И. Г. Черноруцкий

МЕТОДЫ ОПТИМИЗАЦИИ

Компьютерные технологии

К оптимизации сводится огромное количество задач компьютерного моделирования из различных предметных областей. Широко распространено мнение о том, что для успешного решения таких задач достаточно иметь стандартное программное обеспечение и современные аппаратные средства. Однако, как показывает практика, прежде всего необходима высокая профессиональная подготовка специалистов в этой весьма сложной области компьютерных вычислений. Слишком много проблем возникает как при проведении реальных вычислений, так и при интерпретации получаемых результатов.

Научное Издание

И. Г. Черноруцкий

МЕТОДЫ ОПТИМИЗАЦИИ
Компьютерные технологии

Рекомендовано Учебно-методическим объединением по университетскому политехническому образованию в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлениям подготовки магистров «Системный анализ и управление» и «Информатика и вычислительная техника»

Санкт-Петербург

«БХВ-Петербург»

2011

УДК 681.3.06
ББК 32.973.26-018.2
Ч-49

Черноруцкий И. Г.

Ч-49 Методы оптимизации. Компьютерные технологии. — СПб.: БХВ-Петербург, 2011. — 384 с.: ил.

ISBN 978-5-9775-0784-4

В книге изложены теория, методы и основные элементы компьютерных технологий оптимизации. Наиболее подробно описаны методы решения конечномерных задач с учетом таких особенностей, как невыпуклость и плохая обусловленность минимизируемых функционалов. Рассмотрены многопараметрические и многокритериальные задачи. В качестве модельной предметной области выбраны задачи управления. Рассматриваемый материал иллюстрируется многочисленными примерами.

Для научных работников, преподавателей и студентов технических вузов

УДК 681.3.06
ББК 32.973.26-018.2

Группа подготовки издания:

Главный редактор	*Екатерина Кондукова*
Зам. главного редактора	*Татьяна Лапина*
Зав. редакцией	*Григорий Добин*
Редактор	*Анна Кузьмина*
Компьютерная верстка	*Натальи Караваевой*
Корректор	*Виктория Пиотровская*
Оформление обложки	*Елены Беляевой*
Зав. производством	*Николай Тверских*

Лицензия ИД № 02429 от 24.07.00. Подписано в печать 29.07.11.
Формат 70×100 1/16. Печать офсетная. Усл. печ. л. 30,96.
Тираж 1000 экз. Заказ №
"БХВ-Петербург", 190005, Санкт-Петербург, Измайловский пр., 29.

Санитарно-эпидемиологическое заключение на продукцию
№ 77.99.60.953.Д.005770.05.09 от 26.05.2009 г. выдано Федеральной службой
по надзору в сфере защиты прав потребителей и благополучия человека.

Отпечатано с готовых диапозитивов
в ГУП "Типография "Наука"
199034, Санкт-Петербург, 9 линия, 12.

ISBN 978-5-9775-0784-4

Оглавление

Предисловие

Эта книга представляет собой расширенную и в значительной степени модифицированную версию более ранних изданий: "Методы оптимизации"[1] и "Методы оптимизации в теории управления"[2].

В последнее время термин "оптимизация" часто употребляется "всуе", например, в области интернет-технологий или даже в парикмахерском деле — "оптимизация прически". Мы будем использовать его в обычном математическом смысле. В книге рассматриваются проблемы построения "наилучших" систем с помощью компьютерного моделирования. Оптимизируемые системы могут принадлежать к различным предметным областям. Например, это могут быть технические, экономические, экологические, программные, физические и другие системы. Важно лишь, чтобы существовали компьютерные модели, позволяющие вычислять "качество" или показатели эффективности системы в зависимости, например, от выбираемых (управляемых) параметров. В последнем случае имеем важный класс задач "параметрической оптимизации". Могут возникать и более сложные ситуации, когда требуется оптимальным образом выбрать не набор чисел (значений параметров), а некоторую функцию или функции. В особо сложных случаях речь не может идти о построении оптимальной системы. Требуется решить более "скромную" задачу — построить механизм последовательного улучшения характеристик оптимизируемой системы.

Соответствующие компьютерные технологии и являются предметом дальнейшего изложения.

В основу книги положен курс лекций по методам оптимизации, который читается автором на протяжении ряда лет на факультете технической кибернетики Санкт-Петербургского государственного политехнического университета. Тем не менее, книга позиционируется как монография, т. к. является "научным трудом, углубленно разрабатывающим ограниченный круг вопросов" (см. Современный словарь иностранных слов. — М.: Рус. Яз., 1992). Обсуждаемые в книге вопросы теории и практики оптимизации, действительно, не содержат многих традиционных разделов. С одной стороны, это вызвано ограниченным объемом, т. к. не ставилась задача создания "энциклопедии" или стандартного всеохватывающего учебника по оп-

[1] Черноруцкий И. Г. Методы оптимизации. — СПб.: Изд-во СПбГТУ, 1998.
[2] Черноруцкий И. Г. Методы оптимизации в теории управления. — СПб.: Питер, 2004.

тимизации. С другой стороны, и это главное, в книге отражены личные пристрастия автора, основанные на более чем тридцатилетнем опыте решения реальных задач. Многие суждения и выводы являются субъективными и основаны все на том же опыте. По-видимому, это нормальная ситуация, тем более что автор не претендует на "истину в последней инстанции" и всегда готов корректировать свою точку зрения.

В частности, в книге не рассмотрены специальные методы выпуклой оптимизации, методы линейного программирования, методы учета ограничений, основанные на операции проектирования на допустимое множество. Соответствующий материал присутствует во многих учебниках и монографиях по оптимизации, и автор здесь счел возможным его опустить. Не рассмотрены и многие другие вопросы, например, методы, основанные на теории двойственности. Так уж случилось, что автору в своей практической деятельности обычно "не везло". Задачи, которые приходилось решать, оказывались нелинейными, невыпуклыми, негладкими, плохо обусловленными *(жесткими)*. Часто встречались многопараметрические задачи с большим числом аргументов минимизируемых функционалов или бесконечномерные задачи. Поэтому как раз эти вопросы в книге обсуждаются достаточно подробно (за исключением бесконечномерных задач), вплоть до конкретных алгоритмов и примеров решения реальных задач. Красной нитью через все изложение проходит проблема *жесткости*, или *овражности*, присущая многим практическим оптимизационным (или вариационным) задачам. Один из видных разработчиков и исследователей *методов случайного поиска* Л. А. Растригин определял овражные оптимизационные задачи, как задачи, в которых вероятность убывания целевой функции при случайном выборе направления спуска чрезвычайно мала. И именно поэтому в данной книге из рассмотрения исключены также алгоритмы случайного поиска и, в частности, генетические алгоритмы. Их относительно невысокая эффективность для жестких или овражных задач подтверждается и практически.

Несмотря на кажущуюся неполноту изложения, на основе представленного материала, вообще говоря, можно решать достаточно разнообразные прикладные оптимизационные задачи, выбирая соответствующую траекторию в том иерархическом арсенале методов и средств, которые все-таки содержатся в книге. Эта траектория сама по себе может и не быть оптимальной, но, как правило, она приводит к результату.

Книга предназначена для многих категорий читателей. Во-первых, это студенты технических вузов, имеющих разделы по теории оптимизации в своих учебных планах. Во-вторых, это компьютерные математики-практики (современные инженеры). В-третьих, это могут быть профессиональные математики, разрабатывающие теорию оптимизации в функциональных пространствах. Дело в том, что в книге представлены некоторые результаты собственной научной деятельности автора и соответствующие новые подходы к решению жестких конечномерных задач. Обобщение этих результатов на бесконечномерный случай может представлять определенный интерес для математиков-теоретиков.

При изложении методов, алгоритмов и сопутствующих вычислительных технологий автор следовал системе взглядов Н. Н. Моисеева, который говорил, например, в предисловии редактора перевода к книге Ж. Сеа [63]: "Утверждение, что сходи-

мость — это главное качество предлагаемого алгоритма, казалось вначале очевидным. Однако очень скоро обнаружилось, что это качество не является ни достаточным, ни необходимым для эффективного окончания вычислений, т. е. главной цели, ради которой создавался алгоритм". И еще: "Понимание того факта, что классическое представление о сходимости алгоритмов, как об основном содержании теории вычислительных процессов, не соответствует требованиям к анализу, которые выдвигает практика, является основным стимулом научных поисков". При этом Н. Н. Моисеев, являясь в первую очередь математиком, ни в коей мере не принижал значение выработанных в математике принципов.

Обсуждаемые в книге методы, технологии и алгоритмы, в основном, также получили соответствующее теоретическое обоснование вплоть до доказательства сходимости для некоторых стандартных случаев. Но наибольшую роль здесь играют исследования по обоснованию различных аспектов практического применения методов, отличных от установления самого факта сходимости.

Книга условно состоит из двух частей: *глава 1* и *главы 2—8*.

В *главе 1* рассмотрены общие математические структуры функционального анализа, позволяющие сформулировать основные положения теории оптимизации в банаховых и гильбертовых пространствах. Получены необходимые и достаточные условия экстремума. Обсуждаются с тех же единых позиций некоторые понятия вариационного исчисления. Например, уравнения Эйлера вариационного исчисления получаются как необходимые условия экстремума, вытекающие из равенства нулю градиента соответствующего функционала. Вводится определение минимизирующей последовательности в функциональном пространстве, как основного понятия практической оптимизации. Рассмотрены общие методы построения минимизирующих последовательностей (методы минимизации) в гильбертовых пространствах. При желании материал *главы 1* может быть опущен, если читателя интересуют исключительно прикладные вопросы оптимизации в конечномерных пространствах. Но этот материал оказывается совершенно необходимым в бесконечномерном случае и, в частности, при решении задач оптимального управления. Изложение является замкнутым, элементарным и формально не требует предварительной подготовки в этой области. Однако неявно предполагается, что читатель достаточно хорошо знаком с теорией конечномерных векторных пространств.

Главы 2—8 посвящены конечномерным задачам оптимизации (математическому программированию).

Основное внимание в этих главах уделено практическим вопросам оптимизации, тесно связанным с работой в различных предметных областях. При этом 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 руководств по теории и методам оптимизации, целесообразно вести изложение "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 в области компьютерного моделирования систем. Таким образом, эта предметная область уже знакома большинству читателей данной книги. 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, связанные как с нашей повседневной деятельностью, так и возникающие внутри технических, экономических, программных, эко и других систем. Обычно речь идет о некотором наилучшем в определенном смысле, т. е. "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 сводится к решению целой цепочки различных оптимизационных задач. В этом смысле можно утверждать, что методы оптимизации являются важным алгоритмическим фундаментом современной теории управления.

Специфика формулируемых 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 оптимизационных вычислений, особенно в конечномерных пространствах, показывает, что, как правило, невозможно при создании прикладных программных систем эффективно использовать универсальное алгоритмическое обеспечение "в чистом виде" из-за резкого понижения скорости сходимости предлагаемых оптимизационных 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 с уже упомянутым специальным случаем явления плохой обусловленности, когда целевой функционал имеет "жесткий" или "овражный" характер, т. е. резко возрастает по одним направлениям и слабо изменяется вдоль других, что и вызывает указанные трудности.

Высокие степени жесткости возникают не в исключительных "паталогических" случаях, а отражают обычную ситуацию при решении практически любой прикладной задачи конечномерной оптимизации. Наиболее остро проблема жесткости стоит при решении задач параметрической идентификации объектов компьютерно-

го моделирования и управления, а также в задачах оптимального параметрического синтеза как реальных, т. е. 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" [76] были, по существу, рассмотрены все основные признаки овражной ситуации, включая явление "заклинивания" и понятие многомерного дна оврага. Широкую известность проблема овражности приобрела после работ И. М. Гельфанда и М. Л. Цетлина (1961—1962 г.), посвященных методам управления техническими системами и применению поисковых процедур в системах автоматической оптимизации. Эти исследования привели к созданию известного "метода оврагов". В 1967 г. 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 [54]. Значительный вклад в теорию плохо обусловленных экстремальных задач внесли работы Ю. В. Ракитского, возглавлявшего ленинградскую школу по изучению проблемы жесткости и указавшего на принципиальную связь явления овражности с концепцией жесткости систем обыкновенных дифференциальных уравнений. Эти исследования были поддержаны будущим академиком РАН Н. С. Бахваловым и нашли отражение в брошюре Ю. В. Ракитского, С. М. Устинова и И. Г. Черноруцкого "Численные методы решения жестких систем обыкновенных дифференциальных уравнений" [56] и монографии тех же авторов "Численные методы решения жестких систем" [57]. В этих работах впервые была разработана общая теория систем, описываемых жесткими дифференциальными моделями, включающая асимптотическую теорию жестких систем и основы теории минимизации жестких функционалов.

Однако, несмотря на важность полученных результатов, проблема ими не исчерпывалась. Оставался нерешенным целый ряд существенных вопросов, связанных с развитием теории жестких (овражных) оптимизационных задач, а также общих принципов построения методов и алгоритмов конечномерной 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 аргументов. Такие исследования 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. Данная книга, по мнению автора, позволяет частично восполнить этот пробел.

Дадим рекомендации читателям-непрофессионалам, собирающимся использовать книгу в качестве учебного пособия по соответствующим дисциплинам или разделам дисциплин. 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 для многих, изучающих теорию оптимизации, особенно студентов вузов, 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 работы. В книге даются ссылки на учебную и научную литературу. Многие из этих публикаций 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 публикаций.

Полное понимание представленных в книге материалов предполагает знакомство со стандартными курсами математики для высших технических учебных заведений (содержащими, в частности, достаточные сведения по линейной и матричной алгебре), численного анализа, теории вероятностей (в частности, корреляционной теории) и основ теории управления. Если вы владеете этим материалом в обычном для технических вузов объеме, то никаких проблем возникнуть не должно.

Для облегчения работы над книгой и согласования применяемых терминов и обозначений в *главе 2* даны необходимые для данного курса общие сведения из теории управления и сформулированы характерные постановки оптимизационных задач.

Представленные в книге элементы теории, методы и алгоритмы, которые можно отнести к разряду новых или, говоря более скромно, нетрадиционных, появились не вчера. В разное время и при разных обстоятельствах они обсуждались с акад. АН СССР (впоследствии РАН) Н. Н. Моисеевым, чл. корр. РАН П. А. Бутыриным, профессорами В. А. Городецким, В. Я. Катковником, А. А. Первозванским, Ю. В. Ракитским, Л. А. Растригиным, С. М. Устиновым и др. Всем им автор благодарен за сделанные замечания и поддержку.

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
$D = x \in H \mid P\ x$	Подмножество элементов множества H, обладающих свойством $P\ x$
$\varnothing$	Пустое множество
$C^k\ D$	Множество k раз непрерывно дифференцируемых на множестве D функций
$.^+$	Псевдообратная матрица
МНК	Метод наименьших квадратов
$\mathrm{diag}\,\lambda_i$	Диагональная матрица
$\mathrm{cond}\ A$	Спектральное число обусловленности матрицы A
J', J''	Вектор градиента и матрица вторых производных функционала $J\ x$
$\arg\min J\ x$	Минимизатор функционала $J\ x$
$\mathrm{Arg}\min J\ x$	Множество всех минимизаторов функционала $J\ x$

Сделаем также некоторые терминологические замечания по тексту книги. Согласно общепринятым математическим канонам функционального анализа (см., например, [43]) 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 тоже 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, т. е. 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.

Введение.
Постановка задачи оптимизации

Исследование способов получать те или иные результаты (в широком смысле слова) наилучшим, наивыгоднейшим способом есть задача теории оптимизации.

Математически проблема оптимизации описывается следующим образом. Рассматривается некоторое непустое множество U элементов (вообще говоря, произвольной природы), называемое множеством допустимых элементов. Рассматривается некоторая ограниченная снизу функция (функционал) J, ставящая в соответствие каждому элементу множества допустимых элементов какое-либо действительное (вещественное) число. Требуется найти *минимизатор* — такой элемент $u^* \in U$, которому соответствует минимальное значение J:

$$J\ u^*\ \le J\ u\ ,\quad \forall u \in U. \tag{1}$$

Заданная функция J формализует понятие качества оптимизируемого объекта и часто (не совсем строго) называется критерием. На практике J может также отражать стоимость, время и т. п. Если заменить J на $-J$, то задача минимизации превратится в задачу максимизации и наоборот.

Приведенная постановка задачи выглядит достаточно общей, однако необходимо сделать некоторые уточнения и замечания.

Во-первых, здесь имеется в виду задача оптимизации с одним критерием оптимальности, задаваемым функционалом J. Возможны и более общие *многокритериальные* постановки задачи. Для конечномерного случая такие задачи будут рассмотрены далее достаточно подробно.

Во-вторых, задача (1) в общем случае может и не иметь решения. Искомый элемент u^* *(минимизатор)* может отсутствовать среди допустимых элементов. Например, функция $\exp(x)$, $x \in R$ не имеет минимизаторов на вещественной оси (множестве действительных чисел) R, хотя и ограничена снизу (инфимум равен нулю). Целесообразно поэтому рассматривать более общую задачу — задачу построения минимизирующих последовательностей: требуется найти последовательность элементов u_n множества U, удовлетворяющих условию:

$$\lim_{n\to\infty} J\ u_n\ = \inf_{x\in U} J\ u\ . \tag{2}$$

Последовательность u_n называется *минимизирующей*. Здесь также возникают определенные трудности. Если минимизатор u^* существует (понятно, что он не обязан быть единственным), то минимизирующая последовательность может к нему не сходиться. Приходим к так называемым *некорректным задачам оптимизации*. И так как большинство численных методов оптимизации являются средством построения именно минимизирующих последовательностей, то необходимо совершенно четко понимать, что задача поиска хороших оценок соответствующих минимизаторов в случае некорректности не может быть решена без дополнительного исследования сходимости. Мы здесь должны различать известные из анализа типы сходимости: *по аргументу* и *по функционалу*. Чаще всего для практических приложений достаточно иметь сходимость по функционалу, т. к. неважно при каких допустимых значениях аргумента *u* достигается приемлемо малое значение *J* ("задачи аппроксимации"). И наоборот, иногда сама задача оптимизации ставится лишь для того, чтобы получить механизм определения хороших оценок минимизатора u^* ("задача идентификации"). В последнем случае совершенно необходимо исследовать сходимость по аргументу.

И наконец, в-третьих. Иногда в качестве множества допустимых элементов *U* выступает некоторое подмножество основного (базового) множества *V*, которое также задается. В этом случае имеем дело с т. н. задачей оптимизации с ограничениями, позволяющими выделить множество *U* из общего пространства *V*. Обычно ограничения задаются с помощью системы уравнений (например, дифференциальных) и неравенств.

Для математического анализа основных задач (1), (2) приходится наделять объекты, участвующие в постановке задач, определенными математическими структурами. Так множество *U* или *V* обычно считаются наделенными структурой банахова, гильбертова или конечномерного евклидова пространства. Это, как правило, оказывается достаточным обобщением структур, с которыми мы имеем дело при разработке и применении теории оптимизации. То же относится и к критериальной функции *J*, которая часто наделяется такими свойствами, как ограниченность, непрерывность, гладкость, выпуклость и т. д.

Отметим теперь несколько конкретных реализаций приведенных постановок задач. Для определенности будем говорить о задаче (1). Рассмотрим две альтернативные теории: конечномерные задачи оптимизации и бесконечномерные задачи.

Конечномерные задачи. Задачи математического программирования. *Математическим программированием* (МП) называется теория конечномерных оптимизационных задач (1). Минимизатор ищется среди элементов множества $U \subset R^n$. R^n наделяется структурой линейного *n*-мерного векторного пространства *(см. далее)*. Это серьезное ограничение, т. к. результаты теории математического программирования непосредственно неприменимы, если множество *U* является множеством объектов, не описываемых набором из *n* чисел, а является, например, множеством непрерывных на промежутке a, b функций. Множество допустимых элементов *U* в МП задается различными способами, например, соответствующей

системой неравенств (или равенств и неравенств). Для формулировки задачи математического программирования вводится целевой функционал

$$J : R^n \to R$$

и функционалы ограничений

$$G_i : R^n \to R,\ i = 1,\ ...,\ m.$$

Множество допустимых элементов U может задаваться системой неравенств, например, вида:

$$U = \{ u \mid u \in R^n;\ G_i(u) \geq 0,\ i = 1,\ ...,\ m \}.$$

Существуют и иные способы задания множества U.

Замечание. Равенство можно представить в виде двух разнонаправленных неравенств, но обычно целесообразно работать непосредственно с равенствами, если они изначально заданы в явном виде.

Если ограничения отсутствуют, то имеем задачу математического программирования без ограничений. Такие задачи в последующих разделах будут рассмотрены подробно, т. к. они часто оказываются алгоритмическим фундаментом при решении более общих задач.

В случае, если основные функционалы задачи МП являются линейными, то такая задача называется задачей *линейного программирования* (ЛП). Нелинейные задачи МП называются также задачами *нелинейного программирования* (НП). Существуют и другие классы задач МП.

Бесконечномерные задачи оптимизации. Типичными представителями бесконечномерных задач являются хорошо изученные задачи оптимального управления. Рассмотрим характерный пример.

Пусть управляемый процесс описывается системой обыкновенных дифференциальных уравнений (ОДУ)

$$\frac{dx}{dt} = f(t,\ x,\ u),\quad x(t_0) = x_0.$$

Здесь $x(t)$ — n-мерный вектор состояний системы; $u(t)$ — r-мерный вектор управлений, принадлежащий заданному множеству U допустимых вектор-функций. Например, U — множество кусочно-непрерывных на промежутке $[t_0, t_1]$ вектор-функций. Требуется определить такое управление $u(t) \in U$, чтобы минимизировать функционал

$$J(x,\ u) = \int_{t_0}^{t_1} F(t,\ x,\ u)\, dt.$$

Функция F задана. Естественно, в оптимальном управлении используются функционалы и с иной, более сложной структурой, а также более сложные системы ограничений.

Сформулированную задачу можно интерпретировать по-разному. Например, поиск ведется в пространстве функций $u\ t$. Тогда для вычисления J по заданному аргументу $u\ t$ решается заданная система ОДУ и полученная в результате зависимость $x\ t$ подставляется в выражение для J. Во втором варианте поиск ведется в пространстве x, u , а система ОДУ выступает, как ограничение-равенство.

В книге даются необходимые элементарные сведения из функционального анализа, позволяющие без обращения к дополнительной литературе ознакомиться с важнейшими для теории оптимизации математическими структурами. Методы функционального анализа оказываются особенно важными при изучении задач бесконечномерной оптимизации. Как будет далее показано, с помощью относительно простых и достаточно общих конструкций функционального анализа можно с единых позиций охватить обширный материал, связанный с классическим вариационным исчислением, необходимыми и достаточными условиями экстремума и т. д. Основные выводы и понятия конечномерной оптимизации оказываются частными случаями полученных общих соотношений.

При рассмотрении элементов функционального анализа ставилась задача максимально разгрузить материал от деталей и в то же время дать логически завершенные математические конструкции, достаточные для формулировки необходимых для наших целей результатов теории оптимизации.

Глава 1

Математические основы. Элементы функционального анализа

1.1. Множества

В 1872 г. Георг Кантор, создатель теории множеств, определил множество как "объединение в одно целое объектов, хорошо различимых нашей интуицией или нашей мыслью". Это были времена наивной теории множеств. Такое положение долго не сохранилось. Интуитивное понимание множества оказалось в некоторых случаях логически порочным из-за антиномий или парадоксов теории множеств.

Антиномия Рассела (1903 г.). Для произвольного множества можно поставить вопрос о том, является ли оно своим собственным элементом. Например, множество студентов студентом не является и потому — не собственный элемент. В то же время множество всех множеств естественно считать содержащим себя как элемент. Рассмотрим теперь множество S всех множеств, не являющихся собственными элементами. Тогда получаем следующий обескураживающий результат: если предположить, что $S \in S$, то по определению S имеем $S \notin S$. С другой стороны, предположение $S \notin S$ приводит к следствию $S \in S$. Получили явное противоречие. Данный парадокс можно разрешить с помощью следующих более аккуратных рассуждений. Действительно, на самом деле вначале мы предположили существование множества S всех множеств, не являющихся собственными элементами. После этого получили, что $S \in S$ в том и только в том случае, если $S \notin S$. Из этого мы должны заключить, что такое множество S не существует. Указанное объяснение справедливо. Однако это мало способствует уменьшению парадоксальности рассмотренной ситуации. Действительно, тот факт, что не может существовать множество с четко очерченными свойствами (не содержащими себя в качестве элемента), столь же настораживает, как и прямое противоречие.

Чтобы исключить подобные антиинтуитивные результаты, в теории множеств был использован аксиоматический метод, позволяющий исправить ситуацию с помощью сознательного сужения (обеднения) интуитивного понятия множества.

Мы далее будем использовать дескриптивную (наглядную) теорию множеств. Нижеследующее введение в теорию множеств не является сколько-нибудь полным. Предполагается, что читатель знаком с элементами теории множеств из общих курсов по математике, и нам остается лишь согласовать обозначения, напомнить основные соотношения и сделать некоторые замечания.

Множества будем обозначать прописными буквами, а их элементы — строчными:

- $a \in A$ — элемент a принадлежит множеству A;
- $a \notin A$ — элемент a не принадлежит множеству A.

Если справедливо $\forall x \in A \Rightarrow x \in B$, то множество B включает множество A: $A \subset B$. (Здесь и далее двойная стрелка означает логическое следование.) $\forall$ — квантор общности. $\forall x$ читается: для любого x.

Равенство множеств: $A \subset B \ \wedge \ B \subset A \ \Rightarrow A = B$. ($\wedge$ — знак логического "И".)

Если X — множество, а P — некоторое свойство, то подмножество X, состоящее из тех x, для которых справедливо $P\ x$, обозначается как $x \in X / P\ x$. Пустое множество (не содержащее элементов) обозначается как $\varnothing$. Множество всех подмножеств некоторого множества A будем обозначать как $\Im\ A$.

1.1.1. Операции над множествами и их свойства

Объединение множеств:

$$A \cup B = \{ x / \ x \in A \ \vee \ x \in B \} .$$

Здесь $\vee$ — знак логического "ИЛИ".

Пересечение множеств:

$$A \cap \mathrm{B} = \{ x / \ x \in A \ \wedge \ x \in B \} .$$

Коммутативность операций объединения и пересечения множеств:

$$A \cup B = B \cup A; \ A \cap B = B \cap A.$$

Ассоциативность операций $\cup$ и $\cap$:

$$A \cup B \ \cup C = A \cup \ B \cup C \ ;$$

$$A \cap B \ \cap C = A \cap \ B \cap C \ .$$

Дистрибутивность:

$$A \cup B \ \cap C = \ A \cap C \ \cup \ B \cap C \ ;$$

$$A \cap B \ \cup C = \ A \cup C \ \cap \ B \cup C \ .$$

Разность множеств:

$$A \setminus B = \{ x \in A / x \notin B \} .$$

Декартово произведение множеств:

$$A \times B = \{ (a, b) / a \in A, \ b \in B \} .$$

Здесь множества A и B могут совпадать.

Подмножество $R \subset A \times B$ называется *бинарным отношением*, или просто *отношением*, заданным на множествах A и B.

Переходим к важнейшим определениям, связанным с понятием функции или отображения.

1.1.2. Функции и отображения

Отношение $F \subset X \times Y$ называется *отображением* X в Y или *функцией*, определенной в X и принимающей значения в Y, если для $\forall x$:

$$x,\ y_1\ \in F\ \wedge\ \ x,\ y_2\ \in F\ \Rightarrow y_1 = y_2.$$

Если $x,\ y\ \in F$, то элемент y называется значением F в x и обозначается как $y = F\ x$.

Используются также обозначения:

$$F: X \to Y,$$

$$F: x \to F\ x\ .$$

Принципиально важным свойством отображения является то, что любому значению "аргумента" x ставится в соответствие единственный элемент y. Такое понятие, как "многозначная" функция, здесь не рассматривается. Вполне правомерно, конечно, определить отображение, значениями которого являются подмножества некоторого данного множества, состоящие более чем из одного элемента. Но такое определение практически бесполезно, т. к. не удается разумным образом определить алгебраические операции над значениями таких функций. Например, операция извлечения корня из вещественного числа $\sqrt{z}$ приводит к двум значениям со знаками "плюс" и "минус". Но тогда как понимать равенство: $\sqrt{z} + \sqrt{z} = 2\sqrt{z}$? Левая часть имеет три разных значения, а правая — только два. (Хотя, конечно, существует и понятие Римановой поверхности.)

Сделаем еще одно замечание. Обычно (в "школьной" математике) различают понятия функции и ее графика. В данном выше определении эти понятия совпадают.

В современной математике важнейшую роль играет рассмотрение отображения (функции) как единого объекта (такого же, как точка или число) и проведение ясного различия между отображением F и любым из его значений $F\ x$. Первое есть элемент множества отображений X в Y, обозначаемого как $\Im\ X \times Y$, второе — элемент множества Y, причем

$$F = \ \ x,\ y\ \in X \times Y\ /\ y = F\ x\ \ .$$

Таким образом, отображение F есть некоторое множество упорядоченных пар $x,\ y$.

Пусть $F: X \to Y$. Пусть также $A \subset X$. Тогда множество

$$F\ A\ =\ y \in Y /\ \exists x \in A\quad y = F\ x$$

называется *образом множества A* при отображении *F*. Здесь $\exists$ — квантор существования. $\exists x$ читается: существует *x*. Прообразом множества $B \subset Y$ при отображении *F* называется множество

$$F^{-1}\ B\ =\ x \in X / F\ x\ \in B\ .$$

Пусть $F: X \to Y$. Пусть $A \subset X$. Тогда множество

$$F \cap (A \times Y) \subset A \times Y$$

называется *сужением отображения F* на множестве *A*.

1.1.3. Виды отображений

Пусть $F: X \to Y$. Тогда:

1. Отображение F называется *сюръективным* (накрытием или *"отображением на"*), если $F\ X\ = Y$, т. е. для $\forall y \in Y\ \ \exists x \in X$ такой, что $y = F\ x$.
2. Отображение *F* называется *инъективным* (вложением), если $F\ x\ = F\ x' \Rightarrow x = x'$.
3. Отображение *F* называется *биективным* (наложением, биекцией), если оно одновременно является сюръективным и инъективным. Будем называть биекцию взаимно-однозначным соответствием.

Отображение произвольного множества во множество действительных (вещественных) чисел называется вещественным *функционалом*. Иногда рассматриваемое множество чисел является множеством комплексных чисел, и мы приходим к понятию комплекснозначного функционала. Далее будут рассматриваться только вещественные функционалы. В этом смысле обычные "школьные" вещественные функции от вещественного аргумента также являются функционалами.

Пусть даны три множества *X*, *Y*, *Z* и два отображения

$$F: X \to Y,$$
$$G: Y \to Z.$$

Тогда $x \to G\ F\ x$ есть отображение $H: X \to Z$, которое называется *композицией* отображений *G* и *F* (в указанном порядке) или сложным отображением и обозначается символом $\circ$: $H = G \circ F$.

1.1.4. Семейства элементов

Пусть *L* и *X* — два множества. Отображение $L \to X$ иногда называется также *семейством элементов* множества *X*, имеющим *L* множеством индексов и обозначается $\lambda \to x_\lambda$, или $x_\lambda\ _{\lambda \in L}$, или, когда это не может привести к недоразумению, просто x_λ .

Чаще всего в качестве множества индексов L фигурирует множество целых положительных чисел N или его подмножество (конечное или бесконечное). В этом случае семейство элементов называется *последовательностью* (конечной или бесконечной). Иногда последовательности элементов будут изображаться с применением фигурных скобок. Следует отличать семейство $x_{\lambda\ \lambda\in L}$ элементов множества X от подмножества множества X, состоящего из элементов этого семейства.

1.1.5. Счетные множества

Сравнение конечных множеств по числу элементов не вызывает затруднений. Можно сосчитать количество элементов в обоих множествах и сравнить результаты. Можно поступить и по-другому. Именно попытаться установить между сравниваемыми множествами взаимно-однозначное соответствие (биекцию). Ясно, что биекция между двумя конечными множествами может быть установлена тогда и только тогда, когда число элементов в них одинаково. Второй способ сравнения пригоден и для бесконечных множеств.

X называется *равномощным* множеству Y, если существует биективное отображение $X \to Y$. Множество называется *счетным*, если оно равномощно множеству N натуральных чисел. Бесконечное множество, не являющееся счетным, называется *несчетным* множеством. Можно доказать, что если бесконечное множество X счетно, то $\Im\ X$ — несчетно ("имеет мощность континуума").

Счетность множества четных чисел $n \to 2n$ представляет пример того, что бесконечные множества могут быть равномощны своему подмножеству. Всякое бесконечное множество содержит счетное подмножество. Таким образом, среди бесконечных множеств счетные множества являются "самыми маленькими".

Теорема 1.1. Множество действительных чисел из промежутка $0, 1$ несчетно.

Доказательство (диагональный процесс Кантора). Допустим противное и предположим, что удалось "перенумеровать" все действительные числа следующим образом:

$$\alpha_1 = 0.a_{11}a_{12}a_{13}\ldots$$
$$\alpha_2 = 0.a_{21}a_{22}a_{23}\ldots$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\alpha_n = 0.a_{n1}a_{n2}a_{n3}\ldots$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$

Тогда всегда можно построить действительное число β

$$\beta = 0.b_1b_2\ldots b_n\ldots$$
$$a_{nn} = 1 \to b_n = 2$$
$$a_{nn} \neq 1 \to b_n = 1, \quad n = 1,\ 2,\ 3,\ \ldots,$$

не входящее в перенумерованный список действительных чисел. Таким образом, доказано, что множество действительных чисел не может быть счетным. Мощность множества действительных чисел называется мощностью континуума.

Упражнение 1.1
Постройте биекцию между множеством действительных чисел из отрезка [0, 1] и множеством точек квадрата со сторонами [0, 1].

1.2. Метрические пространства

В теории метрических пространств развивается геометрический язык, на котором выражаются результаты анализа. Этот язык позволяет придать этим результатам достаточную общность и вместе с тем дать наиболее простые и отражающие суть дела доказательства. Нас будут интересовать топологические аспекты теории метрических пространств, связанные с концепцией предельного перехода, а также алгебраические аспекты при изучении основных операций над элементами метрических пространств. Метрические пространства являются частным видом более общих топологических пространств.

Пусть E — некоторое множество. *Расстояние* в E есть отображение (функционал) d произведения $E \times E$ во множество действительных чисел R:

$$d : E \times E \to R .$$

Предполагается, что функционал d обладает следующими свойствами:

1. $\forall\ x, y :\ \ d(x, y) \geq 0.$
2. $d\ x, y\ = 0 \Leftrightarrow x = y.$
3. $\forall x, y :\ \ d\ x, y\ = d\ y, x\ .$
4. $\forall x, y, z :\ \ d\ x, z\ \leq d\ x, y\ + d\ y, z\ ,\ \ \ x, y, z \in E.$

Свойство 4 называется неравенством треугольника.

Множество E с заданным в нем расстоянием d называется *метрическим пространством*. Обычно это пространство, т. е. пару E, d , обозначают одной буквой E.

Примеры метрических пространств.

1. Множество $C\ a, b$ всех непрерывных действительных функций, определенных на отрезке a, b с расстоянием

$$d\ f, g\ = \max_{t \in a, b} \left| f\ t\ - g\ t\ \right|.$$

2. В этом же множестве функций можно ввести расстояние

$$d\ f, g\ = \left(\int_a^b f\ t\ - g\ t\ ^2 dt \right)^{1/2} .$$

В результате получим пространство *непрерывных функций с квадратичной метрикой* $C^2\ a, b$.

Обычно та или иная метрика для построения метрического пространства выбирается в соответствии с особенностями решаемой задачи. Первая метрика в приведенном примере отражает достаточно жесткое требование к близости функций. Ее применяют, например, при решении задачи равномерного приближения функций (равномерная аппроксимация), когда требуется гарантировать, чтобы на всем отрезке a, b отклонение функций f и g было меньше некоторой заданной величины. Вторая метрика отражает менее жесткое требование. Оно заключается в том, что для "подавляющего" большинства (но не для всех) значений аргумента t из a, b абсолютная величина $|f - g|$ также мала. Во многих случаях, например, при обработке результатов наблюдений квадратичное приближение является наиболее приемлемым, т. к. оно позволяет сглаживать отдельные локальные выбросы аппроксимируемой (экспериментальной) функции с помощью некоторой теоретической расчетной зависимости. В итоге можно получить достаточно точное общее представление о характере протекающего процесса даже при наличии ошибок при измерении экспериментальных зависимостей.

1.2.1. Изометрия

Пусть $E,\ E'$ — два метрических пространства с расстояниями, соответственно, $d,\ d'$. Тогда биективное отображение f: $E,\ E'$ называется *изометрией*, если для $\forall\ x, y\ \in E \times E$ имеем

$$d'\ f\ x\ ,\ f\ y\ = d\ x, y\ .$$

Если f: $E,\ E'$ — биективное отображение, то в соответствии с определением биективности отношение $y = f\ x$ функционально не только по y, но и по x, т. е. это отношение определяет не только функцию $E,\ E'$, но и $E' \to E$. Последнее отображение называется *обратным* к f и обозначается $x = f^{-1}\ y$. Оно само биективно. Если f — изометрия $E \to E'$, то тогда f^{-1} будет изометрией $E' \to E$. Важность понятия изометрии заключается в том, что любая теорема, доказанная в E и в формулировке которой участвуют только расстояния между элементами E, немедленно дает соответствующую теорему в любом изометрическом пространстве E'.

Если $E,\ E'$ — два множества, причем E — метрическое пространство с расстоянием d, то при наличии биекции f: $E \to E'$ множество E' также становится метрическим пространством с индуцированным расстоянием d'. Говорят, что расстояние d' перенесено в E' при помощи отображения f.

В теории метрических пространств удобен геометрический язык. В частности, элементы метрического пространства часто называются точками.

1.2.2. Шары, сферы, диаметр, окрестности

Пусть заданы: метрическое пространство E, d, $a \in E$ и действительное число $r > 0$. Тогда множества

$$B\ a; r\ = \ x \in E / d\ a, x\ < r\ ,$$

$$B'\ a; r\ = \ x \in E / d\ a, x\ \leq r\ ,$$

$$S\ a; r\ = \ x \in E / d\ a, x\ = r$$

называются, соответственно, *открытым шаром* (с центром *a* и радиусом *r*), *замкнутым шаром* и *сферой*.

Пусть E — метрическое пространство и $A \subset E$, $B \subset E$, $A \neq \varnothing$, $B \neq \varnothing$. Тогда *расстоянием* от *A* до *B* называется число

$$d\ A, B\ = \inf_{\substack{x \in A \\ y \in B}} d\ x, y\ .$$

Диаметром множества *A* называется число

$$\delta\ A\ = \sup_{\substack{x \in A \\ y \in A}} d\ x, y\ .$$

Диаметр может быть и бесконечным.

Ограниченным множеством в *E* называется непустое множество, диаметр которого конечен.

Открытым множеством в метрическом пространстве E, d называется множество $A \subset E$ такое, что

$$\forall x \in A\ \ \exists r > 0 \Rightarrow B\ x; r\ \subset A\ .$$

Таким образом, любая точка открытого множества *A* входит в это множество вместе с некоторым, содержащим ее открытым шаром.

Пусть $A \subset E$, $A \neq \varnothing$. Тогда *открытой окрестностью* множества *A* называется любое открытое множество, содержащее *A*, а *окрестностью* множества *A* — любое множество, содержащее открытую окрестность *A*.

Если множество *A* состоит из одной точки $A = \ x\ $, то мы говорим об *окрестностях точки x*, а не множества $A = \ x\ $.

Фундаментальной системой окрестностей множества *A* называется семейство U_λ окрестностей *A*, обладающее тем свойством, что любая окрестность *A* содержит хотя бы одно из множеств семейства U_λ .

Точка $x \in A$ называется *внутренней точкой* множества *A*, если *A* является окрестностью точки *x*. Множество всех внутренних точек множества *A* называется *внутренностью* множества *A* и обозначается $\overset{\circ}{A}$.

Внутренняя точка множества $E \setminus A$ называется *внешней точкой* множества A.

Точкой прикосновения множества A называется такая точка $x \in E$, любая окрестность которой имеет с A непустое пересечение, т. е. содержит хотя бы одну точку из A. Множество всех точек прикосновения множества A называется *замыканием* A и обозначается символом $\overline{A}$. По определению, *замкнутое множество* в E есть дополнение открытого множества, т. е. B замкнуто, если $E \setminus B$ открыто. $A \subset E$ замкнуто, если $A = \overline{A}$. Например, замкнутый шар есть замкнутое множество. $\overline{A}$ — наименьшее замкнутое множество, содержащее A.

Точка $x \in E$ называется *предельной точкой* множества $A \subset E$, если любая ее окрестность содержит бесконечно много точек из A. Сама предельная точка не обязана принадлежать множеству A. Можно дать определение замкнутого множества на языке предельных точек. Именно множество A замкнуто, если оно содержит все свои предельные точки.

Точка $x \in E$ называется *граничной точкой* множества A, если она является точкой прикосновения как A, так и $E \setminus A$. Граничная точка x множества A характеризуется тем свойством, что в любой ее окрестности содержится по крайней мере одна точка множества A и по крайней мере одна точка множества $E \setminus A$. Множество $Fr\ A$ всех граничных точек множества A называется *границей* множества A.

1.2.3. Сепарабельные пространства, подпространства, непрерывные отображения

Множество A в метрическом пространстве E называется *плотным* относительно множества B (или в B), если любая точка $x \in B$ есть точка прикосновения множества A, т. е. если $B \subset \overline{A}$. В этом случае любая окрестность любой точки $x \in B$ содержит точку множества A.

Множество A, плотное в E, называется *всюду плотным*. В этом случае $\overline{A} = E$. Метрическое пространство E называется *сепарабельным*, если в E существует не более чем счетное (т. е. конечное или счетное) всюду плотное множество. Например, множество рациональных чисел всюду плотно в множестве действительных чисел R, поэтому пространство действительных чисел сепарабельно. Точно так же сепарабельны пространства $C\ a, b\ ,\ C^2\ a, b$ из приведенных ранее примеров.

Пусть $A \neq \varnothing,\ A \subset E$. Тогда сужение на $A \times A$ отображения d: $E \times E \to R$ является расстоянием в A и называется расстоянием, индуцированным в A расстоянием d пространства E. Метрическое пространство, определяемое этим индуцированным расстоянием, называется *подпространством* A метрического пространства E.

Пусть E и E' — два метрических пространства, d и d' — расстояния в них. Отображение f: $E \to E'$ называется *непрерывным в точке* $x_0 \in E$, если для $\forall \varepsilon > 0$ $\exists \delta > 0$ такое, что

$$d\ x, x_0\ < \delta \Rightarrow d'\ f\ x_0\ , f\ x\ < \varepsilon .$$

Можно сказать нагляднее. Отображение f: $E \to E'$ называется непрерывным в точке $x_0 \in E$, если точка $f\ x$ сколь угодно близка к точке $f\ x_0$, как только точка x достаточно близка к x_0.

Отображение f называется *непрерывным в пространстве* E (или просто непрерывным), если оно непрерывно в каждой точке пространства E.

Интересен следующий факт. Сужение отображения f: $E \to E'$ на подпространстве F метрического пространства E может оказаться непрерывным и в случае, когда f не является непрерывным ни в одной точке $x \in E$. Примером служит отображение f: $R \to R$, равное нулю на множестве Q рациональных чисел и единице на его дополнении (функция Дирихле). Сужение отображения f на Q постоянно и потому непрерывно.

Отображение f: $E \to E'$ называется *равномерно непрерывным*, если для $\forall \varepsilon > 0$ $\exists \delta > 0$ такое, что

$$\forall x,\ y:\ d\ x,\ y\ < \delta \Rightarrow d'\ f\ x\ ,\ f\ y\ \ < \varepsilon .$$

(Здесь важно, что ε и δ не зависят от x и y.) Ясно, что из равномерной непрерывности следует непрерывность. Обратное, вообще говоря, неверно. Например, отображение $x \to x^2$ непрерывно на R, но не равномерно непрерывно.

Упражнение 1.2
Докажите последнее утверждение.

1.2.4. Гомеоморфизмы, пределы, полные пространства

Отображение f: $E \to E'$ называется *гомеоморфизмом*, если f — биекция, а f и f^{-1} непрерывны. Два метрических пространства E и E' называются гомеоморфными, если существует гомеоморфизм $E \to E'$.

Упражнение 1.3
Докажите, что изометрия всегда будет гомеоморфизмом, но не наоборот.

Понятие гомеоморфизма, в частности, оказывается полезным в том смысле, что можно оперировать не с исходным метрическим пространством, а с гомеоморфным ему, а затем перенести результаты на основное пространство.

Пусть x_n — последовательность точек в метрическом пространстве E. Тогда $a \in E$ называется *пределом последовательности* x_n, если для $\forall \varepsilon > 0$ $\exists n_0 \in N$ такое, что $n \geq n_0 \Rightarrow d\ a,\ x_n\ < \varepsilon$.

Используется запись

$$a = \lim_{n\to\infty} x_n \text{, или } \lim_{n\to\infty} d\ a,\ x_n\ = 0 .$$

x_n называется сходящейся последовательностью.

1.2.5. Последовательности Коши, полные пространства

Последовательностью Коши в метрическом пространстве E называется последовательность x_n, удовлетворяющая условиям: для $\forall\varepsilon > 0$ $\exists n_0$ такое, что

$$p \ge n_0,\ q \ge n_0 \ \Rightarrow d\ x_p,\ x_q\ < \varepsilon .$$

Иногда последовательности Коши называются фундаментальными последовательностями.

Теорема 1.2. Любая сходящаяся последовательность есть последовательность Коши.

Доказательство. Пусть $a = \lim_{n\to\infty} x_n$, тогда для $\forall\varepsilon > 0$ $\exists n_0 \in N$ такое, что

$$n \ge n_0 \Rightarrow d\ a,\ x_n\ < \varepsilon/2 .$$

В силу неравенства треугольника имеем: если $p \ge n_0$, $q \ge n_0$, то

$$d\ x_p,\ x_q\ \le d\ a,\ x_p\ + d\ a,\ x_q\ < \varepsilon .$$

Что и требовалось доказать.

Метрическое пространство E называется *полным*, если любая последовательность Коши сходится в E (разумеется, к точке пространства E).

Упражнение 1.4

Докажите, что промежуток a,b, пространство $C^2\ a,\ b$ — неполные пространства, а пространство $C\ a,\ b$ — полное.

Фундаментальная важность полных пространств основывается на том, что для доказательства сходимости некоторой последовательности в таком пространстве достаточно установить, что она является последовательностью Коши (т. е. удовлетворяет критерию Коши). Главное различие между проверкой этого критерия и непосредственной проверкой определения сходящейся последовательности состоит в том, что в критерии Коши не нужно заранее знать значение предела. Сходимость оказывается внутренним свойством последовательности.

Теорема 1.3. В полном метрическом пространстве E любое замкнутое множество $F \subset E$ является полным подпространством.

Доказательство. Так как E полно, то последовательность Коши x_n точек множества F сходится к $a \in E$, и так как $x_n \in F$, то a является точкой прикосновения множества F, т. е. $a \in \bar{F}$. Из замкнутости F следует, что $F = \bar{F}$, т. е. $a \in F$. Что и требовалось доказать.

В приведенном рассуждении заключается важность понятия замкнутости.

1.2.6. Принцип сжимающих отображений

Данный принцип используется при доказательстве теорем существования и единственности в различных разделах функционального анализа.

Дадим необходимые предварительные определения и утверждения.

1. Пусть E, d — метрическое пространство и f: $E \to E$ — отображение E в себя.

   Тогда отображение f называется *сжимающим*, если существует такое число α: $0 < \alpha < 1$, что для $\forall x, y \in E$ выполняется неравенство

   $$d\ f\ x\ ,\ f\ y\ \leq \alpha d\ x, y\ .$$

   В этом случае отображение f называется также *сжатием*.

2. Элемент $x \in E$ называется *неподвижной точкой* отображения f, если $f\ x\ = x$, т. е. неподвижные точки — это решения уравнения $f\ x\ = x$.

3. Любое сжимающее отображение непрерывно. Действительно, для $\forall x_0 \in E$ имеем

   $$d\ x, x_0\ < \delta \Rightarrow d\ f\ x\ ,\ f\ x_0\ < \alpha\delta = \varepsilon,$$

   т. е. достаточно для $\forall \varepsilon > 0$ положить $\delta = \varepsilon/\alpha$.

Теорема 1.4 (принцип сжимающих отображений). Всякое сжимающее отображение, определенное в полном метрическом пространстве E, имеет одну и только одну неподвижную точку.

Доказательство.

1. Существование.

   Пусть x_0 — произвольный элемент E. Положим $x_{n+1} = f\ x_n\ ,\ n = 0, 1, \ldots$ Докажем, что x_n — последовательность Коши. Так как f — сжатие, имеем

   $$d\ x_2, x_1\ = d\ f\ x_1\ ,\ f\ x_0\ \leq \alpha d\ x_1, x_0$$
   $$d\ x_3, x_2\ \leq \alpha d\ x_2, x_1\ \leq \alpha^2 d\ x_1, x_0$$
   $$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
   $$d\ x_{n+1}, x_n\ \leq \alpha^n d\ x_1, x_0\ .$$

Применяя многократно неравенство треугольника, получим

$$d\ x_{n+p},\ x_n\ \le d\ x_{n+p},\ x_{n+p-1}\ + d\ x_{n+p-1},\ x_{n+p-2}\ +\ldots+$$
$$+\ d\ x_{n+1},\ x_n\ \le \alpha^{n+p-1} d\ x_1,\ x_0\ + \alpha^{n+p-2} d\ x_1,\ x_0\ +\ldots+$$
$$+\ \alpha^n d\ x_1,\ x_0\ = \alpha^n d\ x_1,\ x_0\ \ \alpha^{p-1} + \alpha^{p-2} +\ldots+ 1\ \le$$
$$\le \alpha^n d\ x_1,\ x_0\ \ \frac{1}{1-\alpha} = \frac{\alpha^n}{1-\alpha} d\ x_1,\ x_0\ \ \to \underset{n\to\infty}{0}.$$

Таким образом, x_n — последовательность Коши. Так как E — полное пространство, то последовательность $x_n \underset{n\to\infty}{\to} a \in E$. Из непрерывности f следует, что последовательность $f\ x_n$ сходится к $f\ a$. Переходя к пределу в равенстве $x_{n+1} = f\ x_n$, получим $f\ a\ = a$. Существование доказано.

2. Единственность неподвижной точки.

   Пусть a и b — две неподвижные точки. Тогда

$$d\ f\ a\ ,\ f\ b\ \ = d\ a,\ b\ \le \alpha d\ a,\ b\ \Rightarrow d\ a,\ b\ = 0.$$

   Действительно, если $d\ a,\ b\ \ne 0$, имеем $d\ a,\ b\ \le \alpha d\ a,\ b\ \Rightarrow 1 \le \alpha$, что невозможно по условию. Теорема доказана полностью.

Пример. Теорема существования и единственности для дифференциальных уравнений.

Пусть дано дифференциальное уравнение с начальными условиями (задача Коши):

$$\frac{dx}{dt} = f\ x\ ,\ \ x\ t_0\ = x_0. \tag{1.1}$$

Везде здесь имеются в виду обычные вещественные функции от вещественной переменной.

Предполагается, что функция f непрерывна в области G, содержащей точку x_0, и удовлетворяет условию Липшица:

$$\left|f\ x_1\ - f\ x_2\ \right| \le M \left|x_1 - x_2\right|.$$

Докажем, что тогда на некотором сегменте $\left|t - t_0\right| \le \delta$ существует и только одно решение $x = \varphi\ t$ сформулированной задачи Коши, которое может быть найдено последовательными приближениями (теорема Пикара).

Из (1.1) следует

$$x\ t\ = x_0 + \int_{t_0}^{t} f\ x\ t\ \ dt. \tag{1.2}$$

Выберем $\delta > 0$ из условия $M\delta < 1$. Обозначим через C пространство непрерывных на сегменте $|t - t_0| \le \delta$ функций с метрикой (расстоянием)

$$d\ (\varphi_1, \varphi_2\ ) = \max_t |\varphi_1\ t\ - \varphi_2\ t\ |.$$

Такое пространство уже рассматривалось ранее. Можно доказать, что это пространство полно. Рассмотрим отображение F: $C \to C$, $\psi = F\ \varphi$, определяемое формулой

$$\psi\ t\ = x_0 + \int_{t_0}^{t} f\ \varphi\ t\ \ dt,\ |t - t_0| \le \delta.$$

Докажем, что F — сжатие.

Имеем:

$$\forall t:\ |\psi_1\ t\ - \psi_2\ t\ | = \left| \int_{t_0}^{t} f\ \varphi_1\ dt - \int_{t_0}^{t} f\ \varphi_2\ dt \right| \le \int_{t_0}^{t} |f\ \varphi_1\ - f\ \varphi_2\ | dt \le$$

$$\le M \int_{t_0}^{t} |\varphi_1 - \varphi_2| dt \le M\delta \max_t |\varphi_1 - \varphi_2| = M\delta d\ \varphi_1, \varphi_2\ .$$

Отсюда следует, что и

$$d\ \psi_1, \psi_2\ = \max_t |\psi_1 - \psi_2| \le M\delta d\ \varphi_1, \varphi_2\ = \alpha d\ \varphi_1, \varphi_2\ , \alpha = M\delta < 1.$$

Таким образом, уравнение (1.2) вида $x = F\ x$ имеет одно и только одно решение в пространстве C.

Теорема доказана.

1.2.7. Компактные пространства

Метрическое пространство E называется *вполне ограниченным*, если для $\forall \varepsilon > 0$ существует конечное множество $F \subset E$ такое, что $\forall x \in E: d\ x, F\ < \varepsilon$.

В теории метрических пространств это понятие выражает свойство метрического пространства быть *"приблизительно конечным"*. Ясно, что вполне ограниченное метрическое пространство является ограниченным множеством. Обратное, вообще говоря, неверно.

Метрическое пространство E называется *компактным*, если оно является полным и вполне ограниченным.

Пусть в метрическом пространстве E задана последовательность x_n. Тогда точка $b \in E$ называется *предельной точкой* последовательности x_n, если для любой окрестности V точки b и любого m $\exists n \ge m$ такое, что $x_n \in V$.

Ясно, что в любой окрестности предельной точки последовательности содержится бесконечно много точек последовательности (это утверждение может рассматриваться как альтернативное определение предельной точки).

Если последовательность x_n имеет предел, то он является единственной предельной точкой этой последовательности. Обратное, вообще говоря, неверно. Единственная предельная точка не обязана быть пределом последовательности.

Пример. Последовательность x_n действительных чисел, где $x_{2n} = 1/n$, $x_{2n+1} = n$, имеет нуль своей единственной предельной точкой, но не сходится к нулю.

Можно доказать, что если у последовательности x_n в компактном метрическом пространстве есть только одна предельная точка, то она является пределом этой последовательности. (Здесь видна роль компактности. Множество действительных чисел не является компактным.) Если E — компактное метрическое пространство, то любая бесконечная последовательность в E имеет по крайней мере одну предельную точку.

Компактным множеством в метрическом пространстве E называется такое множество A, для которого подпространство A пространства E компактно. Можно показать, что любое компактное множество в метрическом пространстве замкнуто.

Кроме того, в компактном пространстве E всякое замкнутое множество компактно. Действительно, такое множество, очевидно, вполне ограничено и как замкнутое подмножество полного пространства полно, т. е. компактно.

Относительно компактным множеством в метрическом пространстве E называется множество $A \subset E$, замыкание $\overline{A}$ которого компактно. Для того чтобы множество A в метрическом пространстве E было относительно компактно, необходимо и достаточно, чтобы любая последовательность точек множества A имела предельную точку в E.

Упражнение 1.5

Покажите, что на множестве действительных чисел относительная компактность эквивалентна ограниченности.

Изучая метрические пространства, мы изучаем множества, между элементами которых введено расстояние. По существу, мы изучаем свойства понятия расстояния между объектами произвольной природы.

1.3. Линейные пространства

Часто приходится встречаться с объектами, над которыми производятся операции сложения и умножения на числа. Например, векторы в геометрии в трехмерном пространстве умножаются на числа и складываются. Вещественные функции вещественных аргументов умножаются на числа и складываются и т. п. Одни и те же операции производятся над совершенно разными объектами. Для того чтобы изучить все такие примеры с единой точки зрения, вводится понятие линейного (или векторного) пространства.

Пусть на множестве L элементов $x, y, z, \ldots$ заданы два отображения:

$$L \times L \to L,$$
$$L \times R \to L,$$

где R — множество действительных чисел ("вещественная прямая"). Обозначим эти отображения как:

$$x, y \to x + y \in L,$$
$$x, \alpha \to \alpha x \in L$$

соответственно.

Тогда множество L называется *действительным линейным пространством*, если для введенных отображений выполнены следующие требования:

1. $x + y = y + x$ (коммутативность).
2. $x + y + z = x + y + z$ (ассоциативность).
3. $\exists \theta \in L: \forall x \in L \Rightarrow x + \theta = x$ (существование нуля).
4. Для $\forall x \in L \ \exists -x \in L \Rightarrow x + -x = \theta$ (существование противоположного элемента).
5. $1 \cdot x = x \quad 1 \in R$.
6. $\alpha \beta x = \alpha\beta x$.
7. $\alpha + \beta x = \alpha x + \beta x$.
8. $\alpha x + y = \alpha x + \alpha y$.

Часто говорят о линейном *векторном* пространстве. Сами элементы L также называются векторами.

Если вместо множества R действительных чисел используется множество C комплексных чисел, то получим *комплексное линейное пространство*.

Непустое подмножество линейного пространства L называется *подпространством*, если оно само является линейным пространством по отношению к определенным в L операциям.

Пусть x_α — произвольное непустое *семейство элементов* линейного пространства L (счетность множества x_α не предполагается). Определение понятия семейства элементов было дано выше. Рассмотрим все подпространства линейного пространства L, содержащие заданную систему векторов x_α . Пересечение этих подпространств, очевидно, тоже будет подпространством. Это так называемое *наименьшее* подпространство, содержащее x_α . Оно называется подпространством, *порожденным* множеством x_α , или *линейной оболочкой* семейства элементов x_α .

Примеры линейных пространств.

1. Множество R действительных чисел с обычными операциями сложения и умножения (L совпадает с R).
2. Совокупность систем n действительных (или комплексных) чисел $x = x_1, x_2, ..., x_n$, где сложение и умножение на число определяются формулами

$$x + y = x_1 + y_1, x_2 + y_2, ..., x_n + y_n \ , \ \alpha x = \alpha x_1, \alpha x_2, ..., \alpha x_n \ .$$

   называется n-мерным *арифметическим пространством* и обозначается R^n (для действительного пространства) или C^n (в комплексном случае).
3. Множество непрерывных на отрезке a, b функций с обычными операциями сложения и умножения на числа образует векторные пространства $C\ a, b$, $C^2\ a, b$.

Конечное множество векторов x_i называется *линейно зависимым*, если существует множество чисел α_i, из которых не все равны нулю, такое, что $\sum \alpha_i x_i = 0$. Если

$$\sum \alpha_i x_i = 0 \Rightarrow \forall i : \alpha_i = 0 ,$$

то конечное множество x_i называется *линейно независимым*.

Замечание. Важно, что линейная зависимость и линейная независимость — свойства множества векторов. Однако соответствующие прилагательные часто условно применяются к самим векторам, которые называются линейно зависимыми или линейно независимыми.

Множество X (не обязательно конечное) называется линейно независимым, если линейно независимо любое его конечное подмножество. В противном случае множество X — линейно зависимо.

Если x_i — *конечное* множество и для некоторого $x \in L$ справедливо представление $x = \sum \alpha_i x_i$, то говорят, что x является *линейной комбинацией* векторов x_i . Линейное пространство L называется *конечномерным*, если в нем существует n линейно независимых векторов, а любые $n+1$ векторы — линейно зависимы. В этом случае говорят, что L имеет *размерность* n. Любой набор из n линейно независимых векторов n-мерного пространства L называется *базисом* этого пространства. Если существует любое количество линейно независимых векторов, то пространство L называется *бесконечномерным*. Понятие базиса бесконечномерного пространства здесь не обсуждается.

Примеры.

1. Можно доказать, что пространства R^n, C^n имеют размерность n, поэтому и были названы ранее n-мерными пространствами.
2. Пространства $C\ a, b$, $C^2\ a, b$ бесконечномерны.

3. Базисом в пространстве R^1 является любое действительное число, отличное от нуля.

4. Базис в пространстве R^n образует, например, система векторов $1, 0, ..., 0$, $0, 1, ..., 0$, ..., $0, 0, ..., 1$.

n-мерные пространства изучаются в курсах по линейной алгебре и являются основой для задач нелинейного программирования. Пространства с бесконечным числом измерений изучаются в функциональном анализе и представляют основной интерес для бесконечномерных оптимизационных задач, например, для задач теории оптимального управления.

1.3.1. Линейные функционалы

Еще раз напомним, что отображение f: $X \to R$, где X — произвольное множество, называется (вещественным) *функционалом*. Таким образом, вещественный функционал — это функция, принимающая свои значения во множестве действительных (вещественных) чисел. Далее мы будем рассматривать только вещественные функционалы, специально не оговаривая это. В литературе, в том числе и математической, встречаются другие определения функционала, связанные с дополнительными предположениями о составе множества X. Мы далее везде будем иметь в виду данное выше каноническое определение.

Функционал f, определенный на линейном пространстве L, называется *линейным*, если

$$\forall x, y \in L: \quad f(x+y) = f(x) + f(y),$$
$$\forall \alpha: \quad f(\alpha x) = \alpha f(x).$$

Примером линейного функционала в пространстве $C[a, b]$ является интеграл

$$I(x) = \int_a^b x(t)\,dt.$$

1.3.2. Выпуклые множества

Пусть L — линейное действительное пространство и $x_1 \in L$, $x_2 \in L$. *Замкнутым отрезком* (или *сегментом*) в L, соединяющим точки x_1 и x_2, называется совокупность элементов вида

$$W = \vartheta x_1 + (1-\vartheta)x_2 = x_2 + \vartheta(x_1 - x_2), \quad 0 \le \vartheta \le 1.$$

В случае $0 < \vartheta < 1$ имеем *открытый отрезок*.

Множество $A \subset L$ называется *выпуклым*, если выполняется следующее соотношение

$$x_1, x_2 \in A \Rightarrow \vartheta x_1 + (1-\vartheta)x_2 \in A, \quad \forall \vartheta: 0 \le \vartheta \le 1.$$

Таким образом, выпуклое множество вместе с любыми своими двумя точками должно содержать и отрезок, соединяющий эти точки. Очевидно, само пространство L — выпукло.

Ядром $J\ E$ произвольного подмножества $E \subset L$ называется множество, удовлетворяющее следующим условиям: $x \in J\ E$ тогда и только тогда, когда для $\forall y \in L$ существует число $\varepsilon_y > 0$ такое, что для всех t,

$$|t| < \varepsilon_y \Rightarrow x + ty \in E.$$

Здесь y задает направление, а число t — продвижение по этому направлению. Очевидно, элементы ядра принадлежат E, но не все элементы E входят в ядро.

Множество $E \subset L$ называется телом, если $J\ E \neq \varnothing$. Если выпуклое множество E — тело, то оно называется *выпуклым телом*.

Примеры.

1. В пространстве R^3 куб и шар являются примерами выпуклых тел. Отрезок, плоскость, треугольник в том же пространстве — выпуклые множества, но не выпуклые тела.
2. Рассмотрим множество C непрерывных на отрезке a, b функций и его подмножество, удовлетворяющее условию $G = \{f \in C / |f\ t\ | \le 1\}$. (Сами функции из G вовсе не обязаны быть выпуклыми в обычном смысле.) Можно доказать, что G — выпуклое тело.

Докажем утверждение из примера 2. Множество G — выпукло. Действительно, если $g, f \in G$, то $|f| \le 1, |g| \le 1$. При $0 \le \vartheta \le 1$ имеем $|\vartheta f + (1-\vartheta)\ g| \le |\vartheta + (1-\vartheta)| = 1$. Таким образом, отрезок, соединяющий точки g, f, также входит в множество G и, следовательно, G — выпукло. Проверим, является ли это множество выпуклым телом. Для этого необходимо и достаточно доказать, что $J\ G \neq \varnothing$. Рассмотрим во множестве G функцию $\bar{f}$, тождественно равную нулю. Покажем, что $\bar{f} \in J\ G$ и, следовательно, $J\ G \neq \varnothing$. Действительно, рассмотрим $\forall g \in C$ и выражение $\bar{f} + tg \equiv tg$. Ясно, что $|tg| \le 1$ при $|t| \le \varepsilon = 1$. Таким образом, G — выпуклое тело.

Теорема 1.5. Пересечение любого числа выпуклых множеств есть выпуклое множество.

Доказательство ясно.

Пересечение выпуклых тел, будучи выпуклым множеством, также будет выпуклым множеством, но выпуклым телом может и не быть. Например, если рассечь куб в трехмерном пространстве плоскостью, то обе части куба будут выпуклыми телами, а само сечение — нет.

1.3.3. Выпуклые функционалы

Концепция выпуклости (выпуклые множества и выпуклые функционалы) занимает важное место в общей теории оптимизации. Причина заключается в следующем. Для многих алгоритмов оптимизации, рассматриваемых в литературе, их глобальная сходимость может быть доказана только в случае определенных предположений о выпуклости основных объектов в формулировке задачи. К сожалению, на практике условия выпуклости слишком часто нарушаются, и соответствующие теоремы сохраняют лишь теоретический интерес. Обычно "теоретики" указывают, что в достаточно малой окрестности оптимума выпуклость восстанавливается. Но для "практиков" это уже не имеет значения, т. к. попадание в достаточно малую окрестность оптимума и означает для них решение задачи.

Для функции $R \to R$ понятие выпуклости можно определить следующим образом. Функция f: $R \to R$ называется выпуклой на отрезке a, b, если для любых двух точек $x_1, x_2 \in a, b$ выполняется соотношение:

$$\forall \vartheta \in 0, 1: \quad f\ \vartheta x_1 + 1 - \vartheta\ x_2 \ \le \vartheta f\ x_1\ + 1 - \vartheta\ f\ x_2\ .$$

Данное неравенство имеет прозрачный геометрический смысл. Значения функции на отрезке x_1, x_2 не больше, чем ординаты отрезка, соединяющего точки $f\ x_1$, $f\ x_2$. Если в неравенстве знак неравенства изменить на обратный, то получим определение *вогнутой* функции.

Перейдем теперь к общему случаю.

Функционал g: $L \to R$ называется *калибровочной функцией* на линейном пространстве L, если

$$\forall x, y \in L: \ g\ x + y\ \le g\ x\ + g\ y\ ,$$

$$\forall \alpha \ge 0, \ \alpha \in R: \ g\ \alpha x\ = \alpha g\ x\ .$$

Функционал $\mu: L \to R$ называется *выпуклым* на L, если

$$\forall \vartheta \in 0, 1: \ \mu\ \vartheta x_1 + 1 - \vartheta\ x_2\ \le \vartheta \mu\ x_1\ + 1 - \vartheta\ \mu\ x_2\ .$$

Ясно, что калибровочная функция есть выпуклый функционал. При наличии строгих неравенств говорят о *строгой выпуклости*.

Пусть $L_\circ \subset L$ — некоторое подпространство линейного пространства L. Пусть также заданы линейные функционалы:

$$f_\circ: L_\circ \to R, \quad f: L \to R\ .$$

Если $\forall x \in L_\circ: \ f\ x\ = f_\circ\ x\ $, то функционал f называется *продолжением* функционала $f_\circ$.

Теорема 1.6 (теорема *Хана — Банаха* о продолжении). Пусть:

1. L — вещественное линейное пространство, $L_\circ \subset L$ — некоторое подпространство линейного пространства L.
2. g — калибровочная функция на L.
3. $f_\circ : L_\circ \to R$ — линейный функционал, удовлетворяющий неравенству $f_\circ\ x\ \le g\ x\ ,\ x \in L_\circ$.

Тогда функционал $f_\circ$ может быть продолжен до линейного функционала f на L, для которого выполняется неравенство: $f\ x\ \le g\ x\ ,\ x \in L$.

Доказательство основано на лемме Цорна теории множеств и здесь не приводится (его можно найти в любом стандартном курсе функционального анализа). Существуют различные эквивалентные формулировки этой теоремы, отличающиеся от приведенной ранее.

Теорема Хана — Банаха отражает один из основных принципов функционального анализа. В частности, на этой теореме основаны важные факты выпуклого анализа, связанные с понятиями отделимости для выпуклых множеств. В свою очередь, понятия отделимости позволяют исследовать почти все условия оптимальности и соотношения двойственности.

1.3.4. Отделимость выпуклых множеств в линейном пространстве

Основное свойство выпуклых множеств, которое делает их таким ценным орудием при исследовании оптимизационных задач, связано с понятием отделимости. Далее мы сформулируем теорему об отделимости.

Пусть $M \subset L,\ \ N \subset L$. Говорят, что линейный функционал f: $L \to R$ разделяет множества M, N, если $\exists C \in R$ такое, что

$$x \in M \Rightarrow f\ x\ \ge C,$$
$$x \in N \Rightarrow f\ x\ \le C.$$

Используя теорему Хана — Банаха, можно получить следующий результат.

Теорема 1.7. Пусть:

1. $M,\ N \subset L$;
2. $M \cap N = \varnothing$;
3. M, N — выпуклые множества и хотя бы одно из них является выпуклым телом.

Тогда существует ненулевой линейный функционал f: $L \to R$, разделяющий M и N.

Доказательство теоремы здесь не приводится.

Замечание. Нулевой функционал существует в указанном смысле всегда. Достаточно тогда положить $C = 0$.

1.4. Нормированные пространства

В теории метрических пространств сформулировано понятие расстояния между элементами произвольного множества. В этом смысле метрические пространства являются частным случаем более общих *топологических пространств*. Концепция же линейных пространств позволяет наделить множество некоторой алгебраической структурой с помощью определения операций сложения элементов и умножения их на числа. Нормированные пространства являются одновременно линейными и метрическими пространствами и относятся к важному классу топологических линейных пространств. Развитие теории нормированных пространств связано с именем Стефана Банаха и целого ряда других авторов. Структура нормированных пространств оказывается чрезвычайно удобной для изучения основных конструкций теории оптимизации в достаточно общем виде.

1.4.1. Банаховы пространства

Линейное пространство L называется *нормированным*, если на L задан функционал f: $L \to R$, удовлетворяющий следующим четырем условиям для $\forall x,\ y \in L,\ \forall \lambda \in R$:

1. $f\ x\ \geq 0.$
2. $f\ x\ = 0 \Leftrightarrow x = 0.$
3. $f\ \lambda x\ = |\lambda| f\ x\ .$
4. $f\ x+y\ \leq f\ x\ + f\ y\ $ (неравенство треугольника).

Такой функционал f называется *нормой* в L. Значение $f\ x\ $ обозначается $\|x\|$ и называется *нормой элемента x*. *Нормированным пространством* называется линейное пространство L с заданной в нем нормой.

Если $x \to \|x\|$ — норма в L, то функционал $L \times L \to R$ вида $d\ x,\ y\ = \|x - y\|$ есть расстояние в L. Норма индуцирует соответствующую метрику. Справедливость аксиом метрики легко проверяется.

Таким образом, для нормированных пространств имеют смысл все понятия, введенные для метрических пространств. *Полное нормированное пространство* называется *банаховым* пространством или *B*-пространством.

Пример 1. В пространстве $C\ a, b\ $ непрерывных функций определим норму формулой:

$$\|f\| = \max_{t \in a,\ b} |f\ t\ |.$$

Порожденная этой нормой метрика совпадает с ранее рассматриваемой метрикой для этого множества. Как мы указывали, это множество функций является полным относительно своей метрики, и, следовательно, пространство $C\ a, b\ $ является банаховым пространством.

Подпространством нормированного пространства L называется подпространство *линейного пространства L* (линейное подпространство), которое является *замкнутым* множеством относительно расстояния, индуцированного заданной нормой. Иначе говоря, подпространство нормированного пространства есть линейное подпространство, содержащее все свои предельные точки. Еще раз отметим, что только замкнутые линейные подпространства нормированного пространства будут называться подпространствами нормированного пространства.

Например, в пространстве $C\ a, b$ непрерывных функций с указанной нормой многочлены образуют подпространство соответствующего линейного пространства, но не замкнутое. Следовательно, это подпространство не будет подпространством нормированного пространства $C\ a, b$. В конечномерном нормированном пространстве ситуация проще. Там любое линейное подпространство обязательно замкнуто.

Совокупность элементов (не обязательно замкнутую), содержащую вместе с x, y их произвольную линейную комбинацию $\alpha x + \beta y$ (подпространство линейного пространства), будем в случае нормированных пространств называть *линейным многообразием*.

Для нормированных пространств, являющихся частным случаем линейных пространств, сохраняются все определения и результаты, полученные для линейных пространств. Например, такие понятия, как наименьшее подпространство, порожденное подпространство, линейная оболочка и т. д.

В теории нормированных пространств *замыкание* линейной оболочки произвольного непустого множества x_α называется подпространством, порожденным элементами x_α . (Можно доказать, что указанное замыкание действительно будет линейным подпространством.)

Система элементов x_α нормированного пространства *L* называется *полной*, если порождаемое ею линейное многообразие имеет замыкание, совпадающее со всем пространством *L*. Иначе говоря, если порождаемое ею подпространство совпадает с *L*.

Пример 2. Система функций $1,\ t,\ t^2,\ t^3,\ ...,\ t^n,\ ...$ полна в пространстве непрерывных функций $C\ a, b$.

1.4.2. Евклидовы пространства

В евклидовых пространствах вводится понятие скалярного произведения, а уже на его основе определяется норма.

Пусть в действительном линейном пространстве *L* задан функционал $L \times L \to R$. Значение этого функционала называется *скалярным произведением* и обозначается $x, y\ ,\ x \in L,\ y \in L$, если выполняются следующие условия:

1. $\forall x \in L:\ \ x, x\ \geq 0.$

2. $x, x\ = 0 \Leftrightarrow x = 0.$

3. $\lambda x,\ y\ = \lambda\ x,\ y\ ,\ \lambda \in R.$

4. $x_1 + x_2,\ y\ =\ x_1,\ y\ +\ x_2,\ y\ .$

5. $x,\ y\ =\ y,\ x\ .$

Линейное пространство с заданным в нем скалярным произведением называется *евклидовым* пространством. Иногда евклидовы пространства называются *предгильбертовыми* пространствами, а для скалярного произведения используется обозначение x/y .

Норма в евклидовом пространстве вводится формулой

$$\|x\| = \sqrt{x,\ x}\ .$$

Можно проверить, что все аксиомы нормы оказываются при этом выполненными.

В евклидовом пространстве может быть задан угол между векторами. Для ненулевых векторов $x \in L,\ \ y \in L$ угол φ определяется выражением

$$\cos\varphi = \frac{x,\ y}{\|x\| \cdot \|y\|}.$$

Можно доказать, что, как и должно быть, правая часть равенства не превосходит единицы.

Если для *ненулевых* векторов $x \in L,\ \ y \in L$ имеем $x,\ y\ = 0$, то $\varphi = \pi/2$, а векторы x и y называются *ортогональными*.

Система *ненулевых* векторов x_α из L называется *ортогональной*, если они попарно ортогональны:

$$\alpha \neq \beta \Rightarrow\ x_\alpha,\ x_\beta\ = 0.$$

Счетность множества x_α не предполагается.

Упражнение 1.6

Доказать, что если векторы x_α ортогональны, то они линейно независимы.

Если система векторов (семейство элементов) x_α ортогональна и полна в L, то она называется *ортогональным базисом* пространства L. Если при этом $\forall\alpha : \|x_\alpha\| = 1$, то имеем *ортогональный нормированный* базис или *ортонормальный* базис. Если x_α — ортогональная система, то $x_\alpha / \|x_\alpha\|$ — ортонормальная.

Примеры евклидовых пространств.

1. Пространство действительных чисел R. Скалярное произведение — обычное произведение действительных чисел.

2. R^n — *n-мерное арифметическое пространство* с элементами вида $x = x_1, x_2, ..., x_n$, где x_i — действительные числа. Операции сложения и умножения на числа — общеизвестны, а скалярное произведение задается соотношением $x, y = \sum_{i=1}^{n} x_i y_i$.

3. Линейное пространство $C^2 a, b$ непрерывных на a, b действительных функций со скалярным произведением

$$f, g = \int_a^b f t g t dt$$

является евклидовым пространством. Можно установить, что все аксиомы скалярного произведения оказываются выполненными.

В этом случае норма, очевидно, задается выражением

$$\|f\|^2 = f, f = \int_a^b f t ^2 dt.$$

Индуцированная этой нормой метрика имеет вид

$$d f, g = \|f - g\| = \left(\int_a^b f t - g t ^2 dt \right)^{1/2},$$

что совпадает с ранее введенной метрикой при наделении данного множества функций структурой метрического пространства и выборе обозначения $C^2 a, b$.

Одним из ортогональных базисов пространства $C^2 a, b$ является тригонометрическая система функций:

$$1/2, \cos n \frac{2\pi t}{b-a}, \sin n \frac{2\pi t}{b-a} \quad n = 1, 2, ...$$

4. Ранее рассматривалось также метрическое пространство $C^2 a, b$ с метрикой

$$d f, g = \max_{t \in a, b} |f t - g t|.$$

Норма определялась формулой

$$\|f\| = \max_{t \in a, b} |f t|.$$

(Было установлено, что это банахово пространство.)

Поставим вопрос, можно ли наделить данное нормированное пространство структурой евклидова пространства. Для этого достаточно задать вышеприведенную норму с помощью некоторого скалярного произведения:

$$\|x\| = \sqrt{x, x}.$$

Можно показать, что ответ будет отрицательным. Норму пространства $C\ a, b$ нельзя задать с помощью скалярного произведения. Таким образом, не все нормированные пространства можно "превратить" в евклидовы пространства. Евклидовы пространства составляют лишь часть нормированных пространств. Еще раз отметим, что пространство $C\ a, b$ дает пример банахова, но не евклидова пространства.

Можно доказать следующее утверждение (характеристическое свойство евклидовых пространств).

Теорема 1.8. Для того чтобы нормированное пространство L было евклидовым, необходимо и достаточно, чтобы для $\forall x,\ y \in L$ выполнялось равенство

$$\|x+y\|^2 + \|x-y\|^2 = 2\ \|x\|^2 + \|y\|^2 \quad .$$

Доказательство опускаем.

Теорема 1.9 (процесс ортогонализации Шмидта). Пусть

$$f_1,\ f_2, \ldots, f_n, \ \ldots$$

есть линейно независимая (счетная) система векторов в евклидовом пространстве L. (Ясно, что в эту систему не могут входить нулевые векторы, иначе получим линейную зависимость). Тогда в L существует система векторов

$$\varphi_1,\ \varphi_2, \ldots, \varphi_n, \ldots,$$

Такая, что:

1. φ_i — ортонормальна.
2. $\varphi_n = \sum_{i=1}^{n} \alpha_{ni} f_i$, причем $\alpha_{nn} \neq 0$.

(Переход от линейно независимой системы векторов к ортогональной называется процессом ортогонализации.)

Доказательство. Положим

$$\psi_1 = f_1 ,$$

$$\psi_{n+1} = f_{n+1} - \sum_{i=1}^{n} \frac{f_{n+1},\ \psi_i}{\|\psi_i\|^2} \psi_i = f_{n+1} - \sum_{i=1}^{n} \beta_i f_i . \tag{1.3}$$

В этом случае вектор ψ_{n+1} будет ортогонален всем векторам $\psi_i,\ i = 1, \ldots, n$. Действительно,

$$\psi_{n+1},\ \psi_j \ = \ f_{n+1},\ \psi_j \ - \sum_{i=1}^{n} \frac{f_{n+1},\ \psi_i}{\|\psi_i\|^2} \ \psi_i,\ \psi_j \ .$$

Для $n = 1$ имеем

$$\psi_2 = f_2 - \frac{f_2,\ \psi_1}{\|\psi_1\|^2} \psi_1 .$$

Но $\psi_1 = f_1$, поэтому

$$\psi_2 = f_2 - \frac{f_2,\ f_1}{\|f_1\|^2} f_1,$$

$$\psi_1,\ \psi_2 \ = \ f_2,\ \psi_1 \ - \frac{f_2,\ f_1}{\|f_1\|^2} \ f_1,\ \psi_1 \ = \ f_2,\ f_1 \ - \frac{f_2,\ f_1}{\|f_1\|^2} \|f_1\|^2 = 0.$$

Установлена ортогональность ψ_1 и ψ_2. Покажем по индукции, что система векторов $\psi_1, ..., \psi_n, ...$, построенная согласно выражению (1.3), ортогональна.

Пусть $\psi_1, ..., \psi_n$ уже построены и ортогональны. Покажем, что тогда вектор ψ_{n+1} будет ортогонален всем ψ_i, $i = 1, ..., n$.

Имеем (для любого фиксированного $j = 1, ..., n$:

$$\psi_{n+1},\ \psi_j \ = \ f_{n+1},\ \psi_j \ - \sum_{i=1}^{n} \frac{f_{n+1},\ \psi_i}{\|\psi_i\|^2} \ \psi_i,\ \psi_j \ =$$

$$= \ f_{n+1},\ \psi_j \ - \frac{f_{n+1},\ \psi_j}{\|\psi_j\|^2} \ \psi_j,\ \psi_j \ = 0.$$

(Очевидно, в данном выражении только одно слагаемое суммы не равно нулю, а именно слагаемое с индексом $i = j$. Остальные слагаемые равны нулю, т. к. ψ_i, $i = 1, ..., n$ ортогональны и $\psi_i, \psi_j = 0,\ i \neq j$.)

Последнее соотношение может выполняться и если в результате процедуры (1.3) будут все время получаться нулевые векторы. Покажем, что это невозможно из-за линейной независимости системы векторов $f_1, f_2, ..., f_n, ...$ Действительно, пусть получили $\psi_{n+1} = 0$, тогда

$$f_{n+1} = \sum_{i=1}^{n} \frac{f_{n+1},\ \psi_i}{\|\psi_i\|^2} \psi_i = \sum_{i=1}^{n} \gamma_i f_i,$$

что противоречит линейной независимости $f_1, f_2, ..., f_n, ...$ Следовательно $\psi_{n+1} \neq 0$. Для окончательного доказательства теоремы достаточно положить

$$\varphi_n = \frac{\psi_n}{\|\psi_n\|}.$$

Теорема доказана.

1.4.3. Ряд Фурье. Коэффициенты Фурье

Как уже говорилось, линейное пространство L называется конечномерным, если в нем существует n линейно независимых векторов, а любые $n+1$ векторов — линейно зависимы (говорят, что L имеет *размерность* n). Бесконечномерным линейным пространством мы называли такое L, где существует любое количество линейно независимых векторов. Будем далее говорить об евклидовых пространствах, т. е. о линейных пространствах со скалярным произведением.

В n-мерном евклидовом пространстве L^n выберем ортонормальный базис $e_1, ..., e_n$. Он всегда существует в силу теоремы об ортогонализации. Тогда любой вектор $x \in L^n$ представляется в виде:

$$x = \sum_{i=1}^{n} c_i e_i, \quad c_i = \langle x, e_i \rangle .$$

(Правомерность почленного взятия скалярного произведения следует из четвертой аксиомы скалярного произведения.)

Попытаемся делать подобные разложения и в случае бесконечномерных евклидовых пространств. Попытаемся произвольный вектор f из L представить в виде

$$f = \sum_{i=1}^{\infty} c_i \varphi_i ,$$

где $\varphi_1, \varphi_2, ..., \varphi_n, ...$ — счетная ортонормальная система векторов в евклидовом пространстве L. Заметим, что система векторов $\varphi_1, \varphi_2, ..., \varphi_n, ...$ базисом не является, т. к. полнота этой системы не предполагалась. Будем поступать следующим образом.

По вектору $f \in L$ строим набор чисел

$$c_k = \langle f, \varphi_k \rangle ,$$

которые будем называть *коэффициентами Фурье* элемента f относительно системы $\varphi_1, \varphi_2, ..., \varphi_n, ...$ Рассмотрим ряд

$$\sum_{i=1}^{\infty} c_i \varphi_i .$$

Этот ряд называется *рядом Фурье* для элемента f по системе векторов φ_i :

$$f \sim \sum_{i=1}^{\infty} c_i \varphi_i .$$

Мы здесь не ставим знака равенства, т. к. о сходимости этого ряда к элементу f ничего не известно (ряд может не сходиться или сходиться к другому элементу).

Рассмотрим частичную сумму ряда:

$$s_n = \sum_{i=1}^{n} \alpha_i \varphi_i .$$

Подберем α_i так, чтобы норма $\|f - s_n\|$ или, что эквивалентно, квадрат этой нормы были минимальными. Имеем:

$$\begin{aligned}
\|f - s_n\|^2 = \ & f - s_n,\ f - s_n \ = \left(f - \sum_{i=1}^{n} \alpha_i f_i,\ f - \sum_{i=1}^{n} \alpha_i f_i \right) = \\
= \ & f,\ f \ - 2\left(f,\ \sum_{i=1}^{n} \alpha_i f_i \right) + \left(\sum_{i=1}^{n} \alpha_i f_i,\ \sum_{i=1}^{n} \alpha_i f_i \right) = \\
= \ & \|f\|^2 - 2\sum_{i=1}^{n} \alpha_i \ \varphi_i,\ f \ + \sum_{i=1}^{n} {\alpha_i}^2 = \\
= \ & \|f\|^2 - 2\sum_{i=1}^{n} \alpha_i c_i + \sum_{i=1}^{n} {\alpha_i}^2 + \sum_{i=1}^{n} {c_i}^2 - \sum_{i=1}^{n} {c_i}^2 = \\
= \ & \|f\|^2 - \sum_{i=1}^{n} {c_i}^2 + \sum_{i=1}^{n} \ \alpha_i - c_i \ ^2 .
\end{aligned}$$

(Здесь мы использовали формулу "квадрат суммы".) Ясно, что необходимо положить $\alpha_i = c_i$. В этом случае справедливо соотношение:

$$0 \le \|f - s_n\|^2 = \|f\|^2 - \sum_{i=1}^{n} {c_i}^2 ,$$

или

$$\forall n: \ \sum_{i=1}^{n} {c_i}^2 \le \|f\|^2 .$$

Таким образом, последовательность частичных сумм числового ряда

$$\sum_{i=1}^{\infty} {c_i}^2$$

является неубывающей и ограниченной сверху последовательностью чисел. Поэтому она имеет предел и сам ряд сходится. Получили известное *неравенство Бесселя* для коэффициентов Фурье:

$$\sum_{i=1}^{\infty} {c_i}^2 \le \|f\|^2 .$$

Доказав сходимость ряда из квадратов коэффициентов Фурье, можно перейти к пределу при $n \to \infty$ в полученном ранее тождестве:

$$\|f - s_n\|^2 = \|f\|^2 - \sum_{i=1}^{n} {c_i}^2 .$$

Правая (а значит и левая) часть этого равенства стремится к пределу

$$\|f\|^2 - \sum_{i=1}^{\infty} c_i^{\,2} .$$

Отсюда следует, что необходимым и достаточным условием сходимости $s_n \to f$, $n \to \infty$ (сходимость понимается в смысле стремления к нулю нормы $\|s_n - f\| \to 0$, $n \to \infty$) является выполнение так называемого *равенства Парсеваля*

$$\sum_{i=1}^{\infty} c_i^{\,2} = \|f\|^2 .$$

Если такое равенство выполняется для $\forall f \in L$, то соответствующая ортонормальная счетная система φ_i называется *замкнутой*. Работая с различными евклидовыми пространствами, полезно знать в них замкнутые системы векторов, т. к. в этом случае мы сможем любой элемент пространства представить в виде ряда Фурье по такой системе. Легко понять, что равенство Парсеваля может рассматриваться как обобщение известной теоремы Пифагора на случай бесконечной ортонормальной системы векторов.

Теорема 1.10 (о связи полноты и замкнутости). В сепарабельном евклидовом пространстве всякая полная ортонормальная система является замкнутой и обратно.

Доказательство можно найти в любом учебнике по функциональному анализу. Напомним, что понятие сепарабельности было введено нами для метрических пространств.

1.4.4. Гильбертовы пространства

Гильбертовым пространством называется *полное евклидово пространство*. (Напомним, что понятие полноты уже вводилось нами ранее для метрических пространств.) Таким образом, гильбертовы пространства — это частный случай евклидовых (предгильбертовых) пространств. В то же время гильбертово пространство дает пример полного линейного нормированного пространства (со специфической нормой) и поэтому является частным случаем банахова пространства. Ясно, что евклидово пространство не обязано быть банаховым, т. к. оно может быть неполно.

Иногда при определении гильбертова пространства выдвигается дополнительное требование бесконечномерности, т. е. существования любого числа линейно независимых элементов. Мы таких предположений не делаем и, следовательно, допускаем существование конечномерных гильбертовых пространств. Кроме того, часто при изучении гильбертовых пространств явно или неявно предполагают их сепарабельность.

Можно показать, что любое конечномерное евклидово пространство является гильбертовым (т. е. полно).

Изоморфизмом евклидова пространства L на евклидово пространство L' называется биективное отображение f: $L \to L'$ такое, что

$$\forall x,\ y \in L: \quad f\ x+y\ = f\ x\ + f\ y\ ,$$
$$f\ \alpha x\ = \alpha f\ x\ ,$$
$$f\ x\ , f\ y\ \ = \ x,\ y\ .$$

Таким образом, изоморфизм есть такое взаимно однозначное соответствие, которое сохраняет как линейные операции, определенные в пространствах L, L', так и скалярное произведение. Евклидовы пространства называются *изоморфными*, если можно построить соответствующий изоморфизм f.

Рассмотрим конечномерные (n-мерные) евклидовы пространства, т. е. такие линейные пространства со скалярным произведением, для которых существует конечное число n линейно независимых векторов, а любые $n+1$ векторов — линейно зависимы. Можно доказать, что любые два n-мерных евклидова пространства изоморфны между собой. Следовательно, любое такое пространство изоморфно арифметическому пространству R^n (состоящему из наборов n действительных чисел). Если в каком-нибудь n-мерном евклидовом пространстве L^n доказана теорема, сформулированная в терминах операций сложения, умножения на числа и скалярного произведения векторов, то эта же теорема верна и в любом изоморфном ему пространстве. В качестве такого "стандартного" пространства часто используется пространство R^n.

Евклидовы пространства бесконечного числа измерений не обязаны быть изоморфными между собой.

Для гильбертовых пространств в общем случае справедливо следующее утверждение.

Теорема 1.11. Если два бесконечномерных гильбертова пространства сепарабельны, то они изоморфны между собой.

(Напомним, что метрическое пространство E называется *сепарабельным*, если в E существует не более чем счетное (т. е. конечное или счетное) всюду плотное множество.)

Доказательство этой теоремы проводить не будем. Отметим только, что оно заключается в установлении изоморфизма любого бесконечномерного сепарабельного гильбертова пространства евклидову пространству l_2 бесконечных последовательностей действительных чисел:

$$x = \ x_1,\ x_2,\ ...,\ x_n,\ ...\ , \quad \sum_{i=1}^{\infty} x_i^{\,2} < \infty,$$

$$\forall x,\ y \in l_2: \quad x,\ y\ = \sum_{i=1}^{\infty} x_i y_i\ .$$

(Здесь ряд в правой части выражения для скалярного произведения сходится. Можно показать, что евклидово пространство l_2 является примером сепарабельного бесконечномерного гильбертова пространства.)

Аналогично пространству R^n в случае конечного числа измерений, в бесконечномерном случае пространство l_2 является "стандартным" гильбертовым пространством. Можно сказать более определенно. С точностью до изоморфизма существует *только одно* сепарабельное бесконечномерное гильбертово пространство l_2. Концепция сепарабельности в теории гильбертовых пространств настолько важна, что требование сепарабельности часто изначально включается в определение гильбертова пространства и отдельно затем не оговаривается.

1.4.5. Ортогональное дополнение

Пусть E — гильбертово пространство и $M \subset E$ — его подпространство. (Напомним, что подпространством нормированного пространства L называется подпространство линейного пространства L, которое является замкнутым множеством относительно расстояния, индуцированного заданной нормой. Иначе говоря, подпространство нормированного пространства есть линейное подпространство, содержащее все свои предельные точки. Еще раз отметим, что только замкнутые линейные подпространства нормированного пространства называются подпространствами нормированного пространства.) Обозначим через $M^{\perp}$ множество элементов E, ортогональных ко всем элементам $x \in M$, т. е.

$$M^{\perp} = \{ x \in E / \ (x, y) = 0 \ \forall y \in M \} .$$

Можно доказать, что множество $M^{\perp}$ будет подпространством. $M^{\perp}$ называется *ортогональным дополнением* подпространства M.

Теорема 1.12. Если $M \subset E$ — подпространство сепарабельного гильбертова пространства E, то любой $x \in E$ единственным образом представим в виде

$$x = y_1 + y_2 , \quad y_1 \in M, y_2 \in M^{\perp}.$$

Доказательство опускается.

Иногда говорят, что пространство E есть *прямая сумма* подпространств M и $M^{\perp}$. Обозначение:

$$E = M \oplus M^{\perp} .$$

1.5. Линейные операторы в нормированном пространстве

Пусть E и E' — два нормированных пространства и $A: E \to E'$ — отображение или *оператор*, определенный на пространстве E с областью значений в E'. Термин "оператор" в теории нормированных пространств используется как синоним назва-

ний "отображение" или "функция". Ничего качественно нового здесь не появилось. Для отображений уже вводилось обозначение вида $y = A\ x$. Часто будем писать $y = Ax$.

Оператор A называется *линейным*, если

$$\forall x_1,\ x_2 \in E,\ \forall \alpha_i \in R:\quad A\ \alpha_1 x_1 + \alpha_2 x_2\ = \alpha_1 A x_1 + \alpha_2 A x_2\ .$$

Примеры линейных операторов.

1. Рассмотрим пространство непрерывных на a, b действительных функций. Введем оператор в этом пространстве, определяемый формулой

$$\varphi\ t\ = \int_a^b K\ t, \tau\ \psi\ \tau\ d\tau,$$

где $K\ t, \tau$ — заданная непрерывная функция двух переменных. Оператор имеет вид $\varphi = A\psi$ и переводит пространство непрерывных на a, b функций в себя (т. к. φ непрерывна, если непрерывна ψ). Нетрудно показать, что оператор A линеен.

2. Пусть E — сепарабельное гильбертово пространство и $M \subset E$ — его подпространство. Рассмотрим прямую сумму $E = M \oplus M^{\perp}$. Тогда любой элемент $x \in E$ представляется в виде

$$x = y_1 + y_2\ ,\quad y_1 \in M,\ y_2 \in M^{\perp}.$$

Введем оператор $P_M x = y_1$. Этот оператор называется *ортопроектором* E на M. Легко проверить, что это линейный оператор.

1.5.1. Непрерывность и ограниченность

Пусть E и E' — два нормированных пространства. Линейный оператор $A: E \to E'$ называется *ограниченным*, если он определен на всем E и любое ограниченное множество переводит снова в ограниченное множество (т. е. множество, диаметр которого конечен). Сформулируем без доказательства следующую важную теорему.

Теорема 1.13. Любой непрерывный линейный оператор ограничен и наоборот.

(Понятие непрерывного отображения было введено при рассмотрении метрических пространств.) Можно дать следующее эквивалентное определение ограниченности. Линейный оператор A называется ограниченным, если

$$\exists M \in R:\quad \forall x \in E\quad \|Ax\|_{E'} \le M \|x\|_E\ .$$

Наименьшее из таких чисел M называется *нормой линейного ограниченного оператора* и обозначается как $\|A\|$. Таким образом, имеем соотношение:

$$\|Ax\|_{E'} \le \|A\| \cdot \|x\|_E\ .$$

Теорема 1.14. Пусть заданы два нормированных пространства E, E' и линейный ограниченный оператор $A\colon E \to E'$. Тогда справедливо соотношение:

$$\|A\| = \sup_{\|x\|_E=1} \|Ax\|_{E'} = \sup_{x\neq 0} \frac{\|Ax\|_{E'}}{\|x\|_E}.$$

Доказательство. Докажем вначале второе равенство. Так как согласно аксиомам нормы имеем

$$\|\lambda x\| = |\lambda|\|x\|,$$

то, обозначая $\lambda = 1/\|x\|_E$, получим:

$$\sup_{x\neq 0} \frac{\|Ax\|_{E'}}{\|x\|_E} = \sup_{x\neq 0} \lambda \|Ax\|_{E'} = \sup_{x\neq 0} \|\lambda Ax\|_{E'} = \quad A \text{ — линейный оператор } =$$

$$= \sup_{x\neq 0} \|A\lambda x\|_{E'} = \sup_{x\neq 0} \left\| A \frac{x}{\|x\|_E} \right\|_{E'} = \left\{ \frac{x}{\|x\|_E} \triangleq \xi,\ \|\xi\|_E = 1 \right\} = \sup_{\|\xi\|_E=1} \|A\xi\|_{E'}.$$

Второе равенство доказано. Теперь установим справедливость первого равенства.

По определению нормы $\|A\|$ имеем

$$\forall x \in E\colon \quad \|Ax\|_{E'} \le \|A\| \cdot \|x\|_E.$$

Ясно, что тогда и

$$\alpha \triangleq \sup_{\|x\|_E=1} \|Ax\|_{E'} \le \|A\| \cdot \|x\|_E = \|A\|.$$

Теперь достаточно показать, что

$$\alpha \ge \|A\|.$$

Имеем

$$\forall x \neq 0\colon \quad \|Ax\|_{E'} = \frac{\|Ax\|_{E'}}{\|x\|_E} \|x\|_E.$$

Отсюда

$$\|Ax\|_{E'} \le \sup_{x\neq 0} \frac{\|Ax\|_{E'}}{\|x\|_E} \|x\|_E = \alpha \|x\|_E.$$

И так как $\|A\|$ есть минимальное из чисел M, удовлетворяющих последнему неравенству вида

$$\|Ax\|_{E'} \le M \|x\|_E,$$

то получаем

$$\alpha \geq \|A\|.$$

Следовательно, $\alpha = \|A\|$. Что и требовалось доказать. Теорема доказана полностью.

1.5.2. Пространство ограниченных линейных операторов

Пусть заданы два нормированных пространства E, E' и два линейных ограниченных (и поэтому непрерывных) оператора $A, B : E \to E'$.

Определим сложение операторов посредством формулы

$$\forall x \in E: \quad A + B \;\; x = Ax + Bx,$$

а умножение линейного оператора на число — формулой

$$\forall x \in E: \quad \lambda A \;\; x = \lambda Ax.$$

Легко проверить, что $A + B$ и λA также будут линейными и ограниченными операторами. Линейность проверяется тривиально. Докажем ограниченность оператора $A + B$. Надо доказать:

$$\forall x \in E: \quad \| \; A + B \;\; x\| \leq Mx.$$

Имеем:

$$\forall x \in E: \quad \| \; A + B \;\; x\| = \|Ax + Bx\| \leq \;\; \|A\| + \|B\| \;\; \|x\|.$$

Таким образом, ограниченность доказана. Кроме того, доказана формула

$$\|A + B\| \leq \|A\| + \|B\|.$$

Совокупность $L \;\; E; E'$ всех непрерывных линейных операторов $E \to E'$ образует по отношению к введенным операциям сложения операторов и умножения на числа *линейное пространство*. Это пространство является нормированным согласно определению нормы оператора.

1.5.3. Сопряженное пространство

В частном случае, когда $E' = R$, т. е. когда оператор $E \to E'$ является линейным функционалом, пространство $L \;\; E; R$ называется пространством, *сопряженным* с E, и обозначается E^*. Понятие сопряженного пространства оказывается важным из-за следующей теоремы.

Теорема 1.15. Пространство E^* полно (независимо от того, полно или нет пространство E).

Доказательство. Пусть f_n — любая фундаментальная последовательность линейных функционалов из пространства E^*, т. е.

$$\forall \varepsilon > 0 : \exists N : \quad n,\ m \geq N \Rightarrow \|f_n - f_m\| < \varepsilon .$$

Для доказательства полноты E^* необходимо доказать, что эта фундаментальная последовательность сходится к элементу того же пространства E^*.

Имеем:

$$\forall x \in E : \quad \left| f_n\ x\ - f_m\ x\ \right| = \left|\ f_n - f_m\quad x\ \right| \leq \|f_n - f_m\| \cdot \|x\| < \varepsilon \|x\|.$$

Следовательно, при $\forall x \in E$ числовая последовательность $f_n\ x$ сходится по классическому критерию Коши. Введем функционал

$$f\ x\ \triangleq \lim_{n\to\infty} f_n\ x\ .$$

Покажем, что f является линейным и ограниченным функционалом $E \to R$. Докажем линейность:

$$f\ \alpha x + \beta y\ = \lim_{n\to\infty} f_n\ \alpha x + \beta y\ = \lim_{n\to\infty} \left[\alpha f_n\ x\ + \beta f_n\ y\ \right] = \alpha f(x) + \beta f(y) .$$

Для доказательства ограниченности перейдем к пределу при $m \to \infty$ в уже полученном неравенстве

$$\left| f_n\ x\ - f_m\ x\ \right| < \varepsilon \|x\|.$$

Отсюда имеем:

$$\left| f_n(x) - f(x) \right| = \left| f(x) - f_n(x) \right| \leq \varepsilon \|x\| .$$

Следовательно, функционал $f - f_n$ ограничен. Сумма двух ограниченных операторов (а значит и функционалов) также ограничена, поэтому функционал

$$f - f_n\ + f_n = f$$

ограничен (а значит и непрерывен). Поэтому $f \in E^*$. Из неравенства

$$\left| f\ x\ - f_n\ x\ \right| \leq \varepsilon \|x\|$$

следует

$$\|f - f_n\| \leq \varepsilon,$$

т. е. фундаментальная последовательность функционалов f_n сходится к f. Таким образом, показано, что любая фундаментальная последовательность линейных ограниченных функционалов сходится к некоторому функционалу, который также является линейным и ограниченным.

Обращаем внимание на то, что мы здесь везде отличали последовательности *значений* функционалов $f_n\ x$ от последовательностей функционалов f_n. В первом случае это числовые последовательности элементов из R, а во втором — последовательности элементов пространства E^*. Теорема доказана.

1.5.4. Второе сопряженное пространство. Рефлексивность

С помощью понятия сопряженного пространства мы каждому нормированному пространству E поставили в соответствие другое нормированное пространство — пространство линейных ограниченных функционалов E^*. Считая, что исходным пространством является пространство E^*, можно так же построить пространство, сопряженное относительно E^*. По отношению к пространству E это новое пространство называется *вторым сопряженным* пространством и обозначается E^{**}. Следовательно, элемент пространства E^{**} есть непрерывный ограниченный линейный функционал, определенный в нормированном пространстве E^*.

Любому элементу x_0 пространства E можно следующим образом поставить в соответствие элемент из второго сопряженного пространства E^{**}. Положим

$$\psi_{x_0}\ f\ = f\ x_0\ ,\quad x_0 \in E, f \in E^*.$$

Если x_0 — фиксированный элемент пространства E, а f пробегает все пространство E^*, то последнее равенство любому элементу f из E^* ставит в соответствие число $f\ x_0$, т. е. определен функционал

$$\psi_{x_0} : E^* \to R.$$

Функционал ψ_{x_0} является линейным:

$$\psi_{x_0}\ \alpha f_1 + \beta f_2\ = \alpha f_1\ x_0\ + \beta f_2\ x_0\ = \alpha\psi_{x_0}\ f_1\ + \beta\psi_{x_0}\ f_2\ .$$

Можно также доказать, что всякий такой функционал непрерывен на E^*.

В результате построено отображение вида

$$x_0 \to \psi_{x_0},\quad x_0 \in E,\ \psi_{x_0} \in E^{**}.$$

Это отображение обозначается как

$$\pi : E \to E^{**}$$

и называется *естественным отображением* пространства E во второе сопряженное пространство. Можно доказать, что отображение π является изометрией.

π отображает пространство E в некоторое подмножество пространства E^{**}. Если π обладает свойствами:

1. π — биекция E на E^{**};
2. π — непрерывно,

тогда пространство E называется *рефлексивным*.

Таким образом, каждый элемент x из E мы можем рассматривать еще и как элемент E^{**} (если E — рефлексивно, то и наоборот). Поэтому для значений функционала $f \in E^*$ вместо обозначения $f\ x$ часто удобно использовать более симметричное обозначение:

$$f\ x = \langle f, x \rangle .$$

Иногда вместо угловых скобок используются круглые. При фиксированном $f \in E^*$ это функционал, определенный на E, а при фиксированном $x \in E^{**}$ — это функционал на E^*.

Было доказано, что если E — нормированное пространство, то E^* полно. Отсюда можно доказать, что всякое рефлексивное пространство полно.

Гильбертовы пространства (конечномерные и бесконечномерные) — простейшие примеры рефлексивных пространств. Для гильбертовых пространств можно даже доказать, что с точностью до изоморфизма линейных пространств $E = E^*$. Позже это утверждение будет существенно использовано.

1.5.5. Произведение операторов

Пусть даны линейные операторы

$$A : E \to E'$$
$$B : E' \to E''.$$

Произведением BA называется оператор $C : E \to E''$ такой, что $x \to B\ Ax$.

Если операторы A и B ограничены, то оператор BA также ограничен, причем

$$\|BA\| \le \|B\| \cdot \|A\| .$$

Действительно,

$$\|B(Ax)\| \le \|B\| \cdot \|Ax\| \le \|B\| \cdot \|A\| \cdot \|x\| .$$

Ограниченность BA доказана.

1.5.6. Обратный оператор

Пусть $A: E \to E'$ — оператор (не обязательно линейный), отображающий нормированное пространство E на подмножество $R_A \subset E'$ нормированного пространства E'.

Оператор A называется *обратимым*, если для $\forall y \in R_A$ уравнение

$$Ax = y$$

имеет единственное решение.

Если оператор A обратим, то $\forall y \in R_A$ можно поставить в соответствие единственный элемент $x \in E$, являющийся решением уравнения $Ax = y$. Оператор, осуществляющий это соответствие, называется обратным к A и обозначается A^{-1}.

Теорема 1.16. Для того чтобы линейный оператор A имел обратный, необходимо и достаточно, чтобы $Ax = 0$ только при $x = 0$. (Здесь оба нуля принадлежат разным пространствам.)

Доказательство. Докажем достаточность. Пусть $Ax = 0$ только при $x = 0$. Тогда, если x_1 и x_2 есть решения уравнения $Ax = y$, то $Ax_1 = y$ и $Ax_2 = y$. Отсюда следует $A\,(x_1 - x_2) = 0$ и, согласно предположению, $x_1 = x_2$. Единственность решения при $\forall y \in R_A$ установлена и оператор A обратим. Докажем необходимость. Пусть A обратим и A^{-1} существует. Тогда уравнение $Ax = 0$ имеет только одно решение, которым и будет $x = 0$.

Теорема 1.17. Если оператор A — линейный, то оператор A^{-1} (если он существует) тоже линеен.

Доказательство. Достаточно проверить выполнение равенства

$$A^{-1}\,(\alpha_1 y_1 + \alpha_2 y_2) = \alpha_1 A^{-1} y_1 + \alpha_2 A^{-1} y_2 .$$

Пусть $y_1 = Ax_1$, $y_2 = Ax_2$. Тогда, по определению обратного оператора,

$$x_1 = A^{-1} y_1, \quad x_2 = A^{-1} y_2 ,$$

т. е.

$$\alpha_1 A^{-1} y_1 + \alpha_2 A^{-1} y_2 = \alpha_1 x_1 + \alpha_2 x_2 .$$

В силу линейности оператора A имеем

$$A\,(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 y_1 + \alpha_2 y_2 ,$$

или, по определению обратного оператора,

$$A^{-1}\,(\alpha_1 y_1 + \alpha_2 y_2) = \alpha_1 x_1 + \alpha_2 x_2 .$$

Таким образом, требуемое равенство доказано.

Оператор, обратный к линейному непрерывному (т. е. ограниченному) оператору, может не быть непрерывным (т. е. быть неограниченным). Однако справедливо следующее утверждение.

Теорема 1.18 (теорема Банаха об обратном операторе). Пусть линейный ограниченный оператор A является биекцией банахова пространства E на банахово пространство E'. Тогда обратный оператор A^{-1} ограничен (ясно, что он линеен).

Доказательство опускаем.

1.5.7. Сопряженные операторы

Пусть $A: E \to E'$ — непрерывный (ограниченный) линейный оператор, отображающий нормированное пространство E в такое же пространство E'. Пусть $g \in E'^*$ (E'^* — сопряженное к E'), т. е. g является некоторым линейным функционалом на E'. Введем обозначение для произведения операторов: $f \triangleq gA \in E^*$, т. е. оператор f является линейным функционалом на E. Таким образом, получили некоторый оператор

$$A^*: E'^* \to E^*,$$

ставящий в соответствие каждому линейному функционалу g, определенному на E', линейный функционал f, определенный на E.

Этот оператор A^* называется *сопряженным* к оператору A. Связь рассмотренных операторов поясняется следующей диаграммой:

$$E \xrightarrow{A} E' \xrightarrow{g} R$$

$$E \xrightarrow{f=gA} R$$

$$g \in E'^* \xrightarrow{A^*} f \in E^*.$$

Обозначая значение произвольного функционала s на элементе t (т. е. некоторое число) символом $\langle s,\ t\rangle$, получим $\langle g,\ Ax\rangle = \langle f,\ x\rangle$ или

$$\langle g,\ Ax\rangle = \langle A^*g,\ x\rangle.$$

(Здесь угловые скобки не являются обозначением скалярного произведения.)

Еще раз заметим, что оператор A определен в пространстве E, а сопряженный оператор A^* определен в пространстве, сопряженном к E'. Поэтому, даже если пространства E и E' совпадают (как это часто бывает), оператор и его сопряженный оператор, строго говоря, определены в разных пространствах.

Можно доказать следующие свойства сопряженных операторов:

1. A^* — линеен;
2. $(A+B)^* = A^* + B^*$;
3. $(kA)^* = kA^*,\ k \in R$.

1.5.8. Сопряженные операторы в гильбертовом пространстве. Самосопряженные операторы

Пусть H — сепарабельное гильбертово пространство. Напомним, что гильбертово пространство является частным случаем евклидовых пространств (линейных векторных пространств со скалярным произведением). Пусть также A — произвольный ограниченный линейный оператор, заданный на H и принимающий значения в H. Определим с помощью скалярного произведения линейный функционал f_y на H соотношением

$$f_y\ x\ = \langle x,\ y \rangle,$$

где y — произвольный фиксированный элемент из H. Тогда можно доказать, что для любого линейного оператора A и любого $y \in H$ существует и при том единственный элемент $y^* \in H$ такой, что *при всех* $x \in H$

$$\langle Ax,\ y \rangle =\ x,\ y^*\ .$$

Таким образом, любому $y \in H$ ставится в соответствие некоторый единственный $y^* \in H$. Такой оператор называется *эрмитово-сопряженным* к оператору A и обозначается $\tilde{A}^*$. Теперь последнее равенство принимает вид

$$\langle Ax,\ y \rangle = \left(x,\ \tilde{A}^*\ y \right).$$

Чтобы не усложнять обозначения, оператор $\tilde{A}^*$ в теории гильбертовых пространств обозначается A^* и называется *сопряженным оператором*. Из последнего равенства получаем основное для сопряженного оператора соотношение

$$\langle Ax,\ y \rangle = \langle x,\ A^* y \rangle.$$

Повторное введение понятия сопряженного оператора указанным способом оправдывается тем, что *с точностью до изоморфизма* пространства H на сопряженное пространство H^* оператор $\tilde{A}^*$ действительно является сопряженным к оператору A в смысле общего определения. В теории гильбертовых пространств оператор и сопряженный к нему, в отличие от общего случая, действуют *в одном и том же* пространстве H. Если выполняется равенство $A = A^*$, то оператор A называется *самосопряженным*.

Пример. В вещественных конечномерных евклидовых пространствах (мы уже указывали, что любое конечномерное евклидово пространство является гильбертовым) самосопряженные операторы задаются симметричными матрицами, а матрицы оператора и сопряженного к нему связаны операцией транспонирования.

В некотором смысле, понятия сопряженности и самосопряженности обобщают понятия и свойства транспонированной и симметричной матриц операторов в конечномерном анализе.

1.5.9. Спектр оператора

Понятие спектра является важнейшим в теории операторов и обобщает понятие собственного числа матрицы в конечномерном анализе. А роль собственных чисел хорошо известна.

Обратимся вначале к конечномерному случаю и напомним основные соотношения. Пусть задан линейный оператор A в конечномерном комплексном пространстве C^n. Матрицу этого оператора (при каком-то фиксированном базисе) тоже обозначим через A. *Собственным значением (числом)* матрицы (и соответственно, оператора) A называется такое число $\lambda \in C$, для которого уравнение

$$Ax = \lambda x \text{ или } A - \lambda I \ x = 0$$

имеет ненулевое решение $x \neq 0$, $x \in C^n$. (Здесь I — единичная матрица.)

Замечание 1. В конечномерном случае доказывается, что собственные значения являются корнями некоторого алгебраического уравнения. Алгебраическое уравнение, вообще говоря, может не иметь корней. Соответственно, существуют линейные операторы, не имеющие собственных значений.

Перейдем к рассмотрению общего случая.

Далее везде, где речь идет о спектре, подразумеваются только линейные операторы и комплексные линейные пространства.

Оператор $I : E \to E$, где E — нормированное пространство, называется *тождественным* (или *единичным*), если $I \ x \ = x$.

Пусть $I : E \to E$ есть линейный оператор в нормированном пространстве, $\lambda \in C$, где C — множество комплексных чисел. Уравнение

$$Ax - \lambda x = 0 \text{ или } A - \lambda I \ x = 0$$

всегда имеет решение $x = 0$, которое называется тривиальным решением.

Будем говорить, что $\lambda \in C$ является *регулярным значением линейного оператора* $A : E \to E$, если оператор $A - \lambda I$ обратим на всем E для этого λ. Иначе говоря, если существует оператор $A - \lambda I \ ^{-1}$, определенный на всем пространстве E. Как было доказано, для этого необходимо и достаточно, чтобы уравнение $A - \lambda I \ x = 0$ имело только тривиальное решение.

Совокупность всех значений $\lambda \in C$, не являющихся регулярными, называется *спектром* оператора A. Отдельные составляющие спектра называются *спектральными значениями*.

Число $\lambda \in C$ называется *собственным значением* оператора A, если уравнение

$$Ax - \lambda x = 0 \text{ или } A - \lambda I \ x = 0$$

имеет нетривиальные решения. Ясно, что в этом случае обратный оператор $A - \lambda I \ ^{-1}$ не существует, т. е. если λ — собственное значение, то оно принадлежит

спектру. Однако могут существовать спектральные составляющие, не являющиеся собственными значениями. Спектр состоит из чисел двух видов:

1. Собственные значения, т. е. λ, для которых $A-\lambda I^{-1}$ не существует.
2. Числа λ, для которых $A-\lambda I^{-1}$ существует, т. е. уравнение $A-\lambda I\ x=0$ имеет только тривиальное решение, но оператор $A-\lambda I^{-1}$ определен не на всем E (и поэтому, возможно, неограничен).

Составляющие спектра, являющиеся собственными значениями, образуют так называемый *"точечный спектр"*. Числа второго типа называются *непрерывным спектром*.

В случае конечномерного линейного пространства спектр состоит только из собственных значений. Их число не более, чем размерность пространства. В бесконечномерном случае могут существовать спектральные значения, не являющиеся собственными значениями. Наличие второй составляющей спектра (непрерывного спектра) — признак бесконечномерности пространства.

Сформулируем без доказательства следующее утверждение.

Теорема 1.19. Пусть $A: E \to E$ — ограниченный линейный оператор в банаховом пространстве E и $|\lambda| > \|A\|$. Тогда λ — регулярная точка.

Отсюда следует, что спектр оператора A содержится в круге радиуса $\|A\|$ на комплексной плоскости с центром в нуле.

Замечание 2. Относительно всех рассматриваемых пространств мы везде неявно предполагали, что они содержат хотя бы один ненулевой элемент.

1.6. Дифференциальное исчисление. Производная непрерывного отображения

В теории оптимизации вводимые далее понятия производных играют первостепенную роль, т. к. позволяют, помимо всего прочего, обобщить на бесконечномерный случай многие традиционные конструкции конечномерного анализа, такие, как градиентные методы, методы Ньютона, матрицы Якоби и т. д.

Сделаем вначале общие замечания о дальнейшем изложении.

В функциональном анализе дифференциальное исчисление может рассматриваться для отображений одного аффинного пространства в другое, причем первое считается нормированным. При этом вводится специальное понятие аффинного пространства и присоединенного к нему векторного пространства. Грубо говоря, аффинное пространство — это пространство "свободных" векторов, имеющих начало и конец в любых точках. В векторном же пространстве все векторы начинаются в нуле. Нетрудно показать, что векторное пространство является частным случаем аффинного пространства. Для этого достаточно сопоставить любым двум векторам вектор их разности.

В соответствии с последним замечанием можно рассматривать (менее общую) теорию дифференциального исчисления для отображений векторного пространства в векторное же пространство. Чаще всего эти векторные пространства наделяются структурой банахова пространства. Далее мы будем следовать именно такому подходу.

В функциональном анализе рассматривают два вида дифференцируемости: *сильную*, или дифференцируемость по Фреше, и *слабую* — дифференцируемость по Гато. Первый случай соответствует понятию *полной* производной, а второй — понятию производной *по направлению* (или *частной производной вдоль вектора*). Мы далее в основном изложим лишь теорию сильной дифференцируемости, а потому вообще не будем употреблять прилагательных "сильная" и "слабая" для производных.

Пусть E и F — банаховы (действительные) пространства, $A \subset E$ — открытое подмножество в E. Пусть также $f : A \to F$, $g : A \to F$ — заданные непрерывные отображения. Отображения f и g называются *касательными* в точке $x_0 \in A$, если

$$\lim_{\substack{x \to x_0 \\ x \neq x_0}} \frac{\| f(x) - g(x) \|}{\| x - x_0 \|} = 0 .$$

(Отсюда следует, что $f(x_0) = g(x_0)$.)

Легко показать, что это отношение эквивалентности во множестве непрерывных отображений $A \to F$, т. е., в частности, если f и g касательны в точке $x_0 \in A$ и g и h касательны в точке $x_0 \in A$, то f и h касательны в точке $x_0 \in A$.

Непрерывное отображение $f : A \to F$ называется *дифференцируемым* в точке $x_0 \in A$, если существует линейное отображение $u : E \to F$, такое, что отображение $E \to F$ вида

$$x \to f(x_0) + u(x - x_0)$$

касательно к f в точке x_0. Такое линейное отображение называется *производной* (или полной производной) отображения f *в точке* x_0 и обозначается символом $f'(x_0)$ или $Df(x_0)$.

Можно доказать, что если производная существует, то она единственна. Действительно, пусть найдутся два таких отображения:

$$x \to f(x_0) + u_1(x - x_0) ,$$
$$x \to f(x_0) + u_2(x - x_0) .$$

Эти отображения по предположению касательны к $f(x)$ при $x = x_0$. (Они также будут касательными между собой.) Для линейного отображения $v = u_1 - u_2$ это означает, что

$$\lim_{\substack{y \to 0 \\ y \neq 0}} \frac{\| v(y) \|}{\| y \|} = 0 , \text{ где } y = x - x_0 . \qquad (1.4)$$

Если $u_1\ x-x_0\ \neq u_2\ x-x_0\ $, т. е. тождественного равенства нет, то для некоторого $x\neq x_0$

$$\frac{\left\|u_1\ x-x_0\ -u_2\ x-x_0\ \right\|}{\left\|x-x_0\right\|}=c\neq 0.$$

Отсюда следует, что при $\forall\varepsilon\neq 0$

$$\frac{\left\|u_1\left[\varepsilon\ x-x_0\ \right]-u_2\left[\varepsilon\ x-x_0\ \right]\right\|}{\left\|\varepsilon(x-x_0)\right\|}=\frac{\left\|v\ \varepsilon y\ \right\|}{\left\|\varepsilon y\right\|}=c\neq 0$$

и соотношение (1.4) не может быть выполнено.

Теорема 1.20. Если непрерывное отображение $f:A\to F$ дифференцируемо в точке x_0, то производная $f'\ x_0$ является непрерывным линейным отображением $E\to F$.

Доказательство опускаем.

Замечание.

1. Из вышеизложенного следует, что производная непрерывного отображения $f:A\to F$ в точке $x_0\in A$ (если она существует) является элементом банахова пространства $L\ E;F$ непрерывных линейных отображений $E\to F$, а не элементом пространства F. Основная идея дифференциального исчисления как раз и связана с *локальной аппроксимацией функций линейными функциями*. Таким образом, и в классическом анализе производная функции в точке — это не число ("тангенс угла наклона касательной"), а соответствующая линейная функция. Другое дело, что в теории вещественных функций вещественного переменного между линейными функциями и числами существует взаимно однозначное соответствие. Введенное выше общее определение производной совершенно проясняет ситуацию.
2. Опять же из вышеизложенного следует, что понятие производной можно считать введенным для любых нормированных пространств — не обязательно банаховых. Это замечание мы будем использовать в последующих примерах.
3. Производная $f'\ x_0$ является линейным отображением вида $E\to F$. Выражение $f'\ x_0\ \cdot h$ (это элемент пространства F), где $h\in E$, называется *дифференциалом* (дифференциалом Фреше) отображения f в точке x_0.

Легко видеть, что производная непрерывного линейного отображения (оператора) $u:E\to F$ существует для любой точки $x\in E$, и при этом $Du\ x\ =u\ x\ $. Действительно, в силу линейности u

$$u\ x_0\ +u\ x-x_0\ =u\ x\ .$$

Если отображение $f:A\to F$ дифференцируемо в любой точке открытого множества A, то оно называется дифференцируемым в A.

Отображение $x \to f'\ x \triangleq Df\ x$ вида $A \to L\ E; F$ обозначается символом f' или Df и называется *производной отображения f в A*. Ранее мы ввели понятие производной *в точке* с аналогичными обозначениями.

Если отображение Df непрерывно, то f называется *непрерывно-дифференцируемым* в A.

1.6.1. Формальные правила дифференцирования

Далее будет показано, что для произвольных банаховых пространств мы получаем тот же самый формальный аппарат, что и для вещественной функции вещественного переменного.

Может быть доказано следующее утверждение.

Теорема 1.21 (производная сложной функции). Пусть:

1. E, F, G — банаховы пространства.
2. $f : A \to F$, где A — открытая окрестность точки $x_0 \in E$, отображение f — непрерывно.
3. $g : B \to G$, где B — открытая окрестность точки $y_0 = f\ x_0\ \in F$, отображение g непрерывно.

Тогда если f дифференцируемо в точке x_0, а g дифференцируемо в точке $y_0 = f\ x_0\ $, то отображение

$$h = g \circ f$$

вида $h : A \to G$ (которое определено и непрерывно в некоторой окрестности точки x_0) дифференцируемо в x_0 и

$$h'\ x_0\ = g'\ y_0\ \circ f'\ x_0\ .$$

Производная композиции функций есть композиция производных.

Пример. Пусть в одномерном векторном пространстве имеем:

$$f : x \to x^2,$$

$$g : x \to \sin x,$$

$$h = g \circ f : x \to \sin\ x^2\ .$$

Тогда

$$f'\ x_0\ : x \to 2x_0 \cdot x,$$

$$g'\ y_0\ : x \to \cos y_0 \cdot x,$$

где $y_0 = f\ x_0\ = {x_0}^2$. Композиция производных $g'\ y_0\ \circ f'\ x_0$ имеет вид:

$$h'\ x_0\ x = \cos y_0 \cdot 2x_0 \cdot x\ .$$

Коэффициент $\cos y_0 \cdot 2x_0$ соответствует классическому выражению для производной. На самом же деле, как уже говорилось, производными являются линейные функции переменной *x*.

Приведем без доказательств несколько важных утверждений.

Теорема 1.22. Пусть отображения

$$f : A \to F ,$$

$$g : A \to F$$

непрерывны. ($A \subset E$ — открытое подмножество в *E*.) Если *f* и *g* дифференцируемы в точке x_0, то $f + g$ и αf (α — скаляр) дифференцируемы в точке x_0, причем

$$f + g \,'\, x_0 = f' \, x_0 + g' \, x_0 ,$$

$$\alpha f \,'\, x_0 = \alpha f' \, x_0 .$$

Теорема 1.23. Пусть $A \subset E, B \subset F$ — открытые множества, а $f : A \to B$ — гомеоморфизм. Пусть также $g : B \to A$ — обратный к *f* гомеоморфизм. Если отображение *f* дифференцируемо в точке x_0 и $f' \, x_0$ есть также гомеоморфизм вида $E \to F$, то тогда отображение *g* дифференцируемо в точке $y_0 = f \, x_0$ и $g' \, y_0$ является отображением, обратным к отображению $f' \, x_0$:

$$f^{-1} \,'\, f \, x_0 = f' \, x_0 \,^{-1} .$$

Следующее утверждение носит название *теоремы о среднем* значении. Некоторые не без оснований считают ее самой полезной теоремой анализа.

Теорема 1.24. Пусть:

1. *E*, *F* — банаховы пространства.
2. *W* — замкнутый отрезок, соединяющий точки x_0 и $x_0 + t$ пространства *E* ($W = x_0 + \vartheta t$). (Понятие замкнутого отрезка рассматривалось при изучении выпуклых множеств.)
3. Отображение $f : A \to F$ непрерывно. Здесь *A* — некоторая окрестность множества *W*.

Тогда, если *f* дифференцируемо в $\forall s \in W$, то

$$\left\| f \, x_0 + t - f \, x_0 \right\| \le \|t\| \cdot \sup_{0 \le \vartheta \le 1} \left\| f' \, x_0 + \vartheta t \right\| .$$

Классическая формулировка теоремы о среднем имеет вид

$$f \, b - f \, a = f' \, c \cdot b - a$$

или

$$f \, x_0 + t - f \, x_0 = f' \, c \cdot t .$$

Замечание. Недостаток этой классической формулировки состоит в следующем:

1. Не существует аналогичной записи в общем случае.
2. О числе *c* ничего не известно, кроме того, что оно лежит между *a* и *b*.

Между тем, для большинства приложений нужно знать только, что $f'\ c$ есть число, заключенное между нижней и верхней границами производной f' на a, b, а вовсе не то, что это и в самом деле значение производной. Иными словами, истинная природа теоремы о среднем значении выявляется, если записать ее не в виде равенства, а в виде неравенства, как это сделано ранее.

1.6.2. Частные производные

Пусть E_1, E_2 — линейные нормированные пространства. Тогда множество $E = E_1 \times E_2$ можно сделать линейным (векторным) пространством, если ввести следующие операции

$$x_1,\ x_2\ +\ y_1,\ y_2\ =\ x_1 + y_1,\ x_2 + y_2\ ,$$
$$\lambda\ x_1,\ x_2\ =\ \lambda x_1,\ \lambda x_2\ .$$

Пространство *E* можно считать нормированным, если в качестве нормы взять отображение (функционал)

$$x_1,\ x_2\ \to \sup\ \|x_1\|, \|x_2\|\ = \|\ x_1,\ x_2\ \|\ .$$

Аксиомы нормы легко проверяются. Нормированное пространство *E* называется *произведением нормированных пространств* E_1, E_2. Это определение распространяется на случай произведения любого конечного числа нормированных пространств. Если пространства E_1, E_2 — банаховы пространства (т. е. выполняется условие полноты), то легко видеть, что линейное нормированное пространство *E* будет также банаховым.

Пусть $A \subset E = E_1 \times E_2$ — открытое подмножество *E* и $f : A \to F$ — дифференцируемое в любой точке множества *A* отображение. Здесь *F*, как и раньше, является банаховым пространством. Для любой точки $a_1,\ a_2\ \in A$ рассмотрим частные отображения открытых подмножеств пространств E_1, E_2 в *F*

$$x_1 \to f\ \ x_1,\ a_2\ ,$$
$$x_2 \to f\ \ a_1,\ x_2\ .$$

Будем говорить, что *f* дифференцируемо в точке $a_1,\ a_2$ по первой переменной, если частное отображение $x_1 \to f\ \ x_1,\ a_2$ дифференцируемо в точке a_1. Производная этого отображения является элементом пространства непрерывных линей-

ных отображений $L\ E_1;\ F$ и называется *частной производной* отображения f в точке $a_1,\ a_2$ по первой переменной. Эта частная производная обозначается $D_1 f\ a_1,\ a_2$. Аналогично вводится вторая частная производная $D_2 f\ a_1,\ a_2$.

Установим (без доказательства) связь между производной (полной) отображения $f: A \to F$ и частными производными $D_1 f$ и $D_2 f$, т. е. связь между элементом пространства $L\ E_1 \times E_2;\ F$ и элементами пространств $L\ E_1;\ F$ и $L\ E_2;\ F$.

Теорема 1.25 ("о полном дифференциале"). (При старых предположениях относительно $f,\ A,\ E,\ E_1,\ E_2$.) Для того чтобы f было непрерывно дифференцируемо в A, необходимо и достаточно, чтобы:

1. f дифференцируемо в любой точке множества A и по первой и по второй переменной;
2. отображения

$$x_1,\ x_2\ \to D_1 f\ x_1,\ x_2\ ,\quad A \to L\ E_1;\ F\ ;$$
$$x_1,\ x_2\ \to D_2 f\ x_1,\ x_2\ ,\quad A \to L\ E_2;\ F$$

непрерывны в A. В этом случае производная f в каждой точке $x_1,\ x_2$ множества A существует и справедливо равенство

$$Df\ x_1,\ x_2\ \cdot\ t_1,\ t_2\ = D_1 f\ x_1,\ x_2\ \cdot t_1 + D_2 f\ x_1,\ x_2\ \cdot t_2 .$$

Здесь использованы обозначения $u \cdot t$ вместо $u\ t$.

Получили аналог известной теоремы классического анализа.

Замечание. Можно доказать, что если f дифференцируемо, то частные производные существуют и выполняется приведенное соотношение.

Теорема о полном дифференциале обобщается на произведение любого конечного числа банаховых пространств.

1.6.3. Производные функций одной переменной

Будем рассматривать отображение $f: R \to F$. Отождествим пространство $L\ R;\ F$ с пространством F.

Любому $b \in F$ соответствует элемент $L\ R;\ F$ вида $\xi \to b \cdot \xi,\quad \xi \in R$. И обратно, каждый элемент $L\ R;\ F$ является линейным отображением и имеет вид

$$\xi \to f\ \xi\ = f\ \xi \cdot 1\ = \xi f\ 1\ = \xi a ,$$

где $a = f\ 1\ ,\ a \in F$. Установленное соответствие является изометрией.

Если отображение $f: A \to F$ (где $A \subset R$ — открытое множество) дифференцируемо в точке $\xi_0 \in A$, то

$$f'\ \xi_0\ \in L\ R,\ F\ ,$$

т. е. производная также отождествляется с элементом пространства *F*. Если само $F = R$, то мы приходим к классическому понятию производной как числа и из всех вышеприведенных формул получаем соответствующие классические соотношения.

1.6.4. Матрица Якоби

1. Применим теорему о полном дифференциале к случаю $E = R^n$. Пусть отображение $f: A \to F$ дифференцируемо ($A \subset R^n$ — открытое множество). Тогда частная производная $D_k f\ x_1, ..., x_n$ существует и отождествляется с элементом пространства *F*. Поэтому можно говорить не о непрерывности отображений вида

$$A \to L\ E_i;\ F\ ,$$
$$A \to L\ A;\ F$$

(в данном случае $A \to L\ R;\ F$), а о непрерывности самих производных. Из дифференцируемости *F* следует, что частные производные существуют и выполнено равенство

$$Df\ x_1, ..., x_n\ \cdot\ t_1, ..., t_n\ = \sum_{k=1}^{n} D_k f\ x_1, ..., x_n\ \cdot t_k\ .$$

Здесь в правой части точка означает обычный знак умножения, а не обозначение $u\ t\ = u \cdot t$. Имеем в данном случае отображение вида $R^n \to F$:

$$t_1, ..., t_n\ \to \sum_{k=1}^{n} D_k f\ x_1, ..., x_n\ \cdot t_k\ .$$

2. Пусть теперь $E = R^n$, $F = R^m$. В этом случае имеем $f = \varphi_1, ..., \varphi_m$, где φ_i — скалярные функции, определенные в R^n.

Можно доказать, что для дифференцируемости *f* необходимо и достаточно дифференцируемости φ_i и в этом случае

$$f'\ x_0\ =\ \varphi_1'\ x_0\ , ..., \varphi_m'\ x_0\ ,\quad x_0 \in A \subset R^n\ .$$

Получим выражение для φ_i'. Это отображение вида $R^n \to R$ и задача сводится к случаю 1 при $F = R$. Поэтому φ_i' есть отображение вида

$$D\varphi_i\ x_1, ..., x_n\ \cdot\ t_1, ..., t_n\ = \sum_{k=1}^{n} D_k \varphi_i\ x_1, ..., x_n\ t_k\ .$$

Иными словами, производной $f'(x_0)$ (отображение $R^n \to R^m$) соответствует матрица

$$D_k\varphi_i(x_1, ..., x_n),$$

называемая *матрицей Якоби* отображения f в точке $(x_1, ..., x_n)$. Когда $m = n$, определитель этой матрицы называется *якобианом* отображения f или функций $\varphi_1, ..., \varphi_n$.

1.6.5. Производные высшего порядка

Пусть $f: A \to F$ есть непрерывно дифференцируемое отображение открытого множества A банахова пространства E в банахово пространство F. Тогда отображение $Df: A \to L(E; F)$ непрерывно (оно называется производной отображения f во множестве A). Если это отображение Df дифференцируемо в точке $x_0 \in A$ (соответственно, в любой точке x_0 множества A), то мы будем говорить, что f дважды дифференцируемо в точке x_0 (соответственно, в A), а производную отображения Df в точке x_0 будем называть *второй производной* отображения f в точке x_0. Вторая производная обозначается $f''(x_0)$ или $D^2 f(x_0)$. Эта производная является элементом пространства $L(E; L(E; F))$:

$$D^2 f: A \to L(E; L(E; F)).$$

Введем некоторые вспомогательные понятия.

Отображение

$$u: E_1 \times E_2 \times ... \times E_n \to F,$$

где E_i, F — банаховы пространства, называется *полилинейным*, если для любого целого $k = 1, ..., n$ и любой системы элементов $a_i \in E_i$ $(i \neq k)$ "частное" отображение

$$x_k \to u(a_1, ..., a_{k-1}, x_k, a_{k+1}, ..., a_n)$$

(т. е. отображение вида $E_k \to F$) линейно. Другими словами, если зафиксировать все переменные кроме одной, то отображение u линейно зависит от оставшейся переменной.

Обозначим через $L(E_1, ..., E_n; F)$ пространство всех непрерывных полилинейных отображений вида $E_1 \times E_2 \times ... \times E_n \to F$. В пространстве $L(E_1, ..., E_n; F)$ можно ввести норму. Именно, $\|u\|$ называется нормой полилинейного отображения u, если $\|u\| = \min M$, таких, что

$$\|u(x_1, ..., x_n)\| \le M\|x_1\| \cdot ... \cdot \|x_n\|.$$

Можно показать, что если F — банахово пространство, то и нормированное пространство L полно, т. е. тоже является банаховым пространством.

Между банаховыми пространствами $L\ E_1, E_2; F$ и $L\ E_1; L\ E_2; F$ можно построить биекцию, которая является изометрией. Именно каждому $u \in L\ E_1; L\ E_2; F$ ставится в соответствие элемент из $L\ E_1, E_2; F$ вида $s, t \to u, s \cdot t \in F$, или $s, t \to u \cdot s, t$. Здесь $s \in E_1, t \in E_2$.

Пояснение. $u \cdot s \triangleq u\ s$ — элемент пространства $L\ E_2; F$. Применив его к $t \in E_2$, получим элемент пространства F.

Таким образом, т. к. $D^2 f\ x_0 \in L\ E; L\ E; F$, то в соответствии с только что сказанным имеем $D^2 f\ x_0 \in L\ E, E; F \triangleq L_2\ E; F$. Следовательно, производную можно считать некоторым непрерывным билинейным отображением вида $E \times E \to F$ или, более точно,

$$s, t \to D^2 f\ x_0 \cdot s \cdot t \in F.$$

Определим теперь по индукции "p" раз дифференцируемое отображение $f: A \to F$ как "$p-1$" раз дифференцируемое отображение, $p-1$ -я производная $D^{p-1} f$ которого дифференцируема в A. Производная $D\ D^{p-1} f$ называется p-й производной отображения f и обозначается $D^p f$ или $f^{\ p}$. Производная $D^p f\ x_0$ отождествляется с элементом пространства

$$L_p\ E; F \triangleq L\ E, ..., E; F,$$

т. е. является отображением вида

$$t_1, ..., t_p \to D^p f\ x_0 \cdot t_1, ..., t_p.$$

Случай, когда $E = E_1 \times E_2$. Имеем формулу

$$Df\ x_1, x_2 \cdot t_1, t_2 = D_1 f\ x_1, x_2\ t_1 + D_2 f\ x_1, x_2\ t_2.$$

Здесь:

$$f: A \to F,\ A \subset E_1 \times E_2 = E,$$
$$Df\ x_1, x_2 \in L\ E_1 \times E_2; F,$$
$$D_i f\ x_1, x_2 \in L\ E_i; F.$$

Напишем ту же формулу не для f, а для Df. Получим

$$Df: A \to L\ E; F,$$
$$Df\ x_1, x_2 \in L\ E_1 \times E_2; F,$$
$$D^2 f\ x_1, x_2 \in L\ E_1 \times E_2; L\ E_1 \times E_2; F,$$
$$D_i Df\ x_1, x_2 \in L\ E_i; L\ E_1 \times E_2; F.$$

Имеем теперь

$$D^2 f\ x_1,\ x_2\ \cdot\ t_1,\ t_2\ = D_1 Df\ x_1,\ x_2\ t_1 + D_2 Df\ x_1,\ x_2\ t_2\ .$$

Здесь левая и правая части равенства являются элементами пространства $L\ E_1 \times E_2;\ F\ $. Отсюда получаем окончательно

$$D^2 f\ x_1,\ x_2\ \cdot\ t_1,\ t_2\ \cdot\ h_1,\ h_2\ = (D_1 Df\ x_1,\ x_2\ t_1 + D_2 Df\ x_1,\ x_2\ t_2)\cdot\ h_1,\ h_2\ .$$

Это же выражение обобщается на случай $E = E_1 \times ... \times E_n$:

$$D^2 f\ x_1,\ ...,\ x_n\ \cdot\ t_1,\ ...,\ t_n\ \cdot\ h_1,\ ...,\ h_n\ = \sum_{k=1}^{n} D_k Df\ x_1,\ ...,\ x_n\ \cdot t_k\ \cdot\ h_1,\ ...,\ h_n\ .$$

1.6.6. Формула Тейлора

Пусть $f: A \to F$ — "p" раз непрерывно дифференцируемое отображение открытого множества A банахова пространства E в банахово пространство F. Тогда если замкнутый отрезок, соединяющий точки x и $x+h,$ содержится в A, то для любого $\varepsilon > 0$ существует $r > 0$ такое, что при $\|h\| \le r$ имеем соотношение (формулу Тейлора для отображений):

$$\left\| f\ x+h\ - f\ x\ - f'\ x\ \cdot h - ... - \frac{1}{n!} f^{\ n}\ \ x\ \cdot\ h^{\ n} \right\| \le \varepsilon \|h\|^n ,$$

где $h^{\ n}$ означает $h, ..., h\ $ — n раз. Эту же формулу можно записать в более привычном виде:

$$f\ x+h\ = f\ x\ + f'\ x\ \cdot h + ... + \frac{1}{n!} f^{\ n}\ \ x\ \cdot\ h^{\ n} + \omega\ x,\ h\ .$$

Для любого $\varepsilon > 0$ существует $r > 0$ такое, что при $\|h\| \le r$ имеем $\|\omega\ x,\ h\ \| \le \varepsilon \|h\|^n$. Последнее условие часто записывают в виде $\|\omega\ x,\ h\ \| = o\ \|h\|^n\ ,\ \|h\| \to 0$.

1.7. Необходимые условия экстремума

1.7.1. Производная и градиент функционала

Пусть $f: A \to R,\ A \subset E$ — нелинейный функционал, заданный на подмножестве A банахова пространства E. Допустим, что отображение f дифференцируемо в точке $x \in A$, т. е. существует нужное линейное отображение вида $E \to R$ (линейный функционал). Тогда производная этого отображения в точке $x \in A$ (в случае функционалов) называется *градиентом функционала f* в точке x и обозначается

не $Df\ x$, как обычно, а $\operatorname{grad} f\ x$. $f\ x$ называется *потенциалом* оператора $F\ x = \operatorname{grad} f\ x$.

Замечание. В случае $E = R^n$, $F = R^m$ (у нас сейчас $m = 1$) мы уже показали, что производной отображения $f: A \to R^m$ соответствует матрица Якоби размером $n \times m$. В нашем случае $m = 1$ и матрица Якоби вырождается в вектор с компонентами $D_k f\ x_1, ..., x_n$, $k = 1, ..., n$. В классическом анализе этот вектор и называется градиентом. С общих позиций функционального анализа — это не градиент, а некая конструкция, связанная с градиентом и позволяющая вычислять градиент в любой точке.

Еще раз подчеркнем, что матрица Якоби, градиент — это по сути дела объекты одного рода, т. е. производные. Именно поэтому они фигурируют в формулировках различных теорем теории оптимизации.

1.7.2. Теоремы о существовании и единственности минимума функционала

Пусть $f: A \to R$, $A \subset E$ — функционал, заданный на открытом подмножестве A банахова пространства E. Будем говорить, что функционал f имеет *локальный (относительный) минимум* в точке $x_0 \in A$, если существует окрестность $U \subset A$ точки x_0, в которой $\forall x \in U: f\ x \geq f\ x_0$. При выполнении строгого неравенства $f\ x > f\ x_0$ локальный минимум называется *строгим локальным минимумом*. Если $\forall x \in A: f\ x \geq f\ x_0$, то локальный минимум называется *глобальным (абсолютным) минимумом*.

Аналогичные определения вводятся для максимумов. Для того чтобы функционал f в точке $x_0 \in A$ имел *локальный максимум* (соответственно, *строгий локальный максимум*, *глобальный максимум*), необходимо и достаточно, чтобы функционал $-f$ имел в точке $x_0 \in A$ локальный минимум (соответственно, строгий локальный минимум, глобальный минимум).

Теорема 1.26. Если функционал $f: A \to R$, $A \subset E$ имеет локальный минимум в точке $x_0 \in A$ и дифференцируем в этой точке, то $f'\ x_0 = 0$, или, в новых обозначениях,

$$\operatorname{grad} f\ x = 0.$$

Доказательство. Пусть $h \in E$ — произвольный вектор. Рассмотрим функционал

$$g\ t = f\ x_0 + th \qquad g: R \to R$$

для $|t| < \varepsilon$, $t \in R$. Тогда функционал g имеет локальный минимум при $t = 0$. Следовательно, в соответствии с результатами классического анализа, $g'\ 0 = 0$.

Так как $g'(t) = f'(x_0 + th) \cdot h$, то для $\forall h \in E: f'(x_0) \cdot h = 0$. Другими словами, отображение $f'(x_0): E \to R$ является нулевым. Теорема доказана.

Точки, в которых выполняется равенство нулю градиента, называются *стационарными*. Точки, в которых достигается локальный максимум или минимум, называются *экстремальными*.

Пояснение. Дадим из методических соображений дополнительные подробные пояснения по получению соотношения $g'(t) = f'(x_0 + th) \cdot h$. Рассмотрим два отображения

$$\varphi: t \to x_0 + th \qquad R \to E,$$
$$f: x \to f(x) \qquad E \to R.$$

Тогда отображение $g(t) = f(x_0 + th)$ типа $R \to R$ будет композицией $g = f \circ \varphi$. Согласно общей теореме о производной композиции имеем

$$g'(t) = f'(\varphi(t)) \circ \varphi'(t).$$

Надо найти $\varphi'(t)$ и $f'(\varphi(t))$. Получим необходимые соотношения.

Отображение $\varphi'(t)$ для $\forall t$ является линейным отображением вида $R \to E$ (как и отображение φ). Покажем, что это отображение имеет вид

$$\varphi'(t): s \to sh, \quad \forall t \in R,\ s \in R.$$

Для этого, пользуясь определением производной, построим отображение вида

$$s \to \varphi(t + h(s - t)) = x_0 + th + hs - ht = x_0 + hs.$$

То есть это отображение вида $s \to x_0 + hs$. Оно совпадает с самим отображением φ и поэтому касательно к нему. Имеем

$$f'(\varphi(t)) = f'(x_0 + th).$$

Используя теорему о композиции, получим ($g'(t): R \to R$):

$$\forall s \in R: g'(t) \cdot s = f'(x_0 + th) \circ \varphi'(t) \cdot s = f'(x_0 + th) \cdot h \cdot s.$$

В правой части кружок перед h заменили точкой, т. к. все отображения линейны. Для $t = 0$ получим

$$g'(0) \cdot s = f'(x_0) \cdot h \cdot s.$$

Так как отображение $g'(0): R \to R$ является нулевым, то отображение

$$f'(x_0) \cdot h \cdot s: R \to R$$

также будет нулевым для $\forall h, s$. Положив $s = 1$, получим, что отображение $f'(x_0): E \to R$ является нулевым для $\forall h \in E$ $f'(x_0): h \to f'(x_0) \cdot h$. Что и требовалось доказать.

Пример. Пусть

$$F\ x\ = \int_a^b f\ t, x\ t\ \ dt,$$

где f — непрерывно дифференцируемая (в классическом смысле) функция, $x\ t\ \in C\ a, b\ .$ Здесь $C\ a, b\ $ — пространство непрерывных на $a, b\ $ функций. Это нормированное, но не банахово пространство.

Вычислим производную функционала F.

Согласно общему определению производной отображения $f: E \to F$ имеем следующее. Если f непрерывна и дифференцируема в точке $x \in A \subset E$, то существует линейное отображение $u_x : E \to F$, которое обладает свойством

$$\forall h \in E: \ \frac{\left\| f\ x+h\ - f\ x\ - u_x \cdot h \right\|}{\left\| h \right\|} \xrightarrow[\substack{\left\| h \right\| \to 0 \\ h \neq 0}]{} 0$$

или

$$f\ x+h\ - f\ x\ = u_x \cdot h + \mathrm{o}_{\substack{x \\ \left\| h \right\| \to 0}} \left\| h \right\| \ .$$

Для функционала F имеем при $\forall h \in C\ a, b$

$$F\ x+h\ - F\ x\ = \int_a^b \left[f\ t, x+h\ - f\ t, x\ \right] dt = \int_a^b f'_x\ t, x\ h\ t\ \ dt + \mathrm{o}\ h\ .$$

Следовательно, дифференциал F в точке x равен

$$dF = \int_a^b f'_x\ t, x\ h\ t\ \ dt.$$

Равенство нулю этого дифференциала для $\forall h \in C\ a, b\ $ есть, как было доказано, необходимое условие экстремума, а для этого необходимо и достаточно выполнения равенства

$$f'_x\ t, x\ = 0.$$

Это уравнение определяет функцию $x\ t\ $, на которой функционал F может достигать экстремума. Более точных утверждений мы здесь не можем делать.

1.7.3. Уравнение Эйлера

Исследуем на экстремум функционал

$$F\ x\ = \int_a^b f\ t, x\ t\ , x'\ t\ \ dt,$$

определенный в пространстве $C^1\ a, b$ непрерывно дифференцируемых на отрезке a, b функций. Здесь

$$x'\ t\ = \frac{dx\ t}{dt},\ f\ t, x, x'$$

есть дважды дифференцируемые функции своих аргументов. Воспользуемся теоремой о необходимых условиях экстремума. Именно, если функция $x\ t$ доставляет функционалу $F\ x$ экстремум, то $F'\ x\ = 0$ или

$$\forall h\ t\ \in C^1\ a,b\ : F'\ x\ \cdot h\ t\ = 0\,.$$

Выберем $\forall h\ t\ \in C^1\ a, b$ и найдем дифференциал функционала $F\ x$:

$$F\ x+h\ - F\ x\ = \int_a^b \left[f\ t, x+h, x'+h'\ - f\ t, x, x'\ \right] dt =$$

$$= \int_a^b f'_x \cdot h + f'_{x'} \cdot h'\ dt + \underset{\|h\|\to 0}{\mathrm{o}}\ \|h\|\ = dF + \underset{\|h\|\to 0}{\mathrm{o}}\ \|h\|\ .$$

Таким образом, необходимое условие имеет вид

$$\int_a^b f'_x \cdot h + f'_{x'} \cdot h'\ dt = 0\,.$$

После стандартных преобразований получаем для необходимого условия

$$\int_a^b f'_x \cdot h + f'_{x'} \cdot h'\ dt = \int_a^b f'_x \cdot h\, dt + \int_a^b f'_{x'}\, dh = \int_a^b f'_x \cdot h\, dt + \ f'_{x'} \cdot h\ \Big|_a^b - \int_a^b h \cdot \frac{df'_{x'}}{dt} dt =$$

$$= \int_a^b \left(f'_x - \frac{d}{dt} f'_{x'} \right) \cdot h\ t\ dt + \ f'_{x'} \cdot h\ \Big|_a^b = 0.$$

Последнее равенство должно выполняться для $\forall h\ t\ \in C^1\ a, b$, в том числе и для h со свойством $h\ a\ = h\ b\ = 0$. Для всех таких h имеем

$$\int_a^b \left(f'_x - \frac{d}{dt} f'_{x'} \right) \cdot h\ t\ dt = 0\,.$$

Отсюда следует, что необходимо выполнение равенства

$$f'_x - \frac{d}{dt} f'_{x'} = 0\,.$$

Последнее выражение уже не зависит от h и, следовательно, должно выполняться для любых h. Полученное уравнение называется *уравнением Эйлера* и рассматрива-

ется в курсах вариационного вычисления. Здесь оно было элементарно выведено исходя из общих условий экстремума.

Из общего необходимого условия следует, что помимо уравнения Эйлера для $\forall h\ t \in C^1\ a, b$ должно выполняться равенство

$$f'_{x'} \cdot h \Big|_a^b = 0.$$

Выбрав h вида $h\ a = 0,\ h\ b \neq 0$, получим

$$f'_{x'}\ _{t=b} = 0.$$

Это равенство не зависит от h и поэтому выполняется для любого h. Точно так же, полагаем $h\ a \neq 0,\ h\ b = 0$ и получаем аналогично

$$f'_{x'}\ _{t=a} = 0.$$

Последние два равенства называются *естественными граничными условиями*. Таким образом, для выполнения необходимого условия экстремума функционала $F\ x$ необходимо и достаточно, чтобы искомая функция $x\ t$ удовлетворяла уравнению Эйлера и естественным граничным условиям. Уравнение Эйлера является уравнением второго порядка, и его общее решение содержит две произвольные постоянные. Для их определения в нашем распоряжении как раз оказывается нужное количество граничных условий.

Замечание.

1. Заранее мы не можем быть уверенными в том, что функция $x\ t$, удовлетворяющая необходимым условиям (уравнению Эйлера и естественным граничным условиям), действительно существует.
2. Если даже можно утверждать существование такой функции $x\ t$, то ничего в общем случае не доказывает ее единственности.
3. Вообще говоря, теорема о необходимом условии экстремума была сформулирована и доказана для функционала, определенного в открытом множестве A. Однако часто множество A бывает не открытым, а замкнутым. В этом случае, если функционал и имеет экстремум (например, локальный максимум или минимум), то он может не удовлетворять теореме о необходимом условии экстремума, которая применима только к открытым множествам. Экстремальные значения могут достигаться на границе замкнутого множества A. Кроме того, в экстремальной точке функционал может не иметь производной, и общая теорема снова неприменима.
4. Если множество A открыто, производные существуют и для какой-то функции выполнены уравнение Эйлера и естественные граничные условия, то все равно, это только необходимые условия. Соответствующая функция $x\ t$ может соответствовать, например, седловой точке (точке перегиба), а не максимуму или минимуму. Для установления характера такой точки (в данном случае — функции $x\ t$) необходим дополнительный анализ на основе разложения Тейлора.

Таким образом, точками подозрительными на экстремум, являются все точки, удовлетворяющие уравнению Эйлера и граничным условиям, а также те точки, к которым неприменима теорема о необходимых условиях экстремума. Специалистам по оптимальному управлению известно, что для анализа на экстремум функционалов, определенных в замкнутых множествах, в ряде случаев целесообразно использовать формализмы на основе так называемого *принципа максимума* или принципа оптимальности динамического программирования.

1.8. Достаточные условия экстремума

1.8.1. Однородные полиномы

Пусть E и F — банаховы пространства и $n \geq 1$ — целое число. Отображение $\varphi : E \to F$ называется *однородным полиномиальным отображением* степени *n*, если существует такое полилинейное (*n*-линейное) отображение ("полярное" к φ)

$$f : E \times E \times ... \times E \to F \text{ (или } E^n \to F \text{)},$$

что $\varphi\ x = f\ x, ..., x$. В этом случае говорят, что φ есть *однородный полином* степени *n*. Однородный полином второй степени называется *квадратичной формой*, если $F = R$, т. е. если φ — функционал.

В общем случае под *полилинейным отображением* понимается отображение вида

$$f : E_1 \times E_2 \times ... \times E_n \to F$$

(здесь E_i, F — линейные векторные пространства), линейное по всем переменным. Точнее говоря, если зафиксировать все переменные кроме одной, то отображение *f* будет линейно зависеть от оставшейся свободной переменной. Понятие линейного отображения (или линейного оператора) было введено ранее. Из определения полилинейного отображения следует, что $f\ x_1, x_2, ..., x_n = 0$, если в нуль обращается хотя бы одна из переменных x_i.

Можно доказать, что если φ есть однородный полином степени *n*, то существует не только полилинейное отображение $\varphi\ x = f\ x, ..., x$, но и единственное *симметрическое* отображение

$$g : E \times E \times ... \times E \to F,$$

$$\varphi\ x = g\ x, ..., x\ ,$$

для которого выполняется свойство симметричности:

$$\forall i, j : g\ x_1, ..., x_i, ..., x_j, ..., x_n = g\ x_1, ..., x_j, ..., x_i, ..., x_n\ .$$

Еще раз подчеркнем, что симметрическое отображение *g* устанавливает между элементами множеств E^n и F то же самое соответствие, что и φ.

Таким образом, каждой квадратичной форме $\varphi: E \to R$ соответствует билинейное симметрическое отображение $\varphi_B: E \times E \to R$ такое, что $\varphi\ x\ = \varphi_B\ x, x\ .$

Примером билинейной симметрической формы является скалярное произведение в евклидовом пространстве.

Квадратичная форма называется *положительно определенной (или просто положительной)*, если

$$\forall x \in E: \varphi\ x\ \geq 0\ .$$

Билинейную форму также будем называть положительно определенной, если

$$\forall x \in E: \varphi_B\ x, x\ \geq 0\ .$$

Пример. Пусть $E = R^n$. Тогда любой элемент из E есть набор из n вещественных чисел $t_1, ..., t_n$. Примером положительно определенной квадратичной формы является квадратичная форма, которой соответствует следующее отображение $\varphi_B: R^n \times R^n \to R$:

$$\varphi_B:\ x, y\ \to x_1 y_1 + x_2 y_2 + ... + x_n y_n\ ,$$

т. е.

$$x_1, ..., x_n\ ,\ x_1, ..., x_n\ \to x_1^2 + ... + x_n^2\ .$$

Переходим непосредственно к формулировке достаточных условий экстремума.

Как мы уже видели, второй производной отображения $f: A \to F$ является отображение вида

$$D^2 f: A \to L\ E; L\ E; F\ \ .$$

Значение второй производной $D^2 f\ x_0$ в некоторой точке $x_0 \in A$ является элементом множества $L\ E; L\ E; F\ \ .$ Уже отмечался тот факт, что отображение $D^2 f\ x_0$ задает отображение вида $E \times E \to F$ с помощью соотношения

$$s, t\ \to\ D^2 f\ x_0\ \cdot s\ \cdot t \in F\ .$$

Можно доказать, что это отображение является симметрической билинейной формой (которой соответствует своя квадратичная форма).

Непрерывная квадратичная форма $\varphi: E \to R$ называется *невырожденной*, если отображение

$$\varphi_B \in L\ E; L\ E; R\ \ = L\ E, E^*$$

является линейным гомеоморфизмом $E \to E^*$. (Напомним, отображение $f: E_1 \to E_2$ называется гомеоморфизмом, если оно взаимно-однозначно и взаимно-непрерывно.)

Теорема 1.27. Чтобы квадратичная форма φ была невырожденной, необходимо выполнение условия

$$\forall y \in E: \varphi_B\ x,\ y\ = 0\ \Rightarrow x = 0\ .$$

Если $E = R^n$, то это условие является и достаточным.

Доказательство ясно.

Следующее утверждение является обычным обобщением известной теоремы классического анализа.

Теорема 1.28 (достаточное условие строгого минимума). Пусть $f: A \to R$ — дважды дифференцируемый функционал в точке $x_0 \in A$. Если:

1. $f'\ x_0\ = 0$;

2. $f''\ x_0$ — положительная и невырожденная квадратичная форма,

тогда функционал *f* имеет в этой точке x_0 строгий относительный (локальный) минимум.

Доказательство опускаем.

Пояснение. Как мы видели, $f''\ x_0$ есть симметрическое билинейное отображение вида

$$s,\ t\ \to\ f''\ x_0\ \cdot s\ \cdot t\ ,$$

а любой билинейной форме соответствует квадратичная форма. Вот о ней-то и идет здесь речь.

Мы привели две теоремы (необходимые и достаточные условия) для характеристики точек, в которых достигается локальный (относительный) минимум (или максимум). Теперь займемся глобальными вопросами. Здесь важную роль играют уже изученные нами понятия выпуклости для функционала. Именно, функционал *f* называется выпуклым в банаховом пространстве *E*, если

$$\forall \vartheta \in\ 0,\ 1: f\ \ \vartheta x_1 + \ 1 - \vartheta\ \ x_2\ \ \le \vartheta f\ \ x_1\ \ + \ 1 - \vartheta\ \ f\ \ x_2\ \ .$$

Если вместо знака ≤ имеем <, то функционал называется строго выпуклым.

Теорема 1.29. Если выпуклый функционал $f: E \to R$ имеет локальный минимум в некоторой точке $x_0 \in E$, то x_0 — глобальный минимум $f\ \ x\ \ .$

Доказательство. Положим $\varphi\ \ h\ \triangleq f\ \ x_0 + h\ \ - f\ \ x_0\ \ .$ Функционал φ также будет выпуклым. Действительно,

$$\varphi\ \ \vartheta x_1 + \ 1 - \vartheta\ \ x_2\ \ = f\ \ x_0 + \vartheta x_1 + \ 1 - \vartheta\ \ x_2\ \ - f\ \ x_0\ \ =$$

$$= f\ \ \vartheta\ \ x_1 + x_0\ \ + \ 1 - \vartheta\ \ \ x_2 + x_0\ \ \ - f\ \ x_0\ \ \le$$

$$\le \vartheta f\ \ x_1 + x_0\ \ + \ 1 - \vartheta\ \ f\ \ x_2 + x_0\ \ - f\ \ x_0\ \ = \vartheta\varphi\ \ x_1\ \ + \ 1 - \vartheta\ \ \varphi\ \ x_2\ \ .$$

По условию, $\varphi\ h$ имеет локальный минимум в нуле пространства, равный нулю. Это значит, что существует окрестность нуля U такая, что если $h \in U$, то $\varphi\ h\ \geq \varphi\ 0\ = 0$.

Покажем, что $\forall h \neq 0$: $\varphi\ h\ \geq 0$. То есть покажем неотрицательность φ уже для любого h, не требуя принадлежности h какой-то малой окрестности U. Отсюда будет следовать, что $f\ x\ \geq f\ x_0$ при любом x, т. е. утверждение теоремы.

Возьмем $\forall h \in E,\ h \neq 0$. Тогда при некотором малом ε и выборе λ из условия $0 < \lambda < \varepsilon < 1$ будем иметь $\lambda h \in U$ и $\varphi\ \lambda h\ \geq \varphi\ 0\ = 0$. Используя свойство выпуклости φ, получим

$$\varphi\ \lambda h\ = \varphi\ \lambda h + \ 1-\lambda\ \cdot 0\ \leq \lambda\varphi\ h\ +\ 1-\lambda\ \varphi\ 0\ = \lambda\varphi\ h$$

или

$$0 = \varphi\ 0\ \leq \varphi\ \lambda h\ \leq \lambda\varphi\ h\ .$$

Так как $\lambda > 0$, то отсюда следует $\varphi\ h\ \geq 0$. Что и требовалось доказать.

Приведем теперь теорему о существовании глобального минимума. Мы ее сформулируем для гильбертовых пространств, а не для более общего случая банаховых пространств, т. к. здесь существенно используется свойство рефлексивности, а гильбертово пространство автоматически рефлексивно. Об этом говорилось ранее.

Теорема 1.30. Пусть E — гильбертово пространство, функционал $f : E \to R$ непрерывен, $U \subset E$ — ограниченное и замкнутое подмножество. Тогда существует по крайней мере один глобальный минимум в U.

Доказательство опускаем.

1.9. Минимизирующие последовательности

Сформулируем теперь опять основную проблему оптимизации. Задано некоторое подмножество U банахова пространства E и функционал на нем $f : U \to R$. Требуется найти $x_0 \in U$, в котором функционал f принимает глобальный минимум. Элемент x_0 в этом случае называется оптимальным элементом. Сформулированная задача часто имеет решение, например, для конечного множества U. Однако в общем случае, как указывалось во *введении*, она может и не иметь решения, поэтому целесообразно рассматривать более общую постановку задачи. Именно, требуется найти минимизирующую последовательность x_n элементов из U, удовлетворяющую условию

$$\lim_{n\to\infty} f\ x_n\ = \inf_{x\in U} f\ x\ .$$

Задача будет иметь решение, если функционал ограничен снизу на множестве U допустимых элементов. Нас будет интересовать вопрос построения минимизирую-

щих последовательностей и доказательства их сходимости к точке минимума (если она существует). При изучении этого вопроса известную роль играют выпуклые функционалы, обладающие определенными дополнительными свойствами. Однако даже строго выпуклый функционал $f\ t = \exp\ t$ не имеет минимума на $U = R$. Если минимум все-таки существует, то минимизирующая последовательность может к нему не сходиться. Легко привести соответствующие примеры для невыпуклых вещественных функций вещественного переменного. Следующий пример является более тонким и иллюстрирует существоание расходящейся минимизирующей последовательности для строго выпуклого функционала.

Пример. Пусть $x = \xi_1,\ \xi_2,\ \ldots$ — произвольный вектор из вещественного ∞-мерного пространства l^2. Это сепарабельное гильбертово пространство со скалярным произведением

$$x,\ y\ = \sum_{i=1}^{\infty} x_i y_i\ .$$

Вектор **x** можно представить в виде

$$x = \sum_{k=1}^{\infty} \xi_k e_k\ ,$$

где $e_1 =\ 1,\ 0,\ 0,\ \ldots\ ,\ e_2 =\ 0,\ 1,\ 0,\ \ldots\ ,\ \ldots$ Рассмотрим функционал

$$f\ \ x\ = \sum_{k=1}^{\infty} \frac{\xi_k^2}{k^2}\ .$$

Можно доказать, что это строго выпуклый функционал. Он имеет в нуле пространства единственную точку минимума с нулевыми координатами. Последовательность

$$x_n = \sqrt{n} e_n \qquad n = 1,\ 2,\ 3,\ \ldots\ .$$

является минимизирующей, т. к.

$$f\ \ x_n\ = \frac{n}{n^2} = \frac{1}{n} \to 0 \text{ при } n \to \infty\ .$$

Однако

$$\left\| x_n \right\| = \sqrt{n} \to \infty \text{ при } n \to \infty$$

вместо того, чтобы стремиться к нулю.

В общем случае можно доказать, что если существует непрерывная строго возрастающая вещественная функция вещественного аргумента $c\ t\ ,\ \ t \geq 0,\ \ c\ \ 0\ = 0\ ,$ удовлетворяющая условию

$$f\ \ x\ - f\ \ x_0\ \geq c\ \left\| x - x_0 \right\|\ ,$$

то любая минимизирующая последовательность для такого функционала сходится к точке x_0, которая является точкой минимума.

Для конечномерных пространств можно доказать, что последнее условие будет выполнено, если функционал $f\ x$ растет вдоль любого луча, выходящего из точки x_0.

А. Н. Тихонову принадлежит следующее определение.

Определение. Задача минимизации функционала, заданного на некотором множестве векторного пространства, поставлена *корректно*, если:

1. Она разрешима (минимизатор существует).
2. Имеет единственное решение (минимизатор единственен).
3. *Любая* минимизирующая последовательность сходится к точке минимума.

Замечание. В данном случае понятие *корректности* определяет вполне конкретные математические особенности задачи и не является синонимом обиходных значений слов типа "неправильная задача", "неудачно поставленная задача", "задача, не имеющая решения" и т. п., как это иногда утверждается даже в современной учебной литературе по оптимизации и исследованию операций. Основным в приведенном определении корректной задачи является взаимоотношение минимизатора и минимизирующей последовательности, т. е. пункт 3.

Существуют также дополнительные теоремы, где сформулированы достаточные условия корректности (по Тихонову) задачи о безусловном минимуме функционалов, заданных в вещественных банаховых пространствах.

Далее займемся методами построения минимизирующих последовательностей и выяснением условий их сходимости. Мы рассмотрим три метода: метод наискорейшего спуска, метод Ритца и метод Ньютона. Эти методы будут сформулированы для общих функциональных пространств (гильбертовых или банаховых) и поэтому могут непосредственно применяться для решения бесконечномерных задач оптимизации. Их конечномерные версии хорошо известны, особенно это касается метода наискорейшего спуска и метода Ньютона. Соответствующие вопросы будут рассмотрены далее в книге.

1.10. Дифференциалы Гато. Метод наискорейшего спуска

1.10.1. Дифференциалы Гато

Пусть E, F — банаховы пространства. Согласно данному ранее общему определению, производной в точке x для непрерывного отображения

$$f: A \to F, \quad A \subset E$$

называется линейное отображение $u_x : E \to F$, для которого выполнено равенство

$$f\ x+h\ - f\ x\ = u_x \cdot h + \mathrm{o}\ h\ , \quad \forall h \in E\ .$$

Здесь $\mathrm{o}\ h$ означает величину, для которой выполнено соотношение

$$\frac{\|\mathrm{o}\ h\|}{\|h\|} \to 0, \quad \|h\| \to 0.$$

Отсюда следует, что

$$\forall h \in E: \quad f\ x+th\ - f\ x\ = u_x \cdot th + \mathrm{o}\ th\ , \quad t \in R,$$

или, деля на t,

$$\frac{f\ x+th\ - f\ x}{t} = u_x \cdot h + \frac{\mathrm{o}\ th}{t}, \qquad \frac{\mathrm{o}\ th}{t} \to 0,\ t \to 0.$$

Таким образом, существует предел

$$\lim_{t\to\infty} \frac{f\ x+th\ - f\ x}{t} = u_x \cdot h.$$

Вычислив этот предел, мы тем самым определим дифференциал в точке x от h. Такой предел называется *дифференциалом Гато* или *слабым дифференциалом*.

Если дифференциал Гато существует для $\forall h \in E$, то соответствующая функция называется *слабо дифференцируемой* в точке x.

Из слабой дифференцируемости еще не следует дифференцируемость, а обратное мы только что доказали.

Если f — функционал, то u_x является линейным функционалом и обозначается как $\mathrm{grad}\, f\ x\ $. Дифференциал $u_x \cdot h$ обозначается как $\mathrm{grad}\, f\ x\ \cdot h$ или, более симметрично,

$$\langle \mathrm{grad}\, f\ x\ ,\ h \rangle.$$

Мы уже говорили, что через $\langle \varphi,\ x \rangle$ будем обозначать значение линейного функционала $\varphi: E \to F$ на векторе $x \in E$. (Такое же обозначение иногда применяют и в случае линейных операторов, а не только функционалов.)

Таким образом, для функционала f имеем

$$\langle \mathrm{grad}\, f\ x\ ,\ h \rangle = \lim_{t\to 0} \frac{f\ x+th\ - f\ x}{t} = \left[\frac{d}{dt} f\ x+th \right]_{t=0}.$$

В квадратных скобках стоит производная в классическом смысле. Под знаком предела в числителе и в знаменателе — вещественные числа.

Пример. Пусть E — вещественное гильбертово пространство. Вычислим $\mathrm{grad}\ x,\ x\ $, где $f\ x\ =\ x,\ x\ $ — функционал, заданный скалярным произведением. Имеем

$$f\ x+th\ - f\ x\ =\ x+th, x+th\ -\ x,\ x\ =\ x,\ x\ +2\ x,\ th\ +\ th,\ th\ -\ x,\ x\ =$$

$$=2\ x,\ th\ +t^2\ h,\ h\ =2t\ x,\ h\ +t^2\ h,\ h\ .$$

Отсюда

$$\frac{f\ x+th\ -f\ x}{t}=2\ x,\ h\ +t\ h,\ h\ \underset{t\to 0}{\to}2\ x,\ h\ =\left\langle \text{grad}\ x,\ x\ ,\ h\right\rangle .$$

В правой части полученного соотношения — дифференциал.

Приведем без доказательства следующее важное утверждение.

Теорема 1.31. Пусть E — действительное гильбертово пространство. Для любого непрерывного линейного функционала f на E существует единственный элемент $x_0 \in E$ такой, что

$$f\ x\ =\ x,\ x_0\ ,x\in E .$$

Обратно, если $x_0 \in E$, то последнее равенство определяет такой непрерывный линейный функционал f, что

$$\|f\|_{E^*}=\|x_0\|_E .$$

Таким образом, равенство $f\ x\ =\ x,\ x_0$ определяет изоморфизм $f\to x_0$ между пространствами E и E^*.

Напомним, что здесь E^* есть пространство, сопряженное к гильбертову пространству E. (E^* — это пространство непрерывных линейных функционалов на E.)

В силу сформулированной теоремы, между линейными непрерывными функционалами (например, grad), определенными на гильбертовом пространстве E, и элементами E существует взаимно однозначное соответствие, поэтому, работая с гильбертовыми пространствами, часто вместо линейных непрерывных функционалов указывают те $x_0 \in E$, которые порождают данные линейные функционалы согласно равенству

$$f\ x\ =\ x,\ x_0\ ,\ \ x\in E .$$

Для последнего примера часто пишут

$$u_x=\text{grad}\ x,\ x\ =2x ,$$

где $2x \in E$, а не E^*. Чтобы вычислить значение этого линейного отображения, скажем, на векторе $h \in E$, необходимо воспользоваться соотношением $h\to\ 2x,\ h\ $.

1.10.2. Метод наискорейшего спуска

Метод наискорейшего спуска является классическим методом построения минимизирующих последовательностей в достаточно общих пространствах. Он применяется как в конечномерном, так и бесконечномерном случаях. Как далее будет показано, его практическая значимость в конечномерных задачах невелика, т. к. существуют гораздо более эффективные алгоритмы. Мы на этом подробно остановимся в следующих главах, посвященной конечномерной оптимизации. Алгоритмический

арсенал для бесконечномерных задач существенно беднее, и этот метод продолжает сохранять свою актуальность, хотя, конечно, и здесь его вычислительная эффективность обычно невысока.

Для функционала *f* имели соотношение

$$\left\langle \operatorname{grad} f\ \ x\ ,\ h\right\rangle = \left[\frac{d}{dt} f\ \ x+th\ \right]_{t=0}.$$

Нетрудно также показать, что

$$\left\langle \operatorname{grad} f\ \ x+th\ ,\ h\right\rangle = \frac{d}{dt} f\ \ x+th\ \ .$$

В последнем равенстве отсутствует условие $t=0$.

Правая часть предпоследнего выражения (дифференциал) характеризует скорость изменения функционала *f* вдоль направления, задаваемого вектором *h*. В зависимости от *h* эти скорости будут разными и их можно вычислять по указанной формуле для любых *h*, сохраняя, например, $\|h\|=1$.

Пусть на вещественном гильбертовом пространстве *H* задан дифференцируемый ограниченный снизу нелинейный функционал $f: H \to R$. Положим

$$d \triangleq \inf_{x\in H} f\ \ x\ ,\quad F\ \ x\ \triangleq \operatorname{grad} f\ \ x\ \ .$$

Здесь $F\ \ x\ $ — линейный функционал вида $H \to R$. В силу сделанных ранее замечаний, считаем $F\ \ x\ \in H$ и тогда $F\ \ x\ \cdot h = \ F\ \ x\ ,\ h\ \ $ — скалярное произведение.

Выберем произвольный элемент $x_1 \in H$ такой, что $F\ \ x_1\ \neq 0$ и некоторый ненулевой $h \in H$. Пронормируем *h* так, чтобы $\|h\| = \|F\ \ x_1\ \|$. Выясним, как нужно выбрать направление *h*, чтобы производная

$$\left[\frac{d}{dt} f\ \ x_1+th\ \right]_{t=0} = \left\langle F\ \ x_1\ ,\ h\right\rangle = \ F\ \ x_1\ ,\ h$$

имела наименьшее (алгебраически) значение, т. е. чтобы *h* было направлением наибольшего убывания $f\ \ x\ $ в точке x_1.

Согласно неравенству Коши — Буняковского

$$\left|\ x,\ y\ \right| \le \|x\| \cdot \|y\|,\quad \|x\| = \sqrt{\ x,\ x\ }\ ,$$

имеем

$$\left|\ F\ \ x_1\ ,\ h\ \right| \le \|F\ \ x_1\ \| \cdot \|h\| = \|F\ \ x_1\ \|^2 .$$

Из последнего неравенства следует

$$-\|F\ \ x_1\ \|^2 \le \ F\ \ x_1\ ,\ h\ \le \|F\ \ x_1\ \|^2 .$$

Наименьшее значение, очевидно, достигается при $h = -F\ x_1\ $. Положим $h_1 = -F\ x_1\ $. Это направление наибольшего убывания функционала $f\ x\ $ в точке x_1.

Рассмотрим вещественную функцию φ вещественного переменного *t*

$$\varphi\ t\ = f\ x_1 + th_1\ ,\ t \geq 0\,.$$

Пусть t_1 — наименьшее положительное значение *t*, для которого $\varphi\ t_1\ = \min\limits_{t \geq 0} \varphi\ t\ $. Положим

$$x_2 = x_1 + t_1 h_1 = x_1 - t_1 F\ x_1\ .$$

По построению

$$f\ x_2\ = f\ x_1 + t_1 h_1\ < f\ x_1\ .$$

Последнее неравенство будет строгим, т. к. $F\ x_1\ \neq 0$. Предполагая, что $F\ x_2\ \neq 0$, можно продолжить процесс построения точек x_k. Таким образом, если $F\ x_k\ \neq 0$, приходим к следующей последовательности

$$x_{n+1} = x_n - t_n F\ x_n\ ,\quad n = 1,\ 2,\ ...$$

Полученная формула определяет известный *метод наискорейшего (градиентного) спуска (МНС)*. Последовательность, построенная с помощью МНС, будет обладать свойством *релаксационности*, т. к.

$$f\ x_{n+1}\ < f\ x_n\ .$$

Однако эта последовательность не обязана быть минимизирующей (в смысле сходимости к точке минимума). Множители t_n в МНС называются *релаксационными множителями*. Их выбор на практике может представлять определенные трудности. Поэтому при проведении реальных вычислений релаксационные множители заменяются другими положительными числами, выбираемыми априорно либо вычисляемыми из каких-то иных соображений. В результате мы приходим к методам *типа градиентного спуска*. Простейший подход связан с выбором $t_n = \lambda = \text{const}$. Соответствующий метод часто называется методом *простого градиентного спуска*.

Замечание 1. Методы градиентного спуска при незначительном усложнении рассуждений и обозначений могут быть обобщены на рефлексивные банаховы пространства.

Возвращаемся к гильбертовым пространствам. Докажем сходимость метода простого градиентного спуска при выполнении некоторых предположений. Тем самым будет доказана сходимость соответствующей минимизирующей последовательности.

Предварительно введем некоторые вспомогательные понятия и докажем необходимые утверждения.

Докажем равенство

$$f\ x+y\ = f\ x\ + \int_0^1 F\ x+\vartheta y\ ,\ y\ d\vartheta\ .$$

Имеем

$$F\ x\ ,\ y\ = \left[\frac{d}{dt} f\ x+ty\ \right]_{t=0},$$

или

$$F\ x+ty\ ,\ y\ = \frac{d}{dt} f\ x+ty\ .$$

Интегрируя по t от 0 до 1, получаем

$$\int_0^1 F\ x+ty\ ,\ y\ dt = \int_0^1 \frac{d}{dt} f\ x+ty\ dt = \left[f\ x+ty\ \right]_0^1 = f\ x+y\ - f\ x\ ,$$

что и требовалось доказать. Это равенство будет использовано далее при доказательстве теоремы о сходимости метода простого градиентного спуска.

Определение. Функционал $f\ x\ $, заданный в нормированном пространстве E, называется возрастающим, если

$$\lim_{\|x\|\to\infty} f\ x\ = +\infty .$$

Сформулируем без доказательства следующее утверждение.

Теорема 1.32. Возрастающий строго выпуклый функционал $f\ x\ $, заданный в гильбертовом пространстве H, имеет абсолютный минимум в некоторой точке x_0. Других точек локального минимума нет.

Основное утверждение выглядит следующим образом.

Теорема 1.33. Пусть функционал $f\ x\ $ дифференцируем в гильбертовом пространстве H, его градиент $F\ x\ $ удовлетворяет условиям:

1. $\left\|F\ x+y\ - F\ x\ \right\| \le l\|y\|,\quad l = \text{const} > 0$ (условие Липшица);

2. $F\ x+y\ - F\ x\ ,\ y\ \ge l_0\|y\|^2,\quad l_0 = \text{const} > 0$ (условие строгой монотонности).

Тогда существует единственная точка абсолютного минимума x_0.

Последовательность $x_n\ $, построенная методом простого градиентного спуска с $\lambda \in\ 0,2/l\ $, является релаксационной при $F\ x_k\ \ne 0$ и сходится к x_0 при любой начальной точке $x_1\quad f\ x_1\ < +\infty\ $.

Доказательство. Для доказательства утверждения 1 достаточно доказать, что функционал $f\ x$ является строго выпуклым и возрастающим. Строгая выпуклость непосредственно следует из условия строгой монотонности. Из этого же условия следует, что функционал $f\ x$ является возрастающим. Действительно, имеем

$$f\ x\ -f\ 0\ =\int_0^1 F\ \vartheta x\ ,\ x\ d\vartheta =\ F\ 0\ ,\ x\ +\int_0^1 F\ \vartheta x\ -F\ 0\ ,\ x\ d\vartheta \geq$$

$$\geq\ F\ 0\ ,\ x\ +\int_0^1 \frac{1}{\vartheta}\vartheta^2\|x\|^2\,d\vartheta =\ F\ 0\ ,\ x\ +\frac{1}{2}l_0\|x\|^2 \geq$$

$$\geq -\|F\ 0\ \|\cdot\|x\|+\frac{1}{2}l_0\|x\|^2=\|x\|\cdot\left(\frac{1}{2}l_0\|x\|-\|F\ 0\ \|\right)\underset{\|x\|\to\infty}{\to}\infty .$$

Докажем релаксационность последовательности $x_n\ :\ f\ x_{k+1}<f\ x_k$. Имеем, при использовании условия Липшица:

$$f\ x_{k+1}\ -f\ x_k\ =\int_0^1 F\ x_k-\vartheta\lambda F\ x_k\ \ ,-\lambda F\ x_k\ \ d\vartheta =$$

$$=\left|F_\Delta \triangleq F\ x_k-\vartheta\lambda F\ x_k\ \ ,\ F_k \triangleq F\ x_k\ \right| =$$

$$=-\lambda\int_0^1 F_\Delta,F_k\ \ d\vartheta =+\lambda\int_0^1 F_k-F_\Delta-F_k\ \ d\vartheta =$$

$$=-\lambda\|F_k\|^2+\lambda\int_0^1\|F_k-F_\Delta\|\cdot\|F_k\|d\vartheta\leq-\lambda\|F_k\|^2+\lambda\int_0^1\|F_k\|\cdot l\cdot\vartheta\cdot\lambda\|F_k\|d\vartheta =$$

$$=-\lambda\|F_k\|^2+\lambda^2\|F_k\|^2 l\int_0^1\vartheta\,d\vartheta=\lambda\|F_k\|^2\left(\frac{l\lambda}{2}-1\right)<0,$$

т. к. $0<\lambda<\frac{2}{l}$, $F\ x_k\ \neq 0$. Релаксационность доказана. Попутно доказано, что $\|F\ x_k\ \|\to 0,\ k\to\infty$.

Действительно из полученного неравенства получаем

$$\|F\ x_k\ \|^2\leq\frac{f\ x_k\ -f\ x_{k+1}}{\lambda\left(1-\frac{l\lambda}{2}\right)}.$$

Так как последовательность $f\ x_k$ является невозрастающей и ограниченной снизу значением $f\ x_0$, $f\ x_{k+1}\ -f\ x_k\ \underset{k\to\infty}{\to}0$. Сама последовательность является сходящейся: $f\ x_k\ \underset{k\to\infty}{\to}f^0\in R$. Здесь f^0 — некоторое вещественное число.

Докажем теперь сходимость $x_k \to x_0$. Иначе говоря, докажем "сходимость по аргументу". "Сходимость по функционалу" $f\ x_k \underset{k\to\infty}{\to} f\ x_0$ уже доказана. Имеем, с использованием условия монотонности и необходимого условия минимума $F\ x_0 = 0$:

$$l_0 \|x_k - x_0\|^2 \le F\ x_k - F\ x_0\ , x_k - x_0\ = F\ x_k\ , x_k - x_0\ \le \|F\ x_k\ \| \cdot \|x_k - x_0\| .$$

Отсюда

$$\|F\ x_k\ \| \ge l_0 \|x_k - x_0\| .$$

И так как было показано, что $\|F\ x_k\ \| \to 0$, то $\|x_k - x_0\| \to 0$.

Из сходимости $x_k \to x_0$ следует, что $f^0 = f\ x_0$, т. к. f — непрерывный (раз он дифференцируемый). Непрерывность следует и из условия Липшица.

Теорема доказана полностью.

Замечание 2. Можно доказать, что предыдущая теорема справедлива при выборе λ_k любым способом из промежутка $\left(0, \frac{2}{l}\right)$ на каждом шаге k (точнее — из $\alpha,\ \beta \subset \left(0,\ \frac{2}{l}\right)$).

1.11. Метод Ритца

Продолжаем изучение методов построения минимизирующих последовательностей и исследование их сходимости.

Пусть H — сепарабельное гильбертово пространство. (Понятие полноты далее, по существу, не используется, и можно считать H вещественным сепарабельным нормированным пространством.) Пусть $f\ x$ — ограниченный снизу на H вещественный функционал. Основная идея метода Ритца (для построения минимизирующей последовательности) заключается в сведении задачи минимизации $f\ x$ на H к последовательности конечномерных оптимизационных задач, решения которых сходятся к решению исходной задачи. Заметим, что существование минимизатора функционала $f\ x$ здесь не утверждается. Предполагается лишь существование точной нижней границы $d = \inf f\ x\ ,\ x \in H$.

Собственно метод Ритца сводится к следующему.

Из условия сепарабельности следует, что существует счетная линейно-независимая система векторов из H, φ_1, ..., φ_n, ..., которая полна в H. По выбранной системе векторов φ_1, ..., φ_n, ... строится последовательность конечномерных пространств H_n ,

где H_n — n-мерное пространство, натянутое на векторы $\varphi_1, ..., \varphi_n, ...$. Иначе говоря, H_n состоит из всевозможных линейных комбинаций векторов $\varphi_1, ..., \varphi_n$.

Из ограниченности снизу $f\ x$ на H следует, что $f\ x$ ограничен снизу на H_n. Пусть $d_n = \inf f\ x\ ,\ x \in H_n\quad n = 1, 2, 3, ...\ .$

По построению

$$d_1 \geq d_2 \geq ... \geq d_n \geq ....$$

Допустим, что для любого n существует $x_n \in H_n$, такой, что $f\ x_n\ = d_n$. Так как $x_n \in H_n$, то

$$x_n = \sum_{k=1}^{n} \alpha_k \varphi_k, \quad \alpha_k = \alpha_k\ \ n\ .$$

В результате, определение x_n сводится к отысканию n неизвестных числовых коэффициентов $\alpha_k\ \ n\ $. Придавая n возрастающие натуральные значения, получаем последовательность приближенных решений x_n, называемых *приближениями Ритца*.

Пусть функционал $f\ x$ дифференцируем на H и $F\ x\ \triangleq \operatorname{grad} f\ x\ $. В этом случае $f\ x$ будет дифференцируем и на H_n, причем если точка $x_n \in H_n$ является точкой абсолютного минимума $f\ x$ на H_n, то необходимо $\operatorname{grad} f\ x_n\ = 0$ (согласно теореме о необходимом условии экстремума). То есть

$$\forall h \in H_n: \quad F\ x_n\ ,\ h\ = 0\,.$$

Легко видеть, что для выполнения этого соотношения необходимо и достаточно выполнение следующих равенств:

$$F\ x_n\ ,\ \varphi_i\ = 0, \quad i = 1, ..., n\,.$$

Действительно, необходимость следует из того, что в качестве h можно брать любой элемент H_n, в том числе и φ_i. С другой стороны, т. к. для любого $h \in H_n$ имеем

$$h = \sum_{k=1}^{n} \beta_k \varphi_k\,,$$

то

$$F\ x_n\ ,\ h\ = \sum_{i=1}^{n} \beta_i\ \ F\ x_n\ ,\ \varphi_i\ \ ,$$

и достаточность доказана.

Обратим теперь внимание на равенства

$$F\ x_n\ ,\ \varphi_i\ = 0, \quad i = 1, ..., n\,.$$

Так как

$$x_n = \sum_{k=1}^{n} \alpha_k \varphi_k , \quad \alpha_k = \alpha_k \quad n \quad ,$$

то последние равенства перепишутся в виде следующей основной системы уравнений:

$$\left(F\left(\sum_{k=1}^{n} \alpha_k \varphi_k \right), \varphi_i \right) = 0, \quad i = 1, \ ..., \ n \ .$$

Получили так называемую *систему Ритца* для определения неизвестных коэффициентов α_k. Эквивалентная запись системы Ритца имеет вид

$$\left[\frac{d}{dt} f \ \ x_n + t\varphi_i \ \right]_{t=0} = 0, \quad i = 1, \ ..., \ n \ .$$

Еще раз напоминаем, что система Ритца отражает лишь необходимые условия экстремума. Формулировка метода Ритца закончена.

Итак, согласно методу Ритца, вначале определяется система "базисных" элементов $\varphi_1, ..., \varphi_n, ...$ гильбертова пространства H. Затем строится последовательность систем Ритца относительно векторов $\alpha_1, \alpha_2, ..., \alpha_n$ из конечномерных пространств H_n. При определенных условиях последовательность решений

$$x_n = \sum_{k=1}^{n} \alpha_k \varphi_k , \quad x_n \in H$$

сходится при $n \to \infty$ к минимизатору $f \ \ x$ на H (если он существует).

Теоретическое исследование сформулированного метода предполагает ответы на целый ряд вопросов. Перечислим некоторые из них:

1. Разрешимы ли системы Ритца?
2. Будет ли решение системы Ритца (если оно существует) единственным?
3. Является ли это решение минимизатором $f \ \ x$ на H_n? (Ведь системы Ритца отражают лишь необходимые условия экстремума.)
4. Образуют ли приближения Ритца минимизирующую последовательность?
5. Сходятся ли минимизирующие последовательности, построенные методом Ритца, к точке абсолютного минимума $f \ \ x$ (если она существует)?

И т. д.

В качестве примера приведем без доказательства следующее утверждение.

Теорема 1.34. Пусть функционал $f \ \ x$ на H:

1. Дифференцируем и ограничен снизу на H.
2. Является выпуклым и непрерывным.
3. Является возрастающим.

Тогда:

1. Существует единственный минимизатор $f\ x$ на H.
2. Система Ритца разрешима при любом n и имеет единственное решение.
3. Приближения Ритца образуют минимизирующую последовательность.

Заметим, что о сходимости минимизирующей последовательности к точке минимума (т. е. о корректности соответствующей задачи минимизации) здесь ничего не утверждается.

Более подробное изложение теории приближений Ритца выходит за рамки данной книги.

1.11.1. Решение уравнений методом Ритца

Приведем пример практического использования метода Ритца для решения нелинейных уравнений вида

$$\Phi\ x\ =0,\quad \Phi\colon H\to H$$

вариационными методами, т. е. методами сведения к задачам минимизации функционалов. Здесь H — гильбертово пространство.

Заметим, что мы здесь немного отвлекаемся от основной темы изложения метода Ритца. А именно мы демонстрируем, как можно решать уравнения в функциональных пространствах, сводя задачу к проблеме минимизации некоторого функционала. А уж как этот функционал минимизировать — методом Ритца или как-нибудь иначе — это уже другой вопрос. Часто (но не всегда!) это делается методом Ритца.

Существуют два основных способа сведения задачи решения нелинейного уравнения к задаче минимизации. Первый способ заключается в минимизации функционала вида $f\ x\ =\left\|\Phi\ x\ \right\|_H$. Второй подход состоит в том, что по отображению Φ строится функционал f, такой, что его критические точки (в которых $\operatorname{grad} f=0$) являются корнями $\Phi\ x$ или связаны с ними.

Например, из линейной алгебры известно, что система линейных алгебраических уравнений с симметричной матрицей A, записанная в виде $Ax=b, x\in R^n$, очевидно, совпадает с необходимыми условиями экстремума квадратичного функционала

$$J\ x\ =0{,}5\ \ Ax,\ x\ -\ b,\ x\ ,\ J\colon R^n\to R\,.$$

Вместо решения исходной системы уравнений можно искать критические точки функционала $J\ x$. В функциональном анализе указанный способ непосредственно и почти дословно обобщается на системы уравнений с самосопряженными отображениями (операторами) A, заданными в гильбертовых пространствах. Мы далее так поступать не будем и несколько иначе воспользуемся вторым подходом для

приближенного решения обыкновенного дифференциального уравнения второго порядка вида

$$\frac{d}{dt}\ p\ t\ x'\ - q\ t\ x = f\ t$$

на промежутке t_0, t_1 с граничными условиями $x\ t_0\ = x\ t_1\ = 0$. Здесь p, q, f — заданные вещественные функции. Причем p, q, p', f — непрерывны, а функция p строго положительна на t_0, t_1 . Штрих означает производную в обычном смысле. Это уравнение вида $L\ x\ - f = 0$ или $\Phi\ x\ = 0$. Элементы $x\ t$ будем считать принадлежащими некоторому множеству допустимых функций D, состоящему из непрерывных на промежутке t_0, t_1 функций, для которых $x\ t_0\ = x\ t_1\ = 0$.

Замечание 1. Приведенное уравнение называется уравнением Штурма — Лиувилля. Оно является одним из основных уравнений математической физики. К данному виду можно привести любое линейное дифференциальное уравнение второго порядка. Действительно, имеем в общем случае

$$x'' + A\ t\ x' + B\ t\ x + C\ t\ = 0.$$

Из уравнения Штурма — Лиувилля получаем

$$px'' + p'x' - qx - f = 0,$$

$$x'' + \frac{p'}{p}x' - \frac{q}{p}x - \frac{f}{p} = 0.$$

Сравнивая коэффициенты, последовательно находим: p из условия $p'/p = A$, q из условия $-q/p = B$ (при известном уже p), f из условия $-f/p = C$.

Легко видеть, что введенное множество допустимых функций D есть линейное нормированное пространство с нормой, порожденной скалярным произведением

$$x,\ y\ = \int_{t_0}^{t_1} x\ t\ y\ t\ dt\ .$$

Можно показать, что D — сепарабельное евклидово пространство (но не гильбертово, полноты нет). Этого достаточно, чтобы в дальнейшем применить (если это потребуется) процедуру Ритца, т. к. в сепарабельном пространстве можно указать счетную ортонормальную систему элементов, обладающую свойством полноты.

Сведем задачу решения уравнения Штурма — Лиувилля к задаче поиска минимума некоторого функционала.

Рассмотрим функционал

$$J\ x\ = \int_{t_0}^{t_1} p\ x'\ ^2 + qx^2 + 2fx\ dt\ .$$

Построим уравнение Эйлера в качестве необходимых условий экстремума для этого функционала. В общем виде для

$$F\ x\ = \int_a^b f\ t,\ x,\ x'\ dt$$

имели следующий вид уравнения Эйлера:

$$f_x' - \frac{d}{dt} f_{x'}' = 0,$$

или, в нашем случае,

$$2qx + 2f - \frac{d}{dt}\ 2px'\ = 0,$$

что совпадает с уравнением Штурма — Лиувилля.

Таким образом, уравнением Эйлера для функционала J является наше уравнение, и если мы найдем минимум (локальный или глобальный), точку перегиба (седло), т. е. критическую точку функционала J, где градиент равен нулю, то решим задачу. Условие $\mathrm{grad} J\ x\ = 0$ есть необходимое и достаточное условие того, что x удовлетворяет уравнению Эйлера. Этот пример иллюстрирует тот часто встречающийся случай, когда сам минимизатор (даже если он существует) нам не нужен, а нужна именно критическая точка функционала J.

В качестве системы "базисных" элементов $\varphi_1, ..., \varphi_n, ...$ из метода Ритца можно выбрать одну из многих рекомендуемых систем, обладающих свойством полноты, например, тригонометрическую систему или систему полиномов вида

$$\varphi_k = \ t_1 - t\ \ t - t_0\ ^k, k = 1, 2, ...$$

При таком выборе автоматически обеспечивается выполнение граничных условий сформулированной задачи для уравнения Штурма — Лиувилля.

Далее можно поступить следующим образом. Решение ищется в виде

$$x_n\ t\ = \sum_{k=1}^{n} \alpha_k \varphi_k\ \ t\ .$$

Подставляя это выражение в формулу для $J\ x\ $, получим фунционал J_1 от числового вектора $\alpha_1, ..., \alpha_n$

$$J_1\ \alpha\ \triangleq J\left(\sum_{i=1}^{n} \alpha_i \varphi_i\right) =$$

$$= \int_{t_0}^{t_1} \left(p\ t\ \cdot \left(\sum_{i=1}^{n} \alpha_i \varphi_i'\ \ t\ \right)^2 + q\ t\ \cdot \left(\sum_{i=1}^{n} \alpha_i \varphi_i\ \ t\ \right)^2 + 2 \cdot f\ \ t\ \cdot \sum_{i=1}^{n} \alpha_i \varphi_i\ \ t\ \right) dt.$$

Используя необходимые условия экстремума для $J_1\ \alpha$ или процедуру прямого поиска минимума в конечномерном пространстве, определяются приближенные

значения вектора $\alpha_1, ..., \alpha_n$. Затем оценивается полученная точность решения основной задачи непосредственной подстановкой построенной функции времени $x_n\ t$ в исходное уравнение Штурма — Лиувилля. При необходимости можно пытаться улучшить ситуацию и повысить точность с помощью увеличения количества n использованных базисных функций. Все коэффициенты α_i в этом случае пересчитываются заново.

Замечание 2.

1. Основным в изложенном подходе является сам "принцип параметризации". Согласно этому принципу задача бесконечномерной оптимизации заменяется последовательностью конечномерных задач. Полезность такой процедуры состоит в том, что для решения конечномерных задач существует достаточно мощный арсенал методов и алгоритмов. В функциональном анализе строится теория метода Ритца с подробным анализом свойств приближений Ритца и характеристик их сходимости. Соответствующие проблемы были обозначены нами ранее. Все эти свойства выполняются асимптотически при достаточно больших n. Поэтому для оптимизатора-практика несколько неожиданно выглядят последующие рекомендации выбирать на практике $n=1$ или $n=2$. Приводятся и численные примеры с целью подтвердить эффективность метода Ритца, скажем, для $n=1$. Подобные заявления выглядят малоубедительно, если не проводить дополнительный анализ погрешности аппроксимации. Теория метода Ритца в общем случае может являться хорошим теоретическим дополнением к собственно принципу параметризации, объясняя его внутренние структурные особенности. Она имеет особенно большое самостоятельное значение для некоторых специальных классов задач механики и математической физики, поддающихся аналитическому анализу.
2. Практически к тем же вычислительным схемам, что и метод Ритца, часто приводит известный метод Галеркина и его разновидности. Метод Галеркина основан на иных идеях и применяется в основном для решения операторных уравнений, в том числе краевых задач для дифференциальных уравнений. Он не требует прямого построения оптимизационной задачи, эквивалентной исходной — в этом его особенность и, если угодно, преимущество. Для вычислителей-практиков метод Галеркина отражает все тот же принцип параметризации.

1.12. Метод Ньютона. Общая схема методов поиска минимума

1.12.1. Метод Ньютона

В гильбертовом пространстве H задан $f\ x$ — вещественный ограниченный снизу и трижды дифференцируемый функционал. Пусть $F\ x \triangleq \operatorname{grad} f\ x$ и x_0 — точка абсолютного минимума $f\ x$, если она существует. Алгоритм метода Ньютона

строит последовательность элементов пространства H в соответствии со следующими правилами. Пусть x_1 — заданный первый элемент этой последовательности ("начальное приближение"). По формуле Тейлора

$$f(x+h) = f(x) + f'(x)\cdot h + \ldots + \frac{1}{n!} f^{(n)}(x)\cdot h^{(n)} + o\left(\|h\|^n\right)_{\|h\|\to 0}$$

получим

$$f(x) = f(x_1) + f'(x_1)(x-x_1) + \frac{1}{2} f''(x_1)(x-x_1)^{(2)} + o\left(\|x-x_1\|^2\right)_{\|x-x_1\|\to 0},$$

или

$$f(x) = f(x_1) + \langle F(x_1), x-x_1\rangle + 2^{-1}\langle F'(x_1)(x-x_1), x-x_1\rangle + o\left(\|x-x_1\|^2\right).$$

(Здесь угловые скобки использованы, как обычно, для выражения значений линейных функционалов.)

Отбрасывая слагаемое $o(\ldots)$, получим квадратичный функционал

$$\varphi_1(x) = f(x_1) + \langle F(x_1), x-x_1\rangle + 2^{-1}\langle F'(x_1)(x-x_1), x-x_1\rangle.$$

Метод Ньютона состоит в построении следующей точки x_2 из необходимых условий минимума (равенство нулю градиента) квадратичной аппроксимации φ_1 исходного функционала f в точке x_1. Найдем выражение для $\operatorname{grad}\varphi_1(x)$. Имеем

$$\langle \operatorname{grad}\varphi_1(x), h\rangle = \left[\frac{d}{dt}\varphi_1(x+th)\right]_{t=0} =$$

$$= \frac{d}{dt}\left[f(x_1) + \langle F(x_1), x+th\rangle - \langle F(x_1), x_1\rangle + 2^{-1}\langle F'(x_1)(x+th-x_1), x+th-x_1\rangle\right]_{t=0} =$$

$$= \frac{d}{dt}\left[\langle F(x_1), th\rangle + 2^{-1}\langle F'(x_1)(x-x_1), th\rangle + \right.$$

$$\left. + 2^{-1}\langle F'(x_1)(th), x-x_1\rangle + 2^{-1}\langle F'(x_1)(th), th\rangle\right]_{t=0} =$$

$$= \langle F(x_1), h\rangle + 2^{-1}\langle F'(x_1)(x-x_1), h\rangle + 2^{-1}\langle F'(x_1)h, x-x_1\rangle.$$

Так как $D^2 f(x_1)h_1h_2$ — симметричная билинейная форма, то $\langle F'(x_1)h, x-x_1\rangle = \langle F'(x_1)(x-x_1), h\rangle$. Поэтому

$$\langle \operatorname{grad}\varphi_1(x), h\rangle = \langle F(x_1), h\rangle + \langle F'(x_1)(x-x_1), h\rangle = \langle F(x_1) + F'(x_1)(x-x_1), h\rangle.$$

Следовательно,

$$\operatorname{grad}\varphi_1(x) = F(x_1) + F'(x_1)(x-x_1)$$

и x_2 находится из уравнения

$$F\ x_1\ + F'\ x_1\ \ x_2 - x_1\ = 0\,.$$

В случае существования обратного оператора $\Gamma\ x_1\ = \left[F'\ x_1\ \right]^{-1}$ имеем

$$x_2 = x_1 - \Gamma\ x_1\ \ F\ x_1\ \ .$$

Далее по x_2 точно так же строится x_3 и т. д. Общая формула метода Ньютона имеет вид

$$x_{n+1} = x_n - \Gamma\ x_n\ \ F\ x_n\ \ ,\quad n = 1,\ 2,\ \ldots\,.$$

Процесс Ньютона обрывается, если $F\ x_n\ = 0$. Это будет либо точка минимума (локального или глобального), либо "точка перегиба" (седло). В последующих главах метод Ньютона будет исследован более подробно применительно к задачам конечномерной оптимизации.

Известно много предложений о сходимости метода Ньютона к решению уравнения $F\ x\ = 0$. Если $f\ x\ $ — выпуклый функционал, то необходимые условия минимума $F\ x\ = 0$ являются и достаточными. Более детальное исследование метода Ньютона в функциональных пространствах выходит за рамки нашего рассмотрения, и мы отсылаем читателя к обширной литературе на эту тему.

1.12.2. Общая схема методов минимизации

Изученные нами метод градиентного спуска и метод Ньютона относятся к общим методам минимизации (с использованием производных). Общая схема построения всех подобных методов такова. В гильбертовом или банаховом пространстве задан ограниченный снизу функционал $f : H \to R$. Требуется построить минимизирующую последовательность x_n . Определение минимизирующей последовательности было дано ранее. Обычно минимизирующая последовательность обладает свойством релаксационности $\forall m : f\ x_{m+1}\ < f\ x_m\ $. Если существует минимизатор $x_0 \in H$, может потребоваться обеспечить сходимость минимизирующей последовательности к минимизатору. Если x_m — m-я итерация, то сначала определяется направление w_m продвижения из точки x_m. Затем выбирается вещественное число ρ_m и полагается

$$x_{m+1} = x_m - \rho_m w_m\,.$$

Правила выбора w_m и ρ_m для метода градиентного спуска и метода Ньютона были даны ранее. Существуют другие принципы задания этих элементов, что позволяет говорить о целом спектре методов минимизации указанного вида. Соответствующие примеры методов минимизации в конечномерных пространствах будут рассмотрены в следующих главах. Обычно в качестве w_m выбирается "направление

спуска", т. е. такое направление, чтобы при достаточно малом положительном ρ_m выполнялось условие релаксационности $f\ x_m - \rho_m w_m\ < f\ x_m\ $.

Итак, построение каждой итерации в рассматриваемых методах разбивается на два этапа: выбор w_m и выбор ρ_m. Для поиска w_m, а иногда и ρ_m может потребоваться вычисление некоторых производных. В этой связи различают несколько семейств методов:

1. *Прямые методы* или методы *нулевого порядка* — производные не вычисляются. Чаще всего эти методы используются при решении задач в конечномерных пространствах. В качестве примера можно указать процедуры покоординатного спуска.
2. Используются первые производные от минимизируемого функционала. Например, в методе градиентного спуска используются производные при вычислении w_m. Это *методы первого порядка*.
3. *Методы второго порядка*, например, метод Ньютона. В следующих главах будут рассмотрены и иные методы второго порядка.

Замечание. Приведенная классификация достаточно условна. Например, производные могут аппроксимироваться конечными разностями и соответствующие методы становятся методами нулевого порядка.

Глава 2

Задачи конечномерной оптимизации в теории управления

Как уже указывалось, теория и методы оптимизации в данной книге будут иллюстрироваться на примере конкретной предметной области — теории управления в широком смысле. Поэтому далее даются необходимые общие сведения из теории управления и формулируются типовые задачи, приводящие к проблеме конечномерной оптимизации.

2.1. Основные понятия теории управления

Обычно в работах 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 будут интересовать понятия *управляемой системы и управления* как некие общие концепции *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. 2.1.

Объект управления рассматривается как сколь угодно сложная система, преобразующая входные управляющие воздействия $U(t)$ в выходные сигналы (траектории) $V(t)$, которые характеризуют *состояние объекта управления* в момент времени t.

Очевидно, что реальный объект управления может иметь множество входов и выходов, определяющих его функциональное взаимодействие с внешней средой. Все эти каналы связи со средой на рис. 2.1 не показаны. Изображены лишь те входы и выходы, которые являются существенными для формулировки задачи управления.

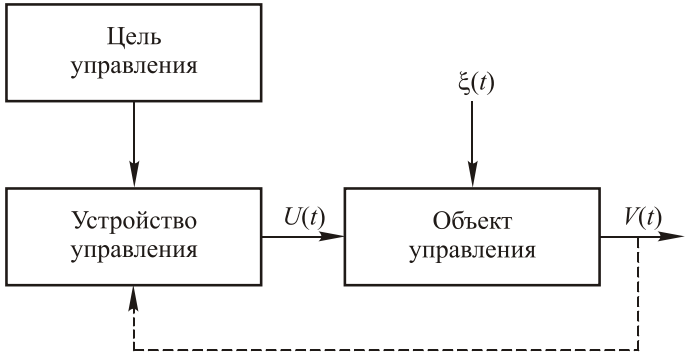


Рис. 2.1. Система управления

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. 2.1, 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. Достаточно 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. 2.1 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), связанной, например, с качеством и количеством выпускаемой продукции, а также, возможно, определенными требованиями к срокам выпуска этой продукции.

Как справедливо указывается в [59], 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 бы как можно более точно аппроксимировала 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 маховиков, изменяя ориентацию ствола по углу и по азимуту (реализуя соответствующую траекторию $V\ t$). Аналогичные задачи слежения возникают, когда антенна радиолокатора отслеживает непредвиденные движения какого-то 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 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. Автоматические системы экстремального управления называются часто *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 необходимы как минимум два *поисковых* измерения целевого функционала как основного выхода объекта. При настройке путем поиска — специальным образом организованных пробных движений — находится заранее неизвестное 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 классический, пример задачи об оптимальном в смысле расхода горючего режиме набора высоты и скорости летательным аппаратом, например, самолетом [16]. На рис. 2.2 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 $V0H$.

Существует множество траекторий в фазовом пространстве, соединяющих начальную точку (A) и конечную — *целевую* — точку (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 рас-

считана 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.

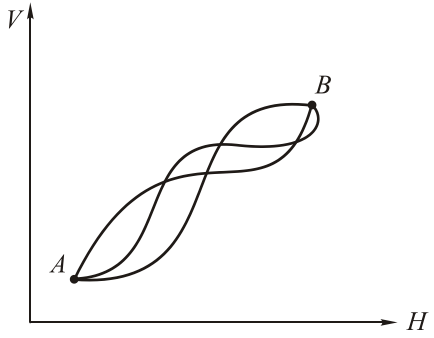


Рис. 2.2. Задача о наборе высоты и скорости самолетом

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. 2.1). Такие системы управления без обратной связи называются *разомкнутыми*. Чаще всего они применяются для целей программного управления.

Принцип жесткого управления проиллюстрирован на рис. 2.3, где представлена система жесткого управления, решающая задачу программного управления (задачу выполнения заданной программы).

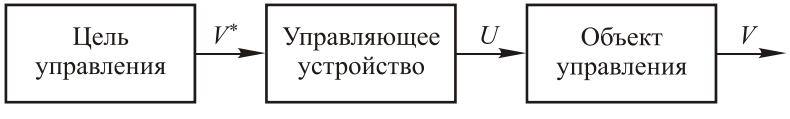


Рис. 2.3. Система жесткого программного управления

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.

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, т. е. 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. 2.4).

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.

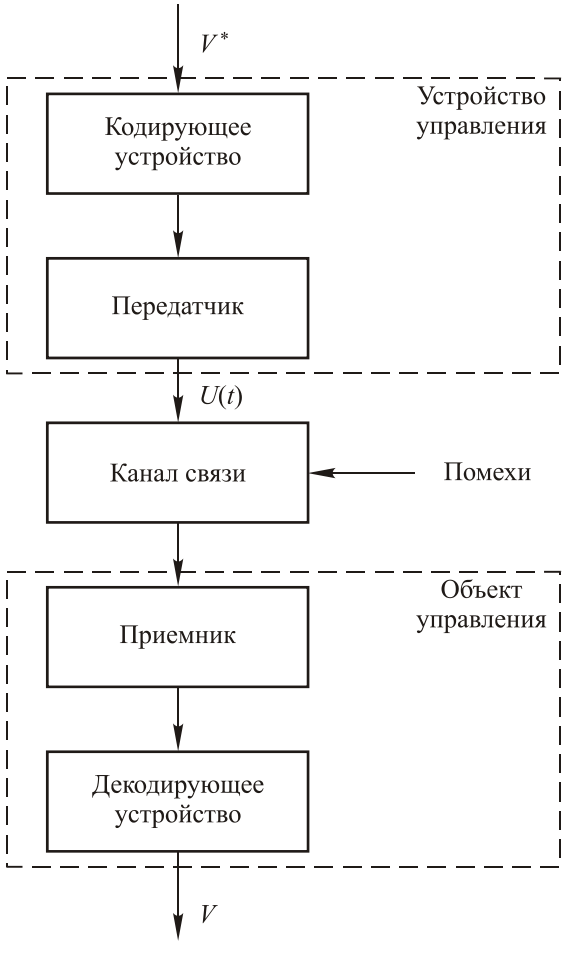


Рис. 2.4. Система радиосвязи

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 воздействий 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. 2.5.

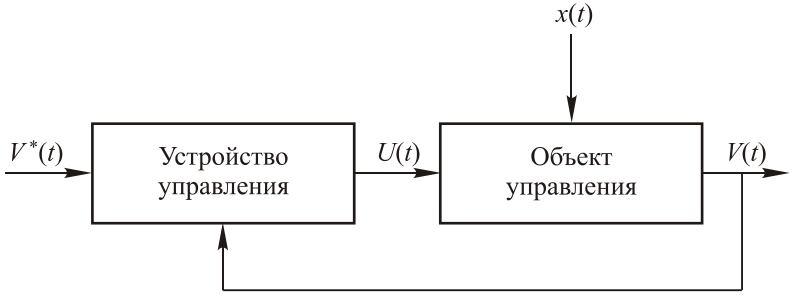


Рис. 2.5. Замкнутая система программного управления

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

$$U(t) = f(V(t) - V^*(t)).$$

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 начать 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*, т. е. 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 простой по поведению.

Схема взаимодействия объекта управления с окружающей средой представлена на рис. 2.6.

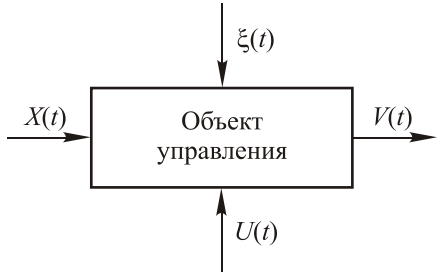


Рис. 2.6. Взаимодействие объекта со средой

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.

2.2. Система управления сложным объектом

На рис. 2.7 представлена укрупненная алгоритмическая структура современной системы управления сложным объектом [61]. На вход объекта управления поступают: вектор управляющих воздействий U, вектор возмущений ξ_1 и дополнительные *"идентифицирующие"* входные воздействия d. Измерительная система позво-

ляет 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.

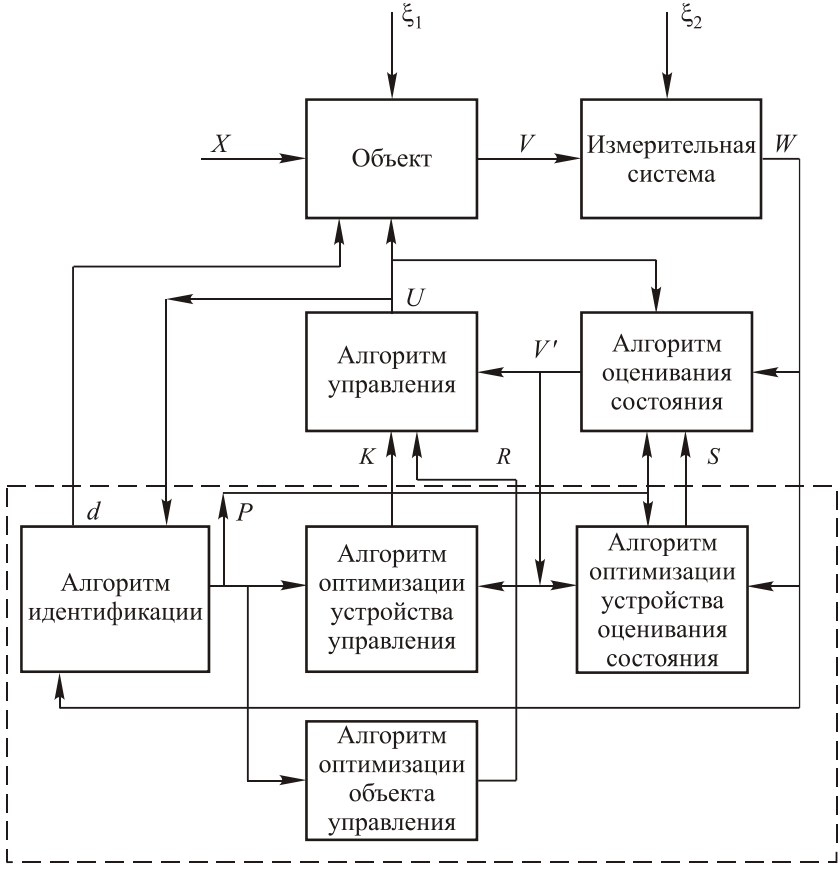


Рис. 2.7. Стратегия управления сложным объектом

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 объ-

екта 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.

Обведенная на рис. 2.7 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. 2.7 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.

2.2.1. 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, т. к. 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. 2.6):

$$V = F\ \ X, U\ \ .$$

Оператор модели, как правило, задается алгоритмически, т. е. 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 бы был 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 задана 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 вышеприведенном 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 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. 2.7.

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 [59]:

$$V = F'\ X,\ P\ .$$

Управляемый вход U объекта здесь отсутствует.

Сама идентификация, т. е. определение параметров P, осуществляется на основе информации о наблюдениях входов X и выходов V объекта в режиме нормальной эксплуатации. После получения необходимой информации о поведении объекта формируется функция невязки ψ выходов модели и объекта. Например, в простейшем случае можно положить:

$$\psi\ t,\ P\ = \sum_{i=1}^{m} \left[V_i\ t\ - V_i^{\mathrm{M}}\ t,\ P\ \right]^2.$$

Здесь через $V_i\ t$ обозначена реакция реального объекта на заданное входное воздействие $X\ t$ по i-му выходу, а через $V_i^{\mathrm{M}}\ t,\ P$ — 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^{M}. Далее задача параметрической идентификации сводится к задаче поиска минимума некоторого целевого функционала, например, вида:

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

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 мы подробнее рассмотрим важнейшие схемы параметрической идентификации и их реализации. Более подробно и полно проблема идентификации изложена в [60]. 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 по-

зволяет 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.

2.2.2. Оценивание состояний объектов управления

Алгоритм оценивания состояния (см. рис. 2.7) 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 посвящена обширная литература.

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

$$\frac{dz}{dt} = f\ t,\ z\ ,\ z\ t_0\ = x,\ t_0 \le t \le T,$$

где $\mathbf{z} = z_1, ..., z_r$ — вектор переменных состояния, f — известная функция. Предполагается, что компоненты вектора состояния z непосредственно не измеряются. Измеряемым и непосредственно наблюдаемым является другой вектор $y = y_1, ..., y_s$, связанный с z линейной зависимостью:

$$y\ t\ = H\ t\ z\ t\ ,$$

где $H\ t$ — известная прямоугольная матрица размером $s \times r$. 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:

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

Очевидно, для однократного вычисления значения функционала при заданном x необходимо вначале проинтегрировать вышеприведенную 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.

2.2.3. 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 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.

2.3. Примеры задач конечномерной оптимизации в теории управления

Под 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 *разд. 2.1* и *2.2* и как это будет продемонстрировано на примерах, приводимых далее, 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 конечномерной оптимизации далее будут характеризоваться совокупностью непрерывных параметров (предполагается, что ограничения типа дискретности отсутствуют), которые условно можно разделить на три группы: входные (управляемые) x, внешние ξ и выходные y (рис. 2.8).

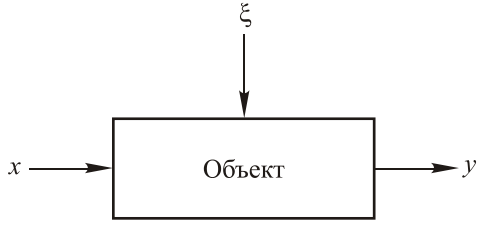


Рис. 2.8. Объект оптимизации

Входными параметрами $x = x_1, x_2, ..., x_n$ называются изменяемые в процессе оптимизации параметры, играющие роль управляющего воздействия при рассмотрении объекта оптимизации как объекта управления. *Внешние параметры* $\xi = \xi_1, \xi_2, ..., \xi_n$ 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:

- влияние случайных отклонений при установке заданных (номинальных) значений x_i на реальном объекте;
- случайные воздействия внешней среды на объект, изменяющие реальные выходные характеристики по отношению к расчетным;

- влияние изменяющихся, как правило, неслучайным, но заранее неизвестным образом, условий функционирования объекта, например таких, как температура, влажность, вибрации, уровень радиации и т. д.

Вектор *выходных параметров* $y = y_1, ..., y_m$ 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

$$y = \varphi\ x, \xi\ , \tag{2.1}$$

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 *разд. 2.2*.

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

$$y_i\ x, \xi\ \to \min_x, i \in [1:k], \quad x \in D \subset R^n, \tag{2.2}$$

$$D = \left\{ x \in R^n \,\middle|\, g_i\ x, \xi\ \le 0, i \in 1:m\ , g_j\ x, \xi\ = 0, j \in m+1:S \right\}.$$

Множество $y_1, ..., y_k$ образует множество *критериальных* выходных параметров, имеющих смысл частных критериев оптимальности и характеризующих качество объекта оптимизации. Наличие нескольких частных критериев по существу отражает ту *неопределенность цели*, которая явно или неявно присутствует при оптимизации любого сколько-нибудь сложного объекта.

Интуитивный смысл задачи (2.2) состоит в выборе такого вектора x из допустимого множества D, чтобы каждый из критериальных выходных параметров принял по возможности меньшее значение. Математическое решение задачи (2.2) в виде конкретного вектора $x^* \in D$, вообще говоря, не существует, т. к. критериальные выходные параметры отражают противоречивые (конфликтные) требования к объекту оптимизации, и минимумы соответствующих функционалов достигаются в различных точках. Это предположение правомерно, т. к. 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 (2.2) с помощью функций g_i, $i \in 1:S$.

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

$$a_i \le x_i \le b_i;\ a_i,\ b_i \in R^1. \tag{2.3}$$

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

$$a_i\ x_j\ \le x_i \le b_i\ x_j\ ;\ i \ne j. \tag{2.4}$$

Ограничения (2.3), (2.4) 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

$$y_i \le t_i;\ t_i \in R^1;\ i \in 1:S\ . \tag{2.5}$$

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

$$y_l \le t_l;\ t_l \in R^1;\ l \in 1:K \tag{2.6}$$

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 (2.5) с большим запасом не требуется; важно только гарантировать их выполнение. С другой стороны, в силу критериального характера выходных параметров y в (2.6) 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 по теории автоматического управления, часто необходимо учитывать стохастический (статистический) характер воздействий, приложенных к системе в различных точках. В некоторых пунктах этого раздела будем предполагать, что входные и выходные сигналы системы являются *случайными функциями*, изучаемыми в курсах по 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.

2.3.1. Идентификация нелинейных детерминированных объектов

В *разд. 2.3.1—2.3.3* рассматриваются некоторые варианты схем идентификации в *режиме нормальной работы* (т. е. 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:

$$H_{\mathrm{M}}\ t\ = F\left[G\ t\ ,\ x\right], \tag{2.7}$$

где $G\ t\ $ — входной сигнал модели и объекта; $H_{\mathrm{M}}\ t\ $ — выходной сигнал модели; $H\ t\ $ — выходной сигнал объекта (рис. 2.9).

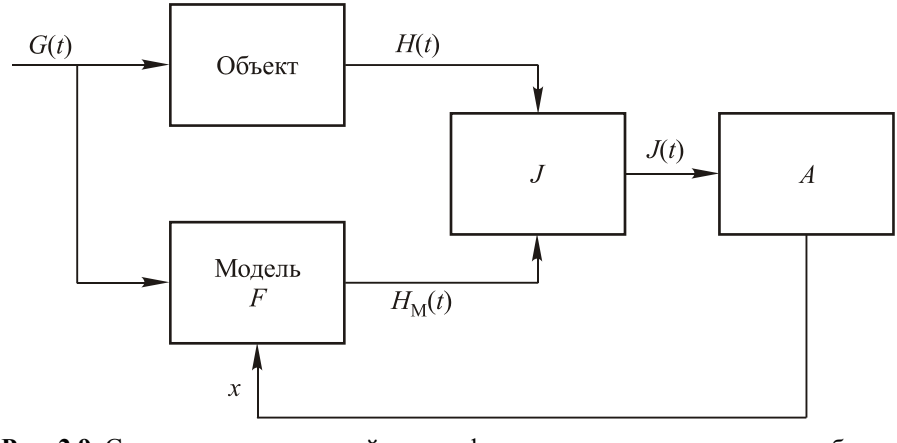


Рис. 2.9. Схема параметрической идентификации детерминированного объекта

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 *разд. 2.3.1* и *2.3.2* 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 подраздела). 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:

$$J(x) = \int_0^{\tau_1} \left\| F\left[G(t), x\right] - H(t) \right\|^2 dt. \qquad (2.8)$$

В (2.8) 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

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. 2.9).

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 [60]:

$$H(t) = \sum_{i=0}^{\infty} \int_0^{\infty} \dots \int_0^{\infty} \omega_i(\tau_1, \dots, \tau_i) \prod_{j=1}^{i} G(t - \tau_j)\, d\tau_1 \dots d\tau_i, \qquad (2.9)$$

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

$$H(t) = \int_0^T \omega_1(\tau)\, G(t - \tau)\, d\tau + \int_0^T \int_0^T \omega_2(\tau_1, \tau_2)\, G(t - \tau_1)\, G(t - \tau_2)\, d\tau_1 d\tau_2. \qquad (2.10)$$

Параметризация задачи приводит к следующим разложениям:

$$\begin{aligned}\omega_1\ t_1\ &= \sum_{\nu=1}^{N_1} c_{1\nu}\varphi_{\nu 1}\ t_1\ ,\\ \omega_2\ t_1, t_2\ &= \sum_{\nu=1}^{N_2} c_{2\nu}\varphi_{\nu 2}\ t_1, t_2\ ,\end{aligned} \tag{2.11}$$

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

Интегральная невязка выходов модели и объекта имеет вид квадратичного функционала:

$$J\ x\ = \int_0^T \left[\sum_{\nu=1}^{N_1} c_{1\nu} y_{\nu 1}\ t\ + \sum_{\nu=1}^{N_2} c_{2\nu} y_{\nu 2}\ t\ - H\ t\ \right]^2 dt, \tag{2.12}$$

где коэффициенты

$$\begin{aligned} y_{\nu 1}\ t\ &= \int_0^T \varphi_{\nu 1}\ \tau\ G\ t-\tau\ d\tau;\\ y_{\nu 2}\ t\ &= \int_0^T\int_0^T \varphi_{\nu 2}\ \tau_1, \tau_2\ G\ t-\tau_1\ G\ t-\tau_2\ d\tau_1 d\tau_2 \end{aligned}$$

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 $\varphi_{\nu i}$ необходимо учитывать, что ядра ω_i рядов Вольтерра для физически реализуемых операторов должны удовлетворять требованиям:

1. $\omega_i\ t_1, ..., t_i\ = 0,\ \forall t_i < 0,\ i \in 1:S$;

2. $\lim\limits_{t_i \to \infty} \omega_i\ t_1, ..., t_i\ = 0,\ \forall t_i < 0,\ i \in 1:S$. (2.13)

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

$$H\ n\ = \sum_{i=0}^{\infty} \sum_{m_1=0}^{\infty} ... \sum_{m_i=0}^{\infty} \omega_i\ m_1, ..., m_i\ \prod_{j=1}^{i} G\left[n - m_j \right]. \tag{2.14}$$

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

$$\begin{aligned} \omega_1\ m_1\ &= \sum_{\nu=1}^{N_1} c_{1\nu}\varphi_{\nu 1}\ m_1\ , \\ \omega_2\ m_1, m_2\ &= \sum_{\nu=1}^{N_2} c_{2\nu}\varphi_{\nu 2}\ m_1, m_2\ , \end{aligned} \tag{2.15}$$

где $\varphi_{\nu 1}$, $\varphi_{\nu 2}$ — опорная система решетчатых функций, получим выражение, аналогичное (2.12):

$$J\ x\ = \sum_{n=1}^{N_0} \left[\sum_{\nu=1}^{N_1} c_{1\nu} y_{\nu 1}\ n\ + \sum_{\nu=1}^{N_2} c_{2\nu} y_{\nu 2}\ n\ - H\ n\ \right]^2, \tag{2.16}$$

где

$$x = \ c_{11},\ ...,\ c_{1N_1};\ c_{21},\ ...,\ c_{2N_2}\ \ ;$$

$$y_{\nu 1}\ n\ = \sum_{m_1=0}^{N_0} \varphi_{\nu 1}\ m_1\ G\ n - m_1\ \ ;$$

$$y_{\nu 2}\ n\ = \sum_{m_1=0}^{N_0} \sum_{m_2=0}^{N_0} \varphi_{\nu 2}\ m_1, m_2\ G\ n - m_1\ G\ n - m_2\ .$$

Таким образом, при использовании моделей на основе ряда Вольтерра задача идентификации сводится к задаче параметрической оптимизации по квадратичным критериям типа (2.12) и (2.16). При этом ограничения на варьируемые параметры отсутствуют, и мы имеем задачу параметрической оптимизации без ограничений. Рассмотренные подходы естественным образом обобщаются на случай $l > 2$.

2.3.2. Идентификация стохастических объектов

Методы, основанные на процедурах сглаживания

Схема связи модели и объекта в рассматриваемом случае показана на рис. 2.10.

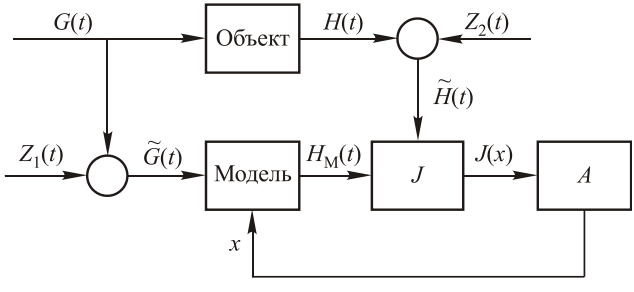


Рис. 2.10. Схема идентификации стохастического объекта

Предполагается, что стохастичность является следствием ошибок измерений входного $G\ t$ и выходного $H\ t$ сигналов детерминированного объекта. Предполагается 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 [60]. 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

$$J_1\left[\ c^{(1)}\ \right] = \sum_{j=1}^{N_1}\left[\ \sum_{i=1}^{m_1} c_i^{(1)}\varphi_i^{(1)}\ \ t_j\ \ - \tilde{G}\ \ t_j\ \right]^2 \to \min_{c^{(1)}}, \tag{2.17}$$

$$J_2\left[\ c^{(2)}\ \right] = \sum_{k=1}^{N_2}\left[\ \sum_{i=1}^{m_2} c_i^{(2)}\varphi_i^{(2)}\ \ t_k\ \ - \tilde{H}\ \ t_k\ \right]^2 \to \min_{c^{(2)}}, \tag{2.18}$$

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:

$$G\ t\ \cong \sum_{i=1}^{m_1} \hat{c}_i^{(1)}\varphi_i^{(1)}\ \ t\ = \left\langle \hat{c}^{(1)},\ \varphi^{(1)}\ \ t\ \right\rangle, \tag{2.19}$$

$$H\ t\ \cong \sum_{i=1}^{m_2} \hat{c}_i^{(2)}\varphi_i^{(2)}\ \ t\ = \left\langle \hat{c}^{(2)},\ \varphi^{(2)}\ \ t\ \right\rangle, \tag{2.20}$$

где $\hat{c}^{(1)}$, $\hat{c}^{(2)}$ — векторы, являющиеся решением задач (2.17), (2.18). Указанное приближение в среднем позволяет использовать избыточность информации для сглаживания случайных ошибок $Z_i\ t$. Как отмечается в [34], 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 независимого переменно-

го на подобласти и построения для каждой из них своих аппроксимаций (2.19), (2.20). Если $\varphi^{(i)}\ t$ — полиномы, то это может соответствовать построению кусочно-полиномиальной аппроксимации. Другие стратегии усреднения, основанные на *методе локальной аппроксимации* (МЛА), представлены в [34]. В МЛА реализована идея использования "скользящего" интервала усреднения, определяемого как некоторая окрестность текущей точки *t*. Например, для входного сигнала $G\ t$ согласно МЛА имеем:

$$\begin{aligned} & G\ t\ \cong\left\langle c\ t\ ,\ \varphi\ 0\ \right\rangle; \\ & \hat{c}\ t\ =\arg\min_{c} J_N\ t,\ c,\ \delta\ ; \\ & J_N\ t,\ c,\ \delta\ =N^{-1}\sum_{i=1}^{N}\rho\left[\frac{t-t_i}{\delta}\right]F\left[\tilde{G}\ t\ -c^T\varphi\ t-t_i\ \right], \end{aligned} \tag{2.21}$$

где $\varphi\ t$ — заданная система координатных функций; $F\ \cdot$ — неотрицательная функция потерь; $\rho\ U$ — функция локальности (например, $\exp\left[-\left|U\right|\right]$); δ — параметр усреднения, определяющий размеры локальной окрестности текущей точки *t*, в которой строится аппроксимирующая зависимость (2.21). Выбор параметра δ может быть основан на *методе перекрестного* (скользящего) *экзамена*, связанного с построением аппроксимаций, которые имеют наилучшие интерполяционные свойства для узлов, не вошедших в базовый набор $t_1,\ t_2,\ ...,\ t_N$. 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[1]

Рассмотрим простейший случай линейного объекта со стационарным эргодическим по отношению к корреляционным функциям случайным входным сигналом $G\ t$ (рис. 2.11).

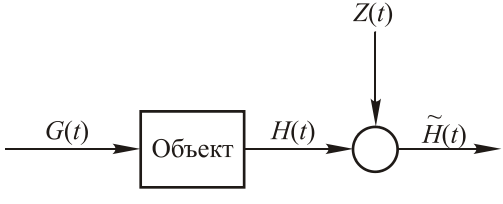


Рис. 2.11. Схема объекта идентификации

[1] См. [60].

Помеха $Z\ t$ аддитивна:

$$\tilde{H}\ t = H\ t + Z\ t\ , \tag{2.22}$$

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

Предполагая, что все используемые реализации случайных процессов центрированы, получаем из (2.22):

$$\begin{aligned} R_{\tilde{H}G} &= \lim_{T\to\infty} T^{-1} \int_0^T G\ t-\tau \left\{ \int_0^\infty \omega\ \lambda\ G\ t-\lambda\ d\lambda \right\} dt + R_{ZG}\ \tau\ = \\ &= \int_0^\infty \omega\ \lambda\ \left\{ \lim_{T\to\infty} T^{-1} \int_0^T G\ t-\tau\ G\ t-\lambda\ dt \right\} d\lambda + R_{ZG}\ \tau\ = \\ &= \int_0^\infty \omega\ \lambda\ R_{GG}\ \tau-\lambda\ d\lambda + R_{ZG}\ \tau\ . \end{aligned} \tag{2.23}$$

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

$$\begin{aligned} R_{\tilde{H}G}\ \tau &\triangleq \lim_{T\to\infty} T^{-1} \int_0^T \tilde{H}\ t\ G\ t-\tau\ dt; \\ R_{GG}\ \tau &\triangleq \lim_{T\to\infty} T^{-1} \int_0^T G\ t\ G\ t-\tau\ dt; \\ R_{ZG}\ \tau &\triangleq \lim_{T\to\infty} T^{-1} \int_0^T Z\ t\ G\ t-\tau\ dt, \end{aligned} \tag{2.24}$$

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:

$$\hat{R}_{\tilde{H}G} = \int_0^{T_\omega} \omega\ \lambda\ \hat{R}_{GG}\ \tau-\lambda\ d\lambda, \tag{2.25}$$

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 $\left|\omega\ t\right| < \varepsilon\omega_{\max}$ $\forall t > T_\omega$, где, например, $\varepsilon = 0{,}05$. Так как $\omega\ t$ первоначально неизвестна, то обоснованный выбор T_ω может быть произведен итеративно при решении нескольких

задач (2.25) для различных T_ω. Аналогично выбираются рабочие промежутки интегрирования при получении оценок корреляционных функций.

Рассмотрим один из возможных способов решения интегрального уравнения (2.25). Перейдем в (2.25) к дискретному времени

$$\hat{R}_{\tilde{H}G}\ qj\ = \sum_{i=1}^{N} \omega_i \hat{R}_{GG}\left[q\ \ j-i\ \right],\ \ 1 \le j \le N, \tag{2.26}$$

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 (2.26) имеет вид

$$J\ x\ = \sum_{j=1}^{N} \left[\sum_{i=1}^{N} \omega_i \hat{R}_{GG}\left[q\ \ j-i\ \right] - \hat{R}_{\tilde{H}G}\ qj\ \right]^2, \tag{2.27}$$

где $x = \ \omega_1, \omega_2, ..., \omega_N$. Выбирая x из условия минимума (2.27), приходим к схеме идентификации, показанной на рис. 2.12, где *БУ* — блоки умножения; ↗ — блок переменной задержки; *И* — интеграторы.

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

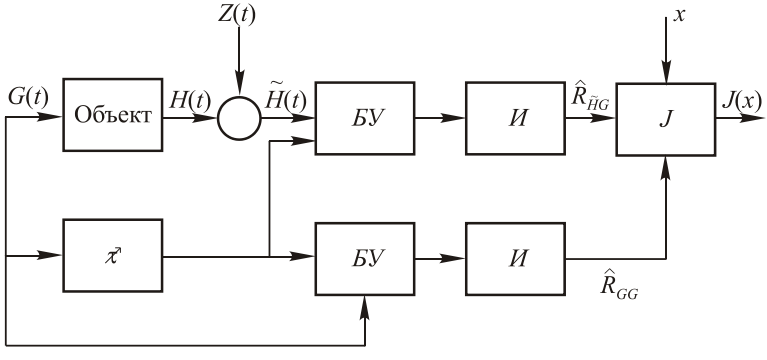


Рис. 2.12. Схема реализации корреляционного метода идентификации

2.3.3. 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:

$$H_{\mathrm{M}}\ \ t\ = F\left[G\ \ t\ ,\ \alpha\ \ t\ \right]. \tag{2.28}$$

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

$$\alpha\ t\ = \left[\alpha_1\ t\ , \ldots \alpha_2\ t\ , \ldots, \alpha_m\ t\ \right]. \tag{2.29}$$

Параметризация $\alpha\ t$ приводит к следующему представлению:

$$\alpha_i\ t\ = \sum_{j=1}^{N_i} c_{ij} \varphi_{ij}\ t\ , \quad i \in 1:m\ , \tag{2.30}$$

где φ_{ij} ($i \in 1:m$; $j \in 1:N_i$) — заданная система координатных функций. Подставляя (2.30) в (2.28), получим

$$H_{\text{M}}\ t\ = F\left[G\ t\ , x\right],$$

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.

2.3.4. 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. 2.13). 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. 2.13 не показаны. В результате строится модельная зависимость $I\ x, y\ $, аппроксимирующая на некотором отрезке времени реальную зависимость $J\ x, y\ $. В дискретные моменты времени τ_i (текущие или упреждающие) модель объекта фиксируется: $J_{\text{M}}\ x \triangleq I\ x, \tau_i\ $. Далее с помощью управляющего устройства *УУ* находится оптимальный режим из условия

$$x = \arg \min_{x_{\text{M}} \in D} J_{\text{M}}\ \ x_{\text{M}}\ . \tag{2.31}$$

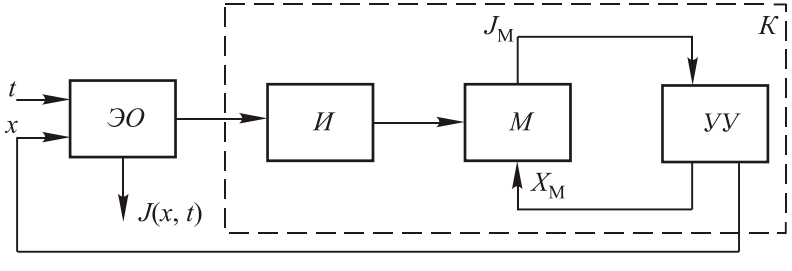


Рис. 2.13. Система экстремального регулирования
(*К* — часть системы, реализованная в управляющем компьютере)

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. 2.14, где $\Delta t_{\text{и}}$ означает время одного цикла идентификации, а $\Delta t_{\text{п}}$ — время поиска экстремума функционала $J_{\text{М}}$. В моменты времени τ_i происходит обновление модели за счет изменения вида функционала $J_{\text{М}}$. В точках t_i по результатам решения задачи (2.31) корректируется вектор x управляющих параметров. Предполагается, что в течение одного цикла $\Delta t_{\text{и}}$ идентифицируемые параметры с достаточной точностью можно считать постоянными *(см. разд. 2.3.3)*.

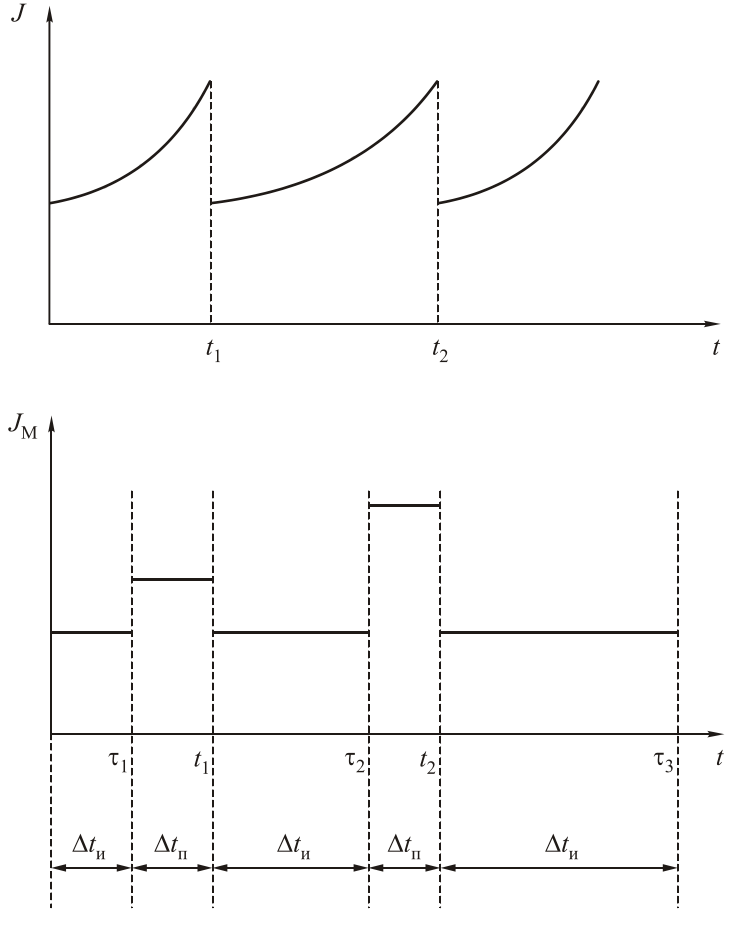


Рис. 2.14. Процесс экстремального регулирования с подстройкой управляющих параметров в дискретные моменты времени

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 $\Delta t_{\text{и}}$ и $\Delta t_{\text{п}}$). 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 вышеизложенной процедуры и требованиями к максимально допустимой величине отклонения процесса от оптимального режима.

В схеме управления (см. рис. 2.13) 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.

2.3.5. Синтез адаптивных систем автоматического управления

Изложенный в предыдущем разделе метод обучающейся модели может рассматриваться как некоторый общий *принцип адаптации*. Он допускает конкретные реализации, отличающиеся способом формирования функционала $J_{\text{М}}\ x$. 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. 2.15.

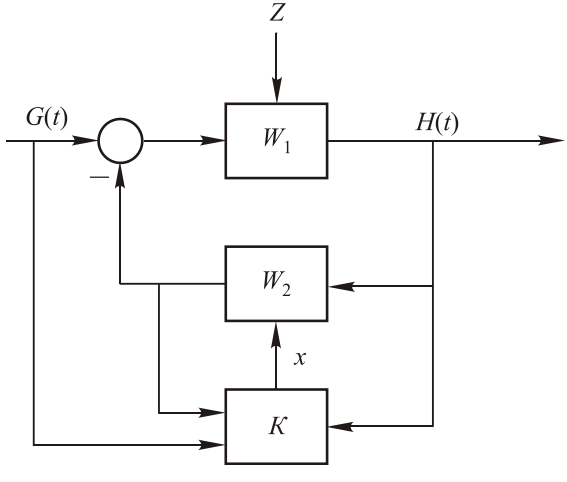


Рис. 2.15. Схема адаптации с корректирующим звеном в цепи обратной связи

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 (K) в виде некоторой *эталонной модели* (в данном случае — цифровой). Идентификатор, реализованный в управляющем компьютере K, определяет неизвестные (дрейфующие во времени) параметры W_1. В результате получаем полную модель M системы управления, характеризуемую функциями W_1, W_2. Далее управляемые параметры x корректирующего звена W_2 определяются как решение задачи параметрической оптимизации вида:

$$J_{\text{М}} = \left\| H_{\text{Э}}\ t\ - H_{\text{М}}\ x, t\ \right\|_t^2 \to \min_{x \in D}, \tag{2.32}$$

где $H_{\text{Э}}\ t$ — реакция эталона на заданный отрезок входного сигнала $G\ t$; $H_{\text{М}}\ x, t$ — реакция модели M 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 предыдущему пункту, однако вектор корректирующих параметров x выбирается из условия совпадения модельных коэффициентов $a_i\ x$ передаточной функции $W = W_1 / \ 1 + W_1 W_2$ реальной системы с соответствующими эталонными коэффициентами b_i:

$$J_{\text{М}}\ x\ = \sum_{i=0}^{N} \left[a_i\ x\ - b_i \right]^2 \to \min_{x \in D}. \tag{2.33}$$

Множество D определяет множество физически реализуемых коэффициентов корректирующего звена W_2.

Пример. Пусть $W_1\ \ p\ = 1 / \ p^2 + g_1 p + g_0$, $W_2\ \ p\ = x_1 p + x_0$, причем помеха Z изменяет коэффициенты g_1, g_0 заранее неизвестным образом. Тогда на каждом шаге идентификации необходимо строить оценки $\hat{q}_1$, $\hat{q}_0$ и далее определять $x = \ x_0, x_1$ из условия (2.33), где $W = 1 / \ p^2 + a_1 p + a_0$, $a_0\ \ x\ = x_0 + g_0$, $a_1\ \ x\ = x_1 + g_1$. Если

эталонная модель имеет достаточно простой вид, например, $W_Э\ \ p\ =1\Big/\ p^2+b_1p+b_0$, то параметры корректирующего звена, доставляющие минимум функционалу $J_М\ \ x$, могут быть найдены непосредственно:

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

2.3.6. Синтез статистически оптимальных систем автоматического управления

Задача определения оптимальной весовой функции линейной стационарной системы автоматического управления[2]

На рис. 2.16 представлена расчетная схема рассматриваемой задачи.

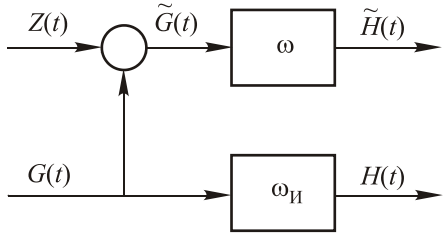


Рис. 2.16. Взаимодействие сигналов
в задаче определения оптимальной весовой функции

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 реализации заданного алгоритма преобразования входного сигнала с помощью идеальной весовой функции $\omega_И\ \ \tau$. Ошибка системы может быть представлена в виде

$$E\ t\ =H\ t\ -\tilde{H}\ t\ .$$

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

$$D_E=M\ \ E^2\ \ t\ \ . \tag{2.34}$$

[2] См. [3].

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 (2.34). В курсах по теории автоматического управления показывается, что при выполнении некоторых дополнительных условий оптимальная в смысле критерия (2.34) весовая функция $\omega_*\ \tau\ $ удовлетворяет интегральному уравнению Винера — Хопфа

$$\int_0^\infty R_{\tilde{G}\tilde{G}}\ \ \tau - \lambda\ \ \omega_*\ \ \lambda\ \ d\lambda = R_{\tilde{G}H}\ \ \tau\ . \qquad (2.35)$$

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

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

Для решения уравнения Винера — Хопфа (2.35) можно воспользоваться методами из *разд. 2.3.2*. В результате приходим к задаче минимизации квадратичного функционала типа (2.27).

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[3]

Пусть структурная схема замкнутой системы автоматического управления с передаточной функцией $W\ \ p, x\ $ задана с точностью до вектора $x = x_1, x_2, ..., x_n$ управляемых параметров. На систему согласно рис. 2.17 воздействуют стационарные и стационарно связанные полезный сигнал $G\ t\ $ и аддитивная помеха $Z\ t\ $. Через $W_{\text{И}}$ обозначена желаемая (идеальная) передаточная функция.

[3] См. [3].

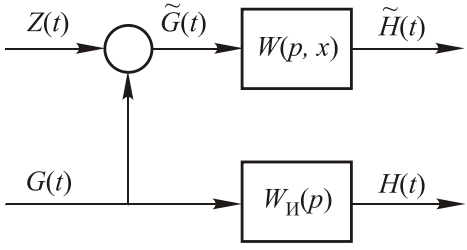


Рис. 2.17. Взаимодействие сигналов в задаче оптимизации системы с заданной структурой

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

$$D_E = D_E\ x\ = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left|W\ \ j\omega,\ x\ \right|^2 S_{\tilde{G}\tilde{G}}\ \ \omega\ \ -$$

$$-2W\ \ j\omega,\ x\ W_{\text{И}}\ -j\omega\ S_{\tilde{G}G}\ \ \omega\ + \left|W_{\text{И}}\ \ j\omega\ \right|^2 S_{GG}\ \ \omega\ \ d\omega, \qquad (2.36)$$

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.

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[4]

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.

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

$$G\ \ t\ = \sum_{i=0}^{m} G_i t^i + G'\ \ t\ , \qquad (2.37)$$

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

[4] См. [3].

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. 2.16). 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:

$$\int_0^T \left\{ \sum_{i=0}^m \alpha_i \tau^i \lambda^i + R_{X'X'}\ \tau, \lambda\ \right\} \omega\ \tau, \lambda\ d\lambda - R_{X'H}\ \tau, t\ = 0, \qquad (2.38)$$

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 (2.38) при фиксированном t_j имеет вид (2.25), и его решение $\omega\ t_j, \lambda\ ,$ $0 \le \lambda \le T,\ \ j \in 1:N$ может быть получено рассмотренными в *разд. 2.3.2* методами на основе алгоритмов параметрической оптимизации. Учет нестационарности приводит в данном случае к необходимости решения множества указанных задач параметрической оптимизации на сетке $t_j,\ j \in 1:N\ ;\ t_j \in 0, T\ ,$ что выдвигает дополнительные требования к эффективности соответствующих оптимизирующих процедур.

2.3.7. 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

$$y_i \; x, \xi \; \le t_i, \; t_i \in R^1, \; i \in 1:m \; , \tag{2.39}$$

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*, т. е. задача построения вектора x из условия

$$x \in \arg\min_x J \; x \; ,$$

где, например, функционал $J \; x$ характеризует качество решения системы неравенств (2.39). Возможны и другие оптимизационные постановки задачи оптимального проектирования. Некоторые конкретные формы критериев оптимальности будут рассмотрены в *разд. 3.8*.

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. 2.18.

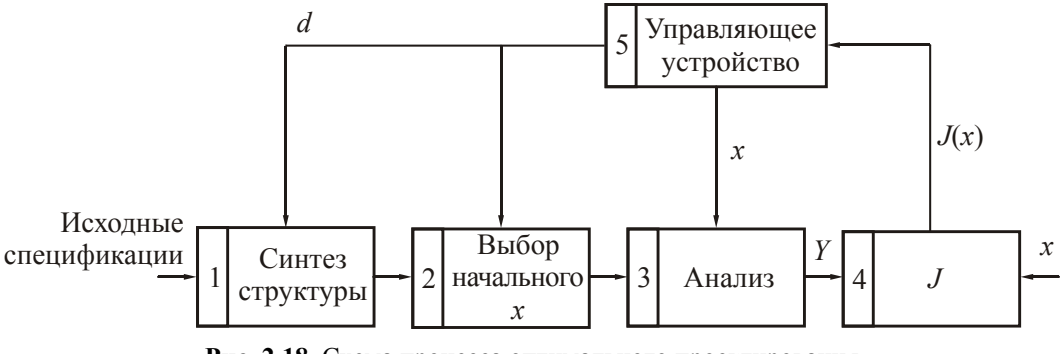


Рис. 2.18. Схема процесса оптимального проектирования

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, т. е. 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 *разд. 3.5*.

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 [83].

Реальные процессы проектирования имеют сложную иерархическую структуру, состоящую из нескольких уровней проектирования [17]. На самом высшем уровне формируется *облик объекта*, т. е. 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. 2.18 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.

2.4. 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 может быть сформулировано 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 3

Математические модели теории конечномерной оптимизации

3.1. Задачи конечномерной оптимизации

Формулировка задачи *математического программирования*, как задачи конечномерной оптимизации, была дана во *введении*. Задается целевой функционал

$$J : R^n \to R \ .$$

Тем или иным способом выделяется множество допустимых элементов $U = \left. u \right| u \in R^n$. Далее ставится задача построения минимизатора (минимизаторов) целевого функционала на допустимом множестве или, в более общем случае, задача построения минимизирующих последовательностей.

Вообще говоря, термины "математическое программирование" и "конечномерная оптимизация" являются синонимами. Однако мы будем употреблять оба термина, полагая, что стандартными задачами математического программирования являются все-таки однокритериальные задачи, тогда как термин "конечномерная оптимизация" в нашем понимании включает и задачи оптимизации при наличии нескольких целевых функционалов (многокритериальные задачи). Эти задачи далее будут рассмотрены достаточно подробно.

Во *введении* говорилось, что при практических приложениях теории оптимизации принято различать "задачи аппроксимации" (аппроксимационные задачи оптимизации) и "задачи идентификации" (идентификационные задачи оптимизации). Разделение этих задач, как и указывалось, связано с двумя типами сходимости минимизирующих последовательностей — по аргументу и по функционалу. Приведем соответствующий содержательный пример.

Пример. Дана задача Коши для системы обыкновенных дифференциальных уравнений

$$\frac{d\,x}{dt} = f\ \ x,\ p,\ t\ ,\ \ \ x\ \ 0\ \ = x_0. \tag{3.1}$$

Предполагается, что система описывает некоторый химико-технологический процесс полимеризации, протекающий в лабораторных условиях. Вектор p является

вектором скоростей "элементарных реакций", принимающих при постоянной температуре конкретные, но неизвестные значения. Здесь $t \in 0, T$ — время; p — k-мерный вектор параметров, x — l-мерный вектор "фазовых" переменных. Задан также набор из m экспериментальных кривых, полученных в лабораторных условиях и характеризующих реальную систему при разных начальных условиях x_0:

$$x^1(t), x^2(t), \ldots, x^m(t), t \in 0, T. \tag{3.2}$$

Предположим, что в результате процесса оптимизации найден вектор p^* — минимизатор функционала

$$J(p) = \sum_{i=1}^{m} \sum_{j=1}^{N} \left\| x^i(t_j) - x^i(p, t_j) \right\|^2, \tag{3.3}$$

где $x(p, t)$ — численное решение уравнения (3.1), полученное при текущем выборе p и соответствующем начальном условии. Задача минимизации функционала (3.3) (задача параметрической идентификации) ставится и решается с единственной целью — получение достаточно точной оценки компонентов вектора p, имеющих конкретный физический смысл скоростей химических реакций, протекающих при заданной температуре.

Далее полученный вектор p^* предполагается использовать при моделировании и оптимизации в промышленных условиях режима управления каскадом реакторов по производству полимера (например, ударопрочного сополимера стирола с каучуком). Сам химико-технологический процесс производства полимера описывается другими математическими моделями, отличными от (3.1):

$$\frac{dz}{dt} = \varphi(z, p, q, t). \tag{3.4}$$

Здесь p — найденный ранее (на этапе идентификации модели (3.1)) вектор оценок скоростей элементарных реакций; t — время; z — вектор фазовых переменных процесса; q — искомый вектор управляющих (режимных) параметров, определяемый как решение некоторой дополнительной задачи оптимизации вида

$$F(q) = \sum_{k=1}^{s} \left\| z(t_k, q) - z(t_k) \right\|^2 \to \min_q, \tag{3.5}$$

где $z(t_k)$ — желаемые значения вектора фазовых переменных в заданный момент времени; $z(t_k, q)$ — соответствующие расчетные значения.

Ясно, что если при решении задачи минимизации функционала (3.3) нас интересует сходимость по аргументу, то в случае (3.5) мы должны обеспечить лишь достаточно быструю сходимость по функционалу.

Заметим, что приведенный пример носит чисто иллюстративный (модельный) характер и не отражает всех аспектов моделирования реальных процессов полимеризации.

Многочисленные примеры оптимизационных задач, где требуется лишь сходимость по функционалу (аппроксимационные задачи), дают не только собственно задачи аппроксимации, но также теория и практика оптимального проектирования. В таких задачах целевой функционал J отражает качество проекта, а аргументом является некоторый вектор конструктивных параметров. При этом по-прежнему часто оказывается неважным, с помощью какого вектора параметров (из допустимой области) удалось обеспечить заданное качество.

3.2. Терминологические замечания. Классификация задач

В классе задач математического программирования (МП) выделяются следующие задачи.

3.2.1. Нелинейное программирование

В задачах *нелинейного программирования* (НП) предполагается, что множество допустимых значений аргументов U или D минимизируемого функционала J задается с помощью систем равенств и неравенств. Таким образом, имеем следующее общее представление для НП-задач:

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

$$D = \{x \in R^n \mid g_i(x) \le 0,\ i = 1, \ldots, l;\ h_j(x) = 0,\ j = 1, \ldots, s\},$$

где J, g_i, h_j — функционалы, реализующие отображения $R^n \to R$.

3.2.2. Линейное программирование

Если целевой функционал и функционалы ограничений могут быть заданы с помощью линейных функций, то такие задачи НП называются задачами *линейного программирования* (ЛП).

Так называемая *основная задача ЛП* (ОЗЛП) ставится следующим образом.

Требуется найти минимизатор линейной функции

$$J(x) = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n = \sum_{i=1}^{n} c_i x_i = \langle c, x \rangle$$

в допустимом множестве D, элементы которого удовлетворяют ограничениям:

- $Ax = b$ (ограничения-равенства);
- $x_i \ge 0,\ i = 1, \ldots, n$ (ограничения-неравенства).

Здесь A — матрица вещественных чисел размером $m \times n$; b — m-мерный вектор; $\langle \cdot, \cdot \rangle$ — знак скалярного произведения.

Кратко ОЗЛП может быть записана в виде

$$\min \ cx \mid x \geq 0, \ Ax = b \ ,$$

где cx — краткая запись скалярного произведения вектора c на вектор x.

Существует несколько эквивалентных форм ЛП-задач. Например, все следующие виды ЛП-задач эквивалентны в смысле сводимости друг к другу:

$$\min \ cx \mid x \geq 0, \ Ax = b \ ,$$

$$\min \ cx \mid x \geq 0, \ Ax \geq b \ ,$$

$$\min \ cx \mid Ax \geq b \ ,$$

$$\max \ cx \mid x \geq 0, \ Ax = b \ ,$$

$$\max \ cx \mid x \geq 0, \ Ax \leq b \ ,$$

$$\max \ cx \mid Ax \leq b \ .$$

Таким образом, любой метод решения одной из представленных задач может быть преобразован для решения всех остальных. (Существуют и другие эквивалентные формулировки, основанные, в частности, на теореме двойственности.)

ЛП-задачи образуют важный класс НП-задач по крайней мере в двух отношениях. Во-первых, такие задачи характерны для многих практических (особенно экономических) приложений. Во-вторых, для решения ЛП-задач созданы специальные и достаточно эффективные методы, позволяющие в большом числе случаев получать решение за конечное число шагов.

Кроме того, существуют вычислительные технологии решения общих НП-задач на основе их последовательной аппроксимации соответствующими ЛП-задачами.

Теория и методы линейного программирования представлены в большинстве публикаций по методам оптимизации, исследованию операций, теории принятия решений и т. п. Стандартные пакеты прикладных программ позволяют решать эти задачи достаточно эффективно, и обычно никаких проблем не возникает. Поэтому в данной книге эти вопросы не рассматриваются.

3.2.3. Выпуклое программирование

Задачи *выпуклого программирования*, так же как и ЛП-задачи, оказываются достаточно удобными для точного математического анализа основных ситуаций. Так называются задачи, в которых целевой функционал и допустимое множество D являются выпуклыми. Существуют свои специфические для данного класса задач методы минимизации, например, методы возможных направлений, линеаризации, двойственные и др. Мы не будем далее рассматривать эти методы. Отметим только, что уже такая простая задача, как задача оценки скалярного параметра a скалярного дифференциального уравнения

$$\frac{dx}{dt} = ax, \quad x\ 0\ = x_0$$

по известному экспериментальному значению $x\ t_1 = x_z$ из условия минимума функционала

$$J\ a\ =\ x\ t_1,\ a\ - x_z\ ^2 \to \min_a ,$$

оказывается невыпуклой. Здесь $x\ t,\ a$ — решение дифференциального уравнения при заданном произвольном значении параметра *a*.

Существует еще несколько "программирований", например "квадратичное", "недифференцируемое" и т. д. Мы не будем здесь приводить соответствующие формулировки, имея в виду представленный выше общий случай НП-задач.

В заключение отметим важный класс задач нелинейного программирования, допускающих весьма специфические и принципиально важные методы решения. Это еще одно "программирование" — *динамическое*. Методы динамического программирования ориентированы на НП-задачи, в которых целевой функционал имеет специальную сепарабельную структуру, что позволяет трактовать задачу построения минимизатора как некоторую многоэтапную процедуру, в которой результирующее значение целевого функционала является суммой его значений на отдельных этапах. На методах динамического программирования основаны, в частности, многие сетевые методы, когда, например, требуется отыскать минимальный путь на графе и т. п.. Общая теория динамического программирования имеет гораздо более широкий спектр приложений. При некоторых предположениях из теории динамического программирования вытекает и "принцип максимума" теории оптимального управления.

3.3. Канонические задачи

Канонической задачей параметрической оптимизации далее называется конечномерной задачей однокритериальной оптимизации без возмущений и без ограничений на возможные значения аргументов минимизируемого функционала:

$$J\ x\ \to \min_x,\ \ x \in R^n . \tag{3.6}$$

Здесь R^n означает конечномерное (*n*-мерное) линейное векторное пространство векторов, составленных из *n* действительных чисел.

Такие задачи в курсах по теории оптимизации называются также задачами *безусловной оптимизации*, т. е. 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, [26].

Конечномерные задачи безусловной оптимизации достаточно часто возникают в приложениях. Например, как было показано в *разд. 2.3.1*, 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 конечномерной оптимизации к конечному (или бесконечному) множеству задач (3.6). При этом мы следуем достаточно традиционному в компьютерной математике подходу, связанному со сведением "сложных" задач к последовательности "простых". Например, существуют технологии, позволяющие сводить нелинейные задачи оптимизации к последовательности линейных, неквадратичные — к последовательности квадратичных, задачи с ограничениями — к последовательности задач без ограничений, многокритериальные задачи — к некоторому набору однокритериальных 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 и использующих операцию проектирования на допустимое множество. Однако, если это возможно, везде в этой книге будет отдаваться предпочтение принципу сведения сложного к простому, сформулированному ранее.

Основной итог *глав 2* и *3* должен состоять в обосновании важной роли канонических оптимизационных задач в алгоритмическом обеспечении современной теории оптимизации.

3.4. Многокритериальные задачи

Когда мы говорим о многокритериальных задачах оптимизации, то имеем в виду достаточно 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 был 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 конечномерная многокритериальная задача с ограничениями формулируется следующим образом:

$$f_i\ x \to \min_x,\ x \in D,\ i \in 1:k \tag{3.7}$$

$$D = \{ x \in R^n \mid g_i\ x \le 0,\ i \in 1:m\ ;\ g_i\ x = 0,\ i \in m+1:s\ ;\ a_j \le x_j \le b_j,\ j \in 1:q \}\ .$$

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, т. к. 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, т. к. максимизация функции $\varphi\ x$ эквивалентна минимизации $f\ x = -\varphi\ x$. Кроме этого мы учитываем здесь, что замена знака у левых частей неравенства $p\ x \ge 0$ меняет знаки неравенств на противоположные и приводит их к стандартному виду $g\ x \le 0$, где $g\ x = -p\ x$.

Задача (3.7) 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 [58], [27].

Существуют различные способы сведения исходной многоцелевой задачи (3.7) к задачам с единым критерием. Формулировка скалярного критерия оптимальности (целевой функции) должна 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, т. е. 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 [83].

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 этого, 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:

$$J(x) = \sum_{i=1}^{k} \alpha_i f_i(x) \to \min_{x \in D},\ \alpha_i > 0,\ \sum_{i=1}^{k} \alpha_i = 1. \tag{3.8}$$

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: $\partial J / \partial f_i = \alpha_i$, $i \in 1:k$. При наличии существенно разнохарактерных частных критериев обычно бывает достаточно 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

$$J(x) = \prod_{i=1}^{k} f_i^{\alpha_i}(x) \to \min_{x \in D} \tag{3.9}$$

принципиально ничем не отличаются от конструкции (3.8), т. к. от функционалов J, $f_1, ..., f_k$ можно перейти к их логарифмам, и тогда (3.9) 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* $t_1, ..., t_k$, фигурирующих в правых частях критериальных ограничений:

$$f_i(x) \le t_i. \tag{3.10}$$

В этом случае в качестве скалярного критерия можно использовать условия вида

$$J(x) = \min_i \alpha_i (t_i - f_i(x)) \to \max_{x \in D} \tag{3.11}$$

или

$$J(x) = \max_i \alpha_i (f_i(x) - t_i) \to \min_{x \in D}, \tag{3.12}$$

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.

3.5. Парето-оптимальные решения

Пусть решается задача

$$f_i \ x \ \to \min_{x \in D}, \ i \in 1:k \ , \ D \subset R^n. \tag{3.13}$$

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

$$f_i \ x' \ \le f_i \ x'' \tag{3.14}$$

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 (3.14). Множество, включающее все эффективные решения, обозначается $P \ D$ и называется *множеством Парето* для векторного отображения f: $D \to R^n$, $f = \ f_1, ..., 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* [53].

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 (3.13), если не существует такой точки $x'' \in D_1$, 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 3.1

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 [53]. Действительно, нетрудно доказать, что решение, которое слабо эффективно по сокращенному набору критериев, будет слабо эффективным и по расширенному набору.

Упражнение 3.2

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. 3.1 $P\ f = b, c$; $S\ f = a, b \cup b, c \cup c, d$. 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.

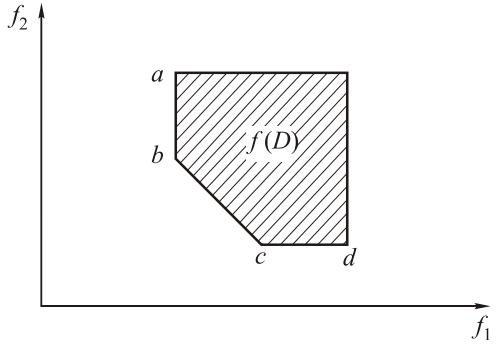


Рис. 3.1. Эффективные и слабо эффективные оценки задачи $f_i \to \min,\ i = 1,\ 2$

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

Теорема 3.1. Пусть заданы произвольные $\alpha_i > 0,\ i \in 1:k$, тогда решение задачи

$$\min_i \alpha_i\ t_i - f_i\ \to \max_{x \in D} \tag{3.15}$$

при любых фиксированных t_i есть слабо эффективный вектор. Наоборот, любой слабо эффективный вектор может быть получен как решение задачи (3.15) при некоторых $\alpha_i > 0$ и $t_i > f_i\ x$, $i \in 1:k$, $x \in D$.

Доказательство. Пусть x^0 есть решение задачи (3.15) 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, $t_i - f_i\ x^0 > 0$. Положим $\alpha_i' = 1 \Big/ t_i - f_i\ x^0$ и покажем, что

$$\max_{x \in D} \min_i \alpha_i'\ t_i - f_i\ x = \min_i \alpha_i'\ t_i - f_i\ x^0 = 1. \tag{3.16}$$

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

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

Но тогда и $\min\limits_i \alpha_i'\ t_i - f_i\ x' \leq 1$, что означает выполнение (3.16), т. к. x' — любой элемент *D*. Теорема доказана.

Замечание 1. Обратное утверждение теоремы 3.1 справедливо в следующей формулировке: любой слабо эффективный вектор x^0 может быть получен как решение задачи (3.15) при $t_i > f_i\ x$,

$$\alpha_i > 0,\ \sum_{i=1}^{k} \alpha_i = 1. \tag{3.17}$$

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

Замечание 2. Из доказательства второй части теоремы 3.1 следует, что для любого слабо эффективного вектора x^0 могут быть выбраны такие $\alpha_i > 0$, для которых он получается как решение задачи (3.15), причем

$$\alpha_1\ t_1 - f_1\ x^0 = \alpha_2\ t_2 - f_2\ x^0 = \ldots = \alpha_k\ t_k - f_k\ x^0. \tag{3.18}$$

Теорема 3.2. Пусть $\alpha_i > 0$, $i \in 1:k$; D' — произвольное подмножество *D*, содержащее хотя бы один эффективный вектор, тогда решение задачи

$$\sum_{i=1}^{k} \alpha_i f_i\ x \to \min_{x \in D'} \tag{3.19}$$

есть эффективный вектор.

Доказательство. Пусть x^0 есть решение задачи (3.19) 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 3. Обратное утверждение без дополнительных предположений неверно. Существуют эффективные векторы, не являющиеся решением задачи (3.19).

Теорема 3.3. Пусть T_α — непустое множество решений задачи (3.15). Тогда существует вектор $x \in T_\alpha$, являющийся эффективным.

Доказательство. Допустим, что существует такой $x' \in D$, что $f_i\ x' \le f_i\ x^0$, $i \in 1:k$, где x^0 — произвольный элемент множества T_α. Тогда

$$\min_i \alpha_i\ t_i - f_i\ x' \ge \min_i \alpha_i\ t_i - f_i\ x^0 = \max_{x \in D} \min_i \alpha_i\ t_i - f_i\ x,$$

т. е. $x' \in T_\alpha$ и, следовательно, вне множества T_α не может существовать решений x', для которых $f_i\ x' \le f_i\ x^0$, $i \in 1:k$. Теорема доказана.

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

По схеме теоремы 3.1 могут быть получены все слабо эффективные решения, если вектор $\alpha = \alpha_1, ..., \alpha_k$ пробегает все множество векторов (3.17). Напротив, вычислительная схема (3.19) позволяет получать только эффективные решения, но не все. А. Джоффриону принадлежит идея выделения эффективных точек из множества $S\ D$, построенного согласно (3.15) с помощью процедуры (3.19). Справедливо следующее утверждение [20].

Теорема 3.4. Пусть $\alpha_i > 0$, $i \in 1:k$, тогда решение задачи

$$\begin{aligned} &\sum_{i=1}^{k} f_i\ x \to \min_{x' \in T_\alpha}, \\ &T_\alpha = \left\{ x' \in D \,\middle|\, x' = \arg\max_{x \in D} \min_i \alpha_i\ t_i - f_i \right\} \ne \varnothing \end{aligned} \tag{3.20}$$

есть эффективный вектор. Наоборот, любой эффективный вектор x^0 может быть получен как решение задачи (3.20) при некоторых $\alpha_i > 0$ и $t_i > f_i\ x$, $i \in 1:k$.

Доказательство. Первое утверждение теоремы есть следствие теорем 3.2 и 3.3. Действительно, согласно теореме 3.3 множество T_α будет содержать эффективный вектор, а согласно теореме 3.2 решение задачи (3.19) 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 3.1 при $\alpha = \alpha'$ он получается как решение задачи (3.15) в составе множества $T_{\alpha'}$ с выполнением соотношений (3.18). 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 (3.18) мы получили бы

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

что противоречит свойству $x^0 \in T_{\alpha'}$. Следовательно, x^0 есть решение задачи (3.20). Теорема доказана.

Доказанные теоремы позволяют понять сущность методов максиминной, минимаксной и линейной сверток, обсуждавшихся в *разд. 3.3*. Становится ясной роль весовых коэффициентов, параметризующих множество слабо эффективных решений. С точки зрения приложений представляет интерес соотношение (3.18), 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 работе [83].

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

Теорема 3.5. Решение задачи

$$f_1\ x \to \min_{x \in D'}, \tag{3.21}$$

где множество

$$D' = \{ x \in D \mid f_i\ x \le t_i,\ i \in 2:k \} \tag{3.22}$$

не пусто, есть слабо эффективный вектор.

Доказательство. Пусть x^0 есть решение задачи (3.21) и существует $x' \in D$, такой, что

$$f_i\ x' < f_i\ x^0 \ , \ i \in 1:k \ . \tag{3.23}$$

Тогда $x' \in D'$, т. к. в противном случае это противоречило бы свойству $f_1\ x^0 \le f_1\ x$ для $x \in D'$. Следовательно, $x' \in D'$, и поэтому существует такой $i = i_0$, для которого $f_{i_0}\ x' > t_{i_0}$, что противоречит (3.23).

Теорема 3.6. Пусть T_t — непустое множество решений задачи (3.21). 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, $f_1\ x' = f_1\ x^0$, т. к. $x^0 = \arg\min\limits_{x \in D'} f_1\ x$. 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 (3.21) единственно, то оно эффективно.

Теорема 3.7. Любой эффективный вектор может быть получен как решение задачи (3.21) при некоторых t_i, $i \in 2:k$.

Доказательство. Пусть $x^0 \in P\ D$; положим $t_i = f_i\ x^0$, $i \in 2:k$ и покажем, что

$$f_1\ x^0 = \min f_1\ x,\ x \in D'. \tag{3.24}$$

Пусть $x' \in D'$; тогда $f_1\ x' \le f_1\ x^0$, $i \in 2:k$. Если предположить, что $f_1\ x' < f_1\ x^0$, то это будет противоречить эффективности вектора x^0. Следовательно, $f_1\ x' \ge f_1\ x^0$, что эквивалентно (3.24).

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 (3.21) 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* [85].

3.6. Методы исключения ограничений

Рассмотрим методы учета ограничений в однокритериальной задаче

$$J\ x\ \to \min_{x},\ x \in D, \tag{3.25}$$

где D задано в (3.7).

Формально наиболее просто снимаются прямые ограничения. Для этого достаточно выполнить замену переменных по одной из формул табл. 3.1, где z_i означают новые независимые переменные. Однако, как указано, например, в [21], 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 3.1

№	Ограничение	Преобразование
1	$x_i > d_i$	$x_i = d_i + \exp\ z_i$
2	$x_i \geq d_i$	$x_i = d_i + z_i^2$
3	$x_i \geq x_j,\ i \neq j$	$x_j = z_j,\ x_i = z_j + z_i^2$
4	$a_i \leq x_i \leq b_i$	$x_i = b_i + \ a_i - b_i\ \sin^2 z_i$ $x_i = 0{,}5\ \ a_i + b_i\ \ + 0{,}5\ \ b_i - a_i\ \ \sin z_i$
5	$a_i < x_i < b_i$	$x_i = b_i + \ a_i - b_i\ \ \frac{1}{\pi}\text{arcctg}\ \ z_i$ $x_i = b_i + \ a_i - b_i\ \ \frac{\exp\ z_i}{1 + \exp\ z_i}$
6	$a \leq x_i \leq b$ $a \leq x_j \leq b$ $x_i \geq x_j,\ i \neq j$	$x_j = b + \ a - b\ \sin^2 z_j$ $x_i = b + \ a - b\ \sin^2 z_j \sin^2 z_i$
7	$a < x_i < b$ $a < x_j < b$ $x_i > x_j,\ i \neq j$	$x_j = b + \ a - b\ \ \frac{1}{\pi}\text{arcctg}\, z_j$ $x_i = b + \ a - b\ \ \frac{1}{\pi^2}\text{arcctg}\, z_i\ \text{arcctg}\, z_j$

Таблица 3.1 *(окончание)*

№	Ограничение	Преобразование
8	$a_i\ x_j\ \le x_i \le b_i\ x_j$ $i \ne j$	$x_j = z_j$ $x_i = b_i\ z_j\ + a_i\ z_j\ - b_i\ z_j\ \sin^2 z_i$
9	$a_i\ x_k\ \le x_i \le b_i\ x_j$ $i \ne j,\ i \ne k$	$x_j = z_j$ $x_k = z_k$ $x_i = b_i\ z_j\ + a_i\ z_k\ - b_i\ z_j\ \sin^2 z_i$

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 [21].

Существуют 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 позволяют 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 [23]. Далее рассматривается второй, в определенном смысле более перспективный, подход.

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

$$J\ x\ \to \min_{x \in D},\ \ D = \ x \in R^n \Big| h_i\ x\ = 0,\ i \in 1:q\ \ , \tag{3.26}$$

не содержащую ограничений в виде неравенств; тогда вместо (3.26) решается последовательность задач безусловной минимизации однопараметрического семейства функционалов J_k , где

$$J_k\ x\ = J\ x\ + \sigma_k \sum_{i=1}^{q} h_i^2\ x\ ,\ \ \ \sigma_k \to \infty,\ k \to \infty. \tag{3.27}$$

Второе слагаемое в (3.27) имеет смысл "штрафа" за нарушение ограничений, что и определяет название метода. Справедлива следующая теорема.

Теорема 3.8. Пусть задача (3.26) 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 $\lim x_k = x^*$, $k \to \infty$.

Доказательства различных утверждений, аналогичных сформулированной теореме, содержатся, например, в [32].

Наиболее распространенный вариант метода штрафных функций для решения общей задачи

$$J\ x\ \to \min_x,\ x \in D, \tag{3.28}$$

$$D = \{\ x \in R^n \mid g_i\ x\ \le 0;\ i \in 1:m\ ;\ g_i\ x\ = 0;\ i \in m+1:s\}$$

состоит в применении вспомогательных функционалов

$$J_k\ x\ = J\ x\ + \sigma_k \sum_{i=1}^{s} \left[\ g_i^+\ x\ \right]^p, \tag{3.29}$$

где

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

Если функционалы J, g_i являются r раз непрерывно дифференцируемыми на множестве D, то при любом $p > r$ этим же свойством будут обладать функционалы $J_k\ x$ [15].

Основной недостаток метода штрафных функций заключатся в ухудшении обусловленности вспомогательных задач при больших σ_k. Соответствующие вопросы рассмотрены в *главе 4*.

Наиболее перспективным общим методом учета ограничений считается *метод модифицированных функций Лагранжа* [21]. Применительно к задаче (3.26) он формулируется следующим образом.

Введем в качестве обобщенного критерия оптимальности функционал

$$M\ x, \lambda, \sigma\ = J\ x\ + \langle \lambda, h\ x\ \rangle + \frac{\sigma}{2} \| h\ x\ \|^2, \tag{3.30}$$

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

$$\begin{aligned} x^{k+1} &= \arg\min_x M\ \ x, \lambda^k, \sigma\ , \\ \lambda^{k+1} &= \lambda^k + \sigma h\ x^{k+1}\ , \ h\ x\ = \left[\ h_1\ x\ , \ldots, h_q\ x\ \right], \end{aligned} \tag{3.31}$$

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, т. к. надежные методы априорного задания σ в настоящее время отсутствуют.

3.7. Влияние неопределенных факторов на процесс оптимизации

Рассмотренные в *разд. 2.3* примеры оптимизационных задач 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

$$y = F\ x, \xi\ , \tag{3.32}$$

включающую неопределенный вектор ξ. В этих условиях с методологической точки зрения целесообразно различать три основные ситуации:

- ξ — случайный вектор с известным законом распределения; вектор выходных параметров *y* реализуется многократно для различных значений вектора ξ;
- вектор *y* реализуется однократно при заранее неизвестном векторе ξ;
- выходные параметры *y* реализуются многократно; вектор ξ изменяется неизвестным образом, но не носит случайный характер, либо статистические характеристики ξ оказываются неизвестными.

В качестве примера рассмотрим задачу оптимального проектирования. Тогда первый случай характерен для анализа технологического разброса параметров при серийном производстве продукции. 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 компонентов 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 (3.32) к зависимости

$$y = F\ x,\ \overline{\xi}\ , \tag{3.33}$$

где $\overline{\xi} = M\ \xi\ $ — математическое ожидание случайного вектора ξ. В модели (3.33) 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 (3.33) гарантируется лишь оптимальность "в среднем" для достаточно большого числа реализаций ξ.

Второй возможный подход основан на использовании представления

$$y = \overline{F\ x,\ \xi\ } = M\ \ F\ \ x,\ \xi\ \ . \tag{3.34}$$

Переход от (3.32) к (3.33) или (3.34) 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 (3.34) однокритериальная задача параметрической оптимизации может быть сформулирована следующим образом:

$$\begin{aligned} &F_0\ \ x\ \ = M\ \ J\ \ x,\ \xi\ \ \ \to \min_{x \in D}, \\ &D = \ \ x \in R^n \Big| F_i\ \ x\ \ = M\ \ g_i\ \ x,\ \xi\ \ \ \le 0,\ i \in\ 1:l\ \ . \end{aligned} \tag{3.35}$$

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

Задача (3.35) является задачей *стохастического программирования* [46]. 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 информации о конкретных реализациях функционалов $J\ \ x,\ \xi\ \ ,\ \ g_i\ \ x,\ \xi\ \ ,$ отвечающих различным значениям вектора ξ.

Один из вариантов подобных методов сводит задачу (3.35) к последовательности детерминированных задач

$$\begin{aligned} &J\ \ x,\ \xi^k\ \ \to \min_{x \in D_k} \\ &D = \ \ x \in R^n \Big| g_i\ \ x,\ \xi^k\ \ \le 0,\ i \in\ 1:l \end{aligned} \tag{3.36}$$

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 (3.36), отвечающее вектору ξ^k, обозначим через x^k. Тогда последовательность векторов $\tilde{x}^k$, сходящаяся к решению исходной задачи (3.35), строится следующим образом: $\tilde{x}^{k+1} = \tilde{x}^k + \alpha_k \left(x^{k+1} - \tilde{x}^k \right)$, $\alpha_k > 0$, где $\tilde{x}^2 = x^1 + \alpha_1 \left(x^2 - x^1 \right)$.

Для сходимости процесса необходимо выполнение условий

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

Этим требованиям удовлетворяет, например, последовательность $\alpha_k = 1/k$. 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:

$$\begin{aligned} & F_0(x) = \max_{\xi \in G_\xi} J(x, \xi) \to \min_{x \in D}, \\ & D = \left\{ x \in R^n \,\middle|\, F_i(x) = \max_{\xi \in G_\xi} g_i(x, \xi) \le 0,\ i \in 1:l \right\}. \end{aligned} \tag{3.37}$$

Из (3.37) видно, что вычисление функционалов, задающих критерий и ограничения, сопряжено с решением вспомогательных задач оптимизации. В результате трудоемкость процедуры в целом оказывается достаточно высокой.

Как видно из изложенного, регулярный учет случайных факторов в общей задаче параметрической оптимизации достаточно 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.

3.8. 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

$$x^* \in \operatorname{Arg} \min_{x \in D \subset R^n} J(x),$$
$$\operatorname{Arg}\min_{x \in D} J(x) = \{x \in D \mid J(x) = \min_{x' \in D} J(x')\}. \tag{3.38}$$

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 (3.38) стандартными методами нелинейной оптимизации оказывается невозможным из-за нереальных вычислительных затрат.

3.8.1. Метод агрегирования

Метод агрегирования допускает следующее формальное представление.

Задача параметрической оптимизации формулируется в виде (3.38). Далее вводятся *агрегированные характеристики*

$$z_i = \varphi_i(x_1, x_2, ..., x_n), \tag{3.39}$$

где вектор $z = (z_1, z_2, ..., z_s)$ должен иметь существенно меньшую по сравнению с *x* размерность. Функции φ_i, по предположению, могут быть построены таким образом, что:

- целевой функционал $J(x)$ может быть представлен в виде суперпозиции отображений

$$J(x) = F(\varphi(x)) = F(z),\ z \in D_z = \varphi(D); \tag{3.40}$$

- 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 (3.39) при выполнении сформулированных требований, то процесс минимиза-

ции $J\ x$ может быть реализован следующим образом. На первом этапе решается задача построения вектора

$$z^* \in \operatorname{Arg}\min_{z\in D_z} F\ z \tag{3.41}$$

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

$$x^* = \varphi^{-1}\ z^*\ . \tag{3.42}$$

Легко видеть, что вектор (3.42) есть решение исходной задачи (3.38). Экономия вычислительных затрат достигается за счет уменьшения размерности пространства поиска при решении задачи (3.41), 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"* [38].

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

$$F\ z\ \to \min_{z\in D_z}, \tag{3.43}$$

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 (3.39). А далее определяется одно из решений системы так называемых *компонентных* уравнений

$$\varphi_i\ x_1, x_2, ..., x_n\ = z_i^*,\ i \in 1:s\ . \tag{3.44}$$

Разрешимость (3.44) 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 [62] 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 достаточно 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 (3.39), получим выигрыш за счет декомпозиции приблизительно в n/s раз. Дополнительные примеры, иллюстрирующие целесообразность изложенного подхода, содержатся в книге [46].

3.8.2. Метод вспомогательных частных критериев

Данный метод связан с сокращением множества D конкурирующих решений [46]. Это может быть выполнено на основе идеи введения вектора *вспомогательных частных критериев* $u\ x\ = \left[u_1\ x\ ,\ u_2\ x\ ,\ ...,\ u_m\ x\ \right]$.

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 $J\ x'\ \geq J\ x''$. 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 (3.38) может быть заменена следующей:

$$J\ x\ \to \min_{x \in P_u\ D}\ , \tag{3.45}$$

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 (3.45) 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

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

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_{τ} — последовательностей, обладающих свойством равномерного заполнения заданной многомерной области [Соболь, 1985].

3.9. 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 достаточно 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 пунктах 5, 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 [47], 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 достаточно высокая степень работоспособности рассматриваемых методов никак не противоречит результатам уже упоминавшейся работы [47]. Дело заключается в том, что на практике достаточно редко реализуются те специальные структуры невыпуклых задач, которые приводят к пессимистическим теоретическим оценкам.

3.10. Некоторые стандартные схемы конечномерной оптимизации

В зависимости от реальных ситуаций, примеры которых были рассмотрены в *разд. 2.3*, 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.

3.10.1. 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:

$$J_1(x) = \sum_{i=1}^{N} \alpha_i^2 \left[W(x, s_i) - W^*(s_i) \right]^2 \to \min_{x \in D}; \tag{3.46}$$

$$J_2(x) = \max_{i=1,N} \alpha_i \left| W(x, s_i) - W^*(s_i) \right| \to \min_{x \in D}, \tag{3.47}$$

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 достаточно прост и обладает свойством "гладкости". 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 (3.47), однако он не сохраняет характеристики гладкости функции $W(x, s)$, что требует привлечения специальных методов оптимизации.

Как показывает практика, достаточно 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 (3.47) ищется минимум функционала со *среднестепенной структурой* [60]:

$$J_3(x) = \sum_{i=1}^{N} \varphi_i^{\nu}(x) \to \min_{x \in D}, \ \nu = 2, 3, \ldots \tag{3.48}$$

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

При достаточно больших значениях ν решения задач (3.47), (3.48) 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 $\varphi_i \geq 0$, $i \in 1:N$ — произвольные числа. Кроме этого, можно показать, что операция извлечения корня ν-й степени не влияет на локализацию точки минимума.

Функционал J_3 совмещает в себе достоинства 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 $\nu = 2$. 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 *разд. 3.6*.

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

3.10.2. Системы неравенств

Наиболее часто задачи с неравенствами возникают при формализации основных спецификаций к оптимизируемому объекту *(см. разд. 2.3.7)*. 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, т. е. $W\ x, s_i\ \in \left[W\ s_i\ - \delta_i, W\ s_i\ + \delta_i \right]$, $i \in 1:N$, то требование $W\ x, s_i\ \cong W\ s_i$ очевидно эквивалентно двум неравенствам:

$$W\ x, s_i\ \geq W\ s_i\ - \delta_i;$$

$$W\ x, s_i\ \leq W\ s_i\ + \delta_i.$$

В общем случае имеем

$$y_i\ x\ \leq t_i,\ i \in 1:m\ . \qquad (3.49)$$

С помощью задания системы (3.49) 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:

$$\varphi_{\text{КН}} \leq \varphi_k\ x\ \leq \varphi_{\text{КВ}},$$

что приводит к двум неравенствам вида (3.49):

$$y_i\ x\ \triangleq \varphi_k\ x\ \leq t_i \triangleq \varphi_{\text{КВ}};$$

$$y_j\ x\ \triangleq -\varphi_k\ x\ \leq t_j \triangleq -\varphi_{\text{КН}}.$$

Поэтому трактовка задачи (3.49) как многокритериальной вида $y_i\ x\ \to \min$ с применением соответствующих процедур свертки обычно нецелесообразна. Как пра-

вило, необходимо обеспечить *безусловное выполнение всех неравенств* (3.49), не допуская больших значений $y_i\ x$ (превышающих соответствующие t_i) из-за очень малых значений $y_i\ x$ других выходных параметров.

Рациональный подход связан с применением критериев вида

$$J\ x\ = \min_i \gamma_i^{-1}\ t_i - y_i\ x\ \to \max_{x\in R^n} \tag{3.50}$$

или

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

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 (3.49). Согласно (3.50) максимизируется минимальный из запасов.

Функционал (3.50), так же как и (3.47), 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,

$$\min_i z_i\ x\ = \max_i \left[\exp\ -z_i\ x\ \right].$$

Поэтому задача (3.50) эквивалентна задаче

$$\max_i \left[\exp\ -z_i\ x\ \right] \to \min_x. \tag{3.51}$$

Для задачи (3.51) применима среднестепенная свертка (3.48), если положить $\varphi_i\ x\ \triangleq \exp\left[-z_i\ x\ \right]$, $i \in 1:m$. В результате приходим к следующему критерию оптимальности:

$$J\ x\ = \sum_{i=1}^{m} \exp\left[-\nu z_i\ x\ \right] \to \min_x,\ \nu = 1,\ 2,\ \ldots \tag{3.52}$$

Точно так же могут решаться и многокритериальные задачи вида

$$f_i\ x\ \to \min_x, i \in 1:l\ ;\ x \in D \subset R^n;$$

$$D = \left\{ x \in R^n \middle| g_i\ x\ \le t_j,\ j \in l+1:s \right\},$$

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

$$z_i\ x\ = \alpha_i \left[\ \hat{f}_i - f_i\ x\ \right] \Big/ \gamma_i ;\ i \in 1:l\ ;$$

$$z_j\ x\ = \alpha_j \left[\ t_j - g_j\ x\ \right] \Big/ \gamma_j ;\ j \in l+1:s\ ,$$

приходим к следующему целевому функционалу, аналогичному (3.52):

$$J_l\ x\ = \sum_{i=1}^{s} \exp\left[-\nu z_i\ x\ \right] \to \min_x .$$

Весовые коэффициенты α_i, α_j должны выбираться из условия $\alpha_i \ll \alpha_j$ ($i \in 1:l$, $j \in l+1:s$). В этом случае (при достаточно больших ν) критерий $J\ x$ эквивалентен минимаксному критерию (3.51), 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 $J_l\ x$ резко возрастает при нарушении хотя бы одного из условий $g_j\ x\ \le t_j$ (т. к. 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 ($j \in 1:s$) 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.

В ряде случаев рассмотренный подход в силу своего единообразия и простоты оказывается наиболее предпочтительным при решении реальных задач параметрической оптимизации.

Как показывает вычислительная практика, трудности оптимального синтеза параметров по критериям типа максимума минимального запаса (3.50) в ряде случаев вызываются исключительно негладкостью критериев, приводящей к преждевременной остановке поисковой процедуры. Целесообразно поэтому сразу обращаться к модифицированным критериям (3.52) 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 4* 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 5* и *6* методы позволяют получить

удовлетворительные результаты для невыпуклых овражных функционалов при условии их гладкости. При этом удается использовать структурные особенности функционалов (3.48), (3.50) для увеличения эффективности соответствующих вычислительных процедур.

3.10.3. Решение систем неравенств в условиях неопределенности

Здесь мы рассматриваем специальный случай неопределенности обстановки. Предполагается, что вектор ξ, входящий в левые части решаемых неравенств

$$y\ x, \xi\ \le t,\ y = \ y_1, ..., y_m\ ,\ t = \ t_1, ..., t_m\ , \tag{3.53}$$

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 (3.53):

$$J\ x\ = P\left[y\ x, \xi\ \le t \right] \to \max_x . \tag{3.54}$$

В данном случае все возможные ограничения на компоненты вектора x считаются учтенными за счет расширения системы неравенств (3.53). 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* и *5* методов параметрической оптимизации.

Альтернативный подход к решению задачи (3.54) рассмотрен в учебнике [48]. Он заключается в следующем. Потребуем, чтобы каждое неравенство (3.53) выполнялось с некоторым запасом

$$y_i\ \ x, \overline{\xi}\ + \delta_i \le t_i,\ \delta_i > 0, \tag{3.55}$$

где δ_i характеризует величину рассеяния i-го выходного параметра за счет статистических вариаций компонентов вектора ξ относительно своих средних (как правило, нулевых) значений $\overline{\xi}_i$. Требования (3.55) эквивалентны неравенствам

$$z_i\ \ x\ \triangleq \left[\frac{t_i - y_i}{\delta_i} - 1 \right] \ge 0. \tag{3.56}$$

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

На практике получила распространение максиминная форма целевого функционала

$$J\ x\ = \min_i z_i\ x\ \to \max_{x \in R^n}, \tag{3.57}$$

аналогичная (3.50). От представления (3.57) можно по методике, изложенной в *разд. 3.9.2*, перейти к выражению (3.52), более удобному для практических расчетов.

В ряде случаев в выражение (3.57) вводятся дополнительные весовые коэффициенты α_i, позволяющие регулировать степень "перевыполнения" требований (3.55) 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.

3.10.4. Сигномиальная оптимизация

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

$$J\ x\ = \sum_{i=1}^{k} s_i\ x\ , \tag{3.58}$$

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 (3.58) называется *позиномом*, а соответствующая задача параметрической оптимизации — *задачей геометрического программирования* [69].

В *разд. 5.4* будут рассмотрены процедуры сигномиальной оптимизации, основанные на простоте вычисления первых и вторых производных функционалов (3.58). 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.

3.11. Основные результаты и выводы

1. Общая постановка задачи конечномерной 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 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 конкретных предметных областях критериев качества необходимо учитывать при построении более эффективных методов и алгоритмов решения соответствующих классов задач по сравнений с традиционными универсальными алгоритмами нелинейного программирования;
   - применяемые в оптимизационных задачах 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 конечномерной оптимизации. Большое число прикладных задач 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 *данной главе* рассмотрена единая стратегия решения сформулированных в пункте 4 задач на основе среднестепенных целевых функционалов, традиционно применяемых при решении задач аппроксимации непрерывных зависимостей.

Глава 4

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 (жестких) экстремальных задач в практической оптимизации.

4.1. Явление жесткости (овражности)

Рассмотрим следующий пример критерия оптимальности, зависящего от двух управляемых параметров $x_1,\ x_2$ [57]:

$$J\ x_1,\ x_2\ = g_0^2\ x_1,\ x_2\ + \sigma g_1^2\ x_1,\ x_2\ \to \min_x, \tag{4.1}$$

где σ — достаточно большое положительное число. Рассмотрим также уравнение

$$g_1\ x_1,\ x_2\ = 0, \tag{4.2}$$

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\ = \partial J/\partial x_1,\ \partial J/\partial x_2\ $, 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. 4.1. Там же стрелками показано векторное поле антиградиентов, определяющее локальные направления наискорейшего убывания $J\ x$.

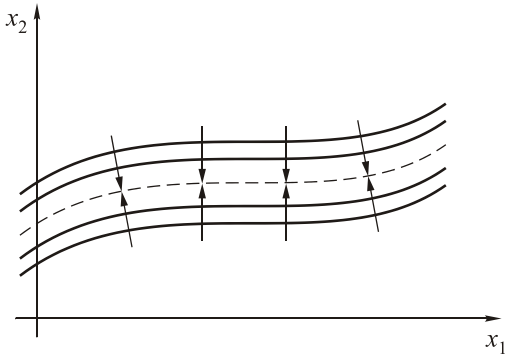


Рис. 4.1. Явление жесткости

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 (4.1) 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 (жесткой) 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. 4.1 следует, что направления поиска, задаваемые антиградиентами, оказываются неэффективными. Приводя достаточно 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 (4.2) 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:

$$J\ x = g_0^2\ x + \sigma \sum_{i=1}^{m} g_i^2\ x\ ,\ x \in R^n,\ n > 2,\ m < n. \qquad (4.3)$$

В этом случае дно оврага задается системой уравнений

$$g_i\ x\ = 0,\ i \in 1:m\ , \tag{4.4}$$

что в принципе позволяет выразить m параметров x_i через оставшиеся $n-m$ переменных. Предположим, не ограничивая общности, что уравнения (4.4) определяют следующие зависимости:

$$x_1 = \varphi_1\ x_{m+1},\ ...,\ x_n\ ,$$

. . . . . .

$$x_m = \varphi_m\ x_{m+1},\ ...,\ x_n\ .$$

Аналогично предыдущему случаю устанавливаем, что задача минимизации (4.3) при достаточно больших σ эквивалентна минимизации функции $g_0^2\ \varphi_1,\ ...,\ \varphi_m,\ x_{m+1},\ ...,\ x_n$ от $r \triangleq n-m$ переменных $x_{m+1},\ ...,\ x_n$. 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 (4.1), (4.3), явно содержащие большой параметр σ. Можно привести следующий пример квадратичного функционала:

$$f\ x_1,\ x_2\ = 0,250025x_1^2 + 0,49995x_1x_2 + 0,250025x_2^2 - x_1 - x_2. \tag{4.5}$$

Линии уровня $f\ x\ = \text{const}$ функционала (4.5) 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 (4.5) нельзя, хотя овражная ситуация налицо, и так же как и в предыдущих случаях явно выделяется дно оврага (прямая *ab* на рис. 4.2), имеющее уравнение $x_2 = 2 - x_1$. Подставляя выражение для x_2 в (4.5), 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 (жестких) оптимизационных задач в отдельный класс обусловлена, с одной стороны, значительными вычислительными трудностями при их решении стандартными для компьютерного моделирования мето-

дами, 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:

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

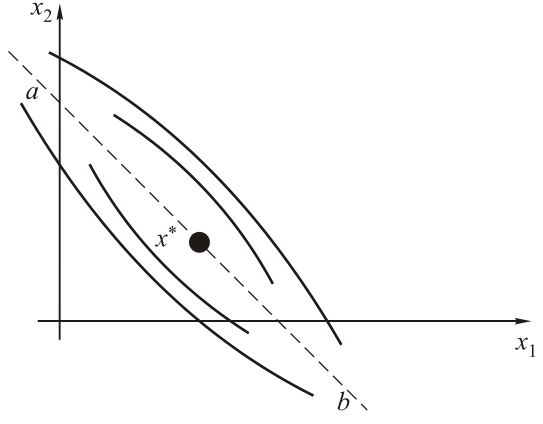


Рис. 4.2. Жесткий квадратичный функционал

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 жестких и одновременно невыпуклых функционалов, т. к. именно в этой ситуации отказывает большинство методов поисковой оптимизации.

4.2. 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

$$\frac{dx}{d\tau} = -J'\ x\ ,\ J'\ x\ \triangleq \left[\frac{\partial J}{\partial x_1}, \ldots, \frac{\partial J}{\partial x_n}\right]. \tag{4.6}$$

Эти траектории в случае жестких функционалов обладают рядом специфических черт. Например, для функционала (4.5) имеем

$$x\ \tau\ = \sum_{i=1}^{2}\ \alpha_i^* + \ \alpha_i^0 - \alpha_i^*\ \exp\ -\lambda_i \tau\ \ u_i, \tag{4.7}$$

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 (4.7) 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 $dx_i\ \tau\ /d\tau$ и означает спуск на дно оврага. На дне выполняются условия типа (4.2) и норма вектора градиента, а с ней и производные $dx_i\ \tau\ /d\tau$ 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 [57] предложено следующее общее определение.

Определение 1. Функционал $J\ x\ $ называется *овражным*, или *жестким*, если отвечающая ему система дифференциальных уравнений (4.6) — жесткая.

Однако с точки зрения приложений к области оптимизации более конструктивным оказывается приведенное далее определение жесткого функционала. Это определение не содержит, в частности, таких неестественных для задач оптимизации требований, как необходимость задания промежутка интегрирования уравнения (4.6) [81].

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

$$\begin{aligned} &1)\ \forall x\in Q_\delta,\ \lambda_1\left[J''\ x\ \right]\geq\sigma\left|\lambda_n\left[J''\ x\ \right]\right|;\\ &2)\ \forall x\in Q,\ \mathrm{Arg}\min_{x'\in X_\delta\ x}\ J\ x'\ \subset Q;\\ &3)\ \forall x\in Q,\ L\left[X_\delta\ x\ \cap Q\right]\leq\sigma^{-1}L\left[X_\delta\ x\ \right], \end{aligned} \tag{4.8}$$

где $X_\delta\ x\ =\ x'\in R^n\left|\left\|x'-x\right\|\leq\delta\right.$; $Q_\delta=\bigcup_{x\in Q}X_\delta\ x$; $\lambda_i\ A$ — собственные числа матрицы $A=J''\ x$, упорядоченные по убыванию:

$$\lambda_1\ A\ \geq\lambda_2\ A\ \geq\ldots\geq\lambda_n\ A\ ;$$

$L\ S$ — константа Липшица в соотношении

$$\left\|J'\ x'\ -J'\ x\ \right\|\leq L\ S\ \left\|x'-x\right\|,\ \forall x',\ x\in S\subset R^n.$$

Множество Q называется *дном оврага*.

Основным является условие 1, констатирующее резко несимметричное расположение спектра матрицы вторых производных $J''\ x$ относительно начала координат: $\lambda_i\in\ -m,\ M\ ,\ M\gg m>0$. Условия 2 и 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 [81].

Пусть в $D\subset R^n$ задана r-мерная поверхность (конфигурационное пространство) $Q=\ x\in D\left|g_i\ x\ =0;\ i\in\ 1:n-r\right.\ ,\ g_i\in C^2\ D\ .$

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

$$\begin{aligned} 1)\ \forall x\in Q_\delta,\ \lambda_1\left[J''\ x\ \right]\geq\ldots\geq\lambda_{n-r}\left[J''\ x\ \right]\geq\sigma\left|\lambda_{n-r+1}\left[J''\ x\ \right]\right|\geq\\ \geq\ldots\geq\sigma\left|\lambda_n\left[J''\ x\ \right]\right|; \end{aligned}$$

$$\begin{aligned} &2)\ \forall x \in Q,\ \operatorname*{Arg\,min}_{x' \in X_\delta\ x} J\ \ x' \ \subset Q; \\ &3)\ \forall x \in Q,\ L\left[X_\delta\ \ x\ \cap Q \right] \le \sigma^{-1} L\left[X_\delta\ \ x\ \right]. \end{aligned} \tag{4.9}$$

Число r называется размерностью оврага (дна оврага) Q.

Пример. Рассмотрим квадратичный функционал

$$f\ \ x\ = \frac{1}{2}\left\langle Ax,\ x \right\rangle - \left\langle b,\ x \right\rangle + c, \quad c = \text{const}. \tag{4.10}$$

Пусть собственные числа $\lambda_i\ \ A\ ,\ i \in 1:n$ удовлетворяют неравенствам (4.9) (при $\forall x \in R^n$). Построим множество Q, являющееся по определению дном оврага. Предположим, без существенного ограничения общности, что $\det A \ne 0$, и обозначим через x^* решение уравнения $Ax = b$. Тогда

$$f'\ \ x\ = Ax - b = A\ \ x - x^*\ = \sum_{i=1}^{n} \beta_i \lambda_i u_i, \tag{4.11}$$

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

Положим

$$Q = \ \ x \in R^n \,\middle|\, \beta_i = \left\langle x - x^*,\ u_i \right\rangle = 0,\ i \in\ 1:n-r\ \ . \tag{4.12}$$

Тогда, очевидно, $Q_\delta = R^n$, $\sigma \cong \lambda_1 / \left| \lambda_{n-r+1} \right|$ и условие 1 из (4.9) выполнено.

Обратимся к проверке выполнения условия 2 из (4.9). Рассмотрим сферическую окрестность

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

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

$$\min_{\beta \in S_\beta} \varphi\ \beta\ = \min_{\overline{\beta} \in \overline{S}_\beta} \overline{\varphi}\ \overline{\beta}\ , \tag{4.13}$$

где

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

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

Из (4.13) следует условие 2 в соотношении (4.9). 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

$$\left\| J'\ x'\ - J'\ x''\ \right\| = \left\| A\ x' - x''\ \right\| = \left\| \sum_{i=1}^{n} \beta'_i - \beta''_i\ \lambda_i u_i \right\| \le \max_i \lambda_i \left\| \sum_{i=1}^{n} \beta'_i - \beta''_i\ u_i \right\| =$$

$$= \lambda_1 \left\| x' - x'' \right\|.$$

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

$$\left\| J'\ \overline{x}'\ - J'\ \overline{x}''\ \right\| \le \left| \lambda_{n-r+1} \right| \cdot \left\| \overline{x}' - \overline{x}'' \right\|.$$

Поэтому $L\ X_\delta = \lambda_1 \gg L\ X_\delta \cap Q = \lambda_{n-r+1}$ и условие 3 выполнено при $\sigma = \lambda_1 / |\lambda_{n-r+1}|$.

Таким образом, для квадратичных функционалов при сдвинутом в точку x^* начале координат дно оврага Q совпадает с линейной оболочкой (4.12) собственных векторов, отвечающих "малым" собственным числам. Это согласуется с интуитивными представлениями, развитыми в *разд. 4.1*.

На этом же примере можно проиллюстрировать роль требований 1—3 в определениях 2 и 3. Действительно, для квадратичного функционала (4.10) условия 1 и 2 из соотношения (4.9) могут выполняться для всего пространства $Q = R^n$ и любого $\sigma > 0$. Необходимая линейная оболочка собственных векторов может быть выделена только при дополнительном требовании, эквивалентном условию 3. В то же время требования 1 и 3 без условия 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 условий 1 и 3, а условие 2 при этом нарушится.

Рассмотренные ранее модели явления жесткости не являются исчерпывающими. Однако они описывают наиболее существенные стороны большинства практических ситуаций, связанных с решением задач конечномерной оптимизации в различных предметных областях.

Определение 4. Пусть $\forall x \in Q$, $\det J''\ x \neq 0$. Наименьшее из чисел σ, удовлетворяющих определению 2, называется *степенью жесткости (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 ∞.

Если $J''\ x > 0$, то $\eta\ x = \mathrm{cond}\left[J''\ x \right] \triangleq \max_i \lambda_i\ x \Big/ \min_i \lambda_i\ x$. В общем случае справедливо неравенство $1 \leq \eta \leq \mathrm{cond}\left[J''\ x \right]$. При наличии больших по модулю отрицательных собственных чисел $\lambda_i\ x$, т. е. при отсутствии овражной ситуации, возможно неравенство $\eta\ x \ll \mathrm{cond}\left[J''\ x \right]$. Следовательно, из высокой степени жесткости $J\ x$ в точке x следует плохая обусловленность матрицы $J''\ x$, а обратное неверно. Действительно, пусть спектр матрицы $J''\ x$ расположен в множестве $-M, -m \cup m, M$, $M \gg m > 0$, включая граничные значения. Тогда $\eta\ x = 1$, а $\mathrm{cond}\left[J''\ x \right] = M/m \gg 1$. Данный функционал не будет относиться к классу жестких, что естественно, поскольку трудностей при его минимизации, например, методом наискорейшего спуска, не возникает.

Различие между двумя характеристиками $\eta\ x$ и $\mathrm{cond}\left[J''\ x\right]$ функционала $J\ x$ часто игнорируется, и жесткими (овражными) называются функционалы с большим числом $\mathrm{cond}\left[J''\ x\right]$. Подобная точка зрения не оправдывается с позиций основных вычислительных трудностей, возникающих при решении экстремальных задач. Однако, учитывая указанную выше связь между $\eta\ x$ и $\mathrm{cond}\left[J''\ x\right]$, задачи с жесткими целевыми функционалами 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

$$\log_2 \eta > t, \tag{4.14}$$

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, становится неэффективным.

4.3. Критерии жесткости

Рассмотрим практические методы распознавания овражной ситуации, играющие роль *критериев жесткости (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

$$x\ k+1\ = x\ k\ - hJ'\ x\ k \tag{4.15}$$

с постоянным шагом $h \in R^1$.

Принадлежность $J\ x$ к классу жестких 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 (4.15) протекать без полной остановки, то по результатам работы метода можно количественно оценить величину η.

Для этого процесс (4.15) продолжается до тех пор, пока отношение $\|J'\ x\ k+1\ \|/\|J'\ x\ k\ \|$ не застабилизируется около некоторого значения μ. Тогда справедливо следующее равенство:

$$\eta \cong \frac{2}{|1-\mu|}. \tag{4.16}$$

Соотношение (4.16) справедливо независимо от выпуклости функционала $J\ x$ и является основным для грубой практической оценки степени жесткости решаемой задачи в окрестности текущей точки. Доказательство соотношения (4.16) дано в *разд. 6.2.1*.

В силу вышеизложенного можно рекомендовать процесс оптимизации начинать с помощью метода ПГС. Если задача простая и степень жесткости невелика, то уже этот метод достаточно 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 вычисление матрицы Гессе функционала и решению для нее полной проблемы собственных значений. Тогда на основе непосредственной проверки выполнения неравенств (4.9) 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 $\delta = \varepsilon_м = 2^{-t}$, где $\varepsilon_м$ — относительная машинная точность, а t — длина разрядной сетки мантиссы числа, получим оценку для абсолютных искажений собственных чисел за счет ошибок округления:

$$|d\lambda_i| \le n\varepsilon_м|\lambda_1|. \tag{4.17}$$

Параметр $\varepsilon_м$ известен для каждой вычислительной системы.

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

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 достаточно 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.

4.4. Источники плохо обусловленных оптимизационных задач

Несмотря на то, что типичность овражной ситуации может считаться установленным экспериментальным фактом, представляет определенный интерес выяснение на качественном уровне основных причин появления оврагов в прикладных задачах. Рассмотрим два типа жесткой ситуации — так называемую *естественную жесткость* и *внесенную жесткость*. 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 жесткость 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.

4.4.1. Естественная жесткость

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. разд. 2.2)*

$$\int_0^T \varphi(\tau - \lambda)\, \omega(\lambda)\, d\lambda = R(\tau), \qquad (4.18)$$

играющему важную роль во многих задачах компьютерного моделирования.

Как известно [60], 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 [70]. Для построения решения уравнения (4.18) применяются алгебраические методы, основанные на минимизации невязки.

Пример соответствующего целевого функционала был дан в *разд. 2.3*:

$$J\ x\ =\sum_{j=1}^{N}\left\{\sum_{i=1}^{N}\omega_i\varphi\left[q\ \ j-i\ \right]-R\ qj\right\}^2\rightarrow\min_x, \tag{4.19}$$

где $\omega_i=\omega\ qi\ ,\ i\in 1:N\ ;\ x=\ \omega_1,\ \omega_2,\ ...,\ \omega_N\ .$

Свойство некорректности исходной задачи приводит к плохой обусловленности матрицы $J''\ x$ с резко выраженным овражным характером соответствующих поверхностей уровня. Характер решения задачи (4.19) показан на рис. 4.3. Как отмечается в [24], 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 конечномерной оптимизации, не учитывающих особенностей оптимизационных задач, отмеченных в *разд. 3.9*, амплитуда паразитных колебаний может быть весьма значительной.

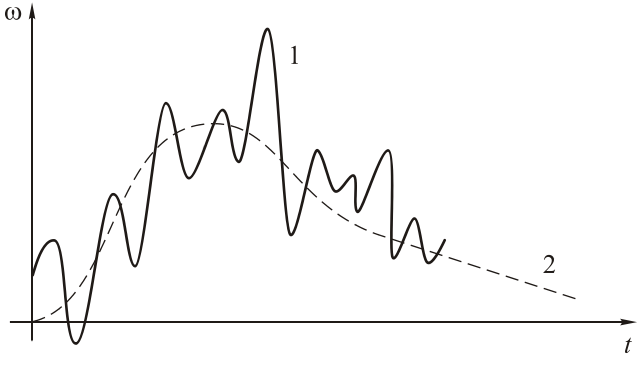


Рис. 4.3. 1 — решение задачи (4.19);
2 — точное решение уравнения Винера — Хопфа

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

$$J_1\ x\ =J\ x\ +\alpha\sum_{j=2}^{N}\ \omega_i-\omega_{i-1}\ ^2\rightarrow\min_x, \tag{4.20}$$

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

Существуют и другие формы регуляризации, связанные с расширением исходных целевых функционалов (2.34) за счет сглаживающих (регуляризующих) функционалов [24]:

$$I\left[\omega\ t\ \right]=M\ \ E_2\ t\ \ +\alpha\Omega\left[\omega\ t\ \right],$$

где

$$\Omega\left[\omega\ t\ \right]=\int_0^T\left\{k\ t\ \left[\frac{d\omega\ t}{dt}\right]^2+p\ t\ \omega^2\ t\right\}dt. \tag{4.21}$$

Условие минимума $I\left[\omega\ t\ \right]$ 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 (4.19) 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 (4.20), также сопряжено с известными трудностями, вызванными некорректностью исходной задачи. Действительно, регуляризующие добавки помимо эффекта сглаживания дают искажение решения, приводя к увеличению невязки $J\ x\ $ [60]. 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 модно 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 $x_1,\ x_2,\ ...,\ x_n,$ а от s агрегатов

$$y=\Phi\ \ z_1,\ z_2,\ ...,\ z_s\ \ , \tag{4.22}$$

где

$$z_i=\varphi_i\ \ x\ \ ,\ i\in 1:s\ ;\ s<n. \tag{4.23}$$

Как показано, например, в [81], 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

$$\varphi_i\ \ x\ =z_i^*,\ i\in 1:s\ , \tag{4.24}$$

где z_i^* — оптимальные значения агрегатов, доставляющих минимум целевому функционалу $I\ z\ =I\left[\varphi\ x\ \right]\triangleq J\ x\ $. Поэтому уравнения (4.24) фактически являются уравнениями дна оврага.

Образование агрегатов можно пояснить на примере описания динамических свойств оптимизируемого объекта. Распространенные методы синтеза линейных оптимальных систем управления 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 (4.24) [37].

Пример 1. Рассмотрим в качестве иллюстрации задачу параметрического синтеза частотного фильтра, схема которого представлена на рис. 4.4 [81].

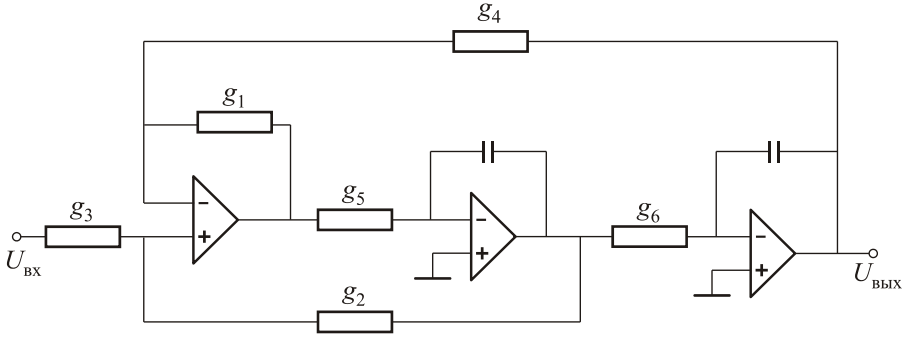


Рис. 4.4. Фильтр нижних частот

Передаточная функция фильтра имеет вид

$$W(j\omega) = \frac{\dot{U}_{вых}}{\dot{U}_{вх}} = \frac{1}{z_3 p^2 + z_2 p + z_1}, \quad p = j\omega, \tag{4.25}$$

где

$$z_0 = g_3 g_5 g_6 (g_1 + g_4); \quad z_1 = \frac{g_4 g_5 g_6 (g_2 + g_3)}{z_0};$$

$$z_2 = \frac{g_2 g_5 (g_1 + g_4) c_2}{z_0}; \quad z_3 = \frac{g_1 c_1 c_2 (g_2 + g_3)}{z_0}.$$

Требуется так подобрать компоненты вектора $x = (c_1, c_2, ..., g_6)$ неизвестных параметров элементов схемы, чтобы оптимизировать отклонение характеристики затухания $a(\omega) = 20 \lg \frac{1}{|W(j\omega)|}$ 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_3. Выберем N_n равноотстоящих точек на промежутке $0, \omega_n$ и N_3 равноотстоящих точек в полосе задерживания $\omega_н, \omega_в$. Тогда задача аппроксимации может быть сформулирована в виде следующей задачи параметрической оптимизации:

$$J\ z\ = \max \left| a\ z, \omega_c\ - a_n \right| \to \min_z, z \in Z \subset R^3, \omega_i \in\ 0, \omega_n\ , i \in 1 : N_n$$

$$Z = \ z \in R^3 \left| z_i \geq 0; a\ z, \omega_j\ \geq a_3, \omega_j \in\ \omega_н, \omega_в\ , j \in 1 : N_3\ \right. .$$

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 (4.25), конкретизирующим (4.24). В результате приходим к следующей задаче параметрической оптимизации:

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

$$D = \ x \in R^8 \left| x_i \geq 0, i \in 1 : 8 \right. .$$

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 вышеизложенному, 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

$$J\ x\ = I\ z_1, z_2, ..., z_s\ + \varepsilon g\ x$$

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 пара-

метра 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 динамических систем 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 [57].

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

Пример 2. Пусть функционирование объекта описывается вектор-функцией

$$u\ t\ = \left[u_1\ t\ + u_2\ t\ \right], \tag{4.26}$$

$$u_1\ t\ = t\left[x_1 \exp\ -At\ +\ x_1 + x_2\ \exp\ -at\ \right],$$

$$u_2\ t\ = t\left[x_1 \exp\ -Bt\ +\ x_1 + x_2\ \exp\ -bt\ \right],$$

являющейся при $A \gg a > 0,\ B \gg b > 0$ 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 $t_0 > \tau_{\text{пс}}$, где $\tau_{\text{пс}}$ — длина пограничного слоя, определяющего практически полное затухание экспонент $\exp\ -At$, $\exp\ -Bt$, то из (4.26) 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

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

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

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

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

4.4.2. Внесенная жесткость

Учет ограничений

Овражная ситуация может быть внесена в задачу оптимизации в силу используемой конструкции обобщенного критерия оптимальности. Рассмотрим эффект жесткости, 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 1. 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 $a_1 = \frac{1}{\sqrt{1+\sigma}};\ a_2 = 1;\ b_1 = \frac{2\sigma}{\sigma+1};\ b_2 = 0;\ d = \frac{4\sigma}{\sigma+1}.$

Уравнение линии уровня $J_0\ x = \text{const}$ является уравнением эллипса с центром в точке $b_1,\ b_2$ и длинами полуосей, относящимися как $a_2 / a_1 = \sqrt{1+\sigma}$. При больших значениях σ, обеспечивающих относительно точное выполнение ограничения $x_1 = 2$, линии уровня оказываются сильно вытянутыми (рис. 4.5). Чем точнее выполняются ограничения, тем ярче выражен эффект жесткости. В данном случае степень жесткости равна $\eta = 1 + \sigma$ и стремится к бесконечности при $\sigma \to \infty$. Заметим, что линии уровня исходного функционала являются сферами, и явление жесткости отсутствует.

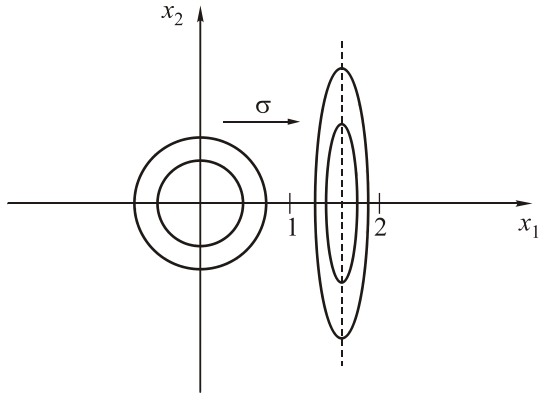


Рис. 4.5. Метод штрафных функций

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 конечномерной оптимизации.

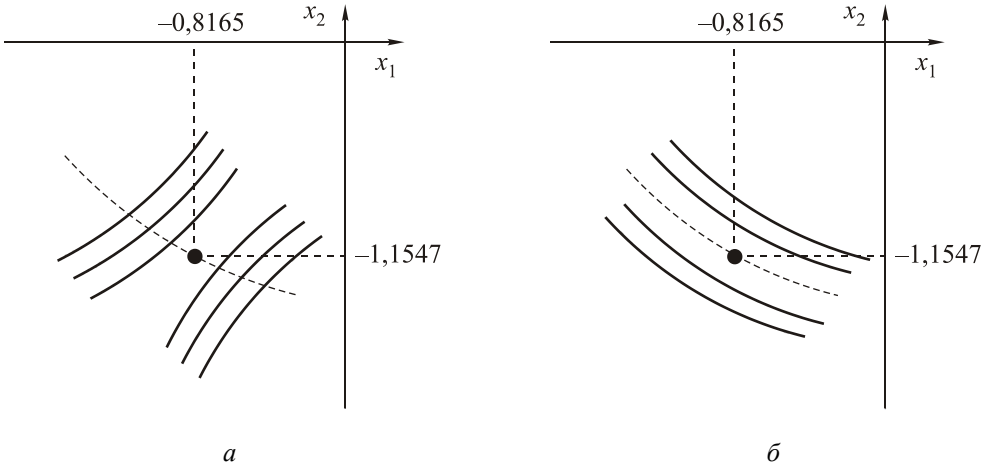
Например, при наличии вырожденного минимума, когда не выполняется условие линейной независимости градиентов ограничений $g_i'\ x$, могут потерять работоспособность все методы учета ограничений, основанные на обычной и модифицированной функции Лагранжа, а также на линеаризации ограничений. Метод же штрафных 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 $\left\| g\ x \right\| \to \min$, для которого величина $J\ x$ минимальна. Другие методы либо теряют смысл, либо заведомо не будут сходящимися.

Более подробное обсуждение этих достаточно тонких вопросов содержится в [54].

В силу вышеизложенного наличие алгоритмов оптимизации, сохраняющих работоспособность при достаточно высокой степени жесткости 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 ослаблен-

ной форме. Как показано в [21], 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 отсутствуют, то становится совершенно ясна необходимость применения специальных жестко-ориентированных процедур для решения вспомогательных канонических задач.



а *б*

Рис. 4.6. Линии уровня модифицированной функции Лагранжа:
а — $\sigma = 0,075$; *б* — $\sigma = 100$

Пример 2. Решается задача [21]

$$J\ x\ = x_1 x_2^2 \to \min,\ x \in D,$$

$$D = \ x \in R^n \Big| g\ x\ = 2 - x_1^2 - x_2^2 = 0\ .$$

На рис. 4.6 показаны линии уровня вспомогательного функционала

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

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, обуславливающим возникновение овражной ситуации. Изложенные в *разд. 3.5* 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 нежесткий 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. 4.7, 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.

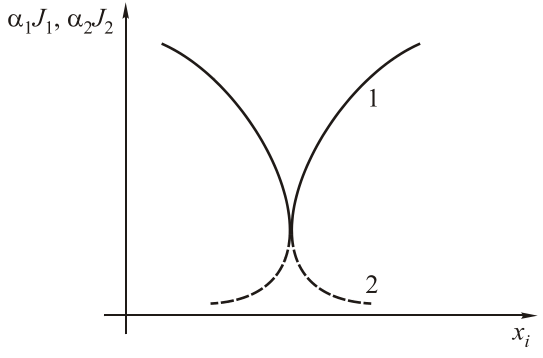


Рис. 4.7. Объединение частных критериев: 1 — $\alpha_1 J_1$; 2 — $\alpha_2 J_2$

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

$$J\ x\ = \sum_{i=1}^{N} \alpha_i f_i\ x\ ,\ f_i\ x\ \triangleq \left[\sum_{j=1}^{n} x_j \varphi_j\ t_i\ - y_i \right]^2, \qquad (4.27)$$

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

$$y\ t\ \cong \sum_{j=1}^{n} x_j \varphi_j\ t\ ,\ 0 \le t \le T, \qquad (4.28)$$

характерных для многих разделов прикладного численного анализа. Например, необходимость в применении параметризации (4.28) возникает в задачах сглаживания экспериментальных зависимостей при идентификации стохастических объектов, а также в задачах идентификации нелинейных объектов на основе функциональных рядов Вольтерра *(см. разд. 2.3)*. 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 (4.27) 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

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

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 — матрица Гильберта размером $n \times n$, являющаяся также матрицей линейной системы так называемых *нормальных уравнений* $J'\ x\ = 0$. Спектральное число обусловленности положительно определенной матрицы H_n, совпадающее со степенью жесткости квадратичного сильно выпуклого функционала $J\ x\ $, оказывается достаточно высоким: матрицы Гильберта являются стандартным примером очень плохо обусловленных матриц. При этом $\mathrm{cond}\ H_n$ быстро растет с увеличением n:

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

В результате уже для небольших n минимизация (4.27) происходит в условиях сильно выраженной овражной ситуации.

4.5. Методы конечномерной оптимизации

В *разд. 3.9* сформулированы особенности канонических оптимизационных задач, возникающих во многих практических ситуациях. В связи с этим далее обсуждаются возможности некоторых популярных методов конечномерной 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, т. к. этот материал достаточно полно представлен в литературе.

4.5.1. Ньютоновские методы

Ньютоновские методы (Н-методы) генерируют последовательность точек x^k по правилу

$$x^{k+1} = x^k + h_k p_N^k,\ h_k \in R^1, \tag{4.29}$$

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 (4.29) предполагается, что все матрицы G_k положительно определены ($G_k > 0$). Последнее условие гарантирует разрешимость задачи вычисления p_N^k исходя из системы уравнений

$$G_k p_N^k = -g^k. \tag{4.30}$$

Вектор x^{k+1}, построенный согласно (4.29) и (4.30) при $h_k = 1$, является минимизатором аппроксимирующего квадратичного функционала *(параболоида)*, являющегося отрезком соответствующего ряда Тейлора:

$$F_k\ \ x\ \ = J\ \ x^k\ \ + \left\langle g^k,\ x - x^k \right\rangle + \frac{1}{2}\left\langle G_k\ \ x - x^k\ \ ,\ x - x^k \right\rangle. \tag{4.31}$$

Поэтому метод (4.29) называется также методом параболоидов [50].

Н-методы (4.29) 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 (4.31) может не иметь конечного минимизатора и, как правило, будет неограничен снизу. Поэтому для невыпуклых целевых функционалов процедура (4.30) 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 [21], "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

$$\bar{G}_k p^k = -g^k, \tag{4.32}$$

где $\bar{G}_k$ — 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 $\bar{G}_k$ на основе различных матричных разложений, позволяющих учитывать знаки собственных чисел G_k.

Методы, основанные на спектральном разложении

Спектральное разложение имеет вид

$$G_k = U\Lambda U^T = \ g_{ij}\ ,\ g_{ij} = \sum_{m=1}^{n} \lambda_m u_i^m u_j^m, \tag{4.33}$$

где $U = \left[u^1,\ u^2,\ ...,\ u^n \right]$; $\Lambda = \mathrm{diag}\ \lambda_i\ ;\ \ u^m,\ m \in 1:n$ — 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

$$\bar{G}_k = U\bar{\Lambda}U^T,\ \bar{\Lambda} = \mathrm{diag}\left[\max\ \left|\lambda_i\right|,\ \delta\ \right], \tag{4.34}$$

где $\delta > 0$ — параметр метода, определяющий границу *"существенной положительности"* любого из собственных чисел. Отмечаемый в литературе недостаток метода (4.32) с матрицей (4.34) связывается с трудоемкостью процедуры построения спектрального разложения (4.33), требующей до $4n^3$ арифметических операций. Кроме того, 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

$$B = LDL^T, \ D = \text{diag } d_i \ , \tag{4.35}$$

где L — нижняя треугольная матрица с единицами на диагонали; D — положительная диагональная матрица. Факторизация (7) непосредственно неприменима к знаконеопределенной симметричной матрице $B = G_k$. В [21] предложена модифицированная процедура, позволяющая равномерно ограничить рост элементов треугольного фактора L на уровне $|r_{ik}| \le \beta \ \ i > k, \ r_{ik} \triangleq l_{ik}\sqrt{d_k}$ и гарантирующая "существенную" положительность $d_i > \delta > 0$ диагональных элементов матрицы D. В результате получаем

$$\bar{G}_k = LDL^T = G_k + C, \quad \bar{G}_k > 0, \tag{4.36}$$

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

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

где $\varepsilon_м$ — машинное эпсилон; ξ — максимальный модуль недиагонального элемента G_k; γ — значение максимального из диагональных элементов G_k.

Основная цель построения разложения (4.36) заключается в сокращении вычислительных затрат по сравнению с (4.34) 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 и многих работах других исследователей. Большое число публикаций, посвященных рассматриваемым вопросам, с одной стороны, указывает на актуальность проблемы, а с другой — на отсутствие метода, полностью отвечающего предъявляемым практикой требованиям.

4.5.2. Методы доверительной окрестности

Основная идея методов доверительной окрестности (МДО) сводится к следующему [21]. Мы рассматриваем *квадратичные* методы, т. е. 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 иначе, 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

$$\frac{1}{2}\langle G_k p,\ p\rangle + \langle g^k,\ p\rangle \to \min,\ p \in D \subset R^n,$$
$$D = \left\{ p \in R^n \,\middle|\, \langle p,\ p\rangle^2 \le \Delta \right\}. \tag{4.37}$$

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 (4.37) приводит к методу определения p из системы линейных уравнений

$$G_k + \beta E\ \ p = -g^k,\ \beta > 0, \tag{4.38}$$

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 (4.38). В ряде случаев [27] вначале полагают $\beta = 0$. Если в процессе решения (4.38) при $\beta = 0$ выясняется, что G_k знаконеопределена или $\|p\| > \Delta$, где Δ — установленное пороговое значение, то β определяется как решение нелинейного уравнения вида

$$\left\| G_k + \beta E^{-1} g^k \right\| = \Delta. \tag{4.39}$$

Далее по найденному β согласно (4.38) определяется искомый вектор p.

Основной недостаток всех методов рассматриваемого класса состоит в необходимости решения на каждом шаге итерационного процесса нелинейных алгебраических уравнений типа (4.39), что требует привлечения различных процедур регуляризации. Кроме того, найденному оптимальному значению может отвечать очень плохо обусловленная матрица $G_k + \beta E$, что приводит к значительным вычислительным трудностям при решении системы (4.38). Аналогичные недостатки присущи и ньютоновским методам, рассмотренным в 4.5.1.

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 $\bar{G}_k$, генерируемые векторы p^k, оставаясь направлениями спуска, оказываются практически случайными, что резко замедляет сходимость соответствующей поисковой процедуры.

4.5.3. Квазиньютоновские методы

Квазиньютоновские методы (КН-методы), также 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 приводит 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 [54], показывают, что стандартные методы СГ сходятся по закону геометрической прогрессии со знаменателем q, близким к единице: $q \cong 1 - 2/\sqrt{\eta}$, где η — степень овражности минимизируемого функционала. Там же имеются указания на достаточно высокую скорость сходимости метода СГ "по функционалу", независимо от величины η:

$$J\ x^k\ - J\ x^*\ \le \frac{L\left\|x^0 - x^*\right\|^2}{2\ 2k+1\ ^2},\ L = \text{const}. \qquad (4.40)$$

Однако оценки типа (4.40) получены в предположении $G\ x\ >0$. Кроме этого, согласно (4.40) 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. 4.8. 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 [21].

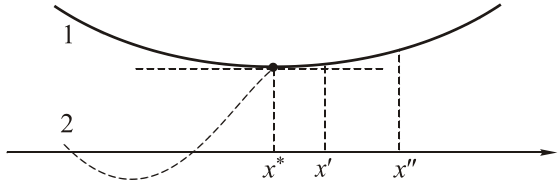


Рис. 4.8. Метод СГ: 1 — аппроксимирующий на участке x', x'' параболоид; 2 — целевой функционал

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.

4.5.4. 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 порядка. Эти трудности могут стать непре-

одолимыми, если функционал характеризуется высокой степенью жесткости 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 других 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, т. к. 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 [21]. Однако при отсутствии выпуклости функционала в окрестности текущей точки генерируемые Н-методами направления спуска оказываются локально неэффективными. Возникающие трудности связаны с указанными в *разд. 4.5.2* 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

$$\bar{G}_k p^k = -g^k \tag{4.41}$$

с разреженной матрицей $\bar{G}_k$. Дело заключается в том, что решение (4.41) обычно проводится на основе некоторой процедуры факторизации положительно определенной матрицы $\bar{G}_k$. Наиболее часто применяется симметричный вариант гауссова исключения, основанный на факторизации Холесского. При этом матрица L в представлении $\bar{G}_k = LL^T$ может оказаться заполненной, несмотря на разреженность исходной матрицы $\bar{G}_k$. Развитые в [25] 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. разд. 4.5.1)*, т. к. в данном случае непосредственно не удается воспользоваться результатами упомянутой работы [25].

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 практических задач конечномерной оптимизации.

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.

4.5.5. Глобальная оптимизация

Все рассмотренные ранее методы являются "локальными", т. к. с их помощью может быть найден один из локальных минимумов функционала $J\ x$. С точки зрения практики представляет интерес поиск глобального минимума, т. е. 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

$$\left|J\ x'\ -J\ x''\ \right| \le L\left\|x'-x''\right\|,\ L=\text{const}\ge 0.$$

Соответствующие методы глобальной оптимизации изложены в [15], [54]. 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 [66].

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 [64].

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 работах [54], [75].

4.5.6. Анализ сложившейся ситуации

Указанные в *разд. 3.9* особенности канонических оптимизационных задач 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 жесткостью. В ряде случаев присоединяется третий фактор, определяемый высокой размерностью вектора аргументов.

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 компонентов, 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 определенного числа компонентов 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, например, в таких задачах управления, как:

- задачи идентификации нелинейных детерминированных объектов с применением рядов Вольтерра;
- задачи идентификации стохастических объектов на основе методов сглаживания;
- задачи параметризации при идентификации существенно нестационарных объектов,

а также в других задачах, основанных на параметрическом представлении искомых непрерывных зависимостей. Регулярные методы решения указанного класса задач в условиях жесткости $J\ x$ будут изложены в последующих главах.

Учитывая изложенное, а также основные характеристики стандартных методов конечномерной оптимизации, рассмотренных в *разд. 4.5.1—4.5.3*, можно поставить следующие важные в прикладном аспекте задачи по совершенствованию имеющегося алгоритмического и программного обеспечения:

- разработка общих процедур конечномерной оптимизации, сохраняющих эффективность в условиях невыпуклости целевых функционалов в предположении их гладкости и "кусочной квадратичности";
- разработка проблемно-ориентированных реализаций алгоритмов, рассчитанных на конкретные схемы конечномерной оптимизации, характерные для отдельных стандартных классов прикладных задач;

- разработка методов решения задач конечномерной оптимизации большой размерности;
- разработка элементов структурного синтеза и принципов построения "минимальных" параметрических представлений искомых непрерывных функций.

Решению указанных задач в предположении высокой степени жесткости целевых функционалов, а также рассмотрению вопросов, связанных с созданием необходимого программного обеспечения, посвящено дальнейшее изложение. Предполагается, что все решаемые задачи конечномерной оптимизации разделены на два класса по степени обусловленности матриц Гессе минимизируемых функционалов. В первый класс входят оптимизационные задачи с "умеренной" степенью жесткости, 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 жесткости. 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 жесткости являются квадратичными и имеют ньютоновский характер, т. к. 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 жесткости 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 жесткости, что и приводит к необходимому вычислительному эффекту.

4.6. Основные результаты и выводы

В данной главе представлены следующие основные результаты.

1. Обоснована необходимость выделения плохо обусловленных (жестких) оптимизационных задач в отдельный класс. Эта необходимость объясняется, с одной стороны, значительными вычислительными трудностями при их решении стандартными средствами, а с другой стороны, экспериментально установленным фактом типичности овражной ситуации для большинства реальных задач.
2. Выявлены и в явном виде сформулированы следующие основные причины появления плохо обусловленных экстремальных задач:
   - частные структурные особенности решаемой задачи, например, некорректность задачи решения уравнения Винера — Хопфа, играющего важную роль во многих разделах общей теории компьютерного моделирования;
   - наличие фактора агрегированности аргументов минимизируемого функционала и отсутствие регулярных методов выделения агрегатов;
   - жесткость систем обыкновенных дифференциальных уравнений, описывающих динамику нестационарных объектов 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, оказывающееся более конструктивным для теории оптимизации по сравнению с общим определением, опирающимся на теорию жестких систем. Введено новое понятие степени жесткости, 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 жесткости 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 жестких 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 жесткости. Построена оценка степени жесткости, позволяющая по результатам работы метода ПГС указать нижнюю границу для степени жесткости, что имеет первостепенное значение при оценке трудности решаемой задачи и при

автоматизации процесса выбора алгоритма конечномерной оптимизации. Кроме того, предложена альтернативная методика оценки степени жесткости 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 (жестких) задач, сохраняющих эффективность в условиях невыпуклости целевых функционалов;
   - построение проблемно-ориентированных реализации алгоритмов, рассчитанных на конкретные схемы конечномерной оптимизации, характерные для отдельных классов задач конкретной предметной области;
   - построение методов решения плохо обусловленных задач конечномерной оптимизации большой размерности;
   - разработка элементов структурного синтеза и принципов построения минимальных параметрических представлений искомых непрерывных зависимостей на основе методов удаления переменных.

Глава 5

Покоординатные стратегии

5.1. 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 (жестких) экстремальных задач этот метод применим в специальных случаях ориентации оврагов вдоль координатных осей. Трудности применения процедур, построенных на основе метода ЦПС, проиллюстрированы на рис. 5.1. 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.

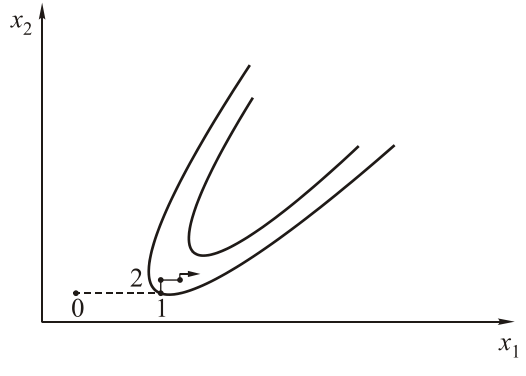


Рис. 5.1. Траектория спуска метода ЦПС при наличии овражной ситуации

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 счисления (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. 5.2.

Ситуация заклинивания представлена на рис. 5.3. 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.

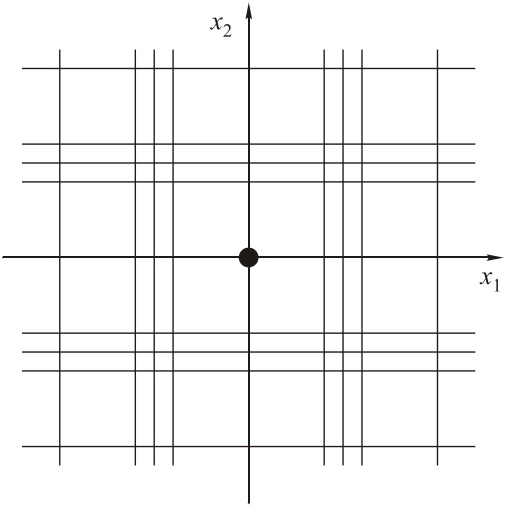


Рис. 5.2. Представление плоскости R^2 в памяти компьютера

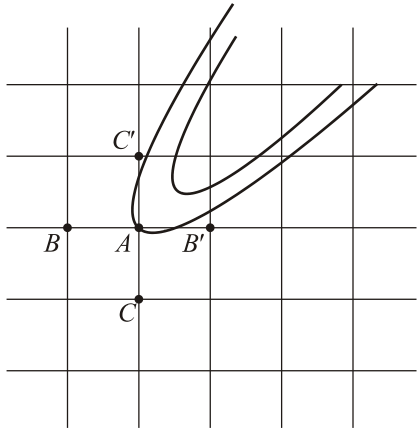


Рис. 5.3. Заклинивание метода ЦПС

Вероятность заклинивания резко возрастает при использовании негладких функционалов, имеющих минимаксную или максиминную структуру. Типичный случай показан на рис. 5.4.

Такого типа ситуации, по существу, не являются следствием высокой степени жесткости и могут быть устранены переходом к гладким аппроксимациям, построенным по методике, изложенной в *разд. 3.10*. В дальнейшем будет приведен пример практической ситуации, иллюстрирующий данное замечание.

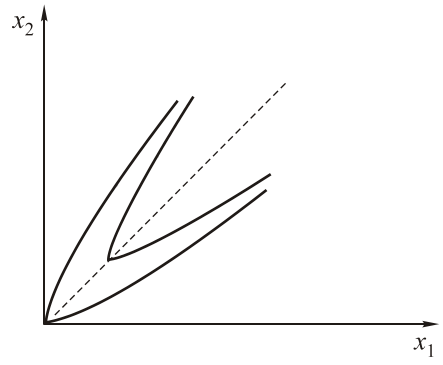


Рис. 5.4. Заклинивание в точке "излома" линии уровня

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, т. е. 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 [62].

Задается вектор начальных шагов $h = h_1, ..., h_n$ продвижений из точки x в направлении координатных ортов $e^1, e^2, ..., e^n$. 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 [83].

Шаг 1. Ввести начальную точку $x = x_1, ..., x_n$ и шаг s; положить $F := J\ x$.

Шаг 2. Положить $h_i := s,\ i \in 1:n$.

Шаг 3. Положить $m := 1$.

Шаг 4. Положить $x_m := x_m + h_m$; вычислить $F_1 = J\ x$.

Шаг 5. Если $F_1 \le F$, положить $h_m := 3h_m$, $F := F_1$ и перейти к шагу 7; иначе — перейти к шагу 6.

Шаг 6. Положить $x_m := x_m - h_m$, $h_m := -0,5h_m$.

Шаг 7. Положить $m := m + 1$. 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 [62]. 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, т. к. 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.

5.2. Методы обобщенного покоординатного спуска

Пусть решается задача $J\ x\ \to \min$, $x \in R^n$, $J \in C^2\ R^n$ с жестким (в смысле определения из *разд. 4.2*) 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. Давались и объяснения этих неудовлетворительных результатов. Они основаны 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 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 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 [83, 85].

Теорема 5.1. Пусть A — симметричная матрица размером $n \times n$, u^i — ортонормированные собственные векторы, λ_i — собственные значения. Тогда при $\lambda_i \neq \lambda_j$ с точностью до величин второго порядка малости имеем

$$\left\langle u^i + du^i,\, u^j \right\rangle = \left(\lambda_i - \lambda_j\right)^{-1} \left\langle\, dA\ u^i,\, u^j \right\rangle, \tag{5.1}$$

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

$$dA\ u^i + A du^i = \lambda_i du^i + d\lambda_i u^i.$$

Отсюда, умножая обе части равенства скалярно на u^j, имеем

$$\left\langle\, dA\ u^i,\, u^j \right\rangle + \left\langle A du^i,\, u^j \right\rangle = \left\langle \lambda_i du^i,\, u^j \right\rangle + \left\langle d\lambda_i u^i,\, u^j \right\rangle$$

В силу равенств $\left\langle u^i,\, u^j \right\rangle = 0$, $\left\langle A du^i,\, u^j \right\rangle = \lambda_j \left\langle du^i,\, u^j \right\rangle$ имеем

$$\lambda_i - \lambda_j \left\langle du^i,\, u^j \right\rangle = \left\langle\, dA\ u^i,\, u^j \right\rangle,$$

откуда следует (5.1). Теорема доказана.

Пусть теперь

$$M_1 = \sum_{i=1}^{n-r} \alpha_i u^i, \quad M_2 = \sum_{j=n-r+1}^{n} \alpha_j u^j$$

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 $|\lambda_i| \gg |\lambda_j|$, то из (5.1) следует $\langle u^i + du^i, u^j \rangle \cong 0$ при достаточно малой величине $\|dA\|/|\lambda_i|$. 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 (5.1), 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

$$f\ x\ = \frac{1}{2}\langle Ax,\ x\rangle - \langle b,\ x\rangle + c \tag{5.2}$$

исходного функционала $J\ x$ в окрестности точки $x \in Q$. Допустим, что известны матрица *A* и ортогональная матрица *U*, приводящая ее к диагональному виду $U^T AU = \mathrm{diag}\lambda_i$. Тогда замена переменных $x = Uy$ приводит квадратичный функционал к сепарабельному виду

$$f\ x\ = f\ Uy\ = \sum_{i=1}^{n} f_i\ y_i\ , \tag{5.3}$$

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, ибо связь между аргументами y_i фактически исчезает. В оказанной 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 диагонализации. 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 (5.3). Для изучения создавшейся ситуации важное значение имеет следующая теорема.

Теорема 5.2. Пусть собственные значения λ_i и отвечающие им ортонормированные собственные векторы u^i , $i\in 1:n$ некоторой симметричной матрицы A разделены произвольным образом на p групп $\lambda_i^t,\ u^{it},\ i\in 1:k_t$, $t\in 1:p$, $\sum_{t=1}^{p}k_t=n$ так, что $u^{it'}\neq u^{it''},\ t'\neq t'',\ \ i\in 1:k_{t'}$, $j\in 1:k_{t''}$, где $\lambda_j^t,\ u^{jt}$ 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 $\overline{W}^T A\overline{W},\ W=\ w^{11},\ \ldots,\ w^{k_1 1},\ \ldots,\ w^{k_p p}$, $\overline{W}=WP$ приводит матрицу A к блочно-диагональному виду $\overline{W}^T A\overline{W}=\mathrm{diag}\ A_1,\ A_2,\ \ldots,\ A_p$, $\overline{W}^T=\overline{W}^{-1}$ с квадратными $(k_t\times k_t)$ матрицами A_t на главной диагонали;
2. собственные значения матрицы A_t есть $\lambda_i^t,\ i\in 1:k_t$, $t\in 1:p$.

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:

$$\overline{W}^T A\overline{W}=\mathrm{diag}\ \lambda_1^1,\ \lambda_1^2,\ \ldots,\ A_m,\ \ldots,\ \lambda_1^p\ .$$

Учитывая, что преобразование подобия не изменяет спектр матрицы, приходим к требуемому заключению. Теорема доказана.

Теорема 5.3. Пусть $V = \Lambda\ A$. Тогда:

1. Замена переменных $x = Vy$ с точностью до нумерации компонентов вектора y приводит функционал $f\ x$ (5.2) к блочно-сепарабельному виду

$$f\ x = f\ Vy = f_s\ y = \sum_{t=1}^{p} f_t\ y^t\ , \tag{5.4}$$

где

$$y = y_1, ..., y_n = y^1, ..., y^p\ ,\ y^t = y_1^t, ..., y_{k_t}^t\ ,$$

$$f_t\ y^t \triangleq \frac{1}{2}\left\langle A_t y^t, y^t \right\rangle - \left\langle b^t, y^t \right\rangle + c_t,\ c_t \in R^1.$$

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

$$f\ x = \frac{1}{2}\left\langle V^T AVy, y \right\rangle - \left\langle V^T b, y \right\rangle + c = \frac{1}{2}\left\langle P^T \overline{W}^T A\overline{W}Py, y \right\rangle - \left\langle P^T \overline{W}^T b, y \right\rangle + c =$$
$$= \frac{1}{2}\left\langle \overline{W}^T A\overline{W}z, z \right\rangle - \left\langle \overline{W}^T b, z \right\rangle + c,\ \ z \triangleq Py.$$

Согласно предыдущей теореме, матрица $\overline{W}^T A\overline{W}$ имеет блочно-диагональную структуру, что и доказывает первое утверждение. Второе утверждение есть прямое следствие второго утверждения той же теоремы.

Следствие. Пусть собственные числа матрицы A удовлетворяют неравенствам $\lambda_1 \geq \lambda_{n-r} \gg |\lambda_{n-r+1}| \geq ... \gg |\lambda_n|$. Тогда:

1. Замена переменных $x = Vy,\ V = \Lambda\ A$, где $V = v^{11}, ..., v^{1n-r}, v^{21}, ..., v^{2r}\ ^T$,

$$v^{1i} = \sum_{m=1}^{n-r} \alpha_{mi}^1 u^m;\ \ v^{2i} = \sum_{m=1}^{r} \alpha_{mi}^2 u^{n-r+m}$$

с точностью до нумерации компонентов вектора y приводит $f\ x$ к виду

$$f_s\ y = f_1\ y^1 + f_2\ y^2\ ,\ y = y^1, y^2\ , \tag{5.5}$$

где $y^1 = y_1, ..., y_{n-r}\ ,\ y^2 = y_{n-r+1}, ..., y_n\ .$

2. $\eta_1 \ll \eta,\ \eta_2 \ll \eta,$ где η_i — показатели жесткости 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 . Представле-

ние (5.5) реализует принцип частично локальной декомпозиции и является аналогом идеализированного соотношения (5.3).

Если собственные числа матрицы квадратичного функционала разделяются более чем на две группы, то будет справедливо представление (5.5), содержащее соответствующее число слагаемых.

Согласно (5.5), появляется возможность независимого решения не связанных между собой оптимизационных задач для функционалов f_i с невысокими показателями жесткости.

Полученные результаты носят локальный характер и справедливы в рамках квадратичной аппроксимации исходного функционала $J\ x$. Для неквадратичных функционалов приближенное выполнение соотношений типа (5.5) 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 [6].

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 $d^1, ..., d^n$, начиная из точки x^k. Предполагается, что матрица $D = d^1, ..., d^n$ 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, т. к. сходится базовый алгоритм A.

Лемма 5.1. Пусть $J\ x \in C^1\ R^n$ и пусть задана последовательность точек $x^k,\ k = 1, 2, ...$ 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 [31]. В лемме утверждается, что если последовательность значений функционала J_k является монотонно невозрастающей и ее под-

последовательность J_k, $k \in K$ сходится к некоторому значению J, то вся последовательность J_k сходится к этому же значению.

Теорема 5.4. 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 $x \in X^*$, т. е. $J'\ x\ = 0$.

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 доказанной леммы 5.1 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 $q^1, ..., q^n$. Поэтому, согласно четвертому предположению из равенства $J(x) = J(x')$ следует $x = x'$. Это значит, что убывания J нет ни в одном из направлений q^i, т. е. 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 $q^1, ..., q^n$, которые позволяют получить точку x' из x с помощью последовательной минимизации по этим направлениям.

Пусть D_k — матрица размером $n \times n$, столбцами которой являются построенные на k-й итерации направления спуска d^{ik}, $i \in 1:n$; т. е. $x^{k+1} = x^k + D_k \lambda^k$, где $\lambda^k = (\lambda_{k1}, ..., \lambda_{kn})^T$ — вектор значений шагов вдоль направлений $d^{1k}, ..., d^{nk}$, 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

$$J(y^{j+1,k}) \le J(y^{jk} + \tilde{\lambda} d^{jk}), \quad \tilde{\lambda} \in R^1, \quad j \in 1:n. \tag{5.6}$$

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 $\det Q = 1 \ne 0$, т. к. для всех k $\det D_k = 1$. Имеем, таким образом,

$$x^k \to x,\ x^{k+1} \to x',\ D_k \to Q,\ k \in K''. \tag{5.7}$$

Поэтому из соотношения $x^{k+1} = x^k + D_k \lambda^k$ имеем

$$\lambda^k \to \lambda = Q^{-1}(x' - x). \tag{5.8}$$

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

$$J(y^{j+1}) \le J(y^j + \tilde{\lambda} q^j), \quad \tilde{\lambda} \in R^1, \quad j \in 1:n. \tag{5.9}$$

Из (5.7) и (5.8) следует, что $y^{jk} \to y^j$, $j \in 1:n+1$ при $k \in K''$. Поэтому в силу непрерывности J неравенство (5.9) следует из (5.6). 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, например в [54], 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.

5.3. Реализация методов обобщенного покоординатного спуска

При реализации методов ОПС в первую очередь необходимо учитывать три фактора, связанных с нормализацией основных переменных задачи, построением аппроксимации матрицы $G(x)$, а также с выбором метода диагонализации матрицы $G(x)$.

5.3.1. 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 $x = x_1, ..., x_n$ к новым $y = y_1, ..., y_n$ осуществляется с помощью замены вида

$$x = Dy, \quad D = \text{diag } d_i \ . \tag{5.10}$$

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

$$d_i = \min\ \delta_2, \max\ |\tilde{x}_i|, \delta_1 \quad , \delta_i > 0, \tag{5.11}$$

где $\tilde{x}_i$ — лучшее значение i-ой переменной, полученное к началу следующего цикла; $\delta_1 \cong \varepsilon_м$ — заданное пороговое значение, исключающее появление нулевых d_i; $\delta_2 \cong \varepsilon_м^{-1}$. 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

$$d_i = k\gamma_i, \quad \gamma_i = \left(\left| \frac{\partial J}{\partial x_i} \right| + \varepsilon_м \right)^{-1} + \varepsilon_м, \tag{5.12}$$

где, например, $k = 1$. В этом случае (при отсутствии влияния $\varepsilon_м$) будем иметь

$$\frac{\partial J\ \ Dy}{\partial y} = \frac{D\partial J}{\partial x} = \ 1, \ ..., \ 1\ ^T.$$

При реализации масштабов (5.12) 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 конечномерной 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, целесообразно 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 (5.11), затем промасштабировать значения $J\ x$, а далее перейти к уравниванию производных (5.12).

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. Укажем два характерных случая.

Первый случай связан с нормализацией некоторой независимой переменной *t* в задачах аппроксимации по среднестепенным критериям *(см. разд. 3.10)* 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 остановок алгоритма, катастрофически теряется точность. Проблема снимается, если нормализовать переменную *t* (а тем самым и все x_i) с помощью соотношения $t = T\tau$, $T \gg 1$, где τ — новая независимая переменная, $\tau \in \tau', \tau''$, $\tau' = t'/T$, $\tau'' = t''/T$. 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 $\Delta J = \Delta I = I\ x + \Delta x\ - I\ x\ \ll c.$ Существенно, что общая нормализация значений $J\ x$ согласно *разд. "Нормализация значений минимизируемого функционала" ранее в этой главе* не изменяет ситуацию к лучшему.

Нормализация ограничений

Рассмотрим общий вид ограничений — неравенств

$$g\ x\ \le 0,\ g\ x\ =\ g_1\ x\ ,\ g_2\ x\ ,\ ...,\ g_s\ x\ \ .$$

Основная цель нормализации ограничений состоит в достижении их сбалансированности. Предполагается, что нормализация управляемых параметров x произведена. Переход к нормализованному варианту ограничений $\alpha^T g\ x\ \le 0,\ \ \alpha_i > 0$ ($i \in\ 1:s\ $) может быть выполнен, согласно условию, аналогичному (5.12):

$$\alpha_i = \left[\left\| \frac{\partial g_i}{\partial x} \right\| + \varepsilon_{\text{м}} \right]^{-1} + \varepsilon_{\text{м}}.$$

Оценка соответствующих градиентов может производиться на основе грубых односторонних приращений независимых переменных x_i.

Более тонкие проблемы, связанные с масштабированием ограничений, рассмотрены в [21].

5.3.2. Методы диагонализации

В качестве основной процедуры приведения симметричной матрицы к главным осям в методах ОПС может быть использован метод Якоби [72], несмотря на наличие конкурирующих, вообще говоря, более эффективных вычислительных схем [51]. Данный 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, т. к. предотвращается возможность вырождения базиса, существующая, например, в методе Пауэлла [80]. 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 и ненадежных вычислительных схемах. В-третьих, алгоритмы, основанные на методах Якоби, часто оказываются наиболее доступными, т. к. соответствующие реализации имеются в большинстве вычислительных систем. И, наконец, определенное влияние на выбор алгоритма оказала простота логики метода Якоби, что приводит к компактности и надежности 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 [51].

В методе Якоби исходная симметричная матрица A приводится к диагональному виду с помощью цепочки ортогональных преобразований вида

$$A_{k+1} = U_k^T A_k U_k,\ A_0 = A,\ k = 1,\ 2,\ ..., \tag{5.13}$$

являющихся преобразованиями вращения. В результате надлежащего выбора последовательности U_k получаем $\lim A_k = D = U^T A U,\ k \to \infty$, где $D = \mathrm{diag}\ \lambda_i$ — диагональная матрица; $U = U_0 U_1 U_2 ...$ — ортогональная матрица. Так как (5.13) 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 *разд. 5.2*, 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 (5.13) 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 q_{pq} матрицы A нулем. Вопросы сходимости различных численных схем, реализующих метод Якоби, рассмотрены в [74].

За основу может быть взят алгоритм jacobi из сборника алгоритмов линейной алгебры [73], 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 q_{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 $n^2 - n \,/2$ 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 невели-

ко и сравнимо со 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, представленных в *разд. 5.3*. Различные версии алгоритмов, учитывающие специфику решаемых классов прикладных задач, отличаются, в основном, методами построения аппроксимаций матриц $G\ x\ $. Соответствующие вопросы излагаются в следующем разделе.

5.3.3. 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, т. е. методами, не использующими в своей схеме производные от *J*. Действительно, заменяя $\partial J/\partial x_i$, например, конечно-разностным отношением $\left[J\ x+se^i\ -J\ x\ \right]\Big/s$, где $e^i=\ 0, ..., 1, ..., 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

$$\frac{\partial f}{\partial x_i}=\frac{f\ x+se^i\ -f\ x-se^i}{2s} \tag{5.14}$$

оказывается точным при любом $s\neq 0$.

Упражнение 5.1
Докажите последнее утверждение.

Точно так же точным оказывается следующее представление для вторых производных квадратичного функционала

$$\frac{\partial f}{\partial x_i \partial x_j} =$$

$$= \frac{f\ x + se^i + se^j\ - f\ x - se^i + se^j\ - f\ x + se^i - se^j\ + f\ x - se^i - se^j}{4s^2}. \qquad (5.15)$$

При использовании соотношений (5.14), (5.15) 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 (5.14), (5.15) 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 $\left\| x^i - x^{i-1} \right\|$ 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 (5.15) применительно к квадратичным зависимостям достигается при работе с максимально возможным s, т. к. в этом случае вклад погрешностей задания значений J в окончательный результат становится наименьшим. Поэтому чем дальше удалось продвинуться на основе построенной квадратичной аппроксимации функционала, т. е. 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 дальнейшее существенное улучшение ситуации не ожидается. В изу-

чаемых 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 приведенной далее последовательности шагов.

Алгоритм SPAC1.

Шаг 1. Ввести исходные данные: x, s.

Шаг 2. Вычислить матрицу $B = b_{ij}$ по формулам

$$b_{ij} = J\ x + se^i + se^j\ - J\ x - se^i + se^j\ - J\ x + se^i - se^j\ + J\ x - se^i - se^j\ ;$$

$$i,\ j \in 1{:}n\ ;\ e^i = 0, ..., 1, ..., 0\ . \qquad (5.16)$$

Шаг 3. С помощью процедуры jacobi построить ортогональную матрицу U, приводящую матрицу B к диагональному виду $U^T BU$.

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

$$s_{i+1} = 0{,}1\left\|x^{i+1} - x^i\right\|.$$

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 $T^T BT$.

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 компонентов 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 $I''\ y = U^T J''U$. Таким образом, если необходимо работать с функционалом, имеющим гессову матрицу вида $U^T J''U$, то для этого достаточно в качестве базиса взять столбцы матрицы U. Если требуется изменить матрицу Гессе и привести ее к виду $T^T\ U^T J''U\ T = U_1^T J''U_1$, 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-то 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 изменение 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 для SPAC1, т. к. 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 (5.15). При применении же базового алгоритма SPAC1 пригоден любой из описанных далее методов. Последнее обстоятельство обычно оказывается решающим при выборе метода.

5.3.4. 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

$$F_N\ c \triangleq \sum_{i=1}^{N}\left[J\ x^i\ - f\ x^i, c\ \right]^2 \to \min_{c\in R^m},$$
$$x^i = x_1^i, x_2^i, \ldots, x_n^i\ ,\ c = c_1, c_2, \ldots, c_m\ ,\ m = \frac{n^2}{2} + \frac{3n}{2} + 1,\ N \geq m. \tag{5.17}$$

Здесь $f\ x, c$ означает аппроксимирующий функционал с неизвестными коэффициентами c_i:

$$\begin{aligned} f\ x, c\ = {} & c_1 x_1^2 + c_2 x_1 x_2 + c_3 x_1 x_2 + \ldots + c_n x_1 x_n + \\ & + c_{n+1} x_2^2 + c_{n+2} x_2 x_3 + \ldots + c_{2n-1} x_2 x_n + \\ & \ldots \\ & + c_{\frac{n^2+n}{2}} x_n^2 + c_{\frac{n^2+n}{2}+1} x_1 + \ldots + c_{m-1} x_n + c_m. \end{aligned} \tag{5.18}$$

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 (5.18) в виде

$$f\ x, c\ = \langle c, y \rangle = y^T c,$$

что позволяет говорить о линейной регрессионной задаче оценки параметров $c_1, \ldots, c_m$. Система нормальных уравнений, отвечающая МНК-функционалу (5.17), имеет вид

$$Y_N Y_N^T c \triangleq \sum_{k=1}^{N} y^k\ \ y^k\ ^T c = Y_N J^N \triangleq \sum_{k=1}^{N} J_k y^k, \tag{5.19}$$

где $Y_N = y^1, y^2, \ldots, y^N$ — матрица размером $m \times N$; $y^k \triangleq y\ x^k\ $; $J^N \triangleq J_1, J_2, \ldots, J_N\ ^T$. Заметим, что так как квадратичный функционал $F\ c\ \geq 0$ и $F''\ c\ = YY^T$, то матрица YY^T неотрицательно определена. Можно рассчитать коэффициенты квадратичной модели непосредственно из системы линейных алгебраических уравнений (5.19). 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

$$Y^T\ ^+ = \lim\ \delta E + YY^T\ ^{-1} Y,\ \ \delta \to 0,\ \delta > 0.$$

Отсюда следует, что вместо системы уравнений

$$Y_r Y_r^T c = Y_r J^r \tag{5.20}$$

можно рассматривать систему

$$\delta E + Y_r Y_r^T \; c = Y_r J^r, \tag{5.21}$$

решение которой при $\delta \to 0$ сходится к решению (5.20) с минимальной нормой среди всех векторов, минимизирующих величину

$$\left\| Y_r^T c - J^r \right\|^2 \to \min_c,$$

что при $r = N$ совпадает с выражением (5.17). Далее на основе известных рекуррентных алгоритмов оценивания, соответствующих случаю $\delta = 0$, будут построены методы решения регуляризованных систем (5.21) 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 (5.20) может наталкиваться на существенные вычислительные трудности, т. к. при $N < m$ матрицы $Y_N Y_N^T$ будут вырождены, а при $N > m$ — плохо обусловлены. Введем обозначение

$$P_r^{-1} \triangleq \sum_{k=1}^{r} y^k \; y^{k\;T} + \delta E = P_{r-1}^{-1} + y^r \; y^{r\;T}. \tag{5.22}$$

Используя так называемую вторую лемму об обращении матриц (эквивалентную формуле Шермана — Моррисона — Вудбери [29])

$$\left[K^{-1} + B^T R^{-1} B \right]^{-1} = K - KB^T \left[BKB^T + R \right]^{-1} BK,$$

получим рекуррентное соотношение [4], [22]

$$P_r = \left[P_{r-1}^{-1} + y^r \; y^{r\;T} \right]^{-1} = P_{r-1} - P_{r-1} y^r \left[\; y^{r\;T} P_{r-1} y^r + 1 \right]^{-1} \; y^{r\;T} P_{r-1}. \tag{5.23}$$

Так как, очевидно, $P_1^{-1} = y^1 \; y^{1\;T} + \delta E,$ то из (5.22) следует, что необходимо положить $P_0^{-1} = \delta E$ 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 [72] немедленно следует из представления (5.22), т. к. матрица $y^r \; y^{r\;T}$ 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 (5.23) существуют.

Построим рекуррентное соотношение для определения оценок c_i^r. Уравнение (5.21) имеет вид

$$P_r^{-1}c^r = \sum_{k=1}^{r} J_k y^k, \tag{5.24}$$

откуда

$$P_r^{-1}c^r = \sum_{k=1}^{r} J_k y^k + y^r J_r = P_{r-1}^{-1}c^{r-1} + y^r J_r. \tag{5.25}$$

Здесь c^r означает оценку вектора c по r вычислениям функционала $J\ x$. Прибавляя и вычитая $y^r\ \ y^r\ ^T c^{r-1}$ в правой части (5.25), получим

$$P_r^{-1}c^r = P_{r-1}^{-1}c^{r-1} + y^r\left[J_r - \ y^r\ ^T c^{r-1}\right]. \tag{5.26}$$

Из (5.26) имеем окончательное выражение

$$c^r = c^{r-1} + P_r y^r\left[J_r - \ y^r\ ^T c^{r-1}\right]. \tag{5.27}$$

По формуле (5.27) может быть вычислена новая, оценка c^r вектора параметров при условии, что известны: предыдущая оценкам 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 (5.21) операцией вычисления скаляра, обратного к заданному $y^r\ ^T P_{r-1}y^r + 1$, выполняемой на каждом шаге итерационного процесса (5.23).

Непосредственно из построения уравнения видно, что результат для $r = N$, полученный согласно (5.23), (5.27) приводит к оценке, которая получается из решения полной системы (5.21). При этом необходимо положить $c^0 = 0$. Последнее следует из соотношений (5.24), (5.27), записанных для $r = 1$.

Действительно, согласно (5.24) имеем оценку $c^1 = P_r J_1 y^1$, полученную в результате решения системы (5.21). Из (5.27) следует $c^1 = c^0 + P_1 y^1\left[J_1 - \ y^1\ ^T c^0\right]$. 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, $i \in \left[1 : n^2/2 + n\right]$, задающих аппроксимацию матрицы $J''\ x$, может быть реализована процедура ОПС. При этом возможны раз-

личные 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 довольно 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 (5.23). 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* [12].

Используя представление (5.18), записанное в виде

$$f = y^T c, \tag{5.28}$$

где c — вектор оцениваемых параметров, получим следующую рекуррентную процедуру уточнения оценок c_i параметров c_i^*:

$$c^k = d^k + \frac{f_k - \left\langle d^k,\ y^k \right\rangle}{\left\langle y^k,\ y^k \right\rangle} y^k; \tag{5.29}$$

$$d^k = \begin{cases} d^{k-1}, & \text{если выполнено неравенство (5.30),} \\ c^{k-1}, & \text{в противном случае;} \end{cases}$$

$$\frac{f_k - \left\langle y^k,\ d^{k-1} \right\rangle^2}{\left\langle y^k,\ y^k \right\rangle} > \frac{f_{k-1} - \left\langle y^{k-1},\ d^{k-1} \right\rangle^2}{\left\langle y^{k-1},\ y^{k-1} \right\rangle} + \frac{f_k - \left\langle y^k,\ c^{k-1} \right\rangle^2}{\left\langle y^k,\ y^k \right\rangle}. \tag{5.30}$$

Геометрически алгоритм (5.29) реализует операцию проектирования вектора d^k на k-ю гиперплоскость (5.28), 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 (5.29) заключаются в небольшом количестве вычислений для реализации соотношений (5.29), (5.30), а также в существенно меньших объемах используемой памяти компьютера по сравнению с рекуррентным методом наименьших квадратов. Кроме того, алгоритм (5.29) 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: $x \in R^n$, $c \in R^N$ $(N = n^2/2 + 3n/2 + 1)$, $s \in R^1$; принять $F\ x := J\ x$, $d := c$; $y^0 = y\ x$, $F_0 := F$.

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

$$\frac{F_1 - \left\langle y^1,\ d \right\rangle^2}{\left\langle y^1,\ y^1 \right\rangle} \le \frac{F_0 - \left\langle y^0,\ d \right\rangle^2}{\left\langle y^0,\ y^0 \right\rangle} + \frac{F_1 - \left\langle y^1,\ c \right\rangle^2}{\left\langle y^1,\ y^1 \right\rangle},$$

принять $d = c$.

Шаг 6. Принять

$$c := d + \frac{F_1 - \left\langle d,\ y^1 \right\rangle}{\left\langle y^1,\ y^1 \right\rangle} y^1;\quad F_0 := F_1;\quad y^0 := y^1.$$

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 $m \le n$, 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, $i \in \left[1:\ n^2+n\ \middle/ 2\right]$ 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.

5.4. 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.

5.4.1. Задачи аппроксимации

Характерные для теории управления целевые функционалы, отражающие меру "близости" расчетных и делаемых зависимостей, могут быть представлены в виде:

$$J\ x\ = \nu^{-1} \sum_{k=1}^{n} \varphi_k^{\nu}\ x\ ,\quad \nu = 2,\ 3,\ \ldots, \tag{5.31}$$

где φ_k — алгоритмически заданные функции ($R^n \to R$) вектора управляемых параметров. Как отмечалось в *разд. 3.10.1*, класс (5.31) включает в себя наиболее часто используемые на практике МНК-критерии, а также минимаксные целевые функционалы.

Из (5.31) имеем следующие выражения для составляющих вектора-градиента:

$$\frac{\partial J}{\partial x_i} = \sum_{k=1}^{m} \varphi_k^{\nu-1}\ x\ \frac{\partial \varphi_k\ x}{\partial x_i},\quad i \in 1:n\ . \tag{5.32}$$

Естественный способ аппроксимации вторых производных, принятый, в частности, в процедурах оптимизации типа Гаусса — Ньютона [50], а также в теории чувствительности систем автоматического управления, состоит в линеаризации функций φ_k вблизи текущей точки x':

$$\varphi_k\ x\ \cong \varphi_k\ x'\ + \left\langle \frac{\partial \varphi_k\ x'}{\partial x},\ x - x' \right\rangle. \tag{5.33}$$

Используя (5.33), получим

$$\frac{\partial^2 J}{\partial x_i \partial x_j} \cong \nu - 1 \sum_{k=1}^{m} \varphi_k^{\nu-2}\ x\ \frac{\partial \varphi_k\ x}{\partial x_i} \frac{\partial \varphi_k\ x}{\partial x_j}; \quad (5.34)$$

$$\nu = 2,\ 3,\ ...;\ \ j \geq i;\ \ i \in 1:n\ .$$

В векторно-матричных обозначениях использование аппроксимаций (5.34) эквивалентно отбрасыванию второго слагаемого в представлении

$$G_i\ x\ =\ \nu - 1\ F^T\ x\ \mathrm{diag}\left[\ \varphi_i^{\nu-2}\ x\ \right] F\ x\ + Q\ x\ , \quad (5.35)$$

где $F\ x$ — матрица Якоби размером $m \times n$ вектор-функции $\varphi\ x\ = \left[\ \varphi_1\ x\ ,\ ...,\ \varphi_m\ x\ \right]$; $G_i\ x$ — 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 (5.31) [21].

Метод ОПС с вычислением аппроксимации матрицы Гессе $G\ x$ на основе формул (5.34) 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 $B = F^T \mathrm{diag}\left[\ \varphi_i^{\nu-2}\ \right] F.$

Шаг 4. С помощью процедуры jacobi построить ортогональную матрицу U, приводящую матрицу B к диагональному виду $U^T B U\ .$

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 $\varphi\ x\ = \left[\ \varphi_1\ x\ ,\ ...,\ \varphi_m\ x\ \right]$.

Далее рассматриваются вопросы применения алгоритмов типа SPAC1, SPAC5 для решения конкретных классов задач теории управления.

5.4.2. Идентификация нелинейных детерминированных объектов на основе функциональных рядов Вольтерра

Согласно выражению (2.16) имеем следующее представление для целевого функционала, отражающего ошибку идентификации

$$J\ x\ = 0{,}5\sum_{k=1}^{N_0}\left[\ \sum_{\nu=1}^{N_1} x_\nu y_{\nu 1}\ k\ + \sum_{\nu=1}^{N_2} x_{N_1+\nu} y_{\nu 2}\ k\ - H\ k\ \right]^2, \tag{5.36}$$

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 (5.36) имеет вид (5.31) с параметром $\nu = 2$, где

$$\varphi_k\ \ x\ = \left\langle x,\ y\ k\ \right\rangle - H\ k\ ; \tag{5.37}$$

$$x = \ x_1,\ x_2,\ ...,\ x_n\ ;\ n = N_1 + N_2;\ y\ k\ =\ y_{11}\ k\ ,\ ...,\ y_{N_1 1}\ k\ ,\ y_{12}\ k\ ,\ ...,\ y_{N_2 2}\ k\ \ .$$

Из (5.37) следует, что в данном случае приближенные равенства (5.33), (5.34) выполняются точно.

Элементы гессовой матрицы g_{ij} равны

$$g_{ij} = \sum_{k=1}^{N_0} y_i\ k\ \ y_j\ k\ ;\ \ i,\ j \in 1{:}n\ . \tag{5.38}$$

В силу линейной зависимости φ_k от x функционал (5.36) 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 (5.38). Однако с вычислительной точки зрения более эффективным обычно оказывается подход, основанный на непосредственной минимизации (5.36). Дадим необходимые разъяснения.

Как отмечалось, задачи минимизации функционалов типа (5.36) оказываются очень плохо обусловленными. Это приводит к известным вычислительным трудностям при решении линейных систем нормальных уравнений

$$Gx = b, \tag{5.39}$$

отражающих необходимые условия экстремума. Здесь $G = Y^T Y$, где Y — матрица размером $N_0 \times n$ вида

$$Y = \begin{bmatrix} y_1\ 1 & ... & y_n\ 1 \\ ... & ... & ... \\ y_1\ N_0 & ... & y_n\ N_0 \end{bmatrix}. \tag{5.40}$$

Задача решения (5.39) оказывается некорректной по Тихонову [70], 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 (5.39) можно проследить из представления решения в виде

$$x^* = \left[u^1,\ u^2,\ \ldots,\ u^n \right] \begin{bmatrix} \lambda_1^{-1} \left\langle b,\ u^1 \right\rangle \\ \lambda_2^{-1} \left\langle b,\ u^2 \right\rangle \\ \ldots \\ \lambda_n^{-1} \left\langle b,\ u^n \right\rangle \end{bmatrix}, \tag{5.41}$$

где $J\ x = 1/2 \langle Gx,\ x \rangle - \langle b,\ x \rangle + c,\ c \in R^1,\ u^i,\ \ \lambda_i$ — соответственно, собственные векторы (ортонормированные) и собственные числа матрицы G. Данное соотношение следует из очевидного матричного представления

$$x^* = G^{-1} b = U \text{diag}\ \ \lambda_i^{-1}\ \ U^T b,$$

где $U = \left[u^1,\ u^2,\ \ldots,\ u^n \right]$.

Из (5.41) 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 (5.39), содержащей в явном виде матрицу G, недопустимо, т. к. из-за ограниченной точности представления элементов g_{ij}

$$g_{ij} = \sum_{k=1}^{n} u_i^k u_j^k \lambda_k \tag{5.42}$$

матрицы G информация о "малых" собственных числах $\lambda_{n-r+1},\ \ldots,\ \lambda_n$ теряется на фоне "больших" $\lambda_1,\ \ldots,\ \lambda_{n-r}$. Указанное обстоятельство и приводит к некорректности задачи (5.39). Приведем конкретный численный пример.

Пример [83]. Рассмотрим квадратичный функционал с высокой степенью овражности $\eta = 10^{12}$:

$$J = 0{,}5 \sum_{i=1}^{4} \lambda_i \left\langle x,\ u^i \right\rangle^2 - \langle b,\ x \rangle;\ \ b = \ 1;\ 1;\ 1;\ 1\ .$$

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:

$$u^1 = \frac{1}{\sqrt{3}}\ 1;\ -1;\ 1;\ 0\ ; \qquad u^2 = \frac{q}{\sqrt{6}}\ 1;\ 2\ ;1;\ 0\ ;$$

$$u^3 = \frac{1}{\sqrt{3}}\ 1;\ 0;\ -1;\ 1\ ; \qquad u^4 = \frac{q}{\sqrt{6}}\ 1;\ 0;\ -1;\ -2\ ;$$

$$x^0 = \ 0;\ 0;\ 0;\ 0\ ;$$

$$x^* \cong\ 3333{,}3;\ 13333;\ 10000;\ 6666{,}7\ ;\ J\ x^*\ \cong -16667.$$

Точные значения компонентов вектора x^* задаются выражением (11):

$$x^* = \ u^1;\ u^2;\ u^3;\ u^4\quad \lambda_1^{-1}\langle b,\ u^1\rangle;\ \lambda_2^{-1}\langle b,\ u^2\rangle;\ \lambda_3^{-1}\langle b,\ u^3\rangle;\ \lambda_4^{-1}\langle b,\ u^4\rangle\ \cong$$

$$\cong\ u^1;\ u^2;\ u^3;\ u^4\quad 0;10^4\langle b,\ u^2\rangle;0\ ;10^4\langle b,\ u^4\rangle\ .$$

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 $J = 1/2\langle Ax,\ x\rangle - \langle b,\ x\rangle$, содержащего в явном виде матрицу $A = J''$, недопустимо, т. к. 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 (5.39).

Привлечение методов регуляризации для решения (5.39) позволяет определить "квазирешение" $\tilde{x}^*$, отражающее некоторый компромисс между величиной нормы $\left\|\tilde{x}^*\right\|$ и невязкой $\left\|G\tilde{x}^* - b\right\|$. При этом относительно малым невязкам могут соответствовать относительно большие ошибки, как по аргументу $\left\|\tilde{x}^* - x^*\right\|$, так и по функционалу $J\ \tilde{x}^*\ - J\ x^*\ $. Действительно, рассмотрим одномерный случай параболы $y\ x\ = 1/2\,ax^2$. Тогда невязка будет определяться величиной градиента (в данном случае — производной) $y'\ x\ = ax$. Здесь коэффициент a моделирует влияние неко-

торого "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 "малости" $\left\|\tilde{x}^*\right\|$ можно получить сколь угодно большое значение нормы $\left\|\tilde{x}^* - x^*\right\|$ при малой норме $\left\|\tilde{x}^*\right\|$.

В силу изложенного на практике достаточно редко обращаются к непосредственному решению нормальных уравнений (5.39). Стандартный подход описан в [35] и заключается в построении сингулярного разложения прямоугольной матрицы (5.40)

$$Y = U \Sigma V^T,$$

где U — ортогональная матрица размером $N_0 \times N_0$, V — ортогональная матрица размером $n \times n$; Σ — диагональная матрица размером $N_0 \times n$, у которой $\sigma_{ij} = 0$ при $i \neq j$ и $\sigma_{ii} \triangleq \sigma_i \geq 0$. Вектор $\hat{x}^*$, аппроксимирующий точный минимизатор функционала (5.36), выражается соотношениями $\hat{x}^* = V\hat{z}^*$, где $\hat{z}_i^* = d_i / \sigma_i$, 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 $d = U^T h$, $h = H\ 1\ , ...,\ H\ N_0\ ^T$. Величину δ, отражающую уровень "малости" соответствующего сингулярного числа, целесообразно полагать равной $n\lambda_1 \varepsilon_м$, где $\lambda_1 = \max_i \left|\lambda_i\right|$. 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 $\hat{z}^*$ 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 (5.31) 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 (5.31) является наиболее предпочтительным подходом.

5.4.3. Корреляционные методы идентификации стохастических объектов

В соответствии с результатами из *разд. 2.3.2*, 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 (4.20)

$$J\ x\ = 0{,}5\sum_{j=1}^{N}\left[\sum_{i=1}^{N}\omega_i\hat{R}_{GG}\left[q\ \ j-i\ \right]-\hat{R}_{\hat{H}G}\ qj\ \right]^2+0{,}5\alpha\sum_{j=2}^{N}\left[\omega_j-\omega_{j-1}\right]^2, \quad (5.43)$$

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. 2.11); α — параметр регуляризации.

При отсутствии необходимости проводить регуляризацию имеем $\alpha = 0$, и минимизация функционала (5.43), осуществляется по методике из *разд. 5.4.2* 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:

$$\frac{\partial J}{\partial x_i}=\sum_{j=1}^{N}\left[\sum_{i=1}^{N}\omega_i\hat{R}_{GG}\left[q\ \ j-i\ \right]-\hat{R}_{\hat{H}G}\ qj\ \right]\hat{R}_{GG}\left[q\ \ j-i\ \right]+\alpha\Psi_l,\ l\in 1:N\ , \quad (5.44)$$

$$\Psi_l=\begin{cases}\omega_l-\omega_{l-1}\ -\ \omega_{l+1}-\omega_l\ , & 1<l<N;\\ -\ \omega_2-\omega_1\ , & l=1;\\ \omega_N-\omega_{N-1}\ , & l=N.\end{cases}$$

$$\frac{\partial^2 J}{\partial x_l\partial x_m}=\sum_{j=1}^{N}\hat{R}_{GG}\left[q\ \ j-m\ \right]\hat{R}_{GG}\left[q\ \ j-l\ \right]+\alpha\varphi_{lm},\ \ l,\ m\in 1:N\ , \quad (5.45)$$

где матрица $\varphi=\ \varphi_{lm}$ имеет вид

$$\varphi=\begin{bmatrix}1 & -1 & 0 & \ldots & 0 & 0\\ -1 & 2 & -1 & \ldots & 0 & 0\\ 0 & -1 & 2 & \ldots & 0 & 0\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & 0 & \ldots & 2 & -1\\ 0 & 0 & 0 & \ldots & -1 & 1\end{bmatrix}.$$

5.4.4. 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 нестационарных задач.

Решение последних с помощью методов ОПС может быть проведено по методике из *разд. 5.4.3* и не требует дополнительных разъяснений.

5.4.5. Идентификация нелинейных динамических систем

Пусть траектория $z\ t,\ x$ динамической системы описывается уравнением

$$\dot{z}\ t,\ x\ = f\ t,\ z,\ x\ ,z\ t_0,\ x\ = z^0,\ \ t_0 \le t \le T, \tag{5.46}$$

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

$$y\ t\ = H\ t\ z\ t,\ x\ ,\ t \in\ t_0,\ T \tag{5.47}$$

определить вектор параметров x из условия минимума функционала

$$J\ x\ = 0{,}5\sum_{j=1}^{N}\left\| H\ t_j\ z\ t_j,\ x\ - y\ t_j\ \right\|^2, \tag{5.48}$$

где $H\ t$ — непрерывная матрица размером $l \times r$, которая связывает вектор измерения с вектором состояния ($l < r$); t_j ($j \in 1:N$, $N \ge n$) — дискретные моменты съема измерительной информации.

Функционал (5.48) с точностью до обозначений имеет, очевидно, вид (5.31)

$$J\ x\ = 0{,}5\sum_{k=1}^{N}\sum_{s=1}^{l}\varphi_{sk}^2\ \ x\ , \tag{5.49}$$

где $\varphi_{sk}\ \ x\ = \left\langle h_s\ \ t_k\ \ , z\ \ t_k, x\ \right\rangle - y_s\ \ t_k\ \ ;\ h_s$ — s-я строка матрицы H.

Из (5.49) получаем представление, аналогичное (5.34):

$$\frac{\partial^2 J}{\partial x_i \partial x_j} \cong \sum_{k=1}^{N}\sum_{s=1}^{l} \frac{\partial \varphi_{sk}}{\partial x_i}\frac{\partial \varphi_{sk}}{\partial x_j} =$$

$$= \sum_{k=1}^{N}\sum_{s=1}^{l} \left\langle h_s\ t_k\ , \frac{\partial z\ t_k, x}{\partial x_i} \right\rangle \left\langle h_s\ t_k\ , \frac{\partial z\ t_k, x}{\partial x_j} \right\rangle, \qquad (5.50)$$

$i, j \in 1:n$.

Таким образом, для данного класса задач построение аппроксимации матрицы $G\ x$ сводится к построению матрицы Якоби $F\ t, x = \partial z\ t, x\ /\partial x$ 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 (5.46), при фиксированном $x = \tilde{x}$ решается линейное матричное уравнение

$$\frac{d}{dt} F\ t, \tilde{x}\ = \frac{\partial f\ t, z, \tilde{x}}{\partial z} F\ t, \tilde{x}\ + \frac{\partial f\ t, z, \tilde{x}}{\partial x} \qquad (5.51)$$

с $F\ t_0, \tilde{x} = 0,\ t \in t_0, T$. При этом зависимость $z\ t, x$ при фиксированном x получается непосредственным интегрированием (5.46). Уравнение (5.51) выводится в любом стандартном учебнике по дифференциальным уравнениям.

Важно понимать, что в данном случае квадратичная модель целевого функционала (5.48) вводится лишь для выбора наиболее рациональной системы координат, а минимизации подвергается исходный функционал (5.48).

Пример. Пусть фазовые траектории некоторой динамической системы имеют вид

$$z\ t, x\ = \varphi\ t\ \left[\exp\ -x_1 t\ + \exp\ -x_2 t\ \right],\ t_0 = 0,\ z^0 = 2\varphi\ 0\ ,$$

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

$$J\ x\ = \sum_{j=1}^{2} \left[z\ t_j, x\ - \overline{z}_j \right]^2,$$

где $\overline{z}_j = z\ t_j, x^*$; $t_1 = 1,\ t_2 = 2;\ x_1^* = 1,\ x_2^* = 1{,}5$.

Для начальной точки $x^0 = 0{,}1;\ 0{,}1$, $J\ x^0 = 2{,}6006$ имеем согласно (5.50) (элемент матрицы g_{12} не показан, т. к. матрица симметрична)

$$G\ x \cong \left[\begin{array}{c|c} \dfrac{\partial z\ t_1,\ x}{\partial x_1}^2 + \dfrac{\partial z\ t_2,\ x}{\partial x_1}^2 & \\ \hline \dfrac{\partial z\ t_1,\ x}{\partial x_1} \cdot \dfrac{\partial z\ t_1,\ x}{\partial x_2} + \dfrac{\partial z\ t_2,\ x}{\partial x_1} \cdot \dfrac{\partial z\ t_2,\ x}{\partial x_2} & \dfrac{\partial z\ t_1,\ x}{\partial x_2}^2 + \dfrac{\partial z\ t_2,\ x}{\partial x_2}^2 \end{array}\right] =$$

$$= \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}.$$

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 $J\ \tilde{x} = 1{,}5164 \cdot 10^{-11}$.

Начальное значение шага дискретности *s* полагалось равным 0,01.

5.4.6. 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 $y = y_1, y_2, ..., y_l$, функционально связанные с вектором $z = z_1, z_2, ..., z_r$ 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 *разд. 2.2* терминологии, сформулированная задача является задачей сглаживания.

Пусть движение системы определяется уравнением

$$\dot{z}\ t = f\ t, z\ ,\ z\ t_0 = x,\ t_0 \le t \le T, \tag{5.52}$$

где $f\ t, z$ — известная нелинейная непрерывно дифференцируемая функция своих аргументов, удовлетворяющая в некоторой замкнутой области условиям теоре-

мы существования и единственности решения. Предположим, что векторы z и y связаны линейным соотношением

$$y\ t\ = H\ t\ z\ t\ , \tag{5.53}$$

где $H\ t\ $ — известная матрица размером $l \times r$, определяемая конструкцией измерительного устройства. При сделанных предположениях траектория $z\ t\ $ системы (5.52) однозначно определяется начальным вектором $x = \ x_1, ..., x_n\ $. Оценки $x_i, \ i \in \ 1:n\ $ могут быть определены из условия минимума целевого функционала

$$J\ x\ = 0{,}5 \sum_{j=1}^{N} \left\| H\ t_j\ \ z\ t_j\ - y\ t_j\ \right\|^2. \tag{5.54}$$

Аналогично *разд. 5.4.5* имеем следующие формулы для элементов аппроксимации матрицы Гессе:

$$\frac{\partial^2 J}{\partial x_i \partial x_j} \cong \sum_{k=1}^{N} \sum_{s=1}^{l} \left\langle h_s\ \ t_k\ \ , \frac{\partial z\ t_k}{\partial x_i} \right\rangle \left\langle h_s\ \ t_k\ \ , \frac{\partial z\ t_k}{\partial x_j} \right\rangle, \tag{5.55}$$

$$i, \ j \in \ 1:n\ .$$

В силу известных результатов теории дифференциальных уравнений матрица $F\ t, \ x\ \triangleq \partial z / \partial x$ 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 (5.52).

5.4.7. Идентификация возмущающих воздействий

Задача идентификации возмущающих сил формулируется следующим образом.

Пусть движение некоторой динамической системы описывается уравнением

$$\dot{z}\ t\ = f\ \ t, \ z, \ v\ \ , \ \ z\ \ t_0\ \ = z_0, \tag{5.56}$$

где v — вектор возмущающих сил. Предполагая, что известна параметризация

$$v = v\ \ t, \ x\ \ , \ x = \ x_1, \ x_2, \ ..., \ x_n\ \ , \ t \in \ t_0, \ T\ \ , \tag{5.57}$$

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 из *разд. 5.4.5*. Более того, обе эти задачи могут решаться совместно по единой методике.

Неизвестный вектор x по измерениям y (5.53) может быть найден из условия минимума функционала

$$J\ x\ = 0{,}5\sum_{j=1}^{N}\left\| H\ t_j\ \ z\ t_j,\ x\ - y\ t_j\ \right\|^2, \tag{5.58}$$

где $z\ t_j,\ x\ $ — результат интегрирования (5.56) при фиксированном x. Приближенные выражения для элементов матрицы $G\ x\ $ определяются выражениями (5.55), где матрица $F\ t,\ x\ \triangleq \partial z\ t,\ x\ /\partial x$ при $x = \tilde{x}$ является решением матричного уравнения

$$\frac{d}{dt}F\ t,\ \tilde{x}\ = \frac{\partial f\ t,\ z,\ v}{\partial z}F\ t,\ \tilde{x}\ + \frac{\partial f\ t,\ z,\ \tilde{v}}{\partial v}\frac{\partial v\ t,\ \tilde{x}}{\partial x}, \tag{5.59}$$

где $\tilde{v} = v\ t, \tilde{x}\ ;\ \ F\ t_0, \tilde{x}\ = 0.$

5.4.8. Решение систем неравенств

Задачи с неравенствами весьма часто возникают в теории управления. При этом задача решения системы вида

$$y_j \le t_j,\ j \in 1{:}m \tag{5.60}$$

одним из рассмотренных в *разд. 3.10.2* способов формулируется как оптимизационная. Соответствующие достаточно общему подходу целевые функционалы представимы в виде

$$J\ x\ = \nu^{-1}\sum_{k=1}^{m}\exp\left[-\nu z_k\ \ x\ \right],\ \ \nu = 1,\ 2,\ 3,\ ..., \tag{5.61}$$

где $z_k\ \ x\ $ отражает k-й "запас" в выполнении неравенств (5.60).

Из (5.61) получаем:

$$\frac{\partial J}{\partial x_i} = -\sum_{k=1}^{m}\exp\left[-\nu z_k\ \ x\ \right]\frac{\partial z_k\ \ x}{\partial x_i}; \tag{5.62}$$

$$\frac{\partial^2 J}{\partial x_i \partial x_j} = \sum_{k=1}^{m}\exp\left[-\nu z_k\ \ x\ \right]\left[\frac{\partial z_k}{\partial x_i}\frac{\partial z_k}{\partial x_j}\nu - \frac{\partial^2 z_k}{\partial x_i \partial x_j}\right]. \tag{5.63}$$

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 (5.63) используемую на практике аппроксимацию матрицы Гессе

$$\frac{\partial^2 J}{\partial x_i \partial x_j} \cong \nu \sum_{k=1}^{m} \exp\left[-\nu z_k \;\; x\right] \frac{\partial z_k}{\partial x_i} \frac{\partial z_k}{\partial x_j}, \;\; \nu = 1, \; 2, \; 3, \; ...; \;\; j \geq i, \;\; i \in 1:n \; . \qquad (5.64)$$

Из (5.63) следует, что при увеличении ν точность аппроксимации (5.64), 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.

5.4.9. 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 (3.54)):

$$J \;\; x \;\; \triangleq P \;\; y \;\; x, \; \xi \;\; \leq t \;\; \rightarrow \max_x . \qquad (5.65)$$

Ставится задача выбора такого вектора управляемых параметров $x = \; x_1, \; x_2, \; ..., \; x_n \;$, 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 (5.65) 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 (5.65) может быть представлен в виде [9]:

$$P\ x\ = \int_D \ldots \int \varphi\ \xi\ \Psi\ x, \xi\ d\xi, \tag{5.66}$$

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 (5.66) и перехода к статистическим аналогам соответствующих соотношений могут быть получены выражения для составляющих градиента и матрицы Гессе критерия (5.65). Для случая, когда $\Psi\ x, \xi$ задает многомерное гауссовское распределение вида

$$\Psi\ x, \xi\ = \left[\ 2\pi\ ^{n/2} \prod_{i=1}^{n} \sigma_{\xi_i} \sqrt{R}\ \right]^{-1} \exp\left[-\frac{1}{2R} \sum_{i=1}^{n} \sum_{j=1}^{n} R_{ij} \frac{\xi_i - x_i \quad \xi_j - x_j}{\sigma_{\xi_i} \sigma_{\xi_j}} \right], \tag{5.67}$$

справедливы следующие представления для производных:

$$\frac{\partial P}{\partial x_s} = \frac{1}{NR\sigma_s} \sum_{i=1}^{N} \varphi\ \xi^l\ \sum_{i=1}^{N} R_{si} \frac{\xi_i^l - x_i}{\sigma_i}; \tag{5.68}$$

$$\frac{\partial^2 P}{\partial x_s \partial x_t} = \frac{1}{NR\sigma_s\sigma_t} \left[-kR_{si} + R^{-1} \sum_{i=1}^{N} \varphi\ \xi^l\ \sum_{i=1}^{n} R_{si} \frac{\xi_i^l - x_i}{\sigma_i} \sum_{j=1}^{n} R_{tj} \frac{\xi_i^l - x_j}{\sigma_j} \right], \tag{5.69}$$

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 (5.68), (5.69) следует, что элементы вектора градиента и матрицы Гессе критерия (5.65) 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 достаточно широкого класса прикладных задач теории управления.

5.4.10. Обеспечение максимального запаса работоспособности оптимизируемой системы

Как отмечалось в *разд. 3.10*, основные требования-спецификации к оптимизируемой системе могут быть выражены в виде системы неравенств (5.60). При этом основная задача формулируется следующим образом: обеспечить такой набор управляемых параметров x, при которых наилучшим образом выполняются все спецификации (5.60) 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 (3.56):

$$z_j\ x \triangleq \left[\frac{t_j - y_j\ x}{\delta_j} - 1 \right]. \tag{5.70}$$

В случае использования представлений (5.70), где δ_j, по существу, играет роль весовых коэффициентов, общие соотношения (5.62), (5.63), приближенно реализующие принцип максимума минимального из запасов, принимают следующий вид:

$$\frac{\partial J}{\partial x_i} = \sum_{k=1}^{m} \exp\left[-\nu z_k\ x \right] \delta_k^{-1} \frac{\partial y_k\ x}{\partial x_i}; \tag{5.71}$$

$$\frac{\partial^2 J}{\partial x_i \partial x_j} = \sum_{k=1}^{m} \exp\left[-\nu z_k\ x \right] \left[\nu \delta_k^{-2} \frac{\partial y_k\ x}{\partial x_i} \frac{\partial y_k\ x}{\partial x_j} + \delta_k^{-1} \frac{\partial^2 y_k\ x}{\partial x_i \partial x_j} \right]. \tag{5.72}$$

Выполняя линеаризацию k-го выходного параметра y_k в окрестности текущей точки x'

$$y_k\ x \cong y_k\ x' + \left\langle \frac{\partial y_k\ x'}{\partial x},\ x - x' \right\rangle, \tag{5.73}$$

приходим к следующему представлению, аналогичному (5.64):

$$\frac{\partial^2 J}{\partial x_i \partial x_j} \cong \nu \sum_{k=1}^{m} \exp\left[-\nu z_k\ x \right] \left[\delta_k^{-2} \frac{\partial y_k\ x}{\partial x_i} \frac{\partial y_k\ x}{\partial x_j} \right], \tag{5.74}$$

$$\nu = 1,\ 2,\ 3,\ ...,\quad j \geq i,\quad i \in 1:n\ .$$

При реализации методов ОПС множитель ν в (5.74) может быть опущен. В ряде случаев соотношение (5.73), а вместе с ним и (5.74) 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 (5.72).

5.4.11. Оптимизация систем по сигномиальным целевым функционалам

Так называемые "простые" задачи оптимального параметрического синтеза систем характеризуются наличием известных аналитически заданных связей между вектором аргументов и соответствующими значениями целевых функционалов. При этом, как показывает практика [71], весьма характерна *сигномиальная* структура указанных связей:

$$J\ x\ =\sum_{i=1}^{N}\sigma_i t_i\ x\ , \tag{5.75}$$

$$t_i\ x\ =c_i\prod_{j=1}^{n}x_j^{\alpha_{ij}};\ \ c_i>0;\ \ \alpha_{ij}\in R^1;\ \ \sigma_i=\pm 1;\ \ x_j>0,\ \ j\in 1:n\ .$$

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 (5.75) изучаются в специальном разделе теории нелинейного программирования — *геометрическом программировании*.

Существуют регулярные методы минимизации функционалов (5.75), основанные на теории двойственности в геометрическом программировании в предположении $\sigma_i=1,\ \ i\in 1:N$ [69]. 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 [71]. Указанный подход становится практически наиболее оправданным в общем случае произвольных σ_i, когда основные предпосылки метода двойственного геометрического программирования нарушаются.

Как показано в работе [71], применение Н-методов для решения задачи минимизации (5.75) 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 (5.75). Действительно, выполняя, например, замену переменных $x_j=\exp\ y_j$, получим:

$$\begin{aligned}
&t_i\ x\ =c_i\prod_{j=1}^{n}\left[\exp\ y_j\ \right]^{\alpha_{ij}}=c_i\exp\left[\sum_{j=1}^{n}\alpha_{ij}y_j\right]=\tilde{t}_i\ \ y\ ;\\
&\frac{\partial\tilde{t}_i\ \ y}{\partial y_k}=c_i\alpha_{ik}\exp\left[\sum_{j=1}^{n}\alpha_{ij}y_j\right]=\alpha_{ik}\tilde{t}_i\ \ y\ ;\\
&\frac{\partial^2\tilde{t}_i\ \ y}{\partial y_k\partial y_l}=\alpha_{ik}\alpha_{il}\tilde{t}_i\ \ y\ .
\end{aligned} \tag{5.76}$$

Из (5.76) имеем

$$\frac{\partial \tilde{J}}{\partial y_k} = \sum_{i=1}^{N} \sigma_i \alpha_{ik} c_i \exp\left[\sum_{j=1}^{n} \alpha_{ij} y_j\right]; \tag{5.77}$$

$$\frac{\partial^2 J}{\partial y_k \partial y_l} = \sum_{i=1}^{N} \sigma_i \alpha_{ik} \alpha_{il} c_i \exp\left[\sum_{j=1}^{n} \alpha_{ij} y_j\right]. \tag{5.78}$$

Перспективность указанной реализации методов ОПС определяется достаточно широкой сферой приложений математических моделей типа (5.75), не ограниченной задачами оптимального проектирования систем с функционалами качества, непосредственно сводимых к сигномиальному виду. Как уже указывалось, функциональная зависимость (5.75) часто дает хорошее приближение к эмпирическим данным в широком диапазоне изменения переменных x_i. Кроме того, функционалы (5.75) используются непосредственно в задачах дискретного оптимального управления.

Более детальное обсуждение соответствующих вопросов выходит за рамки данной книги и может составить предмет отдельного изложения.

5.4.12. 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 чрезмерной громоздкостью численных процедур, соответствующих методу динамического программирования Р. Беллмана. 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:

$$R\ x, \omega\ = b\ z_q\ + \sum_{i=1}^{q-1} h_i B\ z_i\ , \tag{5.79}$$

где $\omega = \left[u^1, u^2, ..., u^q\right] \in R^{rq}$ — полный вектор управлений; $x = x\ \omega\ = \left[x^1, x^2, ..., x^q\right] \in R^q$ — 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

$$\frac{dx}{dt} = f\left[x\ t\ ,u\ t\ ,t\right],\ t \in\ 0,\ T\ ,\ x\ 0\ = x_1;$$

$$x\ t\ \in R^n,\ u\ t\ \in R^r;\ x_i = x\ t_i\ ,\ u_i = u\ t_i\ ;\ 0 = t_1 < t_2 < ... < t_q = T. \qquad (5.80)$$

К виду (5.79), 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 (5.79) представлены в работе [29]. 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.

5.5. Основные результаты и выводы

В *данной главе* представлены следующие основные результаты.

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 (жестких) 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 конечной величины $\varepsilon_м$ (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 и SPAC2 в среднем сократить время минимизации приблизительно в $0{,}8n$ раз, где n — размерность пространства поиска. Однако надежность и универсальность алгоритмов SPAC1 и SPAC2 остаются более высокими.
7. Изучены специальные схемы реализация методой ОПС для следующих классов прикладных задач:
   - идентификация нелинейных детерминированных объектов на основе функциональных рядов Вольтерра;
   - идентификация стохастических объектов на основе корреляционных методов;
   - синтез статистически оптимальных систем автоматического управления;
   - идентификация нелинейных динамических систем;
   - оценка состояний динамических систем (задача о наблюдении);
   - идентификация возмущающих сил;
   - управление технологическим процессом серийного выпуска изделий по критерию вероятности выхода годных;

- обеспечение максимального запаса работоспособности системы по заданному списку выходных параметров;
- задачи оптимального управления.

Указанные специальные реализации методов ОПС позволяют 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 6

Градиентные стратегии

В данной главе строятся оптимизирующие процедуры второго порядка ньютоновского типа с учетом основных особенностей прикладных оптимизационных задач и требований к соответствующим методам и алгоритмам *(см. разд. 3.9 и 4.5.6)*. 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 *разд. 4.5.6*.

С учетом представленных в *главе 4* моделей явления овражности, эффективность алгоритмов оценивается по свойствам соответствующих *функций релаксации*, полностью определяющих локальные характеристики рассматриваемых методов.

6.1. Общая схема градиентных методов. Понятие функции релаксации

Пусть решается задача

$$J\ x\ \to \min_{x},\ x \in R^n,\ J \in C^2\ \ R^n\ . \tag{6.1}$$

Рассмотрим класс матричных градиентных методов вида

$$x^{k+1} = x^k - H_k\ \ G_k,\ h_k\ \ g\ \ x^k\ \ ,\ \ \ h_k \in R^1; \tag{6.2}$$

$G_k \triangleq J''\ \ x^k\ ;\ g\ \ x^k\ \triangleq J'\ \ x^k\ ,\ H_k$ — матричная функция от G_k. Предполагается, что в некоторой ζ_k-окрестности $x \in R^n \left\| x - x^k \right\| \le \zeta_k$ точки x^k функционал $J\ x$ достаточно точно аппроксимируется параболоидом

$$f\ \ x\ = \frac{1}{2}\left\langle G_k x,\ x \right\rangle - \left\langle b_k,\ x \right\rangle + c_k, \tag{6.3}$$

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

$$f_1(z) = f(x^* + z) = \frac{1}{2}\langle G_k z,\ z\rangle + \bar{c}_k. \tag{6.4}$$

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 (6.2) обладает свойством инвариантности относительно смещения начала координат: будучи записанной для $f(x)$, она преобразуется в аналогичное соотношение для $f_1(x)$. Именно, имеем для $f(x)$:

$$x^{k+1} = x^k - H_k(G_k x^k - b_k). \tag{6.5}$$

Полагая $z^k = x^k - x^*$, получим из (6.5): $z^{k+1} = z^k - H_k G_k z^k$. А это есть запись метода (6.2) для функционала f_1.

Ставится задача построения таких матричных функций H_k, чтобы выполнялись условия релаксационности процесса $f(x^{k+1}) \le f(x^k)$ и при этом величина нормы $\left\|x^{k+1} - x^k\right\|$ ограничивалась сверху только параметром ζ_k, характеризующим область справедливости локальной квадратичной модели (6.3). При высокой степени овражности $\eta(x^k)$ для большинства классических схем поиска имеем $\left\|x^{k+1} - x^k\right\| \ll \zeta_k$, что в результате приводит к медленной сходимости.

Определение. Скалярная функция $R_h(\lambda) = 1 - H(\lambda, h)\lambda;\ \lambda, h \in R^1$ называется *функцией релаксации* метода (6.2), а ее значения $R_h(\lambda_i)$ на спектре матрицы G_k — *множителями релаксации* для точки x^k [81].

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 (6.2).

Если G — симметричная матрица и

$$G = T\,\mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_n)\,T^T,$$

где T — ортогональная матрица, столбцы которой есть собственные векторы матрицы G, то любая матричная функция $F\ G$ представима в виде

$$F\ G\ = T\,\mathrm{diag}\ F\ \lambda_1\ , F\ \lambda_2\ , ..., F\ \lambda_n\ \ T^T.$$

Матричная функция $F\ G$ имеет смысл, если скалярная функция $F\ \lambda$ определена в точках $\lambda_1, \lambda_2, ..., \lambda_n$.

Теорема 6.1. Для выполнения условия

$$f\ x^{k+1}\ \le f\ x^k \tag{6.6}$$

при $\forall x^k \in R^n$ необходимо и достаточно, чтобы

$$\left|R\ \lambda_i\ \right| \ge 1,\ \lambda_i < 0;\ \ \left|R\ \lambda_i\ \right| \le 1,\ \lambda_i > 0 \tag{6.7}$$

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 после разложения x^k по векторам базиса получим

$$x^k = \sum_{i=1}^{n} \xi_{i,k} u^i,$$

$$x^{k+1} = x^k - H_k\ G_k, h_k\ f'\ x^k\ = E - H_k G_k\ x^k =$$

$$= \sum_{i=1}^{n} \xi_{i,k} \left[1 - H_k\ \lambda_i, h_k\ \lambda_i\right] u^i = \sum_{i=1}^{n} \xi_{i,k} R\ \lambda_i\ u^i.$$

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

$$\begin{aligned} f\ x^k\ &= 0{,}5 \sum_{i=1}^{n} \xi_{i,k}^2 \lambda_i; \\ f\ x^{k+1}\ &= 0{,}5 \sum_{i=1}^{n} \xi_{i,k+1}^2 \lambda_i = 0{,}5 \sum_{i=1}^{n} \xi_{i,k}^2 \lambda_i R^2\ \lambda_i \end{aligned} \tag{6.8}$$

следует, что при выполнении (6.7) каждое слагаемое суммы в представлении $f\ x^k$ не возрастает. Достаточность доказана. Докажем необходимость. Рассмотрим первое условие (6.7). Пусть существует такой $i = i_0$, для которого $\left|R\ \lambda_{i_0}\ \right| < 1$, $\lambda_{i_0} < 0$. 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 противо-

речит условию релаксационности (6.6). Аналогично рассматривается второе неравенство (6.7). Теорема доказана.

Замечание 1. Для строгого выполнения неравенства (6.6) необходимо и достаточно кроме (6.7) потребовать, чтобы существовал такой $i = i_0$, для которого $\xi_{i_0,k} \neq 0$ и соответствующее неравенство (6.7) было строгим.

Замечание 2. Выражения (6.8) позволяют оценить скорость убывания функционала f в зависимости от "запаса", с которым выполняются неравенства (6.7). 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

$$2\left|f\ x^k\ - f\ x^{k+1}\ \right| = \Sigma^+ \xi_{i,k}^2 \lambda_i^+ \left[1 - R^2\ \lambda_i^+\ \right] + \Sigma^- \xi_{i,k}^2 \left|\lambda_i^-\right| \left[R^2\ \lambda_i^-\ - 1\right].$$

Из полученного выражения следует, что наибольшее подавление будут испытывать слагаемые, для которых величина множителя релаксации наиболее существенно отличается от единицы (при выполнении условий (6.7)).

Далее будут рассматриваться в основном зависимости $R_h\ \lambda\ $, обладающие свойством

$$R_h\ \lambda\ \to 1,\ h \to 0. \tag{6.9}$$

В этом случае из равенства

$$\left\|x^{k+1} - x^k\right\| = \sum_{i=1}^{n} \xi_{i,k}^2 \left[R_{h_k}\ \lambda_i\ - 1\right]^2 \tag{6.10}$$

следует, что для $\forall \zeta_k \in R^1$ всегда можно выбрать такой h_k, что $\left\|x^{k+1} - x^k\right\| \le \zeta_k$.

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 (6.3).

Иногда для ограничения нормы (6.10) параметр h может вводиться в схему оптимизации как множитель в правой части (6.2):

$$x^{k+1}\ h\ = x^k - hH_k g\ x^k\ ,\ h \in\ 0, 1\ . \tag{6.11}$$

При этом $\left\|x^{k+1}\ h\ - x^k\right\| = h\left\|x^{k+1} - x^k\right\|$, а второе равенство (6.8) трансформируется к виду

$$f\ x^{k+1}\ = 0{,}5 \sum_{i=1}^{n} \xi_{i,k}^2 \lambda_i \bar{R}^2\ \lambda_i\ ,$$

где $\bar{R}\ \lambda_i\ = 1 - h\ + hR\ \lambda_i\ $. Таким образом, новые множители релаксации $\bar{R}\ \lambda_i$ принимают промежуточные значения между 1 и $R\ \lambda_i\ $, что и требуется для обеспечения свойства релаксационности, определяемого требованиями (6.9).

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 (6.2) построить функцию релаксации, характеризующую область его релаксационности в множестве собственных чисел матрицы Гессе аппроксимирующего параболоида. При этом роль тестового функционала играет квадратичная зависимость (6.3). Требуемый характер функции релаксации представлен на рис. 6.1; заштрихована "запрещенная" область, где условия релаксационности (6.7) не выполняются.

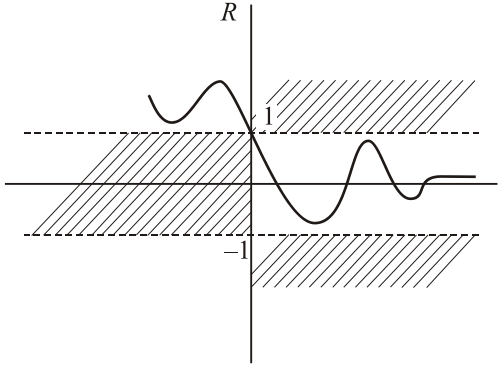


Рис. 6.1. Общий вид функции релаксации

Важное свойство функций релаксации заключается в возможности использования соответствующих представлений для синтеза новых процедур (6.2), обладающих некоторыми желательными свойствами при решении конкретных классов задач параметрической оптимизации.

6.2. Классические градиентные схемы

Рассмотрим некоторые конкретные методы (6.2) и отвечающие им функции релаксации.

6.2.1. Простой градиентный спуск (ПГС)

Формула метода ПГС имеет вид

$$x^{k+1} = x^k - hg\ \ x_k\ \ ,\ h \in R^1. \tag{6.12}$$

Соответствующая функция релаксации линейна (рис. 6.2):

$$R\ \ \lambda\ \ = 1 - \lambda h. \tag{6.13}$$

Пусть собственные значения матрицы G_k расположены в замкнутом интервале $m,\ M$, причем $M \gg m > 0$, так что $\eta = \dfrac{M}{m} >> 1$. В этом случае условие (6.9), очевидно, выполняется, а неравенства (6.7) сводятся к требованию $\left| R\ \lambda_i \right| \le 1,\ i \in 1:n$ или

$$\left|1 - h\lambda_i\right| \le 1,\ i \in 1:n\ . \tag{6.14}$$

Эти оценки иллюстрируются рис. 6.2. Точка пересечения прямой $R\ \lambda$ с осью абсцисс есть точка $\lambda = 1/h$, 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 M/m. Будем считать, что для собственных чисел матрицы G_k выполняются неравенства

$$\lambda_1 \ge \ldots \ge \lambda_{n-r} \ge \sigma\left|\lambda_{n-r+1}\right| \ge \ldots \ge \sigma\left|\lambda_n\right|,\ \ \sigma \gg 1,$$

характерные для овражной ситуации. Тогда для точки $x^k \in Q$, где Q — дно оврага, имеем согласно (6.10)

$$\left\| x^{k+1} - x^k \right\| \cong 4 \sum_{i=n-r+1}^{n} \xi_{i,k}^2 \left(\frac{\lambda_i}{\lambda_1} \right)^2 \le 4\sigma^{-2} \sum_{i=n-r+1}^{n} \xi_{i,k}^2,$$

что может быть существенно меньше ζ_k.

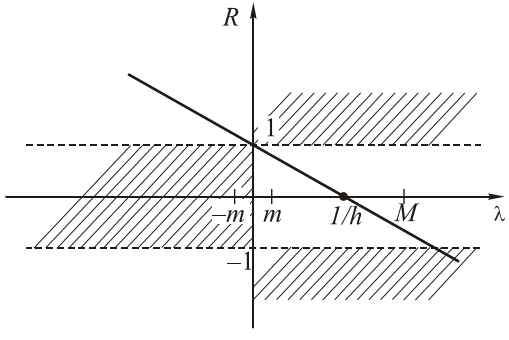


Рис. 6.2. Функция релаксации метода простого градиентного спуска

В результате соответствующие малым собственным значениям из окрестности $\lambda = 0$ слагаемые в (6.8) практически не будут убывать, а продвижение будет сильно замедленным. Это и определяет низкую эффективность метода (6.12).

В области $\lambda < 0$ функция (6.13) 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

$$x^{k+1} \in \operatorname{Arg}\min_{h\geq 0} J\left[x^k - hg\ x^k \right],$$

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. 6.3.

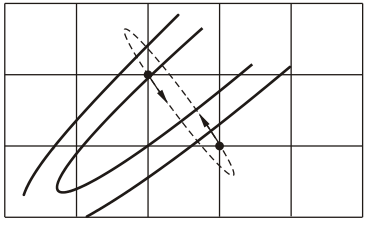


Рис. 6.3. Остановка метода ПГС в овражной ситуации

Как указывалось в *разд. 4.3*, 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 $\left\| x - x^0 \right\| \leq \zeta_0$, и для $J\ x$ справедлива квадратичная аппроксимация

$$f\ x = 0{,}5\langle Gx,\ x\rangle - \langle b,\ x\rangle,\ \ \lambda\ G\ \in\ -m,\ M\ . \tag{6.15}$$

Метод (6.12) для (6.15) примет вид

$$x^{k+1} = x^k - h\ Gx^k - b\ = Bx^k + d, \tag{6.16}$$

где $B \triangleq E - hG,\ d \triangleq hb.$

Записывая (6.16) для двух последовательных номеров k и вычитая полученные равенства, приходим к соотношению

$$y^k \triangleq x^{k+1} - x^k = B\ x^k - x^{k-1}\ = B^k\ x^1 - x^0\ = B^k y^0. \tag{6.17}$$

Согласно хорошо известному из курса численного анализа степенному методу определения максимального собственного числа симметричной матрицы, в результате проведения процесса (6.17) может быть получена оценка максимального собственного числа матрицы B:

$$\frac{\left\| y^{k+1} \right\|}{\left\| y^{k} \right\|} \to \max_i \left| \lambda_i \ \ B \ \right|, \ \ k \to \infty. \tag{6.18}$$

Полагая, что шаг h в итерационном процессе (6.16) выбирается из условия релаксационности (6.1.7), можно заключить, что $h \le 2/M$. Построив график функции релаксации $1 - h\lambda$ как функции от λ при фиксированном h, легко установить, что при $h \le 2/M$ будем иметь:

$$\left| 1 - hM \right| \le 1; \ \ 1 + mh > 1.$$

Таким образом,

$$\max_i \left| \lambda_i(B) \right| = \max_i \left| 1 - h\lambda_i(G) \right| = 1 + mh.$$

Следовательно, для достаточно больших k

$$\frac{\left\| y^{k+1} \right\|}{\left\| y^{k} \right\|} = \frac{\left\| g \ \ x^{k+1} \right\|}{\left\| g \ \ x^{k} \right\|} = \mu_k \cong 1 + mh > 1. \tag{6.19}$$

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 (6.16) до тех пор, пока он не замедлится и отношение $\left\| g^{k+1} \right\| / \left\| g^{k} \right\|$ 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 (6.19) 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 (6.16) 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.

6.2.2. Метод Ньютона

Схема Н-метода основана на формуле (4.29):

$$x^{k+1} = x^k - h_k\ G_k\ ^{-1} g^k, \tag{6.20}$$

где $0 \le h_k \le 1$.

Соответствующая методу (6.20) функция релаксации имеет вид

$$R\ \lambda\ = 1 - H\ \lambda,\ h_k\ \lambda = 1 - h_k,\ \ \lambda \ne 0. \tag{6.21}$$

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 *разд. 5.3* и *5.4*, оказывается менее существенным.

6.2.3. Метод Левенберга

Если известно, что собственные значения матрицы G_k расположены в интервале $-m,\ M\ $, где $M \gg m$, то можно построить метод, имеющий нелинейную функцию релаксации (рис. 6.4)

$$R\ \lambda\ = \frac{h}{h+\lambda},\ h > 0, \tag{6.22}$$

удовлетворяющую требованиям (5.1.7) при $\forall \lambda \in\ -m,\ M\ $, если $h > m$.

Соответствующий метод предложен Левенбергом для специальных классов оптимизационных задач метода наименьших квадратов и имеет функцию

$$H\ \lambda,\ h\ = \frac{1-R}{\lambda} = \ h + \lambda\ ^{-1}. \tag{6.23}$$

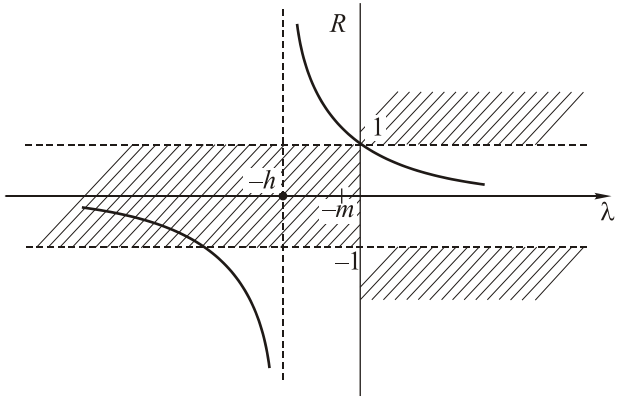


Рис. 6.4. Функция релаксации метода Левенберга

Соответствующий метод называется *методом Левенберга*.

Схема метода (6.2) с функцией (6.23) имеет вид

$$x^{k+1} = x^k - \left(hE + G_k\right)^{-1} g^k \tag{6.24}$$

и реализует некоторый метод доверительной окрестности *(см. разд. 4.5.2)*. 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 $\left\|x^{k+1} + x^k\right\| \le \Delta$, где величина Δ может меняться от итерации к итерации.

Реализация метода (6.23) сводится к решению на каждом шаге по k линейной алгебраической системы

$$\left(hE + G_k\right)\Delta x^k = -g^k,\ \Delta x^k \triangleq x^{k+1} - x^k. \tag{6.25}$$

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

$$h \ge \max\left(\varepsilon_{\text{м}} n \left\|G_k\right\|, \left|\min_i \lambda_i\ G_k\right|\right), \tag{6.26}$$

где $\varepsilon_{\text{м}}$ — машинное эпсилон. При этом оценка для m существенно ухудшается при возрастании размерности n. Правая часть неравенства (6.26) обусловлена тем, что абсолютная погрешность представления любого собственного числа матрицы G_k за счет ограниченности разрядной сетки равна [81]

$$\left|d\lambda_i\right| \le n\lambda_1\varepsilon_{\text{м}} \cong n\left\|G_k\right\|\varepsilon_{\text{м}}.$$

При невыполнении условия $h > m$ система (6.25) 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 (6.25) с различными пробными значениями h.

Легко видеть, что число обусловленности матрицы $hE + G_k$ может превышать $\operatorname{cond} G_k$. 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 $\operatorname{cond}(hE + G_k) \cong \dfrac{9\lambda_{\max}(G_k)}{|\lambda_{\min}(G_k)|}$.

При выборе заведомо больших значений h, что реализуется, например, когда определяющим в (6.26) является первое выражение в скобках, мы имеем $m \ll h$ 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, т. к. 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 $\dfrac{\eta(x^k)\,|\varepsilon_{\text{м}} J(x^k)|}{\|g^k\|}$. 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 (6.25) 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 (6.25) 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 *разд. 4.5.4*.

6.3. Методы с экспоненциальной релаксацией

Начало систематического изучения методов описываемого класса было положено в книге [57]. Далее рассматривается другой подход к построению и анализу алгоритмов, основанный на понятии функции релаксации [81].

Исходя из вышеизложенных требований к функциям релаксации, естественно рассмотреть экспоненциальную зависимость вида (рис. 6.5)

$$R\ \lambda\ = \exp\ -\lambda h\ ,\ \ h > 0, \tag{6.27}$$

для которой условие (5.1.7) 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.

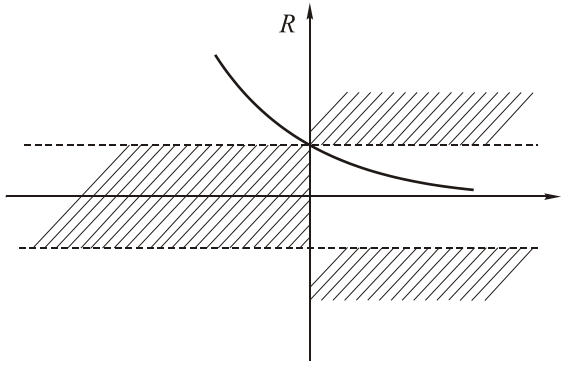


Рис. 6.5. Экспоненциальная функция релаксации

Функция (6.27) обобщает ранее рассмотренные функции релаксации и является в определенном смысле оптимальной. Действительно, раскладывая экспоненту (6.27) в ряд Тейлора и ограничиваясь двумя первыми членами разложениями, получим

$$\exp\ -\lambda h\ = \frac{1}{\exp\ \lambda h} \cong \frac{1}{1+\lambda h} = \frac{h'}{h'+\lambda},\ \ h' = \frac{1}{h},$$

что совпадает с (6.22) при $h = h''$. И аналогично, полагая $\exp\ -\lambda h\ = 1 - \lambda h$, приходим к зависимости (6.13). Для достаточно больших значений параметра h имеем $\exp\ -\lambda h\ \cong 0$ при любых $\lambda \geq m \geq 0$, что позволяет говорить о вырождении метода в классический метод Ньютона без регулировки шага.

Для представления схемы метода в виде (6.2) необходимо определить соответствующую функцию $H\ \lambda,\ h\ $. Имеем

$$\lambda H\ \lambda,\ h\ = 1 - R\ \lambda\ = 1 - \exp\ -\lambda h\ .$$

Полагая $\lambda \neq 0$, получим

$$H\ \lambda,\ h\ = \lambda^{-1}\left[1 - \exp\ -\lambda h\ \right] = \int_0^h \exp\ -\lambda\tau\ d\tau. \tag{6.28}$$

Доопределяя $H\ 0,\ h\ $ из условия непрерывности, получим $H\ 0,\ h\ = h$. В результате схема *метода с экспоненциальной релаксацией (ЭР-метода)* примет вид

$$x^{k+1} = x^k - H\ G_k,\ h_k\ \ g\ \ x_k\ , \tag{6.29}$$

$$H\ \ G,\ h\ = \int_0^h \exp\ -G\tau\ \ d\tau. \tag{6.30}$$

Параметр h_k определяется равенством

$$h_k \in \operatorname{Arg}\min_{h \geq 0} J\left[\ x^k - H\ \ G_k,\ h\ \ g\ \ x^k\ \right], \tag{6.31}$$

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 (6.29), (6.30) сходится практически при тех же ограничениях на минимизируемый функционал, что и метод наискорейшего спуска [32], 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) [57].

Теорема 6.2. Пусть:

1. $J\ \ x\ \in C^2\ \ R^n\ $.

2. Множество $X_* = \left\{\ x^*\ \middle|\ J\ \ x^*\ = \min J\ \ x\ \right.$ непусто.

3. Для любого $\varepsilon > 0$ найдется такое $\delta > 0$, что $\|g\ x\| \geq \delta$, если $x \notin S\ X_*$, где

$$S\ X_* = x \big| d\ x, X_*\ \leq \varepsilon\ ,\ d\ x, X_*\ = \min_{x^* \in X_*} \|x - x^*\|.$$

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 (6.29), (6.30), выполняются предельные соотношения

$$\lim d\ x^k, X_*\ = 0, \tag{6.32}$$

$$\lim J\ x^k\ = J\ x^*\ ,\ k \to \infty. \tag{6.33}$$

Доказательство. Используя соотношения

$$J\ x + y\ = J\ x\ + \int_0^1 \langle g\ x + \vartheta y\ ,\ y \rangle d\vartheta,$$

$$\left| \int_0^1 \langle x\ \vartheta\ ,\ y\ \vartheta\ \rangle d\vartheta \right| \leq \int_0^1 \|x\ \vartheta\ \| \cdot \|y\ \vartheta\ \| d\vartheta$$

и обозначая $J_k = J\ x^k$, $g_k = J'\ x^k$, $G_k = J''\ x^k$, получим

$$\begin{aligned}
&J_k - J\left[x^k - H\ G_k, h\ g^k\right] = \int_0^1 \left\langle g\left[x^k - \vartheta H\ G_k, h\ g^k\right], H\ G_k, h\ g^k \right\rangle d\vartheta = \\
&= \left\langle H\ G_k, h\ g^k, g^k \right\rangle - \int_0^1 \left\langle g^k - J\left[x^k - \vartheta H\ G_k, h\ g^k\right], H\ G_k, h\ g^k \right\rangle d\vartheta \geq \\
&\geq \left\langle H\ G_k, h\ g^k, g^k \right\rangle - l\left\|H\ G_k, h\ g^k\right\|^2 \int_0^1 \vartheta d\vartheta = \\
&= \left\langle H\ G_k, h\ g^k, g^k \right\rangle - \frac{l}{2}\left\|H\ G_k, h\ g^k\right\|^2 \geq \rho\left\|g^k\right\|^2 - \frac{l}{2}\left\|g^k\right\|^2 R^2 = \alpha\left\|g^k\right\|^2,
\end{aligned} \tag{6.34}$$

$$\alpha \triangleq \rho - \frac{l}{2}R^2.$$

При этом использованы неравенства

$$\rho\|y\|^2 \leq \left\langle H\ G_k,\ h\ y,\ y \right\rangle \leq R\|y\|^2, \tag{6.35}$$

где ρ, R — соответственно, минимальное и максимальное собственные числа положительно определенной матрицы $H\ G_k, h$.

Левое неравенство (6.35) следует из представления минимального собственного числа λ любой симметричной матрицы B в виде $\lambda = \min\limits_{x \neq 0} \langle Bx, x\rangle / \langle x, x\rangle$, 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 предположению 5 в теореме 6.2, 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

$$\alpha \ge\ e-1\ \frac{2M - le\ \ e-1}{2M^2 e} > 0. \tag{6.36}$$

Из (6.31), (6.34), (6.36) следует

$$J_k - J_{k+1} \ge J_k - J\left[x^k - H\ \ G_k,\ M^{-1}\ \ g^k\ \right] \ge \alpha \left\| g^k \right\|^2. \tag{6.37}$$

Значит, 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 (6.37) следует $\left\| g^k \right\|^2 \le \alpha^{-1}\ J_k - J_{k+1}$, поэтому $\left\| g^k \right\| \to 0$ при $k \to \infty$. А так как по условию $\left\| g^k \right\| \ge \delta$ при $x^k \in S\ X_*$, то найдется такой номер N, что $x^k \in S\ X_*$ при $k \ge N$ и, следовательно, справедливо утверждение (6.32).

Обозначим через $\bar{x}^k$ проекцию x^k на множество X_*. Тогда по теореме о среднем $J_k - J\ \bar{x}^k = \left\langle g\ x^{kc}\ , x^k - \bar{x}^k \right\rangle$, 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

$$J_k - J\ \bar{x}^k = \left\langle g\ x^{kc} - g\ \bar{x}^k\ , x^k - \bar{x}^k \right\rangle \le$$
$$\le \left\| g\ x^{kc} - g\ \bar{x}^k \right\| \cdot \left\| x^k - \bar{x}^k \right\| \le l d^2\ x^k, X_*\ .$$

И в силу (6.32) получаем (6.33). Теорема доказана.

Замечание 1. Утверждения теоремы, очевидно, выполняются, если h_k выбирать не из условия (6.31), а из условия $J\left[x^k - H\ G_k, h_k\ g^k \right] = \min\limits_{h \in [0,\ \bar{h}]} J\left[x^k - H\ G_k, h\ g^k \right]$, где $\bar{h} > 0$ — произвольное число. Действительно, легко видеть, что равенство (6.37) 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 $h = 1/M$), и в то же время, если при $h = 1/M$ 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 (6.32), (6.33) сохраняются также при замене условия 5 следующим:

$$J_{k+1} = J\left[x^k - H\ G_k, h_k\ g^k \right] \le 1 - \gamma_k\ g^k + \gamma_k \min_{h>0} J\left[x^k - H\ G_k, h\ g^k \right],$$
$$0 < \gamma < \gamma_k \le 1. \qquad (6.38)$$

Действительно, из (6.38) будем иметь

$$J_k - J_{k+1} \ge \gamma_k\ J_k - \min_{h>0} J\left[x^k - H\ G_k, h\ g^k \right] \ge \gamma_k\ J_k - J\left[x^k - H\ G_k, h\ g^k \right]$$

и, согласно (6.34), $J_k - J_{k+1} \ge \gamma_k \alpha \left\| g^k \right\|^2 = \bar{\alpha} \left\| g^k \right\|^2$, $\bar{\alpha} > 0$.

Получено неравенство, аналогичное (6.37), 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 6.3. Пусть:

1. $J\ x \in C^2\ R^n$.

2. Для любых $x, y \in R^n$ выполняются условия $\lambda\|y\|^2 \le \langle G\ x\ y,\ y\rangle \le \Lambda\|y\|^2$, $\|G\ x+y\ -G\ y\ \| \le L\|x\|,\ \lambda > 0,\ L \ge 0.$

Тогда независимо от выбора начальной точки x^0 для метода (6.29) справедливы соотношения (6.32), (6.33), и оценка скорости сходимости

$$\left\|x^{k+1} - x^*\right\| \le \left(\frac{\Lambda}{\lambda}\right)^{1/2} \frac{L\left\|x^k - x^*\right\|^2}{2\lambda}.$$

Доказательство содержится в книге [57].

Таким образом, установлена квадратичная скорость сходимости, характерная для Н-методов.

6.3.1. Реализация методов с экспоненциальной релаксацией

Алгоритм вычисления матричных функций (5.30) может быть основан на использовании известного рекуррентного соотношения

$$H\ G,\ 2h\ = H\ G,\ h\ \left[2E - GH\ G,\ h\ \right]. \tag{6.39}$$

Так как все рассматриваемые матричные функции симметричны и, следовательно, обладают простой структурой, то для доказательства (6.39) достаточно проверить его для соответствующих скалярных зависимостей, что тривиально.

Формула (6.39) используется также для получения обратной матрицы G^{-1}, т. к. 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 (6.31). 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:

$$H\ G_k, h_0\ \cong h_0 \sum_{i=1}^{l} \frac{-G_k h_0\ ^{i-1}}{i!}. \tag{6.40}$$

Далее последовательно наращивают h с помощью соотношения (6.39), вычисляя каждый раз значение $J\left[x^k - H\ G_k, 2^q h_0\ J_k'\right]$, $q = 0, 1, \ldots$ 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 (6.31) 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 (6.39) 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 (6.39) при этом оказывается сравнительно небольшим, что подтверждается опытом практического применения (6.39) в качестве алгоритма построения обратной матрицы.

Сказанное подтверждается следующими рассуждениями для случая $G > 0$.

Соотношение (6.39) может быть преобразовано к виду

$$E - GH\ \ G,\ 2h_0\ = \left[\ E - GH\ \ G,\ h_0\ \right]^2.$$

Из равенств $\left\| E - GH\ \ G,\ h_0\ \right\| = \left\| \exp\ -h_0 G\ \right\| = \exp\ -mh_0\ ,$ при $h_0 = 0{,}1/M$, где $m = \min\limits_i \lambda_i\ \ G\ ,\quad M = \max\limits_i \lambda_i\ \ G$ следует, что $\left\| E - GH\ \ G,\ h_0\ \right\| = \exp\ -0{,}1m/M\ ;$ рассмотренный выбор h_0 аргументируется позже. Поэтому

$$E - GH\ \ G,\ 2^q h_0\ \ =\ \ E - GH\ \ G,\ h_0\ \ ^{2^q}$$

и

$$\left\| E - GH\ \ G,\ 2^q h_0\ \right\| \le \left\| E - GH\ \ G,\ h_0\ \right\|^{2^q} = \exp\left(-0{,}1 \cdot 2^q\ \frac{m}{M} \right).$$

Необходимое число итераций q можно определять из условия выполнения с машинной точностью равенства $\left\| E - GH\ \ G,\ 2^q h_0\ \right\| = 2^{-t}$, 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 (6.39) 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 (6.31) значение h_*. Далее полагаем $\varphi\ h \triangleq J\left[x^k - H\ G_k, h\ g^k\right]$. 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', при выборе нового параметра 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 *разд. 5.3* и *5.4*. Рассмотрим наиболее универсальный алгоритм, основанный на конечноразностных соотношениях. В результате вычислений по формулам (5.15), (5.15) приходим к матрице $G_k=\dfrac{D_k}{\beta_k^2}$ и к вектору $g^k=\dfrac{d^k}{\beta_k}$, 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

$$H\ D_k, h\ = \int_0^h \exp\ -D_k\tau\ d\tau = \beta_k^{-2}\int_0^h \exp\ -G_k\beta_k^2\tau\ d\beta_k^2\tau = \beta_k^{-2}\int_0^{\beta_k^2 h} \exp\ -G_k t\ dt,$$

или

$$\beta_k^2 H\ \beta_k^2 G_k, h\ = H\ G_k, h_k\ ,\quad h_k=\beta_k^2 h. \tag{6.41}$$

С учетом (6.41) основное соотношение (6.29) приводится к виду

$$x^{k+1} = x^k - H\ \ G_k,\ h_k\ \ g^k = x^k - \beta_k^2 H\left(D_k,\ \frac{h_k}{\beta_k^2}\right)\frac{d^k}{\beta_k} = \\ = x^k - 2s_k H\ \ D_k,\ h\ \ d^k,\quad h = \frac{h_k}{4s_k^2}. \tag{6.42}$$

Имеем также

$$H\ \ D_k,\ 2h\ = H\ \ D_k,\ h\ \left[2E - D_k H\ \ D_k,\ h\ \right].$$

Оптимальное значение h находится непосредственно из соотношения:

$$J\ \ x^{k+1}\ = \min_{h>0} J\left[x^{k+1} - 2s_k H\ \ D_k,\ h\ \ d^k\ \right].$$

При использовании разностного уравнения (6.42) укрупненная вычислительная схема метода с экспоненциальной релаксацией может быть сведена к представленной далее 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

$$D_{ij} = J\ \ x + se_i + se_j\ \ - J\ \ x - se_i + se_j\ \ - \\ -J\ \ x + se_i - se_j\ \ + J\ \ x - se_i - se_j\ \ , \tag{6.43}$$

$$i,\ j \in 1:n\ ;$$

$$d_i = J\ \ x + se_i\ \ - J\ \ x - se_i\ \ ,\ \ i \in 1:n\ ,\ \ e_i = \ 0,\ ...,\ 1,\ ...,\ 0\ ; \tag{6.44}$$

принять $h_0 := \dfrac{0{,}1}{\|D\|}$.

Шаг 4. Принять $k = 0$. Вычислить матрицу $H = H\ \ D,\ h_0\ \ :$

$$H = \sum_{i=1}^{7}\ -D\ ^{i-1}\frac{h_0^i}{i!}. \tag{6.45}$$

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 6.2, гарантирует убывание функционала:

$$J\left[x^k - H(G_k),\ h_0\ g^k\right] < J(x^k). \tag{6.46}$$

Действительно, как было сказано ранее, для выполнения неравенства (6.46) достаточно положить $h_0 \le \dfrac{1}{M}$, 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 $h_0 = \dfrac{1}{2{,}3\|G\|} \cong \dfrac{0{,}4}{\|G\|}$. Замена коэффициента 0,4 на 0,1 позволяет более точно реализовать шаг 4 алгоритма, одновременно гарантируя выполнение (6.46).

Параметр s_k может меняться в зависимости, например, от величины $\left\|x^k - x^{k-1}\right\|$ так, как это было описано в *разд. 5.3*. Возможны и другие способы регулировки шага.

6.3.2. Области применения и анализ влияния погрешностей

Обратимся к анализу влияния погрешностей вычислений при реализации ЭР-методов.

Рассмотрим итерационный процесс, определяемый рекуррентным соотношением

$$x^{k+1} = x^k - H_k(G_k),\ h_k\ G_k x^k = q(G_k)\ x^k, \tag{6.47}$$

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. 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, т. к. 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

$$\lambda_1 \geq \ldots \geq \lambda_{n-r} \gg \left|\lambda_{n-r+1}\right| \geq \ldots \geq \left|\lambda_n\right|. \tag{6.48}$$

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 *разд. 5.2*, будем считать, что вариации собственных векторов происходят в пределах линейных оболочек

$$M_1 = \sum_{i=1}^{n-r} \alpha_i u^i, \quad M_2 = \sum_{j=n-r+1}^{n} \alpha_j u^j,$$

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

$$U^T G U = \operatorname{diag}\ \lambda_i\ ,\ U = u^1, u^2, \ldots, u^n\ ; \tag{6.49}$$

$$W^T\ G+dG\ W = \operatorname{diag}\ \lambda_i + d\lambda_i\ ,\ W = w^1, w^2, \ldots, w^n\ .$$

Имеем теперь

$$q\ G\ = E - W D_1 W^T W D_2 W^T = E - W D_3 W^T,$$

где

$$D_1 = \operatorname{diag}\left[\int_0^h \exp\ -\lambda_i - d\lambda_i\ \tau\ d\tau\right];$$

$$D_2 = \operatorname{diag}\ \lambda_1 + d\lambda_i\ ; \quad D_3 = D_1 D_2.$$

Таким образом, матрица $q\ G$ имеет собственные векторы w^i и соответствующие им собственные числа $\lambda_i\ q\ = 1 - \lambda_i\ D_3$.

Полагая

$$x^k = \sum_{i=1}^{n} \xi_{i,k} w^j, \quad w^j = \sum_{j=1}^{n} \alpha_{ji} u^i,$$

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

$$x^{k+1} = q\ \ G\ \ x^k = \sum_{i=1}^{n} \xi_{i,k} \lambda_i\ \ q\ \ w^j = \sum_{i=1}^{n} \xi_{i,k+1} w^j;$$

$$f\ x^{k+1}\ = 0{,}5 \sum_{i=1}^{n} \xi_{i,k+1}^2 \left(\sum_{j=1}^{n} \alpha_{ji} \right)^2 \lambda_i = 0{,}5 \sum_{i=1}^{n} \xi_{i,k}^2 \left(\sum_{j=1}^{n} \alpha_{ji} \right)^2 \lambda_i \lambda_i^2\ \ q\ ,$$

где

$$\lambda_i\ \ q\ = 1 - \ \lambda_i + d\lambda_i\ \ \int_0^{h_k} \exp\ \ -\lambda_i - d\lambda_i\ \ \tau\ d\tau = \exp\ \ -\lambda_i - d\lambda_i\ \ h_k\ . \qquad (6.50)$$

Для выполнения неравенства $f\ x^{k+1}\ \le f\ x^k$ согласно теореме 6.1 должны выполняться условия релаксационности

$$\left|\lambda_i\ \ q\ \right| \le 1,\ \ \lambda_i > 0;\ \ \left|\lambda_i\ \ q\ \right| \ge 1,\ \lambda_i < 0. \qquad (6.51)$$

Теперь легко видеть, что если возмущение $d\lambda_i$ таково, что собственное число меняет знак:

$$\text{sign}\ \ \lambda_i\ \ \ne \text{sign}\ \ \lambda_i + d\lambda_i\ \ , \qquad (6.52)$$

то условия (6.51), вообще говоря, нарушаются. Это приводит к резкому замедлению сходимости процесса оптимизации.

Пусть вариация dG матрицы G вызывается только погрешностями округления. Тогда неравенство (6.52) невозможно, если все малые собственные числа ограничены снизу величиной $n\lambda_1\varepsilon_{\text{м}}$. Действительно, в этом случае $\text{sign}\ \ \lambda_i\ \ \ne \text{sign}\ \ \lambda_i + d\lambda_i\ \ ,$ т. к. $\left|d\lambda_i\right| \le n\lambda_1\varepsilon_{\text{м}} \le \left|\lambda_i\right|$. Отсюда имеем следующее ограничение на степени овражности функционалов, эффективно минимизируемых ЭР-методами:

$$\eta\ \ x^k\ \ \le \frac{1}{n\varepsilon_{\text{м}}}. \qquad (6.53)$$

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 сни-

жение эффективности метода. Из (6.53) следует, что трудности возрастают при увеличении размерности n решаемой задачи и уменьшении длины разрядной сетки компьютера. Вычисления с двойной точностью приводят к оценке $\eta\ x^k\ \le \frac{1}{n\varepsilon_{\text{м}}^2}$ 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.

6.4. 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 *разд. 4.5.4* причин. Методы ОПС, а также ЭР-методы неприменимы, т. к. их вычислительные схемы содержат заполненные матрицы размером $n \times n$, 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 (6.2) 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:

$$y_j\ \ x^j,\ x^q\ \le t_j,\ \ j \in 1:q-1\ ;\ \ y_q\ \ x^q\ \le t_q, \tag{6.54}$$

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 $x = \left[x^1,\ x^2,\ ...,\ x^q \right]$ равна

$$n = \sum_{i=1}^{q} n_i. \tag{6.55}$$

Используя технику оптимизации, представленную в *разд. 3.10.2*, можно привести задачу решения системы неравенств (6.54) к виду

$$J\ x\ = \sum_{j=1}^{q} \Psi_j\ \ x^j,\ x^q\ \ \to \min,\ \ x \in R^n, \tag{6.56}$$

где критерий (6.56) является сглаженным вариантом критерия минимального запаса работоспособности.

Функционалы (6.56) возникают и при других постановках задач оптимального параметрического синтеза, не основанных непосредственно на критериях минимального запаса работоспособности. Поэтому задача (6.56) имеет достаточно общий характер.

Далее будут рассмотрены методы решения задачи (6.56) при следующих предположениях:

1. Критерий (6.56) обладает относительно высокой степенью овражности, а его выпуклость гарантируется только в окрестности точки минимума.
2. Размерность (6.55) 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:

$$G\ x\ = \begin{bmatrix} G_{11} & & 0 & & G_{1q} \\ & G_{22} & & & G_{2q} \\ 0 & & G_{33} & & G_{3q} \\ & & & \ddots & \vdots \\ G_{q1} & G_{q2} & G_{q3} & \dots & G_{qq} \end{bmatrix}.$$

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 [13], [25].

6.4.1. Методы с чебышевскими функциями релаксации[1]

Обратимся снова к классу матричных градиентных схем.

Пусть $\lambda_i\ G_k\ \in -m, M\ ,\ M \gg m > 0$. В силу приведенных ранее предположений и сформулированных в *разд. 6.1* 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 2-го рода $P_s\ \lambda$, удовлетворяющие следующим соотношениям [74]:

$$P_1\ \lambda\ =1,\ P_2\ \lambda\ =2\ 1-2\lambda\ ;\ P_{s+1}\ \lambda\ =2\ 1-2\lambda\ P_s\ \lambda\ -P_{s-1}\ \lambda\ . \qquad (6.57)$$

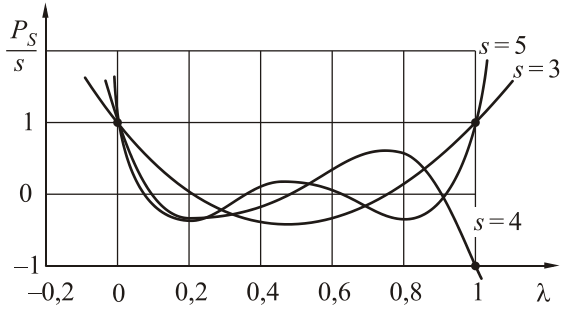


Рис. 6.6. Чебышевские функции релаксации

[1] См. [81].

Графики зависимостей $\frac{P_s\ \lambda}{s}$ для $s = 3,\ 4,\ 5$ представлены на рис. 6.6. Действительно, полагая $R\ \lambda\ = \frac{P_L\ \lambda}{L}$ при достаточно большом значении L получим сколь угодно быструю релаксацию любого слагаемого в представлении

$$f\ \ x^{k+1}\ \ = \frac{1}{2}\sum_{i=1}^{n}\xi_{i,k}^2\lambda_i R^2\ \ \lambda_i\ \ , \tag{6.58}$$

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

$$H\ \lambda\ = \frac{\left[1 - R\ \lambda\ \right]}{\lambda} = \frac{\left[1 - \frac{P_L\ \lambda}{L}\right]}{\lambda}. \tag{6.59}$$

Построение методов (6.2) непосредственно с функцией (6.59) 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 (6.59) 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:

$$x^{k+1} = x^k - \ \alpha_1 E + \alpha_2 G_k + ... + \alpha_{L-1}G_k^{L-2}\ \ g^k = x^k - H\ \ G_k\ \ g^k. \tag{6.60}$$

Реализация метода (6.60) 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 (6.57) можно получить рекуррентное соотношение

$$\begin{aligned} &s+1\ H_{s+1} = 2s\ 1-2\lambda\ H_s - \ s-1\ H_{s-1} + 4s; \\ &H_1 = 0,\ \ H_2 = 2,\ \ s \in\ 2:L-1\ . \end{aligned} \tag{6.61}$$

Из (6.61) 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

$$\begin{aligned} &\vartheta_{s+1} \triangleq x^{k+1}\ s+1\ - x^k = \frac{2s}{s+1}\ E - 2G_k\ \ \vartheta_s - \frac{s-1}{s+1}\vartheta_{s-1} - \frac{4s}{s+1}g^k; \\ &\vartheta_1 = 0,\ \ \vartheta_2 = -2g^k,\ \ s \in\ 2:L-1\ . \end{aligned} \tag{6.62}$$

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 (6.62) 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 (6.58) 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 $\left|R_s\ \lambda\ \right| \le R_s$:

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 (6.58). 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

$$\alpha_s = \frac{1,63}{s^2},\ \ \beta_s = 1 - \alpha_s; \tag{6.63}$$

при этом $R_s < 0,22$.

Соотношение (6.63) получается из следующего представления полиномов Чебышева

$$P_L\ \lambda\ = \frac{\sin L\zeta}{L\sin\zeta},\quad \lambda = \sin^2\frac{\zeta}{2},\quad \lambda,\ \zeta \in [0,\ 1].$$

Действительно, при достаточно малых ζ имеем [39]:

$$P_L\ \lambda\ \cong \Phi\ \xi\ = \frac{\sin\sqrt{\xi}}{\sqrt{\xi}},\ \ \xi \triangleq 4L^2\lambda.$$

График функции $\Phi\ \xi$ представлен на рис. 6.7. Полагая $x = \sqrt{\xi}$, получим

$$\Phi\ \xi\ = \varphi\ x\ = \frac{\sin x}{x}.$$

Последняя функция и ее числовые характеристики показаны на рис. 6.8.

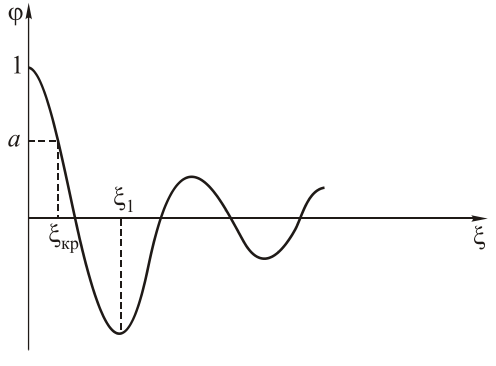


Рис. 6.7. График зависимости $\Phi\ \xi\ = \dfrac{\sin\sqrt{\xi}}{\sqrt{\xi}}$

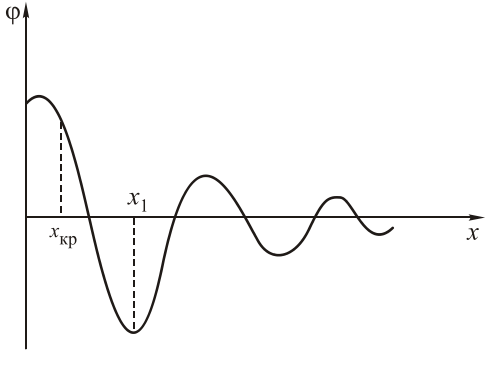


Рис. 6.8. График функции

$$\varphi\ x\ = \frac{\sin x}{x};\ x_1 = 4,4934,\ x_{\text{кр}} = 2,5536,$$

$$\varphi\ x_1\ = -0,2172 \cong -0,22,\ \varphi\ x_{\text{кр}}\ = \left|\varphi\ x_1\right|$$

Из рис. 6.7 и 6.8 имеем

$$\Phi\ \xi\ \le \Phi\ \xi_{\text{кр}} \quad \text{при } \xi \ge \xi_{\text{кр}},$$

где $\xi_{\text{кр}} = x_{\text{кр}}^2 = 6{,}523;\ \ \Phi\ \xi_{\text{кр}}\ = \varphi\ x_{\text{кр}}\ \cong 0{,}22.$

Таким образом, полагая

$$\xi_{\text{кр}} = 4L^2\lambda_{\text{кр}},$$

получим следующее утверждение: если для наименьшего (положительного) собственного числа m выполняется неравенство

$$\xi_{\min} = 4L^2 m \ge \xi_{\text{кр}} = 6{,}523,$$

т. е. если

$$m \ge \frac{6{,}523}{4L^2} = \frac{1{,}63}{L^2}, \tag{6.64}$$

то для всех $\lambda > m$ будем иметь

$$\left|R_L\ \ \lambda\ \right| \le 0{,}22.$$

Из (6.64) следует (6.63).

Укрупненная схема алгоритма, построенного на основе соотношения (6.62), может быть реализована с помощью приведенной далее последовательности шагов. При этом предполагается, что все переменные задачи надлежащим образом нормализованы *(см. разд. 5.3.1)*. 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 (6.62) (об априорном выборе L см. далее).

Шаг 2. Вычислить $g := J'\ \ x\ $, $G := J''\ \ x\ $; принять $g := \dfrac{g}{\|G\|}$; $G := \dfrac{G}{\|G\|}$; $\alpha := 1$.

Шаг 3. По формуле (6.62) построить ϑ_L; положить $x^t := x + \vartheta_L$.

Шаг 4. Вычислить $J_t := J\ \ x^t\ $. Если $J_t > J$, перейти к шагу 5, иначе — к шагу 6.

Шаг 5. Принять $\alpha := \dfrac{\alpha}{2}$, $x^t := x + \alpha\vartheta_L$ 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 (6.62) является параметром, задаваемым пользователем. Согласно (6.63) первоначально целесообразно полагать $L \cong \sqrt{\dfrac{1{,}63}{\alpha_L}} \cong 1{,}3\sqrt{\eta}$, 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 $J_s \triangleq J\ x^k + \vartheta_s$, 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.

6.4.2. Характеристики сходимости и сравнение с методами сопряженных градиентов

Дадим оценку эффективности метода (6.62) 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 [54].

Действительно, рассмотрим случай

$$f\ x = \frac{1}{2}\langle Gx,\ x\rangle, \quad G > 0$$

и оценим скорость сходимости метода СГ к экстремальной точке $x = 0$. Итерация x^k, полученная методом СГ, может быть представлена в виде [74]

$$x^k = E + c_1 G + c_2 G^2 + \ldots + c_k G^k\ x^0 = P_k\ G\ x^0,$$

где $P_k\ G$ — матричный полином k-й степени. При этом из свойств метода СГ следует, что коэффициенты c_1, ..., 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, ..., 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

$$x^k = P_k\ G\ \sum_{i=1}^{n} \xi_{i,0} u^i = \sum_{i=1}^{n} \xi_{i,0} P_k\ \lambda\ u^i,\ P_k\ 0 = 1,$$

$$f\ x^k\ = \frac{1}{2}\left\langle Gx^k,\ x^k \right\rangle = \frac{1}{2}\sum_{i=1}^{n} \xi_{i,0}^2 P_k^2\ \lambda_i\ \lambda_i. \tag{6.65}$$

Отсюда имеем

$$\begin{gathered} \left\| x^0 \right\|^2 = \sum_{i=1}^{n} \xi_{i,0}^2, \\ \left\| x^k \right\|^2 = \sum_{i=1}^{n} \xi_{i,0}^2 P_k^2\ \lambda_i\ \le \max_i P_k^2\ \lambda_i\ \left\| x^0 \right\|^2. \end{gathered} \tag{6.66}$$

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 [8]

$$\tilde{T}_k\ t\ = \frac{T_k\ t}{\cos\ k \arccos t_0} = \frac{T_k\ t}{T_k\ t_0},$$

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,

$$\max_{-1 \le t \le 1} \left| T_k\ t\ \right| = 1,$$

поэтому $L_k = \max_{\lambda} \left| P_k\ \lambda\ \right| = \max_{t} \left| \tilde{T}_k\ t\ \right| = \dfrac{1}{T_k\ t_0}$, $\lambda \in m, M$, $t \in -1, 1$.

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

$$L_k = \frac{2}{\left(\frac{M+m}{M-m}+\sqrt{\left(\frac{M+m}{M-m}\right)^2-1}\right)^k+\left(\frac{M+m}{M-m}-\sqrt{\left(\frac{M+m}{M-m}\right)^2-1}\right)^k} =$$

$$= \frac{2}{\left(\frac{\sqrt{M}+\sqrt{m}}{\sqrt{M}-\sqrt{m}}\right)^k+\left(\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}\right)^k}.$$

При достаточно больших k ($k \geq k_0$) имеем

$$L_k \cong \frac{2}{\left(\frac{\sqrt{M}+\sqrt{m}}{\sqrt{M}-\sqrt{m}}\right)^k} = 2\left(\frac{\sqrt{\eta}-1}{\sqrt{\eta}+1}\right)^k \cong 2\left(1-\frac{2}{\sqrt{\eta}}\right)^k, \quad \eta \triangleq \frac{M}{m}. \tag{6.67}$$

Из (6.66) и (6.67) получаем

$$\left\|x^k\right\| \leq L_k \left\|x^0\right\|$$

или

$$\left\|x^k\right\| \leq 2q^k \left\|x^0\right\|, \quad k \geq k_0, \tag{6.68}$$

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

$$L_k = \frac{2}{\left[\left(1+\frac{2}{\sqrt{\eta}}\right)^k+\left(1-\frac{2}{\sqrt{\eta}}\right)^k\right]}, \quad L_0 = 1.$$

Из (6.68) следует, что при $\eta \gg 1$ сходимость может быть очень медленной.

"Конечность" метода СГ, т. е. 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 (6.65) 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 (6.68) для квадратичной функции общего вида

$$f\ x\ = \frac{1}{2}\langle Gx,\ x\rangle - \langle G,\ x\rangle + c$$

преобразуется к виду

$$\left\|x^k - x^*\right\| \le 2q^k \left\|x^0 - x^*\right\|, \tag{6.69}$$

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 (6.69) скорость сходимости может быть очень малой. Кроме того, каждая итерация метода СГ даже для квадратичного случая требует нового вычисления градиента, т. е. 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 $\left\|x'\right\| \le 0{,}23$ методом СГ и алгоритмом RELCH из начальной точки x^0 с $\left\|x^0\right\| = 1$. По достижении точки x' вся ситуация повторяется, поэтому полученные ниже сравнительные оценки эффективности имеют достаточно общий характер.

Будем предполагать, что для вычисления производных применяются двусторонние конечноразностные соотношения, что в приведенном далее 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 (6.62), не требующих дополнительных вычислений целевого функционала f.

Чтобы получить вектор x', методу СГ потребуется N итераций, где число N определяется из условия (6.69):

$$\left\|x^N\right\| = 2q^N = 0{,}23,$$

т. е. $N \cong -\dfrac{2,2}{\ln q}$. Для выполнения каждой итерации необходимо обновление вектора градиента, что связано с $2n$ вычислениями $f\ x$. Общее число вычислений f равно $-\dfrac{4,4n}{\ln q}$. Относительный выигрыш в количестве вычислений f методом RELCH по сравнению с методом СГ задается функцией $\Psi\ \eta \cong -\dfrac{2,2}{\gamma n \ln q}$. 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 далее:

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 (6.62). 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 положим, 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 $t_y = 3 \cdot 10^{-5}$ секунд. Вычисление значения $f\ x$ занимает при этом $t_f = 0,6$ секунд процессорного времени. Для вычисления f' и f'' с помощью общих конечноразностных формул потребуется, соответственно, $t' = 2nt_f = 4$ мин, $t'' = 2\gamma n^2 t = 20$ мин. Число пересчетов по формуле (6.62) равно $L = 1,3\sqrt{\eta} = 50$. При каждом пересчете производится умножение слабо заполненной матрицы $E - 2G_k$ на вектор ϑ_s, что требует $\gamma n^2 t_y \cong 3 \cdot 10^{-2}$ секунд машинного времени. Время построения вектора ϑ_{50} без учета вычисления f', f'' составит около $50 \cdot 3 \cdot 10^{-2} = 1,5$ секунд 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 $\Psi\ 1500\ \cdot 20 \cong 1,3$ час.

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, т. к. 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.

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. 6.9. 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 (6.39) и (6.62). Характерные зависимости для рассмотренных случаев (для фиксированного $\lambda_i < 0$) и разных $\varepsilon_м$ представлены на рис. 6.10.

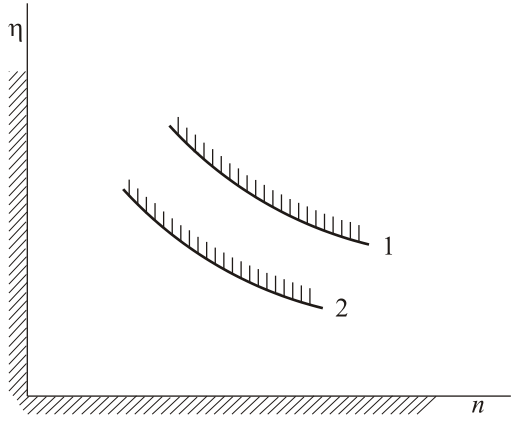


Рис. 6.9. Области работоспособности:1 — $\varepsilon_м = \varepsilon'_м$; 2 — $\varepsilon_м = \varepsilon''_м > \varepsilon'_м$

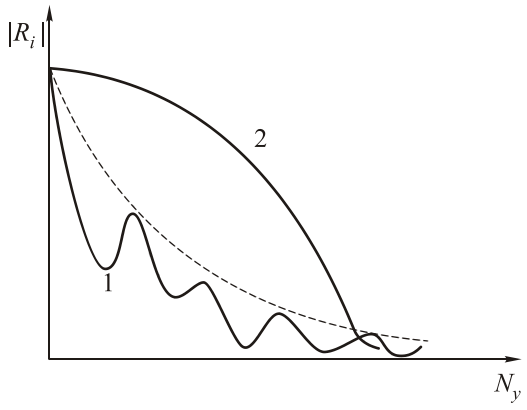


Рис. 6.10. Характер изменения множителей релаксации: 1 — RELCH; 2 — RELEX

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.

6.5. Применение процедур RELEX и RELCH в прикладных задачах теории оптимизации

Рассмотренные процедуры RELEX и RELCH целесообразно реализовывать согласно общим рекомендациям, данным в *разд. 5.3.1* и *6.3*. При этом необходимо использовать специальные приемы вычисления матриц Гессе, основанные на конкретных структурных особенностях отдельных классов задач *(см. разд. 5.4)*. 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. разд. 2.3)*:
   - 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 *разд. 2.3* примеров практических задач 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.

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. 6.11. Один из возможных подходов к формированию соответствующего критерия оптимальности с указанием структуры матрицы Гессе был изложен ранее (см. соотношения (6.54), (6.56)). 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.

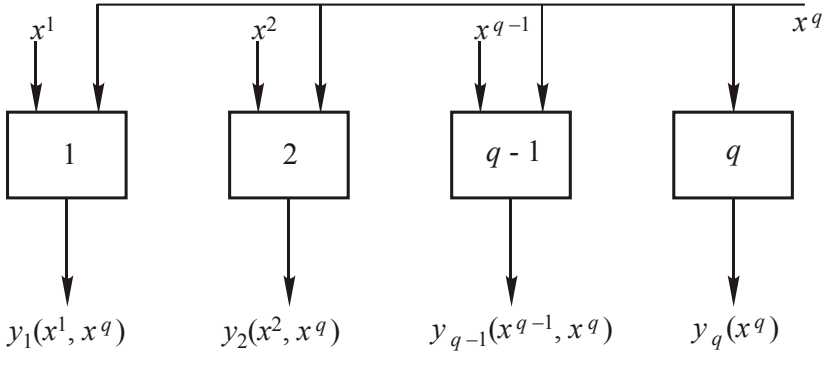


Рис. 6.11. Совокупность взаимосвязанных подсистем

6.6. Тактика решения общей задачи конечномерной оптимизации

На основе представленных в книге методов можно сформулировать общий подход к решению "незнакомой" задачи конечномерной 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. 6.12).

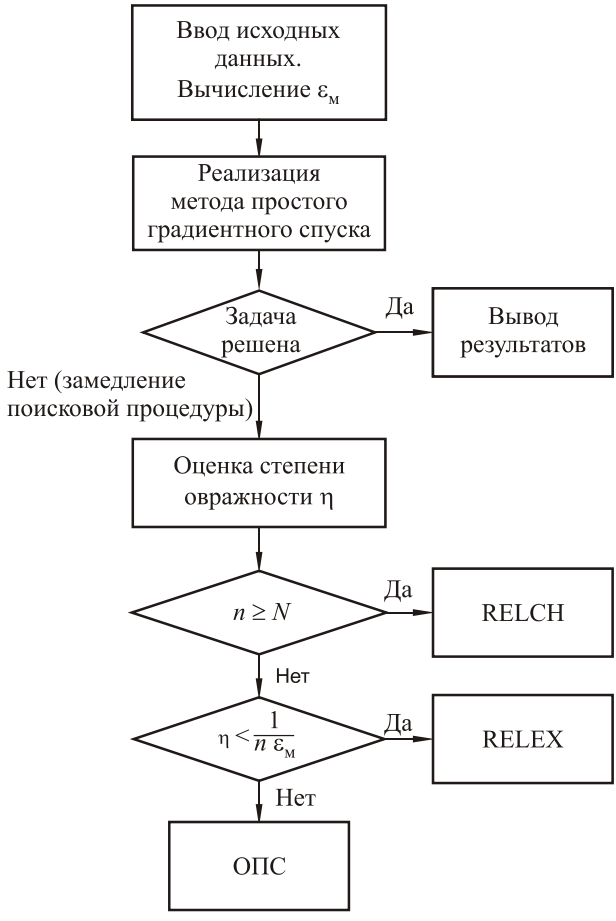


Рис. 6.12. Тактика решения общей задачи конечномерной оптимизации

После ввода исходных данных, определяется машинное эпсилон на основе известных из курсов 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 на методы обобщенного покоординатного спуска как на наиболее мощные и универсальные процедуры.

6.7. Основные результаты и выводы

В *данной главе* представлены следующие основные результаты.

1. Для класса матричных градиентных методов введено новое 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 новый 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.
5. Доказаны теоремы, устанавливающие сходимость ЭР-методов для достаточно широкого класса невыпуклых функционалов и гарантирующие квадратичную

скорость сходимости в предположении сильной выпуклости целевого функционала.

6. 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.
7. Изучены методы конечномерной 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.
8. 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 7

Методы уменьшения размерности вектора аргументов минимизируемых функционалов

В этой главе рассматриваются две причины [84], приводящие к необходимости сокращения размерности пространства управляемых параметров оптимизируемой системы. Первая причина *(см. разд. 7.1)* связана с разработкой методов понижения локальной степени овражности минимизируемых функционалов при решении плохо обусловленных (овражных) оптимизационных задач.

Вторая причина *(см. разд. 7.2)* 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 2*).

7.1. Методы теории жестких систем

Как следует из общего определения овражного (жесткого) 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

$$\frac{dx}{dt} = -g\ x\ ,\ \ g\ x\ \triangleq J'\ x\ \triangleq \frac{\partial J}{\partial x} \tag{7.1}$$

и обобщающего известный *метод квазистационарных концентраций* Семенова — Боденштейна. ПКП для общего нелинейного случая был предложен в [57]. Анализ, приведенный далее, и относящиеся к линейным системам теоремы и вычислительные технологии были получены автором настоящей книги.

7.1.1. Принцип квазистационарности производных для линейных систем с симметричными матрицами

Решение уравнения спуска

$$\frac{dx}{dt} = -Dx + b \tag{7.2}$$

для параболоида

$$f\ x\ = 0{,}5\langle Dx,\ x\rangle - \langle b,\ x\rangle + c,\quad x \in R^n,\ c \in R^1, \tag{7.3}$$

являющегося локальной моделью исходного функционала $J\ x\ $, может быть представлено в виде

$$x\ t\ = \sum_{k=1}^{n}\left[\alpha_k \exp\ -\lambda_k\ \ D\ t\ + \beta_k \int_0^t \exp\ -\lambda_k\ \ D\ \tau\ d\tau\right] u^k, \tag{7.4}$$

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

$$\lambda_1\ \ D\ \geq \ldots \geq \lambda_{n-r}\ \ D\ \geq \sigma\left|\lambda_{n-r+1}\ \ D\ \right| \geq \ldots \geq \sigma\left|\lambda_n\ \ D\ \right|,\quad \sigma \gg 1, \tag{7.5}$$

что соответствует овражной структуре функционала (7.3) в смысле определений, приводимых в *разд. 4.2*.

Умножая обе части равенства (7.4) 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

$$\left\langle x,\ u^i\right\rangle = \alpha_i \exp\left[-\lambda_i\ \ D\ t\right] - \frac{\beta_i}{\lambda_i\ D}\left[\exp\ -\lambda_i\ \ D\ t\ -1\right]_{t>\tau_{\text{пс}}} \cong \frac{\beta_i}{\lambda_i\ D},$$

где $\tau_{\text{пс}}$ — длина *пограничного слоя*, $\tau_{\text{пс}} \sim \dfrac{1}{\sigma}$. 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

$$\langle x,\ u^i \rangle = \lambda_i^{-1}\ \ D\ \langle b,\ u^i \rangle,\ \ i \in 1:n-r\ . \tag{7.6}$$

Полученные соотношения согласно развитым в *разд. 4.2* представлениям определяют дно оврага.

На основе равенств (7.6) для рассматриваемого частного случая (7.2) уравнений спуска ПКП может быть доказан непосредственно.

Теорема 7.1 (ПКП для линейных систем с симметричными матрицами).

Пусть:

1. $x\ \ t$ — решение уравнения (7.2) с постоянной симметричной матрицей *D*.

2. Собственные числа матрицы *D* удовлетворяют неравенствам (7.5).

Тогда существует такой индекс $j \in 1:n$, что уравнение

$$\frac{d^s x_j}{dt^s} = 0 \tag{7.7}$$

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 (7.2) $s-1$ 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

$$c_j^s = A^s e^j,\ \ e^j \triangleq (0,\ ...,\ 1,\ ...,\ 0) = \sum_{k=1}^{n} \langle e^j,\ u^k \rangle u^k \tag{7.8}$$

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

$$\left\langle e^j,\, u^m \right\rangle = \sum_{k=1}^{n} \alpha_k \left\langle u^k,\, u^m \right\rangle = \alpha_m,$$

т. к. $\left\langle u^k,\, u^m \right\rangle = 1$ при $k = m$ и $\left\langle u^k,\, u^m \right\rangle = 0$ — в противном случае. В результате имеем (7.8).

Отсюда

$$\frac{d^s x_j}{dt^s} = \lambda_1^{s-1}\ A\ \sum_{k=1}^{n} u_j^k \left[\frac{\lambda_k\ A}{\lambda_1\ A}\right]^{s-1} \left[\lambda_k\ A\ \left\langle u^k,\, x\right\rangle + \left\langle u^k,\, b\right\rangle\right].$$

Пренебрегая слагаемыми с номерами $k = l+1,\ ...,\ n$, где l — кратность максимального собственного числа $\lambda_1\ D$, получим

$$\frac{d^s x_j}{dt^s} = \lambda_1^s\ A\ \left[\left\langle v^j,\, x\right\rangle + \lambda_1^{-1}(A)\left\langle v^j,\, b\right\rangle\right], \tag{7.9}$$

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 (7.9) видно, что условие квазистационарности (7.7) $(s-1)$-й производной от j-го компонента вектора $x\ t$, соответствующего ненулевому вектору v^j, эквивалентно искомому уравнению связи (7.6) для $i = 1$.

Если выбранный номер j определяет нулевые компоненты $u_j^k,\ k = 1,\ ...,\ l$ 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

$$\forall j \in 1:n\ ,\ \ \forall i \in 1:l\ :\ u_j^i = \left\langle v^j,\, u^i\right\rangle = 0,$$

т. е. $u^i = 0,\ i \in 1:l$, 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 (7.7) 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 (7.9) получим

$$\frac{d^s x_j}{dt^s} = \lambda_1^s \ A \ u_j^1 \left[\left\langle u^1, \ x \right\rangle + \lambda_1^{-1} \left\langle u^1, \ b \right\rangle \right].$$

Ясно, что если для некоторых j имеем $u_j^1 \neq 0$, то соответствующее уравнение (7.7) доставляет искомую линейную связь. Если матрица D не является блочно диагональной матрицей вида

$$D = \begin{bmatrix} L & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & N \end{bmatrix}, \tag{7.10}$$

то пригодно любое уравнение (7.7).

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 некотором s получим

$$c_1^s = \ A^s + \Pi_s \ e^1$$

вместо точного соотношения (7.8).

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 т. к.

$$c_1^{s+m} = A^m c_1^s = \sum_{j=2}^{n} u_1^j \lambda_j^{m+s} \ A \ u^j + \sum_{j=1}^{n} \xi_{s1}^j \lambda_j^m \ A \ u^j,$$

то

$$\frac{d^{s+m} x_1}{dt^{s+m}} = \left\langle c_1^{s+m}, \ x \right\rangle + \left\langle c_1^{s+m-1}, \ b \right\rangle = \sum_{j=2}^{n} u_1^j \lambda_j^{s+m-1} \ A \left[\lambda_j \ A \left\langle u^j, \ x \right\rangle + \left\langle u^j, \ b \right\rangle \right] +$$

$$+ \sum_{j=1}^{n} \xi_{s1}^j \lambda_j^{m-1} \ A \left[\lambda_j \ A \left\langle u^j, \ x \right\rangle + \left\langle u^j, \ b \right\rangle \right].$$

При дальнейшем увеличении m слагаемое

$$\xi_{s1}^1 \lambda_1^m \;\; A \left[\left\langle u^1, \, x \right\rangle + \lambda_1^{-1} \;\; A \;\; \left\langle u^1, \, b \right\rangle \right]$$

начнет быстро возрастать по сравнению с остальными и, в конце концов, станет доминирующим, определяя искомое уравнение связи.

Если матрица D имеет блочную структуру (7.10), то исходная оптимизационная задача распадается на несколько независимых подзадач для разных параболоидов.

На практике целесообразно выбирать уравнение (7.7) 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 (7.9), 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, ибо предполагается, что

$$\lambda_{n-r} \geq \sigma \left| \lambda_{n-r+1} \right|, \quad \sigma \gg 1.$$

Замечание 1. Изложенный метод построения линейных зависимостей (7.6) 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 (7.8) 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 (7.5) 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 *разд. 5.2*, вариации u^1 будут обязательно 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 (7.6).

Аналогично вышеизложенному можно показать, что существуют такие номера j_1, j_2, ..., $j_{n-r} \in 1 : n$, которые определяют асимптотически эквивалентную соотношениям (7.6) подсистему уравнений (7.7). Однако вышеизложенный подход, связанный с последовательным выделением одиночных доминирующих связей, алгоритмически наиболее рационален.

7.1.2. Методы иерархической оптимизации: частный случай

Рассмотрим частный случай задачи минимизации параболоида

$$f\ x\ = 0{,}5\langle Dx,\ x\rangle - \langle b,\ x\rangle + c \tag{7.11}$$

(11)

в предположении

$$\lambda_1\ D\ \geq \sigma\lambda_i\ D\ > 0,\quad \sigma \gg 1,\ i \in\ 2:n\ . \tag{7.12}$$

(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 (7.6) для $i = 1$ выразим компонент x_k вектора x, которому соответствует максимальный по модулю компонент u_k^1 собственного вектора u^1:

$$\begin{aligned} x_k = L\ z\ &= -\frac{1}{u_k^1}\left[\sum_{\substack{j=1\\ j\neq k}}^{n} u_j^1 x_j - \lambda_1^{-1}\ D\ \left\langle b,\ u^1\right\rangle\right], \\ z &\triangleq\ x_1,\ x_2,\ ...,\ x_{k-1},\ x_{k+1},\ ...,\ x_n\ . \end{aligned} \tag{7.13}$$

Определим функционал $f_1\ z\ $:

$$\begin{aligned} f_1\ z\ &\triangleq f\ x_1,\ x_2,\ ...,\ x_{k-1},\ L\ z\ ,\ x_{k+1},\ ...,\ x_n\ = \\ &= 0{,}5\langle f_1'' z,\ z\rangle - \langle d,\ z\rangle + c_1. \end{aligned} \tag{7.14}$$

Теорема 7.2. Пусть заданы функционалы (7.11), (7.14) и выполняются неравенства (7.12). Тогда:

1. Матрица f_1'' положительно определена ($f_1'' > 0$) и минимизаторы x^*, z^* параболоидов f, f_1 связаны соотношениями

$$\begin{aligned} x^* &= z_1^*,\ ...,\ z_{k-1}^*,\ L\ z^*\ ,\ z_k^*,\ ...,\ z_{n-1}^*\ \triangleq\left[L_k\ z^*\ ,\ z^*\right], \\ z^* &= x_1^*,\ ...,\ x_{k-1}^*,\ x_{k+1}^*,\ ...,\ x_n^*\ . \end{aligned} \tag{7.15}$$

2. Спектральные числа обусловленности матриц D, f_1'' удовлетворяют неравенству

$$k\ f_1''\ \leq\ n-1\quad n+1\ \sigma^{-1} k\ D\ . \tag{7.16}$$

Доказательство. Матрицы f'', f_1'', очевидно, не изменятся, если в выражении (7.11) принять $b=0$, $c=0$. При этом для любого $z\neq 0$ по определению $f_1\ z$ будем иметь

$$0{,}5\left\langle f_1''z,\ z\right\rangle = 0{,}5\left\langle Dx,\ x\right\rangle, \tag{7.17}$$

где $x=\left[L_k\ \ z\ ,\ z\right]$.

Выражение справа в равенстве квадратичных форм (7.17) больше нуля из-за условия $D>0$. Отсюда следует неравенство $f_1''>0$.

Минимизатор x^* параболоида $f\ x$, очевидно, удовлетворяет уравнению $x_k^* = L\ z^*$, т. к. имеем

$$\left\langle u^1,\ x^*\right\rangle = \left\langle u^1,\ D^{-1}b\right\rangle = \left\langle D^{-1}u^1,\ b\right\rangle = \lambda_1^{-1}\ D\ \left\langle b,\ u^1\right\rangle.$$

Поэтому $x^*=\left[L_k\ \ \tilde{z}\ ,\ \tilde{z}\right]$, где $\tilde{z}\triangleq\ x_1^*,\ ...,\ x_{k-1}^*,\ x_{k+1}^*,\ ...,\ x_n^*$.

Первое соотношение (7.15) 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

$$f\ \left[L_k\ \ z'\ ,\ z'\right]\ = f_1\ z'\ < f_1\ \tilde{z}\ = f\ x^*\ ,$$

что противоречит свойству минимизатора x^*. Следовательно, $\tilde{z}=z^*$, и первое утверждение теоремы доказано.

Предположим, не ограничивая общности, что $\left|u_1^1\right|\ge\left|u_i^1\right|$, $i\in\ 2:n$. Непосредственным дифференцированием получаем выражение элементов g_{ij} матрицы f_1'' через элементы d_{ij} матрицы D:

$$g_{ij}=d_{ij}=\frac{u_i^1}{u_1^1}d_{1j}-\frac{u_j^1}{u_1^1}d_{1i}-\frac{u_i^1u_j^1}{u_1^{1\ 2}}d_{11};\ \ i,\ j\in\ 2:n\ . \tag{7.18}$$

Учитывая представление

$$d_{ij}=\sum_{k=1}^{n}u_i^ku_j^k\lambda_k\ \ D\ ,$$

получим для диагональных элементов матрицы f_1'':

$$g_{ii}=\sum_{k=1}^{n}\left(u_i^k-\frac{u_i^1}{u_1^1}u_1^k\right)^2\lambda_k\ \ D\ =\sum_{k=2}^{n}\left(u_i^k-\frac{u_i^1}{u_1^1}u_1^k\right)^2\lambda_k\ \ D\ ,\ i\in\ 2:n\ .$$

Суммируя по i, приходим к выражению для следа Sp матрицы f_1'':

$$Sp\ f_1'' = \sum_{i=2}^{n} g_{ii} = \sum_{k=2}^{n} \lambda_k\ D \sum_{i=1}^{n} \left(u_i^k - \frac{u_i^1}{u_1^1} u_1^k \right)^2 = \sum_{k=2}^{n} \lambda_k\ D \left[1 + \left(\frac{u_1^k}{u_1^1} \right)^2 \right], \left(\frac{u_1^k}{u_1^1} \right)^2 \le n.$$

Следовательно,

$$\lambda_1\ f_1'' \le Sp\ f_1'' \le\ n+1 \sum_{k=2}^{n} \lambda_k\ D\ \le\ n+1\ \ n-1\ \sigma^{-1}\lambda_1\ D\ . \qquad (7.19)$$

Пусть $\min\limits_i \lambda_i\ f_1'' \triangleq \lambda_{n-1}\ f_1''$.

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 $x = \left[L\ z\ ,\ z \right]$, получаем согласно равенству (7.17)

$$\frac{\langle Dx,\ x \rangle}{\langle x,\ x \rangle} = \frac{\langle f_1'' z,\ z \rangle}{\langle z,\ z \rangle + L^2\ z} \le \frac{\langle f_1'' z,\ z \rangle}{\langle z,\ z \rangle} = \lambda_{n-1}\ f_1''\ .$$

А т. к.

$$\lambda_n\ D\ = \min_i \lambda_i\ D\ = \min_{x \ne 0} \frac{\langle Dx,\ x \rangle}{\langle x,\ x \rangle},$$

то, очевидно,

$$\lambda_n\ D\ \le \lambda_{n-1}\ f_1''\ . \qquad (7.20)$$

Из (7.19), (7.20) имеем

$$k\ f_1'' = \frac{\lambda_1\ f_1''}{\lambda_{n-1}\ f_1''} \le \frac{n+1\ \ n-1\ \sigma^{-1}\lambda_1\ D}{\lambda_n\ D} =\ n+1\ \ n-1\ \sigma^{-1} k\ D\ , \qquad (7.21)$$

что доказывает неравенство (7.16).

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 (7.13) согласно теореме 7.1 может быть найдена с помощью ПКП.

7.1.3. 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 матрицы D размером $n \times n$ удовлетворяют неравенствам

$$\lambda_1\ D\ \ge \ldots \ge \lambda_{n-r}\ D\ \ge \sigma\left|\lambda_{n-r+1}\ D\ \right| \ge \ldots \ge \sigma\left|\lambda_n\ D\ \right|,$$

так что параболоид

$$f\ x\ = 0{,}5\left\langle Dx,\ x\right\rangle - \left\langle b,\ x\right\rangle + \text{const} \tag{7.22}$$

является *r*-овражным (где *r* — размерность дна оврага) в смысле определений *главы 4*.

Пусть для некоторого $i \in 1:n-r$ задано уравнение

$$\left\langle x,\ u^i\right\rangle = \lambda_i^{-1}\left\langle b,\ u^i\right\rangle. \tag{7.23}$$

Для определенности примем $i = 1$. Пусть

$$k = \arg\max_j \left|u_j^1\right|. \tag{7.24}$$

Тогда согласно (7.23) можно выразить *k*-й компонент вектора *x* через остальные и аналогично (7.14) построить новый функционал

$$\begin{aligned} &f_1\ z\ = 0{,}5\left\langle G_k^D z,\ z\right\rangle - \left\langle d,\ z\right\rangle + \text{const},\\ &G_k^D = f_1''\ z\ ,\\ &z = \ x_1,\ \ldots,\ x_{k-1},\ x_{k+1},\ \ldots,\ x_n\ . \end{aligned} \tag{7.25}$$

Таким образом, определен оператор $D,\ K\ \to G_k^D$, ставящий в соответствие матрице D размером $n \times n$ и номеру $k \in 1:n$ исключаемой переменной матрицу G_k^D размером $n-1\ \times\ n-1$.

Лемма 1. Пусть задано уравнение относительно переменной $\lambda \in R^1$:

$$\lambda + \sum_{i=1}^{k} a_i b_i^2\ \ b_i - \lambda\ ^{-1} = \sum_{i=1}^{k} a_i b_i,\ a_i > 0,\ b_i \ne 0, b_j > b_{j+1},\ j \in 1, k-1\ . \tag{7.26}$$

Тогда:

1. Корни λ_i уравнения (7.26) различны, их число равно $k+1$ и они строго разделяются числами b_i:

$$b_1 < \lambda_1;\ b_i \le \lambda_i \le b_{i-1},\ i \in\ 2:k\ ;\ \lambda_{k+1} < b_k; \tag{7.27}$$

при этом один из корней $\lambda_j = 0,\ j \in 1:k+1$.

2. Если $b_1 > 0,\ b_k < 0,$ то

$$b_k\ (1+s) \le \lambda_i \le b_1\ (1+s)\ ,\quad i \in 1:k+1\ , \tag{7.28}$$

если $b_1 > 0,\ b_k > 0,$ то

$$b_k < \lambda_i \le b_1\ (1+s)\ ,\quad i \in 1:k\ ,\ \lambda_{k+1} = 0; \tag{7.29}$$

если $b_1 < 0,\ b_k < 0,$ то

$$b_k\ (1+s) \le \lambda_i \le b_1,\quad i \in 2:k+1\ ,\ \lambda_1 = 0; \tag{7.30}$$

где $s \triangleq \sum_{i=1}^{k} a_i.$

3. Если $b_1 > |b_i|,\ i \in 2:k$, то корни μ_i уравнения

$$\mu + \sum_{i=2}^{k} a_i b_i^2\ (b_i - \mu)^{-1} = \sum_{i=2}^{k} a_i b_i \tag{7.31}$$

удовлетворяют неравенствам

$$|\mu_j| \ge |\lambda_{j+1}|,\quad j \in 1:k\ . \tag{7.32}$$

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 (7.27).

2. Для доказательства (7.28)—(7.30) достаточно доказать, что для $\lambda_i \ne 0$

$$\underline{\lambda} \le \lambda_i \le \overline{\lambda}, \tag{7.33}$$

где

$$\overline{\lambda} = \begin{cases} b_1, & b_1 < 0, \quad (7.34) \\ b_1\ (1+s)\ , & b_1 > 0; \quad (7.35) \end{cases}$$

$$\underline{\lambda} = \begin{cases} b_k, & b_k > 0, \quad (7.36) \\ b_k\ (1+s)\ , & b_k < 0. \quad (7.37) \end{cases}$$

Пусть

$$\varphi_1(\lambda) = \sum_{i=1}^{k} a_i b_i - \lambda - \sum_{i=1}^{k} a_i b_i^2\ (b_i - \lambda)^{-1}, \tag{7.38}$$

$$\varphi_2(\lambda) = b_1 \sum_{i=1}^{k} a_i - \lambda - b_1^2 \sum_{i=1}^{k} a_i \Big/ (b_1 - \lambda)^{-1}, \tag{7.39}$$

$$\varphi_3(\lambda) = b_k \sum_{i=1}^{k} a_i - \lambda - b_k^2 \sum_{i=1}^{k} a_i \Big/ (b_k - \lambda)^{-1}, \tag{7.40}$$

$$\varphi_4(\lambda) = \sum_{i=2}^{k} a_i b_i - \lambda - \sum_{i=2}^{k} a_i b_i^2 \Big/ (b_i - \lambda)^{-1}. \tag{7.41}$$

Тогда

$$\varphi_1 - \varphi_2 = \sum_{i=1}^{k} a_i (b_1 - b_i) + \sum_{i=1}^{k} a_i \left(\frac{b_1^2}{\lambda - b_1} - \frac{b_i^2}{\lambda - b_i} \right).$$

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 (7.38), (7.39), 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 (7.35) доказано. (7.34) следует из (7.27), т. к. 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 (7.37). Соотношение (7.36) следует из (7.27), т. к. $\lambda_{k+1} = 0$.

3. Докажем третье утверждение леммы. Вычитая (7.38) из (7.41), имеем

$$\varphi_4 - \varphi_1 = a_1 b_1 + \frac{a_1 b_1^2}{b_1 - \lambda} = \frac{a_1 b_1 \lambda}{b_1 - \lambda}. \tag{7.42}$$

На промежутках (b_i, b_{i+1}) для $i \in 1:k-1$, $(-\infty, b_k)$, $(b_1, +\infty)$ $\varphi_1' < 0$, $\varphi_4' < 0$ 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 (7.42) отрицательна и

$$|\mu_i| > |\lambda_{i+1}|.$$

Поэтому в общем случае

$$|\mu_i| > |\lambda_{i+1}|,$$

и неравенство (7.32) доказано.

Теорема 7.3. Пусть заданы параболоиды (7.22), (7.25). Тогда:

1. Для собственных чисел матрицы G_k^D выполняются неравенства

$$\lambda_1\ G_k^D\ \geq \ldots \geq \lambda_{n-r-1}\ G_k^D\ \geq n^{-1}\sigma\left|\lambda_{n-r}\ G_k^D\ \right| \geq \ldots \geq n^{-1}\sigma\left|\lambda_{n-1}\ G_k^D\ \right|; \qquad (7.43)$$

$$\lambda_1\ G_k^D\ \leq n\lambda_2\ D\ ; \qquad (7.44)$$

$$\lambda_{n-r-1}\ G_k^D\ \leq \lambda_{n-r}\ D\ . \qquad (7.45)$$

2. Для "малых" по модулю собственных чисел матриц D и G_k^D выполняются неравенства

$$\left|\lambda_i\ D\ \right| \leq \left|\lambda_{i-1}\ G_k^D\ \right| \leq n\left|\lambda_{n-r+1}\ D\ \right|,\ \ i \in\ n-r+1:n\ . \qquad (7.46)$$

3. Если $D > 0$, то $G_k^D > 0$ и минимизаторы x^*, z^* функционалов f, f_1 связаны соотношениями (7.15).

Доказательство. Принимая без ограничения общности $k = n$, получим представление элементов g_{ij} матрицы G_k^D через элементы d_{ij} матрицы D:

$$g_{ij} = d_{ij} - d_{nj}\frac{u_i^1}{u_n^1} - d_{ni}\frac{u_i^1}{u_n^1} + d_{nn}\frac{u_i^1}{{u_n^1}^2},\ \ i,\ j \in\ 1:n-1\ . \qquad (7.47)$$

При этом, очевидно, $\forall i : g_{in} = 0,\ \forall j : g_{nj} = 0.$

Матрица G_n^D является блоком $(n \times n)$-матрицы $G = \ g_{ij}\ $:

$$G = \begin{bmatrix} G_n^D & 0 \\ 0 & 0 \end{bmatrix}. \qquad (7.48)$$

Введем с помощью преобразования подобия матрицу $F = T^T G T$, где $T = \left[\, u^1,\ u^2,\ \ldots,\ u^n \,\right]$ — ортогональная матрица, составленная из ортонормированных собственных векторов матрицы D.

Непосредственно проверяется, что F является окаймленной диагональной матрицей вида

$$F = \begin{bmatrix} \alpha & a_2 & a_3 & \dots & a_n \\ a_2 & \lambda_2\ D & 0 & \dots & 0 \\ a_3 & 0 & \lambda_3\ D & \dots & 0 \\ . & . & . & \dots & \dots \\ a_n & 0 & 0 & \dots & \lambda_n(D) \end{bmatrix};\ \alpha, a_i \in R^1,\ i \in 2:n\ .$$

Действительно, используя представление (7.47), 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:

$$a_i = f_{i1} = -\frac{1}{u_n^1}\sum_{k=1}^{n} u_k^i d_{nk} = -\frac{1}{u_n^1} u_n^i \lambda_i\ D\ . \qquad (7.49)$$

Из (7.48) 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 $a = a_2, \dots, a_n$ не равны нулю. Тогда, выполняя последовательность преобразований подобия с помощью матриц перестановок [18], затрагивающих последние $n-1$ строк и столбцов матрицы F, получим матрицу

$$\Phi = \begin{bmatrix} \alpha & h^T & 0 \\ h & \operatorname{diag}\ \beta_i & 0 \\ 0 & 0 & \operatorname{diag}\ \gamma_i \end{bmatrix},$$

где компоненты вектора $h = h_1, \dots, h_s$ являются перестановкой ненулевых величин (7.49); $\operatorname{diag}\ \beta_i$ — порядка s; $\operatorname{diag}\ \gamma_i$ — порядка $n-s-1$; диагональ $\operatorname{diag}\ \beta_i, \gamma_i$ 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 diag γ_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

$$Z = \begin{bmatrix} \alpha & h^T \\ h & \text{diag } \beta_i \end{bmatrix}, \quad i \in 1:s \ . \tag{7.50}$$

Характеристический полином матрицы Z равен [72]

$$\varphi\ \lambda\ = \alpha - h \prod_{i=1}^{s} \beta_i - \lambda\ - \sum_{j=1}^{s} h_j^2 \prod_{i \neq j} \beta_i - \lambda\ . \tag{7.51}$$

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:

$$q_1 + q_2 + ... + q_t = s.$$

Тогда характеристическое уравнение

$$\varphi\ \lambda\ = 0 \tag{7.52}$$

будет иметь множитель

$$\prod_{i=1}^{t} \left[\lambda_{m_i}\ D\ - \lambda \right]^{q_i - 1}. \tag{7.53}$$

Следовательно, $\lambda_{m_i}\ D$ являются собственными числами матрицы Z, а значит, и матрицы F с кратностями $q_i - 1$. Поделив (7.52) на

$$\prod_{i=1}^{t} \left[\lambda_{m_i}\ D\ - \lambda \right]^{q_i}, \tag{7.54}$$

приходим к уравнению для оставшихся собственных чисел матрицы Z

$$\alpha - \lambda - \sum_{i=1}^{t} \eta_i^2 \left[\lambda_{m_i}\ D\ - \lambda \right]^{-1} = 0, \tag{7.55}$$

где η_i^2 есть сумма q_i величин h_j^2, связанных с β_i.

Согласно (7.49) 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

$$\lambda_{i_1}\ D\ , ..., \lambda_{i_{n-i-1}}\ D\ ;\ i_k \in 2:n\ ,\ k \in 1:n-t-1\ ;$$

остальные являются корнями уравнения

$$\alpha-\lambda-\sum_{i=1}^{t} c_{m_i}^2 \lambda_{m_i}^2\ D\left[\lambda_{m_i}\ D\ -\lambda\right]^{-1}=0, \tag{7.56}$$

где

$$c_{m_i}^2=\frac{1}{u_n^1\ ^2}\sum_{j=2}^{n}{}'\ u_n^j\ ^2\ . \tag{7.57}$$

Известно, что матрица F имеет нулевое собственное число, не вошедшее в группу γ_i и поэтому сохранившееся у матрицы Z (деление на множитель (7.54) убирает лишь кратные собственные числа, и хотя бы одно нулевое собственное число останется). Подставляя $\lambda=0$ в уравнение (7.56), получим

$$\alpha=\sum_{i=1}^{t} c_{m_i}^2 \lambda_{m_i}\ D\ ,$$

и уравнение (7.56) примет вид

$$\lambda+\sum_{i=1}^{t} c_{m_i}^2 \lambda_{m_i}^2\ D\left[\lambda_{m_i}\ D\ -\lambda\right]^{-1}=\sum_{i=1}^{t} c_{m_i}^2 \lambda_{m_i}\ D\ , \tag{7.58}$$

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 теоремы 7.3 достаточно показать, что число больших корней уравнения (7.58) 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 (7.46).

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 (7.58) согласно (7.27) 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 (7.58) для p больших (положительных) собственных чисел $\lambda_{m_i}\ D$:

$$\mu+\sum_{i=1}^{t-p} \frac{c_{\eta_i}^2 \lambda_{\eta_i}^2\ D}{\lambda_{\eta_i}\ D\ -\mu}=\sum_{i=1}^{t-p} c_{\eta_i}^2 \lambda_{\eta_i}\ D\ , \tag{7.59}$$

где $\left|\lambda_{\eta_i}\ D\right| \le \left|\lambda_{n-r+1}\ D\right|$. При этом $t-p$ ненулевых корней уравнения (7.59) согласно третьему утверждению леммы будет, соответственно, не меньше, чем $t-p$ ненулевых малых корней уравнения (7.58). Следовательно, левая часть неравенств (7.46) доказана. Однако в силу второго утверждения леммы эти $t-p$ корней не могут быть большими.

Действительно, согласно неравенствам (7.28)—(7.30) имеем для корней уравнения (7.59)

$$\left|\mu_j\right| \le \max_{i\in 1:t-p} \left|\lambda_{\eta_i}\ D\right| \left(1+\sum_{i=1}^{t-p} c_{\eta_i}^2\right) \le n \max_{i\in 1:t-p} \left|\lambda_{\eta_i}\ D\right|, \tag{7.60}$$

что доказывает правую часть неравенств (7.46).

Поясним последнее неравенство в (7.60). Согласно (7.57) получим

$$\sum_{i=1}^{t-p} c_{\eta_i} \le \left(u_n^1\right)^{-2} \sum_{i=2}^{n} \left(u_n^i\right)^2, \tag{7.61}$$

так как если $m_i \ne m_j$, то $c_{m_i}^2$ и $c_{m_j}^2$ не содержат одинаковых слагаемых. Следовательно,

$$1+\sum_{i=1}^{t-p} c_{\eta_i}^2 \le \left(u_n^1\right)^{-2} \sum_{i=1}^{n} \left(u_n^i\right)^2 = \left(u_n^1\right)^{-2} \le n, \tag{7.62}$$

т. к. для максимального по модулю компонента нормированного к единице вектора u^1, очевидно, выполняется неравенство

$$\left|u_n^1\right| \ge \frac{1}{\sqrt{n}}.$$

Из (7.62) следует (7.60).

Из рассмотрения уравнения (7.58) получили, что p корней λ_i будут заведомо большими, что эквивалентно (7.45):

$$\lambda_{n-r-1}\ G_k^D \ge \lambda_{n-r}\ D\ .$$

А так как имеем по условию, что

$$\lambda_{n-r}\ D \ge \sigma \left|\lambda_{n-r+1}\ D\right|,$$

то соотношение (7.43) доказано полностью. Неравенство (7.44) очевидным образом следует из соотношений, аналогичных (7.60), записанных для уравнения (7.58).

Для доказательства третьего утверждения теоремы заметим, что x^* удовлетворяет уравнению

$$\left\langle x,\ u^1\right\rangle = \lambda_1\ D^{-1} \left\langle b,\ u^1\right\rangle,$$

поэтому x^* имеет вид

$$x^* = \left[L_k \;\; y^* \;\; , \; y^* \right],$$

где y^* есть подвектор вектора x^* с исключенным k-м компонентом. Доказательство равенства $y^* = z^*$ проводится по схеме доказательства соотношения (7.15) в теореме 7.2. Теорема доказана.

7.1.4. 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 [57].

Общая идея ППИ была предложена Ю. В. Ракитским. Она звучала следующим образом: траектории спуска овражного функционала описываются жесткими дифференциальными уравнениями, и между компонентами их решений существуют асимптотические алгебраические связи типа равенств. Исключив с помощью этих алгебраических связей "лишние" аргументы минимизируемого функционала, мы получим функционал от меньшего числа аргументов и с меньшими степенями жесткости (овражности).

Приводимые далее математические конструкции, интерпретации и выводы принадлежат автору настоящей книги.

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 результатом 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

$$f\ x\ = 0,5\langle Dx,\ x\rangle - \langle b,\ x\rangle + \text{const}, \qquad (7.63)$$

где D — симметричная положительно определенная ($n \times n$)-матрица, b — n-мерный вектор.

Далее применение методов ньютоновского типа, по существу, сводится к определению элементов матрицы D и вектора b и вычислению x^* из решения системы линейных алгебраических уравнений

$$Dx = b. \tag{7.64}$$

Именно эта модель применяемых методов оптимизации будет далее использована.

Если функционал (7.63) овражный, то спектральное число обусловленности $k\ D \gg 1$ и уже малые деформации коэффициентов системы (7.64) могут приводить к сильным вариациям компонентов искомого вектора x^*. При $k\ D > 1/\delta$, где δ — относительная погрешность представления указанных коэффициентов ($\delta \cong \delta_J$), задачу решения системы уравнений (7.64) 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 (7.64) 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

$$\alpha_i = \frac{\beta_i}{\lambda_i\ D};\ x^* = \sum_{i=1}^{n} \frac{\beta_i}{\lambda_i (D)} u^i;\ \frac{\partial x^*}{\partial \lambda_i\ D} = -\lambda_i^{-2}\ D\ \beta_i u^i;$$

$$\left| \frac{\left\langle dx^*, u^i \right\rangle}{\left\langle x^*, u^i \right\rangle} \right| = \Delta_i,$$

где $\Delta_i = \left| \frac{d\lambda_i}{\lambda_i} \right|$ — относительная ошибка в представлении i-го собственного числа матрицы D.

Из последних двух равенств следует, что решение системы (7.64) 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

$$\|D\|_1 = \max_i \sum_{j=1}^{n} |d_{ij}|,$$

$$\|D\| = \max_i |\lambda_i \ D| \triangleq \Lambda_D$$

выполняется неравенство [74]:

$$\|D\|_1 \le \sqrt{n}\|D\| = \sqrt{n}\Lambda_D.$$

Так как, очевидно, $\Lambda_D \le \|D\|_1$, то

$$\|D\| = \Lambda_D \le \max_i \sum_{j=1}^{n} |d_{ij}| \le \sqrt{n}\Lambda_D. \tag{7.65}$$

Эти неравенства показывают, что:

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, т. к. 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 [72], если матрица Π добавляется к D, то все собственные числа матрицы D могут изменяться на величину, лежащую между наименьшим и наибольшим собственным числом Π.

Легко привести пример, когда минимальное собственное число матрицы D изменяется при этом весьма сильно. Рассмотрим матрицу

$$D = \begin{bmatrix} L & L-\varepsilon \\ L-\varepsilon & L \end{bmatrix}, \ |\varepsilon| \ll L, \ L > 0$$

и диагональную матрицу ошибок

$$\Pi = \begin{bmatrix} h & 0 \\ 0 & h \end{bmatrix}.$$

Матрица D имеет спектр:

$$\lambda_1 \ D \cong 2L - \varepsilon \cong 2L,$$

$$\lambda_2 \ D = \varepsilon.$$

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, т. е. $\lambda_1\ D+\Pi\ \cong\lambda_1\ D\ $, если $h\ll L$, 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

$$\max_i \sum_{j=1}^{n}\left|\alpha_{ij}\right|\le n\max_{i,j}\left|d_{ij}\right|.$$

Отсюда, учитывая (7.65), получим (см. теорему 3.5.1 из [44])

$$n^{-1}\Lambda_D\le\max_{i,j}\left|d_{ij}\right|\le\Lambda_D \tag{7.66}$$

или

$$\max_{i,j}\left|d_{ij}\right|\le\Lambda_D\le n\max_{i,j}\left|d_{ij}\right|. \tag{7.67}$$

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 (7.66) на δ, получим

$$\frac{\delta\Lambda_D}{n}\le\max_{i,j}\left|p_{ij}\right|\le\delta\Lambda_D.$$

Записывая неравенства (7.67) 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

$$\frac{\delta}{n}\frac{\Lambda_D}{\left|\lambda_i\ D\ \right|}\le\Delta_i\le n\delta\frac{\Lambda_D}{\left|\lambda_i\ D\ \right|}. \tag{7.68}$$

Согласно правому неравенству (7.68), 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 погрешно-

стей Δd_{ij}. Согласно левому неравенству, погрешность может быть особенно большой для собственных чисел с минимальными модулями в случае

$$\Lambda_D \gg \left|\lambda_i \ D\right|,$$

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 [57] 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^i вычисляются неизвестные коэффициенты линейной системы (7.64). Как уже указывалось, решать систему (7.64) при больших $k\ D$ бессмысленно, т. к. 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 (7.64) переходим к решению линейной системы

$$B\xi = d \tag{7.69}$$

относительно некоторого подвектора ξ вектора x. Система (7.69) 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 мат-

рицы 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, ибо 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 .

Пример. Пусть задан выпуклый квадратичный функционал вида

$$J\ \beta,\ x\ = \beta_0 + \beta_1 x_1^2 + \beta_2 x_1 x_2 + \beta_3 x_2^2 - \beta_4 x_1 - \beta_5 x_2, \tag{7.70}$$

где $\beta = \beta_0, \beta_1, ..., \beta_5$ — 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

$$J\ \hat{\beta}, \hat{x}^*\ = \min_x J\ \hat{\beta}, x\ . \tag{7.71}$$

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 показано 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:

$$\bar{\beta}_0 = 1111,\ \bar{\beta}_1 = 989,\ \bar{\beta}_2 = 984,\ \bar{\beta}_3 = 251,\ \bar{\beta}_4 = 1978,\ \bar{\beta}_5 = 984. \tag{7.72}$$

При этом легко проверить, что

$$x^* = 1,0\ ,\ J\ \bar{\beta},\ x^*\ = 122,0. \tag{7.73}$$

Используемые при получении оценок $\hat{\beta}_i$ экспериментальные данные имитируются при заданных значениях параметров (7.72). Шум измерений моделируется округлением результата измерений до двух цифр мантиссы в представлении числа в форме с плавающей точкой (запятой).

В табл. 7.1 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 7.1

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

$$\bar{\beta}_0 = 1100,\ \bar{\beta}_1 = 1035,\ \bar{\beta}_2 = 975,\ \bar{\beta}_3 = 225,\ \bar{\beta}_4 = 1990,\ \bar{\beta}_5 = 975. \qquad (7.74)$$

Сравнивая с (7.72), видим, что погрешность в определении $\hat{\beta}_i$ лежит во втором-третьем 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 (7.74) 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 (7.73). Причина неудачи заключается в овражном характере $J\ \ x\ $.

Если представить выражение (7.70) с параметрами (7.72) в виде

$$f\ \ x\ = 0{,}5\langle Dx,\ x\rangle - \langle b,\ x\rangle + c, \qquad (7.75)$$

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 (7.75) изображены на рис. 7.1. Овраг направлен параллельно вектору u^2. Кружками обозначены точки плана эксперимента, использованные при получении оценок (7.74).

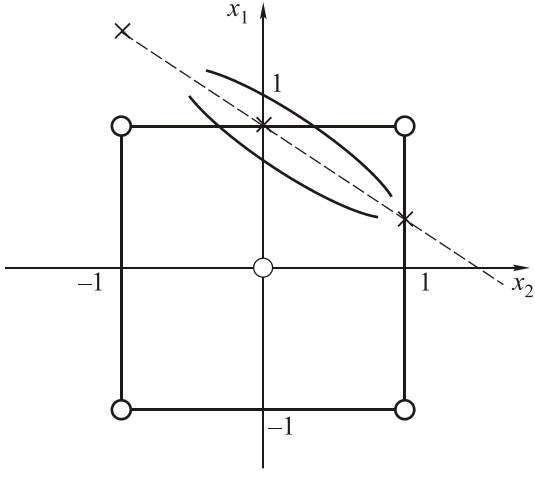


Рис. 7.1. План эксперимента

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 (7.74) не удалось получить приемлемое приближение к вектору x^*, однако построенное на их основе уравнение связи

$$x_1 = 0{,}968 - 0{,}469x_2 \cong 1 - 0{,}5x_2 \qquad (7.76)$$

достаточно точно локализует дно оврага. При получении (7.76) было использовано первое уравнение системы вида (7.7).

Согласно общей методике исключаем переменную x_1 и переходим к новому функционалу вида

$$J_1\ c,\ x_2\ = c_0 + c_1 x_2 + c_2 x_2^2. \tag{7.77}$$

В результате задача свелась к задаче одномерной минимизации по переменной x_2, а соответствующие значения x_1 вычисляются по уравнению (7.76).

На следующем этапе производим независимую от $\hat{\beta}_i$ оценку значений параметров c_i. Используется план эксперимента

$$x_2^1 = 0,\ x_2^2 = 1,\ x_2^3 = -1. \tag{7.78}$$

Для вычисления значений функционала (7.77) в точках плана (7.78) вычисляются соответствующие значения x_1^i по уравнению (7.76). Получается план для исходного функционала, представленный в табл. 7.2, точки которого расположены вдоль оврага (обозначены крестиками на рис. 7.1).

Таблица 7.2

I	x_1^i	x_2^i	$\tilde{J}_i$	$\overline{J}_i$
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. 7.2 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 (7.76)

$$x_1^* = 0,968 - 0,469 x_2^* = 0,968.$$

Таким образом, получен вектор

$$\hat{x}^* = \ 0,968; 0\ ,\ J\ \overline{\beta}, \hat{x}^*\ = 123,64,$$

близкий к истинному решению.

Погрешность в определении оптимальных значений компонентов x^*, как и у экспериментальных значений, лежит во втором знаке мантиссы. Задача решена.

7.1.5. 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

$$I\ y\ \triangleq J\ \bar{x}+y\ . \tag{7.79}$$

Пусть $f\ y$ — параболоид, аппроксимирующий функционал $I\ y$ в окрестности точки $y=0$:

$$f\ y\ =0{,}5\langle Ay,\ y\rangle-\langle d,\ y\rangle+c, \tag{7.80}$$

где

$$A=f''\ 0\ =J''\ \bar{x}\ =I''\ 0\ ;$$
$$d=-f'\ 0\ =-J'\ \bar{x}\ =-I'\ 0\ .$$

Дифференциальное уравнение спуска (7.2) для параболоида (7.80) и уравнение асимптотической связи между компонентами вектора *y* примут соответственно вид

$$\frac{dy}{dt}=-Ay+d,$$
$$\langle y,\ u^1\rangle=\lambda_1^{-1}\ A\ \langle d,\ u^1\rangle. \tag{7.81}$$

Коэффициенты уравнения (7.81) 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

$$A^k y=A^{k-1}d. \tag{7.82}$$

при надлежащем задании параметра *k*.

Укрупненная вычислительная схема метода иерархической оптимизации сводится к представленной далее последовательности шагов.

Алгоритм МИО.

Шаг 0. Принять:

$$y^0:=y;\ I_0\ y^0\ \triangleq I\ y\ ;\ n_0:=n;\ y^0=\ y_1^0,...,y_{n_0}^0\ ;\ l:=1.$$

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 (7.81) выразить k_l-й компонент вектора $y^{l-1} = \ y_1^{l-1}, \ldots, y_{n_{l-1}}^{l-1}$:

$$y_{k_l}^{l-1} = L^l\ \ y^l\ \triangleq -\ u_{k_l}^{l,\,1}\ ^{-1}\left[\ \lambda_1^l\ ^{-1}\left\langle g^{l-1},\, u^{l,\,1}\right\rangle - \sum_{i=1}^{n_{l-1}} y_i^{l-1} u_i^{l,\,1}\right],\ \ i \neq k_l, \qquad (7.83)$$

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

$$I_l\ \ y^l\ \triangleq I_{l-1}\left[L_{k_l}\ \ y^l\ \ ,\ y^l\right],$$

где $\left[L_{k_l}\ \ y^l\ \ ,\ y^l\right] \triangleq\ y_1^l, \ldots, y_{k_l-1}^l, L^l\ \ y^l\ \ , y_{k_l}^l, \ldots, y_{n_l}^l\ .$

Шаг 5. Положить $l := l+1$. 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

$$I_{n-r}\ \ y^{n-r}\ \ = I_0\ \ \tilde{y}\ = I\ \ \tilde{y}\ ;$$

$$\tilde{y} = \left[L_k\ \ y^1\ \ ,\ y^1\right],$$

$$y^1 = \left[L_{k_2}\ \ y^2\ \ ,\ y^2\right],$$

$$\ldots\ldots\ldots\ldots\ldots$$

$$y^{n-r+1} = \left[L_{k_{n-r}}\ \ y^{n-r}\ \ ,\ y^{n-r}\right].$$

(Согласно теореме 7.3 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; положить $\bar{x} := \bar{x} + y$ 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 МІО 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 МІО 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 МІО по сравнению с методами ОПС, в принципе также реализующими идею локальной декомпозиции задачи минимизации, заключается в следующем. В отличие от методов ОПС в алгоритме МІО 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 МІО 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.

7.2. Методы исключения переменных на основе спектрального разложения матрицы Гессе

7.2.1. Постановка задачи

Пусть решается: задача

$$J\ x\ \to \min_{x\in D},\ D\subset R^n=\ x\ \big|\ x=\ x_1,\ x_2,\ ...,\ x_n\ \ ,\ \ J:R^n\to R,\ \ J\in C^2\ \ R^n\ \ . \qquad (7.84)$$

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

$$x_i = b_i + \ a_i - b_i\ \ \sin^2 y_i\ \text{ или }\ x_i = y_i^2.$$

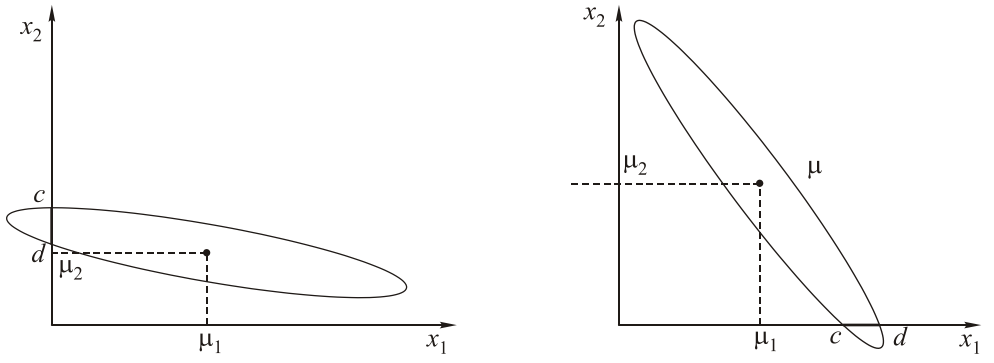
Пусть μ — минимизатор функционала $J\ x\ $. Основная задача формулируется следующим образом. Требуется построить такой вектор z, чтобы вектор $\mu' = \mu + z$ имел максимальное число компонентов, совпадающих с заданным вектором d (как правило, $d=0$), и чтобы при этом

$$\Delta J \triangleq J\ \ \mu'\ \ - J\ \ \mu\ \ \le \delta J, \qquad (7.85)$$

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. 7.2, *а*.



а *б*

Рис. 7.2. Исключение переменных: *а* — исключается переменная x_1; *б* — исключается переменная x_2

Компонент μ_1 больше μ_2, однако из рисунка видно, что исключаться должна не переменная x_2, а переменная x_1, т. к. 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. 7.2, *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 (7.84). В представленных на рис. 7.2 ситуациях этот метод приводит к выбору одной из точек отрезка c, d, т. е. к исключению, соответственно, переменных x_1, x_2 и необходимому пересчету оставшихся компонентов. Превышение допустимого уровня δJ при этом, очевидно, не происходит.

В принципе решение рассматриваемой проблемы может быть получено последовательным решением задач типа (7.84) с различными наборами $x_{i_1}, ..., x_{i_k}$ управляемых параметров. При этом должны варьироваться 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 (7.84) 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

$$J\ x \cong 0{,}5\langle Ax,\ x\rangle - \langle b,\ x\rangle + \text{const}, \tag{7.86}$$

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

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{n-r} \gg \lambda_{n-r+1} \geq \ldots \geq \lambda_n. \tag{7.87}$$

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 [72], что линейные оболочки 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 далее 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 $v = \left[u^{n-r+1}, \ldots, u^n\right]$ — матрица полного ранга размером $n \times r$; $c = c_1, \ldots, c_r$ — вектор неизвестных коэффициентов, удовлетворяющий соотношению $\|c^2\| \leq \delta$.

Значение δ вычисляется из условия (7.85).

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, т. к. 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 (7.86) имеем

$$\Delta J = \langle J' \ \mu \ , \ vc \rangle + 0{,}5 \langle Avc, \ vc \rangle = 0{,}5 \langle Avc, \ vc \rangle = 0{,}5 \sum_{i=n-r+1}^{n} c_{i-n+r}^2 \lambda_i , J' \ \mu \ = 0. \qquad (7.88)$$

Полагая $|\lambda_i| \le \alpha, \ i \in \ n-r+1 : n$, получим $\Delta J \le 0{,}5\alpha \left\| c^2 \right\| \le \delta J$, откуда

$$\|c\|^2 = \sum_{i=1}^{r} c_i^2 \le \delta, \quad \delta \triangleq \frac{2\delta J}{\alpha}. \qquad (7.89)$$

Введенная величина α формализует понятие малого собственного значения и в разных случаях может выбираться из условий, определяемых типом решаемой задачи.

Согласно (7.68) для погрешностей задания собственных значений при ограниченной разрядности применяемых компьютеров выполняются неравенства $|d\lambda_i| \le n\lambda_1 \varepsilon_{\text{м}}$, где $\varepsilon_{\text{м}}$ — 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 $\alpha \le n\lambda_1 \varepsilon_{\text{м}}$, и если в результате решения задачи на собственные значения для некоторых из вычисленных значений выполняются неравенства

$$|\lambda_i| \le n\lambda_1 \varepsilon_{\text{м}}, \qquad (7.90)$$

то соответствующие собственные значения целесообразно классифицировать как "малые", а отвечающие им собственные векторы включать в группу $u^{n-r+1}, \ ..., \ u^n$.

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 $\delta = \dfrac{2\delta J}{\alpha}$, 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 пошаговое уменьшение величины α от некоторого большого значения порядка $\dfrac{\lambda_1}{2}$ до значения $n\lambda_1 \varepsilon_{\text{м}}$ 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

$$vc = f, \quad f \triangleq d - \mu \tag{7.91}$$

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.

7.2.2. Алгоритм исключения

Для приближенного определения оптимальной в смысле величины нормы $\|c\|$ подсистемы (7.91) предлагается воспользоваться геометрически очевидной идеей проектирования точки $f \in R^n$ вдоль линейной оболочки M_2 на ближайшие подпространства R^l, $l = n-1, ..., n-r$. 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 приведенной далее последовательности действий.

Каждое из уравнений системы (7.91) имеет вид

$$\langle w, c \rangle = t, \tag{7.92}$$

где *w* — вектор коэффициентов, $c = \|c_1, ..., c_r\|$.

На первом шаге выберем уравнение, для которого норма вектора *c*, определяющего нормальное решение уравнения (7.92) с компонентами $c_i = \dfrac{tw_i}{\|w\|^2}$, $i \in 1:r$, имеет минимальное значение. Разрешим данное уравнение относительно переменной c_i с наибольшим по модулю коэффициентом, полагая, без ограничения общности, что выбрано первое уравнение и первая переменная c_1. Подставляя выражение для c_1 в остальные уравнения, приходим к системе

$$\begin{aligned} &c_1 = L_1 \|c_2, ..., c_r\|, \\ &v'_{21} f_1 + v'_{22} c_2 + ... + v'_{2r} c_r = f_2; \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &v'_{n1} f_1 + v'_{n2} c_2 + ... + v'_{nr} c_r = f_n, \end{aligned} \tag{7.93}$$

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 (7.93), начиная со второго, также имеет вид (7.92) $\langle \overline{w}, \overline{c} \rangle = \overline{t}$, где $\overline{c} = c_2, ..., c_r$. На втором шаге выбирается уравнение, для которого норма вектора c с компонентами $c_1 = L_1 \; c_2, ..., c_r$, $c_j = \dfrac{\overline{t}\,\overline{w}_j}{\|\overline{w}\|^2}$ имеет минимальное значение. Аналогично предыдущему получаем линейную связь $c_2 = L_2 \; c_3, ..., c_r$, 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 вышеизложенном процессе уравнений.

Для выполнимости сформулированного алгоритма существенно, что если уже исключены переменные $c_1, ..., c_k$, $k < r$, то всегда найдется уравнение вида $\beta_{k+1}c_{k+1} + ... + \beta_r c_r = t_k$, где не все $\beta_i = 0$, и, следовательно, процесс исключения может быть продолжен. Справедливость данного замечания следует из теории гауссовского исключения для прямоугольных матриц [65]. А именно если прямоугольная матрица v размером $n \times r$ 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 однократного решения задачи (7.84) 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 $J\ x$ вида (7.84) 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 (7.84) 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'' μ равны: $\lambda_1 = 143{,}12$; $\lambda_2 = 87{,}633$; $\lambda_3 = 4{,}141 \cdot 10^{-7}$; $\lambda_4 = -3{,}713 \cdot 10^{-6}$.

Согласно (7.90), 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:

$$v^1 = 10^{-1}\ \ 1{,}400280;\ 5{,}601120;\ 7{,}001400;\ 4{,}200840\ \ ^T,$$
$$v^2 = 10^{-1}\ \ -8{,}980265;\ 1{,}796053;\ -1{,}796053;\ 3{,}592106\ \ ^T.$$

Полагая $r = 2$, решали задачу исключения двух переменных — по числу малых собственных значений. Система двух уравнений (7.91) с двумя неизвестными c_1, c_2 при $d_i = 0$ имеет вид

$$c_1 v^1 + c_2 v^2 = 10\ \ 1;\ 1;\ 1;\ 1\ \ ^T. \tag{7.94}$$

На первом шаге согласно базовому алгоритму вычислялись нормы $\dfrac{\|tw\|}{\|w\|^2} = \dfrac{|t|}{\|w\|}$ для всех 4-х уравнений (7.94). 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:

$$c_2 = 1{,}559285 \cdot 10^{-1} c_1 - 11{,}13552. \tag{7.95}$$

Подставляя (7.95) в уравнения 2—4 системы (7.94), получим на втором шаге

$$\begin{aligned} 5{,}881175 c_1 &= 120{,}0 \\ 6{,}721344 c_1 &= 80{,}0 \\ 4{,}760951 c_1 &= 140{,}0. \end{aligned} \tag{7.96}$$

С учетом (7.95) имеем для трех случаев (7.96)

$$\begin{aligned} c_1 &= 20{,}40408, \quad c_2 = -7{,}953943; \\ c_1 &= 11{,}90238, \quad c_2 = -9{,}279600; \\ c_1 &= 29{,}40588, \quad c_2 = -6{,}550306. \end{aligned} \tag{7.97}$$

Наименьшее значение нормы $\|c\|$ достигается во втором уравнении (7.97), что соответствует третьему уравнению исходной системы (7.94). В результате установле-

но, что исключению подлежат компоненты μ_1, μ_3 вектора μ. Новый вектор μ' находится из равенства

$$\mu' = \mu + c_1 v^1 + c_2 v^2 = -10\ \ 1;\ 1;\ 1;\ 1\ ^T + 11{,}90238 \cdot 10^{-1}\ \ 1{,}400280;$$

$$5{,}601120;\ 7{,}001400;\ 4{,}200840\ ^T - 9{,}279600 \cdot 10^{-1}\ \ -8{,}980265;\ 1{,}796053;$$

$$-1{,}79053;\ 3{,}592106\ ^T \cong\ 0{,}0;\ -5{,}00;\ 0{,}0;\ -8{,}33\ ^T.$$

При этом согласно (7.89) имеем оценку

$$\Delta J = J\ \ \mu'\ \ - J\ \ \mu\ \ \le \frac{1}{2}\|c\|^2\,\alpha \le 2 \cdot 10^{-3},$$

что удовлетворяет условию (7.85) допустимости отклонения. Задача решена. В данном случае простое отбрасывание любой из переменных μ_i приводило к многократному превышению заданного значения δJ.

Пример 2. На рис. 7.3 представлена ARC-цепь с пятиузловой RC-схемой, имеющей избыточную структуру [38]. 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

$$J\ \ x\ = f_1^2 + f_2^2 +\ \ f_3 - 1\ ^2 +\ \ f_4 - 1\ ^2 +\ 100 f_5 - 1\ ^2 +\ \ f_6 - 1\ ^2, \qquad (7.98)$$

где $x = x_1, ..., x_{15}$; $y = G_1, ..., G_7, c_1, ..., c_7, k$; $y_i = x_i^2$, $i \in 1{:}15$; $f_i\ \ x$ — известные функции, определяющие коэффициенты передаточной функции

$$T\ \ p\ = \frac{f_1 + f_2 p + f_3 p^2}{f_4 + f_5 p + f_6 p^2}.$$

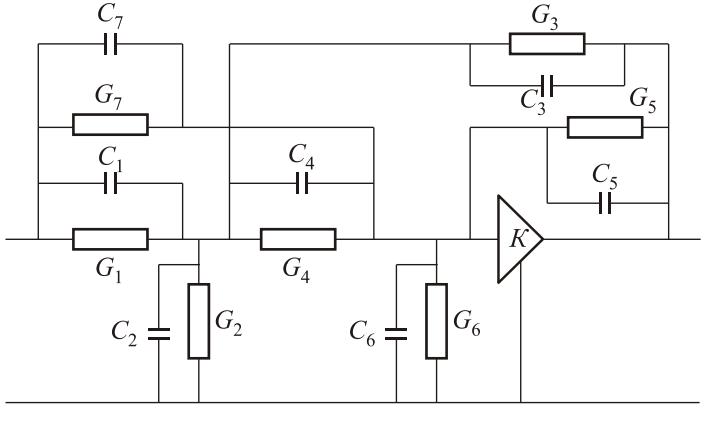


Рис. 7.3. Исходная схема с избыточной структурой

Минимизация (7.98) методом ОПС привела к следующей точке μ:

$$\mu_1 = 0,7240029 \cdot 10^{-5}; \quad \mu_2 = 0,2704296 \cdot 10^{-2}; \quad \mu_3 = 0,2324280 \cdot 10;$$
$$\mu_4 = 0,2430573 \cdot 10^{-7}; \quad \mu_5 = 0,7020994 \cdot 10^{-5}; \quad \mu_6 = 0,4297312;$$
$$\mu_7 = 0,1366784 \cdot 10^{-4}; \quad \mu_8 = 0,8707104; \quad \mu_9 = 0,4076001;$$
$$\mu_{10} = 0,4429756 \cdot 10^{-4}; \quad \mu_{11} = 0,7618611; \quad \mu_{12} = 0,1217431 \cdot 10^{-4};$$
$$\mu_{13} = 0,1281024 \cdot 10^{-1}; \quad \mu_{14} = 0,3692824 \cdot 10^{-6}; \quad \mu_{14} = 0,1507476 \cdot 10$$

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:

$$\lambda_1 = 0,4238 \cdot 10^{7}; \quad \lambda_2 = 0,5342 \cdot 10^{2}; \quad \lambda_3 = 0,230 \cdot 10^{2};$$
$$\lambda_4 = 0,570 \cdot 10; \quad \lambda_5 = 0,167; \quad \lambda_6 = -0,795 \cdot 10^{-4};$$
$$\lambda_7 = 0,192 \cdot 10^{-5}; \quad \lambda_8 = 0,238 \cdot 10^{-6}; \quad \lambda_9 = 0,220 \cdot 10^{-7};$$
$$\lambda_{10} = 0,452 \cdot 10^{-9}; \quad \lambda_{11} = -0,189 \cdot 10^{-11}; \quad \lambda_{12} = 0,585 \cdot 10^{-12};$$
$$\lambda_{13} = 0,616 \cdot 10^{-13}; \quad \lambda_{14} = \lambda_{15} = 0.$$

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 (7.98) в 5-мерном пространстве были получены следующие окончательные результаты (рис. 7.4): $G_3 = 25,301$; $G_6 = 0,039525$; $C_1 = 0,63571$; $C_4 = 1,5715$; $k = 1,0019$;

$$T\ p\ = \frac{1,00096 p^2}{1,000003 + 0,9999997 \cdot 10^{-2} p + 0,9990286 p^2}.$$

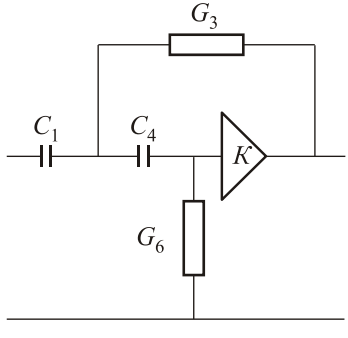


Рис. 7.4. Схема, полученная после упрощения

Последующее упрощение цепи на основе включения λ_5 в блок малых собственных значений оказалось невозможным из-за пропадания необходимых составляющих в выражении для $T \quad p$.

7.2.3. Удаление переменных в задаче наименьших квадратов

Под линейной задачей наименьших квадратов (НК) будем понимать следующую задачу [40].

Пусть заданы действительная $(m \times n)$-матрица A ранга $k \le \min\ m, n$ и действительный m-мерный вектор b. Требуется найти действительный n-мерный 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 $\tilde{A}$ — матрица размером $m \times\ n-1$, 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 2*.

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 задачи НК относительно k указанных переменных x_i, $i \in I_k$, не превосходит значение нормы, получаемой при выборе любого другого набора из k переменных. Сформулированная задача полностью эквивалентна задаче, рассматриваемой в *разд. 7.2.2* в предположении

$$J\ x\ \triangleq 0{,}5\|Ax - b\|^2. \tag{7.99}$$

Так же как и в случае задачи (7.91), 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 выше 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 $\tilde{I}_k \cup j \quad j \notin \tilde{I}_k, j \in 1:n$, для которого реализуется наименьшая невязка $\tilde{\rho}_{k+1}$.

Проведем сравнительный анализ алгоритма пошаговой регрессии и алгоритма из *разд. 7.2.2* на примере решения конкретной задачи НК.

Пример. Рассмотрим задачу НК [40] $Ax \cong b$, (15×5)-матрица A и вектор b которой приведены в табл. 7.3. 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

$$A = USV^T,$$

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 $i \in 1:n \quad s_{ii} \geq 0$, для $i \in 1:k \quad s_{ii} > 0$. Диагональные элементы s называются *сингулярными числами A*.

Таблица 7.3. *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 матрица $A^T A$ размером $n \times n$ допускает спектральное разложение

$$A^T A = VDV^T, \tag{7.100}$$

где $D = d_{ij}$ — диагональная $(n \times n)$-матрица, состоящая из собственных чисел матрицы $A^T A$, причем $d_{ii} = s_{ii}^2$, $i \in 1:n$; столбцы матрицы V суть ортонормированные собственные векторы матрицы $A^T A$.

Далее представлены результаты сингулярного разложения матрицы A из табл. 7.3:

$$10^4 V = \begin{bmatrix} 3742 & -7526 & 3382 & -1981 & -3741 \\ 5196 & -636 & 2301 & 6349 & 5195 \\ 4123 & 6510 & 4741 & -1067 & -4123 \\ 4796 & 689 & -2493 & -6877 & 4797 \\ 4359 & 302 & -7388 & 2707 & -4359 \end{bmatrix},$$

$$s_{11} = 1,00000;\ s_{22} = 0,1000;\ s_{33} = 0,0100;\ s_{44} = 0,9997 \cdot 10^{-5};\ s_{55} = 0,9904 \cdot 10^{-7}.$$

Собственные числа матрицы $J''\ x\ = A^T A$, где $J\ x$ определен равенством (7.99), равны

$$\lambda_i\ \ A^T A\ = s_{ii}^2,\ \ i \in\ 1:5\ ,$$

и, следовательно, cond $A^T A\ \cong 10^{14}$.

Решим задачу удаления трех переменных в сформулированной задаче НК. Точное решение задачи, полученное простым перебором всех возможных комбинаций, согласно табл. 7.4, 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 *разд. 7.2.2* и алгоритма пошаговой регрессии.

Таблица 7.4

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. 7.4 видно, что если обозначить через $\omega^{(i,\ j,\ ...)}$ решение, опирающееся на столбцы i, j, ..., то сначала будет найден вектор $\omega^{(3)}$, которому соответствует наименьшее значение нормы невязки, равное 0,07. Далее будет выбрано наилучшее решение из множества

$$\omega^{(1,\ 3)},\ \omega^{(2,\ 3)},\ \omega^{(3,\ 4)},\ \omega^{(3,\ 5)}. \tag{7.101}$$

В результате получим вектор $\omega^{(1,\ 3)}$ как наилучший среди векторов (7.101) 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 *разд. 7.2.2*, позволяет получить оптимальное решение этой задачи на основе информации о спектральном разложении (7.100), получаемом, например, из сингулярного анализа матрицы A.

Выберем собственные векторы, отвечающие трем наименьшим собственным числам $\lambda_i\ \ A^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 (7.91) примет вид

$$c_1 V^1 + c_2 V^2 + c_3 V^3 = f, \tag{7.102}$$

где $f =\ 2,4857328;\ 0,52913252;\ 0,18414114;\ -1,6156794;\ -3,4547871$, $f = -\mu$, а μ есть построенное в [40] решение задачи НК (опирающееся на все 5 столбцов матрицы A).

Согласно алгоритму, приведенному в *разд. 7.2.2*, 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 (102), получаем новую систему уравнений относительно двух переменных:

$$\begin{aligned}
&(1): \quad -0,1219853c_2 - 7,998515 \cdot 10^{-2} c_3 = 2,423456;\\
&(2): \quad\ \ 0,686858c_2 + 0,7196059c_3 = 0,4867616;\\
&(4): \quad -0,7438069c_2 + 0,2628968c_3 = -1,569773;\\
&(5): \quad\ \ 0,1044271c_2 - 1,078395c_3 = -3,318743.
\end{aligned} \tag{7.103}$$

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 (7.84) в (7.103));

$$(2):\ c_2 = 0,3378435;\quad c_3 = 0,3540399;\quad c_1 = 0,5680655;\qquad \|\cdot\|_2 = 0,7497872;$$

$$(4):\ c_2 = 1,876087;\qquad c_3 = -0,6630985;\ c_1 = 2,970736\cdot 10^{-2};\ \|\cdot\|_4 = 1,990046;$$

$$(5):\ c_2 = -0,2952416;\ c_3 = 3,048893;\qquad c_1 = 2,769158;\qquad \|\cdot\|_5 = 4,129304.$$

Таким образом, выбирается уравнение 2 системы (7.103):

$$0,6866858c_2 + 0,7196059c_3 = 0,4867616.$$

Отсюда имеем вторую линейную связь:

$$c_3 = -0,9542525c_2 + 0,676428 \triangleq L_2\ \ c_2\ \ .$$

Исключая c_3 из уравнений 1, 4 и 5 системы (7.103), получаем

$$(1):\ c_2 = -54,26191;\ c_3 = L_2\ \ c_2\ \ = 52,45599;\ c_1 = L_1\ \ c_2,\ c_3\ \ = 33,59029;\ \|c\| = 82,60927;$$

$$(4):\ c_2 = 1,756956;\quad c_3 = -1,000151;\qquad c_1 = -0,2902209;\qquad \|c\| = 2,042406;$$

$$(5):\ c_2 = -2,284352;\ c_3 = 2,856276;\qquad c_1 = 2,153983;\qquad \|c\| = 4,244551.$$

Наименьшее значение $\|c\|$ соответствует уравнению 4; следовательно, остаются 1-е и 5-е уравнения исходной системы (7.102), 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.

7.3. Основные результаты и выводы

1. 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.
2. 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.

3. 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 МІО. Указаны области рационального применения алгоритма МІО.

4. 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.

5. 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.

6. 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.

7. 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 8

Примеры решения задач

Приводимые далее численные примеры имеют целью проиллюстрировать на небольших и легко обозримых задачах методику применения изученных в книге методов и алгоритмов. 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 *разд. 8.4.2* 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 и построения гладких аппроксимаций исходных негладких функционалов;
- ❒ 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 представленным 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 оптимизации.

8.1. Реализация оптимальной весовой функции линейной стационарной системы

На этапе синтеза статистически оптимальной системы *(см. разд. 2.3.6)* 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 $x = \ x_1, \ x_2, \ ..., \ x_n$ 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 [60]. Поэтому нет необходимости рассматривать бесконечный интервал времени, и достаточно ограничиться значениями $t \le T$, где T определяется из условия:

$$\forall t > T : \left|\omega \ t \ \right| \le \alpha \max_{t \in \ 0, \ T} \left|\omega \ t \ \right|, \ \alpha \in \ 0, \ 05, \ 0{,}1 \ .$$

Форма представления функции $\omega \ x, t$, гарантирующая выполнение условий физической реализуемости, имеет вид [37]

$$\omega \ x, \ t \ = \sum_{i=1}^{r} \exp \ \sigma_i t \quad A_{i1} \cos \omega_i t + A_{i2} \sin \omega_i t \ ,$$

$$x = \ A_{11}, \ ..., \ A_{r1}, \ A_{12}, \ ..., \ A_{r2}, \ \sigma_1, \ ..., \ \sigma_r, \ \omega_1, \ ..., \ \omega_r \ ; \ \ \sigma_i \le 0, \ \ i \in \ 1 : r \ .$$

Наиболее часто в качестве критериев близости $\omega_* \ t$ и $\omega \ x, t$ используются функционалы, построенные с помощью принципов минимакса, наименьших квадратов и обобщающего эти подходы метода среднестепенной аппроксимации *(см. разд. 3.10.1)*.

Далее приводится пример численного решения сформулированной задачи [81].

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. 8.1. Начальный вектор равен $x^0 = A_{11}, A_{12}, \sigma_1, \omega_1 = -1{,}005;\ 0{,}06105;\ -0{,}296;\ 0{,}9355$.

На рис. 8.2 показан график зависимости

$$\Delta_0\ t \triangleq \omega_*\ t - \omega\ x^0, t\ ;\ \max_t \left|\Delta_0\ t\right| = 0{,}6866 \cdot 10^{-2}.$$

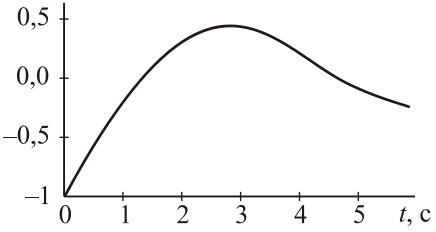


Рис. 8.1. Заданная весовая функция

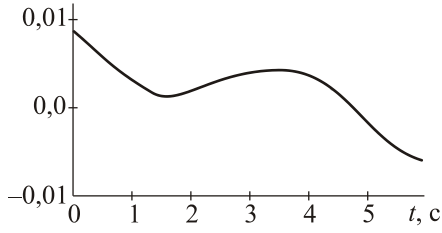


Рис. 8.2. График зависимости $\omega_*\ t - \omega\ x^0, t$

Критерий близости ω_* и ω построен на основе метода наименьших квадратов:

$$J\ x = \sum_{i=1}^{13} \left[\omega_*\ t_i - \omega\ x, t_i\ \right]^2, \tag{8.1}$$

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 накладывались никакие ограничения. 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. 8.3 приведен график $\Delta_*\ t = \omega_*\ t - \omega\ x^*, t$. Максимальная погрешность $\max\limits_t \left|\Delta_*\ t\right| \cong 0,5 \cdot 10^{-3}$ уменьшилась приблизительно в 14 раз.

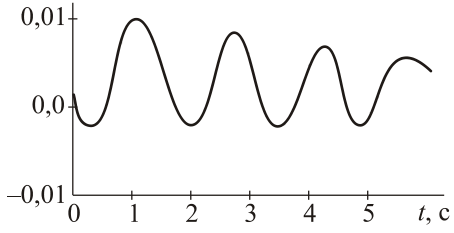


Рис. 8.3. График зависимости $\omega_*\ t - \omega\ x^*, t$

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 величины максимальной погрешности вместо МНК-критерия (8.1) была использована среднестепенная аппроксимация минимаксного критерия

$$J_1\ x = \sum_{i=1}^{13} \left[\omega_*\ t_i - \omega\ x, t_i\right] \cdot 10^{3\ \ 8}. \qquad (8.2)$$

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 (8.2) методом SPAC1 привела к результату $x^* = -0,99796;\ 0,061193;\ -0,28950;\ 0,92546$. При этом максимальная погрешность оказалась равна $0,38 \cdot 10^{-3}$, что хорошо согласуется с результатами работы [37], где минимизировался непосредственно минимаксный критерий методом возможных направлений. В данном случае за счет перехода от критерия (8.1) к (8.2) удалось уменьшить величину погрешности примерно в 18 раз по сравнению с исходной.

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 (т. е. 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 вышеприведенных результатов приблизительно в три раза. Для "больших" задач указанный выигрыш может оказаться принципиальным.

8.2. 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. разд. 3.8.1)*.

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 [42] и заключается в построении схемы с монотонно нарастающей по заданному закону амплитудно-частотной характеристикой.

В качестве примера рассмотрим задачу построения дробно-рационального выражения вида [41]

$$W\ \ p_1,\ f\ \ = \frac{1}{f_1 + f_2 p + f_3 p^2 + f_4 p^3\ \cdot\ f_5 + f_6 p + f_7 p^2 + f_8 p^3},$$

обладающего следующими свойствами. Модуль коэффициента передачи $\left|W\ \ j\omega,\ f\ \right|$ в полосе пропускания $\frac{\omega}{2\pi} \in 0,\ 600$ (Гц) удовлетворяет неравенству $0,7 \le \left|W\ \ j\omega,\ f\ \right| \le 1$.

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-филътру (фильтру нижних частот) и геометрически представлены на рис. 8.4.

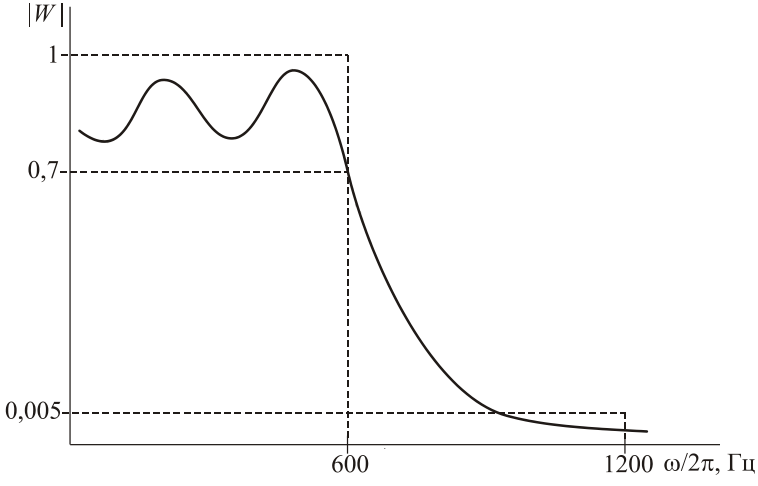


Рис. 8.4. Характеристика НЧ-фильтра

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

$$\sqrt{f_1}\ \ f_3^2 + f_2 f_4\ \ < 10\sqrt{f_3}\Delta_1;\ \ \sqrt{f_5}\ \ f_7^2 + f_6 f_8\ \ < 10\sqrt{f_7}\Delta_2. \qquad (8.3)$$

Представления (8.3) обычно являются более предпочтительными, т. к. не содержат операций деления, что существенно при возможных малых значениях Δ_i в случае выхода в процессе вычислений на границу области устойчивости. Ясно, что т. к. $f_i \ge 0$, то при $\Delta_1 \le 0$ либо при $\Delta_2 \le 0$ неравенства (8.3) не будут выполнены. Поэтому справедливость (8.3) гарантирует и выполнение требований по устойчивости.

Имеем $F\ \ f, \omega\ \ \triangleq |W\ \ j\omega,\ f\ \ | = \frac{1}{\sqrt{c_1 c_2}}$, где

$$c_1 = \ f_1 - f_3\omega^2\ \ ^2 + \omega^2\ \ f_2 - f_4\omega^2\ \ ^2;$$

$$c_2 = f_5 - f_7\omega^2{}^2 + \omega^2 f_6 - f_8\omega^2{}^2; \quad f = f_1, ..., f_8 . \tag{8.4}$$

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 положили значение ω_0 равным 10^4. Далее перешли к нормализованной частоте $\Omega = \dfrac{\omega}{\omega_0}$. Подставляя $\omega = \omega_0\Omega$ в (8.4), 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:

$$\begin{aligned} c_1 &= x_1^2 - \Omega^2 x_3^2{}^2 + \Omega x_2^2 - \Omega^3 x_4^2{}^2; \\ c_2 &= x_5^2 - \Omega^2 x_7^2{}^2 + \Omega x_6^2 - \Omega^3 x_8^2{}^2; \end{aligned} \tag{8.5}$$

где $x = x_1, x_2, ..., x_8$ — вектор нормализованных переменных с компонентами

$$\begin{aligned} &x_1 = y_1; \quad x_2 = \sqrt{\omega_0}\, y_2; \quad x_3 = \omega_0 y_3; \quad x_4 = \omega_0^{3/2} y_4; \\ &x_5 = y_5; \quad x_6 = \sqrt{\omega_0}\, y_6; \quad x_7 = \omega_0 y_7; \quad x_8 = \omega_0^{3/2} y_8. \end{aligned} \tag{8.6}$$

Заданные требования к амплитудно-частотной характеристике фильтра могут быть переформулированы в новых обозначениях. Модуль коэффициента передачи $\Phi\ x, \Omega \triangleq \dfrac{1}{\sqrt{c_1 c_2}}$, где c_1, c_2 рассчитываются по формулам (8.5) в полосе пропускания $\Omega \in \Omega_1, \Omega_2$, $\Omega_1 = 0$, $\Omega_2 = \left(\dfrac{2\pi}{\omega}\right)\cdot 600 = 0{,}377$ 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 (8.3).

Задача оптимального параметрического синтеза фильтра теперь может быть сформулирована следующим образом. Определить такой вектор x, чтобы функционал

$$J\ x = \sum_{i=0}^{35}\left[\Phi\ x, 10^{-2} i - 0{,}85\right]^2 + 500\left[\Phi\ x, 0{,}377 - 0{,}707\right]^2 + \\ + 10^3\left[\Phi\ x, 0{,}754\right]^2 \tag{8.7}$$

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

$$\beta = \frac{0{,}15^2}{0{,}007^2} \cong 500; \quad \gamma = \frac{0{,}15^2}{0{,}005^2} \cong 10^3.$$

Минимизация $J\ x$ осуществлялась алгоритмом GZ1 с начальным шагом дискретности $s = 0{,}1$ из произвольно выбранной начальной точки, удовлетворяющей неравенствам (8.3): $f_1^0 = 1$; $f_2^0 = 10^{-2}$; $f_3^0 = 10^{-4}$; $f_4^0 = 10^{-8}$; $f_5^0 = 1$; $f_6^0 = 10^{-2}$; $f_7^0 = 10^{-4}$; $f_8^0 = 10^{-8}$, или, на языке переменных x_i:

$$x_1^0 = x_2^0 = x_3^0 = x_5^0 = x_6^0 = x_7^0 = 1; \ x_4^0 = x_8^0 = 0{,}1. \tag{8.8}$$

В точках x, для которых не выполнялось хотя бы одно из соотношений (8.3), 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 (8.8), представлен на рис. 8.5.

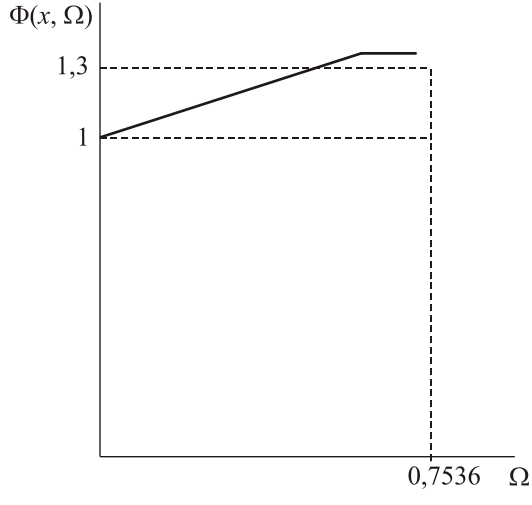


Рис. 8.5. АЧХ в начальной точке

После практической остановки метода была получена точка

$$x_1 = 0{,}84394; \quad x_2 = 2{,}2411; \quad x_3 = 4{,}0624; \quad x_4 = -1{,}5526;$$
$$x_5 = 1{,}2995; \quad x_6 = 1{,}2618; \quad x_7 = 3{,}5938; \quad x_8 = -0{,}73217, \tag{8.9}$$

которой соответствуют значения $J = 0,3855;\ Q_1 = 0,72;\ Q_2 = 3,084$ и коэффициенты передаточной функции

$$f_1 = 0,71223;\quad f_2 = 0,50227 \cdot 10^{-3};\quad f_3 = 0,16503 \cdot 10^{-6};\quad f_4 = 0,24105 \cdot 10^{-11};$$
$$f_5 = 1,6887;\quad f_6 = 0,15921 \cdot 10^{-3};\quad f_7 = 0,12915 \cdot 10^{-6};\quad f_8 = 0,53607 \cdot 10^{-12}.$$

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. 8.6.

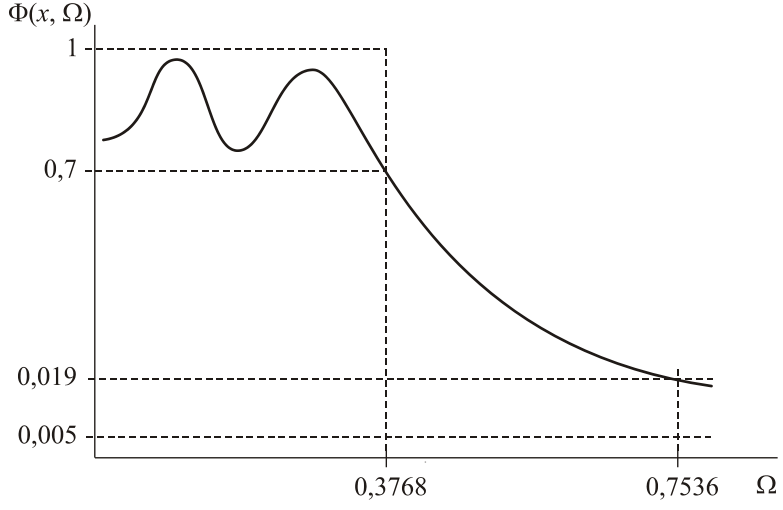


Рис. 8.6. АЧХ фильтра, полученная алгоритмом GZ1

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 *разд. 3.10.1*.

Для устранения отмеченного недостатка функционал (8.7) был заменен на минимаксный с применением его среднестепенной аппроксимации. А именно положили

$$J\ x = \sum_{i=0}^{35}\left[\Phi\ x;\ 10^{-2} i\ -0,85\right]^8 +\ 20\left[\Phi\ x;\ 0,377\ -0,707\right]^{\ 8} +$$
$$+\left[30\Phi\ \ 0;\ \ 0,754\ \right]^8. \qquad (8.10)$$

За начальную точку были взяты значения (8.9). Результаты, полученные методом GZ1 для функционала (8.10), представлены на рис. 8.7. Сходимость опять резко замедлилась до достижения приемлемой точки. При этом в точке $\Omega = 0,754$ имели $\Phi \cong 0,009$, что все еще превышает допустимое значение, равное 0,005.

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 (8.9) для минимизации критерия (8.10) применялся алгоритм SPAC5 с начальным шагом дискретности $s = 0,1$. График полученной зависимости $\Phi\ x,\ \Omega$ представлен на рис. 8.8. Все требования к амплитудно-частотной характеристике оказываются выполненными; в том числе имеем

$$\Phi\ x,\ 0,754\ \cong 0,004 < 0,005.$$

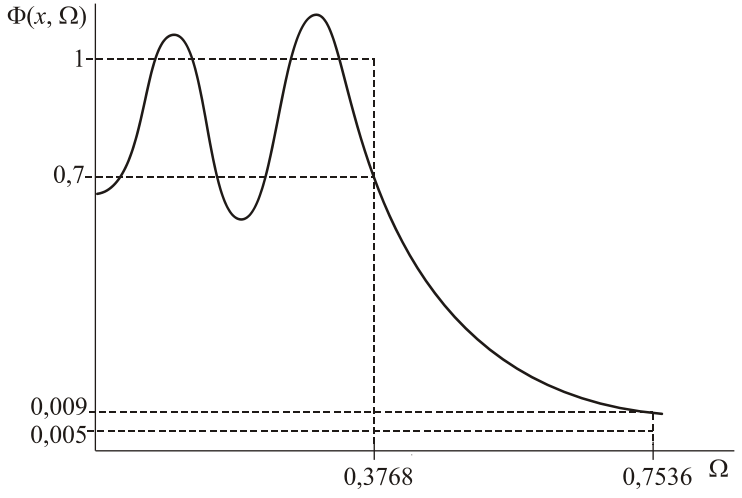


Рис. 8.7. АЧХ, построенная алгоритмом GZ1 при использовании минимаксного критерия

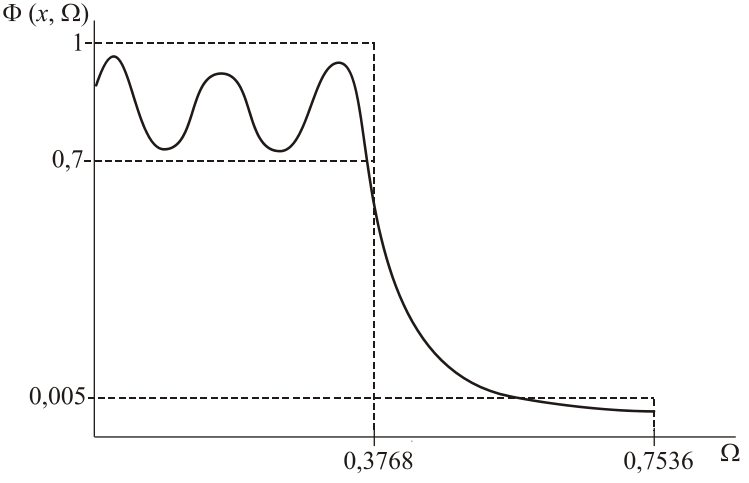


Рис. 8.8. АЧХ, построенная методом SPAC5 для минимаксного критерия

Соответствующие полученным результатам добротности и коэффициенты передаточной функции равны: $Q_1 = 1{,}549;\ Q_2 = 6{,}155;$

$$f_1 = 0{,}65346;\quad f_2 = 0{,}81629 \cdot 10^{-3};\quad f_3 = 0{,}29287 \cdot 10^{-6};\quad f_4 = 0{,}10853 \cdot 10^{-9};$$
$$f_5 = 1{,}6996;\quad f_6 = 0{,}16451 \cdot 10^{-3};\quad f_7 = 0{,}13085 \cdot 10^{-6};\quad f_8 = 0{,}64040 \cdot 10^{-11}.$$

Для решения задачи минимизации функционала (8.10) 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.

8.3. 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. 8.9) [7], [49]. Последние выбраны соответствующими наихудшему сочетанию внешних условий.

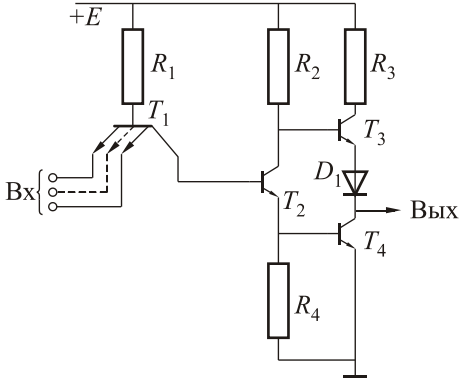


Рис. 8.9. ТТЛ-схема И–НЕ / ИЛИ–НЕ

Зададим условия работоспособности ТТЛ-схемы в виде системы неравенств

$$N \ge 20;\quad \Delta u \ge 0{,}8;\quad P \le 20;\quad t_з \le 15;\quad s \ge 1{,}5, \tag{8.11}$$

где N — коэффициент разветвления по выходу; Δu — допустимый уровень помехи в логическом состоянии "нуль"; P — средняя потребляемая мощность; $t_з$ — задержка распространения сигнала; s — степень насыщения транзистора T_2 . (Здесь

и далее применяются следующие единицы измерения: вольты, милливатты, наносекунды, килоомы.)

В качестве математической модели использованы известные зависимости:

$$N = 2{,}1 \cdot \left[2{,}45 + R_1 \left(\frac{3{,}95}{R_2 + 0{,}011} - \frac{1}{R_4} \right) \right];$$

$$\Delta u = 0{,}9 - 0{,}03 \ln R_4;$$

$$P = \frac{19}{R_1} + \frac{10{,}8}{R_2 + 0{,}017};$$

$$t_з = 0{,}55 + \frac{0{,}087 R_2}{R_4} +$$

$$+ \left[R_2^2 \ \ 6 + 0{,}079 N \ \ ^2 + \left(0{,}69 - \frac{0{,}0024 N R_2}{R_1} \right) \times \ \ 5{,}7 + 0{,}15 N \ \ R_1 R_2 \right]^{\frac{1}{2}} - \qquad (8.12)$$

$$- 0{,}5 \left\{ R_2 \ \ 3 + 0{,}079 N \ \ - 8 \left[1 - 0{,}134 R_2 - \frac{0{,}136 R_2}{R_4} - \frac{0{,}026 N R_2}{R_1} \right] \times \right.$$

$$\left. \times \left[1 - 0{,}134 R_2 + \frac{6{,}9}{8 + 3 R_3} \right]^{-1} - R_2 \ \ 4{,}9 + 0{,}055 N \right\};$$

$$s = \frac{8{,}55 R_2}{R_1}.$$

Наличие явных выражений (8.12) в данном случае вносит непринципиальное упрощение задачи, т. к. 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, ..., R_4, которые составляют вектор R управляемых параметров.

Приведем условия работоспособности (8.11) к стандартному виду:

$$y_i \le t_i, \ i \in 1{:}5 \ , \qquad (8.13)$$

где

$$y_1 = -N; \quad y_2 = -\Delta u; \quad y_3 = P; \quad y_4 = t_з; \quad y_5 = -s;$$
$$t_1 = -20; \quad t_2 = -0{,}8; \quad t_3 = 20; \quad t_4 = 15; \quad t_5 = 1{,}5.$$

Запас работоспособности по каждому из требований (8.13) представим в виде:

$$z_i = a_i \left[\frac{t_i - y_i}{\delta_i} - 1 \right], \qquad (8.14)$$

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.

Положим $\delta_i = 0,2|t_i|$ за исключением параметра $y_2 = -\Delta u$, имеющего меньший разброс порядка $0,05|t_2|$ [49].

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 $z_i \geq 0$ 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 T_2, 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. разд. 3.10)* критерия минимального запаса работоспособности. Для устранения прямых ограничений

$$0,5 \leq R_1 \leq 10;\ \ 0,5 \leq R_2 \leq 10;\ \ 0,05 \leq R_3 \leq 1;\ \ 0,1 \leq R_4 \leq 2,$$

имеющих вид $R_i' \leq R_i \leq R_i''$, 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

$$J\ \ x\ = \sum_{i=1}^{5} \exp\left[-\nu z_i\ \ x\ \right] \to \min_x,\ \ x \in R^4. \qquad (8.15)$$

На первом этапе для получения приемлемого начального приближения функционал (8.15) минимизировался при $\nu = 1$ из произвольно выбранной начальной точки $x_i^0 = 0;\ i \in 1:4$. При этом $J\ x_0\ \cong 3{,}046 \cdot 10^8$. С помощью алгоритма GZ1 была получена точка

$$\begin{aligned} &x_1 = 4{,}234; \quad x_2 = 1{,}30; \quad x_3 = 1{,}51; \quad x_4 = 0{,}794; \\ &R_1 = 5{,}56; \quad R_2 = 1{,}18; \quad R_3 = 0{,}054; \quad R_4 = 1{,}034; \\ &N = 32{,}57; \quad \Delta u = 0{,}899; \quad P = 12{,}44; \quad t_з = 14{,}72; \quad s = 1{,}815, \end{aligned} \tag{8.16}$$

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. 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:

$$\begin{aligned} &R_1 = 4{,}33; \quad R_2 = 0{,}956; \quad R_3 = 0{,}05; \quad R_4 = 0{,}05; \\ &N = 25{,}1; \quad \Delta u = 0{,}92; \quad P = 15{,}5; \quad t_з = 11{,}7; \quad s = 1{,}9. \end{aligned}$$

Полученные результаты согласуются с данными из книги [49], где применялись более сложные и универсальные процедуры оптимального параметрического синтеза. В данном случае удалось ограничиться простейшими средствами в виде метода циклического покоординатного спуска.

В данной ситуации можно было сразу ограничиться значениями (8.16), 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 (8.14), не приводит к цели из-за представленной на рис. 5.4 ситуации заклинивания для негладких целевых функционалов.

Таким образом, приведенные ранее результаты подтверждают целесообразность применения среднестепенных форм минимаксных критериев с точки зрения эффективности последующей оптимизации.

8.4. 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 достаточно глубоко вникать в специфику соответствующих представлений и язык, присущий 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.

На рис. 8.10 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 $T_1 = 100\text{—}130\,^\circ\text{C}$ до конверсии порядка 30—40%. Время пребывания реакционной массы (τ_1) в реакторе составляет около 5 час. 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).

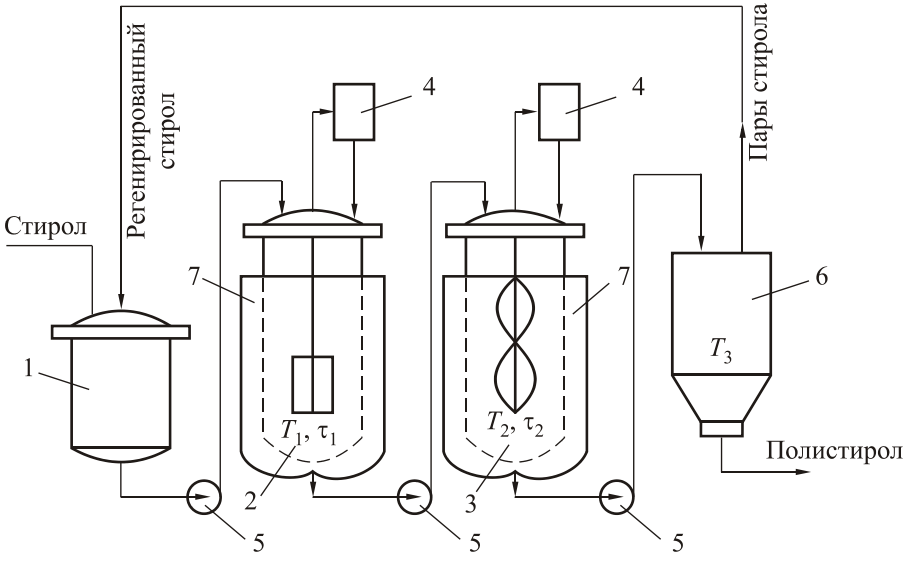


Рис. 8.10. Схема процесса производства блочного полистирола в каскаде аппаратов с перемешиванием (непрерывный способ)

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 варьироваться и также составляет около 5 час.

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. 8.11 занимает *модель кинетики* (МК), позволяющая по заданным режимным параметрам r_i вычислять значения конверсии z, средневесовой $\bar{M}_W$ и среднечисловой $\bar{M}_n$ молекулярных масс, которые и определяют основные характеристики полимера [10].

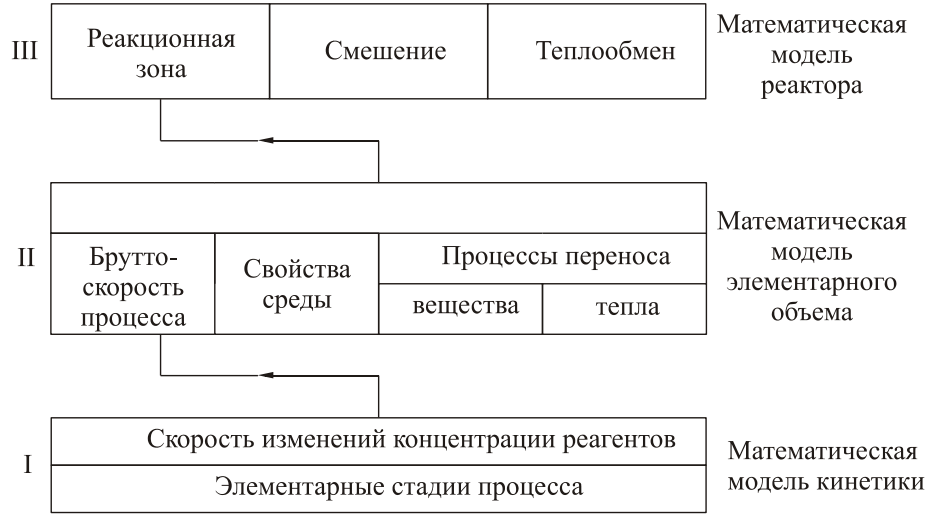


Рис. 8.11. Схема математической модели реактора

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 $T_i \in T_{\text{н}}, T_{\text{в}}$, $i \in 1:s$: $y^{i*} = \arg\min J_i\ \ y\ \ ,$

$$J_i\ \ y\ \ = \frac{1}{2}\sum_{j=1}^{N}\left\| \bar{x}_{\text{Э}}\ \ T_i, t_j\ \ - \bar{x}\ \ y, T_i, t_j\ \ \right\|_G^2,\ t_j \in\ 0,\ \bar{t}\ \ ,\ j \in 1:N\ ;$$

$\bar{x}$ — измеряемый подвектор вектора x; $\bar{x}_{\text{Э}}$ — заданные экспериментальные значения. На основе полученного набора векторов

$$\left[y^{1*},\ y^{2*},\ ...,\ y^{s*} \right]$$

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 $y_i\ \ T\ \ .$

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:

$$r^* \in \arg\min Q\left[x\ \ q^*,\ r\ \ \right],\ \ r \in D_r,$$

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.

8.4.1. 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. 8.12):

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

$$\dot{R}_c + M \xrightarrow{k_1} P_{1,1}.$$

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.

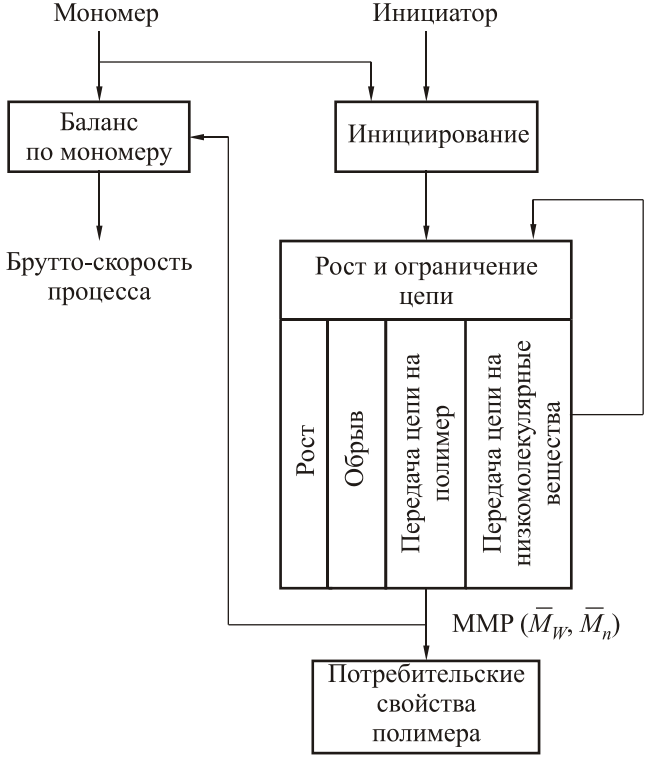


Рис. 8.12. Укрупненная схема модели кинетики полимеризационного процесса

Исходя из заданного множества элементарных стадий, может быть получена следующая система обыкновенных дифференциальных уравнений, моделирующая процесс полимеризации стирола

$$
\begin{aligned}
&\frac{d\dot{R}_c}{dt} = k_d M^2 - \dot{R}_c \;\; k_2' \dot{R}_c + k_1 M \;\; ; \\
&\frac{d\dot{M}}{dt} = M \;\; k_{4M} R - k_{1M} \dot{M} \;\; ; \\
&\frac{dM}{dt} = -M\left[k_d M + k_1 \dot{R}_c + \;\; k_2 + k_{4M} \;\; R + k_{1M} \dot{M} \right]; \\
&\frac{dR}{dt} = M \;\; k_1 \dot{R}_c + k_{1M} \dot{M} \;\; - R \;\; k_{4M} M + k_3 R \;\; ; \\
&\frac{dP}{dt} = k_2 \dot{M} R; \\
&\frac{dN}{dt} = M \;\; k_1 \dot{R}_c + k_{1M} \dot{M} \;\; - 0{,}5 k_3 R^2; \\
&\frac{d\nu_1^1}{dt} = \nu_2^0 \;\; k_{4P} R + k_2 M \;\; - \nu_1^1 \;\; 2k_3 R + k_{4M} M + k_{4P} P - k_3 \nu_2^0 \;\; ; \\
&\frac{d\nu_0^2}{dt} = \nu_1^1 \;\; 2k_2 M + k_3 \nu_1^1 \;\; ; \\
&\frac{d\nu_2^0}{dt} = k_3 \;\; \nu_2^0 \;\; ^2 - 2\nu_2^0 \;\; 2k_3 R + k_{4M} M + k_{4P} P \;\; + \\
&\quad + R\left[2k_3 R + k_{4M} M + 2k_{4P} \;\; P + \nu_1^1 \;\; \right] + M \;\; k_{1M} \dot{M} + k_1 \dot{R}_c \;\; .
\end{aligned}
\tag{8.17}
$$

Конечность системы уравнений (8.17) 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 (8.17):

$$\dot{R}_c = \dot{M} = R = P = N = \nu_1^1 = \nu_0^2 = \nu_2^0 = 0;\ M = 8.$$

Общая методика построения подобных моделей для различных полимеризационных процессов представлена, например, в работе [79].

8.4.2. Методика воспроизведения моделей полимеризационных процессов

Практическое исследование моделей полимеризационных процессов, аналогичных (8.17), 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 приводит к свойству жесткости систем дифференциальных уравнений (8.17), что определяет известные вычислительные трудности при их интегрировании с помощью явных схем численного интегрирования [57]. 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 (8.17) явные схемы по скорости в 10—15 раз. В процессе идентификации возникает необходимость многократного интегрирования систем типа (8.17), и поэтому уже незначительная экономия 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 (8.17) на компьютере, основанная на неявных методах ломаных (НМЛ) и трапеций (НМТ) с аналитическим разрешением возникающих неявных зависимостей.

Для системы уравнений (8.17), записанной в виде

$$\frac{dx}{dt} = f\ \ x\ ,$$

уравнения НМЛ, НМТ получаются при аппроксимации интеграла в представлении

$$x\ \ t_{k+1}\ = x\ \ t_k\ + \int_{t_k}^{t_{k+1}} f\Big[x\ \ \tau\ \Big]d\tau = x\ \ t_k\ + \int_0^h f\Big[x\ \ t_k + \tau\ \Big]d\tau,$$

соответственно, по способу правых прямоугольников

$$x^{k+1} = x^k + hf^{k+1},\ f^{k+1} \triangleq f\ x^{k+1} \tag{8.18}$$

и трапеций

$$x^{k+1} = x^k + \left(\frac{h}{2}\right)\ f^k + f^{k+1}\ . \tag{8.19}$$

Для разрешения уравнения (8.18) относительно x^{k+1} можно воспользоваться следующим приемом. Запишем *i*-е уравнение (8.18) в виде

$$\frac{x_i^{k+1} - x_i^k}{h} = f_i^{k+1} = f_i\ x^{k+1}\ . \tag{8.20}$$

Используя первые члены разложения Тейлора, можно приближенно представить правую часть (8.20) в виде

$$f_i^{k+1} \cong f_i^k + \sum_{j=1}^{n} \frac{\partial f_i\ x^k}{\partial x_j^k}\ x_j^{k+1} - x_j^k\ ,\ i \in 1:n \tag{8.21}$$

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

$$\frac{x_i^{k+1} - x_i^k}{h} = f_i^k + \sum_{j=1}^{n} \frac{\partial f_i\ x^k}{\partial x_j^k}\ x_j^{k+1} - x_j^k\ . \tag{8.22}$$

Аналогичным образом, НМТ (8.19) может быть записан в виде

$$\frac{x_i^{k+1} - x_i^k}{h} = f_i^k + 0{,}5\sum_{j=1}^{n} \frac{\partial f_i\ x^k}{\partial x_j^k}\ x_j^{k+1} - x_j^k\ ,\ i \in 1:n\ . \tag{8.23}$$

Формулы (8.22), (8.23) обладают тем преимуществом перед исходными соотношениями (8.18), (8.19), что x_i^{k+1} входит линейно и уравнения (8.22), (8.23) могут рассматриваться как системы линейных алгебраических уравнений относительно неизвестных $x_i^{k+1},\ i \in 1:n$. Применение соотношений (8.22), (8.23), очевидно, эквивалентно разрешению нелинейных систем уравнений (8.18), (8.19) относительно x^{k+1} с помощью одного шага метода Ньютона.

Для динамических моделей, описывающих кинетику радикальной полимеризации, соотношения (8.22), (8.23) могут быть эффективно разрешены относительно x^{k+1} с получением явных представлений

$$x^{k+1} = G\ x^k\ , \tag{8.24}$$

где вектор-функция *G* определяется видом вектор-функции $f\ x$ и используемым методом (8.22) или (8.23).

Основная идея предлагаемой методики численного интегрирования системы (8.17) для целей ее последующей параметрической идентификации заключается в использовании на первом этапе НМЛ (8.22) 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 (8.23), что позволяет существенно нарастить значение шага h при сохранении устойчивости и точности воспроизведения медленно изменяющихся составляющих (порядок точности метода (8.23) есть $o\ h^3$, тогда как для метода (8.22) имеем ошибку порядка $o\ h^2$).

Заметим, что из условия сохранения устойчивости процесса численного интегрирования при использовании явных схем, процедура наращивания шага h при $t > \tau_{пс}$ не может применяться [57]. 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 по точности, как правило, позволяет выбрать достаточно большой шаг, однако при этом возникают трудности в подавлении быстро затухающих составляющих решения.

На основе приведенных замечаний применялась комплексная методика численного интегрирования системы уравнений (8.17), усложненной дополнительным предположением

$$k_2 := k_2 \exp\left[\beta_2 \ \ M - \bar{M}\ \right];$$
$$k_3 := k_3 \exp\left[\beta_3 \ \ M - \bar{M}\ \right],$$

где $\bar{M}$ — некоторое фиксированное значение M. (При конкретных расчетах полагали $\bar{M} = M_0 = 8$.) Согласно этой методике вводятся следующие параметры блока интегрирования:

- $\tau_{пс}$ — длина пограничного слоя;
- h — начальный шаг интегрирования в пограничном слое неявным методом ломаных;
- E — коэффициент увеличения шага;
- N — целое число, определяющее при $t < \tau_{пс}$ моменты изменения шага;
- T — промежуток интегрирования; $t \in 0, T$.

При $t < \tau_{пс}$ интегрирование ведется НМЛ с постоянным шагом h. По достижении точки $t = \tau_{пс}$ осуществляется переход на НМТ с увеличением шага интегрирования в E раз через каждые N шагов ($h := h \times E$).

Экспериментально установлено, что для моделирования процесса термоинициированной полимеризации стирола (8.17) целесообразно полагать:

$$\tau_{\text{пс}} = 0,5; \quad h = 0,01; \quad E = 2; \quad N = 4;$$

T определяется конкретным вариантом решения (это — десятки часов). НМЛ и НМТ использовались в виде модификаций (8.22), (8.23).

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 4-х знаках мантиссы, что существенно выше точности используемых на этапе параметрической идентификации экспериментальных данных.

Характерное время одновариантного анализа системы классическими явными схемами сравнимо с продолжительностью T реального процесса (порядка 10 часов).

Процедуры Гира и им аналогичные при соответствующей настройке позволяют существенно сократить это время, но недостаточно.

Приведенная методика численного интегрирования по сравнению с "библиотечными" методами, в том числе и методами типа гировских, позволяет получить выигрыш в тысячи раз.

8.4.3. Параметрическая идентификация кинетических моделей полимеризационных процессов (полимеризация стирола)

При параметрической идентификации процесса полимеризации стирола на основе модели (8.17) в качестве экспериментальных данных используются зависимости от времени конверсии z, среднечисловой $\bar{M}_n$ и средневесовой $\bar{M}_W$ молекулярных масс, полученные в изотермическом режиме. Указанные характеристики следующим образом связаны с переменными системы (8.17) [10]:

$$z\ t\ = \frac{M\ 0\ - M\ t}{M\ 0}; \ \bar{M}_n = \frac{P}{N} m_0; \ \bar{M}_W = \frac{\nu_0^2}{P} m_0, \tag{8.25}$$

где $m_0 = 106$ — молекулярный вес стирола; $N \triangleq \nu_0^0$ — общее число макромолекул.

Показатели $\bar{M}_n$, $\bar{M}_W$ 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. 8.13. Блок 1 реализует преобразование вектора y в множество векторов x^j фазовых переменных системы (8.17), соответствующих заданным моментам времени t_j,

$j \in 1:N$ съема экспериментальной информации. Численное интегрирование системы (8.17) производится с помощью изложенной в *разд. 8.4.2* методики. Блок 2 осуществляет перевод фазовых переменных модели в измеряемые экспериментально характеристики (8.25). 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.

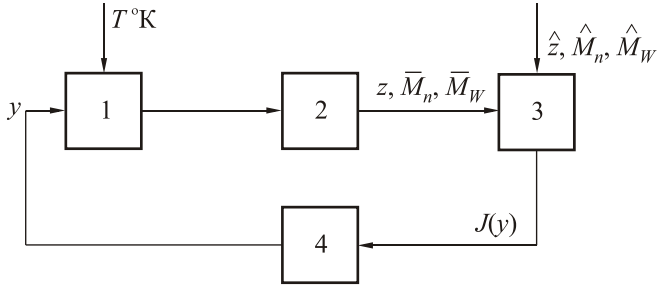


Рис. 8.13. Схема настройки кинетической модели на экспериментальные данные

В качестве критерия оптимальности используется взвешенный метод наименьших квадратов, согласно которому минимизируемый функционал имеет вид

$$J\ y = \frac{1}{2}\sum_{j=1}^{N}\left[\alpha_j\ \bar{M}_{nj} - \hat{M}_{nj}\ ^2 + \beta_j\ \bar{M}_{Wj} - \hat{M}_{Wj}\ ^2 + \gamma_j\ z_j - \hat{z}_j\ ^2\right], \qquad (8.26)$$

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; $\bar{M}_{nj} = M_n\ t_j\ ,\ \bar{M}_{Wj} = M_W\ t_j\ ,\ z_j = z\ t_j$ — расчетные значения; $\hat{M}_{nj},\ \hat{M}_{Wj},\ \hat{z}_j$ — экспериментальные значения в точках t_j, $j \in 1:N$ съема измерительной информации; $\alpha_j, \beta_j, \gamma_j$ — нормирующие коэффициенты, равные обратным квадратам соответствующих экспериментальных значений.

Минимизация функционала (8.26) осуществлялась с помощью алгоритма RELEX при периодическом включении процедуры MIO, изложенной в *разд. 7.1.5*. 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 *разд. 5.4.1* с конечноразностной аппроксимацией коэффициентов чувствительности первого порядка.

Подгонка модели (8.17) производилась при различных температурных режимах. Далее в качестве примера представлены результаты, полученные для температуры 120 °C. Аналогичные результаты получены и для других значений температур на интервале, представляющем технологический интерес (100—170 °C).

В табл. 8.1 приведены выборочные данные лабораторных исследований, взятые в качестве экспериментальных значений при решении задачи параметрической идентификации.

Таблица 8.1

Время (час)	z	$\frac{\bar{M}_n}{10^6}$	$\frac{\bar{M}_W}{10^6}$
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 [80], вращения осей Розенброка, симплексный метод Нелдера — Мида.

Лучший результат $\bar{y}$, полученный для данной задачи с помощью попеременного применения указанных процедур, является неудовлетворительным:

$$\bar{y}_1 = k_d = 0,736586 \cdot 10^{-8};\ \bar{y}_5 = k_3 = 0,693696 \cdot 10^{8};$$
$$\bar{y}_2 = k_3' = 0,287481 \cdot 10^{11};\ \bar{y}_6 = k_{1M} = 0,246245 \cdot 10^{-4};$$
$$\bar{y}_3 = k_1 = 0,220636 \cdot 10^{8};\ \bar{y}_7 = k_{4M} = 0,495080 \cdot 10^{4};$$
$$\bar{y}_4 = k_2 = 0,534836 \cdot 10^{3};\ \bar{y}_8 = k_{4P} = 0,189181;$$
$$J\ \bar{y}\ = 0,29437. \qquad (8.27)$$

В табл. 8.2 представлены соответствующие значения конверсии z, среднечисловой $\bar{M}_n$ и средневесовой $\bar{M}_W$ молекулярных масс в соответствующих табл. 8.1 экспериментальных точках.

Таблица 8.2

Время (час)	z	$\frac{\bar{M}_n}{10^6}$	$\frac{\bar{M}_W}{10^6}$
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:

$$\lambda_1 = 0{,}13710 \cdot 10^3; \qquad \lambda_4 = 0{,}24827 \cdot 10^{-1};$$
$$\lambda_2 = 0{,}27146 \cdot 10^2; \qquad \lambda_5 = 0{,}30857 \cdot 10^{-2};$$
$$\lambda_3 = 0{,}34817 \cdot 10^1; \qquad \lambda_6 = \lambda_7 = \lambda_8 = 0{,}0.$$

Точка $\bar{y}$ была выбрана как начальная при идентификации методом RELEX в указанной выше реализации. Полученные результаты представлены в табл. 8.3.

Таблица 8.3

Время (час)	z	$\frac{\bar{M}_n}{10^6}$	$\frac{\bar{M}_W}{10^6}$
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 (8.27) и имеют следующие значения:

$$y_1 = 0{,}123051 \cdot 10^{-11}; \qquad y_5 = 0{,}206709 \cdot 10^3;$$
$$y_2 = 0{,}287480 \cdot 10^{11}; \qquad y_6 = 0{,}773434 \cdot 10^{-1};$$
$$y_3 = 0{,}296799 \cdot 10^{10}; \qquad y_7 = 0{,}44929 \cdot 10^4; \qquad (8.28)$$
$$y_4 = 0{,}109727 \cdot 10^3; \qquad y_8 = 0{,}410339 \cdot 10^{-1};$$
$$J\ \ y\ = 0{,}57178 \cdot 10^{-1}.$$

Значения параметров (8.28) позволяют весьма удовлетворительно описать экспериментальные данные табл. 8.1. Погрешность оказывается в пределах погрешности эксперимента ($5 \div 10\%$). Напротив, результаты табл. 8.2 значительно отличаются от требуемых значений (максимальная погрешность достигает 30%).

При решении данной задачи идентификации осуществлялась проверка получаемых результатов. 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 по сравнению с методиками, основанными на традиционных схемах численного интегрирования и конечномерной оптимизации.

8.5. 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:

$$\begin{aligned}
&\frac{\partial T_1}{\partial t}+\omega_1\ t\ \frac{\partial T_1}{\partial l}+\kappa_1\ t\ \ T_1-T_3\ =0;\\
&\frac{\partial T_2}{\partial t}-\omega_2\ t\ \frac{\partial T_2}{\partial l}+\kappa_2\ t\ \ T_2-T_3\ =0;\\
&\delta\frac{\partial T_3}{\partial t}+\kappa_3\ t\ \ T_3-T_1\ +\kappa_4\ t\ \ T_3-T_2\ =0.
\end{aligned} \tag{8.29}$$

Здесь T_1, T_2, T_3 — температура 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 (8.29) подчиняется граничным условиям:

$$\begin{aligned}
&T_1\ \ l,\ t\ \big|_{l=0}=T_{1\text{вх}}\ \ t\ ;\\
&T_2\ \ l,\ t\ \big|_{l=L}=T_{2\text{вх}}\ \ t\ .
\end{aligned} \tag{8.30}$$

Для построения системы управления объектом с распределенными параметрами необходимо многократно воспроизводить динамику объекта на основе решения системы (8.29) при условиях (8.30). Однако решение подобных задач аналитически невозможно, а численный подход недостаточно эффективен по временным затратам.

В [67], [68] предложена методика аппроксимации системы нелинейных уравнений в частных производных типа (8.29) 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:

$$\begin{aligned}
&W_1\ \ p\ \triangleq\frac{\Delta T_{1\text{вых}}\ \ p}{\Delta T_{1\text{вх}}\ \ p}=\frac{k_1\ \ 1+\eta_{12}p}{1+\eta_{11}p}\exp\ -p\tau_1\ ;\\
&W_2\ \ p\ \triangleq\frac{\Delta T_{1\text{вых}}\ \ p}{\Delta T_{2\text{вх}}\ \ p}=\frac{k_2}{1+\eta_{21}p\ \ 1+\eta_{22}p};\\
&W_3\ \ p\ \triangleq\frac{\Delta T_{1\text{вых}}\ \ p}{\Delta G_1\ \ p}=\frac{k_3\ \ 1+2\xi_1\eta_{31}p+\eta_{31}^2p^2}{1+\eta_{32}p\ \ 1+\eta_{33}p\ \ 1+\eta_{34}p\ ^2};\\
&W_4\ \ p\ \triangleq\frac{\Delta T_{1\text{вых}}\ \ p}{\Delta G_2\ \ p}=\frac{k_4\ \ 1+2\xi_2\eta_{41}p+\eta_{41}^2p^2}{1+\eta_{42}p\ \ 1+\eta_{43}p\ \ 1+\eta_{44}p\ ^2}.
\end{aligned} \tag{8.31}$$

Температура на выходе первого контура теплообменника $\Delta T_{1\text{вых}}$ описывается уравнением

$$\Delta T_{1\text{вых}}\ p = W_1\ p\ \Delta T_{1\text{вх}}\ p + W_2\ p\ \Delta T_{2\text{вх}}\ p + \\ + W_3\ p\ \Delta G_1\ p + W_{14}\ p\ \Delta G_2\ p\ . \tag{8.32}$$

Второй контур теплообменника описывается аналогичными соотношениями.

Ставится задача построения уточненных значений параметров η_{ij} передаточных функций $W_1 \div W_8$ 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 (8.29).

Для каждой из передаточных функций $W_i\ p$ задача параметрической идентификации формулируется как оптимизационная с критерием, построенным по методу наименьших квадратов:

$$J_i\ \eta_i = \sum_{k=1}^{11} \left[A_i\ \omega_k - \bar{A}_i\ \omega_k \right]^2 + \left[\varphi_i\ \omega_k - \bar{\varphi}_i\ \omega_k \right]^2\ ,\ i \in 1{:}8\ , \tag{8.33}$$

где $\omega_k \in \left[0{,}62832 \cdot 10^{-2};\ 0{,}62832 \right]$; A_i, φ_i — расчетные значения амплитуды и фазы, соответствующие приближенной передаточной функции W_i с параметрами $\eta_i = \eta_{i1}, ..., \eta_{in}$, $\eta_{ik} \geq 0$; $\bar{A}_i$, $\bar{\varphi}_i$ — аппроксимируемые значения, полученные в результате линеаризации исходной системы уравнений в частных производных относительно известных установившихся режимов и построения точных передаточных функций $\bar{W}_i$. Задачи (8.33) решались для каждого из трех рассматриваемых режимов работы теплообменника для значений $0{,}1P_\text{н}$, $0{,}5P_\text{н}$ и $P_\text{н}$, где $P_\text{н}$ — номинальная мощность.

Ограничения устранялись переходом к новым переменным $y_{ik} : \eta_{ik} = y_{ik}^2$. В качестве алгоритма оптимизации использовалась процедура RELAX *(см. разд. 6.3)*. Полученные результаты представлены в табл. 8.4.

Таблица 8.4

Погрешность		**Первый контур**				**Второй контур**			
		W_1	W_2	W_3	W_4	W_1	W_2	W_3	W_4
$P_\text{н}$	$\Delta\varphi_{\max}$, °	0,24	0,10	0,29	0,218	0,20	0,32	0,08	0,07
	$\Delta A_{\max}$, %	0,13	0,371	0,37	0,272	0,79	0,16	0,10	0,12

Таблица 8.4 *(окончание)*

Погрешность		Первый контур				Второй контур			
		W_1	W_2	W_3	W_4	W_1	W_2	W_3	W_4
$0{,}5P_н$	$\Delta\varphi_{max}$, °	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
$0{,}1P_н$	$\Delta\varphi_{max}$, °	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 (8.31) 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 $-\dfrac{1}{\eta_{ik}}$.

Последнее позволяет найти аналитические решения, справедливые в окрестности выбранного режима линеаризации и используемые, например, в процессе управления атомным реактором, как объектом эксперимента [68].

Литература

1. Алберт А. Регрессия, псевдоинверсия и рекуррентное оценивание. — М.: Наука, 1977.
2. Алексеев В. М., Тихомиров В. М., Фомин С. В. Оптимальное управление. — М.: Наука, 1979.
3. Андреев Н. И. Корреляционная теория статистически оптимальных систем. — М.: Наука, 1966.
4. Аоки М. Введение в методы оптимизации. — М.: Мир, 1977.
5. Аттетков А. В., Галкин С. В., Зарубин В. С. Методы оптимизации. — М.: Изд-во МГТУ им. Н. Э. Баумана, 2001.
6. Базара М., Шетти К. Нелинейное программирование. — М.: Мир, 1982.
7. Батищев Д. И. Поисковые методы оптимального проектирования. — М.: Советское радио, 1975.
8. Бахвалов Н. С. Численные методы. — М.: Наука, 1973.
9. Беляков Ю. Н., Курмаев Ф. А., Баталов Б. В. Методы статистических расчетов микросхем на ЭВМ. — М.: Радио и связь, 1985.
10. Берлин А. А., Вольфсон С. А. Кинетический метод в синтезе полимеров. — М.: Химия, 1973.
11. Бесекерский В. А., Попов Е. П. Теория систем автоматического управления. — М.: Наука, 1972.
12. Болнокин В. Е., Чинаев П. И. Анализ и синтез систем автоматического управления на ЭВМ. Алгоритмы и программы. — М.: Радио и связь, 1986.
13. Брамеллер А., Аллан Р., Хэмэм Я. Слабозаполненные матрицы. — М.: Энергия, 1979.
14. Вайнберг М. М. Вариационный метод и метод монотонных операторов. — М.: Наука, 1972.
15. Васильев Ф. П. Численные методы решения экстремальных задач. — М.: Наука, 1980.
16. Вентцель Е. С. Исследование операций. — М.: Сов. радио, 1972.

17. Вязгин В. А., Федоров В. В. Математические методы автоматизированного проектирования. — М.: Высшая школа, 1989.
18. Гантмахер Ф. Р. Теория матриц. — М.: Наука, 1966.
19. Геминтерн В. И., Каган Б. М. Методы оптимального проектирования. — М.: Энергия, 1980.
20. Гермейер Ю. Б. Игры с непротивоположными интересами. — М.: Наука, 1976.
21. Гилл Ф., Мюррей У., Райт М. Практическая оптимизация. — М.: Мир, 1985.
22. Гроп Д. Методы идентификации систем. — М.: Мир, 1979.
23. Гроссман К., Каплан А. Нелинейное программирование на основе безусловной минимизации. — Новосибирск: Наука, 1981.
24. Дейч А. М. Методы идентификации динамических объектов. — М.: Энергия, 1979.
25. Джордж А., Лю Дж. Численное решение больших разреженных систем уравнений. — М.: Мир, 1984.
26. Деннис Дж., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений. — М.: Мир, 1988.
27. Дубов Ю. А., Травкин С. И., Якимец В. Н. Многокритериальные методы формирования и выбора вариантов систем. — М.: Наука, 1986.
28. Дьедонне Ж. Основы современного анализа: Пер. с англ. — М.: Мир, 1964.
29. Евтушенко Ю. Г. Методы решения экстремальных задач и их применение в системах оптимизации. — М.: Наука, 1982.
30. Ермольев Ю. М. Методы стохастического программирования. — М.: Наука, 1976.
31. Зангвилл У. И. Нелинейное программирование. — М.: Сов. радио, 1973.
32. Карманов В. Г. Математическое программирование. — М.: Наука, 1975.
33. Картан А. Дифференциальное исчисление. Дифференциальные формы. — М.: Мир, 1971.
34. Катковник В. Я. Непараметрическая идентификация и сглаживание данных. — М.: Наука, 1985.
35. Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. — М.: Мир, 1998.
36. Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. — М.: Наука, 1972.
37. Ланнэ А. А. Оптимальный синтез линейных электронных схем. — М.: Связь, 1978.
38. Ланнэ А. А., Михайлова Е. Д., Саркисян Б. С., Матвийчук Я. Н. Оптимальная реализация линейных электронных RLC-схем. — Киев: Наукова Думка, 1982.
39. Ланцош К. Практические методы прикладного анализа. — М.: Гос. изд-во физ. мат. лит-ры, 1961.
40. Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов. — М.: Наука, 1986.

41. Лыпарь Ю. И. Автоматизация проектирования избирательных усилителей и генераторов. — Л.: Изд-во ЛГУ, 1983.
42. Лэм Г. Аналоговые и цифровые фильтры. — М.: Мир, 1982.
43. Люстерник Л. А., Соболев В. И. Элементы функционального анализа. — М.: Наука, 1965.
44. Маркус М., Минк Х. Обзор по теории матриц и матричных неравенств. — М.: Наука, 1972.
45. Мину М. Математическое программирование. Теория и алгоритмы: Пер. с фр. — М.: Наука, 1990.
46. Моисеев Н. Н. Математические задачи системного анализа. — М.: Наука, 1981.
47. Немировский А. С., Юдин Д. Б. Сложность задач и эффективность методов оптимизации. — М.: Наука, 1979.
48. Норенков И. П. Введение в автоматизированное проектирование технических устройств и систем. — М.: Высшая школа, 1986.
49. Норенков И. П., Мулярчик С. Г., Иванов С. Р. Экстремальные задачи при схемотехническом проектировании в электронике. — Минск: Изд-во БГУ, 1976.
50. Ортега Дж., Рейнболдт В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными. — М.: Мир, 1975.
51. Парлетт Б. Симметричная проблема собственных значений. — М.: Мир, 1983.
52. Первозванский А. А., Гайцгори В. Г. Декомпозиция, агрегирование и приближенная оптимизация. — М.: Наука, 1979.
53. Подиновский В. В., Ногин В. Д. Парето-оптимальные решения многокритериальных задач. — М.: Наука, 1982.
54. Поляк Б. Т. Введение в оптимизацию. — М.: Наука, 1983.
55. Ракитский Ю. В. Новые численные методы решения систем обыкновенных дифференциальных и разностных уравнений. / В кн. Труды ЛПИ № 332. — Л.: ЛПИ, 1973.
56. Ракитский Ю. В., Устинов С. М., Черноруцкий И. Г. Численные методы решения жестких систем обыкновенных дифференциальных уравнений. — Л.: ЛПИ, 1977.
57. Ракитский Ю. В., Устинов С. М., Черноруцкий И. Г. Численные методы решения жестких систем. — М.: Наука, 1979.
58. Растригин Л. А. Системы экстремального управления. — М.: Наука, 1974.
59. Растригин Л. А. Современные принципы управления сложными объектами. — М.: Сов. радио, 1980.
60. Растригин Л. А., Маджаров Н. Е. Введение в идентификацию объектов управления. — М.: Энергия, 1977.
61. Рей У. Методы управления технологическими процессами. — М.: Мир, 1983.
62. Розенброк Х., Стори С. Вычислительные методы для инженеров-химиков. — М.: Мир, 1968.
63. Сеа Ж. Оптимизация. Теория и алгоритмы. — М.: Мир, 1973.

64. Соболь И. М. Точки, равномерно заполняющие многомерный куб. — М.: Знание, 1985.

65. Стренг Г. Линейная алгебра и ее применения. — М.: Мир, 1980.

66. Стронгин Р. Г. Численные методы в многоэкстремальных задачах. — М.: Наука, 1978.

67. Тарасов В. С., Веренинов И. А., Ерунов В. Я. Моделирование технологических процессов с распределенными параметрами. — Л.: Изд-во ЛПИ, 1982.

68. Тарасов В. С., Савченко Е. С. Параметрическая идентификация объектов эксперимента с распределенными параметрами с использованием системного метода оптимизации / В кн. Труды ЛПИ № 381. — Л.: ЛПИ, 1982.

69. Таха Х. Введение в исследование операций. Кн. 1,2. — М.: Мир, 1985.

70. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. — М.: Наука, 1979.

71. Уайлд Д. Оптимальное проектирование. — М.: Мир, 1981.

72. Уилкинсон Дж. Х. Алгебраическая проблема собственных значений. — М.: Наука, 1970.

73. Уилкинсон Дж. Х., Райнш С. Справочник алгоритмов на языке АЛГОЛ. — М.: Машиностроение, 1976.

74. Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры. — М.: Физматгиз, 1963.

75. Федоренко Р. П. Приближенное решение задач оптимального управления. — М.: Наука, 1978.

76. Фельдбаум А. А. Вычислительные устройства в автоматических системах. — М.: Физматгиз, 1959.

77. Форсайт Дж., Моулер К. Численное решение систем линейных алгебраических уравнений. — М.: Мир, 1969.

78. Форсайт Дж., Малькольм М., Моулер К. Машинные методы математических вычислений. — М.: Мир, 1980.

79. Хохлов В. А., Любецкий С. Г. Особенности применения преобразующих функций для математического описания кинетики полимеризационных процессов. / В кн. Полимеризационные процессы. Аппаратное оформление и математичское моделирование. — Л.: ОНПО "Пластполимер", 1976.

80. Химмельблау Д. Прикладное нелинейное программирование. — М.: Мир, 1975.

81. Черноруцкий И. Г. Оптимальный параметрический синтез: электротехнические устройства и системы. — Л.: Энергоатомиздат, 1987.

82. Черноруцкий И. Г. Методы оптимизации. — СПб.: Изд-во СПбГТУ, 1998.

83. Черноруцкий И. Г. Методы оптимизации и принятия решений. — СПб.: Лань, 2001.

84. Черноруцкий И. Г. Методы оптимизации в теории управления. — СПб.: Питер, 2004.

85. Черноруцкий И. Г. Методы принятия решений. — СПб.: БХВ-Петербург, 2005.

86. Эйкхофф П. Основы идентификации систем управления. — М.: Мир, 1975.

Предметный указатель

(окончание рубрики см. на стр. 368)

Н

О

(окончание рубрики см. на стр. 370)

У

Ф

Ш

Э, Я