega_0 = 10^4$. Далее перешли к нормализованной частоте $\Omega = \dfrac{\omega}{\omega_0}$.

Подставляя $\omega = \omega_0\Omega$ в (7.4), получим

$$c_1 = (f_1 - f_3\Omega^2\omega_0^2)^2 + (\Omega\omega_0 f_2 - \Omega^3\omega_0^3 f_4)^2;$$
$$c_2 = (f_5 - f_7\Omega^2\omega_0^2)^2 + (\Omega\omega_0 f_6 - \Omega^3\omega_0^3 f_8)^2.$$

Учитывая неотрицательность компонентов вектора f и полагая $f_i = y_i^2$, получим окончательно:

$$c_1 = (x_1^2 - \Omega^2 x_3^2)^2 + (\Omega\, x_2^2 - \Omega^3 x_4^2)^2;$$
$$c_2 = (x_5^2 - \Omega^2 x_7^2)^2 + (\Omega\, x_6^2 - \Omega^3 x_8^2)^2; \qquad (7.5)$$

где $x = (x_1, x_2, \ldots, x_8)$ — вектор нормализованных переменных с компонентами

$$x_1 = y_1;\ x_2 = \sqrt{\omega_0}\, y_2;\ x_3 = \omega_0 y_3;\ x_4 = \omega_0^{\frac{3}{2}} y_4;$$
$$x_5 = y_5;\ x_6 = \sqrt{\omega_0}\, y_6;\ x_7 = \omega_0 y_7;\ x_8 = \omega_0^{\frac{3}{2}} y_8. \qquad (7.6)$$

Заданные требования к амплитудно-частотной характеристике фильтра могут быть переформулированы в новых обозначениях. Модуль коэффициента передачи $\Phi(x,\Omega) \triangleq \frac{1}{\sqrt{c_1 c_2}}$, где c_1, c_2 рассчитываются по формулам (7.5) в полосе пропускания $\Omega \in [\Omega_1, \Omega_2]$, $\Omega_1 = 0$, $\Omega_2 = \left(\frac{2\pi}{\omega_0}\right) 600 = 0{,}377$, должен удовлетворять соотношению $\Phi(x, \Omega) = 0{,}85 \pm 0{,}15$. И аналогично, $\Phi(x, \Omega) = 0{,}707 \pm 0{,}007$ при $\Omega = 0{,}377$; $\Phi(x, \Omega) \le 0{,}005$ при $\Omega = 0{,}754$.

По каждому вектору x может быть построен соответствующий вектор f и проверены условия (7.3).

Задача оптимального параметрического синтеза фильтра теперь может быть сформулирована следующим образом. Определить такой вектор x, чтобы функционал

$$J(x) = \sum_{i=0}^{35} [\Phi(x,\ 10^{-2} i) - 0{,}85]^2 + 500[\Phi(x,\ 0{,}377) - 0{,}707]^2 + \\ + 10^3 [\Phi(x, 0{,}754)]^2 \qquad (7.7)$$

принял минимальное значение. Весовые коэффициенты определялись исходя из заданных требований к точности аппроксимации в различных частотных диапазонах $\alpha \cdot 0{,}15^2 = \beta \cdot 0{,}007^2 = \gamma \cdot 0{,}005^2$. Полагая $\alpha = 1$, получим

$$\beta = \frac{0{,}15^2}{0{,}007^2} \cong 500;\ \gamma = \frac{0{,}15^2}{0{,}005^2} \cong 10^3.$$

Минимизация $J(x)$ осуществлялась алгоритмом GZ1 с начальным шагом дискретности $s = 0{,}1$ из произвольно выбранной начальной точки, удовлетворяющей неравенствам (7.3): $f_1^0 = 1$; $f_2^0 = 10^{-2}$; $f_3^0 = 10^{-4}$; $f_4^0 = 10^{-8}$; $f_5^0 = 1$; $f_6^0 = 10^{-2}$; $f_7^0 = 10^{-4}$; $f_8^0 = 10^{-8}$, или, на языке переменных x_i:

$$x_1^0 = x_2^0 = x_3^0 = x_5^0 = x_6^0 = x_7^0 = 1;\ x_4^0 = x_8^0 = 0{,}1. \qquad (7.8)$$

В точках x, для которых не выполнялось хотя бы одно из соотношений (7.3), полагали $J := 10^8 > J(x^0) \cong 2 \cdot 10^3$, что эквивалентно применению метода штрафных функций с достаточно большим коэффициентом штрафа.

График АЧХ, соответствующий начальным значениям (7.8) представлен на рис. 7.5.

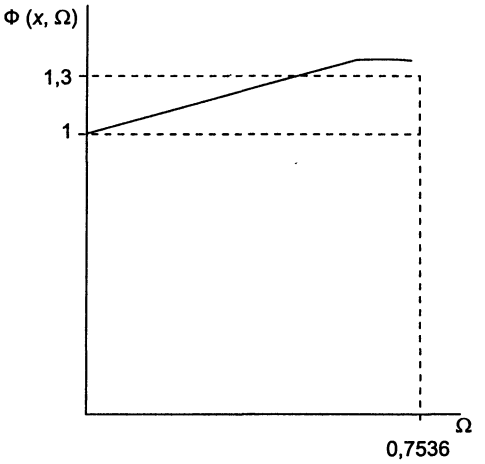

Рис. 7.5. АЧХ в начальной точке

После 1350 обращений к подпрограмме вычисления $J(x)$ была получена точка

$$x_1 = 0{,}84394;\ x_2 = 2{,}2411;\ x_3 = 4{,}0624;\ x_4 = -1{,}5526;$$
$$x_5 = 1{,}2995;\ x_6 = 1{,}2618;\ x_7 = 3{,}5938;\ x_8 = -0{,}73217, \qquad (7.9)$$

которой соответствуют значения $J = 0{,}3855;\ Q_1 = 0{,}72;\ Q_2 = 3{,}084$ и коэффициенты передаточной функции

$$f_1 = 0{,}71223;\ f_2 = 0{,}50227 \cdot 10^{-3};\ f_3 = 0{,}16503 \cdot 10^{-6};\ f_4 = 0{,}24105 \cdot 10^{-11};$$
$$f_5 = 1{,}6887;\ f_6 = 0{,}15921 \cdot 10^{-3};\ f_7 = 0{,}12915 \cdot 10^{-6};\ f_8 = 0{,}53607 \cdot 10^{-12}.$$

При этом условие в точке $\Omega = 0{,}754$, где имеем $\Phi \cong 0{,}02$, не выполнено (должны иметь $\Phi \le 0{,}005$), а сходимость метода GZ1 стала весьма медленной. График $\Phi(x, \Omega)$ изображен на рис. 7.6.

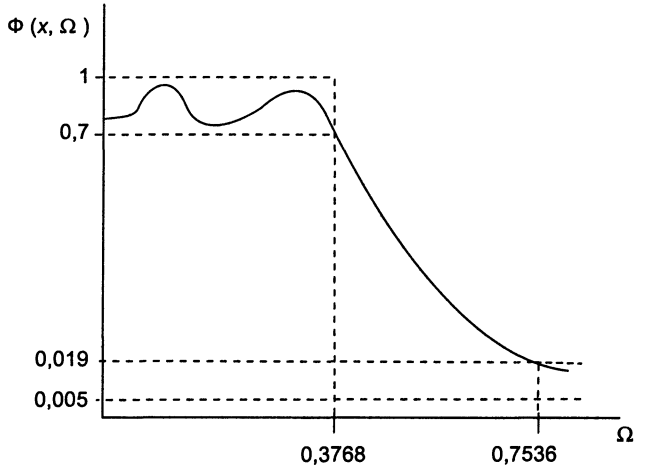

Рис. 7.6. АЧХ фильтра, полученная алгоритмом GZ1

Таким образом, минимизация $J(x)$ привела к наличию «выброса» аппроксимирующей функции в точке $\Omega = 0{,}754$ за счет высокой точности аппроксимации в остальной части диапазона. Данная характерная особенность МНК-критериев рассматривалась в 2.8.1.

Для устранения отмеченного недостатка функционал (7.7) был заменен на минимаксный с применением его среднестепенной аппроксимации. А именно, приняли

$$J(x) = \sum_{i=0}^{35} [\Phi(x; 10^{-2} i) - 0{,}85]^8 + \{20[\Phi(x; 0{,}377) - 0{,}707]\}^8 + \\ + [30\Phi(x; 0{,}754)]^8. \qquad (7.10)$$

За начальную точку были взяты значения (7.9). Результаты, полученные методом GZ1 для функционала (7.10), представлены на рис. 7.7. Сходимость опять резко замедлилась до достижения приемлемой точки, и после $g = 2269$ Горнеров в точке $\Omega = 0{,}754$ имели $\Phi \cong 0{,}009$, что все еще превышает допустимое значение, равное 0,005.

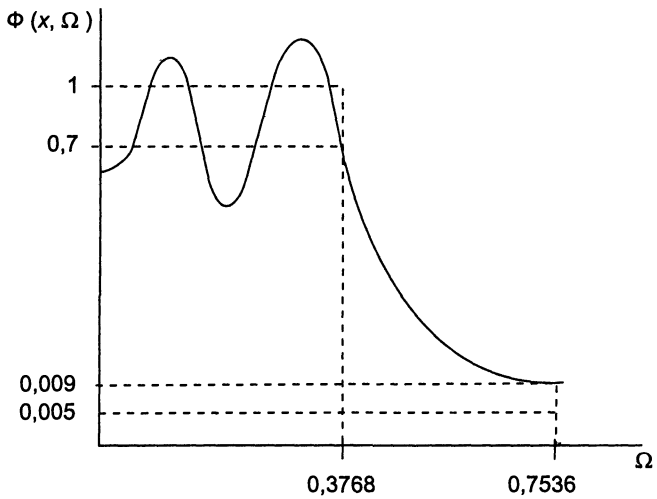

Рис. 7.7. АЧХ, построенная алгоритмом GZ1 при использовании минимаксного критерия

Точность аппроксимации в остальных точках была при этом нарушена. Максимальное уклонение от заданной характеристики с учетом весовых коэффициентов по-прежнему реализуется на границе диапазона при $\Omega = 0{,}754$.

Из начальной точки (7.9) для минимизации критерия (7.10) применялся алгоритм SPAC5 с начальным шагом дискретности $s = 0{,}1$. График полученной зависимости $\Phi(x, \Omega)$ представлен на рис. 7.8. Все требования к амплитудно-частотной характеристике оказываются выполненными; в том числе имеем

$$\Phi(x, 0{,}754) \cong 0{,}004 < 0{,}005.$$

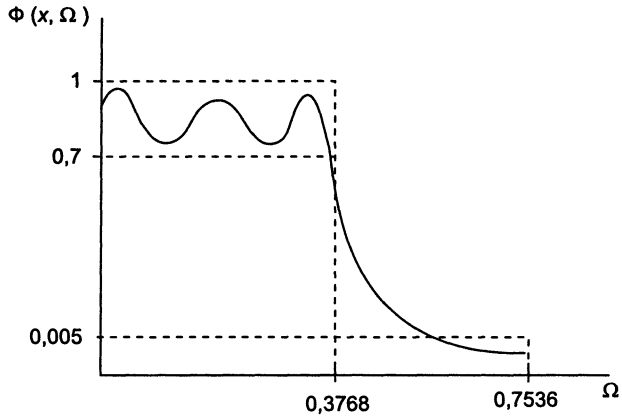

Рис. 7.8. АЧХ, построенная методом SPAC5 для минимаксного критерия

Соответствующие полученным результатам добротности и коэффициенты передаточной функции равны: $Q_1 = 1{,}549$; $Q_2 = 6{,}155$;

$$f_1 = 0{,}65346;\ f_2 = 0{,}81629 \cdot 10^{-3};\ f_3 = 0{,}29287 \cdot 10^{-6};\ f_4 = 0{,}10853 \cdot 10^{-9};$$

$$f_5 = 1{,}6996;\ f_6 = 0{,}16451 \cdot 10^{-3};\ f_7 = 0{,}13085 \cdot 10^{-6};\ f_8 = 0{,}64040 \cdot 10^{-11}.$$

Трудоемкость приведенного результата составила около 600 Горнеров. Для решения задачи минимизации функционала (7.10) в качества эксперимента привлекались стандартные квазиньютоновские процедуры, а также библиотечная версия метода Нелдера—Мида. При этом были достигнуты неудовлетворительные результаты, аналогичные полученным методом GZ1. Работа указанных алгоритмов прерывалась при превышении уровня в 2000 вычислений $J(x)$.

Приведенная технология синтеза фильтров может быть полезна также при проектировании современных систем мобильной связи.

7.3. Оптимизация параметров переключательных электронных схем

Рассмотрим задачу определения оптимальных значений параметров резисторов в интегральной транзисторно-транзисторной логической схеме (ТТЛ) с заданными значениями параметров интегральных компонентов (рис. 7.9) [6], [43]. Последние выбраны соответствующими наихудшему сочетанию внешних условий.

Зададим условия работоспособности ТТЛ-схемы в виде системы неравенств

$$N > 20;\ \Delta u > 0{,}8;\ P \le 20;\ t_3 \le 15;\ s > 1{,}5, \qquad (7.11)$$

где N — коэффициент разветвления по выходу; Δu — допустимый уровень помехи в логическом состоянии «нуль»; P — средняя потребляемая мощность; t_3 — задержка распространения сигнала; s — степень насыщения транзистора $T2$.

(Здесь и далее применяются следующие единицы измерения: вольты, милливатты, наносекунды, килоомы.)

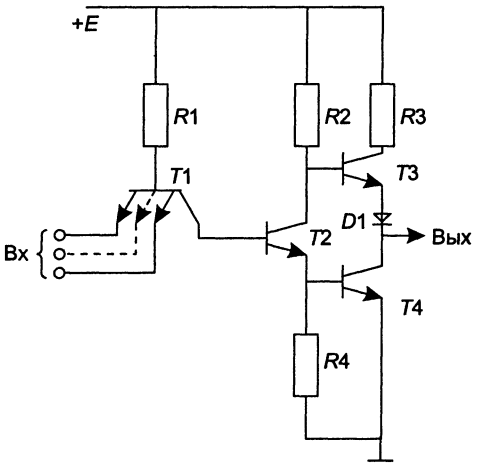

Рис. 7.9. ТТЛ-схема И–НЕ / ИЛИ–НЕ

В качестве математической модели использованы известные зависимости

$$N = 2{,}1\left[2{,}45 + R_1\left(\frac{3{,}95}{R_2 + 0{,}011} - \frac{1}{R_4}\right)\right];$$

$$\Delta u = 0{,}9 - 0{,}03 \ln R_4;$$

$$P = \frac{19}{R_1} + \frac{10{,}8}{R_2 + 0{,}017};$$

$$t_3 = 0{,}55 + \frac{0{,}087\ R_2}{R_4} + \left[R_2^2(6 + 0{,}079N)^2 + \left(0{,}69 - \frac{0{,}0024N\ R_2}{R_1}\right)\times\right.$$

$$\left.\times\ (5{,}7 + 0{,}15N)\,R_1 R_2\right]^{\frac{1}{2}} - 0{,}5\left\{R_2(3 + 0{,}079N) - 8\left[1 - 0{,}134R_2 - \frac{0{,}136R_2}{R_4} - \right.\right. \qquad (7.12)$$

$$\left.\left. - \frac{0{,}026N\ R_2}{R_1}\right]\left[1 - 0{,}134R_2 + \frac{6{,}9}{8 + 3R_3}\right]^{-1} - R_2(4{,}9 + 0{,}055N);\right.$$

$$s = \frac{8{,}55R_2}{R_1}.$$

Наличие явных выражений (7.12) в данном случае вносит непринципиальное упрощение задачи, так как здесь в первую очередь исследуется оптимизационный аспект. В общем случае для вычисления основных характеристик оптимизируемой системы необходимо прибегать к соответствующим программам анализа.

Ставится задача оптимального выбора номиналов резисторов $R_1, \ldots, R_4$, которые составляют вектор R управляемых параметров.

Приведем условия работоспособности (7.11) к стандартному виду:

$$y_i \le t_i,\ i \in [1{:}5], \tag{7.13}$$

где

$$y_1 = -N;\ y_2 = -\Delta u,\ y_3 = P,\ y_4 = t_3,\ y_5 = -s;$$
$$t_1 = -20;\ t_2 = -0{,}8;\ t_3 = 20;\ t_4 = 15;\ t_5 = -1{,}5.$$

Запас работоспособности по каждому из требований (7.13) представим в виде

$$z_i = a_i\left[\frac{t_i - y_i}{\delta_i} - 1\right], \tag{7.14}$$

где δ_i имеет смысл параметра рассеяния выходного параметра y_i.

Параметры рассеяния δ_i для отдельных выходных параметров в данном случае задаются как исходные данные (в процентном отношении к соответствующим значениям t_i), хотя и могут быть рассчитаны с помощью метода статистических испытаний.

Примем $\delta_i = 0{,}2\,|\,t_i\,|$ за исключением параметра $y_2 = -\Delta u$, имеющего меньший разброс, — порядка $0{,}05\,|\,t_2\,|$ [43].

Таким образом, имеем $\delta_1 = 4$; $\delta_2 = 0{,}04$; $\delta_3 = 4$; $\delta_4 = 3$; $\delta_5 = 0{,}3$.

Все весовые коэффициенты выбраны равными $a_i = 1$, кроме $a_5 = 10$. Последнее отклонение вызвано желанием усилить влияние запаса работоспособности z_5 на результирующую целевую функцию, исходя из физического смысла выходного параметра y_5.

Действительно, выбирая достаточно большой весовой коэффициент a_i для некоторого z_i, мы, с одной стороны, при нарушении условия работоспособности имеем существенное ухудшение целевой функции за счет ее умножения на a_i. С другой стороны, уже при незначительном перевыполнении условия $z_i \ge 0$ мы имеем заметный выигрыш, сравнимый с запасами работоспособности по остальным выходным параметрам. Следовательно, увеличение a_i вносит некоторый стабилизирующий фактор, в результате которого соответствующее условие работоспособности с высокой вероятностью будет выполнено, имея в то же время небольшой положительный запас работоспособности. Параметр s характеризует степень насыщения транзистора $T2$, которая обязательно должна достигать предписанного значения, но в то же время значительное увеличение s оказывается нецелесообразным с точки зрения быстродействия устройства; выбор $a_s = 10$ является, таким образом, оправданным.

В качестве критерия оптимальности воспользуемся среднестепенной формой (см. 2.8) критерия минимального запаса работоспособности. Для устранения прямых ограничений

$$0{,}5 \le R_1 \le 10;\ 0{,}5 \le R_2 \le 10;\ 0{,}05 \le R_3 \le 1;\ 0{,}1 \le R_4 \le 2,$$

имеющих вид $R_i' \le R_i \le R_i''$, перейдем к новому вектору управляемых параметров x, где

$$R_i = R_i'' + (R_i' - R_i'')\sin^2 x_i,\ i \in [1:4].$$

В результате получим следующее представление для критерия оптимальности:

$$J(x) = \sum_{i=1}^{5} \exp[-\nu z_i(x)] \to \min_x,\ x \in R^4. \qquad (7.15)$$

На первом этапе для получения приемлемого начального приближения функционал (7.15) минимизировался при $\nu = 1$ из произвольно выбранной начальной точки $x_i^0 = 0;\ i \in [1:4]$. При этом $J(x_0) \cong 3{,}046 \cdot 10^8$. После приблизительно 40 обращений к вычислению J с помощью алгоритма GZ1 была получена точка

$$x_1 = 4{,}234;\ x_2 = 1{,}30;\ x_3 = 1{,}51;\ x_4 = 0{,}794;$$

$$R_1 = 5{,}56;\ R_2 = 1{,}18;\ R_3 = 0{,}054;\ R_4 = 1{,}034; \qquad (7.16)$$

$$N = 32{,}57;\ \Delta u = 0{,}899;\ P = 12{,}44;\ t_3 = 14{,}72;\ s = 1{,}815,$$

соответствующая значению $J \cong 4{,}234$.

Далее для перехода к максиминному критерию приняли $\nu = 15$ и продолжили оптимизацию тем же методом GZ1 из найденной точки. Затратив на поиск около 160 Горнеров, получили: $x_1 = -0{,}88;\ x_2 = 1{,}35;\ x_3 = 1{,}57;\ x_4 = 1{,}075$, что соответствует следующим значениям управляемых и выходных параметров:

$$R_1 = 4{,}33;\ R_2 = 0{,}956;\ R_3 = 0{,}05;\ R_4 = 0{,}05;$$

$$N = 25{,}1;\ \Delta u = 0{,}92;\ P = 15{,}5;\ t_3 = 11{,}7;\ s = 1{,}9.$$

Полученные результаты согласуются с данными из книги [43], где применялись более сложные и универсальные процедуры оптимального параметрического синтеза. В данном случае удалось ограничиться простейшими средствами в виде метода циклического покоординатного спуска.

В такой ситуации можно было сразу ограничиться значениями (7.16), которым соответствуют приемлемые в соответствии с основными требованиями выходные параметры.

Заметим, что применение алгоритма GZ1 непосредственно к максиминному критерию вида

$$\min_j z_j \to \max_x,$$

где z_j задается выражением (7.14), не приводит к цели из-за представленной на рис. 4.4 ситуации заклинивания для негладких целевых функционалов.

Таким образом, приведенные ранее результаты подтверждают целесообразность применения среднестепенных форм минимаксных критериев с точки зрения эффективности последующей оптимизации.

7.4. Управление химико-технологическими процессами производства высокомолекулярных соединений

Представленные в книге методы, методики и алгоритмы применялись для решения задач управления химико-технологическими процессами производства полиэтилена в трубчатых многозонных реакторах, а также при управлении процессами блочной полимеризации стирола.

Здесь используется специальная профессиональная терминология без дополнительных комментариев. Цель изучения данного раздела состоит, помимо всего прочего, еще и в том, чтобы проиллюстрировать на реальной задаче те проблемы, которые возникают при работе системного аналитика и компьютерного математика в конкретной предметной области. Обычно приходится довольно глубоко вникать в специфику соответствующих представлений и в язык, присущие моделируемой проблеме. При изучении представленного материала необходимо основное внимание уделять системным и оптимизационным аспектам, отвлекаясь от непонятной терминологии, — если, конечно, вы не специалист в рассматриваемой предметной области.

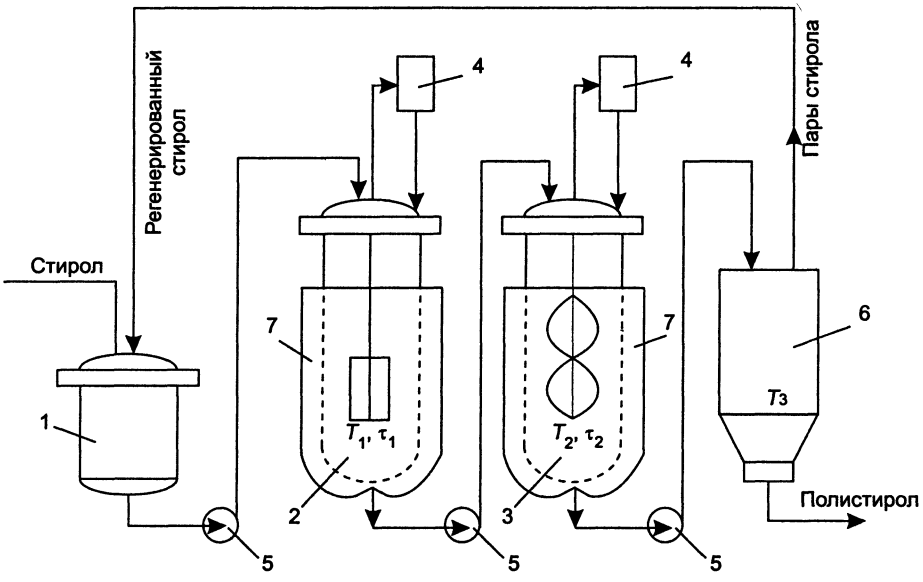

Рис. 7.10. Схема процесса производства блочного полистирола в каскаде аппаратов с перемешиванием (непрерывный способ)

На рис. 7.10 показана схема процесса полимеризации стирола в каскаде из двух реакторов с перемешиванием. Из емкости *1* стирол с помощью насоса *5* непрерывно подается в реактор первой ступени *2*, который имеет вид цилиндра с коническим дном объемом 10–22 м^3. В реакторе *2* процесс полимеризации протекает в изотермическом режиме при температуре T_1 = 110–130 °C до конверсии

порядка 30–40%. Время пребывания реакционной массы (τ_1) в реакторе составляет около 5 ч. Реактор снабжен листовой мешалкой с частотой вращения 30–90 об/мин. Основной съем избыточного тепла осуществляется за счет испарения части мономера из реакционной массы и ее конденсации в холодильнике *4*. Дополнительный теплосъем реализуется через охлаждающие поверхности (рубашки) *7*, питаемые от системы водоохлаждения (или пароохлаждения).

Из реактора *2* реакционная масса поступает в реактор второй ступени *3*. Реактор *3* по конструкции и габаритам полностью аналогичен реактору первой ступени, но снабжен ленточной мешалкой с частотой вращения 2–8 об/мин, что обеспечивает эффективное перемешивание высоковязких растворов. Полимеризация в реакторе *3* протекает до 85–95% конверсии мономера при постоянной температуре T_2 из интервала 140–170 °С. Время пребывания τ_2 может варьировать и также составляет около 5 ч.

Раствор полистирола в стироле из реактора второй ступени подается насосом *5* в вакуум-камеру *6* объемом 10 м^3 с температурой T_3 = 180–200 °С. При этом происходит испарение стирола из расплава, и содержание остаточного мономера снижается до приемлемого уровня 0,1%. Пары стирола после регенерации вновь поступают в емкость *1* либо выводятся из данного технологического цикла.

Из вакуум-камеры расплав полистирола поступает на окрашивание, грануляцию, расфасовку и упаковку.

Задача *управления процессом полимеризации* заключается в выборе и поддержании таких значений режимных параметров процесса $r = (T_1, \tau_1, T_2, \tau_2, \ldots)$, чтобы обеспечить выход продукта с заданными физико-механическими, оптическими, электрическими и другими свойствами. Поэтому основная задача состоит в построении математической модели химико-технологического процесса, с помощью которой по значениям режимных параметров можно надежно прогнозировать марку выпускаемого полимера.

Центральное место в общей математической модели процесса полимеризации (рис. 7.11) занимает *модель кинетики* (МК), позволяющая по заданным режимным параметрам r_i вычислять значения конверсии z, средневесовой $\overline{M}_W$ и среднечисловой $\overline{M}_n$ молекулярных масс, которые и определяют основные характеристики полимера [9].

Полная неформальная МК задается системой обыкновенных дифференциальных уравнений (ОДУ) 9-го порядка (для стирола), определенной с точностью до 8-мерного вектора y неизвестных параметров, соответствующих (при фиксированной температуре) скоростям протекания отдельных элементарных реакций (стадий):

$$\frac{dx}{dt} = f(t,x,y,r),\ t \in [0,\bar{t}],\ y \in D_y \subset R^8,\ x \in R^9,\ r \in D_r \subset R^k.$$

Ставится задача параметрической идентификации процесса по данным лабораторных исследований для дискретной сетки температур $T_i \in [T_н, T_в]$, $i \in [1{:}\ s] : y^{i*} = \arg\min J_i(y)$,

$$J_i(y) = \frac{1}{2}\sum_{j=1}^{N} \left\| \bar{x}_э(T_i, t_j) - \bar{x}(y, T_i, t_j) \right\|_G^2, \; t_j \in [0, \bar{t}], \; j \in [1:N];$$

$\bar{x}$ — измеряемый подвектор вектора x; $\bar{x}_э$ — заданные экспериментальные значения. На основе полученного набора векторов

$$[y^{1*}, y^{2*}, \ldots, y^{s*}]$$

определяются неизвестные параметры a_i, b_i в температурных зависимостях $y_i(T)$ скоростей протекания элементарных реакций:

$$y_i(T) = a_i \exp\left(\frac{-b_i}{RT}\right), \; i \in [1:8],$$

где $R = 1{,}986$ кал/град.моль. Для этого решается дополнительная задача параметрической оптимизации

$$\sum_{j=1}^{s}\left[y_i^{j*} - a_i \exp\left(\frac{-b_i}{RT_j}\right)\right]^2 \to \min_{a_i, b_i}$$

или соответствующая система линейных уравнений относительно переменных $\ln a_i$, b_i. Такая система уравнений получается логарифмированием выражений для $y_i(T)$.

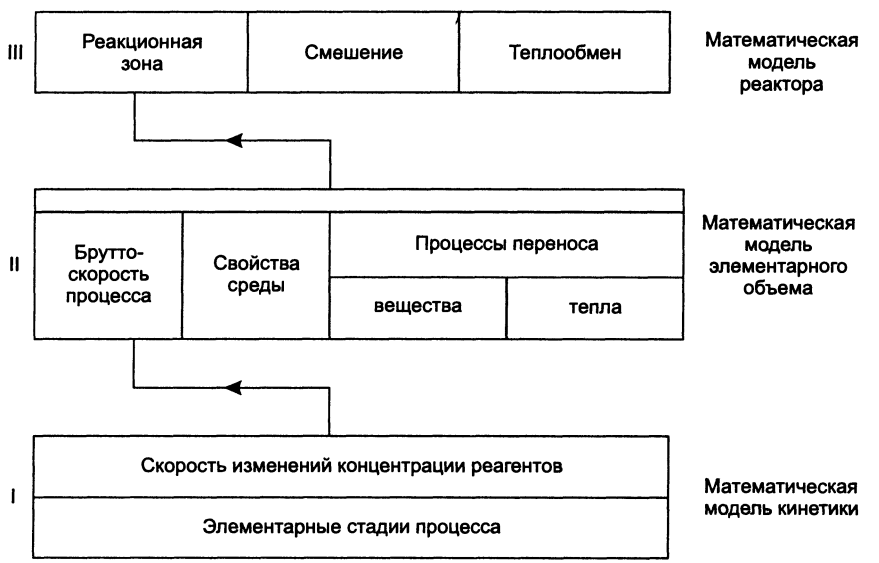

Рис. 7.11. Схема математической модели реактора

В ряде случаев, например, при необходимости учета гель-эффекта, вводятся более сложные зависимости скоростей элементарных реакций от температуры.

Оптимальный режим проведения процесса определяется из условия

$$r^* \in \operatorname{Arg\,min}\, Q[x(q^*, r)], \; r \in D_r,$$

где $q^* = (a_1^*, b_1^*, a_2^*, ..., b_8^*)$; Q — функционал, отражающий критерий оптимальности. Наиболее часто применяются технологические и экономические критерии. В качестве технологических критериев используются такие показатели качества, как производительность (единицы объема катализатора, реактора, каскада реакторов), характеристики продукта (молекулярная масса), объем реактора или каскада реакторов и т. д. Экономические критерии отражают себестоимость продукта или приведенные затраты на производство продукта при выполнении определенных технологических ограничений, связанных с выпуском продукции заданной марки.

В связи с приведенной укрупненной постановкой задачи управления процессом полимеризации возникает сразу несколько проблем, не позволяющих применять стандартное алгоритмическое и программное обеспечение. Эти проблемы рассмотрены далее и связаны, главным образом, с жесткостью систем обыкновенных дифференциальных уравнений, представляющих модель кинетики, а также с высокой степенью овражности целевых функционалов в задачах параметрической идентификации различных кинетических механизмов.

Указанные трудности привели к тому, что даже для хорошо изученных кинетических механизмов при разработке промышленных процессов полимеризации чаще применяются формальные или полуформальные (опирающиеся на упрощенные представления о механизме процесса) математические модели кинетики. Недостаток применения упрощенных моделей заключается в том, что они позволяют рассчитывать лишь брутто-скорость процесса и не дают информации о молекулярной структуре полимеров, определяющей потребительские свойства продукта. Последнее, в свою очередь, не позволяет целенаправленно управлять технологическим процессом для получения материала с заданными свойствами.

7.4.1. Кинетическая модель процесса термоинициированной полимеризации стирола в массе

Рассмотрим процедуру построения МК. При термоинициированной полимеризации стирола характерно протекание процесса по радикальному механизму с учетом следующих элементарных стадий (рис. 7.12).

1. Инициирование

$$2M \xrightarrow{k_d} \dot{R}_c.$$

2. Гибель первичных радикалов

$$2\dot{R}_c \xrightarrow{k_2'} Q.$$

3. Присоединение молекулы мономера к первичному радикалу

$$\dot{R}_c + M \xrightarrow{k_1} P_{1,1}.$$

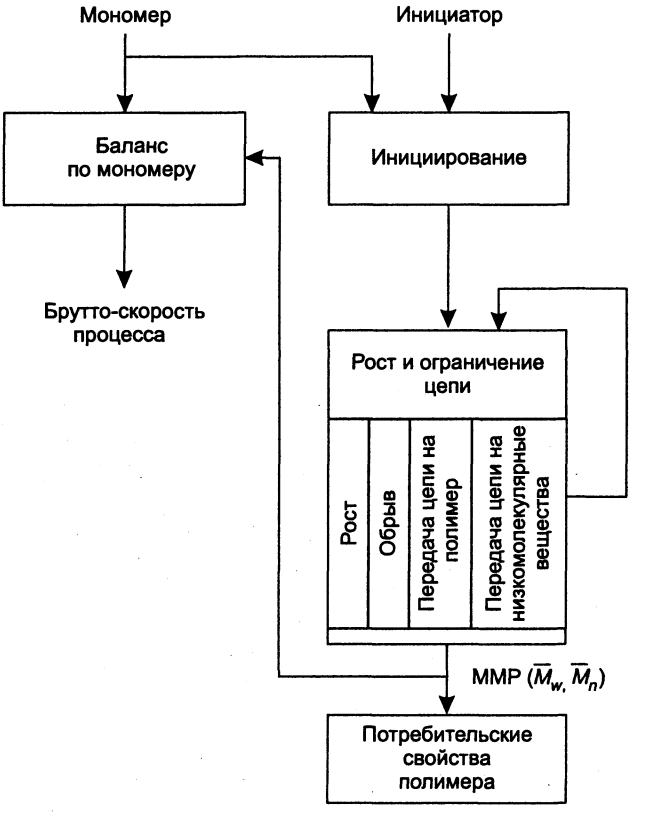

Рис. 7.12. Укрупненная схема модели кинетики полимеризационного процесса

4. Рост цепи

$$P_{n,l} + M \xrightarrow{nk_2} P_{n,l+1}.$$

5. Обрыв цепи рекомбинацией

$$P_{n,l} + P_{m,l'} \xrightarrow{nmk_3} P_{n+m-2,l+l'}.$$

6. Передача цепи на мономер

$$P_{n,l} + M \xrightarrow{nk_{4M}} P_{n-1,l} + \dot{M}.$$

7. Присоединение молекулы мономера к возбужденной молекуле мономера

$$M + \dot{M} \xrightarrow{k_{1M}} P_{1,2}.$$

8. Передача цепи на полимер

$$P_{n,l} + P_{m,l'} \xrightarrow{l' nk_{4P}} P_{n-1,l} + P_{m+1,l'}.$$

Здесь введены следующие обозначения: M — молекула мономера; $\dot{M}$ — возбужденная молекула мономера; $\dot{R}_c$ — первичный радикал; $P_{n,\,l}$ — макромолекула полимера длиной l с n активными центрами; Q — продукт, не принимающий дальнейшего участия в реакции; k_i — константы скоростей элементарных реакций.

Исходя из заданного множества элементарных стадий, может быть получена следующая система обыкновенных дифференциальных уравнений, моделирующая процесс полимеризации стирола

$$\begin{aligned}
\frac{d\dot{R}_c}{dt} &= k_d M^2 - \dot{R}_c(k_3'\dot{R}_c + k_1 M);\\
\frac{d\dot{M}}{dt} &= M(k_{4M}R - k_{1M}\dot{M});\\
\frac{dM}{dt} &= -M[k_d M + k_1\dot{R}_c + (k_2 + k_{4M})R + k_{1M}\dot{M}];\\
\frac{dR}{dt} &= M(k_1\dot{R}_c + k_{1M}\dot{M}) - R(k_{4M}M + k_3 R);\\
\frac{dP}{dt} &= k_2\dot{M}R;\\
\frac{dN}{dt} &= M(k_1\dot{R}_c + k_{1M}\dot{M}) - 0{,}5k_3R^2;\\
\frac{d\nu_1^1}{dt} &= \nu_2^0(k_{4P}R + k_2 M) - \nu_1^1(2k_3R + k_{4M}M + k_{4P}P - k_3\nu_2^0);\\
\frac{d\nu_0^2}{dt} &= \nu_1^1(2k_{2M}M + k_3\nu_1^1);\\
\frac{d\nu_2^0}{dt} &= k_3(\nu_2^0)^2 - 2\nu_2^0(2k_3R + k_{4M}M + k_{4P}P) + R[2k_3R + k_{4M}M +\\
&+2k_{4P}(P + \nu_1^1)] + M(k_{1M}\dot{M} + k_1\dot{R}_c).
\end{aligned} \tag{7.17}$$

Конечность системы уравнений (7.17) вызвана переходом от бесконечной системы уравнений относительно концентраций $P_{n,\,l}(t)$ макромолекул полимера к суммарным характеристикам распределения:

$R(t) = \sum_{n=0}^{\infty} n \sum_{l=1}^{\infty} P_{n,l}$ — суммарная концентрация активных центров;

$P(t) = \sum_{n=0}^{\infty} \sum_{l=1}^{\infty} l P_{n,l}$ — суммарная длина макромолекул;

$\nu_i^j = \sum_{n=0}^{\infty} n^i \int_0^{\infty} l^i P_n(l,t)\,dl$ — моменты концентрационного распределения.

Константы k_i неизвестны и должны определяться в результате решения задачи параметрической идентификации при известных начальных условиях для фазовых переменных модели (7.17):

$$\dot{R}_c = \dot{M} = R = P = N = \nu_1^1 = \nu_0^2 = \nu_2^0 = 0;\ M = 8.$$

Общая методика построения подобных моделей для различных полимеризационных процессов предложена в работе [70].

7.4.2. Методика воспроизведения моделей полимеризационных процессов

Практическое исследование моделей полимеризационных процессов, аналогичных (7.17), показало, что для воспроизведения таких моделей на компьютере должны применяться специальные методы численного интегрирования, отличные, в частности, от широко распространенных процедур Рунге—Кутта, Адамса и т. п.

Существенный разброс (порядка 10^{14}) в значениях констант скоростей элементарных реакций для типичных (порядка 413 °К) значений температур приводят к свойству жесткости систем дифференциальных уравнений (7.17), что определяет известные вычислительные трудности при их интегрировании с помощью явных схем численного интегрирования [49]. В то же время универсальные методики интегрирования жестких систем типа методов Гира, Хиндмарша и аналогичных им в силу своей громоздкости также требуют заметных вычислительных затрат, хотя и превосходят применительно к системе (7.17) явные схемы по скорости в 10–15 раз. В процессе идентификации возникает необходимость многократного интегрирования систем типа (7.17), и поэтому уже незначительная экономия машинного времени при решении задачи одновариантного анализа может оказаться решающей. Возникает проблема разработки специальных методов интегрирования с учетом характерных значений основных переменных и с применением минимального числа арифметических и логических операций на каждом шаге за счет отказа от универсальности общих схем численного интегрирования.

Для данной задачи может быть построена специальная методика численного воспроизведения моделей класса (7.17) на компьютере, основанная на неявных методах ломаных (НМЛ) и трапеций (НМТ) с аналитическим разрешением возникающих неявных зависимостей.

Для системы уравнений (7.17), записанной в виде

$$\frac{dx}{dt} = f(x),$$

уравнения НМЛ и НМТ получаются при аппроксимации интеграла в представлении

$$x(t_{k+1}) = x(t_k) + \int_{t_k}^{t_{k+1}} f[x(\tau)]\,d\tau = x(t_k) + \int_0^h f[x(t_k + \tau)]\,d\tau,$$

соответственно, по способу правых прямоугольников

$$x^{k+1} = x^k + hf^{k+1},\ f^{k+1} \triangleq f(x^{k+1}) \qquad (7.18)$$

и трапеций

$$x^{k+1} = x^k + \left(\frac{h}{2}\right)(f^k + f^{k+1}). \tag{7.19}$$

Для разрешения уравнения (7.18) относительно x^{k+1} можно воспользоваться следующим приемом. Запишем i-е уравнение (7.18) в виде

$$\frac{x_i^{k+1} - x_i^k}{h} = f_i^{k+1} = f_i(x^{k+1}). \tag{7.20}$$

Используя первые члены разложения Тейлора, можно приближенно представить правую часть (7.20) в виде

$$f_i^{k+1} \cong f_i^k + \sum_{j=1}^{n} \frac{\partial f_i(x^k)}{\partial x_j^k}(x_j^{k+1} - x_j^k),\ i \in [1{:}n] \tag{7.21}$$

с точностью до бесконечно малых высшего порядка относительно $x_j^{k+1} - x_j^k$ при $x^{k+1} \to x^k$. В результате формула НМЛ примет вид

$$\frac{x_i^{k+1} - x_i^k}{h} = f_i^k + \sum_{j=1}^{n} \frac{\partial f_i(x^k)}{\partial x_j^k}(x_j^{k+1} - x_j^k). \tag{7.22}$$

Аналогичным образом, НМТ (7.19) может быть записан в виде

$$\frac{x_i^{k+1} - x_i^k}{h} = f_i^k + 0{,}5\sum_{j=1}^{n} \frac{\partial f_i(x^k)}{\partial x_j^k}(x_j^{k+1} - x_j^k),\ i \in [1{:}n]. \tag{7.23}$$

Формулы (7.22), (7.23) обладают тем преимуществом перед исходными соотношениями (7.18), (7.19), что x_i^{k+1} входит линейно и уравнения (7.22), (7.23) могут рассматриваться как системы линейных алгебраических уравнений относительно неизвестных x_i^{k+1}, $i \in [1{:}n]$. Применение соотношений (7.22), (7.23), очевидно, эквивалентно разрешению нелинейных систем уравнений (7.18), (7.19) относительно x^{k+1} с помощью одного шага метода Ньютона.

Для динамических моделей, описывающих кинетику радикальной полимеризации, соотношения (7.22), (7.23) могут быть эффективно разрешены относительно x^{k+1} с получением явных представлений

$$x^{k+1} = G(x^k), \tag{7.24}$$

где вектор-функция G определяется видом вектор-функции $f(x)$ и используемым методом (7.22) или (7.23).

Основная идея предлагаемой методики численного интегрирования системы (7.17) для целей ее последующей параметрической идентификации заключается в использовании на первом этапе НМЛ (7.22) с достаточно малым значением шага дискретности h. Это приведет к достаточно точной аппроксимации основных, относительно медленно изменяющихся составляющих решения и к быстрому затуханию быстро изменяющихся компонентов. Затем, когда пройден

пограничный слой и быстро затухающие составляющие решения окажутся приблизительно равными нулю, осуществляется переход к методу (7.23), что позволяет существенно нарастить значение шага h при сохранении устойчивости и точности воспроизведения медленно изменяющихся составляющих (порядок точности метода (7.23) есть 0 (h^3), тогда как для метода (7.22) имеем ошибку порядка 0 (h^2)).

Заметим, что из условия сохранения устойчивости процесса численного интегрирования при использовании явных схем процедура наращивания шага h при $t > \tau_{\text{П.С}}$ не может применяться [49]. Кроме того, существенно, что НМЛ и НМТ по отдельности также не решают задачи. Действительно, НМЛ обладает (в линейном приближении) идеальной устойчивостью, но при этом невозможно значительно увеличить шаг интегрирования из-за потери точности воспроизведения медленных компонентов. С другой стороны, НМТ, как правило, позволяет выбрать достаточно большой шаг по точности, однако при этом возникают трудности в подавлении быстро затухающих составляющих решения.

На основе приведенных замечаний применялась комплексная методика численного интегрирования системы уравнений (7.17), усложненной дополнительным предположением

$$k_2 := k_2 \exp[\beta_2 (M - \overline{M})];$$
$$k_3 := k_3 \exp[\beta_3 (M - \overline{M})],$$

где $\overline{M}$ — некоторое фиксированное значение M. (При конкретных расчетах полагали $\overline{M} = M_0 = 8$.) Согласно этой методике, вводятся следующие параметры блока интегрирования:

- $\tau_{\text{П.С}}$ — длина пограничного слоя;
- h — начальный шаг интегрирования в пограничном слое неявным методом ломаных;
- E — коэффициент увеличения шага;
- N — целое число, определяющее при $t > \tau_{\text{П.С}}$ моменты изменения шага;
- T — промежуток интегрирования, $t \in [0, T]$.

При $t < \tau_{\text{П.С}}$ интегрирование ведется НМЛ с постоянным шагом h. По достижении точки $t = \tau_{\text{П.С}}$ осуществляется переход на НМТ с увеличением шага интегрирования в E раз через каждые N шагов ($h := h \cdot E$).

Экспериментально установлено, что для моделирования процесса термонаициированной полимеризации стирола (7.17) целесообразно полагать

$$\tau_{\text{П.С}} = 0{,}5;\ h = 0{,}01;\ E = 2;\ N = 4;$$

T определяется конкретным вариантом решения (это — десятки часов). НМЛ и НМТ использовались в виде модификаций (7.22), (7.23).

Изложенная методика проверялась с помощью просчета ряда контрольных вариантов и последующего сравнения с результатами интегрирования, полученными на основе более мощных вычислительных процедур (типа метода Гира) с гарантированной точностью получения решения. В достаточно широком диапазоне

изменения констант скоростей элементарных реакций было получено совпадение решений в четырех знаках мантиссы, что существенно выше точности используемых на этапе параметрической идентификации экспериментальных данных.

Время реализации одного варианта решения составляет менее 1 с при продолжительности T реального процесса порядка 10 ч.

Характерное время одновариантного анализа системы классическими явными схемами совпадает с продолжительностью T реального процесса. Процедуры Гира и им аналогичные при соответствующей настройке позволяют сократить это время до нескольких минут.

7.4.3. Параметрическая идентификация кинетических моделей полимеризационных процессов (полимеризация стирола)

При параметрической идентификации процесса полимеризации стирола на основе модели (7.17) в качестве экспериментальных данных используются зависимости от времени конверсии z, среднечисловой $\overline{M}_n$ и средневесовой $\overline{M}_W$ молекулярных масс, полученные в изотермическом режиме. Указанные характеристики следующим образом связаны с переменными системы (7.17) [9]:

$$z(t) = \frac{M(0) - M(t)}{M(0)}; \overline{M}_n = \frac{P}{N} m_0; \overline{M}_W = \frac{\nu_0^2}{P} m_0, \qquad (7.25)$$

где $m_0 = 106$ — молекулярный вес стирола; $N \overset{\Delta}{=} \nu_0^0$ — общее число макромолекул.

Показатели $\overline{M}_n$, $\overline{M}_W$ в значительной степени определяют технологические, физико-механические и прочие свойства материала. Поэтому оправдан подход, связанный с определением неизвестных параметров кинетической модели из условия совпадения экспериментальных и расчетных значений указанных характеристик. Структурная схема определения вектора кинетических констант приведена на рис. 7.13. Блок *1* реализует преобразование вектора y в множество векторов x^j фазовых переменных системы (7.17), соответствующих заданным моментам времени t_j, $j \in [1: N]$ съема экспериментальной информации. Численное интегрирование системы (7.17) производится с помощью изложенной в 7.4.2 методики. Блок *2* осуществляет перевод фазовых переменных модели в измеряемые экспериментально характеристики (7.25). В блоке *3* производится сравнение расчетных и экспериментальных значений измеряемых переменных. На выходе блока *3* формируется значение функционала $J(y)$, характеризующее невязку выходов модели и реального объекта для текущих значений параметров y_i, настраиваемых в блоке *4*.

В качестве критерия оптимальности используется взвешенный метод наименьших квадратов, согласно которому минимизируемый функционал имеет вид

$$J(y) = \frac{1}{2} \sum_{j=1}^{N} [\alpha_j (\overline{M}_{nj} - \hat{M}_{nj})^2 + \beta_j (\overline{M}_{Wj} - \hat{M}_{Wj})^2 + \gamma_j (z_j - \hat{z}_j)^2], \qquad (7.26)$$

где $y \triangleq (k_d, k'_3, k_1, k_2, k_3, k_{1M}, k_{4M}, k_{4P})$ — 8-мерный вектор неизвестных параметров модели; $\overline{M}_{nj} = M_n(t_j)$, $\overline{M}_{Wj} = M_W(t_j)$, $z_j = z(t_j)$ — расчетные значения; $\hat{M}_{nj}$, $\hat{M}_{Wj}$, $\hat{z}_j$ — экспериментальные значения в точках t_j, $j \in [1: N]$ съема измерительной информации; α_j, β_j, γ_j — нормирующие коэффициенты, равные обратным квадратам соответствующих экспериментальных значений.

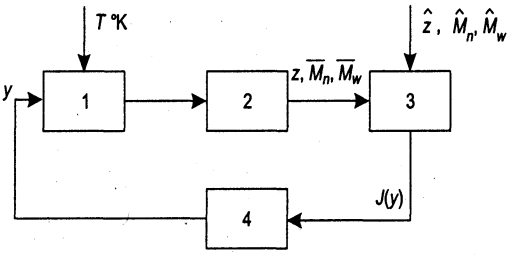

Рис. 7.13. Схема настройки кинетической модели на экспериментальные данные

Минимизация функционала (7.26) осуществлялась с помощью алгоритма RELEX при периодическом включении процедуры MIO, изложенной в 6.1.5. Необходимость перехода к процедуре MIO вызывалась достаточно высокими степенями обусловленности матриц Гессе $J''(y)$ (до 10^{10}) на значительных по продолжительности отрезках траектории спуска. Первые и вторые производные функционала определялись по методике, изложенной в 4.5.1, с конечноразностной аппроксимацией коэффициентов чувствительности первого порядка.

Подгонка модели (7.17) производилась при различных температурных режимах. Далее в качестве примера приводятся результаты, полученные для температуры 120 °C. Аналогичные результаты получены и для других значений температур на интервале, представляющем технологический интерес (100–170 °C).

В табл. 7.1 приведены выборочные данные лабораторных исследований, взятые в качестве экспериментальных значений при решении задачи параметрической идентификации.

Таблица 7.1

Время, ч	z	$\frac{\overline{M}_n}{10^6}$	$\frac{\overline{M}_W}{10^6}$
2,5	0,24	0,16	0,37
5,0	0,45	0,15	0,37
9,0	0,75	0,14	0,42
20,0	0,90	0,14	0,42

Указанная задача использовалась также как тестовая, на которой испытывались различные классические процедуры поисковой и параметрической идентификации: Давидона—Флетчера—Пауэлла, Хука—Дживса [71], вращения осей Розенброка, симплексный метод Нелдера—Мида.

Лучший результат $\bar{y}$, полученный для данной задачи с помощью попеременного применения указанных процедур, является неудовлетворительным:

$$\bar{y}_1 = k_d = 0{,}736586 \cdot 10^{-8}; \quad \bar{y}_5 = k_3 = 0{,}693696 \cdot 10^{8};$$
$$\bar{y}_2 = k_3' = 0{,}287481 \cdot 10^{11}; \quad \bar{y}_6 = k_{1M} = 0{,}246245 \cdot 10^{-4};$$
$$\bar{y}_3 = k_1 = 0{,}220636 \cdot 10^{8}; \quad \bar{y}_7 = k_{4M} = 0{,}495080 \cdot 10^{4};$$
$$\bar{y}_4 = k_2 = 0{,}534836 \cdot 10^{3}; \quad \bar{y}_8 = k_{4P} = 0{,}189181;$$
$$J(\bar{y}) = 0{,}29437. \qquad (7.27)$$

В табл. 7.2 представлены соответствующие значения конверсии z, среднечисловой $\overline{M}_n$ и средневесовой $\overline{M}_W$ молекулярных масс в соответствующих табл. 7.1 экспериментальных точках.

Таблица 7.2

Время, ч	z	$\frac{\overline{M}_n}{10^6}$	$\frac{\overline{M}_W}{10^6}$
2,5	0,28 (0,24)	0,16 (0,16)	0,27 (0,37)
5,0	0,44 (0,45)	0,16 (0,15)	0,30 (0,37)
9,0	0,58 (0,75)	0,16 (0,14)	0,35 (0,42)
20,0	0,76 (0,90)	0,16 (0,14)	0,47 (0,42)

Аппроксимация матрицы Гессе $J''(\bar{y})$ вырождена и имеет следующий спектр:

$$\lambda_1 = 0{,}13710 \cdot 10^{3}; \lambda_4 = 0{,}24827 \cdot 10^{-1};$$
$$\lambda_2 = 0{,}27146 \cdot 10^{2}; \lambda_5 = 0{,}30857 \cdot 10^{-2};$$
$$\lambda_3 = 0{,}34817 \cdot 10^{1}; \lambda_6 = \lambda_7 = \lambda_8 = 0{,}0.$$

Точка $\bar{y}$ была выбрана как начальная при идентификации методом RELEX в указанной ранее реализации. Полученные результаты представлены в табл. 7.3.

Таблица 7.3

Время, ч	z	$\frac{\overline{M}_n}{10^6}$	$\frac{\overline{M}_W}{10^6}$
2,5	0,24 (0,24)	0,15 (0,16)	0,32 (0,37)
5,0	0,48 (0,45)	0,15 (0,15)	0,36 (0,37)
9,0	0,70 (0,75)	0,15 (0,14)	0,40 (0,42)
20,0	0,87 (0,90)	0,15 (0,14)	0,45 (0,42)

Соответствующие значения констант скоростей существенно отличаются от (7.27) и имеют следующие значения:

$$y_1 = 0{,}123051 \cdot 10^{-11}; y_5 = 0{,}206709 \cdot 10^{3};$$
$$y_2 = 0{,}287480 \cdot 10^{11}; y_6 = 0{,}773434 \cdot 10^{-1};$$

$$y_3 = 0{,}296799 \cdot 10^{10};\ y_7 = 0{,}44929 \cdot 10^{4}; \qquad (7.28)$$
$$y_4 = 0{,}109727 \cdot 10^{3};\ y_8 = 0{,}410339 \cdot 10^{-1};$$
$$J(y) = 0{,}57178 \cdot 10^{-1}.$$

Значения параметров (7.28) позволяют весьма удовлетворительно описать экспериментальные данные табл. 7.1. Погрешность оказывается в пределах погрешности эксперимента (5–10%). Напротив, результаты табл. 7.2 значительно отличаются от требуемых значений (максимальная погрешность достигает 30%).

При решении данной задачи идентификации осуществлялась проверка получаемых результатов. А именно, в окрестности полученной точки выбирались контрольные близкие точки, на их основе строились имитированные «экспериментальные» данные и далее решались задачи идентификации по имитированным данным. Получаемые таким образом решения сравнивались с известными контрольными точками.

Основной эффект от применения изложенной методики идентификации получается за счет существенного расширения класса решаемых на стадии моделирования задач, а также в существенной экономии компьютерного времени (в 10–30 раз) по сравнению с методиками, основанными на традиционных схемах численного интегрирования и конечномерной оптимизации.

7.5. Идентификация моделей теплообменников атомных реакторов

Атомные реакторы и теплообменники являются типичными представителями объектов с распределенными параметрами, описываемых дифференциальными уравнениями в частных производных.

Уравнения теплообменника имеют вид

$$\begin{aligned}
&\frac{\partial T_1}{\partial t} + \omega_1(t)\frac{\partial T_1}{\partial l} + \kappa_1(t)(T_1 - T_3) = 0;\\
&\frac{\partial T_2}{\partial t} - \omega_2(t)\frac{\partial T_2}{\partial l} + \kappa_2(t)(T_2 - T_3) = 0;\\
&\delta\frac{\partial T_3}{\partial t} + \kappa_3(t)(T_3 - T_1) + \kappa_4(t)(T_3 - T_2) = 0.
\end{aligned} \qquad (7.29)$$

Здесь T_1, T_2, T_3 — температуры греющей жидкости, нагреваемой жидкости и разделительной стенки; $\omega_1(t)$, $\omega_2(t)$ — скорости теплоносителей, пропорциональные расходам $G_1(t)$, $G_2(t)$ первого и второго контуров; $\kappa_i(t)$ — величины, пропорциональные коэффициентам теплопередачи.

Решение системы (7.29) подчиняется граничным условиям:

$$\begin{aligned}
T_1(l, t)\big|_{l=0} &= T_{1BX}(t);\\
T_1(l, t)\big|_{l=L} &= T_{2BX}(t).
\end{aligned} \qquad (7.30)$$

Для построения системы управления объектом с распределенными параметрами необходимо многократно воспроизводить динамику объекта на основе решения системы (7.29) при условиях (7.30). Однако решение подобных задач аналитически невозможно, а численный подход недостаточно эффективен по временным затратам.

В работах [58], [59] предложена методика аппроксимации системы нелинейных уравнений в частных производных типа (7.29) с помощью системы линейных обыкновенных дифференциальных уравнений. Методика основана на аппроксимации частотных характеристик процессов с распределенными параметрами. Приближенные передаточные функции первого контура теплообменника имеют вид

$$
\begin{aligned}
W_1(p) &\triangleq \frac{\Delta T_{1вых}(p)}{\Delta T_{1вх}(p)} = \frac{k_1(1+\eta_{12}p)}{(1+\eta_{11}p)}\exp(-p\tau_1);\\
W_2(p) &\triangleq \frac{\Delta T_{1вых}(p)}{\Delta T_{2вх}(p)} = \frac{k_2}{(1+\eta_{21}p)(1+\eta_{22}p)};\\
W_3(p) &\triangleq \frac{\Delta T_{1вых}(p)}{\Delta G_1(p)} = \frac{k_3(1+2\xi_{31}\eta_{31}p+\eta_{31}^2p^2)}{(1+\eta_{32}p)(1+\eta_{33}p)(1+\eta_{34}p)^2};\\
W_4(p) &\triangleq \frac{\Delta T_{1вых}(p)}{\Delta G_2(p)} = \frac{k_4(1+2\xi_{2}\eta_{41}p+\eta_{41}^2p^2)}{(1+\eta_{42}p)(1+\eta_{43}p)(1+\eta_{44}p)^2}.
\end{aligned}
\tag{7.31}
$$

Температура на выходе первого контура теплообменника $\Delta T_{1вых}$ описывается уравнением

$$
\begin{aligned}
\Delta T_{1вых}(p) = W_1(p)\,\Delta T_{1вх}(p) + W_2(p)\,\Delta T_{2вх}(p) +\\
+ W_3(p)\,\Delta G_1(p) + W_{14}(p)\,\Delta G_2(p).
\end{aligned}
\tag{7.32}
$$

Второй контур теплообменника описывается аналогичными соотношениями.

Ставится задача построения уточненных значений параметров η_{ij} передаточных функций $W_1 \ldots W_8$ по результатам линеаризации исходной системы уравнений. Таким образом, в данном случае задача параметрической идентификации ставится и решается для упрощения исходной (глобальной) математической модели за счет ограничения области справедливости упрощенной модели окрестностью режима линеаризации модели (7.29).

Для каждой из передаточных функций $W_i(p)$ задача параметрической идентификации формулируется как оптимизационная с критерием, построенным по методу наименьших квадратов:

$$
J_i(\eta_i) = \sum_{k=1}^{11}\left\{[A_i(\omega_k)-\overline{A}_i(\omega_k)]^2 + [\varphi_i(\omega_k)-\overline{\varphi}_i(\omega_k)]^2\right\},\ i \in [1{:}8], \tag{7.33}
$$

где $\omega_k \in [0{,}62832\cdot10^{-2};\ 0{,}62832]$; A_i, φ_i — расчетные значения амплитуды и фазы, соответствующие приближенной передаточной функции W_i с параметрами $\eta_i = (\eta_{i1}, \ldots, \eta_{in})$, $\eta_{ik} \geq 0$; $\overline{A}_i$, $\overline{\varphi}_i$ — аппроксимируемые значения, полученные в результате линеаризации исходной системы уравнений в частных производных

относительно известных установившихся режимов и построения точных передаточных функций $\overline{W}_i$. Задачи (7.33) решались для каждого из трех рассматриваемых режимов работы теплообменника для значений $0{,}1P_н$, $0{,}5P_н$ и $P_н$, где $P_н$ — номинальная мощность.

Ограничения устранялись переходом к новым переменным y_{ik}: $\eta_{ik} = y_{ik}^2$. В качестве алгоритма оптимизации использовалась процедура RELEX (см. 5.3). Полученные результаты представлены в табл. 7.4.

Таблица 7.4

Погрешность		Первый контур				Второй контур			
		W_1	W_2	W_3	W_4	W_1	W_2	W_3	W_4
$P_н$	$\Delta\varphi^\circ_{max}$	0,24	0,10	0,29	0,218	0,20	0,32	0,08	0,07
	ΔA_{max}, %	0,13	0,371	0,37	0,272	0,79	0,16	0,10	0,12
$0{,}5P_н$	$\Delta\varphi^\circ_{max}$	0,53	0,25	—	—	0,43	0,78	0,20	0,18
	ΔA_{max}, %	0,27	0,11	—	—	0,92	0,31	0,40	0,50
$0{,}1P_н$	$\Delta\varphi^\circ_{max}$	0,75	0,30	—	—	0,87	0,91	0,14	0,26
	ΔA_{max}, %	0,41	0,98	—	—	0,71	0,42	0,58	0,67

При исходных значениях η_{ik} в рабочем диапазоне частот передаточные функции (7.31) реализуются с погрешностью по фазе $\Delta\varphi \le 3°$, а по амплитуде $\Delta A \le 2\%$. Согласно данным, приведенным в таблице, в результате уточнения параметров η_{ik} точность аппроксимации существенно возросла: максимальная погрешность по амплитуде на рабочих частотах составляет доли процента, а по фазе — доли градуса.

Далее по построенным передаточным функциям строятся соответствующие системы линейных обыкновенных дифференциальных уравнений (с отклоняющимся аргументом) с известными корнями характеристического полинома, равными $-1/\eta_{ik}$. Последнее позволяет найти аналитические решения, справедливые в окрестности выбранного режима линеаризации и используемые, например, в процессе управления атомным реактором как объектом эксперимента [59].

Литература

1. *Алберт А.* Регрессия, псевдоинверсия и рекуррентное оценивание. М.: Наука, 1977.
2. *Андреев Н. И.* Корреляционная теория статистически оптимальных систем. М.: Наука, 1966.
3. *Аоки М.* Введение в методы оптимизации. М.: Мир, 1977.
4. *Аттетков А. В., Галкин С. В., Зарубин В. С.* Методы оптимизации. М.: Изд-во МГТУ им. Н. Э. Баумана, 2001.
5. *Базара М., Шетти К.* Нелинейное программирование. М.: Мир, 1982.
6. *Батищев Д. И.* Поисковые методы оптимального проектирования. М.: Сов. радио, 1975.
7. *Бахвалов Н. С.* Численные методы. М.: Наука, 1973.
8. *Беляков Ю. Н., Курмаев Ф. А., Баталов Б. В.* Методы статистических расчетов микросхем на ЭВМ. М.: Радио и связь, 1985.
9. *Берлин А. А., Вольфсон С. А.* Кинетический метод в синтезе полимеров. М.: Химия, 1973.
10. *Бесекерский В. А., Попов Е. П.* Теория систем автоматического управления. М.: Наука, 1972.
11. *Болнокин В. Е., Чинаев П. И.* Анализ и синтез систем автоматического управления на ЭВМ. Алгоритмы и программы. М.: Радио и связь, 1986.
12. *Брамеллер А., Аллан Р., Хэмэм Я.* Слабозаполненные матрицы. М.: Энергия, 1979.
13. *Васильев Ф. П.* Численные методы решения экстремальных задач. М.: Наука, 1980.
14. *Вентцель Е. С.* Исследование операций. М.: Сов. радио, 1972.
15. *Вязгин В. А., Федоров В. В.* Математические методы автоматизированного проектирования. М.: Высш. шк., 1989.
16. *Гантмахер Ф. Р.* Теория матриц. М.: Наука, 1966.
17. *Геминтерн В. И., Каган Б. М.* Методы оптимального проектирования. М.: Энергия, 1980.
18. *Гермейер Ю. Б.* Игры с непротивоположными интересами. М.: Наука, 1976.
19. *Гилл Ф., Мюррей У., Райт М.* Практическая оптимизация. М.: Мир, 1985.
20. *Гроп Д.* Методы идентификации систем. М.: Мир, 1979.
21. *Гроссман К., Каплан А. А.* Нелинейное программирование на основе безусловной минимизации. Новосибирск: Наука, 1981.
22. *Дейч А. М.* Методы идентификации динамических объектов. М.: Энергия, 1979.

23. *Джордж А., Лю Дж.* Численное решение больших разреженных систем уравнений. М.: Мир, 1984.
24. *Деннис Дж., Шнабель Р.* Численные методы безусловной оптимизации и решения нелинейных уравнений. М.: Мир, 1988.
25. *Дубов Ю. А., Травкин С. И., Якимец В. Н.* Многокритериальные методы формирования и выбора вариантов систем. М.: Наука, 1986.
26. *Евтущенко Ю. Г.* Методы решения экстремальных задач и их применение в системах оптимизации. М.: Наука, 1982.
27. *Ермольев Ю. М.* Методы стохастического программирования. М.: Наука, 1976.
28. *Зангвилл У. И.* Нелинейное программирование. М.: Сов. радио, 1973.
29. *Карманов В. Г.* Математическое программирование. М.: Наука, 1975.
30. *Катковник В. Я.* Непараметрическая идентификация и сглаживание данных. М.: Наука, 1985.
31. *Каханер Д., Моулер К., Нэш С.* Численные методы и программное обеспечение. М.: Мир, 1998.
32. *Ланнэ А. А.* Оптимальный синтез линейных электронных схем. М.: Связь, 1978.
33. Оптимальная реализация линейных электронных RLC-схем / А. А. Ланнэ, Е. Д. Михайлова, Б. С. Саркисян, Я. Н. Матвийчук. Киев: Наук. Думка, 1982.
34. *Ланцош К.* Практические методы прикладного анализа. М.: Физматгиз, 1961.
35. *Лоусон Ч., Хенсон Р.* Численное решение задач метода наименьших квадратов. М.: Наука, 1986.
36. *Лыпарь Ю. И.* Автоматизация проектирования избирательных усилителей и генераторов. Л.: Изд-во Ленинградского Гос. ун-та, 1983.
37. *Лэм Г.* Аналоговые и цифровые фильтры. М.: Мир, 1982.
38. *Люстерник Л. А., Соболев В. И.* Элементы функционального анализа. М.: Наука, 1965.
39. *Маркус М., Минк Х.* Обзор по теории матриц и матричных неравенств. М.: Наука, 1972.
40. *Моисеев Н. Н.* Математические задачи системного анализа. М.: Наука, 1981.
41. *Немировский А. С., Юдин Д. Б.* Сложность задач и эффективность методов оптимизации. М.: Наука, 1979.
42. *Норенков И. П.* Введение в автоматизированное проектирование технических устройств и систем. М.: Высш. шк., 1986.
43. *Норенков И. П., Мулярчик С. Г., Иванов С. Р.* Экстремальные задачи при схемотехническом проектировании в электронике. Минск: Изд-во БГУ, 1976.
44. *Ортега Дж., Рейнболдт В.* Итерационные методы решения нелинейных систем уравнений со многими неизвестными. М.: Мир, 1975.
45. *Парлетт Б.* Симметричная проблема собственных значений. М.: Мир, 1983.
46. *Первозванский А. А., Гайцгори В. Г.* Декомпозиция, агрегирование и приближенная оптимизация. М.: Наука, 1979.
47. *Подиновский В. В., Ногин В. Д.* Парето-оптимальные решения многокритериальных задач. М.: Наука, 1982.
48. *Поляк Б. Т.* Введение в оптимизацию. М.: Наука, 1983.
49. *Ракитский Ю. В., Устинов С. М., Черноруцкий И. Г.* Численные методы решения жестких систем. М.: Наука, 1979.
50. *Растригин Л. А.* Системы экстремального управления. М.: Наука, 1974.

51. *Растригин Л. А.* Современные принципы управления сложными объектами. М.: Сов. радио, 1980.
52. *Растригин Л. А., Маджаров Н. Е.* Введение в идентификацию объектов управления. М.: Энергия, 1977.
53. *Рей У.* Методы управления технологическими процессами. М.: Мир, 1983.
54. *Розенброк Х., Стори С.* Вычислительные методы для инженеров-химиков. М.: Мир, 1968.
55. *Соболь И. М.* Точки, равномерно заполняющие многомерный куб. М.: Знание, 1985.
56. *Стренг Г.* Линейная алгебра и ее применения. М.: Мир, 1980.
57. *Стронгин Р. Г.* Численные методы в многоэкстремальных задачах. М.: Наука, 1978.
58. *Тарасов В. С., Веренинов И. А., Ерунов В. Я.* Моделирование технологических процессов с распределенными параметрами. Л.: Изд-во ЛПИ, 1982.
59. *Тарасов В. С., Савченко Е. С.* Параметрическая идентификация объектов эксперимента с распределенными параметрами с использованием системного метода оптимизации // Труды ЛПИ № 381. Л.: Изд-во ЛПИ, 1982.
60. *Таха Х.* Введение в исследование операций. Кн. 1, 2. М.: Мир, 1985.
61. *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. М.: Наука, 1979.
62. *Уайлд Д.* Оптимальное проектирование. М.: Мир, 1981.
63. *Уилкинсон Дж. Х.* Алгебраическая проблема собственных значений. М.: Наука, 1970.
64. *Уилкинсон Дж. Х., Райнш С.* Справочник алгоритмов на языке АЛГОЛ. М.: Машиностроение, 1976.
65. *Фаддеев Д. К., Фаддеева В. Н.* Вычислительные методы линейной алгебры. М.: Физматгиз, 1963.
66. *Федоренко Р. П.* Приближенное решение задач оптимального управления. М.: Наука, 1978.
67. *Фельдбаум А. А.* Вычислительные устройства в автоматических устройствах. М.: Физматгиз, 1959.
68. *Форсайт Дж., Моулер К.* Численное решение систем линейных алгебраических уравнений. М.: Мир, 1969.
69. *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений. М.: Мир, 1980.
70. *Хохлов В. А., Любецкий С. Г.* Особенности применения преобразующих функций для математического описания кинетики полимеризационных процессов // Полимеризационные процессы. Аппаратное оформление и математичское моделирование / ОНПО «Пластполимер». Л., 1976.
71. *Химмельблау Д.* Прикладное нелинейное программирование. М.: Мир, 1975.
72. *Черноруцкий И. Г.* Оптимальный параметрический синтез: электротехнические устройства и системы. Л.: Энергоатомиздат, 1987.
73. *Черноруцкий И. Г.* Методы оптимизации. СПб.: Изд-во СПбГТУ, 1998.
74. *Черноруцкий И. Г.* Методы оптимизации и принятия решений. СПб.: Лань, 2001.
75. *Эйкхофф П.* Основы идентификации систем управления. М.: Мир, 1975.

Учебное издание

Черноруцкий Игорь Георгиевич

Методы оптимизации в теории управления: Учебное пособие

Главный редактор	*Е. Строганова*
Заведующий редакцией	*И. Корнеев*
Руководитель проекта	*Е. Саргаева*
Научный редактор	*В. Родионов*
Иллюстрации	*В. Шендерова*
Художники	*Н. Биржаков*
Корректоры	*Н. Рощина, Н. Солнцева*
Верстка	*И. Актанова*

Лицензия ИД № 05784 от 07.09.01.

Подписано к печати 20.09.03. Формат 70×100 $^{1}/_{16}$. Усл. п. л. 20,64.
Тираж 4000. Заказ 418

ООО «Питер Принт», 196105, Санкт-Петербург, ул. Благодатная, д. 67в.

Налоговая льгота — общероссийский классификатор продукции ОК 005-93, том 2; 95 3005 — литература учебная.

Отпечатано с готовых диапозитивов в ФГУП ордена Трудового Красного Знамени «Техническая книга» Министерства Российской Федерации по делам печати, телерадиовещания и средств массовых коммуникаций 198005, Санкт-Петербург, Измайловский пр., 29