И. К. Даугавет

ТЕОРИЯ ПРИБЛИЖЕННЫХ МЕТОДОВ
линейные уравнения
2-е издание

- *Общая схема исследования сходимости и устойчивости приближенных методов решения линейных функциональных уравнений*
- *Проекционные методы*
- *Метод конечных элементов*
- *Метод сеток*

УЧЕБНОЕ ПОСОБИЕ

bhv®

И. К. Даугавет

ТЕОРИЯ ПРИБЛИЖЕННЫХ МЕТОДОВ
линейные уравнения
2-е издание

Рекомендовано УМО по образованию в области инновационных междисциплинарных образовательных программ в качестве учебного пособия по специальности «Математическое обеспечение и администрирование информационных систем» — 010503

Санкт-Петербург
«БХВ-Петербург»
2006

УДК 519.6(075.8)
ББК 22.143я73
Д21

Даугавет И. К.

Д21 Теория приближенных методов. Линейные уравнения. —
2-е изд., перераб. и доп. — СПб.: БХВ-Петербург, 2006. — 288 с.: ил.

ISBN 5-94157-737-0

Книга является вторым, исправленным и дополненным, изданием опубликованного в 1985 году учебника «Приближенное решение линейных функциональных уравнений». Излагается исследование основных приближенных методов решения задач математической физики (проекционные методы, метод сеток, включая метод конечных элементов), основанное на общей схеме, использующей язык функционального анализа. Конкретными объектами исследования являются метод механических квадратур для интегральных уравнений (используется принцип компактной аппроксимации), методы Ритца, Галеркина, метод сеток для эллиптических уравнений, уравнений теплопроводности и колебаний струны. Основное внимание уделяется вопросам сходимости и устойчивости. Некоторые из результатов принадлежат автору.

В новом издании добавлены некоторые результаты, касающиеся метода конечных элементов и устойчивости.

Для студентов технических вузов и математических факультетов университетов, специалистов в области приближенных методов и их приложений

УДК 519.6(075.8)
ББК 22.143я73

Рецензенты:
*Рябов В. М., доктор физико-математических наук, профессор,
заведующий кафедрой методов вычислений Санкт-Петербургского государственного университета;*
*Демьянович Ю. К., доктор физико-математических наук, профессор,
заведующий кафедрой параллельных алгоритмов Санкт-Петербургского государственного университета*

Группа подготовки издания:

Главный редактор	*Екатерина Кондукова*
Зам. главного редактора	*Людмила Еремеевская*
Зав. редакцией	*Григорий Добин*
Компьютерная верстка	*Натальи Смирновой*
Корректор	*Наталия Першакова*
Дизайн серии	*Игоря Цырульникова*
Оформление обложки	*Елены Беляевой*
Зав. производством	*Николай Тверских*

Лицензия ИД № 02429 от 24.07.00. Подписано в печать 22.12.05.
Формат 70×100 1/16. Печать офсетная. Усл. печ. л. 23,22.
Тираж 2000 экз. Заказ №
"БХВ-Петербург", 194354, Санкт-Петербург, ул. Есенина, 5Б.

Санитарно-эпидемиологическое заключение на продукцию № 77.99.02.953.Д.006421.11.04
от 11.11.2004 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 5-94157-737-0

© Даугавет И. К., 2006
© Оформление, издательство "БХВ-Петербург", 2006

Оглавление

Предисловие ко второму изданию **1**

Из предисловия к первому изданию **3**

Глава 1. Общая схема приближенных методов **7**

§ 1. Близкие уравнения 7
§ 2. Общая схема приближенных методов. Сходимость 14
§ 3. Устойчивость 19

Глава 2. Приближенное решение интегральных уравнений **29**

§ 1. Метод замены ядра на вырожденное 29
§ 2. Метод механических квадратур. Теорема о сходимости 36
§ 3. Метод механических квадратур. Оценка погрешности 50
§ 4. Итеративное решение систем уравнений метода механических квадратур 63

Глава 3. Проекционные методы **71**

§ 1. Сущность проекционных методов 71
§ 2. Теорема о сходимости 78
§ 3. Метод Галеркина для уравнений второго рода 85
§ 4. Метод моментов для обыкновенных дифференциальных уравнений 96
§ 5. Метод наименьших квадратов 102
§ 6. Метод Ритца 108
§ 7. Применение метода Ритца к обыкновенным дифференциальным уравнениям 122
§ 8. Применение метода Ритца к дифференциальным уравнениям эллиптического типа 132
§ 9. Метод Галеркина 142

Глава 4. Метод сеток..... 149

§ 1. Сущность метода сеток 149
Метод сеток для обыкновенного дифференциального уравнения..... 149
Аппроксимация дифференциальных выражений 150
Метод сеток для эллиптического уравнения..... 156
Вариационно-разностные схемы..... 161
Особенности применения метода сеток к нестационарным уравнениям 166
Особенности применения общей схемы приближенных методов к исследованию метода сеток 173
§ 2. Вспомогательные сведения из линейной алгебры 177
§ 3. Метод конечных элементов для обыкновенного дифференциального уравнения 184
§ 4. Метод сеток для эллиптических уравнений..... 193
§ 5. Явная разностная схема для уравнения теплопроводности 202
§ 6. О системах уравнений 210
§ 7. Неявная разностная схема для уравнения теплопроводности..... 226
§ 8. Метод сеток для уравнения колебаний струны 240
§ 9. Метод прогонки 257

Литература..... 271

Предисловие ко второму изданию

Дополнения, внесенные во второе издание, касаются в основном вопросов устойчивости проекционных методов и сверхсходимости метода конечных элементов. Изменены некоторые доказательства.

Нумерация теорем, формул в каждом параграфе своя. Ссылки на формулы, теоремы из других параграфов той же главы даются с указанием параграфа, а из другой главы — с указанием параграфа и главы.

Читателям, заинтересованным в более детальном изучении затронутых в книге вопросов, можно рекомендовать книги [2], [14]. В качестве дополнительной литературы к главе 1 можно указать также [10], глава XIV, к главе 2 — [11], глава II, к главе 3 — [15], к главе 4 — [18], [19].

Автор благодарит Ю. К. Демьяновича и В. М. Рябова, замечания которых были учтены при составлении окончательного текста книги.

Из предисловия к первому изданию

Книга написана на основе лекций, которые под названием "Методы вычислений-2" автор читает студентам IV курса математико-механического факультета Ленинградского университета, специализирующимся по отделениям математики и прикладной математики. Содержанием лекций являются численные методы решения линейных интегральных и дифференциальных уравнений. Особенности курса лекций, а тем самым и книги, диктуются следующими четырьмя обстоятельствами: краткостью курса (лекции читаются один семестр по три часа в неделю), достаточно высокой общей математической подготовкой слушателей, различной узкой математической специализацией слушателей, тем, что для большинства специализаций лекции сопровождаются практическими занятиями, на которых основное внимание уделяется чисто алгоритмическим вопросам. Все это предъявляет к курсу лекций, по мнению автора, следующие требования: курс должен быть посвящен основным идеям построения и исследования приближенных методов, уровень строгости изложения должен быть таким же, как в других математических курсах, курс не должен быть перегружен деталями применения тех или иных методов к решению тех или иных классов задач, желательно, чтобы курс был не набором отдельных алгоритмов и фактов, а представлял собой нечто единое целое. Все эти требования автор и старался удовлетворить в своей книге. Неизбежной основой изложения становится при этом функциональный анализ. Вместе с тем автор стремился к тому, чтобы применение общих идей к конкретным задачам и методам было проиллюстрировано на достаточно содержательном уровне. Курс лекций такой направленности создавался на математико-механическом факультете Ленинградского университета на протяжении ряда лет предшественниками автора; наиболее важная роль в его формировании принадлежит М. К. Гавурину [4].

Термин "функциональное уравнение", использованный в заглавии этой книги, двусмыслен. С одной стороны, так называют те уравнения, в которых искомой является функция, например, интегральные или дифференциальные уравнения. С другой стороны, так часто называют уравнения в абстрактных

(линейных нормированных или, более обще, метрических) пространствах; слово "функциональное" в этой связи может иметь непосредственное отношение к функциональному анализу. В книге будут изучаться приближенные методы решения интегральных и дифференциальных уравнений, причем эти уравнения будут рассматриваться как частные случаи уравнений в нормированных пространствах. Если говорить более подробно, то в первой главе книги будет указана некоторая общая схема исследования приближенных методов решения абстрактных уравнений, которая будет дальше использоваться применительно к интегральным и дифференциальным уравнениям.

Приближенные методы решения уравнений предлагались чаще всего не "чистыми" математиками, а механиками, физиками, которые были в первую очередь заинтересованы в доведении результатов до числа. Проходил обычно довольно значительный срок, прежде чем предложенный метод сам становился объектом математического исследования. Так, предложенный в начале века метод Галеркина впервые был исследован лишь в 1942 г. М. В. Келдышем. Первоначально каждый алгоритм численного решения исследовался своими специфическими методами. Однако с накоплением материала в этой области появилась необходимость и одновременно возможность выработать общую точку зрения на приближенные методы. Эта работа была проделана Л. В. Канторовичем в 1948 г. в его знаменитой статье "Функциональный анализ и прикладная математика", где была построена общая теория приближенных методов решения уравнений. Естественной базой построения общей теории явился функциональный анализ, поскольку все многообразие уравнений, которые реально приходится решать, может быть объединено термином "функциональное уравнение" в его втором смысле. В настоящее время существуют различные общие схемы приближенных методов решения линейных уравнений, но идейный исток у них один — схема Л. В. Канторовича.

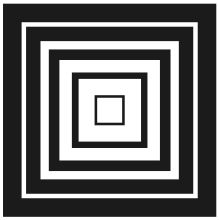
Та конкретная общая схема, которая используется в книге, принадлежит М. К. Гавурину. Вообще следует отметить идейную близость предлагаемой книги книге М. К. Гавурина "Лекции по методам вычислений" (1971 г.). Читатели, знакомые с этой книгой, заметят, что из нее заимствована не только общая схема, но и способ изложения некоторых других вопросов. Концепция устойчивости, которой придерживается автор, также принадлежит М. К. Гавурину. Эта концепция не является общепринятой. Причины, по которым автор отдал ей предпочтение, излагаются в конце § 4 главы 4, когда имеется уже достаточный материал для сопоставления разных понятий устойчивости.

Переходя к характеристике содержания книги, отметим только две особенности, которые являются следствием тех требований, предъявляемых к курсу,

которые были сформулированы ранее. Во-первых, относительно большое место в книге занимают приближенные методы решения интегральных уравнений. Это объясняется тем, что именно на уравнениях второго рода проще и ярче всего можно проиллюстрировать некоторые важные идеи теории приближенных методов. В частности, доказательство теоремы о сходимости метода механических квадратур позволяет познакомить читателя в простейшей форме с важным в современной вычислительной математике принципом компактной аппроксимации. Во-вторых, исследование тех или иных приближенных методов проводится обычно лишь для простейших модельных задач.

Из сказанного следует, что основное назначение книги — быть учебным пособием для студентов-математиков. Вместе с тем автор надеется, что она может быть полезна и другим читателям. Предполагается, что читатель знаком с основными простейшими понятиями и фактами функционального анализа и математической физики.

Глава 1



Общая схема приближенных методов

§ 1. Близкие уравнения

Приведем некоторые основные понятия, которые будут использоваться на протяжении всей книги.

Пусть U и F — два линейных нормированных пространства. Оператор A, заданный на линейном множестве $D(A) \subset U$, со значениями в F будем назвать *линейным*, если он однороден и аддитивен. Область задания $D(A)$ оператора A обычно будем считать плотной в U. Таким образом, ограниченность (или, что то же самое, непрерывность) оператора не входит в понятие линейности. Однако в тех случаях, когда линейный оператор A еще и ограничен, обычно будем считать, что область его задания — все пространство: $D(A) = U$. Символом $L(U, F)$ будем обозначать множество всех линейных ограниченных операторов, действующих из U в F. Как известно, $L(U, F)$ само является линейным нормированным пространством. Наконец, через $R(A)$ будем обозначать множество значений оператора A:

$$R(A) = \{f \mid f \in F, \quad f = Au, \quad u \in D(A)\}.$$

Будем говорить, что линейный оператор A *обратим*, если однородное уравнение $Au = 0$ имеет только нулевое решение. В этом случае оператор A устанавливает между $D(A)$ и $R(A)$ взаимнооднозначное соответствие, и на множестве $R(A)$ определен оператор A^{-1}, называемый обратным по отношению к A, такой, что для любого $u \in D(A)$ будет $A^{-1}Au = u$ и для любого $f \in R(A)$ будет $AA^{-1}f = f$. Оператор A^{-1} также линеен (т. е. аддитивен и однороден), причем

$$D(A^{-1}) = R(A), \quad R(A^{-1}) = D(A).$$

Задача, рассматриваемая в этом параграфе, такова. Пусть дано линейное уравнение

$$Au = f \tag{1}$$

и другое линейное уравнение

$$(A + \Delta A)u = f + \Delta f, \tag{2}$$

в котором линейный оператор ΔA и элемент $\Delta f \in F$ в каком-то смысле малы. Пусть u^* — решение уравнения (1), а $u^* + \Delta u$ — решение уравнения (2). Требуется оценить разность Δu между этими решениями.

Прежде чем перейти к решению этой задачи, отметим, что она бывает интересна в разных отношениях, и ее решение находит применение в различных ситуациях. Перечислим некоторые из них.

1. Довольно частым является тот случай, когда само уравнение, которое требуется решить, известно нам неточно. Например, неточно (или даже таблично) заданы коэффициенты дифференциального уравнения и его правая часть. Вопрос о том, к каким ошибкам в решении приведет неточность в задании уравнения, приводит нас к поставленной задаче.

2. Распространенным приближенным методом является метод замены уравнения на более простое, "близкое". Так, интегральные уравнения часто решают методом замены ядра на вырожденное (этот метод будет подробно изучаться в *главе 2*). Ясно, что оценка погрешности метода замены уравнения на близкое сводится к рассматриваемой задаче.

3. Как будет видно из дальнейшего, большинство методов приближенного решения интегральных и дифференциальных уравнений приводят к системам линейных алгебраических уравнений. Матрица коэффициентов и правые части этой системы строятся по некоторым правилам, исходя из заданного уравнения. Вычисление коэффициентов и правых частей, как правило, не может быть произведено вполне точно хотя бы из-за неизбежных ошибок округления. Вопрос о том, насколько исказится приближенное решение из-за неточного построения системы линейных уравнений, также приводит нас к поставленной задаче.

4. Пусть нам известен некоторый элемент $\overline{u} \in D(A)$, который из каких-либо соображений нам угодно принять за приближенное решение уравнения (1). Для оценки качества этого приближенного решения естественно подставить его в уравнение. Разность между результатом подстановки в левую часть и правой частью $A\overline{u} - f$ называется *невязкой* приближенного

решения $\overline{u}$. Само приближенное решение $\overline{u}$, как очевидно, является "точным" решением уравнения $Au = f + \Delta f$, где $\Delta f = A\overline{u} - f$ есть его невязка. Таким образом, оценка погрешности приближенного решения через его невязку также сводится к поставленной задаче. Некоторая специфика здесь заключается в том, что $\Delta A = 0$. Как видно из сказанного, ошибку, вызванную неточным построением решения, можно трактовать как следствие неточного задания правой части. Это позволяет, в частности, вопрос о том, с какой точностью следует решать упомянутые в пункте 3 системы уравнений, заменять вопросом о допустимой невязке полученных приближенных решений этих систем.

Поставленная задача решается следующей теоремой.

Теорема 1

Пусть выполнены условия: 1) пространство U полно; 2) $R(A) = F$, оператор A обратим и обратный оператор A^{-1} ограничен[1]; 3) оператор ΔA ограничен и $\left\|A^{-1}\Delta A\right\| \le \rho < 1$.

Тогда:

а) $R(A + \Delta A) = F$, оператор $A + \Delta A$ обратим, обратный оператор $(A + \Delta A)^{-1}$ ограничен и

$$\left\|(A + \Delta A)^{-1}\right\| \le \frac{\left\|A^{-1}\right\|}{1 - \rho}; \tag{3}$$

б) для разности решений уравнений (2) и (1) выполняется оценка

$$\|\Delta u\| \le \frac{1}{1-\rho}\left(\rho\left\|u^*\right\| + \left\|A^{-1}\right\| \cdot \|\Delta f\|\right). \tag{4}$$

Доказательство

а) Пусть для некоторого элемента $f \in F$ нашлось решение u_0 уравнения $(A + \Delta A)u = f$, так что $(A + \Delta A)u_0 = f$. Применяя к обеим частям последнего тождества оператор A^{-1}, получаем $(I + A^{-1}\Delta A)u_0 = A^{-1}f$. Так как норма

[1] Тем самым уравнение (1) однозначно разрешимо при любой правой части f.

оператора $A^{-1}\Delta A$ меньше единицы, то по теореме Банаха существует обратный оператор $(I + A^{-1}\Delta A)^{-1}$, причем

$$\left\| (I + A^{-1}\Delta A)^{-1} \right\| \le \frac{1}{1 - \left\| A^{-1}\Delta A \right\|} \le \frac{1}{1 - \rho}.$$

Поэтому $u_0 = (I + A^{-1}\Delta A)^{-1} A^{-1} f$. В частности, если $f = 0$, то непременно и $u_0 = 0$, и потому оператор $A + \Delta A$ обратим. Пусть теперь f — произвольный элемент пространства F. Положим $u_0 = (I + A^{-1}\Delta A)^{-1} A^{-1} f$. Для u_0 справедливо равенство $u_0 = A^{-1} f - A^{-1}\Delta A u_0$, из которого видно, что $u_0 \in D(A) = D(A + \Delta A)$. Применяя к элементу u_0 оператор $A + \Delta A = A(I + A^{-1}\Delta A)$, видим, что $(A + \Delta A)u_0 = f$. Итак, при любом $f \in F$ уравнение $(A + \Delta A)u = f$ имеет решение $u_0 = (I + A^{-1}\Delta A)^{-1} A^{-1} f$. Тем самым показано, что $R(A + \Delta A) = F$ и $(A + \Delta A)^{-1} = (I + A^{-1}\Delta A)^{-1} A^{-1}$. Остается отметить, что из последнего равенства вытекают ограниченность оператора $(A + \Delta A)^{-1}$ и оценка (3). Утверждение а) доказано.

б) В силу наложенных условий и уже доказанного утверждения а) уравнения (1) и (2) однозначно разрешимы. Также однозначно разрешимо и уравнение $Au = f + \Delta f$; его решение обозначим через $u^* + \Delta u_1$. Итак, мы имеем тождества

$$Au^* = f, \tag{5}$$

$$(Au^* + \Delta u_1) = f + \Delta f, \tag{6}$$

$$(A + \Delta A)(u^* + \Delta u) = f + \Delta f. \tag{7}$$

Вычитая из (6) тождество (5), получаем $A\Delta u_1 = \Delta f$, откуда $\Delta u_1 = A^{-1}\Delta f$ и

$$\left\| \Delta u_1 \right\| \le \left\| A^{-1} \right\| \cdot \left\| \Delta f \right\|. \tag{8}$$

Вычитая из тождества (7) тождество (6), находим

$$(A + \Delta A)(\Delta u - \Delta u_1) = -\Delta A(u^* + \Delta u_1),$$

откуда (вспоминая, что $(A+\Delta A)^{-1} = (I + A^{-1}\Delta A)^{-1} A^{-1}$),

$$\begin{aligned} \|\Delta u - \Delta u_1\| &\le \left\|(I + A^{-1}\Delta A)^{-1}\right\| \cdot \left\|A^{-1}\Delta A\right\| \cdot \left(\left\|u^*\right\| + \|\Delta u_1\|\right) \le \\ &\le \frac{\rho}{1-\rho}\left(\left\|u^*\right\| + \|\Delta u_1\|\right) \le \frac{\rho}{1-\rho}\left(\left\|u^*\right\| + \left\|A^{-1}\right\| \cdot \|\Delta f\|\right) \end{aligned} \tag{9}$$

Из оценок (8) и (9) сразу же следует (4). ■

❑ **Замечание 1.** Поскольку $\left\|u^*\right\| = \left\|A^{-1}f\right\| \le \left\|A^{-1}\right\| \cdot \|f\|$, то из неравенства (4) вытекает оценка

$$\|\Delta u\| \le \frac{\left\|A^{-1}\right\|}{1-\rho}\left(\rho\|f\| + \|\Delta f\|\right). \tag{10}$$

❑ **Замечание 2.** Уравнения (1) и (2) можно считать близкими, если малы величины $\left\|A^{-1}\Delta A\right\|$ и $\|\Delta f\|$. В этом случае правые части оценок (4) и (10) будут, вообще говоря, также малы.

❑ **Замечание 3.** В приложениях теоремы 1 обычно одно из уравнений (1), (2) считается "точным" — тем, решение которого нас в действительности интересует, а другое — "приближенным", решение которого нам известно или мы собираемся его реально строить. В оценки (4) и (10) эти уравнения входят неравноправно — вычисление их правых частей требует в основном информации об уравнении (1). Поэтому какое из этих уравнений считать "точным", а какое "приближенным", зависит от той задачи, которую мы решаем при помощи теоремы 1. Например, если по заданному "точному" уравнению мы должны выбрать "приближенное", которое дало бы нам решение с заданной точностью, уравнением (1) удобно считать "точное". Если же у нас уже имеются оба уравнения — "точное" и "приближенное", то обычно о "приближенном" уравнении мы располагаем большей информацией (оно уже решено или будет реально решаться, и значительную информацию мы можем получить в процессе решения), и потому именно "приближенное" уравнение выгодно считать уравнением (1), а "точное" — уравнением (2). Вообще, видимо, первый способ ((1) — "точное", (2) — "приближенное" уравнение) более удобен в теоретических вопросах, например, при исследовании сходимости решений последова-

тельности уравнений к решению предельного. Противоположный же способ рассмотрения более удобен, когда требуется вычислить реальную оценку.

В теореме 1 речь шла об оценке абсолютной погрешности при замене решения уравнения (1) решением уравнения (2) (или наоборот). Иногда больший интерес представляет оценка относительной погрешности $\|\Delta u\| / \|u^*\|$. Задача об оценке относительной погрешности будет решаться в предположении, что все операторы A, A^{-1}, ΔA ограничены.

Определение

Пусть линейный оператор A ограничен вместе с обратным A^{-1}. *Числом обусловленности* оператора A называется величина

$$\mu(A) = \|A\| \cdot \|A^{-1}\|.$$

В случае, если оператор A не имеет обратного (или A^{-1} не ограничен), иногда считают $\mu(A) = +\infty$. Очевидно, что для любого обратимого оператора A $\mu(A) \geq 1$.

Неформальный смысл числа обусловленности выявляется следующей теоремой.

Введем обозначения:

$$\delta A = \frac{\|\Delta A\|}{\|A\|}, \quad \delta f = \frac{\|\Delta f\|}{\|f\|}, \quad \delta u = \frac{\|\Delta u\|}{\|u^*\|}$$

— относительные погрешности в операторе A, правой части f и относительная погрешность в решении уравнения (1), вызванная погрешностями в A и f.

Теорема 2

Пусть пространство U — полное и пусть A, A^{-1}, ΔA — ограниченные линейные операторы, причем $\|A^{-1}\| \cdot \|\Delta A\| < 1$. Пусть u^* и $u^* + \Delta u$ — решения уравнений (1) и (2) соответственно. Тогда

$$\delta u \leq \frac{\mu(A)}{1 - \mu(A) \cdot \delta A} (\delta A + \delta f). \tag{11}$$

Доказательство

Отметим прежде всего, что

$$\left\|A^{-1}\Delta A\right\| \le \left\|A^{-1}\right\| \cdot \left\|\Delta A\right\| = \mu(A) \cdot \delta A = \rho < 1,$$

и поэтому выполнены условия теоремы 1. Значит, справедливо неравенство (4). Но

$$\left\|\Delta f\right\| = \left\|f\right\| \cdot \delta f \le \left\|A\right\| \cdot \left\|u^*\right\| \cdot \delta f.$$

Поэтому

$$\left\|A^{-1}\right\| \cdot \left\|\Delta f\right\| \le \mu(A)\delta f \cdot \left\|u^*\right\|.$$

Подставив эту оценку в правую часть неравенства (4) и поделив на $\left\|u^*\right\|$, получим

$$\delta u \le \frac{1}{1-\rho}\left(\rho + \mu(A)\delta f\right).$$

Остается подставить сюда вместо ρ его значение и вынести $\mu(A)$ за скобку. ■

В предположении, что относительная погрешность δA настолько мала, что величиной $\mu(A)\delta A$ можно пренебречь по сравнению с единицей, качественно оценка (11) выглядит так: относительная погрешность решения оценивается величиной, пропорциональной сумме относительных погрешностей в операторе и в правой части, причем коэффициент пропорциональности — число обусловленности оператора A.

Разумеется, правая часть оценки (11) тем меньше, чем меньше число обусловленности $\mu(A)$. Оператор A и уравнение (1) принято называть *хорошо обусловленными*, если число обусловленности $\mu(A)$ невелико, и *плохо обусловленными*, если $\mu(A)$ большое.

Можно показать, что оценка (11) является точной в том отношении, что для любого обратимого оператора A найдутся такая правая часть f и такие отличные от нулевых погрешности ΔA $\left(\left\|A^{-1}\right\| \cdot \left\|\Delta A\right\| < 1\right)$ и Δf, что отношение правой и левой части оценки (11) сколь угодно близко к единице.

§ 2. Общая схема приближенных методов. Сходимость

Пусть требуется решить уравнение

$$Au = f \tag{1}$$

с линейным оператором A, действующим из нормированного пространства U в пространство F. Приближенные методы решения уравнения (1) обычно требуют решения некоторого другого, более простого линейного уравнения, чаще всего системы линейных алгебраических уравнений. Это обстоятельство и будет положено в основу излагаемой ниже общей схемы приближенных методов[1].

Итак, пусть наряду с уравнением (1) задано другое уравнение (или последовательность уравнений)

$$A_n u_n = f_n \tag{2}$$

с линейным оператором A_n, действующим из пространства U_n в пространство F_n. Нормированные пространства U_n и F_n всегда будем считать полными, а оператор A_n всегда будем предполагать ограниченным. Уравнения (1) и (2) считаются некоторым образом связанными между собой. Будем предполагать, что задан линейный ограниченный оператор ψ_n, действующий из пространства F в пространство F_n, и что правая часть f_n уравнения (2) всегда есть результат применения этого оператора к f: $f_n = \psi_n f$. Решение u^* уравнения (1) и решение u_n^* уравнения (2) лежат, вообще говоря, в разных пространствах; поэтому не всегда u_n^* можно рассматривать как приближенное решение уравнения (1). Будем называть u_n^* *каркасом приближенного решения* и считать, что само приближенное решение $u^{(n)}$ восстанавливается по каркасу с помощью линейного ограниченного оператора $\overline{\varphi}_n$, действующего из U_n в U: $u^{(n)} = \overline{\varphi}_n u_n^*$. Наконец, предположим, что задан еще один линейный ограниченный оператор φ_n, действующий из пространства U в про-

[1] Эта основная идея общей теории приближенных методов принадлежит Л. В. Канторовичу, а та конкретная схема, о которой идет речь далее, — М. К. Гавурину.

странство U_n. Роль этого оператора будет выяснена позже. Пока заметим только, что он не связан с алгоритмом метода; $\varphi_n u$ есть "представитель" элемента u в пространстве U_n.

Все введенные пространства и операторы можно изобразить на схеме (рис. 1).

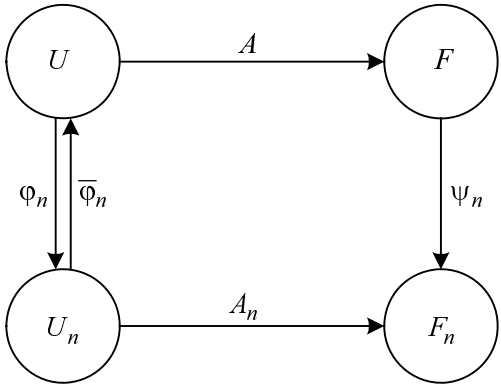


Рис. 1

Перейдем к вопросу о сходимости приближенных решений $u^{(n)}$ к точному u^*. При этом, конечно, считается, что (2) — последовательность уравнений.

Определение 1

Мерой аппроксимации уравнения (1) *уравнением* (2) *на элементе* $u \in D(A)$ называется число

$$\gamma_n(u) = \|A_n \varphi_n u - \psi_n A u\|.$$

Говорят, что последовательность операторов A_n аппроксимирует оператор A на элементе u, если $\gamma_n(u) \xrightarrow[n\to\infty]{} 0$.

Пусть δ_n — некоторая бесконечно малая величина. Условимся говорить, что уравнения (2) аппроксимируют на элементе u уравнение (1) с порядком δ_n, если $\gamma_n(u) = O(\delta_n)$.

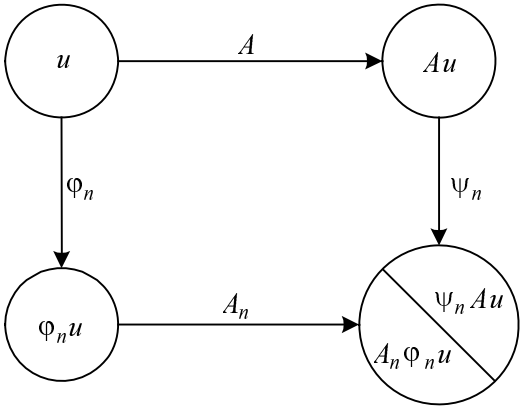


Рис. 2

Формула, определяющая меру аппроксимации, легко запоминается с помощью схемы (рис. 2), которую следует сопоставить со схемой, изображенной на рис. 1: мера аппроксимации есть норма разности двух элементов, указанных в правом нижнем кружочке. Поясним смысл меры аппроксимации еще следующим образом. Каждый элемент $\overline{u} \in D(A)$ есть решение уравнения $Au = \bar{f}$, где $\bar{f} = A\overline{u}$. Соответствующее этому уравнению "приближенное" уравнение есть $A_n u_n = \bar{f}_n$, где $\bar{f}_n = \psi_n \bar{f} = \psi_n A\overline{u}$. Поэтому $\gamma_n(u) = \left\| A_n \varphi_n \overline{u} - \bar{f}_n \right\|$ — норма невязки, которая получается при подстановке в "приближенное" уравнение "представителя" точного решения.

Определение 2

Будем говорить, что *каркасы приближенных решений сходятся*, если выполняется соотношение $\left\| \varphi_n u^* - u_n^* \right\| \to 0$.

Теорема 1
(о сходимости каркасов приближенных решений)

Пусть существует решение u^* уравнения (1) и пусть при некотором n существует ограниченный обратный оператор A_n^{-1}. Тогда

$$\left\| \varphi_n u^* - u_n^* \right\| \le \left\| A_n^{-1} \right\| \gamma_n(u^*), \tag{3}$$

и если ограниченные операторы A_n^{-1} существуют при всех достаточно больших n и

$$\left\|A_n^{-1}\right\| \cdot \gamma_n(u^*) \to 0, \tag{4}$$

то каркасы приближенных решений сходятся.

Доказательство

Сходимость каркасов при условии (4) сразу же следует из (3), поэтому только это неравенство и требуется доказать. Поскольку

$$u_n^* = A_n^{-1} f_n = A_n^{-1} \psi_n f \quad \text{и} \quad f = Au^*,$$

то

$$\varphi_n u^* - u_n^* = A_n^{-1} A_n \varphi_n u^* - A_n^{-1} \psi_n f = A_n^{-1}(A_n \varphi_n u^* - \psi_n A u^*),$$

и потому

$$\left\| \varphi_n u^* - u_n^* \right\| \le \left\|A_n^{-1}\right\| \cdot \left\| A_n \varphi_n u^* - \psi_n A u^* \right\| = \left\|A_n^{-1}\right\| \cdot \gamma_n(u^*). \blacksquare$$

Следствие. Если операторы A_n аппроксимируют оператор A на решении u^* и если при достаточно больших n существуют непрерывные обратные операторы A_n^{-1} и они равномерно ограничены $\left(\left\|A_n^{-1}\right\| \le C\right)$, то каркасы приближенных решений сходятся.

- **Замечание 1.** Полезно иметь в виду, что мера аппроксимации на точном решении $\gamma_n(u^*)$ есть, как об этом уже говорилось ранее, $\left\|A_n \varphi_n u^* - f_n\right\|$. Оценка (3) в связи с этим приобретает смысл оценки погрешности "приближенного решения" $\varphi_n u^*$ уравнения (2) через норму его невязки и может рассматриваться как частный случай оценки, указанной в *теореме 1 из § 1*.

- **Замечание 2.** Неформальный смысл сходимости каркасов определяется трактовкой решения приближенного уравнения в конкретных методах. Поясним это примером. Обычно искомый элемент $u^* \in U$ есть функция, заданная в некоторой области. Часто (так это будет, например, в методе сеток) каркас приближенного решения можно трактовать как таблицу

приближенного решения. Если определить оператор φ_n таким образом, чтобы функции u он ставил в соответствие таблицу значений этой функции в тех же узлах, то сходимость каркасов будет означать близость (при достаточно больших n) таблиц точного и приближенного решений.

Обратимся теперь непосредственно к вопросу о сходимости приближенных решений $u^{(n)}$.

Теорема 2
(о сходимости приближенных решений)

Пусть существует точное решение u^*. Пусть при некотором n существует ограниченный обратный оператор A_n^{-1}. Тогда

$$\left\| u^{(n)} - u^* \right\| \le \left\| \overline{\varphi}_n \right\| \cdot \left\| A_n^{-1} \right\| \cdot \gamma_n(u^*) + \left\| \overline{\varphi}_n \varphi_n u^* - u^* \right\|. \qquad (5)$$

Если ограниченные обратные операторы A_n^{-1} существуют при всех достаточно больших n и выполняются соотношения

$$\left\| \overline{\varphi}_n \right\| \cdot \left\| A_n^{-1} \right\| \cdot \gamma_n(u^*) \to 0, \qquad \overline{\varphi}_n \varphi_n u^* \to u^*,$$

то $u^{(n)} = \overline{\varphi}_n u_n{}^* \to u^*$.

Доказательство

Доказательство этой теоремы сразу же сводится к доказательству неравенства (5). Очевидно, что

$$u^{(n)} - u^* = \overline{\varphi}_n u_n^* - u^* = \overline{\varphi}_n (u_n^* - \varphi_n u^*) + \overline{\varphi}_n \varphi_n u^* - u^*.$$

Отсюда

$$\left\| u^{(n)} - u^* \right\| \le \left\| \overline{\varphi}_n \right\| \cdot \left\| u_n^* - \varphi_n u^* \right\| + \left\| \overline{\varphi}_n \varphi_n u^* - u^* \right\|.$$

Для завершения доказательства остается воспользоваться для оценки величины $\left\| u_n^* - \varphi_n u^* \right\|$ неравенством (3). ■

§ 3. Устойчивость

При решении "точного" уравнения

$$Au = f \tag{1}$$

построение "приближенных" уравнений

$$A_n u_n = f_n \quad (f_n = \varphi_n f), \tag{2}$$

предписываемое алгоритмом решения задачи, обычно не удается осуществить точно. Поэтому реально вместо этих уравнений мы строим и решаем некоторые другие:

$$(A_n + \Delta A_n)u_n = f_n + \Delta f_n. \tag{3}$$

Возникает вопрос, насколько сильные искажения в каркасах приближенных решений u_n^* и в самих приближенных решениях $u^{(n)}$ вызовут ошибки, допущенные при составлении приближенных уравнений. Кроме того, сами приближенные уравнения (3) также обычно не удается решить точно. Вопрос о том, какая невязка приближенного решения является допустимой, как это было показано в *§ 1*, также сводится к вопросу о допустимых возмущениях Δf_n правых частей. Поставленная задача о влиянии допущенных погрешностей на приближенное решение в действительности чрезвычайно существенна. Известны приближенные методы, которые теоретически обладают очень быстрой сходимостью, но практически непригодны, поскольку малейшие искажения в той системе линейных алгебраических уравнений, которую попутно приходится решать, приводят к большим ошибкам в приближенном решении[1].

При каждом фиксированном n сопоставление решений уравнений (2) и (3) осуществляется теоремами, доказанными в *§ 1*. Нас будет интересовать сейчас качественная характеристика метода, сводящего уравнение (1) к последовательности уравнений (2). При этом решения уравнений (3) будем обозначать через $u_n^* + \Delta u_n$, где u_n^* — решения уравнений (2).

[1] Например, метод Ритца с неудачно выбранной координатной системой.

Определение 1

Пусть при достаточно больших n операторы A_n ограничены вместе с обратными A_n^{-1}. Процесс отыскания каркасов приближенных решений называется *устойчивым*, если, каковы бы ни были возмущения ΔA_n операторов A_n и Δf_n правых частей f_n, соотношения

$$\delta A_n \xrightarrow[n\to 0]{} 0 \quad \text{и} \quad \delta f_n \xrightarrow[n\to 0]{} 0$$

влекут за собой существование при достаточно больших n обратных операторов $(A_n + \Delta A_n)^{-1}$ и соотношение $\delta u_n \xrightarrow[n\to 0]{} 0$.

Другими словами, устойчивость процесса отыскания каркасов означает, что если приближенные уравнения мы будем строить с относительными погрешностями в операторах и в правых частях, стремящимися к нулю, то и относительные погрешности в каркасах также будут стремиться к нулю.

Теореме о признаке устойчивости процесса отыскания каркасов предпошлем лемму.

Лемма

Пусть $A \in L(U, F)$ — линейный ограниченный оператор, имеющий ограниченный же обратный $A^{-1} \in L(F, U)$. Тогда существует такой линейный ограниченный оператор $B \in L(U, F)$, что $\|B\| \cdot \|A^{-1}\| < 2$ и оператор $A + B$ не имеет обратного.

Доказательство

Выберем элемент $\bar{f} \in F$ так, чтобы для элемента $\bar{u} = A^{-1}\bar{f}$ выполнялась оценка $\|\bar{u}\| = \|A^{-1}\bar{f}\| > \frac{1}{2}\|A^{-1}\| \cdot \|\bar{f}\|$. Выберем функционал $\Psi \in U^*$ из условий: $\|\Psi\| = 1$, $\Psi(\bar{u}) = \|\bar{u}\|$. Положим теперь $Bu = -(\Psi(u)/\|\bar{u}\|)\bar{f}$. Тогда

$$B\bar{u} = -(\Psi(u)/\|\bar{u}\|)\bar{f} = -\bar{f}, \quad (A + B)\bar{u} = \bar{f} - \bar{f} = 0,$$

так что оператор $A + B$ необратим, и в то же время $\|B\| = \|\bar{f}\|/\|\bar{u}\| < 2/\|A^{-1}\|$ и $\|A^{-1}\| \cdot \|B\| < 2$. ■

Теорема 1
(об устойчивости процесса отыскания каркасов)

Пусть при всех достаточно больших n существуют непрерывные обратные операторы A_n^{-1}. Тогда для того чтобы процесс отыскания каркасов приближенных решений был устойчив, необходимо и достаточно, чтобы числа обусловленности операторов A_n были равномерно ограничены: $\mu(A_n) \le C$.

Доказательство

1. *Достаточность.*

Пусть $\mu(A_n) \le C$ и возмущения ΔA_n и Δf_n таковы, что $\delta A_n \to 0$ и $\delta f_n \to 0$. Тогда

$$\left\| A_n^{-1} \right\| \cdot \left\| \Delta A_n \right\| = \mu(A_n)\delta A_n \to 0,$$

и поэтому по *теореме 1 из § 1* при достаточно больших n существуют операторы $(A_n + \Delta A_n)^{-1}$. Воспользовавшись оценкой из *теоремы 2 (§ 1)*

$$\delta u_n \le \frac{\mu(A_n)}{1 - \mu(A_n) \cdot \delta A_n} \left(\delta A_n + \delta f_n \right),$$

видим, что при наших условиях $\delta u_n \to 0$. Таким образом, процесс отыскания каркасов устойчив. Достаточность доказана.

2. *Необходимость.*

Необходимость будем доказывать от противного. Допустим, что последовательность $\mu(A_n)$ неограничена. Не умаляя общности, можно считать, что $\mu(A_n) \to \infty$ (в противном случае мы перешли бы к подпоследовательности индексов n_m, для которой $\mu(A_{n_m}) \to \infty$). Применяя лемму, построим последовательность операторов $B_n = \Delta A_n$, так что операторы $A_n + \Delta A_n$ не имеют обратных, и в то же время $\left\| \Delta A_n \right\| \cdot \left\| A_n^{-1} \right\| < 2$. Это последнее неравенство означает, что $\left\| \Delta A_n \right\| \mu(A_n) < 2 \left\| A_n \right\|$ или $\delta A_n < 2 / \mu(A_n) \to 0$. Таким образом, нашлись такие возмущающие операторы ΔA_n, что, несмотря на соотношение $\delta A_n \to 0$, обратные операторы $(A_n + \Delta A_n)^{-1}$ не существуют, так что процесс отыскания каркасов неустойчив. ■

- **Замечание 1.** Как видно из доказанной теоремы, будет ли процесс отыскания каркасов устойчив, не зависит от правой части уравнения.
- **Замечание 2.** Как видно из доказательства теоремы, ограниченность чисел обусловленности $\mu(A_n)$ необходима уже только для выполнения первого из условий устойчивости — существования операторов $(A_n + \Delta A_n)^{-1}$ при всех достаточно больших n. Если последовательность операторов A_n такова, что для любой возмущающей последовательности ΔA_n из соотношения $\delta A_n \to 0$ следует существование (при достаточно больших n) операторов $(A_n + \Delta A_n)^{-1}$, то числа обусловленности $\mu(A_n)$ операторов A_n равномерно ограничены и процесс отыскания каркасов устойчив.
- **Замечание 3.** В реально применяемых методах условие ограниченности чисел обусловленности $\mu(A_n)$ не всегда бывает выполнено. Из доказательства достаточности в теореме 1 легко усмотреть, что существование операторов $(A_n + \Delta A_n)^{-1}$ и соотношение $\delta u_n \to 0$ являются следствиями условий

$$\mu(A_n)\delta A_n \to 0, \quad \mu(A_n)\delta f_n \to 0. \tag{4}$$

Поэтому применение методов с неограниченными числами обусловленности $\mu(A_n)$ возможно, если заботиться о достаточно быстром убывании ошибок ΔA_n и Δf_n так, чтобы выполнялись соотношения (4). Выполнение последних реально осуществимо, если числа обусловленности $\mu(A_n)$ растут не слишком быстро. В связи с этим принято различать степенную и показательную неустойчивость. Первая означает, что для $\mu(A_n)$ выполняется оценка вида $\mu(A_n) \le Cn^{\alpha}$ $(\alpha > 0)$. Это — медленный рост чисел обусловленности. Счет по таким методам обычно возможен, хотя и требует ведения вычислений с дополнительными знаками. Показательная неустойчивость означает, что $\mu(A_n) \ge Cq^n$ $(q > 1)$. Это очень быстрый рост $\mu(A_n)$. Удовлетворить при этом соотношениям (4) обычно не удается, и такие методы для счета практически непригодны. Стоит обратить внимание на то, что при определении степенной и показательной неустойчивости знак неравенства направлен в разные стороны. Этому соответствует разная эмоциональная окраска терминов. Степенная неустойчивость — "хорошо", "почти хорошо", "допустимо", а показательная — "плохо", "недопустимо".

Перейдем к изучению тех ошибок, которые возникают в приближенном решении $u^{(n)}$ из-за неточного проведения алгоритма. Источников ошибок здесь два. Во-первых, это неточное построение каркаса приближенного решения — $u_n^* + \Delta u^n$ вместо u_n^* в результате замены уравнения (2) уравнением (3). Во-вторых, возможны ошибки при применении к полученному каркасу оператора $\overline{\varphi}_n$. Последние мы будем рассматривать как результат возмущения оператора $\overline{\varphi}_n$ некоторым линейным оператором $\Delta\overline{\varphi}_n \in L(U_n, U)$. Таким образом, вместо приближенного решения $u^{(n)} = \overline{\varphi}_n u_n^*$ мы получаем $u^{(n)} + \Delta u^{(n)} = (\overline{\varphi}_n + \Delta\overline{\varphi}_n)(u_n^* + \Delta u_n)$.

Определение 2

Процесс построения приближенных решений называется устойчивым, если:

1) из соотношения $\delta A_n \to 0$ следует существование при всех достаточно больших n обратных операторов $(A_n + \Delta A_n)^{-1}$;

2) из соотношений

$$\delta A_n \to 0, \quad \delta f_n \to 0, \quad \delta\overline{\varphi}_n = \frac{\left\|\Delta\overline{\varphi}_n\right\|}{\left\|\overline{\varphi}_n\right\|} \to 0 \tag{5}$$

вытекает соотношение $\delta u^{(n)} = \dfrac{\left\|\Delta u^{(n)}\right\|}{\left\|u^{(n)}\right\|} \to 0.$

Теорему об устойчивости процесса построения приближенных решений будем доказывать для сходящихся методов. При этом естественно считать, что само точное решение u^* — ненулевое.

Теорема 2
(об устойчивости процесса построения приближенных решений)

Пусть выполнены условия: 1) существует решение u^* уравнения (1), причем $f \neq 0$; 2) при достаточно больших n существуют непрерывные обратные операторы A_n^{-1}; 3) приближенные решения $u^{(n)}$ сходятся к точному: $u^{(n)} \to u^*$.

Тогда для устойчивости процесса построения приближенных решений необходимо и достаточно выполнение двух условий: а) процесс отыскания каркасов приближенных решений устойчив; б) числа $\|\overline{\varphi}_n\|\cdot\|u*_n^*\|$ равномерно ограничены: $\|\overline{\varphi}_n\|\cdot\|u_n^*\|\le C$.

Доказательство

1. *Достаточность.*

Пусть выполнены условия а) и б). Требуется проверить условия 1) и 2) определения 2. Первое из них есть следствие устойчивости процесса отыскания каркасов. При проверке второго достаточно показать, что соотношения (5) влекут сходимость к нулю самих погрешностей $\Delta u^{(n)}$, поскольку, как было предположено в условиях теоремы, приближенные решения сходятся к ненулевому элементу u^*, и потому их нормы ограничены снизу положительным числом. Ясно, что

$$\Delta u^{(n)} = (\overline{\varphi}_n + \Delta\overline{\varphi}_n)\Delta u_n + \Delta\overline{\varphi}_n u_n^*.$$

Отсюда

$$\|\Delta u^{(n)}\| \le \left(\|\overline{\varphi}_n\| + \|\Delta\overline{\varphi}_n\|\right)\|\Delta u_n\| + \|\Delta\overline{\varphi}_n\|\cdot\|u_n^*\| = \{(1+\delta\varphi_n)\delta u_n + \delta\varphi_n\}\|\overline{\varphi}_n\|\cdot\|u_n^*\|.$$

Ввиду устойчивости процесса отыскания каркасов и соотношений (5)

$$\delta u_n \to 0, \quad \delta\overline{\varphi}_n \to 0.$$

Кроме того, $\|\overline{\varphi}_n\|\cdot\|u_n^*\|\le C$. Поэтому из построенной оценки сразу же следует, что $\Delta u^{(n)} \to 0$. Достаточность доказана.

2. *Необходимость.*

Согласно замечанию 2 к теореме 1 уже первое требование определения 2 влечет ограниченность чисел обусловленности $\mu(A_n)$ операторов A_n, а значит, и устойчивость процесса отыскания каркасов. Осталось доказать необходимость условия б). Докажем это от противного. Пусть числа $\|\overline{\varphi}_n\|\cdot\|u_n^*\|$ не ограничены. Не умаляя общности, будем считать, что $\|\overline{\varphi}_n\|\cdot\|u_n^*\|\to\infty$. Построим такие возмущения ΔA_n, Δf_n и $\Delta\overline{\varphi}_n$, чтобы, несмотря на выполнение

соотношений (5), не выполнялось требование $\Delta u^{(n)} \to 0$. Именно положим $\Delta A_n = 0,\ \Delta f_n = 0$, так что $\Delta u_n = 0$ и $\Delta u^{(n)} = \Delta\overline{\varphi}_n u_n^*$. Оператор $\Delta\overline{\varphi}_n$ построим следующим образом. Положим $a_n = \sqrt{\|\overline{\varphi}_n\| \Big/ \|u_n^*\|}$ и выберем функционал $\Psi_n \in U_n^*$, так что $\|\Psi_n\| = 1$ и $\Psi_n(u_n^*) = \|u_n^*\|$, и произвольный элемент $\overline{u} \in U,\ \|\overline{u}\| = 1$. Теперь операторы $\Delta\overline{\varphi}_n \in L(U_n, U)$ определим формулой $\Delta\overline{\varphi}_n u_n = a_n \Psi_n(u_n)\overline{u}$.

Тогда

$$\|\Delta\overline{\varphi}_n\| = a_n, \qquad \delta\overline{\varphi}_n = 1 \Big/ \sqrt{\|\overline{\varphi}_n\| \cdot \|u_n^*\|} \to 0,$$

но

$$\|\Delta u^{(n)}\| = \|\Delta\overline{\varphi}_n u_n^*\| = a_n \|u_n^*\| = \sqrt{\|\overline{\varphi}_n\| \cdot \|u_n^*\|} \to \infty.$$

Этим необходимость условия б) доказана. ■

❑ **Замечание.** Будет ли процесс построения приближенных решений устойчив, зависит от правой части f, поскольку от нее зависит u_n^*.

По поводу доказанной теоремы можно сделать замечание, вполне аналогичное замечанию 3 к теореме 1. Если процесс построения приближенных решений неустойчив, но числа обусловленности $\mu(A_n)$ и последовательность $\|\overline{\varphi}_n\| \cdot \|u_n^*\|$ стремятся к бесконечности не слишком быстро, то за счет быстрого убывания относительных погрешностей δA_n, δf_n и $\delta\overline{\varphi}_n$ обычно удается достичь стремления к нулю погрешности $\Delta u^{(n)}$. Как и ранее, можно было бы ввести понятие степенной и показательной неустойчивости процесса построения приближенных решений, но останавливаться на этом мы не будем.

Понятие устойчивости, как оно дается определениями 1 и 2[1], не является общепринятым. Сопоставление этого определения с другим, более распространенным, мы приведем в *§ 4 главы 4*, где такое сравнение можно будет сделать на более содержательном уровне.

---

[1] Это определение принадлежит М. К. Гавурину.

В заключение главы еще раз обратимся к нашей общей схеме, основная идея которой — рассматривать приближенное решение, построенное по предписанному алгоритму, как точное решение некоторого другого уравнения. Схема эта является весьма гибкой и оставляет для того, кто исследует приближенный метод, достаточно широкий произвол. Реально задано уравнение (1) (точнее, видимо, говорить — класс уравнений; исследование того или иного приближенного метода проводится обычно не для одного отдельного уравнения, а для некоторого класса уравнений). Выбор той пары пространств (U, F), в которых рассматривается оператор A, обычно находится в распоряжении исследователя. Произвол в выборе U ограничен лишь тем, что те функционалы, значение которых на решении нас в действительности интересует, должны быть непрерывны в этом пространстве. Тогда из сходимости приближенных решений в пространстве U будет следовать и сходимость этих функционалов, вычисленных на приближенных решениях. Это требование, как правило, не бывает обременительным, и часто выбор пространств U и F осуществляется из тех соображений, чтобы по возможности облегчить исследование метода. Этим будет в основном диктоваться выбор пространств U и F ниже, при исследовании конкретных методов. Впрочем математический анализ располагает рядом теорем, которые позволяют утверждать, что если некоторая последовательность функций сходится в некотором функциональном пространстве достаточно быстро, то она будет сходиться и в другом функциональном пространстве с более сильной нормой. Поэтому и в тех случаях, когда нас интересует сходимость метода не в той "естественной" метрике, которая наиболее удобна для применения общей схемы, а в другой, более сильной, часто предпочтительным оказывается такой путь. Сначала сходимость исследуется на основе общей схемы все же в "естественной" метрике, а затем используется какая-либо из упомянутых выше теорем о последовательностях функций. Простейший пример такого рода соображений читатель найдет в *§ 7 главы 4*.

В распоряжении исследователя всегда находится выбор пространств U_n и F_n. Что касается самих уравнений (2), то произвол в их выборе связан с тем, что требования алгоритма, которые определяют приближенное решение, могут быть записаны в виде разных эквивалентных уравнений. При исследовании сходимости из множества таких уравнений естественно выбрать то, которое наиболее просто приводит к интересующему нас результату. Устойчивость же метода самым тесным образом связана с тем, которое из этих эквивалент-

ных уравнений мы выбрали для реального счета, так что вопросы устойчивости требуют обычно более детального описания алгоритма[1]. При этом конкретный вид уравнений (2) оказывается определенным. Тем самым определены операторы ψ_n и $\overline{\varphi}_n$ — первый из них по правой части f уравнения (1) строит правую часть f_n уравнения (2), а второй по каркасу восстанавливает само приближенное решение. Оператор же φ_n всегда находится в распоряжении исследователя.

Вернемся еще к вопросу о выборе норм в пространствах U_n и F_n. Если уравнения (2) однозначно разрешимы при всех достаточно больших n, то нетрудно выбрать эти нормы так, чтобы процесс отыскания каркасов формально оказался устойчивым. Для этого достаточно, например, произвольно выбрав норму в U_n, положить по определению $\|f_n\|_{F_n} = \left\|A_n^{-1} f_n\right\|_{U_n}$, или же, наоборот, положить $\|u_n\|_{U_n} = \|A_n u_n\|_{F_n}$, и тогда при всех n будет $\mu(A_n) = 1$. Если нас интересуют вопросы сходимости, то, как легко видеть, такой выбор норм просто означает перенесение трудностей из одного места в другое. Те трудности, которые могли бы возникнуть в оценке нормы $\left\|A_n^{-1}\right\|$ при другом выборе норм в пространствах U_n и F_n, переносятся в задачу оценки меры аппроксимации $\gamma_n(u^*)$ или в задачу оценки $\|\overline{\varphi}_n\|$. Если же нас интересуют вопросы устойчивости счета, то при таком определении норм мы вряд ли сможем на основании результатов общей теории дать какие-либо полезные рекомендации, например, относительно допустимых в правой части уравнений (2) ошибок округления. Таким образом, при изучении устойчивости выбор норм в U_n и F_n в значительной мере определяется ценностью тех рекомендаций относительно проведения счета, которые мы сможем дать на основании общих результатов. Утверждение, что такой-то метод устойчив (устойчив процесс отыскания каркасов), всегда должно сопровождаться указанием выбора норм в пространствах U_n и F_n, при котором эта устойчивость имеет место. Уравнения (2) чаще всего суть системы линейных алгебраических уравнений. В этом случае в пространствах U_n и F_n, которые суть ко-

[1] Например, при изучении метода Ритца вопросы сходимости можно решать в предположении, что нам задана последовательность подпространств, в которых мы ищем приближенные решения. Устойчивость же решающим образом зависит от конкретного выбора координатной системы. т. е. выбора базисов этих подпространств.

нечномерные векторные пространства, естественно пользоваться обычными векторными нормами.

Напомним наиболее употребительные в вычислительной математике векторные нормы и одновременно укажем те символы, которые будут использоваться ниже для их обозначения.

Пусть $x = (\xi_1, ..., \xi_n)$ — n-мерный вектор с вещественными компонентами. Мы будем в основном использовать векторные нормы

$$\| x \|_\infty = \max_{1 \le k \le n} |\xi_k|.$$

и евклидову

$$\| x \|_2 = \sqrt{\sum_{k=1}^{n} \xi_k^2}.$$

Векторные пространства с такими нормами будем обозначать соответственно l_n^∞ и l_n^2.

Если

$$A = \begin{pmatrix} a_{11} & ... & a_{1n} \\ ... & ... & ... \\ a_{n1} & ... & a_{nn} \end{pmatrix}$$

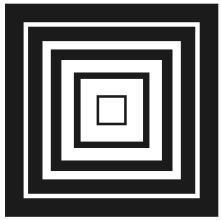
— квадратная матрица порядка n, то через $\| A \|_\infty$ и $\| A \|_2$ будем обозначать ее нормы как оператора, действующего соответственно в l_n^∞ и в l_n^2. Напомним, как вычисляются эти нормы:

$$\| A \|_\infty = \max_k \sum_{j=1}^{n} |a_{kj}|,$$

$$\| A \|_2 = \sqrt{\Lambda}.$$

Здесь Λ — максимальное собственное число матрицы $A'A$ (A' — транспонированная матрица A). Если матрица A симметрична, то $\| A \|_2 = \max |\lambda_k|$ — максимальное по модулю собственное число матрицы A. Наконец, символами $\mu_\infty(A)$ и $\mu_2(A)$ будем обозначать числа обусловленности матрицы A, соответствующие введенным нормам:

$$\mu_\infty(A) = \| A \|_\infty \cdot \left\| A^{-1} \right\|_\infty, \quad \mu_2(A) = \| A \|_2 \cdot \left\| A^{-1} \right\|_2.$$



Глава 2

Приближенное решение интегральных уравнений

§ 1. Метод замены ядра на вырожденное

В этой главе мы будем заниматься вопросами приближенного решения интегральных уравнений Фредгольма второго рода:

$$Au = u(x) - \int_a^b H(x, y)u(y)dy = f(x). \tag{1}$$

Отметим, что, как это обычно принято, одной буквой H мы обозначаем и ядро, и интегральный оператор, этим ядром определяемый. Методы, рассматриваемые в этой главе, специфичны для интегральных уравнений. Они позволяют дать наиболее простую иллюстрацию применения общей теории приближенных методов, изложенной в *главе 1*. В то же время для приближенного решения интегральных уравнений могут применяться и проекционные методы; об этом будет сказано в *главе 3*.

Метод, о котором пойдет речь в этом параграфе, основан на том, что существует класс интегральных уравнений, для которых не трудно построить точное решение. Это — интегральные уравнения с вырожденными ядрами.

Определение

Ядро $\widetilde{H}(x, y)$ называется *вырожденным*, если оно представимо в виде

$$\widetilde{H}(x, y) = \sum_{k=1}^{n} \alpha_k(x)\beta_k(y).$$

В этом представлении функции $\alpha_1(x)$, $\alpha_2(x), ..., \alpha_n(x)$, а также функции $\beta_1(y)$, $\beta_2(y), ..., \beta_n(y)$ можно считать линейно независимыми, т. к. в противном случае ядро $\widetilde{H}(x, y)$ допускает такое же представление, но с меньшим числом слагаемых. В предположении такой линейной независимости (это предположение мы считаем впредь выполненным) число n называют *рангом вырожденного ядра* $\widetilde{H}$.

Решение интегрального уравнения

$$\widetilde{A}u = u - \widetilde{H}u = u(x) - \int_a^b \widetilde{H}(x, y)u(y)dy = f(x) \tag{2}$$

с вырожденным ядром легко сводится к решению системы линейных алгебраических уравнений. Действительно, если $\widetilde{u}^*$ есть решение уравнения (2), то, перенеся в тождестве $\widetilde{u}^* - \widetilde{H}\widetilde{u}^* = f$ член $\widetilde{H}\widetilde{u}^*$ в правую часть, получим

$$\widetilde{u}^*(x) = f(x) + \sum_{k=1}^n \int_a^b \beta_k(y)\widetilde{u}^*(y)dy\alpha_k(x) = f(x) + \sum_{k=1}^n c_k\alpha_k(x), \tag{3}$$

т. е. если уравнение (2) имеет решение, то это решение заведомо представимо в виде суммы правой части уравнения с линейной комбинацией функций $\alpha_k(x)$. При этом коэффициенты упомянутой линейной комбинации таковы:

$$c_k = \int_a^b \beta_k(y)\widetilde{u}^*(y)dy = \int_a^b \beta_k(y)\left(f(y) + \sum_{j=1}^n c_j\alpha_j(y)\right)dy =$$

$$= \sum_{j=1}^n \int_a^b \beta_k(y)\alpha_j(y)dyc_j + \int_a^b \beta_k(y)f(y)dy,$$

т. е. они являются решением системы линейных алгебраических уравнений

$$\overline{A}z = b. \tag{4}$$

Здесь $\overline{A} = \{a_{kj}\}$ — матрица коэффициентов; $b = (b_1, ..., b_n)$ — вектор правых частей; $z = (c_1, ..., c_n)$. Отметим, что коэффициенты системы[1]

$$a_{kj} = \delta_{kj} - \int_a^b \beta_k(y)\alpha_j(y)dy$$

[1] δ_{kj} — символ Кронекера.

зависят лишь от ядра интегрального уравнения, но не от его правой части; от правой части f зависят компоненты вектора b:

$$b_k = \int_a^b \beta_k(y) f(y) dy.$$

Итак, любое решение уравнения (2) имеет вид (3), где коэффициенты c_k удовлетворяют системе уравнений (4). Легко убедиться и в обратном: если числа c_k удовлетворяют системе (4), то функция $\tilde{u}^*(x)$, построенная по формуле (3), является решением уравнения (2). Отсюда, в частности, следует, что для того, чтобы интегральное уравнение (2) было однозначно разрешимо, необходимо и достаточно, чтобы определитель матрицы $\overline{A}$ был отличен от нуля.

Если $\det \overline{A} \neq 0$, то нетрудно построить резольвенту ядра $\tilde{H}$. Обозначим через $D = \{d_{kj}\}$ матрицу, обратную матрице $\overline{A}$: $D = \overline{A}^{-1}$. Тогда коэффициенты c_k вычисляют по формулам

$$c_k = \sum_{j=1}^n d_{kj} b_j,$$

и потому

$$\tilde{u}^*(x) = f(x) + \sum_{k=1}^n c_k \alpha_k(x) = f(x) + \sum_{k=1}^n \sum_{j=1}^n d_{kj} \int_a^b \beta_j(y) f(y) dy \alpha_k(x) =$$

$$= f(x) + \int_a^b \tilde{\Gamma}(x, y) f(y) dy,$$

где $\tilde{\Gamma}(x, y) = \sum_{k=1}^n \sum_{j=1}^n d_{kj} \alpha_k(x) \beta_j(y)$ и есть резольвента ядра $\tilde{H}$.

Метод замены ядра на вырожденное состоит в том, что для ядра H интегрального уравнения (1) мы подбираем "близкое" к нему вырожденное ядро $\tilde{H}$ и решение уравнения (2) $\tilde{u}^*$ принимаем за приближенное решение уравнения (1). В каком смысле следует понимать близость ядер H и $\tilde{H}$, будет видно из теоремы об оценке погрешности. Эту теорему мы будем доказывать, рассматривая уравнения (1) и (2) как уравнения в пространстве

$C = C[a, b]$ непрерывных на $[a, b]$ функций и считая при этом ядра H и $\widetilde{H}$ и правую часть f непрерывными. Впрочем, столь же просто можно было бы получить аналогичные результаты и для других пространств функций.

Теорема 1 (об оценке погрешности)

Пусть интегральное уравнение (2) однозначно разрешимо, его решение есть $\widetilde{u}^*$ и для резольвенты $\widetilde{\Gamma}$ ядра $\widetilde{H}$ выполнена оценка

$$\int_a^b \left|\widetilde{\Gamma}(x, y)\right| dy \le \widetilde{B} \quad (x \in [a, b]).$$

Пусть ядра H и $\widetilde{H}$ связаны неравенством

$$\int_a^b \left|H(x, y) - \widetilde{H}(x, y)\right| dy \le \eta \quad (x \in [a, b]),$$

причем, $(1 + \widetilde{B})\eta < 1$. Тогда уравнение (1) также однозначно разрешимо,

$$\left\|A^{-1}\right\| \le \frac{1 + \widetilde{B}}{1 - (1 + \widetilde{B})\eta} \tag{5}$$

и для его решения u^* при всех $x \in [a, b]$ выполняется оценка

$$\left|u^*(x) - \widetilde{u}^*(x)\right| \le \frac{(1 + \widetilde{B})\eta}{1 - (1 + \widetilde{B})\eta} \left\|\widetilde{u}^*\right\|_C. \tag{6}$$

Доказательство

Запишем уравнение (1) в виде

$$(\widetilde{A} + \Delta\widetilde{A})u = f, \tag{7}$$

где $\Delta\widetilde{A} = A - \widetilde{A} = \widetilde{H} - H$ есть интегральный оператор с ядром $\widetilde{H}(x, y) - H(x, y)$, и потому

$$\left\|\Delta\widetilde{A}\right\| = \max_x \int_a^b \left|\widetilde{H}(x, y) - H(x, y)\right| dy \le \eta.$$

Применим теперь к уравнениям (2) и (7) теорему о близких уравнениях *(теорема 1 из § 1 главы 1)*. Условия этой теоремы выполнены, т. к. пространство C — полное, оператор $\widetilde{A}^{-1}$ существует, причем

$$\left\|\widetilde{A}^{-1}\right\| = \left\|(I-\widetilde{H})^{-1}\right\| = \left\|I+\widetilde{\Gamma}\right\| \le 1 + \left\|\widetilde{\Gamma}\right\| = 1 + \max_x \int_a^b \left|\widetilde{\Gamma}(x,y)\right| dy \le 1 + \widetilde{B},$$

и, наконец,

$$\left\|\widetilde{A}^{-1}\Delta\widetilde{A}\right\| \le \left\|\widetilde{A}^{-1}\right\| \cdot \left\|\Delta\widetilde{A}\right\| \le (1+\widetilde{B})\eta = \rho < 1.$$

Поэтому согласно утверждению этой теоремы существует обратный оператор $A^{-1} = (\widetilde{A} + \Delta\widetilde{A})^{-1}$ и его норма допускает оценку

$$\left\|A^{-1}\right\| \le \frac{\left\|\widetilde{A}^{-1}\right\|}{1-\rho} \le \frac{1+\widetilde{B}}{1-(1+\widetilde{B})\eta}.$$

Что касается оценки погрешности, то ввиду $\Delta\widetilde{f} = 0$ она принимает вид

$$\left\|u^* - \widetilde{u}^*\right\| = \left\|\Delta\widetilde{u}\right\| \le \frac{\rho}{1-\rho}\left\|\widetilde{u}^*\right\| = \frac{(1+\widetilde{B})\eta}{1-(1+\widetilde{B})\eta}\left\|\widetilde{u}^*\right\|. \ \blacksquare$$

Отметим некоторые особенности доказанной теоремы и возможности ее применения. При доказательстве теоремы нигде не использовалось предположение, что ядро $\widetilde{H}$ вырожденное. Поэтому теорема в принципе применима к любым двум интегральным уравнениям с близкими ядрами. Но эффективной оценку погрешности (6) можно считать, видимо, лишь в случае, когда ядро $\widetilde{H}$ вырожденное, т. к. только тогда вычисление $\widetilde{B}$, η и $\left\|\widetilde{u}^*\right\|$ сводится к оценке известных нам функций или функций, которые мы реально можем построить $\left(\widetilde{u}^*(x),\ \ \widetilde{\Gamma}(x,y)\right)$. Оценка (6) имеет апостериорный характер — ее правую часть мы можем вычислить уже после того, как приближенное решение $\widetilde{u}^*$ построено; оценку $\widetilde{B}$ удобно находить в процессе построения приближенного решения. Правая часть оценки (6), вообще говоря, мала, если мало число η. Поэтому "малость" η может служить критерием "близости" ядер H и $\widetilde{H}$. Наконец, стоит заметить, что теорема гарантирует нам существование "точного" решения u^* — этот факт заранее нам может быть и неиз-

вестен. Построив для ядра H близкое к нему вырожденное небольшого ранга, часто можно получить с помощью (5) оценку нормы обратного оператора A^{-1}. Такая оценка бывает чрезвычайно полезна; в частности, как уже отмечалось, она позволяет получить оценку погрешности любого приближенного решения через его невязку.

Рассмотрим теперь случай, когда мы располагаем последовательностью интегральных уравнений

$$A_n u = u(x) - \int_a^b H_n(x, y)u(y)dy = f(x) \tag{8}$$

с вырожденными ядрами H_n (n не обязательно есть ранг ядра H_n). К этому случаю относится следующее утверждение.

Теорема 2 (о сходимости)

Пусть интегральное уравнение (1) однозначно разрешимо, причем $\left\|A^{-1}\right\| \le c$. Пусть при всех n выполняются оценки

$$\int_a^b \left|H(x, y) - H_n(x, y)\right| dy \le \eta_n,$$

причем $\eta_n \to 0$. Тогда при достаточно больших n уравнения (8) также однозначно разрешимы, их решения u_n^* равномерно сходятся к решению u^* уравнения (1) и справедлива оценка

$$\left\| u^* - u_n^* \right\|_C \le \frac{c^2 \eta_n}{1 - c\eta_n} \left\| f \right\|_C. \tag{9}$$

Доказательство

Запишем уравнения (8) в виде

$$(A + \Delta A_n)u = f,$$

где $\Delta A_n = A_n - A = H - H_n$, и, как легко видеть, $\left\|\Delta A_n\right\| \le \eta_n$. Поскольку $\left\|A^{-1}\Delta A_n\right\| \le \left\|A^{-1}\right\| \cdot \left\|\Delta A_n\right\| \le c\eta_n = \rho_n \to 0$, то при достаточно больших n (когда

$\rho_n < 1$) к уравнениям (1) и (8) применима теорема о близких уравнениях *(теорема 1 из § 1 главы 1)*. Из нее непосредственно и следуют все утверждения нашей теоремы, только оценка погрешности (с учетом того, что $\Delta f = 0$) строится в той форме, как она указана в замечании к теореме о близких уравнениях. ■

Оценка (9) в отличие от оценки (6) является априорной, т. е. ее правую часть мы можем вычислить до того, как построено само приближенное решение u_n^*. Поэтому эта оценка применима для решения следующей задачи: дано уравнение (1), требуется выбрать такое вырожденное ядро H_n, чтобы погрешность приближенного решения не превосходила некоторого заданного числа ε. Впрочем, эффективной оценку (9) можно считать лишь в том случае, если мы располагаем оценкой нормы обратного оператора $\left\|A^{-1}\right\| \le c$. Один из способов построения последней был указан в комментарии к теореме 1. Оценка (9) позволяет также судить о быстроте сходимости приближенных решений u_n^* к точному — погрешность этих решений есть бесконечно малая порядка не ниже, чем η_n.

Доказанные теоремы 1 и 2 являются хорошей иллюстрацией к тому, что говорилось в *главе 1* о применениях теоремы о близких уравнениях *(см. замечание 3 из § 1 главы 1)*.

В заключение параграфа отметим, что бесспорным достоинством метода замены ядра на вырожденное является та простота, с которой (по крайней мере при малом ранге вырожденного ядра) осуществляется оценка погрешности. Метод позволяет также оценить норму обратного оператора A^{-1}. Вместе с тем этот метод трудно поддается постановке на машину. Удачный выбор близкого вырожденного ядра требует известного искусства вычислителя и зависит от его опыта. Вычисление элементов системы (4) — это вычисление довольно большого числа интегралов, так что при высоком ранге вырожденного ядра метод оказывается трудоемким. По этим причинам метод замены ядра на вырожденное применяется обычно, лишь когда требуется получить решение с небольшой точностью, и рассчитан он в основном на ручной счет. Этот метод естественно применять также для получения начальной информации о решении интересующего нас уравнения.

§ 2. Метод механических квадратур. Теорема о сходимости

Метод механических квадратур является основным (наиболее часто употребляемым) способом решения интегральных уравнений. Мы будем рассматривать интегральное уравнение Фредгольма второго рода

$$Au = u(x) - \int_0^1 H(x,y)u(y)dy = f(x), \tag{1}$$

которое коротко можно записать в виде

$$Au = (I - H)u = f,$$

где H — интегральный оператор с ядром $H(x,y)$, I — тождественный оператор; $A = I - H$. Ядро $H(x,y)$ и правую часть $f(x)$ будем считать непрерывными функциями, не оговаривая этого каждый раз в дальнейшем. Сразу же заметим, что все, что будет сказано в этом параграфе, может быть отнесено и к многомерным интегральным уравнениям вида

$$u(x) - \int_D H(x,y)u(y)dy = f(x),$$

где D — ограниченная область в m-мерном пространстве; $x,\ y \in D$. Именно потому, что переход к многомерному случаю не вызывает никаких затруднений, мы будем говорить для определенности об уравнении (1).

По отношению к уравнению (1) метод механических квадратур заключается в следующем. Выберем какую-нибудь последовательность квадратурных формул (квадратурный процесс)

$$\int_0^1 \upsilon(x)dx \approx \sum_{k=1}^n p_k^{(n)} \upsilon(x_k^{(n)}), \tag{2}$$

где $p_k^{(n)}$ — веса, а $x_k^{(n)}$ $(x_k^{(n)} \in [0,1],\ x_k^{(n)} \neq x_j^{(n)}$ при $k \neq j)$ — узлы этих формул. Заменив интеграл в уравнении (1) приближенно на квадратурную сумму, мы получим новое уравнение относительно новой неизвестной функции $u_n(x)$:

$$A_n u_n = u_n(x) - \sum_{k=1}^n p_k^{(n)} H(x, x_k^{(n)}) u_n(x_k^{(n)}) = f(x). \tag{3}$$

Если квадратурная сумма достаточно хорошо приближает интеграл, то есть основания надеяться, что решение $u_n^*(x)$ уравнения (3) близко к решению $u^*(x)$ уравнения (1). Само уравнение (3) решается относительно просто. Действительно, если $u_n^*(x)$ есть его решение, то числа $\zeta_j = u_n^*(x_j^{(n)})$ обязаны удовлетворять системе уравнений

$$\zeta_j - \sum_{k=1}^{n} p_k^{(n)} H(x_j^{(n)}, x_k^{(n)})\zeta_k = f(x_j^{(n)}), \quad j = 1, 2, ..., n$$

или, в матричной записи,

$$\overline{A}_n z_n = g_n, \tag{4}$$

где $\overline{A}_n = \left\{\delta_{jk} - p_k^{(n)} H\left(x_j^{(n)}, x_k^{(n)}\right)\right\}$ — матрица коэффициентов системы; $g_n = \left(f(x_1^{(n)}), ..., f(x_n^{(n)})\right)$; $z_n = (\zeta_1, ..., \zeta_n)$ — искомый вектор.

Если $z_n^* = (\zeta_1^*, ..., \zeta_n^*)$ есть решение системы (4), то решение $u_n^*(x)$ уравнения (3) может быть получено по формуле

$$u_n^*(x) = \sum_{k=1}^{n} p_k^{(n)} H(x, x_k^{(n)})\zeta_k^* + f(x). \tag{5}$$

Таким образом, решение уравнения (3) сводится к решению системы линейных алгебраических уравнений (4).

Подчеркнем, что уравнения (3) и (4) эквивалентны в следующем смысле. Если $u_n^*(x)$ есть решение уравнения (3), то вектор $z_n^* = \{\zeta_j^*\}$, где $\zeta_j^* = u_n^*(x_j^{(n)})$, есть решение (4). Если z_n^* есть решение (4), то функция u_n^*, построенная по формуле (5), удовлетворяет уравнению (3). Однозначная разрешимость одного из этих уравнений влечет однозначную разрешимость другого.

Применяя метод механических квадратур, иногда довольствуются решением системы уравнений (4), рассматривая вектор z_n^* как таблицу значений приближенного решения в узлах $x_k^{(n)}$. Иногда, получив z_n^*, приближенное решение строят не по формуле (5), а с помощью какой-либо интерполяции этой таблицы.

Несомненным достоинством метода механических квадратур является простота его машинной реализации. Действительно, построение элементов сис-

темы уравнений (4) требует лишь вычисления значений ядра H и правой части f в отдельных точках (узлах квадратурной формулы). Столь же легко проводится счет по формуле (5).

Перейдем к исследованию метода механических квадратур с помощью общей схемы, изложенной в *главе 1*. При этом уравнение (1) мы будем рассматривать как уравнение в пространстве C непрерывных на промежутке [0, 1] функций, так что $U = F = C$. Что касается приближенного уравнения и пространств U_n, F_n, то из множества возможных здесь вариантов наиболее естественными представляются следующие два.

1. В качестве приближенного уравнения можно рассматривать (3). Это — уравнение в том же самом пространстве C, так что при всех n $U_n = F_n = C$. Значения оператора A_n определяются формулой (3). Правая часть уравнения (3) совпадает с правой частью (1), поэтому оператор $\psi_n = I$ есть тождественный оператор в C. Точно так же, поскольку само решение уравнения (3) рассматривается как приближенное решение уравнения (1), то и оператор $\overline{\varphi}_n$ есть тождественный оператор в C. Наконец, ввиду совпадения пространств U и U_n в качестве φ_n также естественно взять тождественный оператор.

2. В качестве приближенного уравнения можно рассматривать систему линейных алгебраических уравнений (4). Тогда U_n и F_n суть n-мерные векторные пространства. Нормы в этих пространствах могут быть выбраны разными способами, но наиболее удобный — положить $U_n = F_n = l_n^\infty$ — n-мерное векторное пространство с векторной нормой $\|z_n\|_\infty$, т. к. эта норма лучше всего согласуется с нормой, введенной в пространствах U и F. Оператор $\psi_n \in L(F, F_n)$, ставящий в соответствие правой части уравнения (1) правую часть (4), должен действовать по правилу

$$\psi_n f = \left(f(x_1^{(n)}), f(x_2^{(n)}), ..., f(x_n^{(n)}) \right).$$

Что касается оператора $\overline{\varphi}_n$, то с ним дело обстоит сложнее. Построение приближенного решения по формуле (5) не укладывается в общую схему, т. к. правая часть (5) не есть результат применения линейного оператора к z_n^* — мешает слагаемое $f(x)$. Если же мы условимся строить приближенное решение с помощью интерполяции таблицы $u^{(n)}(x_j^{(n)}) = \zeta_j^*$, то

$\overline{\varphi}_n$ — соответствующий интерполяционный оператор. Наконец, наиболее естественный способ выбора оператора φ_n — положить $\varphi_n = \psi_n$, т. е.

$$\varphi_n u = \left(u(x_1^{(n)}), u(x_2^{(n)}), ..., u(x_n^{(n)}) \right).$$

Естественным это представляется потому, что, например, в определении сходимости каркасов каркас приближенного решения (в нашем случае он является таблицей приближенного решения в узлах $x_k^{(n)}$) сопоставляется с $\varphi_n u^*$, а $\varphi_n u^*$ будет теперь таблицей точного решения с теми же узлами. Довольно часто бывают случаи, когда практически достаточно знать лишь таблицу приближенного решения в узлах $x_k^{(n)}$. Тогда вопрос об определении оператора $\overline{\varphi}_n$ вообще не встает — нужный результат даст нам теорема о сходимости каркасов. Утверждение о сходимости каркасов в нашей схеме будет означать, что при всех достаточно больших n компоненты решения системы (4) отличаются от значений точного решения в соответствующих узлах сколь угодно мало.

Сравним варианты 1 и 2. Отметим прежде всего два преимущества схемы 1. Во-первых, эта схема проще; простота заключается в совпадении пространств U_n и F_n с U и F, простоте операторов ψ_n, $\overline{\varphi}_n$, φ_n, в результате чего, например, приближенное решение совпадает с каркасом. Во-вторых, как уже отмечалось, формула (5) вообще не укладывается в схему 2. Вместе с тем при изучении устойчивости нас в первую очередь интересует вопрос о поведении систем линейных алгебраических уравнений (4), насколько хорошо обусловленными они окажутся; но именно эти системы полностью игнорируются схемой 1, и ответ на вопрос об устойчивости интересен лишь в рамках схемы 2. Однако нужно помнить, что с помощью схем 1 и 2 исследуется в действительности один и тот же метод. Это позволит нам пойти по такому пути. Сходимость метода будет исследоваться по схеме 1, как более простой. Полученные при этом результаты будут использованы для доказательства устойчивости систем уравнений (4) (устойчивости процесса отыскания каркасов в рамках схемы 2).

Наша цель — установить сходимость метода механических квадратур при единственном предположении сходимости квадратурного процесса (2) для любой непрерывной функции. Последнее условие (впредь мы будем выражать его словами "квадратурный процесс сходится") ни в коей мере не явля-

ется жестким — квадратурные процессы, не удовлетворяющие условию сходимости, реально никогда не используются.

Обратим внимание на одну трудность, которая возникает при исследовании метода механических квадратур. Как будет показано далее, операторы A_n, определяемые формулой (3), сходятся к оператору A поточечно, т. е. для любой функции $u \in C$ будет $\|A_n u - Au\|_C \to 0$. Но операторы A_n не сходятся к A по норме, т. е. $\|A_n - A\|$ не стремится к нулю. Это не позволяет нам использовать теорему о близких уравнениях, в которой предполагается близость операторов по норме. Поэтому сейчас мы отвлечемся от нашего интегрального уравнения, чтобы установить некоторые свойства последовательностей операторов, сходящихся поточечно. Далее, когда мы будем говорить "последовательность операторов A_n сходится" и писать $A_n \to A$, мы всегда будем иметь в виду поточечную сходимость. Напомним, что в этом случае операторы A_n равномерно ограничены: $\|A_n\| \le c$ (c не зависит от n).

Теорема 1 (о близких уравнениях второго рода[1])

Пусть H и $\widetilde{H}$ — линейные ограниченные операторы в банаховом пространстве U ($H, \widetilde{H} \in L(U, U)$) и пусть выполняются условия:

1) существует обратный оператор

$$(I - H)^{-1} \in L(U, U),$$

2) $\left\|(I - H)^{-1}\right\| \cdot \left\|(H - \widetilde{H})\widetilde{H}\right\| \le q < 1.$

Тогда оператор $I - \widetilde{H}$ также обратим и[2]

$$\left\|(I - \widetilde{H})^{-1}\right\| \le M/(1-q),$$

где $M = 1 + \left\|(I - \widetilde{H})^{-1}\right\| \cdot \left\|\widetilde{H}\right\|.$

---

[1] Уравнением второго рода обычно называют уравнение с выделенным тождественным оператором, т. е. вида $u - Hu = f$.

[2] Оператор $(I - \widetilde{H})^{-1}$ задан на $R(I - \widetilde{H})$, так что следующее далее неравенство понимается так: для любого $f \in R(I - \widetilde{H})$ $\left\|(I - \widetilde{H})^{-1} f\right\| \le (M/(1-q))\|f\|$.

Доказательство

Возьмем произвольный элемент $u \in U$ и проведем тождественное преобразование

$$u = (I - \widetilde{H})u + \widetilde{H}u = (I - \widetilde{H})u + (I - H)^{-1}(I - H)\widetilde{H}u =$$
$$= (I - \widetilde{H})u + (I - H)^{-1}[\widetilde{H}(1 - \widetilde{H}) + (\widetilde{H} - H)\widetilde{H}]u =$$
$$= [I + (I - H)^{-1}\widetilde{H}](I - \widetilde{H})u + (I - H)^{-1}(\widetilde{H} - H)\widetilde{H}u.$$

Отсюда вытекает неравенство

$$\| u \| \le M \| (I - \widetilde{H})u \| + q \| u \|,$$

или

$$\| u \| \le \frac{M}{1 - q} \| (I - \widetilde{H})u \|. \qquad (6)$$

Таким образом, однородное уравнение $(I - \widetilde{H})u = 0$ имеет только нулевое решение, и оператор $I - \widetilde{H}$ обратим. Наконец, если $f \in R(I - \widetilde{H})$ и $u = (I - \widetilde{H})^{-1} f$, т. е. $(I - \widetilde{H})u = f$, то неравенство (6) может быть прочитано и так:

$$\| (I - \widetilde{H})^{-1} f \| \le \frac{M}{1 - q} \| f \|. \quad \blacksquare$$

❑ **Замечание.** Если оператор $\widetilde{H}$ компактен, то, как известно, обратимость оператора $I - \widetilde{H}$ влечет равенство $R(I - \widetilde{H}) = U$, так что при этом дополнительном предположении оператор $(I - \widetilde{H})^{-1}$ задан на всем U.

Дадим теперь следующее определение.

Определение

Пусть H_α, где α пробегает некоторое множество Q, — семейство компактных операторов, действующих из банахова пространства U в U. Пусть $S = \{u \mid u \in U, \|u\| \le 1\}$ — единичный шар в U. Будем говорить, что операторы H_α *компактны в совокупности*, если множество $\bigcup_{\alpha \in Q} H_\alpha S$ (объединение образов единичного шара) относительно компактно.

Компактные в совокупности операторы H_α тем более равномерно ограничены: $\|H_\alpha\| \le C$ при всех α, т. к. относительно компактное множество $\bigcup H_\alpha S$ ограничено. Ясно также, что если $W \subset U$ — любое ограниченное множество, а операторы H_α компактны в совокупности, то множество $\bigcup H_\alpha W$ относительно компактно.

Прежде чем формулировать следующую теорему, докажем лемму общего характера, которая будет использоваться и в дальнейшем.

Лемма

Пусть $A, A_n \in L(U,F)$ и $A_n \to A$. Тогда эта сходимость равномерна на любом относительно компактном множестве. То есть если $M \subset U$ относительно компактно, то по любому $\varepsilon > 0$ найдется такое N, что при всех $n > N$ для всех $u \in M$ выполняется неравенство $\|A_n u - Au\| < \varepsilon$.

Доказательство

Пусть $M \subset U$ относительно компактно. Возьмем $\varepsilon > 0$. По теореме Банаха—Штейнгауза сходящиеся операторы A_n равномерно ограничены: $\|A_n\| < K$ при всех n. Положим $\varepsilon_1 = \varepsilon/(K + \|A\| + 1)$. Для этого ε_1 найдется конечная ε_1-сеть относительно компактного множества M: $\{u_1, u_2, \ldots, u_m\}$. Выберем теперь N из того условия, чтобы при всех $n > N$ и всех $k = 1,2,\ldots,m$ выполнялось неравенство $\|A_n u_k - Au_k\| < \varepsilon_1$. Пусть элемент $u \in M$ произволен. Для него найдется такой номер j, что $\|u - u_j\| < \varepsilon_1$, и при $n > N$ имеем:

$$\|A_n u - Au\| \le \|A_n u - A_n u_j\| + \|A_n u_j - Au_j\| + \|Au_j - Au\| < K\varepsilon_1 + \varepsilon_1 + \|A\|\varepsilon_1 = \varepsilon.$$

Этим лемма доказана. ■

Теорема 2
(о последовательности операторов, компактных в совокупности)

Пусть H_n ($n = 1,2,\ldots$) и H — линейные ограниченные операторы в пространстве U ($H, H_n \in L(U,U)$) и пусть выполнены условия: 1) $H_n \to H$; 2) H_n компактны в совокупности; 3) существует обратный оператор $(I - H)^{-1} \in L(U,U)$.

Тогда при всех достаточно больших n $(n \geq N)$ существуют обратные операторы $(I - H_n)^{-1}$, и эти операторы равномерно ограничены:

$$\left\|(I - H_n)^{-1}\right\| \leq c \quad (n \geq N).$$

Доказательство

Положим $V = \bigcup_n H_n S$, где S — единичный шар в U. В силу предположения 2) множество V относительно компактно. Поэтому по произвольному $\varepsilon > 0$ найдется такое N_1, что для всех $\upsilon \in V$ и всех $n \geq N_1$ $\left\|H_n \upsilon - H\upsilon\right\| < \varepsilon$. Пусть теперь u — произвольный элемент из S (т. е. $\left\|u\right\| \leq 1$) и $n \geq N_1$. Тогда поскольку $H_n u = \upsilon \in V$, будет $\left\|(H_n - H)H_n u\right\| < \varepsilon$.

Итак, при $n \geq N_1$ $\left\|(H_n - H)H_n\right\| < \varepsilon$, т. е. $\left\|(H_n - H)H_n\right\| \to 0$.

Тем самым

$$q_n = \left\|(I - H)^{-1}\right\| \cdot \left\|(H_n - H)H_n\right\| \to 0$$

и по теореме 1 при достаточно больших n (когда $q_n < 1$) операторы $I - H_n$ обратимы. Ввиду компактности операторов H_n обратные операторы $(I - H_n)^{-1}$ заданы на всем U. Теорема 1 дает нам также оценку

$$\left\|(I - H_n)^{-1}\right\| \leq \frac{M_n}{1 - q_n} \leq C \quad (n \geq N),$$

где $M_n = 1 + \left\|(I - H)^{-1}\right\| \cdot \left\|H_n\right\| \leq M$.

Здесь мы воспользовались тем, что операторы H_n равномерно ограничены. ■

❑ **Замечание.** Факт равномерной ограниченности операторов $(I - H_n)^{-1}$ может быть выражен, как это видно из доказательства, формулой

$$\overline{\lim}\left\|(I - H_n)^{-1}\right\| \leq 1 + \left\|(I - H)^{-1}\right\| \cdot \overline{\lim}\left\|H_n\right\|.$$

Вернемся теперь к рассмотрению метода механических квадратур для интегрального уравнения (1). Как уже было сказано, квадратурный процесс (2) мы предполагаем сходящимся, т. е. для любой непрерывной функции υ пра-

вая часть формулы (2) стремится при $n \to \infty$ к левой. Следствием этого предположения является существование такой постоянной M, что при всех n

$$\sum_{k=1}^{n} \left| p_k^{(n)} \right| \le M. \tag{7}$$

Это утверждение является простым следствием теоремы Банаха—Штейнгауза, если учесть, что $\sum_{k=1}^{n} \left| p_k^{(n)} \right|$ есть норма функционала, заданного в пространстве C, который непрерывной функции υ ставит в соответствие квадратурную сумму $\sum_{k=1}^{n} p_k^{(n)} \upsilon(x_k^{(n)})$, а сходимость квадратурного процесса означает слабую (т. е. поточечную) сходимость этих функционалов к интегральному.

Введем в пространстве C операторы H_n:

$$H_n u = \sum_{k=1}^{n} p_k^{(n)} H(x, x_k^{(n)}) u(x_k^{(n)}),$$

так что операторы A_n можно записать в виде $A_n = I - H_n$ (см. формулу (3)). Кроме того, обозначим через h максимум модуля ядра H:

$$\left| H(x, y) \right| \le h. \tag{8}$$

Лемма 1

Если квадратурный процесс (2) сходится, то операторы H_n компактны в совокупности.

Доказательство

Нам требуется показать относительную компактность в пространстве C множества $V = \bigcup H_n S$, где S — единичный шар в C. Воспользуемся для этого теоремой Арцела—Асколи, согласно которой достаточно показать равномерную ограниченность и равностепенную непрерывность функций из V.

Пусть $\upsilon \in V$. Тогда при некотором n

$$\upsilon = H_n u = \sum_{k=1}^{n} p_k^{(n)} H(x, x_k^{(n)}) u(x_k^{(n)}),$$

где $\left| u(x) \right| \le 1$. Из оценок (7) и (8) сразу же следует, что $\left| \upsilon(x) \right| \le hM$. Таким образом, функции множества V равномерно ограничены. Возьмем теперь произвольное $\varepsilon > 0$. Так как ядро $H(x, y)$ непрерывно, а значит, и равномерно

непрерывно в квадрате $0 \le x,\ y \le 1,$ то выбранному ε соответствует такое $\delta > 0,$ что как только точки (x_1, y_1) и (x_2, y_2) из нашего квадрата удовлетворяют неравенству $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \le \delta,$ так сейчас же

$$\left|H(x_1, y_1) - H(x_2, y_2)\right| < \varepsilon_1 = \varepsilon / M.$$

Для той же функции $\upsilon \in V$ в предположении, что $\left|x_1 - x_2\right| \le \delta,$ теперь имеем

$$\left|\upsilon(x_1) - \upsilon(x_2)\right| = \left|\sum_{k=1}^{n} p_k^{(n)} \left(H(x_1, x_k^{(n)}) - H(x_2, x_k^{(n)})\right) u(x_k^{(n)})\right| \le M\varepsilon_1 = \varepsilon.$$

Этим показано, что функции υ множества V равностепенно непрерывны. Итак, по теореме Арцела—Асколи множество V относительно компактно. ■

Лемма 2

Если квадратурный процесс (2) сходится, то операторы H_n сходятся к интегральному оператору H.

Доказательство

Пусть $u(x)$ — произвольная непрерывная функция, $\upsilon = Hu,$ $\upsilon_n = H_n u.$ Требуется показать, что функции $\upsilon_n(x)$ равномерно сходятся к функции $\upsilon(x)$. Сходимость $\upsilon_n(x)$ к $\upsilon(x)$ в каждой точке $x \in [0,1]$ очевидна — это следствие сходимости квадратурного процесса для каждой непрерывной функции; все дело в равномерной сходимости. Рассмотрим в пространстве C линейные функционалы

$$q(u) = \int_0^1 u(y)dy, \quad q_n(u) = \sum_{k=1}^{n} p_k^{(n)} u(x_k^{(n)}).$$

Сходимость квадратурного процесса означает слабую сходимость функционалов q_n и q. Введем еще функции

$$w_x(y) = H(x, y)u(y),$$

зависящие от параметра $x \in [0,1]$. Здесь $u(x)$ — выбранная нами в начале доказательства и зафиксированная непрерывная функция. Очевидно, что

$$\upsilon(x) = q(w_x), \quad \upsilon_n(x) = q_n(w_x).$$

Поэтому для доказательства равномерной сходимости $\upsilon_n(x)$ к $\upsilon(x)$ достаточно убедиться, что семейство функций $\{\omega_x\}$ относительно компактно в C, и воспользоваться тем, что сходимость функционалов q_n и q на этом множестве равномерна. При любом $x \in [0, 1]$

$$\|w_x\|_C \le h \cdot \|u\|_C,$$

поэтому функции этого семейства равномерно ограничены. Пользуясь равномерной непрерывностью функций u и H, по $\varepsilon > 0$ выберем $\delta > 0$ так, чтобы для точек $y_1, y_2 \in [0,1]$, таких, что $|y_1 - y_2| < \delta$ было

$$|u(y_1) - u(y_2)| < \varepsilon_1, \quad |H(x, y_1) - H(x, y_2)| < \varepsilon_1,$$

где $\varepsilon_1 = \varepsilon /(h + \| u \|)$. Тогда при любом $x \in [0,1]$ и $|y_1 - y_2| < \delta$ окажется

$$|w_x(y_1) - w_x(y_2)| =$$
$$= |H(x, y_1)u(y_1) - H(x, y_1)u(y_2) + H(x, y_1)u(y_2) - H(x, y_2)u(y_2)| \le$$
$$\le |H(x, y_1)| \cdot |u(y_1) - u(y_2)| + |H(x, y_1) - H(x, y_2)| \cdot |u(y_2)| \le h\varepsilon_1 + \| u \| \cdot \varepsilon_1 = \varepsilon.$$

Итак, функции $w_x(y)$ равностепенно непрерывны, и по теореме Арцела—Асколи семейство $\{w_x\}$ относительно компактно в C. ■

❑ **Замечание.** Операторы H_n сходятся к H поточечно, но не сходятся по норме. Можно было бы показать, что при всех n $\|H - H_n\| \ge \|H\|$. Отсюда сразу же вытекают такие же свойства сходимости последовательности операторов A_n (об этом было упомянуто ранее).

Теорема 3
(о сходимости метода механических квадратур)

Пусть выполнены условия: 1) ядро $H(x, y)$ и правая часть $f(x)$ интегрального уравнения (1) непрерывны; 2) интегральное уравнение (1) однозначно разрешимо; 3) квадратурный процесс (2) сходится.

Тогда: а) при достаточно больших n системы линейных алгебраических уравнений (4), к которым приводит метод механических квадратур, однозначно разрешимы; б) числа обусловленности $\mu_\infty(\overline{A}_n)$ матриц этих систем

равномерно ограничены; в) приближенные решения $u_n^*(x)$, построенные по формуле (5), равномерно сходятся к точному решению $u^*(x)$.

Доказательство

а) Ввиду компактности интегрального оператора H в пространстве C однозначная разрешимость уравнения (1) означает существование ограниченного обратного оператора $A^{-1} = (I - H)^{-1}$. Отсюда и из доказанных лемм 1 и 2 следует, что по отношению к операторам H и H_n выполнены все условия теоремы 2. На основании этой теоремы заключаем, что при достаточно больших n существуют ограниченные в совокупности обратные операторы $A_n^{-1} = (I - H_n)^{-1}$,

$$\left\|A_n^{-1}\right\| \le c'. \tag{9}$$

Поэтому при таких n уравнения (3) однозначно разрешимы. Ввиду указанной с самого начала эквивалентности систем (4) уравнениям (3) эти системы также окажутся однозначно разрешимыми.

б) Напомним, что для матрицы $\overline{A}_n = \{a_{jk}\}$ ее норма выражается так:

$$\left\|\overline{A}_n\right\|_\infty = \max_j \sum_{k=1}^n \left|a_{jk}\right|.$$

Учитывая, что $a_{jk} = \delta_{jk} - p_k^{(n)} H(x_j^{(n)}, x_k^{(n)})$, видим

$$\left\|\overline{A}_n\right\|_\infty \le 1 + \max_j \sum_{k=1}^n \left|p_k^{(n)}\right| \cdot \left|H(x_j^{(n)}, x_k^{(n)})\right| \le 1 + Mh.$$

Перейдем к оценке нормы обратной матрицы $\overline{A}_n^{-1}$. Считая n достаточно большим, возьмем произвольный вектор $b_n = (\beta_1, \beta_2, ..., \beta_n)$, такой, что $\left\|b_n\right\|_\infty = \max\left|\beta_j\right| \le 1$. Наша цель — оценить норму вектора $z_n = \overline{A}_n^{-1} b_n =$ $z_n = \overline{A}_n^{-1} b_n = (\zeta_1, ..., \zeta_n)$. Построим непрерывную функцию $\bar{f}(x)$, удовлетворяющую следующим двум условиям:

$$(i)\ \ \bar{f}(x_j^{(n)}) = \beta_j,\ \ j = 1,\ 2,\ ...,\ n,$$

$$(ii)\ \ \left|\bar{f}(x)\right| \le 1 \text{ при всех } x \in [0, 1], \text{ т. е. } \left\|\bar{f}\right\|_c \le 1.$$

Функций, удовлетворяющих этим условиям, много; например, определив значения $\bar{f}$ в узлах $x_j^{(n)}$ из (i) $\left(\left|\beta_j\right| \le 1!\right)$, можно считать ее линейной между каждыми двумя соседними узлами. Рассмотрим теперь уравнение

$$A_n u_n = \bar{f}.$$

Его решение $\bar{u}_n$ в силу (9) удовлетворяет неравенству $\left\|\bar{u}_n\right\|_C \le c'$. Но так как компоненты вектора $z_n = \bar{A}_n^{-1} b_n$ удовлетворяют системе уравнений

$$\zeta_j - \sum_{k=1} p_k^{(n)} H(x_j^{(n)}, x_k^{(n)})\zeta_k = \bar{f}(x_j^{(n)}) = \beta_j,$$

то $\zeta_j = \bar{u}_n(x_j^{(n)})$, и потому $\left|\zeta_j\right| \le \left\|\bar{u}_n\right\|_c \le c'$. Так как последнее неравенство выполняется при всех j, то и $\left\|z_n\right\|_\infty = \max\left|\zeta_j\right| \le c'$. Итак, взяв произвольный вектор b_n, такой, что $\left\|b_n\right\|_\infty \le 1$, мы доказали неравенство $\left\|\bar{A}_n^{-1} b_n\right\|_\infty \le c'$. Поэтому $\left\|\bar{A}_n^{-1}\right\|_\infty \le c'$. Остается перемножить полученные оценки:

$$\mu_\infty(A_n) = \left\|\bar{A}_n\right\|_\infty \cdot \left\|\bar{A}_n^{-1}\right\|_\infty \le (1 + Mh)c'.$$

Этим утверждение б) теоремы доказано.

в) Для доказательства сходимости воспользуемся результатами общей теории, применив ее по схеме 1. Поскольку ψ_n, φ_n и $\bar{\varphi}_n$ — тождественные операторы, и потому, в частности, приближенное решение совпадает с каркасом, а $\varphi_n u^* = u^*$, теорема о сходимости каркасов приближенных решений (теорема 1 из *§ 2 главы 1*) дает оценку

$$\left\|u^* - u_n^*\right\| \le \left\|\bar{A}_n^{-1}\right\| \cdot \gamma_n(u^*).$$

Но, как было уже показано, $\left\|\bar{A}_n^{-1}\right\| \le c'$. Далее

$$\gamma_n(u^*) = \left\|A_n \varphi_n u^* - \psi_n A u^*\right\| = \left\|A_n u^* - A u^*\right\| = \left\|H u^* - H_n u^*\right\|.$$

Поэтому согласно лемме 2 $\gamma_n(u^*) \to 0$. Итак, $\left\|u^* - u_n^*\right\|_C \to 0$. ■

❑ **Замечание 1.** Утверждение б) теоремы означает, что при исследовании метода механических квадратур по схеме 2 процесс отыскания каркасов приближенных решений устойчив. Полезно также обратить внимание на то, что из приведенного доказательства и замечания к теореме 2 легко получить неравенство

$$\overline{\lim}\,\mu_\infty(A_n) \le (1+Mh)[1+\left\|(I-H)^{-1}\right\|Mh].$$

Отметим еще следующий очевидный факт. Если решение z_n^* системы уравнений (4) найдено нами с погрешностью Δz_n (которая может быть получена как вследствие неточного вычисления элементов матрицы $\overline{A}_n$ и правой части f_n, так и вследствие неточного решения самой системы), то в приближенном решении, построенном по формуле (5), это приведет к погрешности, которая будет оцениваться числом $Mh\|\Delta z_n\|_\infty$. Здесь мы предполагали, что сам счет по формуле (5) осуществлялся точно. Ошибки, возникающие при счете по формуле (5), нетрудно учесть непосредственно.

❑ **Замечание 2.** Как видно из доказательства, в условиях теоремы нетрудно также установить порядок сходимости метода. Именно погрешность метода $\left\|u_n^* - u^*\right\|_C$ есть бесконечно малая не меньшего порядка, чем квадратурная погрешность функции $H(x,\cdot)u^*$, зависящей от параметра x. Под квадратурной погрешностью этой функции мы понимаем

$$\left\|q_n(H(x,\cdot)u^*) - q(H(x,\cdot)u^*)\right\| = \left\|H_n u^* - Hu^*\right\|$$

(по поводу обозначений q_n и q см. доказательство леммы 2). Знание дифференциальных свойств ядра и решения, а также представление остаточного члена квадратурной формулы часто позволяют указать порядок убывания погрешности непосредственно в зависимости от n. Так, если ядро H и правая часть f четырежды непрерывно дифференцируемы (тогда четырежды непрерывно дифференцируемо и решение u^*) и применяется квадратурный процесс Симпсона, то $\left\|u^* - u_n^*\right\| = O(n^{-4})$. Если H и f m раз непрерывно дифференцируемы и применяется квадратурный процесс Гаусса, то $\left\|u^* - u_n^*\right\| = O(n^{-m})$. Для доказательства последнего результата потребовалось бы использовать известную аппроксимационную теорему Д. Джексона.

- **Замечание 3.** Теорема 3 принадлежит И. П. Мысовских (1967 г.), доказательство которого основывалось на тонких оценках погрешности метода механических квадратур, которые интересны и сами по себе. Теорема 2 принадлежит С. Л. Соболеву (1956 г.). Доказательство теоремы 3, основанное на идеях теоремы 2, впервые было предложено Г. М. Вайникко (1969 г.). На примере исследования сходимости метода механических квадратур мы познакомились с простейшим вариантом так называемого "принципа компактной аппроксимации", разработанного Г. М. Вайникко в значительно более общем виде и находящего применение при исследовании других методов и для других задач (в частности, метода сеток для уравнений в частных производных).
- **Замечание 4.** В некоторых более сильных метриках (например, в пространстве C^1 непрерывно дифференцируемых функций с нормой $\|u\|_{C^1} = \|u'\|_C + \|u\|_C$) исследование метода механических квадратур проще, поскольку при соответствующих предположениях о ядре в этом пространстве операторы H_n сходятся к H по норме. Однако такое исследование потребовало бы наложить на ядро H и правую часть f дополнительные условия гладкости.

§ 3. Метод механических квадратур. Оценка погрешности

Теоремы о сходимости каркасов приближенных решений и самих приближенных решений содержат некоторые оценки погрешности. Однако обычно применение этих оценок либо невозможно, либо затруднительно, поскольку в их правые части входят характеристики точного решения задач, которыми обычно мы не располагаем. Все же иногда, если нам удается извлечь из уравнения достаточно богатую информацию о его точном решении, оказывается возможным построить эффективную оценку погрешности. В этом параграфе мы займемся построением оценки погрешности метода механических квадратур для интегрального уравнения[1].

$$Au = u(x) - \int_0^1 H(x, y)u(y)dy = f(x) \tag{1}$$

[1] Излагаемая далее оценка погрешности принадлежит И. П. Мысовских.

с непрерывными ядром H и правой частью f. При этом будем предполагать, что уравнение (1) однозначно разрешимо и нам известна оценка нормы (в пространстве C) обратного оператора:

$$\left\|A^{-1}\right\| \le B. \tag{2}$$

Последнее предположение представляется на первый взгляд чрезвычайно жестким. В его оправдание отметим следующее. Постоянная B легко находится, если ядро H мало ($\|H\| < 1$, и тогда $B = 1/(1-\|H\|)$). Для получения (2) можно также использовать способ, указанный в *§ 1* и основанный на приближении ядра H вырожденным. Наконец, оценку (2) иногда можно получить, анализируя ту систему линейных алгебраических уравнений, к которой приводит метод механических квадратур. Подробнее об этом будет сказано в конце данного параграфа.

Приближенное решение u_n^*, построенное методом механических квадратур, есть решение уравнения

$$A_n u = u - H_n u = f, \tag{3}$$

где

$$(H_n u)(x) = \sum_{k=1}^{n} p_k H(x, x_k) u(x_k),$$

x_k и p_k — узлы и веса применяемой квадратурной формулы.

Оба уравнения (1) и (3) мы рассматриваем в пространстве C непрерывных функций. Применяя тот же вариант общей схемы, который использовался при доказательстве сходимости метода механических квадратур, мы имеем согласно теореме о сходимости каркасов приближенных решений (и эта оценка уже отмечалась в *§ 2*)

$$\left\|u^* - u_n^*\right\| \le \left\|A_n^{-1}\right\| \cdot \gamma_n(u^*). \tag{4}$$

Оценим сначала меру аппроксимации $\gamma_n(u^*)$, которая в нашем случае принимает вид $\gamma_n(u^*) = \left\|(H - H_n)u^*\right\|$. Для оценки этой величины, в которую входит неизвестная нам функция u^*, удобно воспользоваться равенством $u^* = Hu^* + f$. Тогда

$$\gamma_n(u) \le \left\|(H - H_n)H\right\| \cdot \left\|u^*\right\| + \left\|(H - H_n)f\right\|.$$

Здесь $\|u^*\| \le B\|f\|$, а остальные величины, входящие в правую часть этой оценки, могут быть вычислены. Действительно, поскольку

$$(H^2u)(x) = \int_0^1 H(x, z)\int_0^1 H(z, y)u(y)dydz = \int_0^1\int_0^1 H(x,z)H(z, y)dz \cdot u(y)dy,$$

$$(H_nHu)(x) = \sum_{k=1}^n p_k H(x, x_k)\int_0^1 H(x_k, y)u(y)dy =$$

$$= \int_0^1 \sum_{k=1}^n p_k H(x, x_k)H(x_k, y)u(y)dy,$$

то $(H - H_n)H$ есть интегральный оператор

$$(H - H_n)Hu = \int_0^1 E(x, y)u(y)dy,$$

ядро которого E задается формулой

$$E(x, y) = \int_0^1 H(x, z)H(z, y)dz - \sum_{k=1}^n p_k H(x, x_k)H(x_k, y), \tag{5}$$

Поэтому если при всех $x \in [0, 1]$ выполняется оценка

$$\int_0^1 |E(x, y)|dy \le \varepsilon_1, \tag{6}$$

то $\|(H - H_n)H\| \le \varepsilon_1$. Далее, так как

$$(H - H_n)f = \int_0^1 H(x, y)f(y)dy - \sum_{k=1}^n p_k H(x, x_k)f(x_k),$$

то $\|(H - H_n)f\| \le \varepsilon_f$, если при всех $x \in [0, 1]$

$$\left|\int_0^1 H(x, y)f(y)dy - \sum_{k=1}^n p_k H(x, x_k)f(x_k)\right| \le \varepsilon_f. \tag{7}$$

Итак, при выполнении неравенств (6) и (7)

$$\gamma_n(u^*) \le B\|f\| \cdot \varepsilon_1 + \varepsilon_f. \tag{8}$$

Для оценки $\|A_n^{-1}\|$ мы можем использовать теорему о близких уравнениях второго рода (из *§ 2*) согласно которой $\|A_n^{-1}\| \le M/(1 - q)$, если только $q < 1$.

Здесь M и q — оценки величин:

$$\left\|A^{-1}\right\| \cdot \left\|(H - H_n)H_n\right\| \le q,$$

$$1 + \left\|A^{-1}\right\| \cdot \left\|H_n\right\| \le M.$$

Оценка (2) нормы $\left\|A^{-1}\right\|$ предполагается известной. Вычислим норму оператора $(H - H_n)H_n$

$$(HH_n u)(x) = \int_0^1 H(x, z) \sum_{k=1}^n p_k H(z, x_k) u(x_k)\, dz,$$

$$(H_n^2 u)(x) = \sum_{j=1}^n p_j H(x, x_j) \sum_{k=1}^n p_k H(x_j, x_k) u(x_k).$$

Отсюда

$$\begin{aligned}&\big((H - H_n)H_n u\big)(x) = \\ &= \sum_{k=1}^n p_k \left[\int_0^1 H(x, z)H(z, x_k) - \sum_{j=1}^n p_j H(x, x_j)H(x_j, x_k) \right] u(x_k) = \\ &= \sum_{k=1}^n p_k E(x, x_k) u(x_k).\end{aligned}$$

Очевидно, что

$$\left\|(H - H_n)H_n\right\| = \max_x \sum_{k=1}^n \left| p_k E(x, x_k) \right|.$$

Поэтому если при всех $x \in [0, 1]$ выполняется оценка

$$\sum_{k=1}^n \left| p_k E(x, x_k) \right| \le \varepsilon_2, \tag{9}$$

то $\left\|(H - H_n)H_n\right\| \le \varepsilon_2$.

Норма оператора H_n оценивается просто. Если при всех $x \in [0, 1]$

$$\sum_{k=1}^n \left| p_k H(x, x_k) \right| \le K, \tag{10}$$

то и $\left\|H_n\right\| \le K$. Итак, в предположении, что $B\varepsilon_2 < 1$,

$$\left\|A_n^{-1}\right\| \le \frac{1 + BK}{1 - B\varepsilon_2}. \tag{11}$$

Для получения оценки погрешности $\left\| u^* - u_n^* \right\|$ нам остается теперь подставить в (4) оценки (11) и (8). Этим доказана следующая теорема.

Теорема 1

Пусть для ядра H интегрального уравнения (1) и его правой части f выполнены оценки (2), (6), (7), (9) и (10), причем $B\varepsilon_2 < 1$. Тогда верна следующая оценка погрешности приближенного решения u_n^*, построенного методом механических квадратур: при всех $x \in [0, 1]$

$$\left| u^*(x) - u_n^*(x) \right| \le \frac{(1 + BK)(B\left\| f \right\| \cdot \varepsilon_1 + \varepsilon_f)}{1 - B\varepsilon_2}. \qquad (12)$$

- **Замечание 1.** Ядро $E(x, y)$ имеет следующий простой смысл. Это — погрешность, которая возникает при замене интеграла

$$\int_0^1 H(x, z)H(z, y)dz$$

квадратурной суммой применяемой квадратурной формулы. Разумеется, если нам известна оценка $\left| E(x, y) \right| \le \varepsilon$ $(x, y \in [0, 1])$ и если наша квадратурная формула такова, что при всех k $p_k > 0$ и $\sum_{k=1}^{n} p_k = 1$, то в оценке (12) величины ε_1 и ε_2 можно заменить на ε, если только $B\varepsilon < 1$.

- **Замечание 2.** Оценка (12) имеет априорный характер — ее правая часть может быть вычислена до того, как построено само приближенное решение u_n^*.

В неравенствах (6), (7), (9) и (10), используемых в теореме 1, речь идет об оценке известных функций. Все же вычисление правых частей неравенств (6), (7) и (9) может представлять значительные трудности. Иногда, если производные ядра H и правой части f оцениваются достаточно просто, может оказаться более удобным использовать остаточный член квадратурной формулы.

Мы будем предполагать, что веса p_k применяемой квадратурной формулы положительны (квадратурные формулы, среди весов которых имеются отри-

цательные, как правило, не применяются) и что известны такое натуральное число m и такое $r > 0$, что для всех m раз непрерывно дифференцируемых функций υ выполняется оценка

$$\left|R(\upsilon)\right| = \left|\int_0^1 \upsilon(x)dx - \sum_{k=1}^n p_k \upsilon(x_k)\right| \le r \cdot \max\left|\upsilon^{(m)}(x)\right|. \qquad (13)$$

Формула, для которой верна оценка остатка (13), заведомо точна для функции $\upsilon = 1$, и поэтому для нее $\sum_{k=1}^n p_k = 1$.

Предположим еще, что ядро H и правая часть f интегрального уравнения m раз непрерывно дифференцируемы и что нам известны оценки

$$\left|\frac{\partial^l}{\partial x^l} H(x,y)\right| \le K_l^{(1)}, \quad \left|\frac{\partial^l}{\partial y^l} H(x,y)\right| \le K_l^{(2)}, \quad \left|f^{(l)}(x)\right| \le F_l, \qquad (14)$$

где $l = 0, 1, ..., m$, $x, y \in [0, 1]$. Ядро $E(x, y)$ может быть теперь оценено с помощью формулы (13). Действительно, по формуле Лейбница

$$\frac{\partial^m}{\partial z^m}[H(x,z)H(z,y)] = \sum_{l=0}^m C_m^l \frac{\partial^l}{\partial z^l} H(x,z) \frac{\partial^{m-l}}{\partial z^{m-l}} H(z,y),$$

откуда

$$\left|\frac{\partial^m}{\partial z^m}[H(x,z)H(z,y)]\right| \le \sum_{l=0}^m C_m^l K_l^{(2)} K_{m-l}^{(1)} = N,$$

и, согласно (13),

$$\left|E(x,y)\right| = \left|\int_0^1 H(x,z)H(z,y)dz - \sum_{k=1}^n p_k H(x,x_k)H(x_k,y)\right| \le rN,$$

так что в оценке (12) можно положить (см. замечание 1) $\varepsilon_1 = \varepsilon_2 = rN$. Точно так же

$$\left|\frac{\partial^m}{\partial y^m}[H(x,y)f(y)]\right| = \left|\sum_{l=0}^m C_m^l \frac{\partial^l}{\partial y^l} H(x,y) f^{(m-l)}(y)\right| \le \sum_{l=0}^m C_m^l K_l^{(2)} F_{m-l} = F,$$

$$\left\|(H - H_n)f\right\| \le Fr,$$

и можно положить $\varepsilon_f = Fr$. Наконец, в качестве постоянной $K \geq \|H_n\|$ можно взять

$$K_0^{(1)} = K_0^{(2)}.$$

Итак, мы пришли к следующему утверждению.

Теорема 2

Пусть веса p_k применяемой квадратурной формулы положительны и для остатка этой формулы верна оценка (13). Пусть ядро H и правая часть f интегрального уравнения (1) m раз непрерывно дифференцируемы, и для их производных известны оценки (14). Пусть уравнение (1) однозначно разрешимо и известна оценка (2). Если $BNr < 1$, то для погрешности приближенного решения u_n^*, полученного методом механических квадратур, верна оценка

$$\left| u^*(x) - u_n^*(x) \right| \leq \frac{(1 + BK_0^{(1)})(BF_0N + F)r}{1 - BNr},$$

где

$$N = \sum_{l=0}^{m} C_m^l K_l^{(1)} K_{m-l}^{(2)}, \quad F = \sum C_m^l K_l^{(2)} F_{m-l}.$$

Отметим два частных случая теоремы 2, когда применяемая квадратурная формула есть формула трапеций, или Симпсона.

Для квадратурной формулы трапеций

$$\int_0^1 \upsilon(x)dx \approx \frac{h}{2}[\upsilon(0) + \upsilon(1)] + h\sum_{k=1}^{n-1} \upsilon(kh), \qquad h = \frac{1}{n},$$

остаточный член имеет вид

$$R(\upsilon) = -\frac{1}{12} h^2 \upsilon''(\xi),$$

так что в оценке (13) можно положить $m = 2$, $r = \frac{1}{12} h^2$. Это сразу же приводит нас к следующему утверждению.

Следствие 1. Пусть относительно интегрального уравнения (1) выполнены условия теоремы 2 при $m = 2$ и пусть применяемая квадратурная формула есть формула трапеций. Если $\frac{1}{12} BNh^2 < 1,$ то при всех $x \in [0, 1]$

$$\left| u^*(x) - u_n^*(x) \right| \le \frac{h^2}{12} \frac{(1 + BK_0^{(1)})(BF_0 N + F)}{1 - (1/12)BNh^2},$$

где $N = K_0^{(1)} K_2^{(2)} + 2K_1^{(1)} K_1^{(2)} + K_2^{(1)} K_0^{(2)}, \quad F = K_0^{(2)} F_2 + 2K_1^{(2)} F_1 + K_2^{(2)} F_0.$

В случае формулы Симпсона

$$\int_0^1 \upsilon(x)dx \approx \frac{h}{3}\left[\upsilon(0) + \upsilon(1) + 2\sum_{k=1}^{n-1} \upsilon(2kh) + 4\sum_{k=0}^{n-1} \upsilon\big((2k+1)h\big) \right], \quad h = \frac{1}{2n},$$

остаточный член имеет представление

$$R(\upsilon) = -\frac{1}{180} h^4 \upsilon^{(4)}(\xi).$$

Как и ранее, сразу же получаем еще одно следствие.

Следствие 2. Пусть относительно интегрального уравнения (1) выполнены условия теоремы 2 при $m = 4$ и пусть применяемая квадратурная формула есть формула Симпсона. Если $\frac{1}{180} BNh^4 < 1,$ то при всех $x \in [0, 1]$

$$\left| u^*(x) - u_n^*(x) \right| \le \frac{h^4}{180} \frac{(1 + BK_0^{(1)})(BF_0 N + F)}{1 - (1/180)BNh^2},$$

где

$$N = K_0^{(1)} K_4^{(2)} + 4K_1^{(1)} K_3^{(2)} + 6K_2^{(1)} K_2^{(2)} + 4K_3^{(1)} K_1^{(2)} + K_4^{(1)} K_0^{(2)},$$

$$F = K_0^{(2)} F_4 + 4K_1^{(2)} F_3 + 6K_2^{(2)} F_2 + 4K_3^{(2)} F_1 + K_4^{(2)} F_0.$$

Теперь посмотрим, как можно получить оценку нормы обратного оператора $\left\| A^{-1} \right\|$, используя информацию, полученную в процессе применения к интегральному уравнению (1) метода механических квадратур. Напомним, что

решение уравнения (3) сводится к решению системы линейных алгебраических уравнений

$$\zeta_j - \sum_{k=1}^{n} p_k H(x_j, x_k)\zeta_k = f(x_k),$$

записываемой в матричной форме

$$\overline{A}_n z_n = g_n. \tag{15}$$

Будем предполагать, что матрица $\overline{A}_n$ обратима, и что нам известна оценка нормы обратной матрицы:

$$\left\| \overline{A}_n^{-1} \right\|_\infty \le B_n. \tag{16}$$

Теорема 3

Пусть система уравнений (15) метода механических квадратур однозначно разрешима, и выполняется оценка (16). Пусть для ядра H интегрального уравнения выполняются оценки (6) и (10) и пусть

$$\int_0^1 \left| H(x, y) \right| dy \le L$$

при всех $x \in [0, 1]$. Тогда если $(1 + KB_n)\varepsilon_1 < 1$, то интегральное уравнение (1) однозначно разрешимо и

$$\left\| A^{-1} \right\| \le \frac{1 + (1 + KB_n)L}{1 - (1 + KB_n)\varepsilon_1}. \tag{17}$$

Доказательство

Напомним, что однозначная разрешимость системы уравнений (15) равносильна однозначной разрешимости уравнения (3). Решение уравнения (3) дается формулой

$$u_n^*(x) = \sum\nolimits_{k=1}^{n} p_k H(x, x_k)\zeta_k^* + f(x),$$

где ζ_k^* — компоненты решения системы уравнений (15). Учитывая, что при $k = 1, 2, \ldots, n$

$$\left\| \zeta_k^* \right\| \le \left\| z_n^* \right\|_\infty \le B_n \left\| g_n \right\|_\infty \le B_n \left\| f \right\|,$$

имеем

$$\left\| u_n^* \right\| \le \max_x \sum_{k=1}^{n} \left| p_k H(x, x_k) \right| \cdot B_n \left\| f \right\| + \left\| f \right\| \le (KB_n + 1) \left\| f \right\|.$$

Поскольку это неравенство выполняется для решения уравнения (3) при любой правой части f, то $\left\| A_n^{-1} \right\| \le KB_n + 1$.

Для доказательства существования оператора $A^{-1} = (I - H)^{-1}$ и оценки его нормы воспользуемся теперь теоремой о близких уравнениях второго рода (из § 2). Поскольку $\left\| (H - H_n) H \right\| \le \varepsilon_1$, то

$$\left\| (I - H_n)^{-1} \right\| \cdot \left\| (H - H_n) H \right\| \le (1 + KB_n) \varepsilon_1 < 1,$$

и согласно упомянутой теореме оператор A обратим и

$$\left\| A^{-1} \right\| \le \frac{1 + \left\| (1 - H_n)^{-1} \right\| \cdot \left\| H \right\|}{1 - (1 + KB_n) \varepsilon_1} \le \frac{1 + (1 + KB_n) L}{1 - (1 + KB_n) \varepsilon_1}. \blacksquare$$

❑ **Замечание 1.** Как видно из доказательства теоремы, величина $1 + KB_n$ оценивает норму $\left\| A_n^{-1} \right\|$. Можно указать точное значение этой нормы, если считать известной матрицу

$$D_n = \overline{A}_n^{-1} = \begin{pmatrix} d_{11} \ldots d_{1n} \\ \cdots\cdots\cdots \\ d_{n1} \ldots d_{nn} \end{pmatrix}.$$

Действительно,

$$(A_n^{-1} f)(x) = \sum_{k=1}^{n} p_k H(x, x_k) \zeta_k^* + f(x) =$$

$$= \sum_{k=1}^{n} p_k H(x, x_k) \sum_{j=1}^{n} d_{kj} f(x_j) + f(x) =$$

$$= \sum_{j=1}^{n} \left(\sum_{k=1}^{n} p_k H(x, x_k) d_{kj} \right) f(x_j) + f(x),$$

и потому

$$\left\| A_n^{-1} f \right\| \le \max_x \sum_{j=1}^{n} \left| \sum_{k=1}^{n} p_k d_{kj} H(x, x_k) \right| \cdot \left\| f \right\| + \left\| f \right\|,$$

$$\left\|A_n^{-1}\right\| \le \max_x \sum_{j=1}^{n} \left| \sum_{k=1}^{n} p_k d_{kj} H(x, x_k) \right| + 1 = B'_n.$$

Нетрудно было бы показать, что в действительности $\left\|A_n^{-1}\right\| = B'_n$. В оценке (17) постоянную $1 + KB_n$ можно заменить на B'_n.

❑ **Замечание 2.** В оценке (17) величину ε_1 можно заменить на $\varepsilon = \max\left|E(x, y)\right|$ или (в условиях теоремы 2) на Nr.

Приводим теперь простой пример, иллюстрирующий применение построенных оценок.

Пример

Пусть интегральное уравнение

$$Au = u - Hu = u(x) - \int_0^1 \sin(xy) \cdot u(y) dy = 1$$

решается методом механических квадратур, причем применяется квадратурная формула Гаусса с двумя узлами. Норму оператора A^{-1} в нашем примере мы легко можем оценить по теореме Банаха. Действительно,

$$\left\|H\right\| = \max_x \int_0^1 \left|\sin(xy)\right| dy = \int_0^1 \sin(y) dy = 1 - \cos(1),$$

$$\left\|A^{-1}\right\| \le 1/(1 - \left\|H\right\|) = 1/\cos(1),$$

и можем положить[1]

$$B = 1/\cos(1) \approx 1{,}851.$$

Имея в виду применить сначала для оценки погрешности теорему 2, отметим, что остаточный член квадратурной формулы Гаусса с двумя узлами для промежутка [0, 1] имеет вид

$$R(\upsilon) = r\upsilon^{(4)}(\eta),$$

$$r = \frac{1}{4320} \approx 0{,}000232.$$

---

[1] Здесь и далее все вычисления производились в действительности с бóльшим числом знаков, причем в приводимых постоянных округления всегда делались в бóльшую сторону.

Все производные ядра $H(x, y) = \sin(xy)$ легко оцениваются:

$$K_l^{(1)} = K_l^{(2)} = \begin{cases} \sin(1) & \text{при } l \text{ четном,} \\ 1 & \text{при } l \text{ нечетном,} \end{cases}$$

и поэтому

$$N = \sum_{l=0}^{4} C_4^l K_l^{(1)} K_{4-l}^{(2)} = 8(\sin^2(1) + 1) \approx 13{,}67.$$

Наконец,

$$F_0 = 1,\ F_l = 0\ (l = 1, 2, \ldots),\quad F = \sum_{l=0}^{4} C_4^l K_l^{(2)} F_{4-l} = K_4^{(2)} F_0 = \sin(1) \approx 0{,}842,$$

и по теореме 2

$$\left| u^*(x) - u_2^*(x) \right| \le 0{,}0156. \tag{18}$$

Для того чтобы подчеркнуть априорный характер оценки погрешности, мы намеренно строили ее до вычисления самого приближенного решения. Узлы применяемой квадратурной формулы Гаусса суть $x_1 \approx 0{,}21132;\ x_2 \approx 0{,}78868$, а веса $p_1 = p_2 = 0{,}5$. Система уравнений относительно значений $\zeta_i = u^*(x_i)$ $(i = 1, 2)$ имеет вид

$$\left.\begin{aligned} \zeta_1 - \frac{1}{2}\left[\sin(x_1^2)\zeta_1 + \sin(x_1 x_2)\zeta_2\right] &= 1, \\ \zeta_2 - \frac{1}{2}\left[\sin(x_1 x_2)\zeta_1 + \sin(x_2^2)\zeta_1\right] &= 1. \end{aligned}\right\}$$

Матрица коэффициентов этой системы есть

$$\overline{A}_2 \approx \begin{pmatrix} 0{,}97768 & -0{,}08295 \\ -0{,}08295 & 0{,}70867 \end{pmatrix}.$$

В дальнейшем нам потребуется обратная матрица

$$\overline{A}_2^{-1} \approx \begin{pmatrix} 1{,}03309 & 0{,}12092 \\ 0{,}12092 & 1{,}42526 \end{pmatrix}.$$

Решение системы уравнений: $\zeta_1^* \approx 1{,}15401, \quad \zeta_2^* \approx 1{,}54618$, и потому

$$u_2^*(x) \approx 1 + 0{,}57701 \cdot \sin(0{,}21132x) + 0{,}77309 \cdot \sin(0{,}78868x).$$

Максимум этого решения достигается в точке $x = 1$ и равен 1,6695, так что оценка (18) гарантирует, что относительная погрешность приближенного решения менее 1%.

Для применения теоремы 1 построим ядро

$$E(x,y) = \int_0^1 \sin(xz)\sin(yz)dz - \frac{1}{2}[\sin(x_1x)\sin(x_1y) + \sin(x_2x)\sin(x_2y)] =$$

$$= \frac{1}{2}\left[\frac{\sin(x-y)}{x-y} - \frac{\sin(x+y)}{x+y} - \sin(x_1x)\sin(x_1y) - \sin(x_2x)\sin(x_2y)\right].$$

Можно показать, что при всех $x, y \in [0, 1] \quad E(1,1) \le E(x,y) \le 0$, так что

$$\max|E(x,y)| = |E(1,1)| \approx 0{,}000962 = \varepsilon.$$

Далее

$$\varepsilon_f = \max_x \left| \int_0^1 \sin(x,y)dy - \frac{1}{2}\sin(x_1x) - \frac{1}{2}\sin(x_2x) \right| \approx 0{,}000110,$$

$$K = \max_x \frac{1}{2}\left[|\sin(x_1x)| + |\sin(x_2x)|\right] \approx 0{,}460,$$

и теорема 1 с учетом замечания 1 позволяет получить

$$\left|u^*(x) - u_2^*(x)\right| \le 0{,}0036, \tag{19}$$

что примерно в 4 раза лучше оценки (18). Дальнейшее улучшение оценки возможно, если подсчитать величины ε_1 и ε_2. В нашем случае

$$\varepsilon_1 = \max_x \int_0^1 |E(x,y)|dy = \left| \int_0^1 E(1,y)\,dy \right| \approx 0{,}00044,$$

$$\varepsilon_2 = \max_x \frac{1}{2}(|E(x,x_1)| + |E(x,x_2)| = \frac{1}{2}(|E(1,x_1)| + |E(1,x_2)| \approx 0{,}00044,$$

и непосредственное применение теоремы 1 приводит к оценке

$$\left|u^*(x) - u_2^*(x)\right| \le 0{,}0018.$$

Таким образом, относительная погрешность построенного приближенного решения менее 0,11%.

Оценим еще с помощью теоремы 3 норму $\left\|A^{-1}\right\|$. Для применения этой теоремы нам осталось только отметить, что

$$B_2 = \left\|\overline{A}_2^{-1}\right\|_\infty \approx 1{,}547,$$

$$L = \max_x \int_0^1 \sin(xy)dy = \int_0^1 \sin(y)dy = 1 - \cos(1) \approx 0{,}460,$$

и из (17) сразу же следует

$$\left\|A^{-1}\right\| \le 1{,}788,$$

что несколько лучше, чем оценка, полученная по теореме Банаха. При применении теоремы мы могли бы теперь положить $B = 1{,}788$, что привело бы к дальнейшему (впрочем, весьма незначительному) улучшению оценки погрешности. Обе полученные оценки нормы $\left\|A^{-1}\right\|$ являются вполне реалистическими, поскольку, как это видно из построенного приближенного решения, $\left\|A^{-1}\right\| > 1{,}66$.

Отметим в заключение, что построение оценки погрешности потребовало в действительности вычислений, в несколько раз бóльших, чем построение самого приближенного решения, и это явление общее.

§ 4. Итеративное решение систем уравнений метода механических квадратур

Уравнения, которыми приближенно согласно общей схеме заменяется функциональное уравнение, чаще всего представляют собой систему линейных алгебраических уравнений, или их решение сводится к решению системы таких уравнений. Эти системы часто имеют высокий порядок, и их решение само по себе может представлять существенные трудности. В этом параграфе на примере метода механических квадратур мы покажем, как может быть организовано итеративное решение таких систем, если использовать другую, более простую аппроксимацию исходного функционального уравнения.

Согласно методу механических квадратур, интегральное уравнение

$$Au = u - Hu = u(x) - \int_0^1 H(x, y)u(y)dy = f(x) \tag{1}$$

заменяется последовательностью уравнений

$$A_n u = u - H_n u = u(x) - \sum_{k=1}^{n} p_k^{(n)} H(x, x_k^{(n)})u(x_k^{(n)}) = f(x). \tag{2}$$

Для решения уравнения (2) при некотором зафиксированном n выберем еще некоторое натуральное m, вообще говоря, существенно меньшее, чем n, и перепишем уравнение (2) в виде

$$u - H_m u = H_n u - H_m u + f,$$

или, в предположении, что оператор $I - H_m$ обратим,

$$u = (I - H_m)^{-1}(H_n - H_m)u + (I - H_m)^{-1} f.$$

Для решения последнего уравнения применим метод последовательных приближений: выбрав в качестве начального приближения некоторую функцию u^0, последующие приближения будем строить по формулам

$$u^{\nu+1} = (I - H_m)^{-1}(H_n - H_m)u^{\nu} + (I - H_m)^{-1} f.$$

Или, что то же самое, из уравнений

$$u^{\nu+1} - H_m u^{\nu+1} = (H_n - H_m)u^{\nu} + f. \tag{3}$$

Смысл проделанного преобразования состоит в том, что в то время как уравнение (2) сводится к системе линейных алгебраических уравнений порядка n, уравнение (3) сводится к системе уравнений порядка m. Таким образом, на каждом шаге итеративного процесса приходится решать систему уравнений существенно меньшего порядка, чем исходная.

Остановимся подробнее на том, какие выкладки придется нам проделать, чтобы найти следующее приближение из уравнения (3). Для этого введем некоторые обозначения. Для функции υ, заданной на [0, 1], условимся обозначать через $(\upsilon)_n$ и $(\upsilon)_m$ векторы порядков n и m соответственно, составленные из значений этой функции в узлах квадратурных формул $x_k^{(n)}$ и $x_k^{(m)}$. При $\nu = n, m$

$$(\upsilon)_\nu = \left(\upsilon(x_1^{(\nu)}), \upsilon(x_2^{(\nu)}), ..., \upsilon(x_\nu^{(\nu)})\right).$$

Если использовать более громоздкое обозначение, которое было введено в *§ 2* при изложении второго варианта приложения общей схемы к методу механических квадратур, то можно было бы написать $(\upsilon)_n = \varphi_n \upsilon$, $(\upsilon)_m = \varphi_m \upsilon$. Введем в рассмотрение еще четыре матрицы, составленные из значений ядра $H(x, y)$ и имеющие соответственно порядки $n \times n$, $m \times n$, $n \times m$ и $m \times m$. При $\nu, \mu = n, m$

$$H_{\nu\mu} = \begin{pmatrix} p_1^{(\mu)} H\left(x_1^{(\nu)}, x_1^{(\mu)}\right) \cdots p_\mu^{(\mu)} H\left(x_1^{(\nu)}, x_\mu^{(\mu)}\right) \\ \cdots \quad \cdots \quad \cdots \\ p_1^{(\mu)} H\left(x_\nu^{(\nu)}, x_1^{(\mu)}\right) \cdots p_\mu^{(\mu)} H\left(x_n^{(\nu)}, x_\mu^{(\mu)}\right) \end{pmatrix}.$$

Отметим, что поскольку оператор H_n задается формулой

$$(H_n \upsilon)(x) = \sum_{k=1}^{n} p_k^{(n)} H(x, x_k^{(n)}) \upsilon(x_k^{(n)}),$$

то $(H_n \upsilon)_n = H_{nn}(\upsilon)_n$ и $(H_n \upsilon)_m = H_{mn}(\upsilon)_n$. Совершенно аналогично $(H_m \upsilon)_n = H_{nm}(\upsilon)_m$ и $(H_m \upsilon)_m = H_{mm}(\upsilon)_m$.

Обозначим еще через E_n и E_m единичные матрицы порядков n и m соответственно. Как об этом уже говорилось в *§ 2*, решение уравнения (2) сводится к решению системы линейных алгебраических уравнений

$$(E_n - H_{nn})(u)_n = (f)_n. \tag{4}$$

Точно так же решение уравнения (3) равносильно решению системы уравнений

$$(E_m - H_{mm})(u^{\nu+1})_m = (z^\nu)_m, \tag{5}$$

где:

$$z^\nu = (H_n - H_m)u^\nu + f;$$

$$(z^\nu)_m = (H_n u^\nu)_m - (H_m u^\nu)_m + (f)_m = H_{mn}(u^\nu)_n - H_{mm}(u^\nu)_m + (f)_m.$$

После того как вектор $(u^{\nu+1})_m$ найден из системы (5), функция $u^{\nu+1}{}_{(x)}$ находится, как обычно, по формуле

$$u^{\nu+1}{}_{(x)} = \sum_{k=1}^{m} p_k^{(m)} H(x, x_k^{(m)}) u^{\nu+1}(x_k^{(m)}) + z^\nu(x)$$

и, в частности,

$$(u^{\nu+1})_n = H_{nm}(u^{\nu+1})_m + (z^{\nu})_n =$$
$$= H_{nm}(u^{\nu+1})_m + (H_n u^{\nu})_n - (H_m u^{\nu})_m + (f)_n =$$
$$= H_{nn}(u^{\nu})_n + (f)_n + H_{nm}((u^{\nu+1})_m - (u^{\nu})_m).$$

При проведении итеративного процесса (3) нам нет надобности на каждом шаге выписывать приближение u^{ν} как функцию, заданную на [0, 1]: достаточно иметь лишь значения этой функции в узлах $x_k^{(n)}$ и $x_k^{(m)}$, т. е. векторы $(u^{\nu})_n$ и $(u^{\nu})_m$. Тем самым итеративный метод решения уравнения (2) трансформируется в итеративный метод решения системы линейных алгебраических уравнений (4).

Перечислим последовательность действий, которые требуется выполнить для перехода от приближения $(u^{\nu})_n$, $(u^{\nu})_m$ к следующему приближению $(u^{\nu+1})_n$, $(u^{\nu+1})_m$. При этом систему уравнений (5) удобнее решать не относительно $(u^{\nu+1})_m$, а относительно поправки $w^{\nu} = (u^{\nu+1})_m - (u^{\nu})_m$:

$$(E_m - H_{mm})w^{\nu} = (z^{\nu})_m - (u^{\nu})_m + H_{mm}(u^{\nu})_m = H_{mn}(u^{\nu})_n + (f)_m - (u^{\nu})_m.$$

Итак, для вычисления следующего приближения $(u^{\nu+1})_n$, $(u^{\nu+1})_m$ по предыдущему требуется:

1. Построить вектор

$$g^{\nu} = H_{mn}(u^{\nu})_n + (f)_m - (u^{\nu})_m. \tag{6}$$

2. Решить систему линейных уравнений порядка m относительно вектора w^{ν}

$$w^{\nu} - H_{mm}w^{\nu} = g^{\nu}. \tag{7}$$

3. Положить

$$\begin{aligned}(u^{\nu+1})_m &= (u^{\nu})_m + w^{\nu},\\ (u^{\nu+1})_n &= H_{nn}(u^{\nu})_n + (f)_n + H_{nm}w^{\nu}.\end{aligned} \tag{8}$$

В качестве начального приближения могут быть выбраны векторы $(u^0)_n$, $(u^0)_m$, компоненты которых — значения в соответствующих узлах произ-

вольной непрерывной функции u^0. Но наиболее естественный способ построения начального приближения заключается в том, что в качестве функции u^0 выбирается решение уравнения $u - H_m u = f$. Это приводит к формулам

$$(u^0)_m = (E_m - H_{mm})^{-1}(f)_m,$$

$$(u^0)_n = H_{nm}(u^0)_m + (f)_n.$$

Переходя к изучению сходимости изложенного процесса, обратим внимание на два обстоятельства. Во-первых, вопрос, который мы будем рассматривать, не имеет непосредственного отношения к сходимости метода механических квадратур — речь идет о сходимости итеративного метода решения уравнения (2) при фиксированном n. Во-вторых, итеративный процесс (3) решения уравнения (2) и итеративный процесс (6)—(8) решения системы линейных уравнений (4) эквивалентны: (6)—(8) есть просто некоторая реализация метода (3), и поэтому эти процессы сходятся или расходятся одновременно. При формулировке теоремы удобнее говорить о сходимости метода (6)—(8), т. е. именно того метода, который используется для непосредственного счета), а доказательство сходимости удобнее проводить в терминах процесса (3).

Теорема

Пусть выполнены условия[1]: 1) ядро H и правая часть f интегрального уравнения (1) непрерывны; 2) уравнение (1) однозначно разрешимо; 3) применяемый квадратурный процесс сходится для любой непрерывной функции.

Тогда найдется такое N, что при всех $n, m > N$ итеративный процесс (6)—(8) решения системы уравнений (4) сходится.

Доказательство

По теореме о сходимости метода механических квадратур при достаточно больших m $(m \geq N_0)$ операторы $I - H_m$ обратимы, и процесс (3) записывается в виде

$$u^{\nu+1} = D_{nm} u^{\nu} + f_0,$$

[1] Эти условия совпадают с условиями теоремы о сходимости метода механических квадратур *(см. теорему 3 из § 2)*.

где

$$D_{nm} = (I - H_m)^{-1}(H_n - H_m), \quad f_0 = (I - H_m)^{-1} f.$$

Отметим, что поскольку операторы H_n сходятся к H поточечно, но не по норме, то нормы операторов D_{nm} могут не быть малыми ни при каких n и m $(m \neq n)$. Идея доказательства заключается в том, чтобы показать малость норм операторов D_{nm}^2 при достаточно больших n и m.

Покажем, что операторы D^{nm} $(n, m \geq N_0)$ компактны в совокупности. Начнем с очевидного замечания — если множества $M_1, M_2 \subset U$ (U — нормированное пространство) относительно компактны, то и множество $M = \{u_1 + u_2 | u_1 \in M_1, u_2 \in M_2\}$ относительно компактно. Из этого замечания сразу же следует, что операторы $H_n - H_m$ компактны в совокупности, поскольку таковы в силу *леммы 1 из § 2* операторы H_n. Воспользуемся теперь тождеством

$$(I - H_m)^{-1} = I + H_m(I - H_m)^{-1},$$

которое доказывается домножением справа на $(I - H_m)^{-1}$ очевидного тождества $I = (I - H_m) + H_m$. Тогда, обозначая через S единичный шар в пространстве C, мы получаем

$$\bigcup_{n,m \geq N_0} D_{nm} S \subset M = \{u_1 + u_2 | u_1 \in M_1, u_2 \in M_2\},$$

где

$$M_1 = \bigcup_{n,m \geq N_0} (H_n - H_m)S, \quad M_2 = \bigcup_{n,m \geq N_0} H_m(I - H_m)^{-1}(H_n - H_m)S.$$

Множество M_1 относительно компактно ввиду компактности в совокупности операторов $H_n - H_m$. Множество

$$W = \bigcup_{n,m \geq N_0} (I - H_m)^{-1}(H_n - H_m)S$$

ограничено (здесь мы воспользовались равномерной ограниченностью операторов H_n и $(I - H_m)^{-1}$), и поэтому $M_2 \subset \bigcup_{m \geq N_0} H_m W$ относительно ком-

пактно. Итак, множество M, а с ним и $\bigcup D_{nm}S$, относительно компактно, и операторы D_{nm} компактны в совокупности.

Операторы H_n сходятся к оператору H *(лемма 2 из § 2)*. Эта сходимость равномерна на относительно компактном множестве $\bigcup D_{nm}S$. Поэтому[1] $\|(H_n - H)D_{nm}\| \xrightarrow[n\to\infty]{} 0$ равномерно относительно $m \geq N_0$. Совершенно аналогично $\|(H_m - H)D_{nm}\| \xrightarrow[m\to\infty]{} 0$ равномерно относительно n. Теперь имеем

$$\|D_{nm}^2\| = \|(I - H_m)^{-1}(H_n - H_m)D_{nm}\| \leq$$
$$\leq \|(I - H_m)^{-1}\|\big(\|(H_n - H)D_{nm}\| + \|(H_m - H)D_{nm}\|\big).$$

Учитывая равномерную ограниченность операторов $(I - H_m)^{-1}$ и выбирая произвольно положительное $q < 1$, мы можем теперь найти $N \geq N_0$, так чтобы при $n, m > N$ было $\|D_{nm}^2\| \leq q$.

Решение u_n^* уравнения (2) удовлетворяет равенству $u_n^* = D_{nm}u_n^* + f_0$, и потому

$$u_n^* - u^{\nu+1} = D_{nm}(u_n^* - u^{\nu}).$$

Последовательное применение этого равенства дает нам

$$u_n^* - u^{\nu} = D_{nm}^{\nu}(u_n^* - u^0).$$

Отсюда при четном $\nu = 2\sigma$ следует, что (если $n, m > N$)

$$\|u_n^* - u^{\nu}\| \leq q^{\sigma}\|u_n^* - u^0\|,$$

а при нечетном $\nu = 2\sigma + 1$

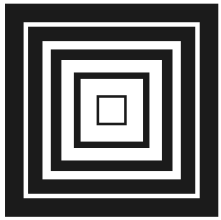
$$\|u_n^* - u^{\nu}\| \leq q^{\sigma}\|D_{nm}\| \cdot \|u_n^* - u^0\|. \quad \blacksquare$$

---

[1] Более подробно это рассуждение проводилось при доказательстве *теоремы 2 из § 2*.

- **Замечание 1.** Как видно из доказательства, итеративный процесс сходится с быстротой геометрической прогрессии, знаменатель которой сколь угодно мал, если только n и m достаточно велики.
- **Замечание 2.** Применение итеративного метода целесообразно в том случае, когда мы можем выбрать m существенно меньшим, чем n, так что решение системы уравнений (7) несравненно проще, чем (4). Поэтому полезно заметить, что теорема гарантирует существование такого m, что метод будет сходиться при всех сколь угодно больших n. При этом быстрота сходимости оценивается вне зависимости от n. Некоторый опыт применения метода показывает, что он часто быстро сходится уже при совсем малых m.
- **Замечание 3.** Реально квадратурные формулы с числом узлов n и m могут быть разных классов. Например, мы решили строить приближенное решение методом механических квадратур, выбрав формулу Симпсона с большим числом узлов, поскольку она позволяет получить сразу же таблицу приближенного решения в равноотстоящих узлах. В качестве вспомогательной квадратурной формулы, используемой для организации итеративного процесса, можно выбрать формулу Гаусса с небольшим числом узлов.

Глава 3



Проекционные методы

§ 1. Сущность проекционных методов

Проекционный метод приближенного решения линейного уравнения

$$Au = f \tag{1}$$

состоит в следующем. Выбираются система элементов $\omega_1, \omega_2, ..., \omega_n$ из области задания оператора A $(\omega_i \in D(A))$, называемая обычно *координатной системой*, и система линейных функционалов в пространстве F $\chi_1, \chi_2, ..., \chi_n$ $(\chi_i \in F^*)$, называемая *проекционной системой*. Приближенное решение u_n ищется в виде $u_n = \sum_{k=1}^{n} c_k \omega_k$, причем коэффициенты c_k находятся из условий

$$\chi_j(Au_n) = \chi_j(f), \quad j = 1, 2, ..., n, \tag{2}$$

что немедленно приводит к системе линейных алгебраических уравнений относительно этих коэффициентов:

$$\sum_{k=1}^{n} c_k \chi_j(A\omega_k) = \chi_j(f), \quad j = 1, 2, ..., n. \tag{3}$$

Основанием для такого способа построения приближенного решения служат следующие тривиальные соображения. Если бы для всех функционалов $\chi \in F^*$ оказалось $\chi(Au_n) = \chi(f)$, то это означало бы, что u_n есть точное решение уравнения (1). Однако, имея в своем распоряжении лишь n параметров — коэффициенты $c_1, c_2, ..., c_n$, мы не можем добиться выполнения этих равенств для всех χ, но мы можем рассчитывать добиться их выполнения для n выбранных функционалов χ_j.

Координатную систему $\{\omega_k\}$ и проекционную $\{\chi_j\}$ в дальнейшем мы всегда считаем линейно независимыми. Легко видеть, что приближенное решение, построенное по проекционному методу, зависит в действительности не от самих координатных элементов и проекционных функционалов, а от натянутых на них координатного подпространства $U^{(n)} \subset U$, $U^{(n)} = \left\{\sum_{k=1}^{n} \alpha_k \omega_k\right\}$, и проекционного подпространства $F^{*(n)} = \left\{\sum_{j=1}^{n} \beta_j \chi_j\right\} \subset F^*$. Действительно, сказанное ранее о том, что такое проекционный метод, может быть выражено словами: приближенное решение, построенное по проекционному методу, есть такой элемент подпространства $U^{(n)}$, что на его невязке образуются в нуль все функционалы подпространства $F^{*(n)}$.

Проекционный метод решения уравнения (1) оказывается связанным с некоторым проекционным оператором. Напомним, что линейный ограниченный оператор P_n, действующий в пространстве F, называется проекционным, проектирующим F на подпространство $F^{(n)} \subset F$, если он удовлетворяет следующим двум условиям: 1) для каждого $f \in F$ оказывается $P_n f \in F^{(n)}$; 2) если $f \in F^{(n)}$, то $P_n f = f$.

Пусть в пространстве F выбрана система элементов $g_1, g_2, ..., g_n$, такая, что определитель[1]

$$\Delta\begin{pmatrix} g_1 \cdots g_n \\ \chi_1 \cdots \chi_n \end{pmatrix} = \begin{vmatrix} \chi_1(g_1) \cdots \chi_1(g_n) \\ \cdots \quad \cdots \quad \cdots \\ \chi_n(g_1) \cdots \chi_n(g_n) \end{vmatrix}$$

отличен от нуля. Обозначим через $F^{(n)}$ подпространство пространства F, натянутое на $g_1, ..., g_n$. Каждому элементу $f \in F$ можно поставить теперь в соответствие тот единственный элемент $f_n \in F^{(n)}$, для которого при $j = 1, 2, ..., n$ выполняются равенства $\chi_j(f_n) = \chi_j(f)$. Существование и единственность такого элемента f_n сразу же следуют из однозначной разрешимости системы уравнений

$$\sum_{k=1}^{n} b_k \chi_j(g_k) = \chi_j(f)$$

[1] Подобным обозначением определителей мы будем пользоваться в дальнейшем без пояснения.

(ее определитель отличен от нуля!). Таким образом, мы определили оператор $P_n : P_n f = f_n$. Этот оператор линеен и является проекционным оператором, проектирующим F на $F^{(n)}$. Про этот проекционный оператор P_n мы будем говорить, что он определен элементами $g_1, g_2, ..., g_n$ и функционалами $\chi_1, \chi_2, ..., \chi_n$ (или подпространствами $F^{(n)}$ и $F^{*(n)}$: очевидно, что оператор P_n зависит только от подпространств $F^{(n)}$ и $F^{*(n)}$, а не от конкретно выбранных в них базисов $\{g_k\}$ и $\{\chi_k\}$).

Если f' и f'' — два элемента из F, то необходимым и достаточным условием выполнения при всех $j = 1, 2, ..., n$ равенств $\chi_j(f') = \chi_j(f'')$ является совпадение проекций $P_n f' = P_n f''$. Поэтому условия (2), которые определяют приближенное решение и из которых вытекает система уравнений (3), могут быть записаны в форме $P_n A u_n = P_n f$. Итак, приближенное решение u_n, полученное проекционным методом, есть решение уравнения

$$A_n u_n = P_n f, \tag{4}$$

где $A_n = P_n A|_{U^{(n)}}$ есть сужение оператора $P_n A$ над подпространство $U^{(n)}$. Очевидно, что A_n есть линейный ограниченный оператор, действующий из $U^{(n)}$ в $F^{(n)}$: $A_n \in L(U^{(n)}, F^{(n)})$.

До сих пор число n координатных элементов считалось фиксированным. При изучении сходимости проекционного метода естественно считать n переменным, и тогда сами координатные элементы, а равным образом и проекционные, могут зависеть от n. Тем самым мы будем считать заданными "треугольные матрицы" координатных элементов и проекционных функционалов:

$$\begin{pmatrix} \omega_1^{(1)} & & & \\ \omega_1^{(2)} & \omega_2^{(2)} & & \\ \dots & \dots & \dots & \dots \\ \omega_1^{(n)} & \omega_2^{(n)} & \dots & \omega_n^{(n)} \\ \dots & \dots & \dots & \dots \end{pmatrix}, \qquad \begin{pmatrix} \chi_1^{(1)} & & & \\ \chi_1^{(2)} & \chi_2^{(2)} & & \\ \dots & \dots & \dots & \dots \\ \chi_1^{(n)} & \chi_2^{(n)} & \dots & \chi_n^{(n)} \\ \dots & \dots & \dots & \dots \end{pmatrix},$$

где при каждом n элементы $\omega_1^{(n)}, ..., \omega_n^{(n)}$ и функционалы $\chi_1^{(n)}, ..., \chi_n^{(n)}$ линейно независимы.

Условимся еще в такой терминологии. Если заданы бесконечные линейно независимые системы

$$\omega_1, \omega_2, ..., \omega_n, ... \quad (\omega_k \in D(A))$$

и

$$\chi_1, \chi_2, ..., \chi_n, ... \quad (\chi_j \in F^*)$$

и при каждом n мы полагаем

$$\omega_k^{(n)} = \omega_k \quad (k = 1, 2, ..., n), \quad \chi_j^{(n)} = \chi_j \quad (j = 1, 2, ..., n),$$

то такой проекционный метод мы будем называть *методом моментов.*

Если $U = F$ — гильбертово пространство, то функционалы $\chi_j^{(n)}$ также являются его элементами; если при этом всегда $\chi_j^{(n)} = \omega_j^{(n)}$, то такой метод будем называть *методом Галеркина.*

Обратим внимание на один недостаток последнего определения. Согласно этому определению будет ли метод решения некоторого уравнения методом Галеркина, зависит не только от алгоритма метода, но и от выбора пары пространств (U, F), в которых мы это уравнение будем рассматривать.

При исследовании проекционного метода с помощью общей схемы *главы 1*, как это было и при исследовании метода механических квадратур, представляются наиболее естественными две возможности, на которых мы сейчас остановимся.

1. В качестве приближенного уравнения можно рассматривать уравнение (4). При этом $U_n = U^{(n)}$ есть подпространство, натянутое на координатные элементы, а $F_n = F^{(n)}$ — подпространство, натянутое на элементы $\left\{g_j^{(n)}\right\}_1^n$. Отметим, что элементы $g_j^{(n)}$ не определяются алгоритмом метода, а в известной мере находятся в распоряжении исследователя. Оператор A_n, как уже отмечалось, есть в этом случае сужение оператора $P_n A$ на $U^{(n)}$. Оператор $\psi_n \in L(F, F_n)$ должен совпадать с P_n, т. к. правая часть (4) есть $P_n f$. Поскольку само решение u_n^* уравнения (4) мы рассматриваем как приближенное решение уравнения (1) (т. е. $u^{(n)} = u_n^*$), то в качестве оператора $\overline{\varphi}_n \in L(U_n, U)$ следует выбрать оператор вложения U_n в U,

т. е. оператор, который каждому элементу $u_n \in U_n$ ставит в соответствие его же самого как элемент пространства U. Оператор $\varphi_n \in L(U, U_n)$, как обычно, находится в нашем распоряжении.

2. В качестве приближенного уравнения можно рассматривать систему линейных алгебраических уравнений (3). В этом случае U_n и F_n — n-мерные векторные пространства, норма в которых может быть выбрана разными способами. Оператор A_n определяется матрицей

$$\overline{A}_n = \begin{pmatrix} \chi_1^{(n)}(A\omega_1^{(n)}) & \cdots & \chi_1^{(n)}(A\omega_n^{(n)}) \\ \cdots & \cdots & \cdots \\ \chi_n^{(n)}(A\omega_1^{(n)}) & \cdots & \chi_n^{(n)}(A\omega_n^{(n)}) \end{pmatrix}.$$

В качестве $\psi_n \in L(F, F_n)$ следует выбрать оператор, который элементу $f \in F$ ставит в соответствие вектор $\left(\chi_1^{(n)}(f), ..., \chi_n^{(n)}(f)\right)$, а оператор $\overline{\varphi}_n$ вектору $(c_1, ..., c_n)$ должен ставить в соответствие элемент $u = \sum_{k=1}^{n} c_k \omega_k^{(n)} \in U$.

Сравним эти две схемы исследования. При изучении вопросов сходимости схема 1 оказывается более простой. Эта простота связана уже с тем, что $U_n = U^{(n)}$ есть подпространство пространства U, и потому приближенное решение совпадает со своим каркасом. Кроме того, эта схема требует фактически изучения свойств лишь подпространств $U^{(n)}, F^{(n)}$ и $F^{*(n)}$, но не свойств тех конкретных базисов $\left\{\omega_k^{(n)}\right\}$ и $\left\{\chi_j^{(n)}\right\}$, которые мы выбрали для проведения счета. Вместе с тем схема 1 малопригодна для изучения вопросов устойчивости счета. Основной вопрос, который нас здесь интересует, — вопрос о поведении систем линейных алгебраических уравнений (3). Оказывается, что при этом решающую роль играет конкретный выбор систем $\left\{\omega_k^{(n)}\right\}$ и $\left\{\chi_j^{(n)}\right\}$, т. е. базисов подпространств $U^{(n)}$ и $F^{*(n)}$ — именно эти базисы игнорируются схемой 1. Поэтому при изучении устойчивости следует обращаться к схеме 2. Нужно только помнить, что в действительности схемы 1 и 2 предназначены для изучения одного и того же метода, и результаты, полученные с помощью одной схемы, могут использоваться в другой. Далее исследование сходимости проекционного метода будет проводиться по схеме 1.

Полученные при этом утверждения позволят установить и некоторые свойства устойчивости систем уравнений (3).

Исследование проекционного метода связано с изучением проекционного оператора P_n. Введем по этому поводу некоторые обозначения и докажем некоторые вспомогательные утверждения.

Пусть $F^{(n)}$ — подпространство банахова пространства F, а $f \in F$ — произвольный элемент. Величину

$$E_{F^{(n)}}(f) = \inf_{f_n \in F^{(n)}} \|f - f_n\|$$

будем называть *наилучшим приближением элемента f элементами подпространства* $F^{(n)}$. Геометрически $E_{F^{(n)}}(f)$ означает расстояние элемента f до подпространства $F^{(n)}$.

Лемма 1

Если P_n — проекционный оператор, проектирующий пространство F на подпространство $F^{(n)}$, то для любого $f \in F$ справедлива оценка

$$\|f - P_n f\| \le (1 + \|P_n\|) E_{F^{(n)}}(f). \qquad (5)$$

Доказательство

Для любого элемента $f_n \in F^{(n)}$ будет $P_n f_n = f_n$. Поэтому

$$\|f - P_n f\| = \|f - f_n - P_n(f - f_n)\| \le (1 + \|P_n\|)\|f - f_n\|,$$

и для доказательства оценки (5) остается перейти в правой части к инфимуму по всем элементам $f_n \in F^{(n)}$. ■

Определение

Систему элементов $\{g_j^{(n)}\}$ $(n = 1, 2, \ldots, \quad j = 1, 2, \ldots, n)$ пространства F будем называть *полной*, если для всякого элемента $f \in F$ по всякому $\varepsilon > 0$ найдется такое N, что при каждом $n > N$ существует линейная комбинация $f_n = \sum_{j=1}^{n} c_j g_j^{(n)}$, приближающая f с точностью до ε: $\|f - f_n\| < \varepsilon$.

Обозначим, как это было и раньше, через $F^{(n)}$ подпространство, натянутое на элементы $g_1^{(n)}, g_2^{(n)}, ..., g_n^{(n)}$. Очевидно следующее утверждение. Для того чтобы система $\left\{g_n^{(n)}\right\}$ была полной, необходимо и достаточно, чтобы для любого элемента $f \in F$ выполнялось соотношение $E_{F^{(n)}}(f) \to 0$.

Пусть еще $\{P_n\}$ — последовательность проекционных операторов, таких, что P_n проецирует F на $F^{(n)}$. Справедливо следующее утверждение.

Лемма 2

Если система $\left\{g_j^{(n)}\right\}$ полна и нормы операторов P_n равномерно ограничены: $\|P_n\| \le C$, то для любого элемента $f \in F$ выполняется соотношение $P_n f \to f$, т. е. операторы P_n поточечно сходятся к тождественному оператору I.

Доказательство

Доказательство этой леммы непосредственно вытекает из неравенства (5). ■

Лемма 3

Пусть выполнены условия леммы 2 и пусть $T \in L(U, F)$ — компактный оператор. Тогда операторы $P_n T$ сходятся к оператору T по норме:

$$\|P_n T - T\| \to 0.$$

Доказательство

Доказательство этой леммы основано на утверждении, которое уже формулировалось в *лемме из § 2 главы 2*: если последовательность линейных операторов сходится поточечно, то эта сходимость равномерна на каждом относительно компактном множестве.

Пусть $S = \{u \,|\, u \in U, \|u\| \le 1\}$ — единичный шар в пространстве U. Его образ TS относительно компактен. Поэтому сходимость P_n к I на TS равномерна, т. е. по $\varepsilon > 0$ найдется такое N, что при $n > N$ для всех элементов $f \in TS$ выполняется неравенство $\|f - P_n f\| < \varepsilon$. Но это неравенство означает, что для

любого $u \in S$ будет $\|Tu - P_n Tu\| < \varepsilon$, и потому $\|T - P_n T\| < \varepsilon$. А это и значит, что операторы $P_n T$ сходятся к T по норме. ■

В заключение параграфа отметим, что если имеется бесконечная система $g_1, g_2, ..., g_n, ...$ элементов пространства F и если система $\left\{g_j^{(n)}\right\}$ образована по правилу $g_j^{(n)} = g_j \quad (n = 1, 2, ..., j = 1, 2, ..., n)$, то полнота системы $\left\{g_j^{(n)}\right\}$ эквивалентна полноте системы $g_1, g_2, ...$ Здесь имеется в виду следующее определение.

Определение

Система $g_1, g_2, ...$ называется *полной*, если для каждого $f \in F$ по всякому $\varepsilon > 0$ найдутся такое N и такие числа $c_1, c_2, ..., c_N$, что

$$\left\| f - \sum_{k=1}^{N} c_k g_k \right\| < \varepsilon.$$

Очевидно также, что полнота системы $g_1, g_2, ...$ означает, что множество линейных комбинаций этих элементов плотно в F.

§ 2. Теорема о сходимости

Мы будем рассматривать линейное уравнение

$$Au = f \tag{1}$$

с ограниченным оператором $A \in L(U, F)$. Предположим еще, что оператор A имеет вид суммы $A = G + T$, причем выполнены условия:

1а) оператор A имеет ограниченный обратный $A^{-1} \in L(F, U)$;

2а) оператор G имеет ограниченный обратный $G^{-1} \in L(F, U)$;

3а) оператор T компактен.

Пусть для решения уравнения (1) применяется проекционный метод с координатной системой $\left\{\omega_k^{(n)}\right\}$ и проекционной системой $\left\{\chi_j^{(n)}\right\}$. Относительно этих систем сделаем прежде всего предположение:

1б) система $\left\{\omega_k^{(n)}\right\}$ полна в пространстве U.

Это предположение равносильно следующему:

1б') система $\left\{G\omega_j^{(n)}\right\}$ полна в пространстве F.

Действительно, для произвольных $f \in F$ и $\varepsilon > 0$ мы, пользуясь полнотой системы $\left\{\omega_k^{(n)}\right\}$, можем найти такое N, что при каждом $n > N$ существует такая линейная комбинация $u_n = \sum_{k=1}^{n} c_k \omega_k^{(n)}$, что $\left\| G^{-1} f - u_n \right\| < \varepsilon_1$, где $\varepsilon_1 = \varepsilon / \|G\|$. Но тогда при $n > N$ линейная комбинация $f_n = G u_n = \sum_{k=1}^{n} c_k G\omega_k^{(n)}$ элементов $\left\{G\omega_k^{(n)}\right\}$ приближает f с точностью до ε:

$$\left\| f - f_n \right\| = \left\| G(G^{-1} f - u_n) \right\| < \|G\| \cdot \varepsilon_1 = \varepsilon.$$

Этим показано, что предположение 1б) влечет 1б'). Совершенно аналогично показывается, что из 1б') следует 1б).

Введем теперь относительно систем $\left\{\omega_k^{(n)}\right\}$ и $\left\{\chi_j^{(n)}\right\}$ еще одно предположение:

2б) при достаточно больших n определители

$$\Delta \begin{pmatrix} G\omega_1^{(n)} & G\omega_2^{(n)} & \cdots & G\omega_n^{(n)} \\ \chi_1^{(n)} & \chi_2^{(n)} & \cdots & \chi_n^{(n)} \end{pmatrix}$$

отличны от нуля и проекционные операторы P_n, определяемые элементами $G\omega_k^{(n)}$ и функционалами $\chi_j^{(n)}$, равномерно ограничены:

$$\left\| P_n \right\| \le C_1 \quad (n \ge N). \tag{2}$$

Далее, как и раньше, мы будем обозначать через $U^{(n)}$ подпространство пространства U, натянутое на координатные элементы $\omega_1^{(n)}, \omega_2^{(n)}, ..., \omega_n^{(n)}$, а через $F^{(n)}$ — подпространство пространства F, натянутое на $G\omega_1^{(n)}$, $G\omega_2^{(n)}, ..., G\omega_n^{(n)}$. Ясно, что $F^{(n)} = G U^{(n)}$ и $U^{(n)} = G^{-1} F^{(n)}$.

Докажем теперь основную теорему.

Теорема 1 (Л. В. Канторович)

Пусть уравнение (1) удовлетворяет условиям 1а)—3а), а для координатной и проекционной систем выполнены условия 1б) и 2б). Тогда:

1) при достаточно больших n системы линейных алгебраических уравнений, к которым приводит проекционный метод, однозначно разрешимы, и операторы

$$A_n^{-1} = \left(P_n A \Big|_{U^{(n)}} \right)^{-1} \in L\left(F^{(n)}, U^{(n)} \right)$$

равномерно ограничены

2) приближенные решения u_n^*, построенные проекционным методом, сходятся к точному решению u^*, причем быстрота сходимости характеризуется неравенством

$$\left\| u_n^* - u^* \right\| \le C E_{U^{(n)}}(u^*).$$

Доказательство

Доказательство теоремы основано на применении схемы 1, изложенной в *§ 1 данной главы*. Введем в рассмотрение операторы $\widetilde{A}_n = G - P_n T \in L(U, F)$. Согласно лемме 3 из *§ 1 данной главы* $\left\| A - \widetilde{A}_n \right\| = \left\| T - P_n T \right\| \to 0$. Поэтому найдется такое N, что при $n > N$ окажется $\left\| A^{-1} \right\| \cdot \left\| A - \widetilde{A}_n \right\| < 1/2$. При таких значениях n, воспользовавшись теоремой о близких уравнениях *(теорема 1 из § 1 главы 1)*, можно утверждать существование обратных операторов $\widetilde{A}_n^1$, причем для них окажется выполненной оценка

$$\left\| \widetilde{A}_n^{-1} \right\| \le \frac{\left\| A^{-1} \right\|}{1 - \left\| A^{-1} \right\| \cdot \left\| A - \widetilde{A}_n \right\|} < 2 \left\| A^{-1} \right\| = c'.$$

Обратимся теперь к уравнению

$$A_n u_n = P_n f, \tag{3}$$

где $A_n = P_n A \Big|_{U^{(n)}} = (G + P_n T) \Big|_{U^{(n)}}$.

При написании последнего равенства мы воспользовались тем, что при $u_n \in U^{(n)}$ будет $Gu_n \in F^{(n)}$, и потому $P_n G u_n = G u_n$. Таким образом, на подпространстве $U^{(n)}$, где определен оператор A_n, он принимает те же значения, что и оператор $\widetilde{A}_n$. Кроме того, если $\widetilde{u}$ есть решение уравнения $\widetilde{A}_n u = f_n$, где $f_n \in F^{(n)}$, то из формулы $\widetilde{u} = G^{-1} f_n - G^{-1} P_n T \widetilde{u}$ видно, что $\widetilde{u} \in U^{(n)}$. Таким образом, оператор A_n обратим и его обратный совпадает с сужением оператора $\widetilde{A}_n^{-1}$ на подпространство $F^{(n)} : A_n^{-1} = \widetilde{A}_n^{-1}\big|_{F^{(n)}}$. Отсюда следует, что при рассматриваемых нами достаточно больших n выполняется неравенство $\left\|A_n^{-1}\right\| \le \left\|\widetilde{A}_n^{-1}\right\| \le c'$. Система уравнений, к которой приводит проекционный метод, равносильна уравнению (3). Поэтому первое утверждение теоремы доказано.

Перейдем к доказательству второго утверждения. Для приближенного решения u_n^* имеем:

$$u_n^* = A_n^{-1} P_n f = A_n^{-1} P_n A u^* = Q_n u^*,$$

где оператор $Q_n = A_n^{-1} P_n A$ и поэтому $\|Q_n\| \le c' C_1 \|A\| = C_2$. Для произвольного элемента $u_n \in U^{(n)}$ имеем $Q_n u_n = A_n^{-1} A_n u_n = u_n$, так что Q_n — проекционный оператор, проектирующий U на $U^{(n)}$, и согласно лемме 1 из предыдущего параграфа

$$\left\|u^* - u_n^*\right\| = \left\|u^* - Q_n u^*\right\| \le \left(1 + \|Q_n\|\right) E_{U^{(n)}}\left(u^*\right) \le \left(1 + C_2\right) E_{U^{(n)}}\left(u^*\right).$$

Этим доказана требуемая оценка, из которой ввиду полноты системы $\left\{\omega_k^{(n)}\right\}$ вытекает сходимость u_n^* к u^*. ■

- **Замечание 1.** Второе утверждение теоремы нетрудно также доказать, используя общую теорему о сходимости приближенных решений, если определить операторы φ_n в общей схеме по формуле $\varphi_n = G^{-1} P_n G$.

- **Замечание 2.** Неравенство $\left\|u^* - u_n^*\right\| = c E_{U^{(n)}}(u^*)$, грубо говоря, означает, что приближенные решения сходятся к точному с такой быстротой, с ка-

кой вообще это точное решение может быть приближено элементами подпространства $U^{(n)}$.

- **Замечание 3.** Из условий теоремы предположение 2б) оказывается довольно жестким. Именно оно диктует обычно выбор метрики пространств U и F при приложении этой теоремы к конкретным проекционным методам решения дифференциальных и интегральных уравнений.

- **Замечание 4.** Отметим коротко некоторые возможные обобщения теоремы о сходимости. Во-первых, можно отказаться от требования равномерной ограниченности проекционных операторов P_n, если наложить более жесткое требование на оператор T. Именно, если положить $\eta_n = \sup\limits_{\|u\|=1} E_{F^{(n)}}(Tu)$, то ввиду компактности оператора T и полноты системы $\left\{\omega_k^{(n)}\right\}$ окажется, что $\eta_n \to 0$. Если это стремление к нулю происходит настолько быстро, что $\eta_n \|P_n\| \to 0$, то, как и раньше, $\|P_n T - T\| \to 0$, и поэтому все основные рассуждения, проведенные при доказательстве теоремы, останутся в силе (при этом утверждение о быстроте сходимости будет выглядеть, разумеется, иначе, и самый факт сходимости доказывается в предположении, что $E_{U^{(n)}}(u^*)$ стремится к нулю достаточно быстро). Во-вторых, можно рассматривать не только проекционные методы, но и методы, в некотором смысле им близкие. Именно можно считать, что уравнение (1) заменяется последовательностью уравнений $\overline{A}_n u = P_n f$, где операторы $\overline{A}_n$ действуют из $U^{(n)}$ в $F^{(n)}$, причем $A_n = G\big|_{U^{(n)}} + \overline{T}_n$ и операторы $\overline{T}_n$ близки к $P_n T\big|_{U^{(n)}}$. Идеи, которые здесь придется использовать, ясны, а в подробности мы вдаваться не будем. Отметим, что теорема о сходимости, учитывающая оба этих обобщения, имеется в общей теории приближенных методов Л. В. Канторовича.

Обратимся теперь к вопросам устойчивости проекционных методов. При этом мы естественно будем использовать схему 2, изложенную в предыдущем параграфе. Таким образом, у нас U_n и F_n — n-мерные векторные пространства, и решаемое приближенное уравнение есть

$$\overline{A}_n z_n = b_n, \tag{4}$$

где оператор $\overline{A}_n$ задается матрицей

$$\overline{A}_n = \begin{pmatrix} \chi_1^{(n)}(A\omega_1^{(n)}) & \cdots & \chi_1^{(n)}(A\omega_n^{(n)}) \\ \cdots & \cdots & \cdots \\ \chi_n^{(n)}(A\omega_1^{(n)}) & \cdots & \chi_n^{(n)}(A\omega_n^{(n)}) \end{pmatrix},$$

а $b_n = \psi_n f = \left(\chi_1^{(n)}(f), ..., \chi_1^{(n)}(f)\right)$. Сделаем еще некоторые предположения. Будем считать, что пространства U и F гильбертовы, и тем самым $\chi_j^{(n)} \in F$. Наложим на координатную и проекционную системы еще такие ограничения:

3б) координатные элементы $\omega_k^{(n)}$ образуют в $U^{(n)}$ ортонормальный базис:

$$(\omega_k^{(n)}, \omega_j^{(n)}) = \delta_{kj};$$

4б) в подпространстве $F^{(n)} = GU^{(n)}$ имеется ортонормальный базис

$$\left\{g_j^{(n)}\right\} \quad \left((g_k^{(n)}, g_j^{(n)}) = \delta_{kj}\right),$$

такой, что $(g_k^{(n)}, \chi_j^{(n)}) = \delta_{kj}$.

Последнее условие означает, что элементы $\chi_j^{(n)}$ могут отличаться от элементов некоторого ортонормального базиса в $F^{(n)}$ лишь на слагаемые, ортогональные всему $F^{(n)}$.

Теорема 2

Пусть пространства U и F гильбертовы, для уравнения (1) выполнены условия 1а) — 3а), а для координатной и проекционной систем — условия 1б) — 4б). Пусть в векторных пространствах U_n и F_n введена эвклидова норма[1]. Тогда для проекционного метода процессы отыскания каркасов и построения приближенных решений устойчивы.

---

[1] То есть $U_n = F_n = l_n^2$.

Доказательство

Введем в рассмотрение оператор ψ_n^0 — сужение оператора ψ_n на подпространство $F^{(n)}$. Тогда для $f_n = \sum_{j=1}^n b_j g_j^{(n)}$ будет (в силу 4б))

$$\psi_n^0 f_n = \left((\chi_1^{(n)}, f_n), ..., (\chi_n^{(n)}, f_n)\right) = (b_1, b_2, ..., b_n).$$

Учитывая, что $g_j^{(n)}$ — ортонормальный базис $F^{(n)}$, так что $\|f_n\|^2 = \sum_{j=1}^n b_j^2$, видим, что ψ_n^0 есть линейная изометрия между $F^{(n)}$ и F_n. Точно так же, если обозначить через $\overline{\varphi}_n^0$ тот же оператор $\overline{\varphi}_n$, если его значения рассматривать как элементы пространства $U^{(n)}$ ($\overline{\varphi}_n^0 \in L(U_n, U^{(n)})$ в отличие от $\overline{\varphi}_n \in L(U_n, U)$), то для $z_n = (c_1, c_2, ..., c_n)$ будет $\overline{\varphi}_n^0 z_n = \sum_{k=1}^n c_k \omega_k^{(n)}$, и ввиду 3б) $\|z_n\|_2 = \left\|\overline{\varphi}_n^0 z_n\right\|$, так что $\overline{\varphi}_n^0$ — линейная изометрия между U_n и $U^{(n)}$. Ясно также, что $\|\overline{\varphi}_n\| = 1$. Итак, у нас

$$\left\|\psi_n^0\right\| = \left\|(\psi_n^0)^{-1}\right\| = \|\overline{\varphi}_n\| = \left\|\overline{\varphi}_n^0\right\| = \left\|(\overline{\varphi}_n^0)^{-1}\right\| = 1.$$

Справедливо равенство $\overline{A}_n = \psi_n^0 A_n \overline{\varphi}_n^0$, где, как и раньше, $A_n = P_n A\big|_{U^{(n)}}$. Действительно, для произвольного вектора $z_n = (c_1, c_2, ..., c_n)$ имеем

$$\overline{\varphi}_n^0 z_n = \sum_{k=1}^n c_k \omega_k^{(n)};$$

$$A\overline{\varphi}_n^0 z_n = \sum_{k=1}^n c_k A\omega_k^{(n)};$$

$$A_n \overline{\varphi}_n^0 z_n = (P_n A)\overline{\varphi}_n^0 z_n = P_n(A\overline{\varphi}_n^0 z_n);$$

$$\psi_n^0 A\overline{\varphi}_n^0 z_n = (\beta_1, \beta_2, ..., \beta_n) = b_n,$$

где $\beta_j = (\chi_j^{(n)}, A\overline{\varphi}_n^0 z_n) = \sum_{k=1}^n (\chi_j^{(n)}, A\omega_k^{(n)} c_k)$,

а это и значит, что $b_n = \psi_n^0 A_n \overline{\varphi}_n^0 z_n = \overline{A}_n z_n$. Теперь сразу же имеем

$$\left\|\overline{A}_n\right\| = \left\|\psi_n^0 A_n \overline{\varphi}_n^0\right\| = \|A_n\| \le \|P_n\| \cdot \|A\| \le c_1 \|A_n\|;$$

$$\left\|\overline{A}_n^{-1}\right\| = \left\|(\overline{\varphi}_n^0)^{-1} A_n^{-1} (\psi_n^0)^{-1}\right\| = \left\|A_n^{-1}\right\| \le c.$$

Последнее неравенство выполняется при достаточно больших n согласно первому утверждению теоремы 1. Итак,

$$\mu(\overline{A}_n) = \left\|\overline{A}_n\right\| \cdot \left\|\overline{A}_n^{-1}\right\| \le c c_1 \|A\|,$$

и согласно *теореме 1* из *§ 3 главы 1* процесс отыскания каркасов устойчив.

Нормы каркасов приближенных решений z_n^* совпадают с нормами самих приближенных решений (ввиду условия 3б)). Из сходимости приближенных решений к точному следует, что нормы приближенных решений, а тем самым и каркасов z_n^*, ограничены. Поэтому $\left\|\overline{\varphi}_n\right\| \cdot \left\|z_n^*\right\| = \left\|z_n^*\right\| \le c$. Поскольку выполнены и все остальные условия *теоремы 2* из *§ 3 главы 1*, то по этой теореме процесс построения приближенных решений устойчив. ■

Чтобы еще раз подчеркнуть ту мысль, которая была высказана в конце *§ 3 главы 1* по поводу нормировки пространств U_n и F_n, отметим следующее. Утверждение доказанной теоремы вполне содержательно. Оно ориентирует нас, в какой метрике (именно, эвклидовой) должны быть малы ошибки, которые мы совершаем при вычислении элементов матрицы $\overline{A}_n$ и правых частей $(\chi_j^{(n)}, f)$, чтобы ошибки в приближенных решениях оказались того же порядка малости.

§ 3. Метод Галеркина для уравнений второго рода

Уравнением второго рода по аналогии с интегральным уравнением Фредгольма второго рода мы будем называть сейчас уравнение вида

$$Au = u + Tu = f, \tag{1}$$

где T — компактный оператор, действующий в гильбертовом пространстве H, $f \in H$. Метод Галеркина решения уравнения (1), как уже говорилось в *§ 1*, состоит в том, что приближенное решение ищется в виде $u_n = \sum_{k=1}^{n} c_k \omega_k^{(n)}$, где $\left\{\omega_k^{(n)}\right\}$ — выбранная координатная система ($\omega_k^{(n)} \in H$, и

при каждом n элементы $\omega_1^{(n)}, \omega_2^{(n)}, ..., \omega_n^{(n)}$ линейно независимы). Коэффициенты c_k находятся из условий

$$(Au_n, \omega_j^{(n)}) = (f, \omega_j^{(n)}), \quad j = 1, 2, ..., n,$$

т. е. из системы линейных алгебраических уравнений

$$\sum_{k=1}^{n} (\omega_k^{(n)} + T\omega_k^{(n)}, \omega_j^{(n)}) c_k = (f, \omega_j^{(n)}), \quad j = 1, 2, ..., n. \tag{2}$$

Из теоремы Л. В. Канторовича (*§ 1*) немедленно вытекает следующая теорема о сходимости метода Галеркина.

Теорема 1

Пусть существует обратный оператор A^{-1} и пусть координатная система $\{\omega_k^{(n)}\}$ полна в H. Тогда при достаточно больших n системы уравнений (2) однозначно разрешимы и приближенные решения u_n^*, построенные по методу Галеркина, сходятся к точному.

Доказательство

Доказательство этой теоремы сводится к проверке условий 1а)—3а) и 1б)—2б) теоремы Канторовича. При этом мы считаем, что $G = I$ есть тождественный оператор в пространстве H, и потому условие 2а) (существование ограниченного обратного G^{-1}) заведомо выполнено. Выполнение условий 1а) (существование ограниченного обратного A^{-1}) и 3а) (компактность оператора T) было нами предположено. Условие 1б) (полнота системы $\{\omega_k^{(n)}\}$) также является условием доказываемой теоремы. Перейдем к проверке условия 2б). Определитель, о котором идет речь в этом условии, в случае метода Галеркина и оператора $G = I$ принимает вид

$$\Delta \begin{pmatrix} \omega_1^{(n)} & \omega_2^{(n)} & \cdots & \omega_n^{(n)} \\ \omega_1^{(n)} & \omega_2^{(n)} & \cdots & \omega_n^{(n)} \end{pmatrix} = \det \{(\omega_k^{(n)}, \omega_j^{(n)})\}$$

и при всех n отличен от нуля как определитель Грама линейно независимой системы $\omega_1^{(n)}, \omega_2^{(n)}, ..., \omega_n^{(n)}$. В нашем случае $F^{(n)} = GU^{(n)} = U^{(n)}$ есть подпространство H, натянутое на координатные элементы $\omega_k^{(n)}$. Проекционный опе-

ратор P_n, определяемый координатной системой $\{\omega_k^{(n)}\}$ и проекционной системой, совпадающей с координатной (метод Галеркина!), ставит в соответствие элементу $f \in H$ такой элемент $f_n \in F^{(n)}$, что $(f, \omega_j^{(n)}) = (f_n, \omega_j^{(n)})$ при $j = 1, 2, ..., n$, т. е. разность $f - f_n$ ортогональна всему подпространству $F^{(n)}$. Таким образом, P_n есть оператор ортогонального проектирования гильбертова пространства H на подпространство $F^{(n)}$, и потому $\|P_n\| = 1$. Этим проверка условий теоремы Канторовича завершена, и остается воспользоваться ее заключением. ■

❑ **Замечание.** При формулировке доказанной теоремы мы ради краткости опустили два важных утверждения, которые также немедленно вытекают из теоремы Канторовича. Во-первых, при всех достаточно больших n существуют обратные операторы $\|A_n^{-1}\| = (P_n A|_{U^{(n)}})^{-1}$ (P_n — оператор ортогонального проектирования на $U^{(n)}$), и эти операторы равномерно ограничены. Во-вторых, быстрота сходимости метода Галеркина характеризуется неравенством

$$\|u^* - u_n^*\| = O\left(E_{U^{(n)}}(u^*)\right).$$

Если предположить, что при каждом n система $\{\omega_k^{(n)}\}$ ортонормальна, т. е. что выполнено условие 3б), наложенное в предыдущем параграфе на координатную систему, то тем самым окажется выполненным и условие 4б). Это позволяет, используя теорему 2 из предыдущего параграфа, доказать устойчивость метода Галеркина[1] в предположении ортонормальности координатной системы. Однако сейчас будет доказано более общее утверждение. Пусть по-прежнему $\overline{A}_n$ есть матрица коэффициентов системы уравнений (2) и пусть Γ_n — матрица Грама координатной системы:

$$\Gamma_n = \begin{pmatrix} (\omega_1^{(n)}, \omega_1^{(n)}) & \cdots & (\omega_1^{(n)}, \omega_n^{(n)}) \\ \cdots & \cdots & \cdots \\ (\omega_n^{(n)}, \omega_1^{(n)}) & \cdots & (\omega_n^{(n)}, \omega_n^{(n)}) \end{pmatrix}.$$

[1] Точный смысл термина "устойчивость метода" ясен из формулировки теоремы, на которую мы только что сослались.

Теорема 2

В условиях теоремы 1 найдутся такие постоянные c_1 и c_2, что при всех достаточно больших n выполняются неравенства

$$\left\|\overline{A}_n\right\|_2 \le c_1\left\|\Gamma_n\right\|_2, \quad \left\|\overline{A}_n^{-1}\right\|_2 \le c_2\left\|\Gamma_n^{-1}\right\|_2. \tag{3}$$

Прежде чем доказывать эту теорему, напомним некоторые известные свойства положительно определенных матриц. Произвольная матрица Γ_n называется положительно определенной, если она симметрична и для любого вектора $x_n \in R^n \quad (x_n \neq 0)$ выполняется неравенство $(\Gamma_n x_n,\ x_n) > 0$. Для того чтобы симметричная матрица была положительно определенной, необходимо и достаточно, чтобы все ее собственные числа были положительны.

Для положительно определенной матрицы Γ_n существует единственная положительно определенная матрица B_n, такая, что $B_n^2 = \Gamma_n$. Действительно, матрица B_n, как всякая симметричная матрица, имеет полную систему собственных векторов. Если z_n — некоторый собственный вектор матрицы B_n, соответствующий собственному числу λ, то из равенства $B_n^2 = \Gamma_n$ вытекает, что λ^2 есть собственное число матрицы Γ_n при том же собственном векторе z_n. Учитывая положительность собственных чисел матрицы B_n, видим, что равенство $B_n^2 = \Gamma_n$ однозначно определяет все ее собственные числа и собственные векторы, а значит, и саму матрицу B_n. Этим мы показали, что может существовать не более одной положительно определенной матрицы B_n, для которой $B_n^2 = \Gamma_n$.

Для того чтобы некоторая матрица была симметричной, необходимо и достаточно, чтобы ее собственные числа были вещественны, а система собственных векторов полна и собственные векторы, отвечающие различным собственным числам, ортогональны. Поэтому если, исходя из матрицы Γ_n, мы определим матрицу B_n требованием, чтобы ее собственные векторы совпадали с собственными векторами матрицы Γ_n, а собственные числа были квадратными корнями из соответствующих собственных чисел матрицы Γ_n,

то такая матрица B_n окажется положительно определенной и будет удовлетворять равенству $B_n^2 = \Gamma_n$. В проведенном рассуждении мы дважды неявно использовали следующее очевидное утверждение. Если $x_n^{(1)}, x_n^{(2)}, ..., x_n^{(n)}$ — линейно независимые n-мерные векторы и $y_n^{(1)}, y_n^{(2)}, ..., y_n^{(n)}$ — произвольные векторы, то существует, и притом единственная матрица D_n, такая, что $D_n x_n^{(k)} = y_n^{(k)}$ $(k = 1, 2, ..., n)$. Матрица B_n называется *квадратным корнем* из матрицы Γ_n и обозначается $B_n = \Gamma_n^{1/2}$.

Из указанного способа построения матрицы $\Gamma_n^{1/2}$ видно, что матрица $\Gamma_n^{-1/2}$, определяемая равенством $\Gamma_n^{-1/2} = \left(\Gamma_n^{1/2}\right)^{-1}$, совпадает с матрицей $\left(\Gamma_n^{-1}\right)^{1/2}$. Норма положительно определенной матрицы как оператора в пространстве l_n^2 есть ее максимальное собственное число, и потому выполняется равенство $\left\|\Gamma_n^{1/2}\right\|_2^2 = \|\Gamma_n\|_2$.

Докажем простую лемму, которая будет неоднократно использоваться далее.

Лемма

Пусть $x_n = (\xi_1, \ldots, \xi_n)$ — n-мерный вектор, и пусть $u_n = \sum_{k=1}^n \xi_k \omega_k^{(k)}$. Тогда $\|u_n\|_H = \left\|\Gamma_n^{1/2} x_n\right\|_2$.

Доказательство

Доказательство вытекает из цепочки равенств:

$$\|u_n\|_H^2 = (u_n, u_n) = \sum_{k,j=1}^n (\omega_k^{(n)}, \omega_j^{(n)}) \xi_k \xi_j = (\Gamma_n x_n, x_n) =$$

$$= (\Gamma_n^{1/2} x_n, \Gamma_n^{1/2} x_n) = \left\|\Gamma_n^{1/2} x_n\right\|_2^2. \blacksquare$$

Доказательство теоремы 2

Будем считать n настолько большим, что матрица $\overline{A}_n$ обратима и что выполняется неравенство $\left\|A_n^{-1}\right\| \le c_2$. Пусть теперь $x_n = (\xi_1, ..., \xi_n)$ и $y_n = (\eta_1, ..., \eta_n)$ — произвольные векторы, связанные соотношением

$$\overline{A}_n x_n = y_n. \tag{4}$$

Наша цель — оценить норму вектора x_n через норму вектора y_n (это даст оценку нормы матрицы $\overline{A}_n^{-1}$) и, наоборот, норму y_n — через норму x_n (это позволит оценить норму матрицы $\overline{A}_n$). Построим элемент $g_n = \sum_{k=1}^n \beta_k \omega_k^{(n)}$, подобрав коэффициенты β_k из условий $(g_n, \omega_j^{(n)}) = \eta_j$, $j = 1, 2, ..., n$. Очевидно, что такие коэффициенты β_k однозначно определяются из системы уравнений $\Gamma_n b_n = y_n$, где $b_n = (\beta_1, ..., \beta_n)$, т. е. $b_n = \Gamma_n^{-1} y_n$. Построим еще элемент $u_n = \sum_{k=1}^n \xi_k \omega_k^{(n)}$. Тогда u_n есть приближенное решение уравнения $u + Tu = g_n$, построенное методом Галеркина, поскольку применение метода Галеркина к этому уравнению как раз приведет нас к системе уравнений (4). Итак, $g_n = P_n g_n = u_n + P_n T u_n$, откуда

$$\|g_n\|_H \le (1 + \|P_n\| \cdot \|T\|) \|u_n\|_H = (1 + \|T\|) \|u_n\|_H = c_1 \|u_n\|_H,$$

и

$$u_n = A_n^{-1} P_n g_n = A_n^{-1} g_n,$$

откуда

$$\|u_n\|_H \le \left\|A_n^{-1}\right\| \cdot \|g_n\|_H \le c_2 \|g_n\|_H.$$

Теперь, учитывая, что $\|g_n\|_H = \left\|\Gamma_n^{1/2} b_n\right\|_2$, легко получаем

$$\|y_n\|_2 = \|\Gamma_n b_n\|_2 = \left\|\Gamma_n^{1/2} \Gamma_n^{1/2} b_n\right\|_2 \le \left\|\Gamma_n^{1/2}\right\|_2 \cdot \|g_n\|_{\mathrm{H}} \le \left\|\Gamma_n^{1/2}\right\|_2 \cdot c_1 \|u_n\|_{\mathrm{H}} =$$

$$= c_1 \left\|\Gamma_n^{1/2}\right\|_2 \cdot \left\|\Gamma_n^{1/2} x_n\right\|_2 \le c_1 \left\|\Gamma_n^{1/2}\right\|_2^2 \cdot \|x_n\|_2 = c_1 \|\Gamma_n\|_2 \cdot \|x_n\|_2$$

(мы воспользовались тем, что $\left\|\Gamma_n^{1/2}\right\|_2^2 = \|\Gamma_n\|_2$). Так как x_n и y_n — произвольные векторы, связанные равенством (4), то

$$\left\|\overline{A}_n\right\|_2 \le c_1\|\Gamma_n\|_2,$$

и первое из неравенство (3) доказано. Далее

$$\|x_n\|_2 = \left\|\Gamma_n^{-1/2}\Gamma_n^{1/2}x_n\right\|_2 \le \left\|\Gamma_n^{-1/2}\right\|_2 \cdot \|u_n\|_{\mathrm{H}} \le c_2\left\|\Gamma_n^{-1/2}\right\|_2 \cdot \|g_n\|_{\mathrm{H}} =$$
$$= c_2\left\|\Gamma_n^{-1/2}\right\|_2 \cdot \left\|\Gamma_n^{1/2}b_n\right\|_2 = c_2\left\|\Gamma_n^{-1/2}\right\|_2 \cdot \left\|\Gamma_n^{-1/2}y_n\right\|_2 \le$$
$$\le c_2\left\|\Gamma_n^{-1/2}\right\|_2^2\|y_n\|_2 = c_2\left\|\Gamma_n^{-1}\right\|_2 \cdot \|y_n\|_2$$

и потому $\left\|A_n^{-1}\right\|_2 \le c_2\left\|\Gamma_n^{-1}\right\|_2$. ■

В отмеченном ранее случае, когда $\left\{\omega_k^{(n)}\right\}$ есть ортонормальная система, $\Gamma_n = E_n$ — единичная матрица, и из теоремы сразу же вытекает ограниченность чисел обусловленности $\mu_2(\overline{A}_n)$. Это и позволило нам назвать теорему 2 более общей, чем упомянутое утверждение об устойчивости. Оценка

$$\mu_2(\overline{A}_n) \le c_3\mu_2(\Gamma_n) \quad (c_3 = c_1c_2)$$

является очевидным следствием оценок (3). Поэтому значение доказанной теоремы состоит в том, что она позволяет заменить оценку числа обусловленности "сложной" матрицы $\overline{A}_n$ (этого требует исследование вопросов устойчивости процесса отыскания каркасов) изучением более "простой" матрицы Γ_n. Вообще из доказанной теоремы легко вытекает следующее утверждение.

Следствие. Если выполнены условия теоремы 1 и числа обусловленности матриц Грама координатной системы равномерно ограничены, то процессы отыскания каркасов и построения приближенных решений устойчивы. При этом считается, что $U_n = F_n = l_n^2$ — n-мерное векторное пространство с эвклидовой нормой.

Доказательство

Поскольку $\mu_2(A_n) \le c_3\mu_2(\Gamma_n) \le c,$ то согласно теореме об устойчивости процесса отыскания каркасов (*теорема 1* из *§ 3 главы 1*) этот процесс устойчив. Из условий теоремы об устойчивости процесса построения приближенных решений *(теорема 2 из § 3 главы 1)* нам остается проверить лишь ограниченность последовательности $\|\overline{\varphi}_n\| \cdot \|z_n^*\|,$ где z_n^* — решение системы (2). Поскольку оператор $\overline{\varphi}_n$ вектору $x_n = (\xi_1, ..., \xi_n)$ ставит в соответствие элемент $u_n = \sum_{k=1}^n \xi_k \omega_k^{(n)},$ то согласно доказанной лемме

$$\|\overline{\varphi}_n x_n\| = \|u_n\| = \left\|\Gamma_n^{1/2} x_n\right\|_2$$

и потому $\|\overline{\varphi}_n\| = \left\|\Gamma_n^{1/2}\right\|.$ В силу того же равенства

$$\left\|\Gamma^{1/2} z_n^*\right\| = \|u_n^*\| \le c',$$

т. к. $u_n^*(x)$ сходится к $u^*,$ так что

$$\|z_n^*\|_2 \le \left\|\Gamma_n^{-1/2}\right\| \cdot \left\|\Gamma_n^{1/2} z_n^*\right\| \le c' \left\|\Gamma_n^{-1/2}\right\|,$$

и окончательно имеем

$$\|\overline{\varphi}_n\| \cdot \|z_n^*\| \le \left\|\Gamma_n^{1/2}\right\| \cdot c' \left\|\Gamma_n^{-1/2}\right\| = c' \mu_2^{1/2}(\Gamma_n) \le c''. \blacksquare$$

В качестве простейшего приложения доказанных теорем рассмотрим метод Галеркина для интегрального уравнения (другое приложение будет указано в *§ 9*). Будем предполагать, что в интегральном уравнении

$$Au = (I + T)u = u(x) + \int_{-1}^{1} T(x, y)u(y)dy = f(x) \tag{5}$$

ядро T и правая часть f суммируемы с квадратом:

$$\int_{-1}^{1}\int_{-1}^{1} [T(x, y)]^2 dxdy < +\infty, \quad \int_{-1}^{1} [f(x)]^2 dx < +\infty. \tag{6}$$

Как известно, первое из этих условий обеспечивает компактность интегрального оператора T в пространстве L_2. Будем для простоты предполагать, что в методе Галеркина координатная система выбирается по принципу метода

моментов: $\omega_k^{(n)} = \omega_k$, где $\{\omega_k\}$ — бесконечная линейно независимая система суммируемых с квадратом функций. Тогда система уравнений (2) в нашем случае принимает вид

$$\sum_{k=1}^{n} a_{jk} c_k = b_j, \quad j = 1, 2, ..., n, \tag{7}$$

где

$$a_{jk} = \int_{-1}^{1} \omega_k(x)\omega_j(x)dx + \int_{-1}^{1}\int_{-1}^{1} T(x,y)\omega_j(x)\omega_k(y)dxdy, \quad b_j = \int_{-1}^{1} f(x)\omega_j(x)dx.$$

Из сказанного выше о методе Галеркина для уравнений второго рода сразу же вытекает следующее утверждение.

Теорема 3

Если интегральное уравнение (5), удовлетворяющее условиям (6), однозначно разрешимо, а координатная система $\{\omega_k\}$ полна в пространстве L_2, то при достаточно больших n системы линейных алгебраических уравнений (7) однозначно разрешимы и приближенные решения $u_n^*(x)$, построенные по методу Галеркина, сходятся к точному $u^*(x)$ в среднем, т. е. в метрике пространства L_2. Если, кроме того, при всех k и j

$$\int_{-1}^{1} \omega_k(x)\omega_j(x)dx = \delta_{kj}, \tag{8}$$

то метод Галеркина устойчив.

То, что в формулировке теоремы ради краткости было названо устойчивостью метода Галеркина, означает следующее. Если в качестве приближенных уравнений рассматривать системы уравнений (7) в метрике пространств l_n^2, то процесс отыскания каркасов приближенных решений устойчив (что равносильно равномерной ограниченности чисел обусловленности матриц $\overline{A}_n = \{a_{jk}\}_1^n$, $\mu_2(\overline{A}_n) \le C$) и процесс построения приближенных решений также устойчив (при этом возмущения приближенных решений рассматриваются в норме L_2).

Легко можно показать, что если ядро T и правая часть f рассматриваемого интегрального уравнения непрерывны, то по полученным методом Галер-

кина приближенным решениям u_n^* можно построить новые приближенные решения

$$u^{(n)}(x) = f(x) - \int_{-1}^{1} T(x, y) u_n^*(y) dy,$$

которые будут сходиться к точному решению равномерно.

В качестве полной координатной системы для решения интегрального уравнения методом Галеркина можно предложить полиноминальную, когда $\omega_k(x)$ есть полином степени в точности $k-1$. С точки зрения сходимости метода (если бы мы имели возможность все вычисления проводить точно), конкретный выбор таких полиномов безразличен, т. к. подпространство, натянутое на первые n координатных функций, все равно будет одним и тем же — это множество всех полиномов степени не выше $n-1$. С точки зрения устойчивости, вопрос о конкретном выборе полиномов ω_k имеет первостепенную важность. Если выбрать ω_k простейшим образом, положив $\omega_k(x) = x^{k-1}$, то системы уравнений (7) окажутся чрезвычайно плохо обусловленными: можно показать [24], что для такой координатной системы $\underline{\lim} \sqrt[n]{\mu_2(\Gamma_n)} \geq 4$, т. е., грубо говоря, числа обусловленности $\mu_2(\Gamma_n)$ стремятся к бесконечности по меньшей мере как общий член геометрической прогрессии со знаменателем 4. Поэтому естественно ожидать в этом случае показательную неустойчивость процесса отыскания каркасов, и такая координатная система для счета практически непригодна даже для не очень больших n. Теорема 3 показывает, что устойчивость будет гарантирована, если в качестве координатных функций мы возьмем ортогональные полиномы Лежандра

$$\omega_{k+1}(x) = P_k(x) = \frac{1}{2^k \cdot k!} \sqrt{\frac{2k+1}{2}} \frac{d^k}{dx^k} (x^2 - 1)^k,$$

удовлетворяющие условию (8).

Остановимся без доказательства еще на некоторых свойствах метода Галеркина для интегральных уравнений. Если координатная система полиноминальна, а ядро T и правая часть f — достаточно гладкие функции (достаточно потребовать, чтобы они были непрерывно дифференцируемы; это требование в действительности можно существенно ослабить), то приближенные решения, построенные по методу Галеркина, будут сходиться к точному не только

в среднем, но и равномерно. Если при этом само точное решение ν раз непрерывно дифференцируемо, то быстрота сходимости характеризуется соотношением

$$\left\| u^* - u_n^* \right\|_C = O(1/n^{\nu-1/2}).$$

Таким образом, если решение имеет много производных, то приближенные решения сходятся очень быстро. Этот результат может быть получен на основе того обобщения теоремы Канторовича, которое отмечалось в замечании 4 в *§ 2*. Подчеркнем, что быстрота сходимости метода Галеркина определяется свойствами гладкости точного решения, но не ядра интегрального уравнения. Это свойство отличает метод Галеркина от метода механических квадратур, в котором для быстрой сходимости существенно, чтобы быстро убывала погрешность замены интеграла от функции $T(x,\cdot)u^*$ квадратурной суммой, а это обычно требует достаточной гладкости не только решения u^*, но и ядра как функции y. Отметим, что интегральные уравнения, которые при ядре малой гладкости (функции Грина) имеют очень гладкие решения, если такова правая часть, встречаются довольно часто. Так бывает при замене дифференциальных уравнений интегральными про помощи функций Грина. Поэтому указанное преимущество метода Галеркина в сравнении с методом механических квадратур является существенным. Основной недостаток метода Галеркина состоит в том, что система линейных алгебраических уравнений (7) строится сложно, т. к. требует вычисления ряда интегралов, в то время как метод механических квадратур обходится лишь вычислением значений ядра и правой части в ряде точек.

Отметим еще, что если интегральное уравнение (5) мы рассматриваем в пространстве C непрерывных функций, но по-прежнему считаем $U_n = F_n = l_n^2$, то процесс отыскания каркасов приближенных решений при использовании в качестве координатных функций полиномов Лежандра будет устойчив (устойчивость этого процесса зависит лишь от выбора норм в пространствах U_n и F_n, но не от норм в U и F). Процесс же построения приближенных решений устойчивым уже не будет, т. к. окажется, что $\left\| \overline{\varphi}_n \right\| \to \infty$. Однако неустойчивость эта слабая (степенная), т. к. выполняется оценка $\left\| \overline{\varphi}_n \right\| \le cn$, так что практически счет можно вести с небольшим числом запасных знаков.

§ 4. Метод моментов для обыкновенных дифференциальных уравнений

Рассмотрим линейное обыкновенное дифференциальное уравнение

$$L(u) = u^{(m)} + a_1 u^{(m-1)} + \ldots + a_m u = f \tag{1}$$

при граничных условиях

$$\Phi_\nu(u) = \sum_{j=0}^{m-1} [\alpha_{\nu j} u^{(j)}(1) + \beta_{\nu j} u^{(j)}(-1)] = 0, \quad \nu = 1, \ldots, m. \tag{2}$$

Будем предполагать выполненными следующие условия:

1) коэффициенты $a_i(x)$ и правая часть $f(x)$ — непрерывные функции;

2) если $f(x) \equiv 0$, то задача (1)—(2) имеет только нулевое решение;

3) из всех полиномов степени не выше $m-1$ граничным условиям (2) удовлетворяет лишь тождественно равный нулю[1].

Мы будем изучать следующий вариант метода моментов. Пусть выбрана бесконечная линейно независимая система функций $\omega_1, \omega_2, \ldots, \omega_n, \ldots$ Каждая из этих функций m раз непрерывно дифференцируема и удовлетворяет граничным условиям (2): $\Phi_\nu(\omega_n) = 0, \quad n = 1, 2, \ldots, \quad \nu = 1, 2, \ldots, m.$ Приближенное решение u_n ищется в виде $u_n = \sum_{k=1}^{n} c_k \omega_k$, причем коэффициенты c_k находятся из условий

$$\int_{-1}^{1} L(u_n)\omega_j^{(m)} dx = \int_{-1}^{1} f\omega_j^{(m)} dx, \quad j = 1, 2, \ldots, n.$$

Прежде чем доказывать сходимость метода, установим некоторые вспомогательные утверждения. Обозначим через U множество ($m-1$) раз непрерывно дифференцируемых функций, ($m-1$)-я производная которых абсолютно непрерывна, а m-я суммируема с квадратом, и которые удовлетворяют граничным условиям (2). Определим на U норму, положив $\|u\|_U = \|u^{(m)}\|_{L^2}$. Чтобы

[1] Формулируемая далее теорема верна и без этого предположения, однако это предположение несколько упрощает доказательство.

убедиться в законности такого определения, достаточно установить, что эта норма обращается в нуль лишь на функции, тождественно равной нулю, т. к. выполнение всех остальных аксиом нормы тривиально. Действительно, если $\| u \|_U = 0$, то $u^{(m)}(x) \equiv 0$ и u есть полином степени не выше $m-1$, а т. к. он удовлетворяет (2), то в силу сделанного предположения 3) тождественно равен нулю. Пространство U гильбертово, скалярное произведение в нем определяется формулой

$$(u_1, u_2) = \int_{-1}^{1} u_1^{(m)} u_2^{(m)} dx.$$

Лемма 1

Существуют такие постоянные $M_0, M_1, ..., M_{m-1}$, что для любой функции u из пространства U при $j = 0, 1, ..., m-1$ выполняются оценки $\left\| u^{(j)} \right\|_C \le M_j \| u \|_U$.

Доказательство

Однородная граничная задача $u^{(m)} = 0$ при граничных условиях (2) имеет только нулевое решение, как это следует из предположения 3). Следовательно, ей соответствует функция Грина $G(x, y)$. Для любой функции $u \in U$ теперь имеем

$$u(x) = \int_{-1}^{1} G(x, y) u^{(m)}(y) dy,$$

откуда

$$\left| u^{(j)}(x) \right| \le \sqrt{2} \max_{x,y} \left| \frac{d^j}{dx^j} G(x, y) \right| \cdot \left\| u^{(m)} \right\|_{L_2} = M_j \| u \|_U . \blacksquare$$

Лемма 2

Если $0 \le j \le m-1$, то оператор $D_j : D_j u = u^{(j)}$, рассматриваемый из пространства U в пространство L_2, компактен.

Доказательство

Пусть $V = D_j S$ — образ единичного шара S пространства U. При $j \le m-2$ все функции $v \in V$ в силу леммы 1 таковы, что $|\upsilon(x)| \le M_j$, $|\upsilon'(x)| \le M_{j+1}$, и потому V относительно компактно в пространстве C и тем более в L_2.

Если $j = m-1$, то $|\upsilon(x)| = \left| u^{(m-1)}(x) \right| \le M_{m-1}$, и, кроме того, для любых x', x'' $(-1 \le x' < x'' \le 1)$, полагая $v(x) = u^{(m-1)}(x)$, где $u \in S$,

$$|\upsilon(x') - \upsilon(x'')| = \left| \int_{x'}^{x''} \upsilon'(x)dx \right| \le \sqrt{x''-x'} \left(\int_{-1}^{1} [\upsilon'(x)]^2 dx \right)^{1/2} =$$

$$= \sqrt{x''-x'} \left(\int_{-1}^{1} [u^{(m)}]^2 dx \right)^{1/2} \le \sqrt{x''-x'},$$

так что функции из множества V равномерно ограничены и равностепенно непрерывны, и по теореме Арцела — Асколи V также относительно компактно в пространстве C. ■

Обозначим через A оператор, действующий из U в L_2 по правилу $Au = L(u)$ для $u \in U$. Из леммы 1 легко следует, что оператор A ограничен: $A \in L(U, L_2)$.

Лемма 3

Оператор A имеет ограниченный обратный.

Доказательство

Из предположения 2) сразу же следует обратимость оператора A, причем обратный оператор есть интегральный оператор, ядром которого является функция Грина граничной задачи (1), (2) $G_A(x, y)$. Итак, если u есть решение уравнения $Au = f$, то

$$u(x) = \int_{-1}^{1} G_A(x, y) f(y) dy.$$

Используя эту формулу и дифференциальное уравнение (1), видим, что

$$u^{(m)}(x) = f(x) - a_1(x)u^{(m-1)}(x) - \ldots - a_m(x)u(x) =$$

$$= f(x) - \int_{-1}^{1}\left(a_1(x)\frac{\partial^{m-1}G_A(x,y)}{\partial x^{m-1}} + \ldots + a_m(x)G_A(x,y)\right)f(y)dy.$$

Из последней формулы видно, что $\|u\|_U = \left\|u^{(m)}\right\|L_2 \le C\|f\|_{L_2}$, где

$$C = 1 + \left(\int_{-1}^{1}\int_{-1}^{1}\left(a_1(x)\frac{\partial^{m-1}G_A(x,y)}{\partial x^{m-1}} + \ldots + a_m(x)G_A(x,y)\right)^2 dxdy\right)^{1/2},$$

и потому оператор A^{-1} ограничен. ■

Теорема 1

Пусть для граничной задачи (1), (2) выполнены предположения 1)—3) и пусть координатная система $\{\omega_k\}$ такова, что система $\left\{\omega_k^{(m)}\right\}$ полна в L_2. Тогда приближенные решения u_n^*, определяемые методом моментов, существуют при всех достаточно больших n и сходятся к точному в том смысле, что

$$\left\|(u_n^* - u^*)^{(j)}\right\|_C \to 0 \text{ при } j = 0, 1, \ldots, m-1, \quad \left\|(u_n^* - u^*)^{(m)}\right\|_{L_2} \to 0.$$

Доказательство

Как уже отмечалось, граничная задача (1), (2) записывается в виде $Au = f$, где оператор $A \in L(U, L_2)$ определен ранее. Определим еще операторы $G, T \in L(U, L_2)$, положив для $u \in U$

$$Gu = u^{(m)}, \quad Tu = a_1u^{(m-1)} + \ldots + a_mu.$$

Тогда $A = G + T$. Покажем теперь, что выполнены все условия теоремы Канторовича (теорема 1 из *§ 2*). Условие 1а) выполнено в силу леммы 3, а условие 2а) очевидно — оператор G есть линейная изометрия из U в L_2. Оператор T представим в виде $T = A_1D_{m-1} + \ldots + A_mD_0$, где $D_ju = u^{(j)}$ — компактные (согласно лемме 2) операторы из U в L_2, а $A_j\upsilon = a_j\upsilon$ — огра-

ниченные операторы в L_2 $\left(A_j \in L(L_2, L_2)\right)$. Поэтому оператор T компактен, т. е. выполнено условие 3а). Далее полнота системы $\left\{\omega_k^{(m)}\right\}$ в L_2 означает выполнение условия 1б). Наконец, из линейной независимости системы $\{\omega_k\}$ и условия 3) вытекает, что система $\left\{\omega_k^{(m)}\right\}$ также линейно независима. Поэтому определитель, фигурирующий в условии 2б) и имеющий для нашего метода вид

$$\Delta\begin{pmatrix} \omega_1^{(m)} & \omega_2^{(m)} & \cdots & \omega_n^{(m)} \\ \omega_1^{(m)} & \omega_2^{(m)} & \cdots & \omega_n^{(m)} \end{pmatrix},$$

отличен от нуля как определитель Грама линейно независимой системы. Оператор P_n, определяемый элементами $G\omega_k = \omega_k^{(m)}$ и функционалами $\chi_j = \omega_j^{(m)}$, есть оператор ортогонального проектирования пространства L_2 на подпространство, натянутое на $\omega_1^{(m)}, \omega_2^{(m)}, ..., \omega_n^{(m)}$, и потому $\|P_n\| = 1$.

Итак, все условия теоремы Канторовича выполнены. Значит, системы линейных уравнений, к которым приводит метод моментов, разрешимы при достаточно больших n, и приближенные решения сходятся к точному: $\left\|u_n^* - u^*\right\|_U \to 0$.

Остается отметить, что сходимость функций в пространстве U означает сходимость в среднем (в L_2) производных порядка m и в силу леммы 1 равномерную сходимость самих функций и всех их производных до порядка $m-1$ включительно. ■

Наложим на координатную систему еще одно дополнительное условие: система $\left\{\omega_k^{(m)}\right\}$ ортонормальна в L_2, т. е.

$$\int_{-1}^{1} \omega_k^{(m)} \omega_j^{(m)} dx = \delta_{kj}.$$

Тогда $(\omega_k, \omega_j)_U = \delta_{kj}$, и потому выполняется условие 3б) из *§ 2*. При этом автоматически выполняется и условие 4б) — следует положить $g_k = \omega_k^{(m)}$. Поэтому из теоремы 2 (*§ 2*) сразу же вытекает утверждение об устойчивости.

Теорема 2

Пусть выполнены условия теоремы 1 и пусть система $\left\{\omega_k^{(m)}\right\}$ ортонормальна в L_2. Пусть, кроме того, системы уравнений метода моментов рассматриваются как уравнения в l_n^2. Тогда процессы отыскания каркасов и построения приближенных решений устойчивы.

Отметим, что устойчивость процесса построения приближенных решений означает, в частности, что малые в метрике l_n^2 ошибки при составлении и решении систем линейных уравнений вызовут равномерно малые ошибки в самих приближенных решениях и их производных до порядка $m-1$ и малые в среднем ошибки в производных порядка m.

Из сделанного в начале параграфа предположения 3) вытекает, что при каждом $n \geq m$ найдется такой полином степени n, который удовлетворяет граничным условиям (2). Координатную систему $\{\omega_k\}$, в которой ω_k есть полином степени $k+m-1$, будем называть полиномиальной. Для такой системы $\omega_k^{(m)}$ есть полином степени $k-1$. Метод моментов с полиномиальной координатной системой имеет очень простой смысл: приближенное решение u_{n+1}^* есть полином степени $n+m$, удовлетворяющий граничным условиям (2) и такой, что невязка ортогональна всем полиномам степени n. Множество линейных комбинаций функций $\omega_k^{(m)}$ есть множество всех полиномов. Значит, система $\left\{\omega_k^{(m)}\right\}$ полна в L_2, и тем самым полиномиальная координатная система обеспечивает сходимость метода моментов. Содержащееся в теореме Канторовича утверждение о быстроте сходимости позволяет в этом случае установить, что если все коэффициенты a_j и правая часть f p раз непрерывно дифференцируемы (и тем самым решение u^* $p+m$ раз непрерывно дифференцируемо), то

$$\left\| u_n^* - u^* \right\|_U = O(1/n^p), \tag{3}$$

т. е. для достаточно "гладких" задач сходимость метода моментов с полиномиальной координатной системой будет достаточно быстрой. Доказательство оценки (3) потребовало бы применения аппроксимационной теоремы Джек-

сона. Заметим еще, что более тонкий анализ позволяет, кроме того, получить и такую оценку:

$$\left\| u_n^* - u^* \right\|_C = O(\ln n / n^{p+m}).$$

Если мы хотим обеспечить устойчивость метода с полиномиальной координатной системой, то в качестве полиномов ω_k достаточно взять решения уравнений

$$\omega_{k+1}^{(m)} = \hat{P}_k \tag{4}$$

при граничных условиях (2). Здесь $\hat{P}_k$ — нормированные полиномы Лежандра:

$$\hat{P}_k(x) = \frac{1}{2^k \cdot k!}\sqrt{\frac{2k+1}{2}}\frac{d^k}{dx^k}(x^2-1)^k, \quad \int_{-1}^{1} \hat{P}_k(x)\hat{P}_j(x)dx = \delta_{kj}.$$

Любопытно отметить, что при $k \geq m$ решение дифференциального уравнения (4) при условиях (2) в действительности не зависит от этих граничных условий и дается формулой

$$\omega_{k+1}(x) = b_k (1-x^2)^m J_{k-m}^{(m,m)}(x),$$

где $J_\nu^{(m,m)}$ — многочлены Якоби, ортогональные с весом $(1-x^2)^m$, а b_k — некоторые коэффициенты, которые легко вычисляются.

§ 5. Метод наименьших квадратов

В этом и нескольких следующих параграфах будут рассмотрены методы, основанные на том, что задача решения уравнения часто может быть сведена к задаче отыскания минимума некоторого функционала. Такие методы называются *вариационными* и, как будет видно из дальнейшего, имеют много общего с проекционными методами. К числу вариационных методов относится метод наименьших квадратов.

Задача решения уравнения

$$Au = f, \tag{1}$$

где A — линейный (вообще говоря, неограниченный) оператор в гильбертовом пространстве H, сводится к задаче отыскания элемента, доставляющего минимум функционалу

$$F(u) = \|Au - f\|^2 = (Au - f, Au - f), \tag{2}$$

заданному на том же множестве $D(A)$, что и оператор A. Действительно, функционал F принимает лишь неотрицательные значения и обращается в нуль в том и только в том случае, когда u — решение уравнения (1). Как и в проекционных методах, в методе наименьших квадратов приближенное решение ищется в виде линейной комбинации координатных элементов: $u_n = \sum_{k=1}^{n} c_k \omega_k^{(n)}$, причем координатные элементы $\omega_k^{(n)}$ выбираются из области задания оператора A $D(A)$. В отличие от проекционных методов коэффициенты c_k находятся из условия, что приближенное решение u_n^* доставляет минимум функционалу F по сравнению со всеми другими линейными комбинациями тех же элементов $\omega_1^{(n)}, ..., \omega_n^{(n)}$:

$$F(u_n^*) = \min_{u_n \in U^{(n)}} F(u_n).$$

Здесь $U^{(n)}$ — подпространство, натянутое на $\omega_1^{(n)}, ..., \omega_n^{(n)}$.

Отметим, что функционал F заведомо достигает своего наименьшего на $U^{(n)}$ значения в некоторой точке $u_n^* \in U^{(n)}$. Действительно, если положить $F^{(n)} = AU^{(n)}$, то минимизировать на $U^{(n)}$ функционал F — значит, найти ближайший к f элемент подпространства $F^{(n)}$. Вне шара

$$S_n = \left\{ f_n \,\middle|\, f_n \in F^{(n)},\ \|f - f_n\| \le \|f\| \right\}$$

все элементы $F^{(n)}$ отстоят от f дальше, чем 0, и ближайший к f элемент из $F^{(n)}$ можно поэтому искать в S_n. Но S_n — компакт (как замкнутое ограниченное множество в конечномерном пространстве), и потому ближайший к f элемент в нем существует.

Поскольку для $u_n = \sum_{k=1}^{n} c_k \omega_k^{(n)}$

$$F(u_n) = \sum_{k,j=1}^{n} c_k c_j (A\omega_k^{(n)}, A\omega_j^{(n)}) - 2\sum_{j=1}^{n} c_j (A\omega_j^{(n)}, f) + (f, f) = \Phi(c_1, c_2, ..., c_n),$$

то для нахождения точки минимума функционала F на $U^{(n)}$ получается система линейных уравнений относительно c_k:

$$\frac{1}{2}\frac{\partial \Phi}{\partial c_j} = \sum_{k=1}^{n} c_k (A\omega_k^{(n)}, A\omega_j^{(n)}) - (f, A\omega_j^{(n)}) = 0, \quad j = 1, 2, ..., n. \tag{3}$$

Очевидно, что эта система уравнений совпадает с той, к которой приводит проекционный метод, если в качестве проекционной системы выбрать $\chi_j^{(n)} = A\omega_j^{(n)}$. Таким образом, метод наименьших квадратов можно рассматривать как частный случай проекционного метода. Но в то же время вариационная основа метода позволяет довольно просто провести его исследование.

Будем предполагать, что оператор A обратим, т. е. что однородное уравнение $Au = 0$ имеет только нулевое решение. Тогда из линейной независимости системы $\left\{\omega_k^{(n)}\right\}$ при заданном n вытекает линейная независимость системы $\left\{A\omega_k^{(n)}\right\}$. А это, в свою очередь, влечет одно важное следствие: матрица коэффициентов $A_n = \left\{(A\omega_k^{(n)}, A\omega_j^{(n)})\right\}$ системы линейных уравнений (3) есть матрица Грама линейно независимой системы и потому положительно определена. Таким образом, система уравнений (3) заведомо однозначно разрешима, и для ее решения можно применять методы, специально предназначенные для решения систем с положительно определенным и матрицами. Заметим, кстати, что решение этой системы действительно минимизирует функцию Φ, поскольку это единственная стационарная точка, а точка минимума существует.

Относительно сходимости метода наименьших квадратов можно утверждать следующее.

Теорема 1

Если система $\left\{A\omega_k^{(n)}\right\}$ полна в H, то для приближенных решений u_n^*, построенных по методу наименьших квадратов, выполняется соотношение $Au_n^* \to f$. Если, кроме того, оператор A^{-1} ограничен (т. е. для любого $f \in R(A)$ $\left\|A^{-1}f\right\| \le c\|f\|$) и существует решение u^* уравнения (1), то и $u_n^* \to u^*$.

Доказательство

Полнота системы $\left\{A\omega_k^{(n)}\right\}$ означает, что для любого элемента пространства H и, в частности, для правой части f нашего уравнения $E_{F^{(n)}}(f)\to 0$. Здесь $F^{(n)}=AU^{(n)}$ — подпространство H, натянутое на $A\omega_1^{(n)}, A\omega_2^{(n)}, ..., A\omega_n^{(n)}$. Но по самому смыслу метода

$$\left\|Au_n^*-f\right\|=\min_{u_n\in U^{(n)}}\left\|Au_n-f\right\|=\min_{f_n\in F^{(n)}}\left\|f_n-f\right\|=E_{F^{(n)}}(f).$$

Этим первое утверждение теоремы доказано. В случае существования точного решения u^* и ограниченности A^{-1} имеем

$$\left\|u_n^*-u^*\right\|=\left\|A^{-1}(Au_n^*-f)\right\|\le c\left\|Au_n^*-f\right\|\to 0.$$

И теорема доказана полностью. ■

Для выбранной координатной системы $\left\{\omega_k^{(n)}\right\}$ проверка полноты системы $\left\{A\omega_k^{(n)}\right\}$ может иногда представлять существенные трудности (систему $\left\{\omega_k^{(n)}\right\}$ мы обычно стараемся выбрать "простой", но если оператор A "сложный", то система $\left\{A\omega_k^{(n)}\right\}$ также может оказаться "сложной"). Поэтому полезным оказывается следующий признак полноты системы $\left\{A\omega_k^{(n)}\right\}$.

Теорема 2

Пусть A и B — два оператора в гильбертовом пространстве H, причем выполнены следующие условия: 1) области определения операторов A и B совпадают: $D(A)=D(B)$; 2) множества значений $R(A)$ и $R(B)$ этих операторов плотны в H; 3) операторы A и B обратимы; 4) оператор AB^{-1} ограничен (этот оператор задан на $R(B)$). Тогда если система $\left\{B\omega_k\right\}$ $(\omega_k^{(n)}\in D(B)=D(A))$ полна в H, то и система $\left\{A\omega_k^{(n)}\right\}$ полна в H.

Доказательство

Возьмем произвольный элемент $f\in H$ и произвольное число $\varepsilon>0$. Так как $R(A)$ плотно в H, то найдется такой элемент $f'\in R(A)$, что $\left\|f-f'\right\|<\varepsilon$. По-

ложим $g = BA^{-1}f'$ и, пользуясь полнотой системы $\left\{B\omega_k^{(n)}\right\}$, найдем такое число N и числа a_k, что $\left\| g - \sum_{k=1}^{N} a_k B\omega_k^{(n)} \right\| < \varepsilon$. Тогда

$$\left\| f' - \sum_{k=1}^{N} a_k A\omega_k^{(n)} \right\| = \left\| AB^{-1}\left(g - \sum_{k=1}^{N} a_k B\omega_k^{(n)} \right) \right\| \le \left\| AB^{-1} \right\| \cdot \varepsilon,$$

откуда

$$\left\| f - \sum_{k=1}^{N} a_k A\omega_k^{(n)} \right\| \le \| f - f' \| + \left\| f' - \sum_{k=1}^{N} a_k A\omega_k^{(n)} \right\| \le \left(\left\| AB^{-1} \right\| + 1 \right)\varepsilon = C\varepsilon.$$

Итак, для произвольных f и $\varepsilon > 0$ нашлась такая линейная комбинация элементов системы $\left\{A\omega_k^{(n)}\right\}$, которая приближает f с точностью до $C\varepsilon$. А это и означает полноту системы $\left\{A\omega_k^{(n)}\right\}$. ■

Приведем простой пример применения доказанной теоремы. Пусть нам дана граничная задача для обыкновенного дифференциального уравнения

$$Au = p(x)u'' + q(x)u' + r(x)u = f(x); \tag{4}$$

$$u(-1) = u(1) = 0, \tag{5}$$

причем, коэффициенты p, q и r дифференциального уравнения непрерывны, $p(x) > 0$. Пусть эта задача однозначно разрешима. Будем ее рассматривать как уравнение $Au = f$ в пространстве L_2. Оператор A задается формулой (4) на множестве $D(A)$ дважды непрерывно дифференцируемых функций, удовлетворяющих условию (5). Это уравнение мы хотим решать методом наименьших квадратов. Функции координатной системы будем выбирать так, что $\omega_k^{(n)}(x) = \omega_k$, где $\omega_1(x), \omega_2(x), \ldots$ бесконечная система функций. Эти функции должны быть дважды непрерывно дифференцируемы и удовлетворять условиям $\omega_k(-1) = \omega_k(1) = 0$.

Выберем в качестве $\omega_k(x)$ полиномы вида $\omega_k(x) = (1 - x^2)P_{k-1}(x)$, где P_{k-1} — полином степени $k-1$ (множитель $(1 - x^2)$ обеспечивает обращение $\omega_k(x)$ в нуль в точках 1 и –1). Нас интересует вопрос, будет ли полна в L_2 система $\left\{A\omega_k\right\}$. Определим на том же множестве $D(A)$ другой оператор: $Bu = u''$. Для операторов A и B выполняются все условия теоремы 2. Действительно, $D(B) = D(A)$ согласно определению оператора B. Множества значе-

ний $R(A)$ и $R(B)$ состоят из всех непрерывных функций (уравнения $Au = f$ и $Bu = f$ разрешимы, если функция f непрерывна) и потому плотны в L_2. Операторы A и B обратимы (первый из них ввиду сделанного предположения об однозначной разрешимости граничной задачи (4), (5), а обратимость B очевидна) и задаются с помощью соответствующих функций Грина:

$$B^{-1}f = \int_{-1}^{1} G_B(x,y)f(y)dy, \quad A^{-1}f = \int_{-1}^{1} G_A(x,y)f(y)dy.$$

Из свойств функций Грина легко получить оценки

$$\left\|B^{-1}f\right\|_{L_2} \le c_0\|f\|_{L_2}, \quad \left\|\left(B^{-1}f\right)'\right\|_{L_2} \le c_1\|f\|_{L_2}.$$

Поскольку для непрерывной функции f $AB^{-1}f = pf + q(B^{-1}f)' + r(B^{-1}f)$,

то

$$\left\|AB^{-1}f\right\|_{L_2} \le \|p\|_C \cdot \|f\|_{L_2} + \|q\|_C \cdot c_1 \cdot \|f\|_{L_2} + \|r\|_C \cdot c_0\|f\|_{L_2} =$$
$$= (\|p\|_C + c_1\|q\|_C + c_0\|r\|_C)\|f\|_{L_2} = c\|f\|_{L_2},$$

т. е. оператор AB^{-1} ограничен. В то же время множество линейных комбинаций функций $B\omega_k$ есть множество всех алгебраических полиномов, и потому система $\{B\omega_k\}$ полна в L_2. Теорема 2 позволяет теперь сделать заключение о полноте системы $\{A\omega_k\}$. Таким образом, метод наименьших квадратов с нашей координатной системой сходится.

Сделаем еще замечание о характере сходимости метода для задачи (4), (5). Совершенно аналогично предыдущему можно показать, что оператор BA^{-1} также ограничен. Поэтому

$$\left\|Bu^* - Bu_n^*\right\| = \left\|BA^{-1}(Au^* - Au_n^*)\right\| \le \left\|BA^{-1}\right\| \cdot \left\|f - Au_n^*\right\|_{L_2} \to 0.$$

Сходимость (в L_2) Bu_n^* к Bu^* означает сходимость в среднем вторых производных приближенных решений ко второй производной точного решения. Учитывая, что и приближенные, и точное решения удовлетворяют граничным условиям (5), отсюда легко получить, что сами приближенные решения и их первые производные сходятся равномерно. Вообще стоит отметить, что для многих задач сходимость к нулю невязок означает более сильную сходимость, чем сходимость приближенных решений в метрике исходного гильбертова пространства H. Обычно это связано с компактностью оператора A^{-1}.

§ 6. Метод Ритца

Метод Ритца, о котором пойдет речь в этом параграфе, также относится к числу вариационных. Прежде чем описывать сам этот метод, изложим некоторые свойства положительно определенных операторов в гильбертовом пространстве. Мы будем рассматривать уравнение

$$Au = f \tag{1}$$

с линейным оператором A, действующим в вещественном гильбертовом пространстве H. Область задания $D(A)$ и множество значений $R(A)$ этого оператора считаем плотными в H. Кроме того, оператор A будем считать симметричным и положительно определенным. Первое из этих требований означает, что для любых элементов $u, \upsilon \in D(A)$ выполняется равенство

$$(Au, \upsilon) = (A\upsilon, u). \tag{2}$$

Симметричный оператор A называется *положительно определенным*, если существует такая постоянная $m > 0$, что для любого $u \in D(A)$

$$(Au, u) \geq m^2 (u, u). \tag{3}$$

Положительно определенный оператор A заведомо имеет обратный A^{-1}, заданный на $R(A)$, поскольку в силу (3) однородное уравнение $Au = 0$ имеет лишь нулевое решение. Из неравенства (3) следует также, что

$$\| u \|^2 \leq \frac{1}{m^2}(Au, u) \leq \frac{1}{m^2}\| Au \| \cdot \| u \|,$$

так что $\| Au \| \geq m^2 \| u \|$ и потому обратный оператор A^{-1} ограничен: $\left\| A^{-1} \right\| \leq 1/m^2$.

На множестве $D(A)$ введем новое скалярное произведение по формуле

$$[u, \upsilon] = (Au, \upsilon).$$

Билинейный функционал $[u, \upsilon]$ действительно обладает всеми свойствами, которыми должно обладать скалярное произведение: он симметричен ввиду симметричности оператора A и $[u, u] \geq 0$, обращаясь в нуль лишь для $u = 0$, ввиду неравенства (3). Если ввести теперь на $D(A)$ норму, положив

$\|u\|_A = \sqrt{[u, u]}$, то множество $D(A)$ превратится в унитарное (предгильбертово) пространство, которое обозначим через H_A^0. Отметим, что для любого элемента $u \in D(A)$ выполняются неравенства

$$m\|u\| \le \|u\|_A \le \frac{1}{m}\|Au\|. \tag{4}$$

Первое из этих неравенств есть непосредственное следствие (3), а второе доказывается следующей цепочкой неравенств:

$$\|u\|_A^2 = (Au, u) \le \|Au\| \cdot \|u\| \le \frac{1}{m}\|Au\| \cdot \|u\|_A;$$

поделив первый и последний члены этого неравенства на $\|u\|_A$, придем к требуемому (если $u = 0$, то (4) очевидно).

Пространство H_A^0, вообще говоря, не является полным. Его пополнение H_A называется *энергетическим пространством* оператора A. Остановимся несколько подробнее на процессе пополнения. Поскольку пространство H_A^0 унитарно, то его пополнением является сопряженное к нему пространство H_A^{0*}. При этом, конечно, каждый элемент $u \in H_A^0$ отождествляется с функционалом $\varphi \in H_A^{0*}$, задаваемым формулой $\varphi(\upsilon) = [u, \upsilon] = (A\upsilon, u)$. Оказывается, что каждый функционал $\varphi \in H_A^{0*}$ может быть представлен аналогичной формулой: $\varphi(\upsilon) = (A\upsilon, u)$, где u — некоторый элемент пространства H.

Лемма

Пусть $\varphi \in H_A^{0^*}$ — произвольный функционал. Тогда найдется такой единственный элемент $u \in H$, что для всех $\upsilon \in H_A^0$ $\varphi(\upsilon) = (A\upsilon, u)$.

Доказательство

С помощью функционала φ на множестве $R(A)$ определим функционал ψ формулой $\psi(v) = \varphi(A^{-1}v)$. При этом выполняется неравенство

$$|\psi(w)| \le \|\varphi\|_A^* \cdot \|A^{-1}w\|_A \le \frac{1}{m}\|\varphi\|_A \cdot \|w\|.$$

Таким образом ψ — ограниченный функционал, заданный на плотном в H множестве; он может быть продолжен на все пространство H. Пользуясь теоремой об общей форме функционала в гильбертовом пространстве, можно утверждать существование такого элемента $u \in H$, что $\psi(w) = (w, u)$. Но для $\upsilon \in D(A)$ $\varphi(\upsilon) = \psi(A\upsilon) = (A\upsilon, u)$, и этим представимость φ в нужной нам форме доказана. Докажем единственность такого элемента u. Если бы их было два — u_1 и u_2, то для всех $\upsilon \in D(A)$ было бы $(A\upsilon, u_1) = (A\upsilon, u_2)$. Но когда υ пробегает $D(A)$, $A\upsilon$ пробегает множество $R(A)$, плотное в H. Поэтому разность $u_1 - u_2$ окажется ортогональной плотному множеству $R(A)$ и потому равной нулю. ■

Этой леммой фактически установлено вложение пространства H_A в пространство H, причем элементам из H_A^0 (т. е. из $D(A)$) оказываются сопоставленными они же сами. Пользуясь этим, мы будем все элементы H_A отождествлять с соответствующими элементами пространства H.

Скалярное произведение в H_A и норму будем обозначать теми же символами [.,.] и $\|\cdot\|_A$. Из леммы немедленно вытекает следующая теорема.

Теорема 1

Пополнение пространства H_A^0 может быть осуществлено за счет элементов пространства H. При этом элемент $u \in H$ принадлежит пополнению H_A тогда и только тогда, когда функционал $\varphi(\upsilon) = (Au, \upsilon)$ ограничен в пространстве H_A^0:

$$(A\upsilon, u) \le c\|\upsilon\|_A \text{ для всех } \upsilon \in D(A).$$

При этом, если $\upsilon \in D(A)$, то и в пространстве H_A $[u, \upsilon] = (A\upsilon, u)$.

❑ **Замечание.** При доказательстве леммы была использована плотность множества $R(A)$ в пространстве H. Без этого предположения утверждение теоремы 1 о возможности пополнения пространства H_A^0 за счет элементов H остается верным, но ограниченность функционала $\varphi(\upsilon) = (Au, \upsilon)$ будет лишь необходимым, но не достаточным признаком принадлежности элемента u пространству H_A.

Отметим некоторые свойства пространства H_A.

1. Если последовательность $\{u_n\}$, где $u_n \in D(A)$, сходится к u в H_A $\left(\|u_n - u\|_A \to 0\right)$, то к тому же пределу она сходится и в H. Действительно, последовательность $\{u_n\}$ сходится в себе в H_A, а т. к. $\|u_n - u_k\| \le$ $\le \frac{1}{m}\|u_n - u_k\|_A$, то и в H, так что для некоторого $u' \in H$ $\|u_n - u'\| \to 0$. Но для всех $\upsilon \in D(A)$ $(A\upsilon, u_n) \to (A\upsilon, u')$ и в то же время

$$(A\upsilon, u_n) = [\upsilon, u_n] \to [\upsilon, u] = (A\upsilon, u),$$

так что для всех $w \in R(A)$ $(w, u) = (w, u')$ и $u' = u$.

2. Для всех $u \in H_A$ выполняется неравенство

$$\|u\| \le \frac{1}{m}\|u\|_A.$$

Это легко доказывается предельным переходом в соответствующем неравенстве для элементов u из $D(A)$.

3. Если последовательность $\{u_n\}$, где на этот раз[1] $u_n \in H_A$, сходится к u в H_A, то к тому же пределу она сходится и в H. Это непосредственно следует из предыдущего свойства.

Вернемся к рассмотрению уравнения (1). Сопоставим этому уравнению функционал

$$F(u) = [u, u] - 2(u, f),$$

заданный на H_A и называемый *функционалом энергии*. Отметим, что для $u \in D(A)$ этот функционал может быть записан также в форме $F(u) = (Au, u) - 2(u, f)$. Функционал энергии связан с уравнением (1) следующим образом.

[1] В этом отличие от свойства (1).

Теорема 2 (о функционале энергии)

1. Существует единственный элемент $u_0 \in H_A$, на котором функционал F достигает своего наименьшего значения.

2. Если $u_0 \in D(A)$, то $Au_0 = f$, т. е. u_0 есть решение уравнения (1).

3. Если уравнение (1) имеет решение u^*, то $u_0 = u^*$.

Доказательство

1. Для любого элемента $u \in H_A$

$$|(u, f)| \le \|u\| \cdot \|f\| \le \frac{1}{m}\|f\| \cdot \|u\|_A.$$

Таким образом, (u, f) есть ограниченный функционал в гильбертовом пространстве H_A. Значит, найдется такой элемент $u_0 \in H_A$, что для любого $u \in H_A$ $(u, f) = [u, u_0]$. Тогда

$$F(u) = [u, u] - 2[u, u_0] = [u - u_0, u - u_0] - [u_0, u_0] =$$
$$= \|u - u_0\|_A - \|u_0\|_A^2 \ge -\|u_0\|_A^2 = F(u_0),$$

причем знак равенства будет здесь лишь в том случае, если $u = u_0$.

2. Если $u_0 \in D(A)$, то для любого $u \in H_A$

$$(u, f) = [u, u_0] = (u, Au_0),$$

а т. к. H_A плотно в H (в H плотно даже $D(A)$ — часть H_A), то это равенство для всех $u \in H_A$ возможно лишь в случае $Au_0 = f$.

3. Если $Au^* = f$, то для любого $u \in H_A$ $[u, u_0] = (u, f) = (u, Au^*) = [u, u^*]$, что опять-таки возможно лишь в случае $u_0 = u^*$. ■

Элемент u_0, доставляющий минимум функционалу F, будем называть *обобщенным решением* уравнения (1). Как это следует из теоремы, в случае выполнения хотя бы одного из условий — u_0 принадлежит области задания оператора A или решение уравнения (1) существует — u_0 является решением (1). Итак, обобщенное решение уравнения (1) всегда существует, а если существует и обычное решение, то обобщенное с ним совпадает.

В математической физике обобщенные решения часто определяются с помощью "интегральных равенств". Как ясно из доказательства теоремы 2, данное ранее определение обобщенного решения эквивалентно следующему: элемент $u_0 \in H_A$ называется *обобщенным решением* уравнения (1), если для всех элементов $u \in H_A$[1] выполняется равенство $[u_0, u] = (f, u)$. Последнее равенство обычно можно трактовать как "интегральное".

Рассмотрим теперь оператор G, который каждому элементу $f \in H$ ставит в соответствие обобщенное решение уравнения $Au = f : Gf = u_0$. Значение этого оператора принадлежит H_A и однозначно определяется условием

$$[Gf, u] = (f, u) \text{ для всех } u \in H_A$$

(см. доказательство теоремы 2). На множестве $R(A)$ оператор G совпадает с A^{-1}.

Поскольку

$$\|Gf\| \le \frac{1}{m}\|Gf\|_A = \frac{1}{m}\sup_{\|u\|_A \le 1}[Gf, u] = \frac{1}{m}\sup_{\|u\|_A \le 1}(f, u) \le \frac{1}{m^2}\sup_{\|u\|_A \le 1/m}(f, u) = \frac{1}{m^2}\|f\|,$$

то оператор G ограничен, и, следовательно, это есть расширение по непрерывности оператора A^{-1} на все пространство H (A^{-1} задан и ограничен на плотном множестве $R(A)$). Если f_1 и f_2 — различные элементы H, то $Gf_1 \ne Gf_2$, поскольку (f_1, u) и (f_2, u) — различные функционалы на H_A. Поэтому оператор G имеет обратный G^{-1}, заданный на $R(G)$. Очевидно, что $H_A \supseteq R(G) \supseteq D(A)$ и на $D(A)$ оператор G^{-1} совпадает с A. Итак, обобщенное решение u_0 уравнения (1) есть решение уравнения $G^{-1}u = f$, где G^{-1} есть некоторое расширение оператора A. Не следует думать, что G^{-1} есть расширение A на все энергетическое пространство. Отметим еще, что G^{-1} — положительно определенный оператор, энергетическое пространство которого совпадает с энергетическим пространством оператора A. Мы не будем доказывать эти два последних утверждения, хотя это делается и несложно. В сущности, эти утверждения потребуются в дальнейшем лишь для того,

[1] Можно ограничиться элементами $u \in D(A)$.

чтобы, не умаляя общности, предполагать в § 9, что для некоторого положительно определенного оператора A будет $R(A) = H$ — такое расширение допускает любой положительно определенный оператор.

Перейдем теперь непосредственно к методу Ритца. Как и в проекционных методах, приближенное решение в этом методе строится в виде линейной комбинации координатных элементов: $u_n = \sum_{k=1}^{n} c_k \omega_k^{(n)}$. Здесь $\omega_k^{(n)}$ — элементы энергетического пространства H_A, и при каждом n $\omega_1^{(n)}, \omega_2^{(n)}, ..., \omega_n^{(n)}$ линейно независимы. Коэффициенты c_k ищутся из условия минимизации функционала энергии F, т. е. приближенное решение u_n^* должно удовлетворять соотношению

$$F(u_n^*) = \min_{u_n \in U^{(n)}} F(u_n).$$

Минимум в правой части берется по множеству $U^{(n)}$ всех линейных комбинаций координатных элементов $\omega_1^{(n)}, ..., \omega_n^{(n)}$. Отметим, что поскольку $F(u_n) = \| u_n - u_0 \|_A^2 - \| u_0 \|_A^2$, то минимум заведомо достигается. При этом

$$F(u_n) = F\left(\sum_{k=1}^{n} c_k \omega_k^{(n)} \right) = \sum_{i,k=1}^{n} [\omega_i^{(n)}, \omega_k^{(n)}] c_i c_k - 2 \sum_{k=1}^{n} c_k (\omega_k^{(n)}, f) = \Phi(c_1, ..., c_n),$$

и для нахождения точки минимума следует приравнять нулю производные функции Φ:

$$\frac{1}{2} \frac{\partial \Phi}{\partial c_k} = \sum_{i=1}^{n} [\omega_i^{(n)}, \omega_k^{(n)}] c_i - (f, \omega_k^{(n)}) = 0.$$

Таким образом, построение приближенного решения по методу Ритца сводится к решению системы линейных алгебраических уравнений

$$\sum_{i=1}^{n} [\omega_i^{(n)}, \omega_k^{(n)}] c_i = (f, \omega_k^{(n)}), \quad k = 1, 2, ..., n. \tag{5}$$

Матрица коэффициентов этой системы уравнений $\Gamma_n^A = \left\{ [\omega_i^{(n)}, \omega_k^{(n)}] \right\}$ есть матрица Грама линейно независимой системы $\left\{ \omega_k^{(n)} \right\}$ в энергетическом про-

странстве H_A (что нашло отражение в обозначении) и потому является положительно определенной. Итак, система уравнений (5) однозначно разрешима, и ее решение действительно доставляет минимум функции Ф.

Если координатная система $\{\omega_k^{(n)}\}$ выбрана из более узкого, чем H_A, множества — области задания оператора A $(\omega_k^{(n)} \in D(A))$, то систему уравнений (5) можно переписать в виде

$$\sum_{i=1}^{n} (A\omega_i^{(n)},\ \omega_k^{(n)}) c_i = (f, \omega_k^{(n)}), \quad k = 1, 2, ..., n.$$

Точно к такой же системе приведет применение к уравнению (1) метода Галеркина с той же самой координатной системой. Поэтому в случае $\omega_k^{(n)} \in D(A)$ метод Ритца совпадает с методом Галеркина.

Вернемся к системе уравнений (5). Поскольку $(f, \omega_k^{(n)}) = [u_0, \omega_k^{(n)}]$, то из (5) непосредственно видно, что приближенное решение u_n^*, построенное по методу Ритца, удовлетворяет равенствам

$$[u_n^*, \omega_k^{(n)}] = [u_0, \omega_k^{(n)}].$$

Поэтому u_n^* имеет простой геометрический смысл — это ортогональная (в смысле энергетического скалярного произведения) проекция обобщенного решения u_0 на подпространство $U^{(n)} \subset H_A$, натянутое на координатные элементы $\omega_k^{(n)}$. Почти очевидно также следующее утверждение.

Теорема 3

Если координатная система $\{\omega_k^{(n)}\}$ полна в энергетическом пространстве H_A, то приближенные решения u_n^*, построенные по методу Ритца, сходятся к обобщенному решению как в энергетической метрике, так и в метрике пространства H:

$$\|u_n^* - u_0\|_A \to 0, \quad \|u_n^* - u_0\| \to 0. \tag{6}$$

Доказательство

Так как u_n^* доставляет минимум функционалу энергии $F(u)$ на подпространстве $U^{(n)}$, а $F(u)=\|u-u_0\|_A^2-\|u_0\|_A^2$, то очевидно, что $\left\|u_n^*-u_0\right\|_A = E_{U^{(n)}}(u_0)_{H_A}$ — наилучшее приближение обобщенного решения u_0 элементами из $U^{(n)}$ в энергетическом пространстве. Но полнота системы $\left\{\omega_k^{(n)}\right\}$ в H_A как раз и означает, что для любого $u\in H_A$, а тем самым и для u_0, $E_{U^{(n)}}(u)_{H_A}\to 0$. Этим доказано первое из соотношений (6). Второе сразу же следует из первого, ибо, как уже отмечалось раньше, сходимость в H_A влечет сходимость в H. ■

Отметим одно свойство метода Ритца в случае, когда координатная система выбрана по принципу метода моментов: $\omega_k^{(n)}=\omega_k$, где $\{\omega_k\}$ — бесконечная линейно независимая система элементов энергетического пространства. В этом случае, как это ясно из геометрического смысла метода, последовательность $\left\{u_n^*\right\}$ будет сходящейся в H_A независимо от того, является ли координатная система $\{\omega_k\}$ полной в H_A. Только пределом этой последовательности в общем случае будет ортогональная в H_A проекция обобщенного решения на подпространство, являющееся замыканием (в H_A) линейной оболочки координатной системы.

Докажем два признака полноты системы $\left\{\omega_k^{(n)}\right\}$ в энергетическом пространстве H_A.

Теорема 4

Если элементы $\omega_k^{(n)}$ принадлежат $D(A)$ и система $\left\{A\omega_k^{(n)}\right\}$ полна в H, то система $\left\{\omega_k^{(n)}\right\}$ полна в H_A.

Доказательство

Множество $D(A)$ плотно в H_A, и потому для произвольного элемента $u\in H_A$ по заданному $\varepsilon>0$ можно найти $u'\in D(A)$, такой, что $\left\|u-u'\right\|_A<\varepsilon$.

Ввиду полноты системы $\left\{A\omega_k^{(n)}\right\}$ при всех достаточно больших n $(n > N)$ найдутся такие числа c_k, что $\left\| Au' - \sum_{k=1}^{n} c_k A\omega_k^{(n)} \right\| < \varepsilon$. Но тогда

$$\left\| u - \sum_{k=1}^{n} c_k \omega_k^{(n)} \right\|_A \le \left\| u - u' \right\|_A + \left\| u' - \sum_{k=1}^{n} c_k \omega_k^{(n)} \right\|_A <$$

$$< \varepsilon + \frac{1}{m} \left\| A(u' - \sum_{k=1}^{n} c_k \omega_k^{(n)}) \right\| < (1 + \frac{1}{m})\varepsilon . \quad \blacksquare$$

Будем предполагать, что наряду с оператором A задан другой положительно определенный оператор B. Через H_B будем обозначать энергетическое пространство этого оператора. Для того чтобы различать энергетические скалярные произведения и нормы в энергетических пространствах этих двух операторов, будем обозначать их соответственно через $[u, \upsilon]_A$, $[u, \upsilon]_B$, $\|u\|_A$, $\|u\|_B$.

Теорема 5

Пусть выполнены следующие условия: 1) области задания операторов A и B совпадают: $D(A) = D(B) = D$; 2) $R(A)$ и $R(B)$ плотны в H; 3) существует такая постоянная $d > 0$, что для любого элемента $u \in D$ выполняется неравенство

$$(Au, u) \le d^2 (Bu, u).$$

Тогда если $\omega_k^{(n)} \in H_B$ и система $\left\{\omega_k^{(n)}\right\}$ полна в H_B, то $\omega_k^{(n)} \in H_A$ и $\left\{\omega_k^{(n)}\right\}$ полна в H_A.

Доказательство

Покажем сначала, что $H_B \subseteq H_A$. Заметим, что для любого $u \in D$ $\|u\|_A \le d \|u\|_B$. Пусть u — произвольный элемент из H_B. Множество D плотно в H_B, и поэтому найдется последовательность $\{u_k\} \subset D$, такая, что $\|u_k - u\|_B \to 0$, а тем самым и $\|u_k - u\| \to 0$. Так как при любых k и m $\|u_k - u_m\|_A \le d \|u_k - u_m\|_B$, то последовательность $\{u_k\}$ сходится в себе в пространстве H_A и потому сходится к некоторому элементу $u' \in H_A$:

$\|u_k - u'\|_A \to 0$. Но тогда и $\|u_k - u'\| \to 0$ и потому $u = u'$, и тем самым $u \in H_A$. Кроме того, переходя к пределу в неравенстве $\|u_k\|_A \le d\|u_k\|_B$, получаем $\|u\|_A \le d\|u\|_B$. Полнота системы $\left\{\omega_k^{(n)}\right\}$ в H_A следует теперь из того, что D плотно в H_A, а для любого $u \in D$ всегда

$$\left\|u - \sum_{k=1}^{n} a_k \omega_k^{(n)}\right\|_A \le d\left\|u - \sum_{k=1}^{n} a_k \omega_k^{(n)}\right\|_B. \blacksquare$$

❑ **Замечание.** Как видно из доказательства, условие $D(A) = D(B)$ можно заменить требованием, чтобы выполнялось включение $D(B) \subseteq D(A)$ и чтобы $D(B)$ было плотно в H_A.

Рассмотрим теперь некоторые вопросы, связанные с устойчивостью метода Ритца.

В качестве приближенного уравнения будем рассматривать систему линейных уравнений (5) как уравнение в пространстве l_n^2.

Теорема 6

Пусть выполнены условия теоремы 3. Для того чтобы процессы отыскания каркасов и построения приближенных решений в методе Ритца были устойчивы необходима и достаточна ограниченность чисел обусловленности матриц Грама координатной системы в энергетическом пространстве: $\mu_2\left(\Gamma_n^A\right) \le C$.[1]

Доказательство

Поскольку Γ_n и есть матрица коэффициентов системы уравнений (5), необходимость и достаточность указанного условия для процесса отыскания каркасов очевидна *(теорема 1 из § 1 главы 1)*. Остается доказать достаточность этого условия для устойчивости процесса построения приближенных решений, т. е. согласно *теореме 2 из § 1 главы 1* ограниченность произведений $\|\overline{\varphi}_n\| \cdot \left\|z_n^*\right\|_2$. Здесь z_n^* — решение системы (5), а $\overline{\varphi}_n$ — линейный оператор, действующий из l_n^2 в H по правилу:

---

[1] Далее для краткости будем обозначать $\Gamma_n^A = \Gamma_n$.

для $z_n = (c_1, ..., c_n) \in l_n^2$ $\overline{\varphi}_n z_n = \sum_{k=1}^{n} c_k \omega_k^{(n)}$.

Используя *лемму из § 3 главы 3*, имеем

$$\left\| z_n^* \right\|_2 \le \left\| \Gamma_n^{-1/2} \right\|_2 \cdot \left\| \Gamma_n^{1/2} z_n^* \right\|_2 = \left\| \Gamma_n^{-1/2} \right\|_2 \cdot \left\| u_n^* \right\|_A \le \left\| \Gamma_n^{-1/2} \right\|_2 \cdot \left\| u_0 \right\|_A$$

и согласно той же лемме для любого $z_n \in l_n^2$

$$\left\| \overline{\varphi}_n z_n \right\| \le \frac{1}{m} \left\| \overline{\varphi}_n z_n \right\|_A = \frac{1}{m} \left\| \Gamma_n^{1/2} z_n \right\|_2 \le \frac{1}{m} \left\| \Gamma_n^{1/2} \right\|_2 \left\| z_n \right\|_2,$$

так что

$$\left\| \overline{\varphi}_n \right\| \le \frac{1}{m} \left\| \Gamma_n^{1/2} \right\|_2 \text{ и } \left\| z_n^* \right\|_2 \cdot \left\| \overline{\varphi}_n \right\| \le \frac{1}{m} \left\| u_0 \right\|_A \left(\mu_2 (\Gamma_n) \right)^{1/2},$$

и этим доказательство завершено. ■

Пусть A и B — два положительно определенных оператора с плотными в H областями задания $D(A)$ и $D(B)$ и плотными множествами значений $R(A)$ и $R(B)$. Операторы A и B будем называть *полусходными*, если совпадают по своему составу их энергетические пространства H_A и H_B и нормы в H_A и H_B эквивалентны, т. е. существуют такие постоянные $d_1 > 0$ и $d_2 > 0$, что для всех $u \in H_A = H_B$

$$\left\| u \right\|_A \le d_1 \left\| u \right\|_B, \quad \left\| u \right\|_B \le d_2 \left\| u \right\|_A. \tag{7}$$

Как ясно из доказательства теоремы 5, для того чтобы операторы A и B были полусходными, достаточно, чтобы совпадали их области задания: $D(A) = D(B) = D$ и чтобы для всех $u \in D$ выполнялись неравенства

$$(Au, u) \le d_1^2 (Bu, u), \quad (Bu, u) \le d_2^2 (Au, u).$$

Но полусходные операторы могут, вообще говоря, иметь и разные области задания.

Из эквивалентности норм в H_A и H_B следует, что для того, чтобы некоторая система $\left\{ \omega_k^{(n)} \right\}$ $\left(\omega_k^{(n)} \in H_A \right)$ была полной в H_A, необходимо и достаточно, чтобы она была полна в H_B.

Пусть $\left\{ \omega_k^{(n)} \right\}$ — некоторая система элементов из H_A и, как всегда, при каждом n $\omega_1^{(n)}, ..., \omega_n^{(n)}$ линейно независимы. Обозначим через Γ_n^A и Γ_n^B матрицы Грама этой системы в гильбертовых пространствах H_A и H_B соответственно:

$$\Gamma_n^A = \left\{ [\omega_i^{(n)}, \omega_k^{(n)}]_A \right\}, \quad \Gamma_n^B = \left\{ [\omega_i^{(n)}, \omega_k^{(n)}]_B \right\},$$

где A и B — полусходные операторы.

Теорема 7

Для матриц Грама Γ_n^A и Γ_n^B справедливы оценки

$$\left\| \Gamma_n^A \right\|_2 \le d_1^2 \left\| \Gamma_n^B \right\|_2, \quad \left\| (\Gamma_n^A)^{-1} \right\|_2 \le d_2^2 \left\| (\Gamma_n^B)^{-1} \right\|_2,$$

где d_1 и d_2 — постоянные из неравенств (7).

Доказательство

Отметим прежде всего, что если $z_n = (c_1, ..., c_n)$ — произвольный n-мерный вектор и $u_n = \sum_{k=1}^n c_k \omega_k^{(n)}$, то выполняется равенство

$$(\Gamma_n^A z_n, z_n) = \sum_{k=1}^n \sum_{j=1}^n [\omega_k^{(n)}, \omega_j^{(n)}]_A c_k c_j = [u_n, u_n]_A,$$

и точно так же

$$(\Gamma_n^B z_n, z_n) = [u_n, u_n]_B.$$

Поэтому для любого вектора z_n в силу (7) будет

$$(\Gamma_n^A z_n, z_n) \le d_1^2 (\Gamma_n^B z_n, z_n), \quad (\Gamma_n^B z_n, z_n) \le d_2^2 (\Gamma_n^A z_n, z_n).$$

Обозначим через $\lambda_{\max}^A, \lambda_{\min}^A, \lambda_{\max}^B, \lambda_{\min}^B$ соответственно максимальные и минимальные собственные числа матриц Γ_n^A и Γ_n^B. Напомним, что если D — симметричная матрица порядка n и $z_n \in R^n$ — произвольный вектор, то выполняется неравенство

$$\lambda_{\min}(z_n, z_n) \le (Dz_n, z_n) \le \lambda_{\max}(z_n, z_n),$$

где $\lambda_{\min}$ и $\lambda_{\max}$ — минимальное и максимальное собственные числа матрицы D. Пусть λ — произвольное собственное число матрицы Γ_n^A и z_n — соответствующий ему собственный вектор, нормированный условием $(z_n, z_n) = 1$. Тогда имеем цепочку неравенств

$$\lambda_{\min}^B / d_2^2 \le (\Gamma_n^B z_n, z_n) / d_2^2 \le (\Gamma_n^A z_n, z_n) = \lambda \le d_1^2 (\Gamma_n^B z_n, z_n) \le d_1^2 \lambda_{\max}^B.$$

Если, в частности, $\lambda = \lambda_{\max}^A$ — максимальное собственное число матрицы Γ_n^A, то из написанных неравенств следует, что

$$\left\| \Gamma_n^A \right\|_2 = \lambda_{\max}^A \le d_1^2 \lambda_{\max}^B = d_1^2 \left\| \Gamma_n^B \right\|_2.$$

Если же $\lambda = \lambda_{\min}^A$, то

$$\left\| (\Gamma_n^A)^{-1} \right\|_2 = 1 / \lambda_{\min}^A \le d_2^2 / \lambda_{\min}^B = d_2^2 \left\| (\Gamma_n^B)^{-1} \right\|_2. \quad \blacksquare$$

Следствие 1. Числа обусловленности матриц Γ_n^A и Γ_n^B связаны с неравенством

$$\mu_2(\Gamma_n^A) \le d_1^2 d_2^2 \, \mu_2(\Gamma_n^B).$$

Из теорем 6 и 7 немедленно вытекает также…

Следствие 2. Пусть оператор B полусходен с оператором A и координатная система полна в пространстве H_B. Тогда для устойчивости процессов отыскания каркасов и построения приближенных решений необходима и достаточна ограниченность чисел обусловленности $\mu_2\left(\Gamma_n^B\right)$.

Заметим, что в частности устойчивость гарантируется выбором полной ортонормированной в H_B координатной системы: $\left[\omega_k^{(n)}, \omega_j^{(n)}\right]_B = \delta_{kj}$.

❑ **Замечание.** Как видно из доказательства теоремы 6, нормы операторов $\overline{\varphi}_n$ равномерно ограничены не только в том случае, когда мы рассматриваем их из l_n^2 в H, но и из l_n^2 в H_A. Поэтому в условиях теоремы 7 можно утверждать устойчивость процесса построения приближенных решений в более сильном смысле — можно оценивать ошибки приближенных решений в энергетической метрике.

§ 7. Применение метода Ритца к обыкновенным дифференциальным уравнениям

Рассмотрим обыкновенное дифференциальное уравнение второго порядка

$$L(u) = -(pu')' + ru = f$$

при одном из следующих граничных условий:

I. $u(-1) = u(1) = 0.$

II. $u'(-1) = u'(1) = 0.$

III. $u'(-1) + \alpha u(-1) = u'(1) + \beta u(1) = 0.$

Функцию p мы считаем непрерывно дифференцируемой и строго положительной: $p(x) \geq p_0 > 0,$ а r — непрерывной. Для каждого из граничных условий I—III введем в рассмотрение оператор A, действующий в пространстве $L_2(-1,1)$. Областью задания оператора A является множество $D(A)$ всех дважды непрерывно дифференцируемых функций, удовлетворяющих соответствующим граничным условиям. Функции $u \in D(A)$ оператор A ставит в соответствие функцию $Au = L(u)$. Займемся изучением свойств оператора A.

Свойство 1

Оператор A симметричен, причем $(Au, \upsilon) = I(u, \upsilon) + Q(u, \upsilon),$ где

$$I(u, \upsilon) = \int_{-1}^{1} (pu'\upsilon' + ru\upsilon)\,dx,$$

$$Q(u, \upsilon) = \begin{cases} 0 & \text{при граничных условиях I и II} \\ \beta p(1)u(1)\upsilon(1) - \alpha p(-1)u(-1)\upsilon(-1) & \text{при граничном условии III.} \end{cases}$$

Доказательство

Интегрированием по частям получаем

$$(Au, \upsilon) = I(u, \upsilon) - pu'\upsilon'\Big|_{-1}^{1}$$

и остается второе слагаемое преобразовать к виду $Q(u, \upsilon)$. Для этого достаточно заметить, что при условии I $\upsilon(-1) = \upsilon(1) = 0,$ при условии II

$u'(-1)=u'(1)=0$, а при условии III воспользоваться формулами $u'(-1)=-\alpha u(-1)$, $u(1)=-\beta u(1)$. Симметричность оператора A следует из симметричности $I(u,\upsilon)$ и $Q(u,\upsilon)$. ■

Свойство 2

Оператор A положительно определен, если выполняются следующие требования (свои для каждого из граничных условий):

1) при условии I $r(x)\geq 0$;

2) при условии II $r(x)\geq r_0>0$;

3) при условии III $r(x)\geq 0$, $\alpha\leq 0$, $\beta\geq 0$ и хоть в одном из двух последних неравенств стоит знак строгого неравенства (иначе это задача II).

Доказательство

1. Если $u\in D(A)$ при условии I, то $u(-1)=0$, и поэтому

$$|u(x)|=\left|\int_{-1}^{x}u'(t)dt\right|\leq\left[\int_{-1}^{x}[u'(t)]^2dt\right]^{1/2}\left[\int_{-1}^{x}dt\right]^{1/2},$$

$$[u(x)]^2\leq\int_{-1}^{1}[u'(t)]^2dt\,(1+x),$$

$$\|u\|^2=\int_{-1}^{1}[u(x)]^2dx\leq\int_{-1}^{1}[u'(t)]^2\,dt\int_{-1}^{1}(1+x)dx=$$

$$=2\int_{-1}^{1}[u'(t)]^2dt\leq\frac{2}{p_0}\int_{-1}^{1}[pu'^2+ru^2]dt=\frac{2}{p_0}I(u,u)=\frac{2}{p_0}(Au,u).$$

Итак, $(Au,u)\leq\dfrac{p_0}{2}(u,u)$, и положительная определенность A при граничном условии I доказана.

2. При граничном условии II $r(x)\geq r_0>0$, и потому

$$(u,u)=\int_{-1}^{1}u^2dx\leq\frac{1}{r_0}\int_{-1}^{1}ru^2dx\leq\frac{1}{r_0}\int_{-1}^{1}(pu'^2+ru^2)dx=\frac{1}{r_0}I(u,u)=\frac{1}{r_0}(Au,u),$$

$$(Au,u)\geq r_0(u,u).$$

3. При рассмотрении граничной задачи III будем для определенности считать $\beta > 0$. Имеем

$$u(x) = u(1) - \int_x^1 u'(t)dt,$$

$$[u(x)]^2 \le 2\left[\left(\int_x^1 u'(t)dt\right)^2 + u^2(1)\right] \le 2\left[2\int_{-1}^1 u'^2\,dt + u^2(1)\right],$$

$$(u,u) = \int_{-1}^1 u^2 dx \le 8\int_{-1}^1 u'^2\,dt + 4u^2(1) \le \frac{8}{p_0}\int_{-1}^1 (pu'^2 + ru^2)dt +$$

$$+ \frac{4}{p(1)\beta}\left[\beta p(1)u^2(1) - \alpha p(-1)u^2(-1)\right] \le c(Au,u),$$

где $c = \max\left(8/p_0,\ 4/(p(1)\beta)\right)$. Этим и при условии III положительная определенность оператора A доказана. ■

В дальнейшем мы всегда будем считать выполненными указанные условия, при которых оператор A положительно определен, не оговаривая этого каждый раз.

Свойство 3

Для любой функции $u \in D(A)$ выполняются неравенства

$$\|u\|_C \le c_1\|u\|_A, \quad \|u'\|_{L_2} \le c_2\|u\|_A. \tag{1}$$

Доказательство

Второе из доказываемых неравенств очевидно при $c_2 = 1/\sqrt{p_0}$. Докажем первое. Из представления

$$u(x) = u(0) + \int_0^x u'(t)dt$$

видно, что

$$|u(x)| \le |u(0)| + \|u'\|_{L_2}, \quad |u(x)| \ge |u(0)| - \|u'\|_{L_2}.$$

Из последнего неравенства вытекает, что

$$\|u\|_{L_2} \ge \sqrt{2}\left(|u(0)| - \|u'\|_{L_2}\right),$$

откуда $|u(0)| \le \frac{1}{\sqrt{2}} \|u\|_{L_2} + \|u'\|_{L_2}$ и $|u(x)| \le \frac{1}{\sqrt{2}} \|u\|_{L_2} + 2\|u'\|_{L_2}$.

Поскольку правая часть этой оценки не зависит от x и каждое слагаемое в ней оценивается через норму $\|u\|_A$, умноженную на некоторую постоянную, доказательство можно считать завершенным. ■

Обозначим через W_2^1 множество абсолютно непрерывных функций, заданных на $[-1, 1]$, производная которых суммируема с квадратом. После введения нормы $\|u\|^2_{W_2^1} = \|u'\|^2_{L_2} + \|u\|^2_{L_2}$ W_2^1 становится полным линейным нормированным пространством. Заметим, что рассуждения, проведенные при доказательстве свойства 3, показывают, что для любой функции $u \in W_2^1$

$$|u(x)| \le \frac{1}{\sqrt{2}} \|u\|_{L_2} + 2\|u'\|_{L_2},$$

и потому $\|u\|_C \le \frac{3}{\sqrt{2}} \|u\|_{W_2^1}$. ■

Свойство 4

При первом граничном условии энергетическое пространство H_A состоит из тех функций класса W_2^1, которые удовлетворяют этому граничному условию, т. е. обращаются в нуль на концах промежутка (это подпространство W_2^1 будем обозначать через $\overset{0}{W}{}_2^1$). При граничных условиях II и III пространство H_A совпадает по своему составу с W_2^1. Нормы в пространствах H_A и W_2^1 эквивалентны. Для всех функций $u, \upsilon \in H_A$ сохраняется равенство

$$[u, \upsilon] = I(u, \upsilon) + Q(u, \upsilon) \tag{2}$$

и оценки (1).

Доказательство

1. Покажем, что любая функция u из H_A принадлежит W_2^1. Пусть $\{u_k\}$ — последовательность функций из $D(A)$, сходящаяся в H_A к u. Из неравенств (1) следует, что последовательность $\{u_k\}$ сходится равномерно (разумеется,

к той же функции u), а последовательность производных $\{u'_k\}$ сходится в L_2 к некоторой функции $\upsilon \in L_2$. Переходя к пределу в равенстве

$$u_k(x) = u_k(0) + \int_0^x u'_k(t)dt,$$

получаем

$$u(x) = u(0) + \int_0^x \upsilon(t)dt.$$

Из этого представления ясно, что $u'(x) = \upsilon(x)$ и $u \in W_2^1$. При граничном условии I, кроме того, $u(\pm 1) = \lim u_k(\pm 1) = 0$, т. к. при всех k $u_k(\pm 1) = 0$, так что $u \in \overset{0}{W}{}_2^1$.

Предельный переход в неравенствах (1), написанных для функций u_k, дает: $\|u\|_C \le c_1\|u\|_A$, $\|u'\|_{L_2} \le c_2\|u\|_A$, т. е. неравенства (1) выполняются не только для функций из $D(A)$, но и для всех функций из H_A. Из неравенств (1) теперь легко следует, что для любой функции $u \in H_A$ будет $\|u\|_{W_2^1} \le \sqrt{2c_1^2 + c_2^2}\,\|u\|_A$. Точно так же предельным переходом, используя не только последовательность $\{u_k\}$, но и $\{\upsilon_k\}$ $(\upsilon_k \in D(A), \|\upsilon_k - \upsilon\|_A \to 0)$, получается равенство (2) для всех $u, \upsilon \in H_A$.

2. Покажем, что при условии I каждая функция u из $\overset{0}{W}{}_2^1$, а при условиях II или III каждая функция из W_2^1 принадлежит H_A. Для этого согласно *теореме 1 из § 6* достаточно убедиться в существовании такой постоянной $c = c(u)$, что для любой функции $\upsilon \in D(A)$ будет $|(A\upsilon, u)| \le c|\upsilon|$. Используя при условии I равенство $u(\pm 1) = 0$, а при граничных условиях II или III то, что функция υ из $D(A)$ удовлетворяет этому условию, и, интегрируя по частям, получаем

$$(A\upsilon, u) = I(u, \upsilon) + Q(u, \upsilon), \tag{3}$$

где $I(u, \upsilon)$ и $Q(u, \upsilon)$ имеют прежнее значение. Ясно, что $I(u, \upsilon) + Q(u, \upsilon)$ — симметричная билинейная форма, заданная на W_2^1 и такая, что $I(u, u) + Q(u, u) \ge 0$.

Поэтому

$$\left|(A\upsilon, u)\right| = \left|I(u,\upsilon) + Q(u,\upsilon)\right| \le \sqrt{I(u,u) + Q(u,u)}\sqrt{I(\upsilon,\upsilon) + Q(\upsilon,\upsilon)} = c(u)\left\|\upsilon\right\|_A.$$

Этим показана принадлежность функции u из W_2^1 (или соответственно $\overset{0}{W}{}_2^1$) пространству H_A.

3. Оценка $\left\|u\right\|_A \le c_2\left\|u\right\|_{W_2^1}$ легко вытекает из равенства $\left\|u\right\|_A^2 = I(u,u) + Q(u,u)$. ■

❑ **Замечание 1.** Подчеркнем существенную разницу между граничным условием I, с одной стороны, и условиями II и III — с другой. Все функции, входящие в энергетическое пространство H_A при условии I, обязаны удовлетворять этому условию, а при условиях II и III — нет. По существу, это вызвано тем, что функционал вычисления производной в конце промежутка (он входит в условия II и III) не является непрерывным в энергетическом пространстве. Разница между граничными условиями при доказательстве вложения $\overset{0}{W}{}_2^1$ или W_2^1 в H_A проявилась в том, что для получения равенства (3) при первом условии мы использовали граничное условие для функции u, принадлежность которой H_A хотели показать, а при условиях II и III — для функции υ, которая из $D(A)$, а потому этому условию заведомо удовлетворяет. Граничные условия, которым удовлетворяют функции из области задания оператора, но не обязаны удовлетворять функции из энергетического пространства, называются *естественными*. Таким образом, II и III — естественные граничные условия. Граничное условие I, сохраняющееся для функций из H_A, называется *основным*.

❑ **Замечание 2.** Следствием доказанного свойства 4 является то, что энергетическому пространству принадлежат кусочно-дифференцируемые функции. Так мы называем непрерывные функции, непрерывно дифференцируемые везде, за исключением, быть может, конечного числа точек, в которых производная может иметь разрывы первого рода. При граничном условии I эти функции должны, конечно, обращаться в нуль на концах промежутка.

Перейдем теперь непосредственно к методу Ритца. Как это следует из *теоремы 3 § 6*, для сходимости метода Ритца достаточно, чтобы координатная

система $\left\{\omega_k^{(n)}\right\}$ была полна в пространстве H_A. Из неравенств (1) при этом следует, что сходимость в H_A означает равномерную сходимость самих приближенных решений к точному и сходимость в L_2 первых производных. Остановимся более подробно на выборе координатной системы для граничной задачи I. Теперь будем считать, что в качестве координатных выбираются первые n функций некоторой бесконечной системы:

$$\omega_k^{(n)} = \omega_k.$$

Свойство 5

В случае граничного условия I оператор A полусходен с оператором B, заданным на множестве дважды непрерывно дифференцируемых функций, удовлетворяющих условию $u(-1) = u(1) = 0$, формулой $Bu = -u''$.

Доказательство

Согласно свойству 4 пространства H_A и H_B по своему составу совпадают с $\overset{0}{W}{}_2^1$, и нормы в этих пространствах эквивалентны норме в W_2^1. ■

Свойство 6

При граничном условии I координатная система, имеющая вид $\omega_k(x) = (1-x^2)p_{k-1}(x)$, где $p_\nu(x)$ — многочлен степени ν, полна в H_A. Такую координатную систему мы будем называть *полиномиальной*.

Доказательство

Как было показано в *§ 5*, система $\{A\omega_k\}$ полна в L_2, и остается воспользоваться *теоремой 4 из § 6*. ■

Свойство 7

При граничном условии I координатная система

$$\omega_k(x) = a_k \sin k\pi \frac{1+x}{2} \quad (a_k \neq 0)$$

полна в H_A. Эту систему мы будем называть *тригонометрической*.

Доказательство

$B\omega_k = -\omega''_k = a_k\left(\frac{k\pi}{2}\right)^2 \sin\frac{k\pi}{2}(1+x)$. Полнота системы $\{B\omega_k\}$ в $L_2(-1,1)$ равносильна полноте системы $\{\sin k\pi\}$ в $L_2(0,\pi)$, которая хорошо известна. Из полноты системы $\{B\omega_k\}$ в L_2 следует полнота системы $\{\omega_k\}$ в H_B *(см. теорему 4 из § 6)*, а значит и в H_A. ■

Отметим частные случаи указанных в свойствах 6 и 7 координатных систем, которые обеспечивают устойчивость метода.

Свойство 8

Если

$$\omega_k(x) = \int_{-1}^{x} \hat{P}_k(t)\,dt,$$

где $\hat{P}_k(x) = \frac{1}{2^k k!}\sqrt{\frac{2k+1}{2}}\left[(x^2-1)^k\right]^{(k)}$ — полином Лежандра, или

$$\omega_k(x) = \frac{2}{k\pi}\sin\frac{k\pi}{2}(1+x),$$

то метод Ритца устойчив в том смысле, как это утверждается в следствии 2 предыдущего параграфа.

Доказательство

Первая из названных систем действительно является частным случаем указанной в свойстве 7 — $\omega_k(x)$ есть многочлен степени $k+1$, обращающийся в нуль в точках ± 1 (в точке 1 ввиду ортогональности $\hat{P}_k$ постоянным на промежутке $[-1,1]$). Указанные системы очевидным образом ортонормальны в пространстве H_B, и наше утверждение сразу же вытекает из упомянутого выше следствия. ■

Полезно заметить, что первая из указанных в свойстве 8 систем допускает представление $\omega_k(x) = \alpha_k(1-x^2)J_{k-1}^{(1,1)}(x)$, где $J_\nu^{(1,1)}(x)$ — многочлены Якоби, ортогональные с весом $(1-x^2)$. Это представление удобно тем, что по-

зволяет вычислять значения координатных функций ω_k по рекуррентным формулам.

Приведем без доказательства свойства полиномиальной и тригонометрической координатных систем, связанных с быстротой сходимости. Эти свойства свидетельствуют об определенном преимуществе полиномиальной системы перед тригонометрической.

Свойство 9

Если коэффициент p дифференциального уравнения $\nu+1$ раз непрерывно дифференцируем, а r и f — ν раз (тем самым решение u^* — $\nu+2$ раза) и если координатная система полиномиальна, то быстрота сходимости метода Ритца характеризуется соотношением

$$\left\| u^* - u_n^* \right\|_A = O\left(\frac{1}{n^{\nu+1}} \right). \tag{4}$$

Доказательство этого факта в действительности требует не новых сведений о методе Ритца, а лишь применения аппроксимационной теоремы Джексона.

Тригонометрическая координатная система обладает тем недостатком, что для всех ее членов $\omega''_k(\pm 1) = 0$. Поэтому при всех n и приближенное решение удовлетворяет условиям $u_n^{*\prime\prime}(\pm 1) = 0$, которым точное решение, вообще говоря, удовлетворять не обязано. Поэтому при такой координатной системе сходимость метода Ритца обычно не может быть очень быстрой.

Свойство 10

Если точное решение u^* граничной задачи I таково, что выполняется хоть одно из неравенств $u^{*\prime\prime}(-1) \neq 0$ или $u^{*\prime\prime}(1) \neq 0$, а применяемая координатная система — тригонометрическая, то существует такая постоянная $c > 0$, что $\left\| u^* - u_n^* \right\|_A \geq c/n\sqrt{n}$.

Не останавливаясь подробно на выборе координатных систем в случае граничных условий II и III, отметим только, что координатные функции здесь не обязаны подчиняться граничным условиям. В частности, полиномиальная координатная система $\omega_k(x) = p_{k-1}(x)$, где $p_\nu(x)$ есть полином степени ν,

обеспечивает сходимость метода Ритца с быстротой, характеризуемой тем же соотношением (4), что и для первой граничной задачи.

Обратим внимание еще на одно обстоятельство, имея в виду для определенности граничную задачу II. Если координатные функции ω_k, которые мы считаем сейчас дважды непрерывно дифференцируемыми, удовлетворяют граничному условию II, то коэффициенты системы линейных уравнений метода Ритца можно вычислять по любой из следующих двух формул:

$$[\omega_k, \omega_j] = \int_{-1}^{1} (p\omega'_k, \omega'_j + r\omega_k\omega_j)\,dx$$

или

$$[\omega_k, \omega_j] = \int_{-1}^{1} [-(p\omega'_k)' + r\omega_k]\omega_j\,dx$$

Правые части этих формул равны. Если же ω_k условию II не удовлетворяют, то считать можно только по первой формуле — вторая даст неверный результат.

При изучении полноты системы $\left\{\omega_k^{(n)}\right\}$, выбранной для второй или третьей граничной задачи, и исследовании устойчивости метода Ритца в этом случае полезно иметь в виду следующее свойство.

Свойство 11

Любые два оператора A_1 и A_2, определяемые соответственно дифференциальными выражениями

$$L_i(u) = -(p_i u') + r_i u$$

и граничными условиями

$$u'(-1) + \alpha_i u(-1) = u'(1) + \beta_i u(1) = 0 \quad (i = 1, 2),$$

являются полусходными, если только выполнены условия положительной определенности этих операторов, указанные в свойстве 2.

Доказательство сразу же следует из того, что H_{A_i} совпадают по своему составу с W_2^1, и нормы в H_{A_i} эквивалентны норме в W_2^1 (см. свойство 4).

Отметим, что в этом утверждении не исключен случай $\alpha_i = \beta_i = 0$, так что граничные условия для одного из операторов A_i или для обоих могут быть типа II.

§ 8. Применение метода Ритца к дифференциальным уравнениям эллиптического типа

Пусть Ω — ограниченная область в m-мерном пространстве с достаточно гладкой границей $\partial\Omega$. Будем рассматривать в Ω эллиптическое дифференциальное уравнение

$$L(u) = -\sum_{i,j=1}^{m} \frac{\partial}{\partial x_i}\left(a_{ij} \frac{\partial u}{\partial x_j} \right) + ru = f. \qquad (1)$$

Не заботясь о минимальности налагаемых на уравнение требований, будем считать, что коэффициенты $a_{ij}(x) = a_{ji}(x)$ непрерывно дифференцируемы, а r и f непрерывны. Эллиптичность (1) означает существование такой постоянной $\nu > 0$, что для любых чисел ξ_i $(i = 1, 2, ..., m)$ и всех точек $x \in \Omega$ выполняется неравенство

$$\sum_{i,j=1}^{m} a_{ij}(x)\xi_i\xi_j \geq \nu \sum_{i=1}^{m} \xi_i^2.$$

Обозначим еще через μ такую постоянную, что

$$\sum_{i,j=1}^{m} a_{ij}(x)\xi_i\xi_j \leq \mu \sum_{i=1}^{m} \xi_i^2.$$

Уравнение (1) будем рассматривать при одном из следующих граничных условий:

I. $u\big|_{\partial\Omega} = 0$ (задача Дирихле).

II. $\partial u / \partial N\big|_{\partial\Omega} = 0$ (задача Неймана).

III. $(\partial u / \partial N + \sigma u)\big|_{\partial\Omega} = 0$ (третья задача).

Здесь $\dfrac{\partial u}{\partial N} = \sum_{i,j=1}^{m} a_{ij} \dfrac{\partial u}{\partial x_i} \cos(n, x_j)$ — производная по конормали (n — внешняя нормаль к границе области), а σ — непрерывная функция точки границы $\partial\Omega$.

Для каждой из трех поставленных граничных задач мы введем оператор A, действующий в пространстве $L_2(\Omega)$. За область его задания возьмем множе-

ство всех дважды непрерывно дифференцируемых функций, удовлетворяющих соответствующему граничному условию, и для $u \in D(A)$ положим $Au = L(u)$. Изучим свойства оператора A.

Свойство 1

Оператор A симметричен, причем для $u, \upsilon \in D(A)$ справедлива формула

$$(Au, \upsilon) = I(u, \upsilon) + Q(u, \upsilon), \tag{2}$$

где

$$I(u,\upsilon) = \int_\Omega \left[\sum_{i,j=1}^{m} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial \upsilon}{\partial x_j} + ru\upsilon \right] dx,$$

$$Q(u,\upsilon) = -\int_{\partial\Omega} \frac{\partial u}{\partial N} \upsilon dS = \begin{cases} 0 & \text{при граничных условиях I, II,} \\ \int_{\partial\Omega} \sigma u \upsilon dS & \text{при граничном условии III.} \end{cases}$$

Доказательство

Формула (2) получается сразу же применением формулы Остроградского к интегралу (Au, υ). Симметричность оператора A следует из (2). ■

Для исследования положительной определенности оператора A нам потребуется следующее известное неравенство Фридрихса[1]: для области Ω существует такая постоянная c_F, что для любой непрерывно дифференцируемой функции u будет

$$\int_\Omega u^2 dx \le c_F \left[\int_\Omega \sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i} \right)^2 dx + \int_{\partial\Omega} u^2 dS \right].$$

Свойство 2

Оператор A положительно определен, если выполнены следующие условия (свои для каждой граничной задачи):

1) при граничном условии I $r(x) \ge 0$;

2) при граничном условии II $r(x) \ge r_0 > 0$;

[1] По поводу доказательства этого и приводимого ниже неравенства Пуанкаре см., например, [15].

3) при граничном условии III либо а) $r(x) \ge r_0 > 0$, $\sigma \ge 0$, либо б) $r \ge 0$, $\sigma \ge \sigma_0 > 0$.

Доказательство

1. Для функции u, удовлетворяющей условию I, второе слагаемое в правой части неравенства Фридрихса обращаеся в нуль. Поэтому имеем

$$(Au, u) = I(u, u) \ge \nu \int_\Omega \sum_{i=1}^m \left(\frac{\partial u}{\partial x_i} \right)^2 dx \ge \frac{\nu}{c_F} \int_\Omega u^2 dx = \frac{\nu}{c_F} (u, u).$$

2. При граничном условии II имеем

$$(Au, u) = I(u, u) \ge \int_\Omega r u^2 dx \ge r_0 \int_\Omega u^2 dx = r_0 (u, u).$$

3 *а.* Доказательство положительной определенности оператора A для задачи III при условии а) доказывается точно так же, как и для задачи II.

3 *б.* Используя опять неравенство Фридрихса, имеем

$$(Au, u) = I(u, u) + Q(u, u) \ge \nu \int_\Omega \sum_{i=1}^m \left(\frac{\partial u}{\partial x_i} \right)^2 dx + \sigma_0 \int_{\partial\Omega} u^2 dS \ge$$

$$\ge \min(\nu, \sigma_0) \left[\int_\Omega \sum_{i=1}^m \left(\frac{\partial u}{\partial x_i} \right)^2 dx + \int_{\partial\Omega} u^2 dS \right] \ge \frac{\min(\nu, \sigma_0)}{c_F} \int_\Omega u^2 dx = c(u, u).$$

■

Сформулированные в свойстве 2 условия не охватывают один важный частный случай — именно задачу Неймана при $r \equiv 0$. Это не случайно, поскольку функция $u \equiv \text{const}$ удовлетворяет (при $r \equiv 0$) как уравнению (1), так и граничному условию II. В то же время, как известно, необходимым условием разрешимости этой задачи Неймана является равенство нулю среднего значения функции f:

$$\int_\Omega f \, dx = 0. \tag{3}$$

Если это условие выполнено, то решение задачи, определенное с точностью до постоянного слагаемого, можно выбрать так, чтобы его среднее значение тоже было равно нулю. Это позволяет нам ввести для этой частной задачи

оператор A специальным образом, не так, как это делалось в других случаях. Обозначим через H_0 подпространство пространства L_2, состоящее из тех функций, которые удовлетворяют условию (3). Очевидно, что H_0 — гильбертово пространство. На множестве $D(A) \subset H_0$ дважды непрерывно дифференцируемых функций, удовлетворяющих условию II, определим теперь оператор A, положив $Au = L(u)$. Оператор A симметричен и удовлетворяет равенству (2) (при $r \equiv 0$ и $Q(u, \upsilon) = 0$), и, кроме того, выполняется следующее свойство.

Свойство 3

Оператор A, заданный, как только что указано, положительно определен.

Доказательство

Доказательство этого утверждения сразу же вытекает из известного неравенства Пуанкаре, которое выполняется для всех непрерывно дифференцируемых функций:

$$\int_\Omega u^2 dx \le c_P \left(\int_\Omega \sum_{i=1}^m \left(\frac{\partial u}{\partial x_i} \right)^2 dx + \left(\int_\Omega u \, dx \right)^2 \right).$$

Для функций из H_0 второе слагаемое в его правой части обращается в нуль, и для $u \in D(A)$

$$(Au, u) = I(u, u) \ge \nu \int_\Omega \sum_{i=1}^m \left(\frac{\partial u}{\partial x_i} \right)^2 dx \ge \frac{\nu}{c_P} (u, u). \quad \blacksquare$$

Говоря впредь о I, II и III граничных задачах, мы всегда будем считать выполненными условия, указанные в свойстве 2 и обеспечивающие положительную определенность оператора A. Специально будем иметь также в виду задачу Неймана при $r \equiv 0$, считая оператор A заданным так, как только что указано.

Сейчас нам потребуется пространство С. Л. Соболева $W_2^1 = W_2^1(\Omega)$. Напомним, что это — пространство функций, имеющих в Ω обобщенные первые

производные, суммируемые с квадратом, причем норма в этом пространстве определяется равенством

$$\|u\|^2_{W_2^1} = \sum_{i=1}^{m} \left\| \frac{\partial u}{\partial x_i} \right\|^2_{L_2} + \|u\|^2_{L_2} = \int_\Omega \left(\sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i} \right)^2 + u^2 \right) dx.$$

Будем обозначать еще через $\overset{0}{W}{}_2^1$ подпространство пространства $\overset{0}{W}{}_2^1$, состоящее из тех функций, которые обращаются в нуль на границе $\partial\Omega$. В каком смысле понимать это обращение в нуль на границе, требует, конечно, уточнения. Не вдаваясь в подробности, отметим, что $\overset{0}{W}{}_2^1$ есть замыкание в пространстве W_2^1 множества непрерывно дифференцируемых функций, обращающихся в нуль на границе.

Свойство 4

Для всех функций $u \in D(A)$ норма в пространстве H_A эквивалентна норме в W_2^1, т. е. существуют такие постоянные $0 < d_1 < d_2$, что для всех $u \in D(A)$

$$d_1 \|u\|_{W_2^1} \le \|u\|_A \le d_2 \|u\|_{W_2^1}. \qquad (4)$$

Доказательство

1. Как уже было показано,

$$\int_\Omega \sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i} \right)^2 dx \le \frac{1}{\nu} I(u,u) \le \frac{1}{\nu} \|u\|^2_A,$$

и ввиду положительной определенности оператора A $\|u\|^2_{L_2} \le c_1 \|u\|^2_A$. Этим неравенство $\|u\|_{W_2^1} \le \frac{1}{d_1} \|u\|^2_A$ (при $\frac{1}{d_1^2} = \frac{1}{\nu} + c_1$) доказано.

2. Имеем $I(u,u) \le \mu \int_\Omega \sum_{i=1}^{m} \left(\frac{\partial u}{\partial x_i} \right)^2 dx + \max_\Omega r \|u\|^2_{L_2} \le c_2 \|u\|^2_{W_2^1}$,

где $c_2 = \max(\mu, \max r)$. Для граничных задач I и II (в том числе и в особом случае $r \equiv 0$) этим неравенство $\|u\|_A \le d_2 \|u\|_{W_2^1}$ доказано, поскольку $\|u\|_A^2 = I(u,u)$. В случае граничной задачи III следует еще отметить, что

$$Q(u,u) \le \max_{\partial\Omega} \sigma \int_{\partial\Omega} u^2 dS \le c_3 \|u\|_{W_2^1}^2 .$$

Это неравенство является частным случаем так называемых *теорем вложения*, и его доказательство выходит за рамки нашей книги. ■

Свойство 5

Для граничной задачи I выполняется равенство $H_A = \overset{0}{W}{}_2^1$, для граничных задач II и III — $H_A = W_2^1$, для граничной задачи II при $r \equiv 0$ — $H_A = W_2^1 \cap H_0$.

Доказательство

1. Поскольку $D(A) \subset W_2^1$ и H_A есть пополнение $D(A)$ по норме, эквивалентной норме в W_2^1, то ввиду полноты W_2^1 включение $H_A \subseteq W_2^1$ очевидно. Ясно также, что в случае задачи Дирихле $H_A \subseteq \overset{0}{W}{}_2^1$ и в случае задачи Неймана с $r \equiv 0$ $H_A \subseteq W_2^1 \cap H_0$.

2. Пусть теперь u — произвольная функция из $\overset{0}{W}{}_2^1$, W_2^1 или $W_2^1 \cap H_0$ в соответствии с рассматриваемым граничным условием. Для того чтобы показать принадлежность u энергетическому пространству, достаточно убедиться *(согласно теореме 1 из § 6 главы 3)* в существовании такой постоянной $c = c(u)$, что для любой функции $\upsilon \in D(A)$ выполняется неравенство $|(A\upsilon, u)| \le c\|\upsilon\|_A$. Заметим, что при вычислении интеграла $(A\upsilon, u)$ для дважды непрерывно дифференцируемой функции υ и функции $u \in W_2^1$ по-прежнему возможно применение формулы Остроградского, так что, как и раньше, имеем

$$(A\upsilon, u) = I(\upsilon, u) + Q(\upsilon, u),$$

где, в частности,

$$Q(\upsilon, u) = -\int_{\partial\Omega} \frac{\partial \upsilon}{\partial N} u dS.$$

В случае граничного условия I $Q(\upsilon,u)=0$, т. к. функция $u \in \overset{0}{W_2^1}$ обращается на $\partial\Omega$ в нуль. Во всех остальных случаях мы используем то обстоятельство, что функция υ из множества $D(A)$ удовлетворяет граничному условию, и потому

$$Q(\upsilon, u) = \begin{cases} 0 & \text{при граничном условии II,} \\ \int_{\partial\Omega} \sigma u \upsilon dS & \text{при граничном условии III.} \end{cases}$$

Далее, если учесть, что $I(u,\upsilon)+Q(u,\upsilon)$ есть симметричная билинейная форма, заданная на W_2^1 и такая, что $I(u,u)+Q(u,u)\geq 0$ для всех u, то мы получим

$$(A\upsilon, u) = I(\upsilon, u) + Q(\upsilon, u) \leq \sqrt{I(u,u)+Q(u,u)}\sqrt{I(\upsilon,\upsilon)+Q(\upsilon,\upsilon)} = c(u)\| \upsilon \|_A.$$

Этим наше утверждение доказано полностью. ■

- **Замечание.** Неравенство (4), доказанное ранее для функций из $D(A)$, очевидным образом переносится теперь на все функции из H_A. Точно так же на все функции u, υ из H_A переносится формула $[u,\upsilon]=I(u,\upsilon)+Q(u,\upsilon)$.

Из *теоремы 3 § 6 главы 3* сразу же следует такое утверждение о сходимости метода Ритца.

Свойство 6

Если при граничном условии I координатная система $\left\{\omega_k^{(n)}\right\}$ полна в $\overset{0}{W_2^1}$, при условиях II и III — в W_2^1, для задачи Неймана при $r\equiv 0$ — в $W_2^1 \cap H_0$, то метод Ритца сходится:

$$\left\| u_n^* - u^* \right\|_{W_2^1} \to 0.$$

Остановимся несколько подробнее на выборе координатной системы для задачи Дирихле. К сожалению, доказательство сформулированных далее утверждений 7 и 8 выходит за рамки наших возможностей.

Свойство 7

Пусть функция $\omega \in \overset{0}{W}{}_2^1$ такова, что $\omega(x) > 0$ для внутренних точек $\omega \in \Omega$. Тогда множество функций вида $\omega(x)P(x)$, где P — всевозможные полиномы от m переменных, плотно в $\overset{0}{W}{}_2^1$.

Доказательство этого утверждения имеется в книге Л. В. Канторовича и В. И. Крылова [11, гл. IV, § 2].

Из свойства 7 вытекает полнота такой координатной системы, которую мы выписываем для определенности при $m = 2$:

$$\omega_1(x,y) = \omega(x,y), \quad \omega_2(x,y) = \omega(x,y)P_1^1(x,y), \quad \omega_3(x,y) = \omega(x,y)P_1^2(x,y),$$
$$\omega_4(x,y) = \omega(x,y)P_2^1(x,y), \quad \omega_5(x,y) = \omega(x,y)P_2^2(x,y),$$
$$\omega_6(x,y) = \omega(x,y)P_2^3(x,y), \quad \omega_7(x,y) = \omega(x,y)P_3^1(x,y), \ldots$$

Здесь $P_n^k(x,y)$ $(k = 1, 2, \ldots, n+1)$ — многочлены степени n, и многочлены 1, $P_1^1, P_1^2, \ldots, P_n^{n+1}$ линейно независимы. Построенную таким образом координатную систему будем называть полиномиальной.

Свойство 8

Пусть граница $\partial\Omega$ области Ω задается уравнением $\omega(x) = 0$, где функция ω обладает такими свойствами:

а) $\omega(x) > 0$ для внутренних точек $x \in \Omega$;

б) $\omega(x)$ ν раз непрерывно дифференцируема $(\nu \geq 2)$;

в) $grad\ \omega(x) \neq 0$ при $x \in \partial\Omega$.

Пусть u_n^* есть приближенное решение задачи Дирихле, построенное методом Ритца по полиномиальной координатной системе в виде ωP_n, где P_n — многочлен степени n от m переменных[1].

[1] Здесь мы отступили от принятого ранее способа нумерации приближенных решений: для построения u_n^* используется N_n координатных функций полиномиальной системы — столько, сколько существует различных одночленов степени не выше n от m переменных.

Тогда если точное решение u^* ν раз непрерывно дифференцируемо, то

$$\left\| u^* - u_n^* \right\|_{W_2^1} = 0\left(\frac{1}{n^{\nu-1}} \right).$$

Высказанное утверждение принадлежит И. Ю. Харрик [22].

Приведем несколько примеров построения функции $\omega(x, y)$, удовлетворяющей условиям, указанным в утверждении 7, для двумерного случая.

1. Пусть Ω есть круг единичного радиуса с центром в начале координат. Тогда можно положить $\omega(x, y) = 1 - x^2 - y^2$.

2. Пусть Ω есть квадрат $x, y \in [0, 1]$. Тогда полагаем $\omega(x, y) = xy(1-x)(1-y)$.

3. Пусть Ω есть криволинейная трапеция $a \le x \le b, \quad g(x) \le y \le h(x)$ (рис. 3). Тогда можно положить

$$\omega(x, y) = (x-a)(b-x)(y-g(x))(h(x)-y).$$

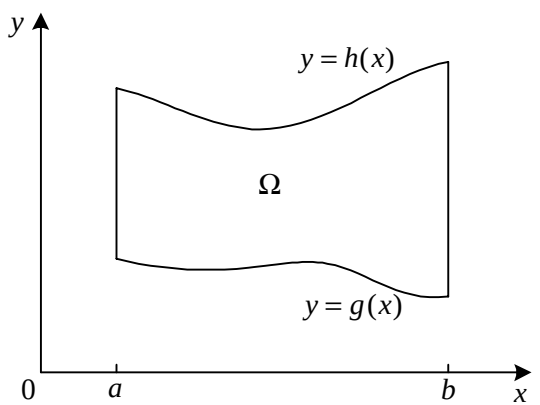


Рис. 3

4. Пусть Ω есть "уголок", изображенный на рис. 4. Трудность построения функции ω заключается здесь в том, что хотя она должна обращаться в нуль при $y = 1$, $1 \le x \le 2$, $\omega(x, y)$ не может содержать множитель $(y-1)$, т. к. при $y = 1$ и $0 < x < 1$ она должна быть положительна. Аналогичная ситуация и с

отрезком $x=1,\ 1\le y\le 2$. Для построения функции ω разобьем уголок прямой $y=x$ на две части Ω_1 и Ω_2 и положим

$$\omega(x,y)=\begin{cases}x(1-x)(2-y) & \text{при } x\le y,\\ y(1-y)(2-x) & \text{при } x\ge y.\end{cases}$$

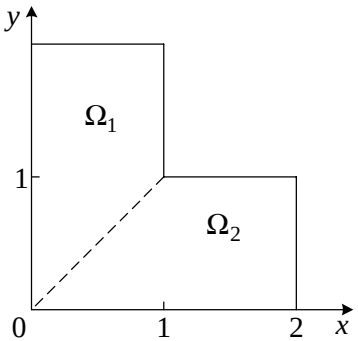


Рис. 4

Очевидно, что функция $\omega(x,y)$ непрерывна и обращается в нуль на границе. Первые производные ω на прямой $y=x$ терпят разрыв, но это не мешает ей принадлежать пространству W_2^1. Можно показать, однако, что разрыв производных функций ω является препятствием для быстрой сходимости метода Ритца, и предпочтительно выбрать функцию ω более гладкой, хотя и более сложной. Это можно сделать, например, так:

$$\omega(x,y)=xy(x-2)(y-2)[(y-1)^2+\kappa(x-1)][(x-1)^2+\kappa(y-1)],$$

где

$$\kappa(t)=\begin{cases}-t^3 & \text{при } t\le 0,\\ 0 & \text{при } t>0.\end{cases}$$

Отметим, что из всех приведенных примеров функций ω лишь первая удовлетворяет всем условиям, сформулированным в утверждении 8 (при любом натуральном ν). В примерах 2—4 нарушается условие в) в угловых точках.

Впрочем для областей с углами точное решение u^* обычно не обладает большой гладкостью, так что условия утверждения 8 не выполняются (при больших ν) еще и по этой причине.

Переходя к граничным задачам II и III, отметим, что координатные функции можно не подчинять (в силу утверждения 5) соответствующим граничным условиям. Это имеет для применения метода Ритца первостепенное значение, т. к. построить полную в H_A систему функций, удовлетворяющих условию II или III, обычно оказывается чрезвычайно сложным. Таким образом, для этих задач можно выбрать функции ω_k в виде многочленов. Подчеркнем, что в случае, например, граничной задачи III коэффициенты системы линейных уравнений метода Ритца можно считать лишь по формуле

$$[\omega_k^{(n)}, \omega_l^{(n)}] = \int_\Omega \left(\sum_{i,j=1}^{m} a_{ij} \frac{\partial \omega_k^{(n)}}{\partial x_i} \frac{\partial \omega_l^{(n)}}{\partial x_j} + r\omega_k^{(n)}\omega_l^{(n)} \right) dx + \int_\Omega \sigma \omega_k^{(n)} \omega_l^{(n)} dS,$$

но не в виде

$$\int_\Omega L(\omega_k^{(n)})\omega_j^{(n)} dx$$

— если координатные функции не удовлетворяют условию III, это выражение не совпадает с предыдущим.

Отметим еще, что координатные функции для задачи Неймана при $r \equiv 0$ можно не подчинять условию (3). Действительно, все элементы системы линейных уравнений (коэффициенты при неизвестных и правые части) получаются при этом такими же, как если бы мы из каждой координатной функции вычли ее среднее значение, а само точное решение все равно определено лишь с точностью до постоянного слагаемого.

§ 9. Метод Галеркина

В этом параграфе мы будем рассматривать уравнение

$$Au = Bu + Cu = f \tag{1}$$

в гильбертовом пространстве H.

Сделаем следующие предположения.

1. B — симметричный положительно определенный оператор с плотной в H областью задания $D(B)$ и множеством значений, совпадающим со всем H.

Предположение, что $R(B) = H$, не является сколько-нибудь жестким. Как было показано в *§ 6*, положительно определенный оператор с плотным в H множеством значений $R(B)$ всегда может быть расширен до оператора, имеющего множеством значений все H. Обратный оператор B^{-1}, как это следует из предположения 1, задан в пространстве H и ограничен. С оператором B связано энергетическое пространство H_B, скалярное произведение в котором — $[u, \upsilon]$, а норма — $\|u\|_B$.

Второе предположение относится к оператору C.

2. Область задания оператора C не у́же, чем оператора B: $D(C) \supseteq D(B)$ и потому $D(A) = D(B)$.

Рассмотрим теперь в пространстве H_B оператор $T_0 = B^{-1}C$. Этот оператор определен на множестве $H_B \cap D(C)$. Поскольку $D(T_0) \supseteq D(B)$, то $D(T_0)$ плотно в H_B.

Сделаем еще одно важное предположение.

3. Оператор T_0 ограничен и оператор T, являющийся его расширением по непрерывности на все пространство H_B, компактен.

Если u^* — решение уравнения (1), то, применяя к левой и правой частям тождества $Bu^* + Cu^* = f$ оператор B^{-1}, легко убедиться, что u^* является также решением уравнения

$$u + Tu = B^{-1}f, \tag{2}$$

рассматриваемого в пространстве H_B.

Может, однако, случиться так, что уравнение (2) имеет решение u_0, которое не входит в область задания оператора B, и потому не является решением уравнения (1). Условимся всякое решение уравнения (2) называть *обобщенным решением* уравнения (1). Как только что было отмечено, каждое решение уравнения (1) является и обобщенным его решением. Легко убедиться

также, что если обобщенное решение u_0 принадлежит $D(B)$, то это есть точное решение уравнения (1). Сделаем, наконец, еще одно предположение.

4. Уравнение (1) имеет единственное обобщенное решение.

Опишем теперь метод приближенного решения уравнения (1), о котором пойдет речь в этом параграфе. Приближенное решение u_n будем искать в виде $u_n = \sum_{k=1}^{n} c_k \omega_k^{(n)}$, где $\{\omega_k^{(n)}\}$ — координатная система (линейно независимая при каждом n), элементы которой принадлежат множеству $H_B \cap D(C)$. Коэффициенты c_k будем находить из условий

$$[u_n, \omega_j^{(n)}] + (Cu_n, \omega_j^{(n)}) = (f, \omega_j^{(n)}), \quad j = 1, 2, ..., n,$$

что немедленно приводит к системе линейных алгебраических уравнений относительно c_k:

$$\sum_{k=1}^{m} \left([\omega_k^{(n)}, \omega_j^{(n)}] + (C\omega_k^{(n)}, \omega_j^{(n)}) \right) c_k = (f, \omega_j^{(n)}), \quad j = 1, 2, ..., n, \tag{3}$$

Если координатные элементы $\omega_k^{(n)}$ принадлежат $D(B)$, то система уравнений (3) переписывается в виде

$$\sum_{k=1}^{m} (A\omega_k^{(n)}, \omega_j^{(n)}) c_k = (f, \omega_j^{(n)}), \quad j = 1, 2, ..., n, \tag{3'}$$

т. е. совпадает с той системой уравнений, к которой приводит метод Галеркина. Это и дало нам основание озаглавить параграф "Метод Галеркина". Однако возможность выбирать координатную систему из более широкого множества $H_B \cap D(C)$ часто оказывается чрезвычайно полезной, так что при исследовании метода ограничиваться случаем $\omega_k^{(n)} \in D(B)$ мы не будем.

Теорема 1 (С. Г. Михлин)

Пусть для уравнения (1) выполнены условия 1—4 и пусть координатная система $\{\omega_k^{(n)}\}$, такая, что $\omega_k^{(n)} \in H_B \cap D(C)$, полна в H_B. Тогда при достаточно больших n системы уравнений (3) однозначно разрешимы, и приближенные решения u_n^* сходятся к обобщенному решению u_0 уравнения (1) как по норме пространства H_B, так и по норме пространства H.

Доказательство

Очевидно, что для любого элемента $g \in H$ будет $(g, \omega_j^{(n)}) = [B^{-1}g, \omega_j^{(n)}]$. Поэтому система уравнений (3) может быть переписана в виде

$$\sum_{k=1}^{n} [(I+T)\omega_k^{(n)}, \omega_j^{(n)}]\, c_k = [B^{-1}f, \omega_j^{(n)}], \quad j = 1, 2, ..., n. \tag{4}$$

Но это не что иное, как система уравнений метода Галеркина для уравнения (2), которое мы рассматриваем, конечно, в пространстве H_B. Согласно предположению 4, оператор $I+T$ обратим. Поэтому для завершения доказательства остается сослаться на теорему о сходимости метода Галеркина для уравнений второго рода *(теорема 2 из § 3)*, которая позволит сделать заключение о сходимости приближенных решений к точному в пространстве H_B. Сходимость же в пространстве H следует из сходимости в H_B. ■

❑ **Замечание.** Быстрота сходимости метода Галеркина для уравнения (1), как и всегда в условиях теоремы Л. В. Канторовича, характеризуется соотношением

$$\left\| u_0 - u_n^* \right\|_B = O(E_{U^{(n)}}(u_0)_{H_B}).$$

Напомним, что

$$E_{U^{(n)}}(u_0)_{H_B} = \inf_{u_n \in U^{(n)}} \left\| u_0 - u_n \right\|_B$$

— наилучшее приближение точного решения u^* элементами подпространства $U^{(n)} \subset H_B$, натянутого на координатные элементы $\omega_1^{(n)}, \omega_2^{(n)}, ..., \omega_n^{(n)}$.

Относительно устойчивости метода Галеркина можно утверждать следующее.

Теорема 2

Пусть выполнены условия теоремы 1. Пусть D — оператор, полусходный с оператором B. Тогда найдутся такие постоянные c_1 и c_2, что при всех достаточно больших n для матрицы $\overline{A}_n$ коэффициентов системы уравнений (3) выполняются оценки

$$\left\| A_n \right\|_2 \le c_1 \left\| \Gamma_n^D \right\|_2, \quad \left\| A_n^{-1} \right\|_2 \le c_2 \left\| (\Gamma_n^D)^{-1} \right\|_2. \tag{5}$$

Здесь $\Gamma_n^D = \{[\omega_k^{(n)}, \omega_j^{(n)}]_D\}$ — матрица Грама координатной системы $\omega_1^{(n)}, ..., \omega_n^{(n)}$ в энергетическом пространстве H_D.

Доказательство

Согласно *теореме 2 из § 3*, при достаточно больших n справедливы неравенства

$$\|A_n\|_2 \le c_1' \|\Gamma_n^B\|_2, \quad \|A_n^{-1}\|_2 \le c_2' \|(\Gamma_n^B)^{-1}\|_2,$$

где $\Gamma_n^B = \{[\omega_k^{(n)}, \omega_j^{(n)}]_B\}$ — матрица Грама координатной системы в H_B. Но по *теореме 7 из § 6*

$$\|\Gamma_n^B\|_2 \le d_1^2 \|\Gamma_n^D\|_2, \quad \|(\Gamma_n^B)^{-1}\|_2 \le d_2^2 \|(\Gamma_n^D)^{-1}\|_2.$$

Здесь d_1 и d_2 — постоянные, связанные с полусходностью операторов B и D. Остается перемножить соответствующие оценки. ■

Переходя непосредственно к вопросу об устойчивости метода, отметим, что в качестве приближенного уравнения мы будем рассматривать систему уравнений (3) как уравнение в пространстве l_n^2. Что касается "точного" уравнения, то здесь представляются две возможности:

А) "точное" уравнение есть (1);

Б) "точное" уравнение есть (2).

Поскольку уравнение (2) есть уравнение второго рода в гильбертовом пространстве H_B, то в случае Б) достаточный признак устойчивости дает следствие из *теоремы 2 § 3* с учетом только что доказанной теоремы: если выполнены условия теоремы 2 и числа обусловленности матриц Γ_n^D ограничены $(\mu_2(\Gamma_n^D) \le C)$, то процессы отыскания каркасов и построения приближенных решений устойчивы. Обратимся к случаю А). Устойчивость процесса отыскания каркасов никак не связана с "точным" уравнением, и для устойчивости каркасов указанные выше условия очевидным образом остаются достаточными. Поскольку уравнения (1) и (2) рассматриваются в разных пространствах, то устойчивость процесса построения приближенных решений в случаях А) и Б) означает разное. Напомним, что оператор $\overline{\varphi}_n$ вектору $x = (\xi_1, ..., \zeta_n) \in l_n^2$ ставит в соответствие элемент $\overline{\varphi}_n x = \sum_{k=1}^n \varsigma_k \omega_k^{(n)}$. Пусть

$\|\overline{\varphi}_n\|$ есть его норма как оператора из l_n^2 в H, а $\|\overline{\varphi}_n\|_B$ — как оператора из l_n^2 в H_B, и так же обозначаются нормы возмущения $\Delta\overline{\varphi}_n$. Итак, в случае А) устойчивость процесса построения приближенных решений означает, что из соотношений $\delta A_n \to 0$, $\delta g_n \to 0$ (здесь g_n — вектор правых частей системы (3)) и $\|\Delta\overline{\varphi}_n\| / \|\overline{\varphi}_n\| \to 0$ вытекает, что $\|\Delta u_n\| / \|u^*\| \to 0$[1]. В случае же Б) из соотношений $\delta A_n \to 0$, $\delta g_n \to 0$ и $\|\Delta\overline{\varphi}_n\|_B / \|\overline{\varphi}_n\|_B \to 0$ должно следовать, что $\|\Delta u_n\|_B / \|u^*\|_B \to 0$. Однако при сделанных предположениях и в случае А) процесс построения приближенных решений оказывается устойчивым. Для доказательства этого следует обратиться к *теореме 2 из § 3 главы 1*. Ввиду устойчивости процесса построения приближенных решений в случае Б) ограничены произведения $\|\overline{\varphi}_n\|_B \cdot \|z_n^*\|$. Но, как легко видеть, $\|\overline{\varphi}_n\| \le \frac{1}{m}\|\overline{\varphi}_n\|_B$ (здесь m — постоянная, характеризующая положительную определенность оператора B), так что произведения $\|\overline{\varphi}_n\| \cdot \|z_n^*\|$ также ограничены, а это влечет устойчивость процесса построения приближенных решений в случае А).

Подведем итоги сказанному.

Теорема 3

Пусть выполнены условия теоремы 2 и пусть числа обусловленности матриц Γ_n^D ограничены: $\mu_2(\Gamma_n^D) \le C$. Тогда как в случае А), так и в случае Б), процессы отыскания каркасов и построения приближенных решений устойчивы.

В заключение параграфа, не вдаваясь в подробности, отметим, что условия 1—3, которые мы налагали на уравнение (1), выполняются, в частности, для "полного" эллиптического уравнения

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{m} b_i \frac{\partial u}{\partial x_i} + cu = f,$$

[1] В этом случае мы вынуждены предполагать существование точного, а не обобщенного решения.

рассматриваемого при одном из следующих граничных условий:

I. $u\big|_{\partial\Omega} = 0.$

II. $\partial u / \partial N\big|_{\partial\Omega} = 0.$

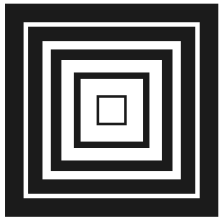
III. $\partial u / \partial N + \sigma u\big|_{\partial\Omega} = 0, \quad \sigma \geq 0.$

Для применения изложенной схемы следует положить

$$Bu = -\sum_{i,j=1}^{m} \frac{\partial}{\partial x_j}\left(a_{ij} \frac{\partial u}{\partial x_j} \right) + ru, \quad r \geq r_0 > 0$$

(в случае граничной задачи I, а также III при $\sigma \geq \sigma_0 > 0$ можно считать $r \equiv 0$). В случае граничных задач II и III функции, входящие в энергетическое пространство H_B, не обязаны подчиняться этим граничным условиям, и потому координатные функции $\omega_k^{(n)}$ также могут им не удовлетворять. Это сильно облегчает построение координатной системы. Подчеркнем еще, что если координатные функции не удовлетворяют граничным условиям (или не принадлежат области задания оператора B по другой причине), то система уравнений (3) не может быть записана в обычной форме метода Галеркина (3').

Глава 4



Метод сеток

§ 1. Сущность метода сеток

Метод сеток, видимо, является в настоящее время основным методом приближенного решения уравнений в частных производных. Этот параграф посвящен "кухне" метода — способам построения тех приближенных уравнений, которыми мы заменяем точное.

Метод сеток для обыкновенного дифференциального уравнения

Сущность метода сеток проще всего пояснить на примере простейшей граничной задачи для обыкновенного дифференциального уравнения

$$L(u) = pu'' + qu' + ru = f, \quad u(0) = a, \quad u(1) = b.$$

Будем считать, что эта задача однозначно разрешима и что ее решение $u^*(x)$ четырежды непрерывно дифференцируемо.

Напомним две известные формулы численного дифференцирования:

$$u'(x) = \frac{u(x+h) - u(x-h)}{2h} - \frac{1}{6}h^2 u'''(\xi) \tag{1}$$

(здесь функция u трижды непрерывно дифференцируема и $\xi \in (x-h,\, x+h)$),

$$u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} - \frac{1}{12}h^2 u^{\mathrm{IV}}(\eta) \tag{2}$$

(u четырежды непрерывно дифференцируема, $\eta \in (x-h, x+h)$). Зададимся некоторым натуральным числом n и введем обозначения

$$h = 1/n, \quad x_k = kh, \quad p_k = p(x_k), \quad q_k = q(x_k), \quad r_k = r(x_k), \quad f_k = f(x_k).$$

Записывая наше дифференциальное уравнение в точках x_k, используя формулы (1) и (2) и учитывая граничные условия, видим, что значения $u^*(x_k)$ решения в точках x_k (называемых *узлами*) удовлетворяют следующей системе линейных алгебраических уравнений:

$$u^*(x_0) = a,$$

$$p_k \frac{u^*(x_{k+1}) - 2u^*(x_k) + u^*(x_{k-1})}{h^2} + q_k \frac{u^*(x_{k+1}) - u^*(x_{k-1})}{2h} + r_k u^*(x_k) =$$

$$= f_k + h^2 \left[\frac{p_k}{12} u^{*\mathrm{IV}}(\eta_k) + \frac{q_k}{6} u^{*'''}(\xi_k) \right], \quad k = 1, 2, ..., n-1.$$

$$u^*(x_n) = b.$$

Правые части большинства из этих уравнений (всех, кроме первого и последнего) нам неизвестны, поскольку неизвестны $u^{*\mathrm{IV}}(\eta_k)$ и $u^{*'''}(\xi_k)$. Однако при достаточно большом n (т. е. достаточно малом h) мы можем считать слагаемые в правых частях, содержащие эти неизвестные нам величины, малыми. Отбрасывая их, мы получим систему линейных алгебраических уравнений относительно новых неизвестных $u_0, u_1, ..., u_n$, которые мы будем рассматривать как приближения к значениям искомой функции ($u_k \approx u^*(x_k)$):

$$\left.\begin{array}{c} u_0 = a, \\ p_k \dfrac{u_{k+1} - 2u_k + u_{k-1}}{h^2} + q_k \dfrac{u_{k+1} - u_{k-1}}{2h} + r_k u_k = f_k, \quad k = 1, 2, ..., n-1, \\ u_n = b. \end{array}\right\} \qquad (3)$$

Таким образом, метод сеток сводит нашу граничную задачу к системе линейных алгебраических уравнений (3). Такую систему уравнений, в каждом из которых неизвестными являются значения искомой функции в нескольких близких точках (узлах), называют *конечно-разностной*.

Аппроксимация дифференциальных выражений

Существенным моментом метода сеток, как это видно из предыдущего раздела, является аппроксимация производных функций или, более обще, диф-

ференциальных выражений посредством линейных комбинаций значений функции в некоторых точках. В этом разделе рассмотрим некоторые из возможных способов такой аппроксимации.

Использование формул численного дифференцирования

Как известно, задача численного дифференцирования решается обычно следующим образом. По выбранным узлам интерполяции для функции u строится интерполяционный полином, и значение интересующей нас производной функции в нужной точке полагается равным значению в этой точке, соответствующей производной интерполяционного полинома. Именно таким способом могут быть получены формулы (1) и (2). Останавливаться подробнее на этом способе мы не будем. Отметим только, что он применим и для аппроксимации любого дифференциального выражения. Пусть L — некоторый линейный дифференциальный оператор (быть может, в частных производных) и пусть P — интерполяционный полином функции u, построенный по узлам $M_1, M_2, ..., M_l$. Тогда можно положить приближенно $Lu\big|_Q = LP\big|_Q$. Ясно, что $LP\big|_Q$ есть линейная комбинация значений в точках $M_1, ..., M_l$ функции u. Коэффициенты этой линейной комбинации зависят лишь от точек M_k и Q и оператора L, но не от u.

Использование разложений в ряд Тейлора

Приведем пример применения этого способа. Пусть u — достаточно гладкая функция двух переменных, и нам требуется приблизить смешанную производную u''_{xy} в точке (x, y) через значения этой функции в четырех точках: $(x+h, y+h)$, $(x+h, y-h)$, $(x-h, y-h)$ и $(x-h, y+h)$. Составим разложения функции u в ряд Тейлора с центром в точке (x, y) для этих четырех точек:

$$u(x+h, y+h) = u + hu'_x + hu'_y + \frac{1}{2}h^2u''_{xx} + h^2u''_{xy} + \frac{1}{2}h^2u''_{yy} +$$

$$+ \frac{1}{6}h^3u'''_{x^3} + \frac{1}{2}h^3u'''_{x^2y} + \frac{1}{2}h^3u'''_{xy^2} + \frac{1}{6}h^3u'''_{y^3} + O(h^4).$$

Здесь все производные функции u вычисляются в точке (x, y), и остаточный член имеет четвертый порядок малости относительно h в предположении, что функция u четырежды непрерывно дифференцируема. Ясно, что разложения функции u в трех остальных точках будут отличаться от выписанного лишь знаками некоторых членов. Укажем эти знаки для каждой точки в табл. 4.1.

Таблица 4.1

Точка	u	u'_x	u'_y	u''_{xx}	u''_{xy}	u''_{yy}	u'''_{x^3}	u'''_{x^2y}	u'''_{xy^2}	u'''_{y^3}
$(x + h, y + h)$	+	+	+	+	+	+	+	+	+	+
$(x + h, y - h)$	+	+	–	+	–	+	+	–	+	–
$(x - h, y - h)$	+	–	–	+	+	+	–	–	–	–
$(x - h, y + h)$	+	–	+	+	–	+	–	+	–	+

План дальнейших действий, в сущности, должен быть таков. Нужно составить линейную комбинацию всех четырех разложений с неопределенными коэффициентами, приравнять полученные коэффициенты при u, u'_x, u'_y, u''_{x^2}, u''_{y^2} и некоторых следующих производных нулю, а коэффициент при u''_{xy} — единице и найти неопределенные коэффициенты из этих условий. Хотя число уравнений здесь окажется существенно больше числа неизвестных, система все же будет разрешимой благодаря большой симметрии разложений. Мы не будем выписывать соответствующую систему уравнений, а сразу отметим, что, как легко в этом убедиться, пользуясь приведенной выше таблицей знаков, справедливо равенство

$$u(x + h, y + h) - u(x + h, y - h) + u(x - h, y - h) - u(x - h, y + h) = \\ = 4h^2 u'''_{xy}(x, y) + O(h^4).$$

Отсюда

$$u''_{xy}(x, y) = \frac{u(x + h, y + h) - u(x + h, y - h) + u(x - h, y - h) - u(x - h, y + h)}{4h^2} + O(h_2).$$

Если бы мы записали остаточные члены разложений в ряд Тейлора более подробно, например, в форме Лагранжа, то и в полученной формуле имели бы не только указание порядка малости, но и выражение остаточного члена через значения четвертых производных функции u в некоторых точках.

Отметим, что приведенные выше формулы (1) и (2) также могут быть получены этим способом.

Так же как и первый метод, способ разложения в ряд Тейлора может применяться не только для аппроксимации каждой производной, входящей в дифференциальное выражение, в отдельности, но и для всего выражения в целом.

Последовательное применение формул

Выведем приближенную формулу вычисления u''_{xy} по тем же точкам, что и выше, другим способом, используя формулу (1):

$$u''_{xy} = \frac{\partial}{\partial x} u'_y \approx \frac{u'_y(x+h, y) - u'_y(x-h, y)}{2h}. \tag{4}$$

Но

$$u'_y(x+h, y) \approx \frac{u(x+h, y+h) - u(x+h, y-h)}{2h}, \tag{5}$$

$$u'_y(x-h, y) \approx \frac{u(x-h, y+h) - u(x-h, y-h)}{2h}. \tag{6}$$

Подставляя две последние приближенные формулы в первую, получаем

$$u''_{xy} \approx \frac{u(x+h, y+h) - u(x+h, y-h) - u(x-h, y+h) + u(x-h, y-h)}{4h^2}, \tag{7}$$

т. е. ту же самую приближенную формулу, что и выше, но только без указания порядка малости погрешности. Произведем теперь учет погрешностей, считая функцию u четырежды непрерывно дифференцируемой. Все формулы (4), (5) и (6) имеют погрешности порядка h^2. Однако при подстановке формул (5) и (6) в формулу (4) происходит деление их погрешностей на h. Поэтому создается впечатление, что полученная формула (7) имеет погрешность всего лишь порядка h, т. е. что происходит потеря точности. Однако эта потеря точности иллюзорна. Действительно, если использовать формулу (1) с

точным указанием погрешности, то мы легко получим, что разность между левой и правой частями приближенной формулы (7) есть

$$R = -\frac{1}{6}h^2 u^{\mathrm{IV}}_{x^3 y}(\xi, y) + \frac{1}{2h}\frac{h^2}{6}\left(u'''_{y^3}(x-h, \eta_1) - u'''_{y^3}(x+h, \eta_2)\right),$$
$$\left(\xi \in (x-h, x+h), \quad \eta_1, \eta_2 \in (y-h, y+h)\right).$$

Так как функция u четырежды непрерывно дифференцируема, то разность $u'''_{y^3}(x-h, \eta_1) - u'''_{y^3}(x+h, \eta_2)$ имеет порядок малости h, и потому $R = O(h^2)$.

Вообще стоит обратить внимание на то, что при последовательном применении формул численного дифференцирования потери порядка точности обычно не происходит.

В качестве еще одного примера повторного применения аппроксимационных формул получим приближение часто встречающегося выражения $(pu')'$:

$$(p(x)u'(x))' \approx \frac{p(x+h/2)u'(x+h/2) - p(x-h/2)u'(x-h/2)}{h},$$

$$u'(x+h/2) \approx \frac{u(x+h)-u(x)}{h}, \quad u'(x-h/2) \approx \frac{u(x)-u(x-h)}{h},$$

откуда

$$(p(x)u'(x))' \approx$$
$$\approx \frac{p(x+h/2)u(x+h) - [p(x+h/2) + p(x-h/2)]u(x) + p(x-h/2)u(x-h)}{h^2}.$$

Предоставляем читателю самому убедиться, что эта формула имеет погрешность порядка малости h^2 (в предположении, что функция u четырежды непрерывно дифференцируема, а функция p — трижды).

Более сложный пример аппроксимации дифференциального уравнения разностным

Целью аппроксимации дифференциальных выражений разностными в методе сеток является построение разностных уравнений, которым значения в узлах точного решения рассматриваемой задачи удовлетворяли бы с достаточной

точностью. Это полезно иметь в виду, поскольку иногда находятся способы построения таких уравнений, не сводящиеся к апроксимации дифференциальных выражений. Приведем такой пример.

Пусть функция $u(x, y)$ удовлетворяет дифференциальному уравнению

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f.$$

Требуется составить линейное уравнение, которому приближенно удовлетворяли бы значения функции u в точках (x, y), $(x-h, y)$, $(x+h, y)$, $(x, y-h)$, $(x, y+h)$.

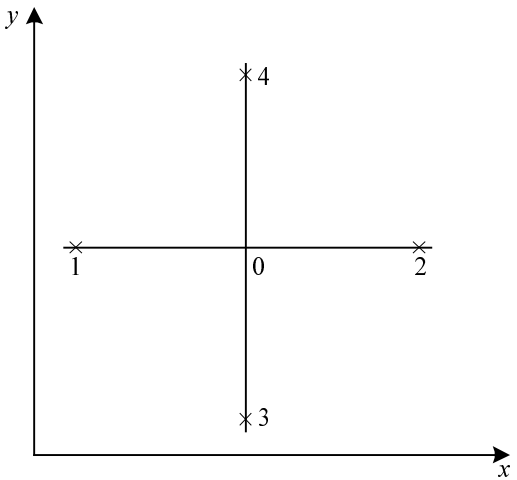


Рис. 5

Для более короткой записи перенумеруем эти точки, как показано на схеме (рис. 5), и значения функции в этих точках будем помечать соответствующим индексом. Функции u и f будем считать имеющими столько производных, сколько потребуется. Используя разложения в ряд Тейлора, легко получаем, что

$$u_1 - 2u_0 + u_2 = h^2 u''_{x^2} + \frac{h^4}{12} u^{\mathrm{IV}}_{x^4} + O(h^6),$$

$$u_3 - 2u_0 + u_4 = h^2 u''_{y^2} + \frac{h^4}{12} u^{\mathrm{IV}}_{y^4} + O(h^6),$$

откуда

$$u''_{x^2} - u''_{y^2} = \frac{u_1 + u_2 - u_3 - u_4}{h^2} - \frac{h^2}{12}(u^{\mathrm{IV}}_{x^4} - u^{\mathrm{IV}}_{y^2}) + O(h^4).$$

Но $u''_{x^2} - u''_{y^2} = f$, и потому

$$u^{\mathrm{IV}}_{x^4} - u^{\mathrm{IV}}_{y^4} = \frac{\partial^2}{\partial x^2}(u''_{x^2} - u''_{y^2}) + \frac{\partial^2}{\partial y^2}(u''_{x^2} - u''_{y^2}) = \Delta f,$$

где Δ — оператор Лапласа. Итак,

$$\frac{u_1 + u_2 - u_3 - u_4}{h^2} = f_0 + \frac{h^2}{12}\Delta f_0 + O(h^4),$$

и с точностью до величин четвертого порядка малости относительно h значения функции u в рассматриваемых точках удовлетворяют уравнению

$$\frac{u_1 + u_2 - u_3 - u_4}{h^2} = f_0 + \frac{h^2}{12}\Delta f_0.$$

Отметим еще, что согласно формуле (2)

$$\Delta f_0 = \frac{f_1 + f_2 + f_3 + f_4 - 4f_0}{h^2} + O(h^2).$$

Поэтому значения функции u удовлетворяют и уравнению

$$\frac{u_1 + u_2 - u_3 - u_4}{h^2} = \frac{2}{3}f_0 + \frac{1}{12}(f_1 + f_2 + f_3 + f_4)$$

также с точностью до величин порядка малости $O(h^4)$.

Метод сеток для эллиптического уравнения

Пусть в ограниченной двумерной области Ω с границей $\partial\Omega$ требуется найти решение эллиптического уравнения

$$-\Delta u + ru = f,$$

удовлетворяющее граничному условию

$$u\big|_{\partial\Omega} = g,$$

где g — функция, заданная на границе области. Выберем как-либо аргументы x_0 и y_0 и шаг $h > 0$ и проведем прямые $x = x_j$, $y = y_k$, где $x_j = x_0 + jh$, $y_k = y_0 + kh$. Точки пересечения этих прямых назовем узлами, так что каждому узлу соответствуют два индекса j и k. Часто вместо того, чтобы говорить об узле как точке, мы будем говорить о соответствующей ему паре индексов, называя эту пару также узлом. Множество I всех узлов, попавших в $\overline{\Omega}$, разобьем на два подмножества: $I = I_1 \cup I_2$. Узел (j, k) отнесем к множеству I_1, если все четыре узла $(j+1, k)$, $(j-1, k)$, $(j, k+1)$, $(j, k-1)$, "соседних" с узлом (j, k), также принадлежат $\overline{\Omega}$, и к множеству I_2 в противном случае. Узлы из I_1 будем называть внутренними (на рис. 6 они помечены крестиками), а из I_2 — узлами граничной полоски (они обозначены кружками).

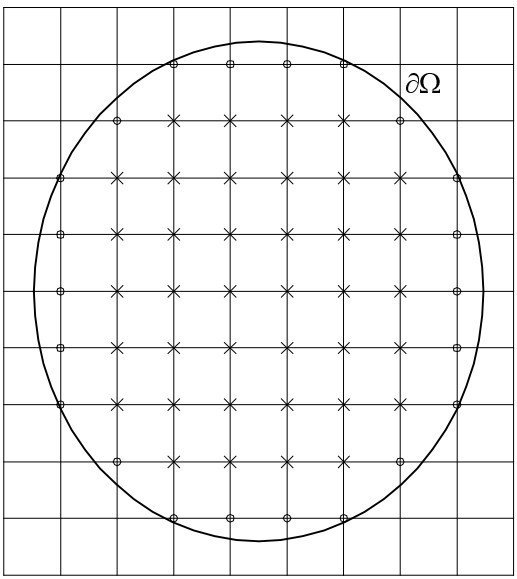


Рис. 6

Решение нашей задачи будем искать только в узлах. Составим уравнения, которые приближенно связывают значения решения в узлах. Для каждого

внутреннего узла (j,k) составим уравнение, используя для приближенной замены вторых производных в дифференциальном уравнении формулу (2):

$$-\frac{u_{jk+1}+u_{jk-1}+u_{j+1k}+u_{j-1k}-4u_{jk}}{h^2}+r_{jk}u_{jk}=f_{jk}. \qquad (8)$$

Здесь $r_{jk}=r(x_j,y_k)$, $f_{jk}=f(x_j,y_k)$, u_{jk} — искомые величины, которые мы принимаем за приближения к значениям решения в узлах (j,k). В предположении, что решение нашей задачи $u^*(x,y)$ четырежды непрерывно дифференцируемо, при подстановке в такое уравнение на место $u_{\mu\nu}$ значений $u^*(x_\mu,y_\nu)$ левая и правая части будут отличаться на величину порядка малости не ниже h^2.

Число уравнений вида (8) равно числу внутренних узлов и потому меньше числа неизвестных, которое есть число всех узлов, лежащих в $\overline{\Omega}$. Недостающие уравнения будем строить, исходя из заданного граничного условия, поставив в соответствие каждому узлу граничной полоски некоторое уравнение. Укажем два способа построения этих уравнений.

1. Для каждого узла граничной полоски $(j,k)\in I_2$ найдем ближайшую к нему точку $M_{jk}\in\partial\Omega$. Так как один из узлов, "соседних" с (j,k), лежит вне области Ω, то расстояние от узла (j,k) до M_{jk} заведомо меньше h. Построим уравнение, соответствующее узлу (j,k), так:

   $$u_{jk}=g(M_{jk}).$$

   Очевидно, что при подстановке в это уравнение значения точного решения невязка будет иметь порядок малости h:

   $$u^*(x_j,y_k)-g(M_{jk})=O(h).$$

   Это — наиболее простой способ, но он приводит к уравнениям, которым точное решение удовлетворяет с большей погрешностью, чем уравнениям (8). Поэтому чаще применяется второй, более точный способ.

2. Для каждого узла $(j,k)\in I_2$ по меньшей мере один из четырех узлов $(j\pm1,k),(j,k\pm1)$ лежит вне области Ω. Будем для определенности считать, что это — узел $(j+1,k)$ и что узел $(j-1,k)$ лежит в Ω. Отрезок, со-

единяющий узлы (j,k) и $(j+1,k)$, пересекает границу $\partial\Omega$ в некоторой точке M_{jk} с координатами $(x_j+\tau_{jk}h, y_k)$ $(0\le\tau_{jk}<1)$ (рис. 7).

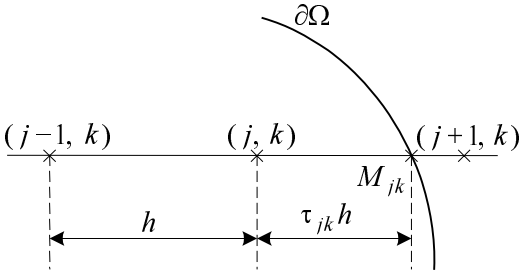


Рис. 7

Выразим значение решения в точке M_{jk} через его значения в узлах $(j-1,k)$ и (j,k) с помощью линейной интерполяции:

$$g(M_{jk})=u^*(x_j+\tau_{jk}h, y_k)=(1+\tau_{jk})u^*(x_j,y_k)-\tau_{jk}u^*(x_{j-1},y_k)+O(h^2).$$

Это и дает нам уравнение, соответствующее узлу граничной полоски (j,k):

$$(1+\tau_{jk})u_{jk}-\tau_{jk}u_{j-1k}=g(M_{jk}). \qquad (9)$$

В случае другого расположения узла (j,k) относительно границы уравнение, составленное для этого узла, выглядело бы точно так же, только вместо неизвестной u_{j-1k} в него вошла бы одна из неизвестных u_{j+1k}, u_{jk-1}, u_{jk+1}. В уравнении (9) известный нам коэффициент τ_{jk} удовлетворяет неравенству $0\le\tau_{jk}<1$ ($\tau_{jk}=0$ соответствует случаю, когда сам узел (j,k) лежит на границе) — это обстоятельство будет использовано при исследовании метода сеток в *§ 4*. Добавив к уравнениям (8) уравнения вида (9), мы получим систему уравнений, в которой число уравнений будет совпадать с числом неизвестных.

Сделаем еще одно общее замечание. Число неизвестных в методе сеток обычно совпадает с числом узлов — неизвестными являются значения искомой функции в этих узлах. Каждое построенное уравнение обычно соотносится некоторому узлу (так мы только что и делали), так что совпадение чис-

ла неизвестных и уравнений обычно означает, что каждому узлу соответствует в точности одно уравнение. В дальнейшем нам часто придется говорить о свойствах матрицы коэффициентов системы линейных уравнений, к которой приводит метод сеток. При этом мы всегда будем считать, что узлы пронумерованы у нас в некотором порядке, а нумерация неизвестных и уравнений строго соответствует нумерации узлов.

Совсем коротко скажем теперь о применении метода сеток к второй и третьей граничным задачам для того же уравнения:

$$-\Delta u + ru = f,$$

$$\text{II.}\quad \partial u/\partial n\big|_{\partial\Omega} = g,$$

$$\text{III.}\quad (\partial u/\partial n + \sigma u)\big|_{\partial\Omega} = g.$$

Здесь $\partial u/\partial n$ — производная по внешней нормали. Каждому внутреннему узлу $(j, k) \in I_1$ поставим в соответствие, как и в случае задачи Дирихле, уравнение (8). Каждому узлу $(j, k) \in I_2$ граничной полоски сопоставим уравнение, "аппроксимирующее граничное условие". Это уравнение может быть построено, например, по способу численного дифференцирования. Именно: а) найдем ближайшую к узлу точку $M_{jk} \in \partial\Omega$; б) выберем несколько ближайших к (j, k) узлов (ν_l, μ_l) $(l = 1, 2, ..., N)$, лежащих в $\overline{\Omega}$; в) построим "полиномы влияния" $\omega_l(x, y)$, такие, что $\omega_l(x_{\nu_s}, y_{\nu_s}) = \delta_{ls}$ $(l, s = 1, 2, ..., N)$. Тогда нужное нам уравнение (для граничной задачи III) примет вид

$$\sum_{l=1}^{N} \left(\frac{\partial \omega_l}{\partial n} + \sigma \omega_l \right)_{M_{jk}} u_{\nu_l \mu_l} = g(M_{jk}).$$

Сделаем по поводу указанного способа аппроксимации граничного условия два замечания. Во-первых, в двумерном (и вообще многомерном) случае интерполяционная задача далеко не всегда разрешима; это следует иметь в виду при выборе узлов (ν_l, μ_l). Во-вторых, если мы хотим добиться хорошей аппроксимации, то число узлов N приходится брать довольно большим; так, если мы строим интерполяционные полиномы второй степени относительно каждой переменной, то $N = 9$. Кроме того, уравнения (10) сильно отличаются от уравнений (8) своей "неправильностью". Все это в действительности в значительной мере затрудняет решение системы уравнений (8), (10).

Вариационно-разностные схемы

Вариационно-разностные схемы за последнее время получили довольно широкое распространение. В сущности, под вариационно-разностными схемами понимаются два разных способа составления конечно-разностных уравнений. Проиллюстрируем оба эти способа на примере обыкновенного дифференциального уравнения.

Способ 1

Как было показано в *§ 7 главы 3*, граничная задача

$$-(pu')'+ru = f \quad (p \geq p_0 \geq 0, \quad r \geq 0),$$
$$u'(0) + \alpha u(0) = u'(1) + \beta u(1) = 0 \quad (\beta > 0, \alpha < 0)$$

эквивалентна задаче отыскания минимума функционала энергии

$$F(u) = \int_0^1 (pu'^2 + ru^2 - 2uf)dx + \beta p(1)u^2(1) - \alpha p(0)u^2(0).$$

Выберем некоторое натуральное n, положим $h = 1/n$, $x_k = kh$ и построим новый функционал $F_n(u)$, заменив в функционале $F(u)$ интеграл от pu'^2 квадратурной суммой формулы средних прямоугольников, а от $ru^2 - 2uf$ — формулы трапеций и заменив значение производной функции u в середине отрезка разностным отношением

$$u'\left(\frac{x_k + x_{k+1}}{2}\right) \approx \frac{u(x_{k+1}) - u(x_k)}{h}.$$

Функционал $F_n(u)$ имеет вид

$$F_n(u) =$$
$$= h\left[\sum_{k=0}^{n-1} p_{k+1/2}\left(\frac{u_{k+1} - u_k}{h}\right)^2 + \frac{1}{2}r_0u_0^2 + \frac{1}{2}r_nu_n^2 - f_0u_0 - f_nu_n + \sum_{k=1}^{n-1}(r_ku_k^2 - 2f_ku_k)\right] +$$
$$+ \beta p_n u_n^2 - \alpha p_0 u_0^2 = \Phi_n(u_0, u_1, ..., u_n).$$

Здесь $u_k = u(x_k)$, $p_{k+1/2} = p((k+1/2)h)$, $r_k = r(x_k)$, $f_k = f(x_k)$.

Функционал F_n на гладких функциях принимает значения, близкие значениям функционала F. Это дает нам основание определить приближенное решение нашей задачи u_n^* как функцию, доставляющую минимум функционалу

F_n. Функционал $F_n(u)$ в действительности зависит лишь от значений функции u в узлах x_k. Поэтому и условие его минимальности определит функцию, доставляющую минимум, лишь в этих узлах; но мы можем довольствоваться этим — приближенное решение будет задано таблицей своих значений в точках x_k. Очевидно, что главная часть функции Φ_n — однородный полином второй степени относительно переменных u_k — есть положительно определенная квадратичная форма. Поэтому функция Φ_n достигает своего минимального значения в единственной точке, которая может быть найдена, если мы приравняем нулю ее частные производные. Это приведет нас к системе линейных алгебраических уравнений относительно u_k — значений приближенного решения в узлах x_k :

$$\left.\begin{aligned}
\frac{1}{2}\frac{\partial \Phi_n}{\partial u_0} &= -p_{1/2}\frac{u_1-u_0}{h}+\frac{1}{2}hr_0u_0-\alpha p_0u_0-\frac{1}{2}hf_0=0,\\
\frac{1}{2}\frac{\partial \Phi_n}{\partial u_j} &= \frac{p_{j-1/2}(u_j-u_{j-1})-p_{j+1/2}(u_{j+1}-u_j)}{h}+hr_ju_j-hf_j=0\\
&\qquad j=1,2,\ldots,n-1\\
\frac{1}{2}\frac{\partial \Phi_n}{\partial u_n} &= p_{n-1/2}\frac{u_n-u_{n-1}}{h}+\frac{1}{2}hr_nu_n+\beta p_nu_n-\frac{1}{2}hf_n=0.
\end{aligned}\right\}$$

Все эти уравнения, кроме первого и последнего, лишь множителем h отличаются от тех, которые получились бы, если бы мы в дифференциальном уравнении, выписанном для точки x_j, заменили выражение $-(p')'$ по той формуле, которая приводилась выше. Что касается первого и последнего уравнений, то они учитывают одновременно как дифференциальное уравнение, так и граничные условия — в них входят значения коэффициентов p, r и f дифференциального уравнения, а также α и β из граничных условий.

Отметим два преимущества такого способа составления конечно-разностных уравнений. Во-первых, нам не пришлось думать о том, как составить уравнения, соответствующие узлам x_0 и x_1, исходя из граничных условий. Это преимущество оказывается особенно существенным при применении подобного метода к второй или третьей граничным задачам для эллиптических дифференциальных уравнений. Уравнения, учитывающие граничные условия, получились у нас довольно простыми — они содержат лишь по два неизвестных. Во-вторых, матрица полученной системы линейных алгебраиче-

ских уравнений симметрична и положительно определена. Это имеет немаловажное значение для численного решения этой системы. Заметим, что если бы мы составляли первое и последнее уравнения исходя только из граничных условий, то добиться симметрии нам бы не удалось.

Способ 2

Другой способ, относящийся к вариационно-разностным, носит еще название *метода конечных элементов*.

Пусть требуется решить дифференциальное уравнение

$$-(pu')'+qu'+ru = f$$

при одном из граничных условий:

I. $u(0) = u(1) = 0.$

II. $u'(0) + \alpha u(0) = u'(1) + \beta u(1) = 0, \quad \alpha \le 0, \beta \ge 0.$

(частным случаем граничного условия II при $\alpha = \beta = 0$ является $u'(0) = u'(1) = 0$). Будем решать нашу граничную задачу методом Галеркина в той форме, как это указано в *§ 9 главы 3*, положив $Bu = -(pu')'$ и выбрав координатные функции специальным образом. Для определенности в случае задачи II мы считаем либо $\alpha < 0$, либо $\beta > 0$ (это обеспечивает положительную определенность оператора B) — если $\alpha = \beta = 0$, следовало бы положить, например, $Bu = -(pu')'+u$, что, впрочем, привело бы к тем же результатам, которые получатся, если в окончательных формулах мы формально положим $\alpha = \beta = 0$. Перейдем к построению координатных функций. Построим прежде всего функцию (рис. 8).

$$\omega(x) = \begin{cases} 0 & \text{при} \quad x \le -1, \\ 1+x & \text{при} \quad -1 \le x \le 0, \\ 1-x & \text{при} \quad 0 \le x \le 1, \\ 0 & \text{при} \quad x \ge 1. \end{cases}$$

Далее, как и раньше, выбрав натуральное n, положим $h = 1/n$, $x_k = kh$ $(k = 0, 1, ..., n)$ и, наконец,

$$\omega_k^{(n)}(x) = \omega\left(\frac{x - x_k}{h}\right).$$

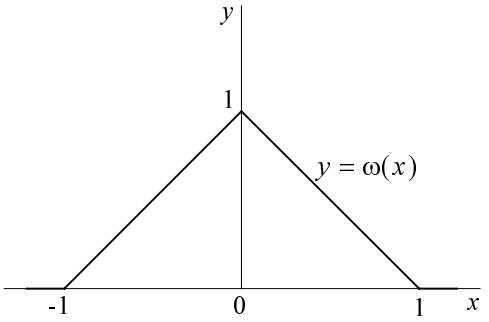


Рис. 8

Здесь для первой граничной задачи $k = 1, 2, ..., n-1,$ а для второй — $k = 0, 1, ..., n.$ Разница между задачами I и II состоит в следующем. В методе Галеркина координатные функции должны принадлежать энергетическому пространству $H_B,$ которое при условии I совпадает с $\overset{0}{W}{}_2^1,$ а при II — с $W_2^1.$ Будучи кусочно-дифференцируемыми, функции $\omega_k^{(n)}$ принадлежат $W_2^1,$ но в случае условия I возникает еще требование $\omega_k^{(n)}(0) = \omega_k^{(n)}(1) = 0,$ которому эти функции удовлетворяют лишь при $k = 1, 2, ..., n-1.$ По поводу сказанного *см. § 7 главы 3*.

Отметим, что мы отошли от принятых раньше обозначений — при каждом n число координатных функций $\omega_k^{(n)}$ есть $n-1$ для задачи I $(k = 1, 2, ..., n-1)$ и $n+1$ для задачи II $(k = 0, 1, ..., n).$ Чтобы объединить изложение обеих граничных задач, положим

$$\sigma = \begin{cases} 1 & \text{в случае задачи I;} \\ 0 & \text{в случае задачи II.} \end{cases}$$

Тогда приближенное решение $u_n(x)$ ищется в виде $u_n(x) = \sum_{k=\sigma}^{n-\sigma} u_k \omega_k^{(n)}(x).$ Искомые коэффициенты мы обозначили через u_k в связи с тем, что они имеют следующий простой смысл: т. к. $\omega_k^{(n)}(x_j) = \delta_{kj},$ то $u_k = u_n(x_k)$ есть значение приближенного решения в узле $x_k.$

Таким образом, систему линейных алгебраических уравнений

$$\sum_{k=\sigma}^{n-\sigma} a_{jk} u_k = b_j, \quad j = \sigma, \sigma+1, ..., n-\sigma, \tag{11}$$

метода Галеркина мы рассматриваем как систему уравнений относительно приближенных значений решения в узлах. Это и позволяет нам относить рассматриваемый метод к методу сеток, хотя формально это — проекционный метод. С методом сеток наш метод сближает еще и то обстоятельство, что каждое из уравнений (11) содержит не более трех неизвестных. Действительно, если $|j-k|>1$, то $\omega_j^{(n)}$ и $\omega_k^{(n)}$ отличны от нуля на непересекающихся промежутках, и потому при таких j и k коэффициенты

$$a_{jk} = [\omega_k^{(n)}, \omega_j^{(n)}]_B + (q\omega_k^{(n)'} + r\omega_k^{(n)}, \omega_j^{(n)})$$

обращаются в нуль. Легко также убедиться, что для $j, k = 1, 2, ..., n-1$

$$[\omega_k^{(n)}, \omega_j^{(n)}]_B = \int_0^1 p\omega_k^{(n)'}, \omega_j^{(n)'} dx = \frac{1}{h} \times \begin{cases} 0 & \text{при} \quad |k-j|>1, \\ p_{k-1/2} + p_{k+1/2} & \text{при} \quad k=j, \\ -p_{k+1/2} & \text{при} \quad j=k+1, \\ -p_{k-1/2} & \text{при} \quad j=k-1, \end{cases}$$

где

$$p_{k+1/2} = \frac{1}{h}\int_{x_k}^{x_{k+1}} p(t)dt$$

— среднее значение коэффициента p на промежутке $[x_k, x_{k+1}]$. Для граничной задачи II, кроме того,

$$[\omega_0^{(n)}, \omega_0^{(n)}]_B = \int_0^1 p(\omega_0^{(n)'})^2 dx - \alpha p(0)[\omega_0^{(n)}(0)]^2 = \frac{1}{h} p_{1/2} - \alpha p(0),$$

$$[\omega_n^{(n)}, \omega_n^{(n)}]_B = \frac{1}{h} p_{n-1/2} + \beta p(1).$$

Мы не будем указывать значения интегралов $(q\omega_k^{(n)'}, \omega_j^{(n)})$, $(r\omega_k^{(n)}, \omega_j^{(n)})$, а также $(f, \omega_j^{(n)}) = b_j$. Отметим только, что если коэффициенты p, q, r и правая часть f на промежутке $[x_{j-1}, x_{j+1}]$ суть постоянные, равные соответственно

p_j, q_j, r_j и f_j, то уравнение с номером j, как в этом легко убедиться, принимает вид

$$-p_j\frac{u_{j+1}-2u_j+u_{j-1}}{h}+q_j\frac{u_{j+1}-u_{j-1}}{2}+hr_j\left(\frac{1}{6}u_{j+1}+\frac{2}{3}u_j+\frac{1}{6}u_{j-1}\right)=hf_j.$$

Это уравнение лишь множителем h и последней скобкой в левой части отличается от уравнения, указанного в *разд. "Метод сеток для обыкновенного дифференциального уравнения"*.

Преимуществом такого способа составления конечно-разностных уравнений является "естественная" аппроксимация граничных условий (речь идет о задаче II). Это преимущество в действительности оказывается очень существенным при решении многомерных задач. Если дифференциальный оператор симметричен (в нашей задаче $q=0$), то и матрица системы уравнений будет симметричной; если оператор положительно определен, то метод Галеркина совпадает с методом Ритца, и матрица окажется также положительно определенной. Важность этого обстоятельства для численного решения системы уравнений уже отмечалась. О преимуществах этого метода в случае малой гладкости искомого решения будет сказано далее в *§ 3*. Наконец, для исследования метода конечных элементов применимы теоремы о проекционных методах — это и будет проделано в одном из следующих параграфов.

Особенности применения метода сеток к нестационарным уравнениям

Применение метода сеток к уравнениям, описывающим процессы, протекающие во времени, имеет существенные особенности, которые мы поясним на примере уравнения теплопроводности. Пусть дана граничная задача для одномерного уравнения теплопроводности

$$\frac{\partial u}{\partial t}-a\frac{\partial^2 u}{\partial x^2}=f, \tag{12}$$

$$u(x,0)=g(x),\quad u(0,t)=\alpha(t),\quad u(1,t)=\beta(t). \tag{13}$$

Здесь a, f, g, α, β — заданные непрерывные функции, причем $a(x,t)>0$, $\alpha(0)=g(0)$, $\beta(0)=g(1)$ (при исследовании метода сеток в следующих пара-

графах мы будем предполагать достаточную гладкость решения, что налагает на все эти функции в действительности несколько большие условия, в т. ч. и лучшее "согласование" функций α и β с f и g). Пусть решение этой задачи требуется найти в прямоугольнике $0 \le t \le T,\ 0 \le x \le 1$.

Как обычно в методе сеток, выберем некоторое натуральное n и положим $h = 1/n$ — шаг по переменной x, выберем также $\tau > 0$ — шаг по переменной t. Точки пересечения прямых

$$x = x_j = jh \quad (j = 0, 1, ..., n)$$

и

$$t = t_k = k\tau \quad (k = 0, 1, ..., m,\ m = [T/\tau] \text{ — целая часть числа } T/\tau)$$

будем называть узлами, и приближенное решение будем искать только в узлах. Для граничных узлов (так мы называем узлы, соответствующие $x = 0$, $x = 1$ или $t = 0$) уравнения метода сеток очевидным образом получаются из начального и граничных условий:

$$u_{j0} = g_j = g(x_j), \quad j = 0, 1, ..., n, \tag{14}$$

$$u_{0k} = \alpha_k = \alpha(t_k), \quad u_{nk} = \beta_k = \beta(t_k), \quad k = 1, 2, ..., m. \tag{15}$$

Для остальных узлов (их мы будем называть внутренними) составим уравнения так. Для узла (j, k) $(1 \le j \le n-1, 1 \le k \le m)$ запишем дифференциальное уравнение (12) в точке (x_j, t_{k-1}) и произведем приближенную замену производных по формулам

$$\left(\frac{\partial u}{\partial t}\right)_{jk-1} \approx \frac{u_{jk} - u_{jk-1}}{\tau}, \quad \left(\frac{\partial^2 u}{\partial x^2}\right)_{jk-1} \approx \frac{u_{j+1k-1} - 2u_{jk-1} + u_{j-1k-1}}{h^2}.$$

Тогда получим уравнение

$$\frac{u_{jk} - u_{jk-1}}{\tau} - a_{jk-1}\frac{u_{j+1k-1} - 2u_{jk-1} + u_{j-1k-1}}{h^2} = f_{jk-1}. \tag{16}$$

Система (14)—(16) представляет собой полную систему уравнений относительно неизвестных u_{jk}, которые мы рассматриваем как приближения к значениям точного решения в узлах: $u_{jk} \approx u^*(x_j, t_k)$.

Значения точного решения u^* в узлах удовлетворяют уравнениям этой системы с точностью до величин порядка малости $O(h^2+\tau)$, если только это решение u^* дважды непрерывно дифференцируемо по t и четырежды по x в рассматриваемом прямоугольнике.

Выписанная система уравнений решается чрезвычайно просто. Действительно, при $k=0$ u_{j0} сразу же получаются по формулам (14). При $k=1$ перепишем уравнения (16) для $j=1,2,...,n-1$ в виде

$$u_{j1}=\left(1-2a_{j0}\frac{\tau}{h^2}\right)u_{j0}+a_{j0}\frac{\tau}{h^2}(u_{j+10}+u_{j-10})+\tau f_{j0}.$$

Все числа, входящие в правые части этих равенств, нам известны, так что по этим формулам мы сразу можем найти все u_{j1} $(j=1,2,...,n-1)$; кроме того, $u_{01}=\alpha_1$, $u_{n1}=\beta_1$. Вообще после того, как для некоторого k мы нашли все значения $u_{0k-1}, u_{1k-1,}, ..., u_{nk-1}$ ("решение на $(k-1)$-м слое"), следующие значения находятся по формулам

$$u_{0k}=\alpha_k,\quad u_{nk}=\beta_k,$$

$$u_{jk}=\left(1-2a_{jk-1}\frac{\tau}{h^2}\right)u_{jk-1}+a_{jk-1}\frac{\tau}{h^2}(u_{j+1k-1}+u_{j-1k-1})+\tau f_{jk-1}$$

$$(j=1,2,...,n-1)$$

(это — "решение на k-м слое"). Таким образом, наша система уравнений оказалась настолько простой, что допускает последовательное вычисление всех неизвестных, одно за другим. Это означает, что при соответствующей нумерации узлов (а тем самым — неизвестных и уравнений) матрица коэффициентов этой системы оказывается нижней треугольной.

Такие разностные схемы, которые позволяют последовательно, одно за другим, находить все искомые значения, называются явными.

Введем еще одно понятие. Конфигурация расположения точек плоскости (x,t), в которых значения приближенного решения связываются одним уравнением, называется шаблоном разностной схемы. Разностная схема, которую мы сейчас построили, имеет шаблон, изображенный на рис. 9. Характерным признаком явной разностной схемы является то, что ее шаблон на "верхнем" временном слое имеет только одну точку. Тогда решение в "по-

следующий момент времени" t_k выражается через значения решения в "предшествующий момент времени" t_{k-1} (быть может, еще и $t_{k-2}, t_{k-3}, \ldots$).

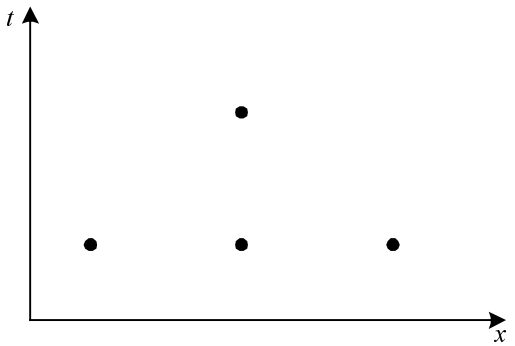


Рис. 9

Отметим еще одну особенность рассматриваемой разностной схемы. Нетрудно убедиться, что приближенное решение u_{jk}, построенное в точке (x_j, t_k), зависит от значений правой части f и функций g, α, β лишь в точках, расположенных ниже прямых $\dfrac{t-t_k}{\tau} = \pm\dfrac{x-x_j}{h}$ (рис. 10).

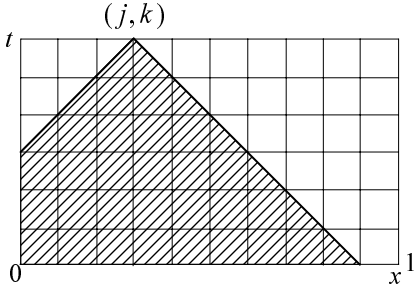


Рис. 10

Между тем точное решение u^* в точке (x_j, t_k) зависит от всех значений g и от всех значений f, α, β при $t < t_k$. Отсюда следует такой вывод. Пусть мы

находим последовательность приближенных решений, устремляя шаги h и τ к нулю. Сходимость приближенных решений к точному можно ожидать лишь в том случае, если отношение τ / h стремится к нулю, т. е. область зависимости приближенных решений от правой части и от начального и граничных условий стремится к области зависимости точного решения. Как будет выяснено в дальнейшем, устойчивость метода налагает на соотношение между шагами τ и h значительно более жесткие ограничения и практически метод применим лишь при очень малом в сравнении с h шагом τ. Однако использование малых шагов τ требует большой затраты труда для получения решения на заданном промежутке изменения $t : t \in [0, T]$. Поэтому представляют интерес другие разностные схемы, которые свободны от указанного недостатка.

Оставив прежней совершенно естественную аппроксимацию начального и граничных условий (14), (15), поставим теперь каждому внутреннему узлу (j, k) в соответствие уравнение

$$\frac{u_{jk} - u_{jk-1}}{\tau} - a_{jk} \frac{u_{j+1k} - 2u_{jk} + u_{j-1k}}{h^2} = f_{jk}, \qquad (17)$$

$$j = 1, 2, ..., n-1; \quad k = 1, 2, ..., m.$$

Мы записали наше дифференциальное уравнение на этот раз в точке (x_j, t_k) и заменили производные указанными разностными отношениями. Шаблон этой разностной схемы изображен на рис. 11.

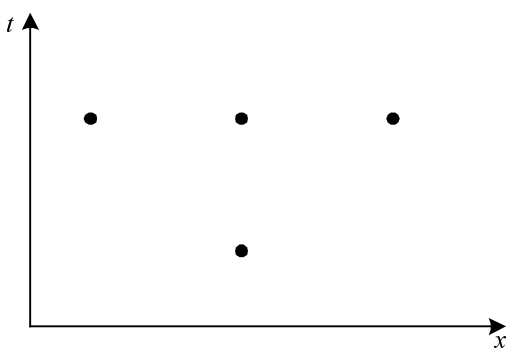


Рис. 11

Как и в первой схеме, все неизвестные u_{j0} сразу же находятся из уравнений (14). Выпишем теперь при $k=1$ уравнения (15) и (17), отнеся в последних в правую часть уже известные нам значения u_{j0}:

$$u_{01}=\alpha_1, \quad u_{n1}=\beta_1,$$

$$\left(1+2a_{j1}\frac{\tau}{h^2}\right)u_{j1}-a_{j1}\frac{\tau}{h^2}(u_{j+11}+u_{j-11})=u_{j0}+\tau f_{j1}, \quad j=1,2,...,n-1.$$

Это — система $n+1$ уравнений относительно $n+1$ неизвестных $u_{01}, u_{11}, ..., u_{n1}$. Решив эту систему, мы можем затем перейти к нахождению $u_{02}, u_{12}, ..., u_{n2}$, т. е. решения на "втором временном слое", и т. д. Вообще после того, как мы нашли компоненты решения $u_{0k-1}, u_{1k1}, ..., u_{nk-1}$, компоненты $u_{0k}, u_{1k}, ..., u_{nk}$ находятся из системы уравнений

$$\left.\begin{aligned} u_{0k}&=\alpha_k, \quad u_{nk}=\beta_k, \\ \left(1+2a_{jk}\frac{\tau}{h^2}\right)u_{jk}-a_{jk}\frac{\tau}{h^2}(u_{j+1k}+u_{j-1k})&=u_{jk-1}+\tau f_{jk}, \quad j=1,2,...,n-1. \end{aligned}\right\} \qquad (18)$$

Итак, система уравнений (14), (15), (17), содержащая $(n+1)(m+1)$ уравнений и столько же неизвестных, разбилась у нас на системы (14) и m систем (18), которые решаются последовательно, причем каждая из них содержит $n+1$ уравнений и неизвестных. Таким образом, наша вторая схема, как и первая, позволяет находить решение последовательно, слой за слоем. Но в отличие от первой здесь решение на каждом слое получается в результате решения системы линейных алгебраических уравнений, а не по явным формулам. Такие схемы, в которых построение решения на каждом "временном слое" требует решения системы линейных алгебраических уравнений, называются неявными.

Возможность находить значения решения последовательно, слой за слоем, вообще является особенностью метода сеток для нестационарных уравнений, и такая возможность связана, конечно, со свойством самих уравнений: точное решение в некоторой точке зависит от значений правой части и функций, входящих в граничные условия, относящихся лишь к предшествующим моментам времени.

В только что описанной разностной схеме область зависимости приближенного решения от правой части и граничных условий та же, что и у точного —

каждая компонента решения системы уравнений (18) зависит от всех ее правых частей. Но эта схема имеет один существенный недостаток — для приближенной замены производной по времени при составлении уравнений (17) (как и для уравнений (16)) мы пользовались очень грубой формулой, и точное решение u^* будет удовлетворять этим уравнениям с погрешностью порядка всего лишь $O(\tau + h^2)$.

Сейчас мы укажем более точный способ составления уравнений для внутренних узлов. Запишем дифференциальное уравнение в точке $(x_j, t_{k-1/2})$, где $t_{k-1/2} = (k-1/2)\tau$:

$$\left(\frac{\partial u}{\partial t}\right)_{jk-1/2} - a_{jk-1/2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{jk-1/2} = f_{jk-1/2}.$$

Произведем замену производных следующим способом:

$$\left(\frac{\partial u}{\partial t}\right)_{jk-1/2} = \frac{u_{jk} - u_{jk-1}}{\tau} + O(\tau^2),$$

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{jk-1/2} = \frac{1}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{jk} + \frac{1}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{jk-1} + O(\tau^2) =$$

$$= \frac{u_{j+1k} - 2u_{jk} + u_{j-1k}}{2h^2} + \frac{u_{j+1k-1} - 2u_{jk-1} + u_{j-1k-1}}{2h^2} + O(h^2) + O(\tau^2),$$

причем при выписывании порядков остаточных членов мы предполагали, что функция u четырежды непрерывно дифференцируема. Это приводит нас к таким уравнениям:

$$\frac{u_{jk} - u_{jk-1}}{\tau} - a_{jk-1/2} \times$$
$$\times\left(\frac{u_{j-1k} - 2u_{jk} + u_{j-1k}}{2h^2} + \frac{u_{j+1k-1} - 2u_{jk-1} + u_{j-1k-1}}{2h^2}\right) = f_{jk-1/2}. \qquad (19)$$

Значения точного решения задачи в узлах будут удовлетворять этим уравнениям с погрешностью порядка $O(h^2 + \tau^2)$, если только это решение обладает нужной гладкостью.

Как и в предыдущем случае, система уравнений (14), (15), (19) может быть переписана в виде системы (14) и систем

$$u_{0k} = \alpha_k, \quad u_{nk} = \beta_k,$$

$$\left(1 + a_{jk-1/2}\frac{\tau}{h^2}\right)u_{jk} - \frac{1}{2}a_{jk-1/2}\frac{\tau}{h^2}(u_{j+1k} + u_{j-1k}) = \left(1 - a_{jk-1/2}\frac{\tau}{h^2}\right)u_{jk-1} +$$

$$+ \frac{1}{2}a_{jk-1/2}\frac{\tau}{h^2}(u_{j+1k-1} + u_{j-1k-1}) + \tau f_{jk-1/2}, \quad j = 1, 2, ..., n-1.$$

Здесь $k = 1, 2, ..., m$. Эти системы уравнений можно решать последовательно. Наша последняя схема — неявная и имеет шаблон, изображенный на рис. 12 (крестиком помечена точка, для которой выписывалось дифференциальное уравнение).

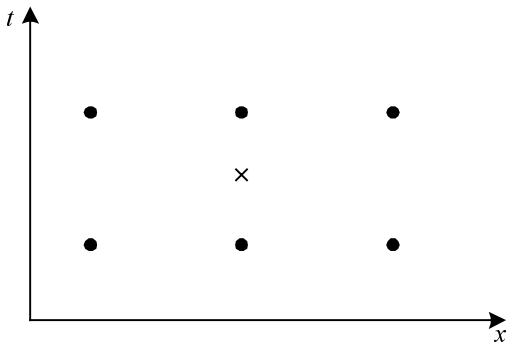


Рис. 12

Особенности применения общей схемы приближенных методов к исследованию метода сеток

На том же примере граничной задачи для обыкновенного дифференциального уравнения

$$L(u) = pu'' + qu' + ru = f, \quad u(0) = a, \quad u(1) = b,$$

который рассматривался в *разд. "Метод сеток для обыкновенного дифференциального уравнения"*, поясним особенности применения общей схемы, изложенной в *главе 1*, к методу сеток.

Нашу граничную задачу можно записать в виде линейного уравнения $Au = g$. Оператор A задан на множестве $D(A)$ дважды непрерывно дифференцируемых на $[0, 1]$ функций, которое мы можем рассматривать как линейное подмножество некоторого линейного нормированного пространства U заданных на $[0, 1]$ функций. Выбор этого пространства мы уточнять не будем — как будет видно из дальнейшего, он чаще всего безразличен. Оператор A каждой функции $u \in D(A)$ ставит в соответствие тройку: $Au = (L(u), u(0), u(1))$. Таким образом, $F = F_0 \times R^2$, где F_0 — некоторое пространство функций, заданных на $[0, 1]$. Правая часть уравнения $Au = g$ есть тройка $g = (f(x), a, b)$.

Прежде чем переходить к выяснению элементов общей схемы, относящихся к приближенному уравнению и связи между приближенным и точным, сделаем одно общее замечание. Систему уравнений метода сеток принято связывать не с параметром n ("номером приближенного уравнения"), а с параметром h, который имеет смысл шага сетки. В случае рассматриваемого дифференциального уравнения шаг h однозначно связан с n: $h = 1/n$. Но в других случаях (например, в методе сеток для эллиптических уравнений) h может меняться непрерывно. В некоторых случаях система уравнений метода сеток кроме шага h может определяться еще и другими независимыми параметрами (например, в методе сеток для уравнения теплопроводности — шагом τ по переменной t); в таких случаях мы всегда будем считать, что все остальные параметры заданы нам как функции h, так чтобы система уравнений полностью определялась этим одним параметром. Поэтому все пространства и операторы, связанные с приближенным уравнением, мы будем теперь помечать значком h, а не n, как это было раньше.

Поведение приближенных уравнений нас будет интересовать при $h \to 0 \quad (h > 0)$, причем иногда h будет принимать лишь счетное множество значений ($h = 1/n$ в нашем примере), а иногда можно считать h меняющимся непрерывно. Разумеется, все теоремы общей теории остаются при этом в силе. Теперь вернемся к рассматриваемой граничной задаче для обыкновенного дифференциального уравнения.

В качестве приближенного уравнения естественно рассматривать систему линейных алгебраических уравнений (3), так что U_h и F_h — $(n+1)$-мерные векторные пространства и A_h есть матрица коэффициентов системы уравнений (3). Оператор $\psi_h \in L(F, F_h)$ в общей схеме определяется тем, что правой части точного уравнения он ставит в соответствие правую часть приближенного, поэтому применительно к нашей задаче оператор ψ_h каждой тройке (f, a, b) должен ставить в соответствие вектор $\psi_h g = (a, f(x_1), ..., f(x_{n-1}), b)$.

Особо стоит остановиться на выборе оператора $\varphi_h \in L(U, U_h)$, который не определяется алгоритмом метода. Каркас приближенного решения — вектор $u_h^* = (u_0, u_1, ..., u_n)$, являющийся решением системы уравнений (3), трактуется как таблица приближенного решения: $u_k \approx u^*(x_k)$. Это делает целесообразным такой выбор оператора φ_h, при котором точному решению u^* он ставит в соответствие таблицу этого точного решения в тех же самых узлах. Тем самым для $\upsilon \in U$ должно быть $\varphi_h \upsilon = (\upsilon(x_0), \upsilon(x_1), ..., \upsilon(x_n))$. При таком выборе φ_h теорема о сходимости каркасов приближенных решений *из § 2 главы 1* дает нам оценку нормы вектора $\varphi_h u^* - u_h^*$, компоненты которого — разности между табличными значениями точного и приближенного решений. Такой информации — оценки погрешности таблицы приближенного решения — обычно бывает достаточно и для того, чтобы судить, не прибегая к общей схеме, и о погрешности приближенного решения — функции, построенной с помощью какой-либо интерполяции таблицы. Поэтому исследование метода сеток проводится обычно лишь на уровне сходимости каркасов приближенных решений и устойчивости процесса отыскания каркасов приближенных решений. Этим уровнем мы будем ограничиваться при исследовании различных вариантов метода сеток в следующих параграфах. При этом отпадет какая-либо необходимость говорить что-либо о выборе пространства U и F, в которых рассматривается точное уравнение.

Выбранные указанным способом операторы ψ_h и φ_h называются обычно в теории метода сеток *операторами сноса*. В тех случаях, когда по полученному "сеточному" решению u_h^* все же строится приближенное решение $u^{(h)}$ как функция, заданная во всей рассматриваемой области, оператор $\overline{\varphi}_h$, осуществляющий такое построение, называется *оператором восполнения*.

В связи со сделанным нами выбором оператора φ_h отметим еще, что мера аппроксимации точного уравнения приближенным, вычисленная на точном решении, имеет очень простой смысл — это норма вектора невязок, которые получаются при подстановке в систему конечно-разностных уравнений значений точного решения в узлах сетки. Так же обстоит дело не только с нашим обыкновенным дифференциальным уравнением, но и со всеми другими вариантами метода сеток для различных задач, которые описывались ранее. Когда при построении конечно-разностных уравнений выше мы отмечали, что точное решение задачи удовлетворяет этим уравнениям с точностью до величин некоторого порядка малости ε ($\varepsilon(h)$ или $\varepsilon(h,\tau)$), то это означало тем самым, что такой порядок малости имеют все компоненты вектора, норма которого и есть мера аппроксимации на точном решении. Это позволит нам в дальнейшем при исследовании метода сеток для конкретных задач каждый раз, когда речь пойдет об оценке меры аппроксимации, ссылаться на соответствующее замечание, сделанное при построении конечно-разностных уравнений.

Напомним, что оценка, даваемая теоремой о сходимости каркасов, $\left\| \varphi_h u^* - u_h^* \right\| \le \left\| A_h^{-1} \right\| \cdot \gamma_h(u^*)$, имеет совсем простой смысл. Вектор u_h^* есть точное решение системы конечно-разностных уравнений; таблица точного решения $\varphi_h u^*$, рассматриваемая как приближенное решение той же системы, дает невязку, норма которой есть $\gamma_h(u^*)$; сама указанная оценка — это просто оценка погрешности приближенного решения через норму обратного оператора и норму невязки. С точки зрения применимости этой оценки нам безразлично, что $\varphi_h u^*$ есть таблица точного решения — это обстоятельство используется лишь для оценки порядка убывания нормы невязки. Это дает нам еще одно основание не говорить подробно о формальной трактовке рассматриваемых далее задач в рамках общей схемы. В частности, ранее мы показали, каким образом граничную задачу для обыкновенного дифференциального уравнения можно рассматривать как уравнение с оператором, действующим из некоторого линейного нормированного пространства в другое; так же можно было бы поступить и с другими граничными задачами, но делать этого мы не будем еще и по указанной только что причине, а не только из-за естественной аналогии с граничной задачей для обыкновенного дифференциального уравнения.

Итак, применение общей схемы приближенных методов к исследованию метода сеток имеет следующие особенности.

1. Исследование проводится лишь на уровне сходимости каркасов приближенных решений и устойчивости процесса отыскания каркасов. Это позволяет не производить выбор той пары пространств (U, F), в которых рассматривается исходная задача.
2. Сходимость каркасов приближенных решений имеет смысл сходимости таблиц приближенного решения к таблицам точного на сгущающихся одновременно сетках узлов.
3. Мера аппроксимации приближенными уравнениями точного, вычисленная на точном решении, есть норма вектора, компоненты которого — невязки, получающиеся при подстановке таблицы точного решения в конечно-разностные уравнения.

В заключение параграфа подведем итоги сказанному о том, что такое метод сеток, отметив его характерные особенности.

- Метод сеток сводит задачу решения линейного дифференциального уравнения к решению системы линейных алгебраических уравнений.
- Неизвестными в этой системе являются приближения к значениям искомой функции в некоторых точках области — узлах (иногда некоторые неизвестные — приближения к значениям производных искомой функции).
- Поскольку хорошее представление о решении обычно может дать лишь множество его значений в достаточно густой сетке узлов, алгебраические системы уравнений, которые приходится решать в методе сеток, как правило, имеют высокий порядок.
- Обычно каждое уравнение системы содержит лишь небольшое число неизвестных, и это значительно облегчает их решение.

§ 2. Вспомогательные сведения из линейной алгебры

Матрицы систем линейных уравнений, которые возникают в методе сеток, часто обладают некоторыми особыми свойствами. В этом параграфе будут изучены два специальных класса матриц, которые часто появляются при применении метода сеток к дифференциальным уравнениям.

Рассмотрим матрицу порядка n

$$R = \begin{pmatrix} 0 & 1 & 0 & . & . & . & 0 \\ 1 & 0 & 1 & . & . & . & 0 \\ 0 & 1 & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & 1 \\ 0 & 0 & . & . & . & 1 & 0 \end{pmatrix}.$$

Лемма 1

Собственными числами матрицы R являются $\lambda_k = 2\cos\dfrac{k\pi}{n+1}$ $(k = 1, 2, ..., n)$ при собственных векторах

$$z_k = \sqrt{\frac{2}{n+1}}\left(\sin\frac{k\pi}{n+1}, \sin\frac{2k\pi}{n+1}, ..., \sin\frac{nk\pi}{n+1}\right) = (\xi_{k1}, \xi_{k2}, ..., \xi_{kn}).$$

Собственные векторы z_k ортогональны и нормированы: $(z_k, z_j) = 1$.

Доказательство

Покажем, что $Rz_k = \lambda_k z_k$. Действительно, при всех $j = 1, 2, \ldots, n$

$$(Rz_k)_j = \sqrt{\frac{2}{n+1}}\left(\sin\frac{(j-1)k\pi}{n+1} + \sin\frac{(j+1)k\pi}{n+1}\right) =$$

$$= 2\cos\frac{k\pi}{n+1}\sqrt{\frac{2}{n+1}}\sin\frac{jk\pi}{n+1} = \lambda_k \xi_{kj}.$$

(Мы воспользовались тем, что при $j = 1$ $\sin\dfrac{(j-1)k\pi}{n+1} = 0$, а при $j = n$ $\sin\dfrac{(j+1)k\pi}{n+1} = 0$.)

Все собственные числа λ_k различны, т. к. на промежутке $[0, \pi]$ функция $\cos t$ монотонно убывает. Поэтому векторы z_k попарно ортогональны как собственные векторы симметричной матрицы, соответствующие различ-

ным собственным числам. Осталось показать, что при всех k $(z_k, z_k) = 1$, т. е. что

$$s = \sum_{j=1}^{n} \left(\sin \frac{jk\pi}{n+1} \right)^2 = \frac{n+1}{2}.$$

Докажем это последнее равенство:

$$s = \frac{1}{2} \sum_{j=1}^{n} \left(1 - \cos \frac{2jk\pi}{n+1} \right) = \frac{n}{2} - \frac{1}{2} \left(\sum_{j=0}^{n} \cos \frac{2jk\pi}{n+1} - 1 \right).$$

Остается заметить, что

$$\sum_{j=0}^{n} \cos \frac{2jk\pi}{n+1} = \operatorname{Re} \sum_{j=0}^{n} e^{2jk\pi i/(n+1)} = \operatorname{Re} \frac{e^{2k\pi i} - 1}{e^{2k\pi i/(n+1)} - 1} = 0. \blacksquare$$

Теорема 1

Пусть B и C — матрицы порядка n:

$$B = \begin{pmatrix} a & b & 0 & \cdots & 0 & 0 \\ b & a & b & \cdots & 0 & 0 \\ 0 & b & a & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & b & a \end{pmatrix}, \quad C = \begin{pmatrix} c & d & 0 & \cdots & 0 & 0 \\ d & c & d & \cdots & 0 & 0 \\ 0 & d & c & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & d & c \end{pmatrix}.$$

Для того чтобы матрица B была неособенной, необходимо и достаточно, чтобы при $k = 1, 2, ..., n$ было $a + 2b \cos \dfrac{k\pi}{n+1} \neq 0$. Если это условие выполнено, то матрица $B^{-1}C$ симметрична и ее собственные числа суть

$$\lambda_k = \frac{c + 2d \cos[k\pi/(n+1)]}{a + 2b \cos[k\pi/(n+1)]}$$

при собственных векторах z_k, указанных в лемме.

Доказательство

Доказательство немедленно вытекает из леммы, поскольку $B = aE + bR$, $C = cE + dR$ (E — единичная матрица). Матрица $B^{-1}C$ симметрична, по-

скольку ее собственные числа вещественны, и она имеет полную систему попарно ортогональных собственных векторов. ■

Перейдем теперь к рассмотрению другого класса матриц.

Пусть

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

— матрица порядка n. Введем некоторые понятия, связанные с индексами — номерами строк и столбцов матрицы A.

- Индекс i будем называть *избыточным*, если $a_{ii} > \sum_{k \neq i} |a_{ik}|$.
- Индекс j будем называть *непосредственно подчиненным* индексу i, если $a_{ij} \neq 0$.
- Индекс j будем называть *подчиненным* индексу i, если найдется такая цепочка индексов $k_0 = j, k_1, k_2, ..., k_m = i$, что каждый индекс k_ν непосредственно подчинен индексу $k_{\nu+1}$ $(\nu = 0, 1, ..., m-1)$.

Определение

Будем говорить, что матрица A есть *матрица монотонного типа*, и писать $A \in M^*$, если выполняются условия:

1) при всех $i = 1, 2, ..., n \quad a_{ii} > 0$ и при всех $k \neq i \quad a_{ik} \leq 0$;

2) при всех $i \quad \sum_{k \neq i} |a_{ik}| \leq a_{ii}$;

3) каждый индекс i либо сам является избыточным, либо ему подчинен некоторый избыточный.

Лемма 2

Пусть матрица A монотонного типа и пусть векторы $x = (\xi_1, \xi_2, ..., \xi_n)$ и $y = (\eta_1, \eta_2, ..., \eta_n)$ связаны равенством $Ax = y$. Тогда если при всех $k = 1, 2, ..., n \quad \eta_k \leq 0$, то при всех k и $\xi_k \leq 0$.

Доказательство

Допустим противное: пусть среди компонент ξ_k вектора k имеются положительные. Назовем максимальными те индексы k, для которых $\xi_k = \max\limits_i \xi_i > 0$. Пусть k — некоторый максимальный индекс (такие заведомо существуют). Тогда, приравнивая компоненты с номером k векторов Ax и y, видим

$$a_{kk}\xi_k = \eta_k - \sum_{j\neq k} a_{kj}\xi_j = \eta_k + \sum_{j\neq k}\left|a_{kj}\right|\xi_j \leq \sum_{j\neq k}\left|a_{kj}\right|\xi_j \leq \sum_{j\neq k}\left|a_{kj}\right|\xi_k \leq a_{kk}\xi_k .$$

Поскольку первый и последний члены этой цепочки равенств и неравенств совпадают, то на месте знаков $\leq$ всюду могут быть только знаки равенства. В последнем случае это означает, что индекс k не может быть избыточным, а в предпоследнем — что непосредственно подчиненные ему индексы также являются максимальными. Итак, каждый максимальный индекс не является избыточным, и все непосредственно подчиненные ему индексы (а значит, и все просто подчиненные) также являются максимальными. Это противоречит тому, что максимальному индексу, как и всякому другому неизбыточному, должен быть подчинен некоторый избыточный. ■

Теорема 2

Пусть $A \in M^*$ и пусть векторы[1] x, y, x' и y' связаны соотношениями $Ax = y$, $Ax' = y'$. Если при всех $k = 1, 2, ..., n$ $\left|\eta_k\right| \leq \eta_k'$, то при всех k и $\left|\xi_k\right| \leq \xi_k'$.

Доказательство

Все компоненты вектора $A(x - x') = y - y'$ неположительны. Значит, по лемме 2 при всех k и $\xi_k - \xi_k' \leq 0$. Точно так же неположительны и компоненты вектора $A(-x - x') = -y - y'$, а значит и $-\xi_k - \xi_k' \leq 0$. Итак, $\xi_k \leq \xi_k'$, $-\xi_k \leq \xi_k'$, $\left|\xi_k\right| \leq \xi_k'$. ■

[1] Далее используются очевидные обозначения компонент этих векторов.

Следствие. Всякая матрица A монотонного типа неособенная.

Доказательство

Положим $x' = 0$ — нулевой вектор, а x — любое решение однородной системы уравнений $Ax = 0$, так что $y = Ax = 0$ и $y' = Ax' = 0$. Применяя теорему, при всех k имеем $|\xi_k| \le 0$, и тем самым $x = 0$. Таким образом, однородная система уравнений $Ax = 0$ имеет только нулевое решение, и определитель матрицы A отличен от нуля. ■

- **Замечание.** После того как мы доказали обратимость матриц монотонного типа, следует обратить внимание на то, что утверждения леммы 2 и теоремы 2 означают, что если $A \in M^*$, то все элементы матрицы A^{-1} неотрицательны.

Рассмотрим теперь один вопрос, связанный с методом последовательных приближений (методом итерации) для решения систем линейных алгебраических уравнений. Пусть все диагональные элементы a_{ii} матрицы A отличны от нуля и пусть нам требуется решить систему линейных алгебраических уравнений

$$Ax = y. \tag{1}$$

Применим для этой цели следующий итеративный метод. Исходя из начального приближения $x^0 = (\xi_1^0, ..., \xi_n^0)$, последующие будем строить по формулам

$$\begin{aligned}
\xi_1^{\nu+1} &= (\eta_1 - a_{12}\xi_2^{\nu} - ... - a_{1n}\xi_n^{\nu})/a_{11},\\
\xi_2^{\nu+1} &= (\eta_2 - a_{21}\xi_1^{\nu} - a_{23}\xi_3^{\nu} - ... - a_{2n}\xi_n^{\nu})/a_{22},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
\xi_n^{\nu+1} &= (\eta_n - a_{n1}\xi_1^{\nu} - a_{n2}\xi_2^{\nu} - ... - a_{nn-1}\xi_{n-1}^{\nu})/a_{nn}.
\end{aligned}$$

Такой метод будем называть методом итерации для системы (1) в стандартной форме. Очевидно, что закон вычисления следующих приближений в этом методе может быть описан также формулой

$$x^{\nu+1} = Qx^{\nu} + z, \tag{2}$$

где элементы матрицы $Q = \{q_{ik}\}$ задаются равенством

$$q_{ik} = \begin{cases} 0 & \text{при} \quad k = i; \\ -a_{ik} / a_{ii} & \text{при} \quad k \neq i, \end{cases}$$

а вектор правых частей $z = (\zeta_1, ..., \zeta_n)$ таков: $\zeta_i = \eta_i / a_{ii}$.

Теорема 3

Если матрица A принадлежит классу M^*, то для системы уравнений (1) метод итерации в стандартной форме сходится.

Доказательство

Как известно, для сходимости метода последовательных приближений (2), начиная с любого начального приближения, необходимо и достаточно, чтобы спектральный радиус (максимальный из модулей собственных чисел) матрицы Q был меньше единицы. Обозначим спектральный радиус через ρ. Очевидно, что ввиду выполнения условия 2) из определения матриц монотонного типа будет

$$\|Q\|_\infty = \max_i \sum_{k=1}^{n} |q_{ik}| = \max_i \sum_{k \neq i} |a_{ik}| / |a_{ii}| \leq 1,$$

откуда сразу же следует, что $\rho \leq 1$. Остается убедиться, что матрица Q не имеет собственных чисел, равных по модулю единице. Допустим противное: пусть λ ($|\lambda| = 1$) есть собственное число (вообще говоря, комплексное) матрицы A, а $x = (\xi_1, ..., \xi_n)$ — соответствующий ему собственный вектор. Тогда при всех k

$$\lambda a_{kk} \xi_k = \sum_{i \neq k} a_{ki} \xi_i .$$

Дальнейшие рассуждения будут почти те же, что и при доказательстве леммы 2, но на этот раз индекс k назовем максимальным, если $|\xi_k| = \max_i |\xi_i|$. Для каждого максимального индекса k имеем

$$a_{kk} |\xi_k| = |\lambda a_{kk} \xi_k| \leq \sum_{i \neq k} |a_{ki}| \cdot |\xi_i| \leq \sum_{i \neq k} |a_{ki}| \cdot |\xi_k| \leq a_{kk} |\xi_k|.$$

Отсюда следует, что максимальный индекс не может быть избыточным и что каждый непосредственно подчиненный ему индекс (а значит, и просто подчиненный) также является максимальным. Это противоречит условию 3) из определения матриц монотонного типа. ■

❑ **Замечание.** Доказанная теорема допускает некоторое усиление — мы нигде при доказательстве не использовали, что $a_{ik} \le 0$ при $k \ne i$.

§ 3. Метод конечных элементов для обыкновенного дифференциального уравнения

Дифференциальное уравнение

$$-(pu')'+qu'+ru = f$$

(здесь $p \ge p_0 > 0$ — непрерывно дифференцируемая, а q, r, и f — непрерывные функции) мы рассматриваем при одном из граничных условий:

I. $u(0) = u(1) = 0$.

II. $u'(0) + \alpha u(0) = u'(1) + \beta u(1) = 0$

($\alpha \le 0, \beta \ge 0$ и хоть один из этих коэффициентов отличен от нуля[1]).

Напомним, что метод конечных элементов есть метод Галеркина при специальном выборе координатных функций: $\omega_1^{(n)}, \omega_2^{(n)}, ..., \omega_{n-1}^{(n)}$ для задачи I и $\omega_0^{(n)}, \omega_1^{(n)}, ..., \omega_n^{(n)}$ для задачи II. Здесь $\omega_k^{(n)}$ есть функция, график которой изображен на рис. 13 ($x_k = kh$, $h = 1/n$).

Желая применить к исследованию метода конечных элементов теоремы о методе Галеркина *(см. § 9 главы 3)*, формализуем прежде всего поставленную задачу. В пространстве $L_2(0,1)$ введем в рассмотрение оператор B: $Bu = -(pu')'$. Оператор B задан первоначально на множестве дважды непре-

[1] В действительности последнее предположение делается лишь для большей простоты рассмотрения — полученный далее результат верен и без него.

рывно дифференцируемых функций, удовлетворяющих граничным условиям; он положительно определен, и мы считаем его расширенным таким образом, что $R(B) = L_2$ (не трудно было бы показать, что после такого расширения $D(B)$ есть множество функций, удовлетворяющих граничным условиям, первая производная которых абсолютно непрерывна, а вторая суммируема с квадратом). Энергетическое пространство H_B оператора B есть $\overset{0}{W}{}_2^1$ при граничном условии I и W_2^1 при условии II *(см. § 7 главы 3)*. Определим еще оператор C, положив $Cu = qu' + ru$ и считая при этом $D(C) = W_2^1$. Ясно, что $D(C) \supseteq H_B \supseteq D(B)$. Поэтому оператор $T = B^{-1}C$ задан на всем энергетическом пространстве H_B.

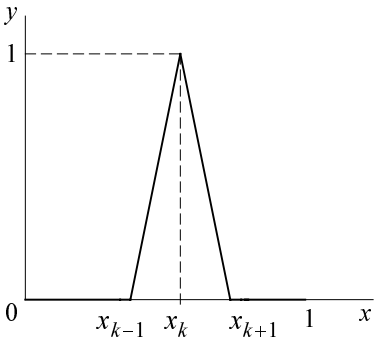


Рис. 13

Лемма 1

Оператор T, рассматриваемый в пространстве H_B, компактен.

Доказательство

Напомним, что

$$\| u \|_B^2 = \int_0^1 pu'^2\, dx + \beta p(1)u^2(1) - \alpha p(0)u^2(0) \geq m^2 \| u \|_{L_2}^2$$

(в случае граничной задачи I два последних слагаемых в правой части равенства отсутствуют). Поэтому оператор C ввиду эквивалентности норм в пространствах H_B и W_2^1 ограничен, если его рассматривать из H_B в L_2, и нам достаточно установить компактность оператора B^{-1}, рассматриваемого из L_2 в H_B. Оператор B^{-1} задается функцией Грина

$$(B^{-1}\upsilon)(x) = \int_0^1 G(x,y)\upsilon(y)dy, \quad (B^{-1}\upsilon)'(x) = \int_0^1 G_x'(x,y)\upsilon(y)dy.$$

Интегральный оператор с непрерывным ядром G компактен как оператор из пространства L_2 в пространство C, а интегральный оператор с ограниченным ядром G_x' компактен из L_2 в L_2. Поэтому из любой последовательности $\{\upsilon_k\}$, ограниченной в L_2, можно выделить такую частичную $\{\upsilon_{k_\nu}\}$, что последовательность $\{B^{-1}\upsilon_{k_\nu}\}$, сходится равномерно, а $\{(B^{-1}\upsilon_{k_\nu})'\}$ — в L_2. Но тогда последовательность $\{B^{-1}\upsilon_{k_\nu}\}$ сходится в H_B. Значит, оператор B^{-1}, рассматриваемый из L_2 в H_B, компактен. ■

Нашу граничную задачу можно записать в виде уравнения в пространстве L_2:

$$Au = f, \tag{1}$$

где оператор $A = B + C$ имеет ту же область задания, что и B. Если предположить еще, что наша граничная задача однозначно разрешима, то для уравнения (1) выполнены все условия, которые налагались на такое уравнение в *§ 9 главы 3*. Перейдем к изучению свойств координатной системы.

Лемма 2

Пусть $u(x)$ — дважды непрерывно дифференцируемая функция, причем

$$|u''(x)| \le M$$

и пусть

$$u_n(x) = \sum_{k=o}^{n} \omega_k^{(n)}(x)u(x_k). \tag{2}$$

Тогда при всех $x \in (0,1)$, отличных от x_k [1],

$$\left| u'(x) - u_n'(x) \right| \le Mh.$$

Доказательство

$u_n(x)$ есть непрерывная кусочно-линейная функция — на каждом промежутке $[x_k, x_{k+1}]$ она линейна, ибо таковы все функции $\omega_j^{(n)}$. Кроме того, $u_n(x_k) = u(x_k)$. Поэтому на каждом промежутке $[x_k, x_{k+1}]$ $u_n(x)$ есть линейный интерполяционный многочлен функции $u(x)$, построенный по узлам x_k и x_{k+1}. Для точки $x \in (x_k, x_{k+1})$

$$u_n'(x) = \frac{u_n(x_{k+1}) - u_n(x_k)}{h} = \frac{u(x_{k+1}) - u(x_k)}{h} = u'(\xi),$$

где $\xi \in (x_k, x_{k+1})$. Поэтому

$$\left| u'(x) - u'_n(x) \right| = \left| u'(x) - u'(\xi) \right| \le M \left| x - \xi \right| \le Mh. \ \blacksquare$$

Обозначим через $U^{(n)}$ подпространство пространства H_B, натянутое на $\omega_1^{(n)}, ..., \omega_{n-1}^{(n)}$ для задачи I и на $\omega_0^{(n)}, ..., \omega_n^{(n)}$ для задачи II. В случае задачи II функция $u(x)$ из леммы заведомо принадлежит пространству H_B, а $u_n \in U^{(n)}$. В случае первой граничной задачи для того, чтобы $u \in H_B$, надо еще потребовать, чтобы $u(0) = u(1) = 0$. Тогда формула (2) может быть переписана в виде

$$u_n(x) = \sum_{k=1}^{n-1} \omega_k^{(n)}(x) u(x_k),$$

и опять $u_n \in U^{(n)}$. Теперь из доказанной леммы могут быть извлечены такие следствия.

Следствие 1. Если в условиях леммы 2 $u \in H_B$ [2], то

$$\left\| u - u_n \right\|_B \le M \sqrt{\max p(x)} \cdot h.$$

---

[1] В узлах x_k. функция $u_n(x)$, вообще говоря, не дифференцируема.

[2] При граничном условии II это заведомо так.

Следствие 2. При тех же условиях

$$E_{U^{(n)}}(u)_{H_B} \le M\sqrt{\max p(x)}\cdot h.$$

Следствие 3. Координатная система $\left\{\omega_k^{(n)}\right\}$ полна в H_B.

Все эти следствия очевидны; при доказательстве третьего следует только учесть, что $D(B)$ плотно в H_B.

Теперь общая теорема о сходимости метода Галеркина *(см. § 9 главы 3)* позволяет нам утверждать сходимость метода конечных элементов в том смысле, что $\left\| u_n^* - u^* \right\|_B \to 0$, причем

$$\left\| u_n^* - u^* \right\|_B \le CE_{U^{(n)}}(u^*)_{H_B} = O(h).$$

В последнем равенстве использовано, что решение u^* нашей задачи дважды непрерывно дифференцируемо. Из сходимости в пространстве H_B вытекает равномерная сходимость. Неизвестными в системе уравнений метода конечных элементов являются значения приближенного решения в узлах x_k (это и есть коэффициенты при $\omega_k^{(n)}$). Поэтому при формулировке результата мы предпочтем говорить о сходимости в узлах. Итак, доказано утверждение.

Теорема 1

Пусть рассматриваемая граничная задача однозначно разрешима. Тогда система линейных уравнений, к которой приводит метод конечных элементов, однозначно разрешима при всех достаточно больших n, и метод сходится в том смысле, что[1]

$$\max_{0\le k\le n}\left| u_k - u^*(x_k) \right| = O(1/n) = O(h).$$

- **Замечание.** Подчеркнем, что оценка быстроты сходимости метода конечных элементов, указанная в теореме, получена, исходя лишь из того, что решение u^* дважды непрерывно дифференцируемо. Это еще одно преимущество метода конечных элементов по сравнению с обычным спосо-

[1] В формуле далее узлы x_k и значения u_k зависят, конечно, от n, что не нашло отражения в обозначениях.

бом составления системы конечно-разностных уравнений. Чтобы при этом обычном способе мера аппроксимации имела порядок убывания $O(h)$, пришлось бы предположить трехкратную дифференцируемость u^*. Сейчас мы сопоставляли, в сущности, оценки быстроты сходимости, но и в действительности для задач с малой гладкостью решения метод конечных элементов сходится быстрее, чем обычный вариант метода сеток.

Обратимся к вопросам устойчивости метода конечных элементов.

Теорема 2

В условиях теоремы 1 для матрицы A_h коэффициентов линейных уравнений, к которой приводит метод конечных элементов, справедливы оценки

$$\|A_h\|_2 \le C_1/h, \quad \|A_h^{-1}\|_2 \le C_2/h.$$

Доказательство

Рассмотрим оператор $Du = -u''+u$ при граничных условиях $u(0) = u(1) = 0$ в случае граничной задачи I и $u'(0) = u'(1) = 0$ в случае граничной задачи II. Этот оператор полусходен с оператором B, и поэтому по *теореме 2 из § 9 главы 3* выполняются оценки

$$\|A_h\|_2 \le C_3\|\Gamma_n^D\|_2, \quad \|A_h^{-1}\|_2 \le C_4\|(\Gamma_n^D)^{-1}\|_2,$$

где $\Gamma_n^D = \{[\omega_k^{(n)}, \omega_j^{(n)}]_D\}$ — матрица Грама координатной системы в энергетическом пространстве H_D. Легко подсчитать, что при граничных условиях II эта матрица такова

$$\Gamma_n^D = \frac{1}{h}\begin{pmatrix} 1+\frac{h^2}{3} & -1+\frac{h^2}{6} & 0 & 0 & \cdots & 0 & 0 \\ -1+\frac{h^2}{6} & 2+\frac{2h^2}{3} & -1+\frac{h^2}{6} & 0 & \cdots & 0 & 0 \\ 0 & -1+\frac{h^2}{6} & 2+\frac{2h^2}{3} & -1+\frac{h^2}{6} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & -1+\frac{h^2}{6} & 1+\frac{h^2}{3} \end{pmatrix},$$

а при граничном условии I она получается из выписанной вычеркиванием первых и последних строк и столбцов; соответствующие формулы для скалярных произведений $[\omega_k^{(n)}, \omega_j^{(n)}]_D$ уже приводились в *§ 1*.

Напомним теорему о кругах Гершгорина. Согласно этой теореме все собственные числа матрицы $A=\{a_{kj}\}$ лежат в объединении кругов на комплексной плоскости

$$S_k = \left\{ z \,\middle|\, |z - a_{kk}| \le \sum\nolimits_{j \ne k} |a_{kk}| \right\}.$$

Учитывая, что собственные числа матрицы вещественны, и применяя эту теорему, получаем, что в случае $h \le 1$ все они принадлежат объединению промежутков $[h/2,\ 2/h + h/6] \cup [h,\ 4/h + h/3]$ (при граничном условии I все собственные числа принадлежат второму из этих промежутков). Отсюда видно, что все собственные числа матрицы Γ_n^D лежат в пределах $h/2 \le \lambda \le 5/h,$ так что

$$\left\| \Gamma_n^D \right\|_2 \le 5/h, \quad \left\| (\Gamma_n^D)^{-1} \right\|_2 \le 2/h$$

и соответственно

$$\left\| A_h \right\|_2 \le 5C_3/h, \quad \left\| (A_h)^{-1} \right\|_2 \le 2C_4/h. \ \blacksquare$$

- **Замечание.** При граничном условии I собственные числа матрицы Γ_n^D легко вычисляются согласно *теореме 1 из § 1*. Наименьшее из них действительно имеет порядок убывания h, а наибольшее — порядок роста $1/h$, так что приводимые в доказанной теореме оценки не являются завышенными.

Доказанная теорема означает, что метод конечных элементов обладает степенной (слабой) неустойчивостью. Такой же неустойчивостью, как будет видно дальше, обладают и многие другие варианты метода сеток для различных задач. Это обстоятельство в действительности не служит препятствием к их применению. Обладая, как правило, медленной сходимостью, метод сеток не дает возможности получать решение с очень большой точностью, и того количества значащих цифр, с которыми ведут вычисления современные ЭВМ, вполне достаточно, чтобы допущенные ошибки округления не проявили себя — внесенный ими в общую погрешность вклад остается малым в

сравнении с погрешностью метода. Кроме того, далеко не все ошибки, допущенные в процессе счета, реально могут возрастать так, как это допускается доказанной теоремой. Подробнее об этом будет сказано в конце *§ 4*.

В теореме 1 сходимость метода конечных элементов и оценка быстроты сходимости были получены при минимальных предположениях о гладкости коэффициентов дифференциального уравнения. Далее для первой граничной задачи мы получим лучшую оценку быстроты сходимости при некоторых дополнительных предположениях о гладкости коэффициентов. Свойство метода конечных элементов, которое будет доказано в теореме 3, можно назвать *сверхсходимостью* этого метода. Не очень устоявшийся термин "сверхсходимость" обычно означает следующее. Для некоторого метода решения какого-нибудь класса уравнений имеется естественная норма, относительно которой доказывается сходимость метода (для метода конечных элементов такой нормой является энергетическая). Если относительно другой, более слабой нормы или полунормы имеет место более быстрая сходимость, то это явление и называется сверхсходимостью.

Напомним, что мы рассматриваем первую граничную задачу. В теореме 2 была получена оценка нормы $\left\|A_h^{-1}\right\|_2$, что составляет обычно наиболее трудную часть проверки условий теоремы о сходимости каркасов. Для применения этой теоремы нам следует еще оценить меру аппроксимации на точном решении. Для этой цели может быть доказано следующее утверждение.

Лемма

Пусть наша задача однозначно разрешима, коэффициент $p(x)$ трижды непрерывно дифференцируем, а коэффициенты $q(x), r(x)$ и правая часть $f(x)$ — дважды. Пусть, далее $z_h = A_h \varphi_h u^*$ — f_h — вектор невязок, которые образуются при подстановке значений в узлах точного решения в сеточные уравнения. Тогда для его компонент справедливы оценки $\left|\zeta_j\right| \le Ch^3 (j = 1, ..., n-1)$, где C — некоторая постоянная.

Доказательство

Доказательство этой леммы связано с довольно громоздкими, но несложными по существу выкладками, которые мы опустим. Заметим лишь, что при

сделанных предположениях о коэффициентах уравнения и его правой части точное решение задачи $u^*(x)$ четырежды непрерывно дифференцируемо и (опуская в обозначении координатных функций $\omega_k^{(n)}$ верхний индекс n):

$$\zeta_j = \sum_{k=j-1}^{j+1} \{[\omega_k, \omega_j] + (q\omega'_k, \omega_j) + (r\omega_k, \omega_j)\} u^*(x_k) - (f, \omega_j),$$

причем

$$(f, \omega_j) = [u^*, \omega_j] + (qu^{*\prime}, \omega_j) + (ru^*, \omega_j),$$

и оценка $|\zeta_j|$ связана с разложением решения u^* и коэффициентов p, q и r в ряды Тейлора в окрестности точки x_j. ■

- **Замечание.** Доказанная лемма не должна создавать впечатление, что метод конечных элементов дает более точную аппроксимацию рассматриваемой задачи, чем обычный способ построения разностных уравнений. Действительно, систему линейных уравнений метода конечных элементов естественно сопоставлять с системой, полученной обычным способом, умноженной на h.

Следствие. В условиях леммы 3 справедливо соотношение

$$\gamma_h(u^*) = \|z_h\|_2 \le Ch^2\sqrt{h}.$$

Обращаясь к общей схеме приближенных методов и используя теорему о сходимости каркасов приближенных решений *(см. § 2 главы 1)*, теперь имеем следующее утверждение.

Теорема 3

Пусть рассматриваемая задача (при первом граничном условии) однозначно разрешима, коэффициент $p(x)$ трижды непрерывно дифференцируем, а коэффициенты $q(x), r(x)$ и правая часть $f(x)$ — дважды. Тогда метод конечных элементов сходится с быстротой, определяемой соотношением

$$\left\|\varphi_h u^* - u_h^*\right\|_2 = \sqrt{\sum_{k=1}^{n-1} (u^*(x_k) - u_k)^2} = O(h\sqrt{h}).$$

❑ **Замечание.** В пространстве каркасов U_h более естественно рассматривать норму $\|u_h\|_{U_h} = \sqrt{h}\|u_h\|_2$, которая является аналогом нормы функций в пространстве $L_2(0,1)$, и для непрерывных функций при $h \to 0$ выполняется соотношение $\|\varphi_h u\|_{U_h} \to \|u\|_{L_2}$. Для этой нормы в условиях теоремы

$$\left\|\varphi_h u^* - u_h^*\right\|_{U_h} = O(h^2).$$

Поскольку в пространстве U_h выполняется очевидное неравенство $\|u_h\|_\infty \le \|u_h\|_2$, то из доказанной теоремы немедленно вытекает утверждение:

Следствие. В условиях теоремы 3 справедливо соотношение

$$\max_{1\le k\le n-1}\left|u^*(x_k) - u_k\right| = O(h\sqrt{h}).$$

❑ **Замечание.** Исследование сверхсходимости для второй граничной задачи тем же способом потребовало бы еще оценки невязки, которая получается при подстановке значений точного решения в первое (нулевое) и последнее уравнение. Соответствующие невязки имеют, вообще говоря, порядок малости лишь $O(h^2)$, так что мера аппроксимации в норме пространства l_{n+1}^2 имеет порядок малости $O(h^2)$, так что $\left\|\varphi_h^* u - u_h^*\right\|_2 = O(h)$, а для более естественной нормы, введенной в предыдущем замечании, $\left\|\varphi_h^* u - u_h^*\right\|_{U_h} = O(h\sqrt{h})$. Доказать более быструю, чем в теореме 1, равномерную сходимость в случае второй задачи на этом пути не удается.

§ 4. Метод сеток для эллиптических уравнений

В этом параграфе мы займемся изучением того варианта метода сеток для эллиптических уравнений, который был изложен в *§ 1*. Напомним, что там рассматривалась граничная задача

$$\left.\begin{aligned} -\Delta u + ru &= f \quad \text{в } \Omega, \\ u\big|_{\partial\Omega} &= g, \end{aligned}\right\} \tag{1}$$

где Ω — ограниченная двумерная область с границей $\partial\Omega$. Будем предполагать, что функция r удовлетворяет условию $r(x,y) \geq 0$ для всех $(x,y) \in \Omega$. При построении разностной схемы в качестве узлов были взяты точки $(x_j, y_k) \in \overline{\Omega}$, где $x_j = x_0 + jh$, $y_k = y_0 + kh$. Каждому узлу было сопоставлено некоторое уравнение. Именно, если узел (j,k) внутренний ($(j,k) \in I_1$), то это уравнение таково:

$$-\frac{u_{jk+1} + u_{jk-1} + u_{j+1k} + u_{j-1k} - 4u_{jk}}{h^2} + r_{jk}u_{jk} = f_{jk},$$

а если (j,k) — узел граничной полоски ($(j,k) \in I_2$), то уравнение имеет вид

$$(1 + \tau_{jk})u_{jk} - \tau_{jk}u_{j-1k} = g(M_{jk}).$$

Здесь $0 \leq \tau_{jk} < 1$, и вместо u_{j-1k} в левой части уравнения может стоять либо u_{j+1k}, либо u_{jk-1}, либо u_{jk+1}, в зависимости от того, какой из "соседних" узлов выходит за пределы области Ω. Итак, построение приближенного решения $u_h = \{u_{jk}\}$ сводится к решению системы линейных алгебраических уравнений, которую мы запишем в виде

$$A_h u_h = f_h. \tag{2}$$

Сделаем одно общее замечание относительно тех обозначений, которые будут использоваться при изучении метода сеток не только в этом параграфе, но и в следующих. Искомый вектор u_h имеет размерность N_h, равную числу узлов, попавших в $\overline{\Omega}$. Мы считаем, как уже говорилось в *§ 1*, что все узлы перенумерованы в некотором порядке, и нумерация неизвестных и уравнений определяется этой нумерацией узлов. Однако в обозначении компонент вектора u_h (а также и вектора f_h) нам удобно сохранить двойную индексацию в качестве указания того узла, которому эта компонента соответствует, т. е., другими словами, нам удобно одновременно рассматривать u_h как функцию, заданную на сетке (т. е. на множестве $I = \{(j,k)\}$ тех пар индексов, для которых $(x_j, y_k) \in \overline{\Omega}$). Поэтому мы будем писать $u_h = \{u_{jk}\}$, причем следует помнить, что, несмотря на наличие у компонент u_h двух индексов, u_h есть вектор, а не матрица.

Изучим теперь свойства матрицы A_h.

Лемма

Матрица A_h принадлежит классу M^*.

Доказательство

Проверим сначала для всех строк матрицы A_h выполнение требований 1) и 2) определения матриц монотонного типа *(см. § 2)*. Если строка матрицы соответствует внутреннему узлу, то ее диагональный элемент есть $4/h_2 + r_{jk} > 0$, а из всех остальных элементов лишь четыре отличны от нуля, и каждый из них есть — $1/h^2$. Таким образом, диагональный элемент положителен, остальные элементы строки не положительны и сумма их абсолютных величин не превосходит диагонального элемента. Заметим, что номер такой строки является избыточным, если $r_{jk} > 0$. Если же строка матрицы соответствует узлу $(j,k) \in I_2$ (т. е. узлу граничной полоски), то диагональный элемент этой строки есть $1 + \tau_{jk}$, а из недиагональных отличен от нуля разве лишь один — он равен $-\tau_{jk}$ и не положителен. Так что и здесь требования 1) и 2) выполнены, причем номер такой строки всегда избыточен. Перейдем к проверке последнего условия принадлежности матрицы классу M^*. Если ν — номер какого-нибудь внутреннего узла (j,k), то ему непосредственно подчинены номера узлов $(j-1,k)$, $(j+1,k)$, $(j,k-1)$, $(j,k+1)$. Поэтому ему подчинены номера узлов $(j,k+1),(j,k+2),...,(j,k+l)$, где $(j,k+l)$ — первый встретившийся узел граничной полоски. Как уже отмечалось, номер узла граничной полоски избыточен. Поэтому каждый индекс либо избыточен, либо ему подчинен некоторый избыточный, и последнее требование определения матриц монотонного типа также выполнено. ■

Следствие. Матрица A_h неособенная.

❑ **Замечание.** Из доказанной только что леммы вытекает также, что для решения системы линейных алгебраических уравнений (2) можно применять метод итерации в стандартной форме *(см. теорему 3 из § 2)*. Если через

u_{jk}^{ν} обозначить приближения, полученные на ν-м шаге метода итерации, то вычислительные формулы метода будут выглядеть так:

$$u_{jk}^{\nu+1} = \frac{u_{j+1k}^{\nu} + u_{j-1k}^{\nu} + u_{jk+1}^{\nu} + u_{jk-1}^{\nu}}{4 + h^2 r_{jk}} + \frac{h^2}{4 + h^2 r_{jk}} f_{jk} \quad ((j,k) \in I_1),$$

$$u_{jk}^{\nu} = \frac{\tau_{jk}}{1 + \tau_{jk}} u_{j-1k}^{\nu} + \frac{1}{1 + \tau_{jk}} g(M_{jk}) \quad ((j,k) \in I_2).$$

Равенства второго типа будут, разумеется, выглядеть несколько иначе для тех узлов (j,k), "соседний" по отношению к которым узел, выходящий за пределы Ω, есть не $(j+1,k)$, а какой-нибудь другой. Тщательный анализ показывает, однако, что при малых h сходимость метода итерации будет медленной (наибольшее по модулю собственное число итерируемой матрицы Q_h стремится по модулю к единице с быстротой $1 - |\lambda_{\max}| = O(h^2)$). Существуют менее трудоемкие методы решения этой системы, останавливаться на которых у нас нет возможности.

Теорема 1

Пусть область Ω заключена в круге $x^2 + y^2 \le R^2$. Тогда при $h \le 1$ выполняется оценка

$$\left\| A_h^{-1} \right\|_{\infty} \le (R+1)^2.$$

Доказательство

Пусть $U = \{u_{jk}\}$ есть решение системы линейных уравнений $A_h U = V$, где $V = \{\upsilon_{jk}\}$ — некоторый вектор, удовлетворяющий единственному условию: $\|V\|_{\infty} = \max |\upsilon_{jk}| < 1$. Введем в рассмотрение функцию $w(x,y) = (R+1)^2 - x^2 - y^2$. Эта функция обладает следующим свойством: $-\Delta w + rw = 4 + rw \ge 4$ для всех точек $(x,y) \in \Omega$. Построим вектор W с компонентами $w_{jk} = w(x_j, y_k)$ и вычислим вектор $Z = A_h W$. Поскольку функция w есть полином второй степени и потому ее вторые производные в точности совпада-

ют со вторыми разностными отношениями, то для $(j,k)\in I_1$ (т. е. в случае, когда (j,k) есть внутренний узел) имеем

$$z_{jk}=\frac{4w_{jk}-w_{j-1k}-w_{j+1k}-w_{jk-1}-w_{jk+1}}{h^2}+r_{jk}w_{jk}=(-\Delta w+rw)_{x_j,y_k}\geq 4.$$

Если $(j,k)\in I_2$ (т. е. (j,k) — узел граничной полоски), то

$$z_{jk}=(1+\tau_{jk})w_{jk}-\tau_{jk}w_{j-1k}=w_{jk}+\tau_{jk}(w_{jk}-w_{j-1k}),$$

или, пользуясь формулой конечных приращений,

$$z_{jk}=w_{jk}+\tau_{jk}hw_x'(\xi,y_k).$$

Но $w_{jk}=w(x_j,y_k)\geq (R+1)^2-R^2$, $\left|w_x'(\xi,y_k)\right|=2|\xi|\leq 2R$, и т. к. $h\leq 1$, $0\leq\tau_{jk}<1$, то $z_{jk}\geq (R+1)^2-R^2-2R=1$. Итак, для всех компонент вектора Z выполняются неравенства $z_{jk}\geq 1\geq|\upsilon_{jk}|$. Поэтому по *теореме 2 из § 2* для всех компонент вектора U будет $|u_{jk}|\leq w_{jk}\leq\max_{\Omega}(x,y)\leq(R+1)^2$, и потому $\|U\|_\infty\leq(R+1)^2$. А т. к. для достижения этого неравенства потребовалось лишь, чтобы было $\|V\|_\infty=\|A_hU\|_\infty\leq 1$, то $\left\|A_h^{-1}\right\|_\infty\leq(R+1)^2$. ■

Теорема 1 позволяет легко доказать сходимость метода сеток.

Теорема 2

Пусть решение u^* задачи (1) четырежды непрерывно дифференцируемо в $\overline{\Omega}$ [1]. Тогда при $h\to 0$ выполняется соотношение

$$\max\left|u^*(x_j,y_j)-u_{jk}\right|=O(h^2). \tag{3}$$

Здесь $\{u_{jk}\}=u_h$ — решение системы уравнений (2), а максимум берется по всем узлам, соответствующим шагу h.

---

[1] Для выполнения этого условия достаточно потребовать, чтобы функции f, r и g, а также граница области $\partial\Omega$ были достаточно гладкими.

Доказательство

Применим теорему о сходимости каркасов приближенных решений *(см. § 2 главы 1)*, взяв в качестве пространств U_h и F_h векторные пространства размерности N_h: $U_h = F_h = l^{\infty}_{N_h}$ (для $u_h = \{u_{jk}\} \in U_h \quad \|u_h\| = \max|u_{jk}|$). При таком выборе пространства U_h левая часть соотношения (3) есть $\left\| \varphi_h u^* - u^*_h \right\|$, и по упомянутой теореме

$$\max\left| u^*(x_j, y_k) - u_{jk} \right| \le \left\| A_h^{-1} \right\|_{\infty} \cdot \gamma_h(u^*).$$

По теореме 1 $\left\| A_h^{-1} \right\|_{\infty} \le (R+1)^2$ (при $h \le 1$), и остается заметить, что $\gamma_h(u^*) = O(h^2)$. Напомним, что $\gamma_h(u^*)$ — норма в пространстве F_h невязки, которая получается при подстановке в сеточные уравнения значений точного решения в узлах. Как отмечалось при построении этих уравнений в *§ 1*, четырежды непрерывно дифференцируемое точное решение удовлетворяет им с точностью до величин порядка малости $O(h^2)$, а норма в F_h — максимальная по модулю компонента. ■

- **Замечание 1.** Сам факт сходимости метода сеток для рассматриваемой задачи может быть установлен при меньших предположениях о гладкости точного решения — достаточно, чтобы функция u^* была дважды непрерывно дифференцируемой в $\overline{\Omega}$. Действительно, для такой функции u^* $\gamma_h(u^*) \to 0$, и рассуждения, проведенные при доказательстве теоремы, останутся в силе, с той только разницей, что получить оценку быстроты сходимости нам не удастся. Вместе с тем бóльшие, чем в теореме, предположения о гладкости решения не позволяют сделать заключение о более быстрой сходимости метода сеток. Как много производных не имело бы решение u^* в $\overline{\Omega}$, погрешность сеточного решения все равно будет лишь порядка $O(h^2)$.

  В теории приближений существует понятие насыщенности аппроксимационного процесса. Не давая точного определения, заметим только, что это означает, что быстрота сходимости зависит от гладкости аппроксимируемой функции лишь до некоторой степени. Усиление предположений о

гладкости не приводит к ускорению сходимости. Таким образом, рассматриваемый вариант метода сеток, как, впрочем, и другие, является насыщенным, и это его орицательное свойство.

❑ **Замечание 2.** Решающим фактором при исследовании метода сеток для нашей задачи оказалось то, что матрица A_h системы уравнений (2) монотонного типа. Однако это же будет верно и при применении метода сеток к задаче Дирихле для более общего эллиптического уравнения:

$$\left.\begin{aligned} -\frac{\partial}{\partial x}\left(a\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(b\frac{\partial u}{\partial y}\right)+c\frac{\partial u}{\partial x}+d\frac{\partial u}{\partial y}+ru=f,\\ u\big|_{\partial\Omega}=g \end{aligned}\right\} \qquad (4)$$

(здесь $a>0,\quad b>0$), если только коэффициент r неотрицателен, а шаг h достаточно мал. Мы считаем при этом, что граничное условие аппроксимируется так же, как и в случае задачи (1), а уравнения, составленные для внутренних узлов, имеют вид

$$-\frac{a_{j+1/2k}(u_{j+1k}-u_{jk})-a_{j-\frac{1}{2}k}(u_{jk}-u_{j-1k})}{h^2}-$$
$$-\frac{b_{jk+1/2}(u_{jk+1}-u_{jk})-b_{jk-\frac{1}{2}}(u_{jk}-u_{jk-1})}{h^2}+$$
$$+c_{jk}\frac{u_{j+1k}-u_{j-1k}}{2h}+d_{jk}\frac{u_{jk+1}-u_{jk-1}}{2h}+r_{jk}u_{jk}=f_{jk}.$$

Малость шага h обеспечивает при этом неположительность всех недиагональных элементов матрицы A_h.

Исследование метода сеток для этой более общей задачи (4) проводится поэтому аналогично тому, как это было сделано для задачи (1). Мы ограничились этой более простой задачей лишь для того, чтобы писать менее громоздкие формулы. Можно было бы также рассмотреть более общий вариант метода, когда шаги h_x и h_y по переменным x и y не обязаны совпадать.

Остановимся теперь на вопросах устойчивости метода сеток для задачи (1). Ясно, что для матрицы A_h выполняется неравенство

$$\|A_h\|_\infty \le \max(8/h^2 + \max r,\ 3),$$

причем если имеется хоть один внутренний узел, то $\|A_h\|_\infty \ge 8/h^2$, так что эта оценка не может быть существенно улучшена. Таким образом (учитывая теорему 1), при малых h

$$\mu_\infty(A_h) \le c/h^2,$$

причем эта оценка опять-таки не допускает улучшения. Точно такая же оценка числа обусловленности системы уравнений метода сеток верна и в случае более общей задачи (4).

Итак, метод сеток для нашей задачи (выражаясь более точно, процесс отыскания каркасов приближенных решений) имеет слабую неустойчивость. Того количества значащих цифр, с которыми ведут вычисления ЭВМ, обычно достаточно, чтобы допускаемые ошибки округления не отражались существенно на полученном приближенном решении. Однако возникает следующий вопрос. Как быть, если сами исходные данные (функции f, g, r) заданы нам неточно? Правые части уравнений (2) — значения функций f и g, и если у нас имеется возможность вычислять их лишь с не очень маленькой относительной погрешностью ε, то согласно *теореме 2 из § 1 главы 1* относительная погрешность в приближенном решении u_h оценивается через $\mu_\infty(A_h)\varepsilon \approx \varepsilon/h^2$. Таким образом, создается впечатление, что в рассматриваемой ситуации метод сеток при малом h вообще не применим. Однако в действительности это не так. Вместо задачи (1) мы решаем методом сеток другую задачу:

$$-\Delta u + (r + dr)u = f + df, \quad u\big|_{\partial\Omega} = g + dg. \tag{5}$$

Если все возмущения dr, df и dg не превосходят некоторого ε, то, как можно показать, разность du^* решений задач (5) и (1) допускает оценку $\|du^*\|_C \le c\varepsilon$. Метод сеток для задачи (5) сходится к $u^* + du^*$, и поэтому при малых h приближенное значение u_h, построенное для возмущенной задачи (5), будет отличаться от u^* на величину, которая оценивается суммой $c\varepsilon + c_1 h^2$. Здесь мы считали возмущения dr, df и dg достаточно гладкими,

такими, что погрешность du^* (а с ней и $u^* + du^*$) четырежды непрерывно дифференцируема. Итак, неточность в задании исходных данных приводит к ошибкам в приближенном решении того же порядка.

В предыдущем абзаце речь шла о "регулярных" возмущениях задачи — при каждом h ошибки, скажем, в правых частях сеточных уравнений были значениями в узлах гладких функций. Разумеется, трактовать таким образом случайные ошибки округления недопустимо. Но и по отношению к таким случайным ошибкам дело обстоит значительно более благополучно, чем можно было бы ожидать. Проще всего исследуется влияние погрешностей, допущенных при вычислении коэффициента r. Если при вычислении значений этого коэффициента мы допускаем относительную погрешность, не превосходящую некоторого ε_h ($\max|dr_{jk}| \Big/ \max|r_{jk}| \le \varepsilon_h$), то вызванная этим относительная погрешность в матрице A_h будет оцениваться числом $\frac{1}{8}\varepsilon_h h^2 \max r$ (т. к. $\|A_h\|_\infty \ge 8/h^2$), и потому ошибка в сеточном решении будет порядка ε_h ($\mu_\infty(A_h)\frac{1}{8}\varepsilon_h h^2 \le c\varepsilon_h$).

Если относительную погрешность ε_h мы допускаем при вычислении правых частей системы (2) (т. е. значений функций f и g), то и абсолютная погрешность в правых частях будет того же порядка. Но тогда согласно теореме 1 тот же порядок малости будет иметь абсолютная погрешность в приближенном решении, а значит, и относительная погрешность, т. к. сеточное решение сходится к точному, и потому норма $\max|u_{jk}|$ ограничена снизу положительным числом. Это свойство сеточных уравнений естественно называть их устойчивостью относительно возмущений правых частей.

Поясним, чем в нашем случае обеспечивается такое благоприятное поведение погрешностей, несмотря на то, что $\mu_\infty(A_h) \to \infty$. Пусть нам дана последовательность "возмущенных" уравнений $A_n u_n = f_n + \Delta f_n$, причем

$$\|A_n^{-1}\| \le c_1, \quad \|A_n\| \to \infty, \quad 0 < c_2 \le \|f_n\| \le c_3, \quad \|\Delta f_n\| \le \varepsilon.$$

Тогда стремление к бесконечности относительных погрешностей реально возможно, но лишь в том случае, если сами решения стремятся к нулю. Однако в случае метода сеток последнее обстоятельство заведомо не имеет места ввиду сходимости метода.

В *§ 3 главы 1* уже говорилось о том, что данное там определение устойчивости не является общепринятым. Высказанные только что соображения являются, видимо, главным доводом в пользу другого определения — *устойчивость есть равномерная ограниченность норм* $\left\|A_h^{-1}\right\|$ (обычно при этом еще добавляются некоторые условия согласования норм в пространствах F и F_h). Однако, возвращаясь к методу сеток для эллиптического уравнения, введем теперь предположение, что при построении матрицы A_h мы допускаем ошибки лишь при вычислении ее диагональных элементов, стоящих в тех строках, которые соответствуют внутренним узлам, — коэффициенты $4/h^2$ заменяем на $(4+h^2)/h^2$. Легко видеть, что тогда решения наших сеточных уравнений будут сходиться к решению другого дифференциального уравнения: $-\Delta u + (r+1)u = f$ при том же граничном условии. Итак, стремящиеся к нулю (с порядком h^2) относительные погрешности в приближенных уравнениях приводят к погрешностям в приближенных решениях, которые к нулю не стремятся. Этот пример явно говорит в пользу определения устойчивости, данного в *главе 1*.

Утверждение о неустойчивости (слабой) предупреждает о том, что возможны малые случайные ошибки, которые существенно скажутся на точности приближенного решения.

Вместе с тем все здесь сказанное позволяет сделать еще один вывод. Исследование устойчивости (или характера неустойчивости) полезно дополнять исследованием, какого рода ошибки являются более существенными, а какого — менее.

§ 5. Явная разностная схема для уравнения теплопроводности

Пусть требуется в прямоугольнике $0 \le x \le 1, \quad 0 \le t \le T$ найти решение дифференциального уравнения теплопроводности

$$\frac{\partial u}{\partial t} - a\frac{\partial^2 u}{\partial x^2} = f \tag{1}$$

(здесь $0 < a_0 \le a(x,t) \le a_1$), удовлетворяющее начальному условию

$$u(x,0) = g(x)$$

и граничным условиям

$$u(0,t) = \alpha(t), \quad u(1,t) = \beta(t).$$

Напомним, что решение этой задачи методом сеток по явной разностной схеме, указанной в *§ 1*, заключается в следующем. Выбираем некоторое натуральное n и шаг по времени τ; шаг по переменной x полагается равным $h = 1/n$. В дальнейшем удобно считать (мы так и будем делать), что шаг τ полностью определяется заданием шага h: $\tau = \tau(h)$, с тем чтобы семейство приближенных решений было однопараметрическим, зависело только от шага h. Функцию $\tau(h)$ будем называть законом зависимости шага τ от шага h (или просто *законом зависимости шагов*). Далее полагаем

$$m = [T/\tau], \quad x_j = jh \quad (j = 0, 1, ..., n), \quad t_k = k\tau \quad (k = 0, 1, ..., m),$$

и приближения u_{jk} к значениям решения в узлах сетки (x_j, t_k) находим из системы уравнений

$$\frac{u_{jk} - u_{jk-1}}{\tau} - a_{jk-1}\frac{u_{j+1k-1} - 2u_{jk-1} + u_{j-1k-1}}{h^2} = f_{jk-1}$$

$$(j = 1, 2, ..., n-1; \quad k = 1, 2, ..., m)$$

(выписанное уравнение считаем сопоставленным узлу (j,k)),

$$u_{j0} = g_j \quad (j = 0, 1, ..., n),$$

$$u_{0k} = a_k, \quad u_{nk} = \beta_k \quad (k = 1, 2, ..., m)$$

(эти уравнения сопоставляем соответственно узлам $(j, 0)$, (j, k), (n, k)). Кратко эту систему уравнений запишем в виде

$$A_h u_h = f_h, \tag{2}$$

где $u_h = \{u_{jk}\}$ — искомый вектор; A_h — матрица коэффициентов; f_h — вектор правых частей[1]. Для записи системы уравнений в такой форме мы, как

[1] Подчеркнем, что компонентами вектора f_h являются значения не только функции f, но и функций g, α, β.

обычно, считаем, что все узлы каким-либо образом перенумерованы и каждая неизвестная u_{jk} и каждое уравнение получают номер соответствующего узла.

В дальнейшем нам потребуется одно существенное ограничение, налагаемое на шаги τ и h. Будем считать закон зависимости шагов таким, что

$$a_1 \frac{\tau}{h^2} \le \frac{1}{2}. \tag{3}$$

Как и в предыдущем параграфе, займемся сначала изучением свойств системы уравнений (2).

Лемма

Если выполнено условие (3), то матрица A_h монотонного типа ($A_h \in M^*$).

Доказательство

Если некоторая строка матрицы A_h соответствует узлу $(j, 0)$, $(0, k)$ или (n, k), то ее диагональный элемент равен единице, а остальные элементы — нули. В строке, соответствующей узлу (j, k) $(1 \le j \le n-1,\ 1 \le k \le m)$, диагональный элемент есть $1/\tau$, и отличными от нуля могут быть еще лишь три элемента, которые равны соответственно $-a_{jk-1}/h^2$, $-1/\tau + 2a_{jk-1}/h^2$, $-a_{jk-1}/h^2$. Все эти элементы неположительны (второй — ввиду условия (3)), и сумма их абсолютных величин есть $1/\tau$. Так что для всех строк диагональный элемент положителен, а остальные элементы строки неположительны, и сумма их абсолютных величин не превосходит диагонального. Избыточными являются номера узлов $(j, 0)$, $(0, k)$, (n, k). При $1 \le j \le n-1$ и $k \ge 1$ номеру узла (j, k) непосредственно подчинены номера узлов $(j-1, k-1)$ и $(j+1, k+1)$, и потому подчинен номер узла $(j-k, 0)$ (если $j \ge k$) или $(0, k-j)$ (если $k > j$), который избыточен. Таким образом, номер каждой строки либо сам является избыточным, либо ему подчинен некоторый избыточный. Итак, все свойства, которым должны удовлетворять матрицы монотонного типа, проверены. ■

Теорема 1

Если выполняется условие (3), то для матрицы A_h справедлива оценка

$$\left\|A_h^{-1}\right\|_\infty \le 1 + 1/(8a_0).$$

Доказательство

Доказательство этой теоремы вполне аналогично доказательству *теоремы 1 из § 3*. Пусть $u_h = \{u_{jk}\}$ — решение системы уравнений $A_h u_h = \upsilon_h$, где $\upsilon_h = \{\upsilon_{jk}\}$ — произвольный вектор, удовлетворяющий условию $\left\|\upsilon_h\right\|_\infty = \max\left|\upsilon_{jk}\right| \le 1$. Зададим в прямоугольнике $0 \le x \le 1, \quad 0 \le t \le T$ функцию $w(x,t) = 1 + [1/(2a_0)]x(1-x)$ (от t она фактически не зависит) и построим вектор $w_h = \{w_{jk}\}$, положив $w_{jk} = w(x_j, t_k)$. Вычислим компоненты вектора $z_h = A_h w_h$. Очевидно, что

$$z_{j0} = w_{j0} = 1 + \frac{1}{2a_0}x_j(1-x_j) \ge 1,$$

$$z_{0k} = w_{0k} = 1, \quad z_{nk} = w_{nk} = 1,$$

а при $1 \le j \le n-1, 1 \le k \le m$

$$z_{jk} = \frac{w_{jk} - w_{jk-1}}{\tau} - a_{jk-1}\frac{w_{j+1k-1} - 2w_{jk-1} + w_{j-1k-1}}{h^2} =$$

$$= -a_{jk-1}\left(-\frac{1}{a_0}\right) = \frac{a_{jk-1}}{a_0} \ge 1.$$

Здесь мы воспользовались тем, что $w(x,t)$ не зависит от t и относительно x есть многочлен второй степени, причем $w''_{x^2}(x,t) = -1/a_0$. Итак, для всех компонент вектора z_h выполняется неравенство $z_{jk} \ge 1 \ge \left|\upsilon_{jk}\right|$. По теореме 2 из *§ 2* можно теперь сделать заключение, что при всех j и k

$$\left|u_{jk}\right| \le w_{jk} \le \max w(x,t) = 1 + 1/(8a_0). \ \blacksquare$$

Теперь легко может быть доказана теорема о сходимости.

Теорема 2

Пусть для закона зависимости шагов выполнено условие (3). Пусть решение $u^*(x,t)$ нашей задачи дважды непрерывно дифференцируемо по t и четырежды — по x. Тогда метод сеток сходится в том смысле, что

$$\max_{j,k}\left|u^*(x_j,t_k)-u_{jk}\right|=O(h^2+\tau). \qquad (4)$$

Здесь u_{jk} — компоненты решения системы уравнений (2).

Доказательство

Левая часть оценки (4) есть $\left\|\varphi_h u^*-u_h^*\right\|_{U_h}$, где $U_h=F_h=l_N^\infty$ есть N-мерное векторное пространство $(N=(n+1)(m+1))$. По теореме сходимости каркасов приближенных решений *(см. § 2 главы 1)* $\left\|\varphi_h u^*-u_h^*\right\|\le\left\|A_h^{-1}\right\|\cdot\gamma_h(u^*)$. По теореме 1 нормы $\left\|A_h^{-1}\right\|$ равномерно ограничены. Сделанные предположения о гладкости решения гарантируют, что при подстановке в сеточные уравнения (2) значений в узлах точного решения эти уравнения будут удовлетворяться с точностью до величин порядка малости $O(h^2+\tau)$. Учитывая, что норма в F_h — максимальная по модулю компонента, имеем $\gamma_h(u^*)=O(h^2+\tau)$. ■

- **Замечание 1.** Ввиду наложенного на закон зависимости шагов условия (3) $\tau=O(h^2)$, и правую часть оценки (4) можно заменить на $O(h^2)$.
- **Замечание 2.** Как и в случае метода сеток для эллиптических уравнений, сходимость рассматриваемого метода можно установить при меньших предположениях о гладкости точного решения — достаточно, чтобы оно было дважды непрерывно дифференцируемо по x и один раз по t, но быстрота сходимости тогда может быть существенно меньше.
- **Замечание 3.** Легко видеть, что

$$\left\|A_h\right\|_\infty\le 2/\tau+4a_1/h^2.$$

Поэтому при условии (3) $\mu_\infty(A_h)=O(1/\tau+1/h^2)$ (мы воспользовались теоремой 1), и метод имеет степенную (слабую) неустойчивость.

❑ **Замечание 4.** Условие (3), наложенное на соотношение между шагами h и τ, является существенным для утверждений теорем 1 и 2. Если это условие не выполнено, то, вообще говоря, имеет место показательная неустойчивость процесса отыскания каркасов приближенных решений, и счет становится практически невозможным. Докажем эту показательную неустойчивость в предположении, что коэффициент a в уравнении (1) есть постоянная.

Теорема 3

Пусть коэффициент a дифференциального уравнения (1) есть постоянная $(a > 0)$. Если закон зависимости шагов таков, что $\tau \to 0$ и

$$2\tau a / h^2 \geq q > 1,$$

то имеет место показательная неустойчивость процесса отыскания каркасов приближенных решений. Более точно: при достаточно малых h выполняется неравенство

$$\left\| A_h^{-1} \right\|_\infty \geq q^{[T/\tau]}.$$

Доказательство

Пусть вектор $u_h = \{u_{jk}\}$ $(j = 0, 1, ..., n, \quad k = 0, 1, ..., m)$ есть решение системы уравнений $A_h u_h = \upsilon_h$, где вектор $\upsilon_h = \{\upsilon_{jk}\}$ мы выберем специальным образом. Прежде всего положим $\upsilon_{jk} = 0$ при всех $k \geq 1$ и $\upsilon_{00} = \upsilon_{n0} = 0$. Таким образом, отличными от нуля могут быть лишь компоненты $\upsilon_{10}, \upsilon_{20}, ..., \upsilon_{n-10}$; их выбор мы укажем позднее, а пока обозначим через V^0 $(n-1)$-мерный вектор $V^0 = (\upsilon_{10}, \upsilon_{20}, ..., \upsilon_{n-10})$. Введем еще $m+1$ векторов размерности $n-1$, связанных с вектором u_h, положив

$$U^k = (u_{1k}, u_{2k}, ..., u_{n-1k}) \quad (k = 0, 1, ..., m).$$

Поскольку из равенства $A_h u_h = \upsilon_h$ следует, что $u_{0k} = \upsilon_{0k}$, $u_{nk} = \upsilon_{nk}$ и $u_{j0} = \upsilon_{j0}$, то $u_{0k} = u_{nk} = 0$ $(k = 0, 1, ..., m)$, $U^0 = V^0$. Явный характер разно-

стной схемы позволяет нам выразить компоненты вектора U^k через компоненты вектора U^{k-1}. Учитывая, что $\upsilon_{jk} = 0$ при $k \geq 1$, имеем

$$u_{jk} = \kappa u_{j+1k-1} + (1-2\kappa)u_{jk-1} + \kappa u_{j-1k-1}, \quad k = 1, 2, ..., m, \quad j = 1, 2, ..., n-1,$$

где $\kappa = \tau a/h^2 \geq q/2 > 1/2$, причем при $j = 1$ u_{0k-1}, а при $j = n-1$ u_{nk-1} можно заменить на нули. Таким образом, векторы U^k и U^{k-1} связаны равенством

$$U^k = BU^{k-1},$$

где B — матрица порядка $n-1$:

$$B = \begin{pmatrix} 1-2\kappa & \kappa & 0 & \cdots & 0 & 0 \\ \kappa & 1-2\kappa & \kappa & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \kappa & 1-2\kappa \end{pmatrix}.$$

По теореме 1 из *§ 2* собственные числа этой матрицы суть

$$\lambda_k = 1 - 2\kappa + 2\kappa\cos(k\pi/n), \quad k = 1, 2, ..., n-1.$$

Наибольшим по абсолютной величине среди этих собственных чисел является λ_{n-1}, причем

$$|\lambda_{n-1}| = \left|1 - 2\kappa\left(1 + \cos\frac{\pi}{n}\right)\right| = 2\kappa\left(1 + \cos\frac{\pi}{n}\right) - 1 =$$

$$= 4\kappa\left(1 - \sin^2\frac{\pi}{2n}\right) - 1 > 2q\left(1 - \frac{\pi^2}{4n^2}\right) - 1.$$

Будем считать теперь h настолько малым, что $\pi^2/(2n^2) = (\pi^2/2)h^2 \leq$ $\leq (q-1)/q$. При таких значениях h будем иметь

$$|\lambda_{n-1}| > q\left(2 - \frac{\pi^2}{2n^2}\right) - 1 \geq q\left(2 - \frac{q-1}{q}\right) - 1 = q.$$

Выберем теперь в качестве вектора V^0 собственный вектор матрицы B, соответствующий собственному числу λ_{n-1}, нормированный условием

$\left\|V^0\right\|_\infty = 1$. Тогда, учитывая равенство $U^0 = V^0$, получаем $U^k = \lambda_{n-1}^k V^0$. Поэтому

$$\left\|u_h\right\|_\infty = \max_{0\le k\le m}\left\|U^k\right\|_\infty \ge \left\|U^m\right\|_\infty = \left|\lambda_{n-1}\right|^m \left\|V^0\right\|_\infty = \left|\lambda_{n-1}\right|^{[T/\tau]} \ge q^{[T/\tau]}.$$

Итак, в то время как $\left\|\upsilon_h\right\|_\infty = \left\|V^0\right\|_\infty = 1$, вектор $u_h = A_h^{-1}\upsilon_h$ таков, что $\left\|u_h\right\| \ge q^{[T/\tau]}$. Значит, $\left\|A_h^{-1}\right\| \ge q^{[T/\tau]}$. ■

Еще раз обратим внимание на то, что ввиду быстрого роста норм матриц A_h^{-1} счет по явной схеме невозможен при малых шагах h и τ, если $\tau a/h^2 > 1/2$. Поясним это примером.

Пример

Пусть мы решаем, используя явную разностную схему, задачу:

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0,$$
$$u(x,0) = g(x), \quad u(0,t) = u(1,t) = 0,$$

считая при этом, что $\max\left|g(x)\right| = 1$, так что в силу принципа максимума для уравнения теплопроводности решение $u^*(x,t)$ при всех x и t также не превосходит по модулю единицы. Пусть, забыв об условии (3), мы выбрали шаги $\tau = h = 0{,}01$. Далее мы будем использовать обозначения, введенные при доказательстве теоремы 3. Вектор $G = \left(g(x_1), ..., g(x_{99})\right)$ имеет разложение $G = \varepsilon V^0 + W$, где вектор W ортогонален вектору $V^0 : (W, V^0) = 0$. Поскольку матрица B симметрична, то при любом j будет

$$(B^j W, V^0) = \lambda_{99}^j (W, V^0) = 0.$$

Поэтому из равенства $U^j = \varepsilon\lambda_{99}^j V^0 + B^j W$ следует, что

$$(U^j, V^0) = \varepsilon\lambda_{99}^j (V^0, V^0) \quad \text{и} \quad \left|(U^j, V^0)\right| \ge \left|\varepsilon\right| \cdot \left|\lambda_{99}\right|^j$$

(мы воспользовались тем, что $\left\| V^0 \right\|_\infty = 1$, так что $(V^0, V^0) \geq 1$), откуда

$$\left\| U^j \right\|_\infty \geq \frac{1}{100} |\varepsilon| \cdot |\lambda_{99}|^j. \tag{5}$$

Займемся оценкой собственного числа λ_{99}. При выбранных шагах τ и h $\varkappa = 100$, и потому

$$\lambda_{99} = 200\big(1 + \cos(\pi/100)\big) - 1 > 398 > 10^{5/2}.$$

Из (5) теперь следует: $\left\| U^j \right\|_\infty > 10^{5j/2-2}\varepsilon$. Допустим, что выполняется неравенство[1] $|\varepsilon| \geq 10^{-12}$. Тогда при $t = 0{,}06$ (что соответствует $j = 6$) среди значений приближенного решения имеются бóльшие по модулю десяти, при том, что точное решение не превосходит единицы. При $j = 100$ (т. е. в момент времени $t = 1$) получается чудовищная оценка $\left\| U^{100} \right\|_\infty > 10^{236}$. Заметим, что при шаге $h = 0{,}01$ согласно (4) мы должны были бы выбрать $\tau \leq 0{,}00005$, т. е. для получения решения на промежутке $t \in [0, 1]$ нам пришлось бы сделать не менее 20 000 шагов по времени, что, впрочем, для компьютера не составляет особого труда.

§ 6. О системах уравнений

В этом параграфе мы займемся изучением в абстрактной форме системы уравнений вида

$$A^{(s)}u = f^{(s)}, \quad s = 1, 2, ..., m, \tag{1}$$

где $A^{(s)}$ — линейные операторы, действующие из одного и того же пространства U в пространства $F^{(s)}$. В форме (1) могут быть записаны граничные задачи для дифференциальных уравнений. Так, граничная задача для уравнения теплопроводности, рассмотренная в *§ 5*, есть, по существу, система

---

[1] Собственный вектор, соответствующий λ_{99}, имеет знакопеременные компоненты, и в случае гладкой функции g коэффициент ε будет очень мал. Однако компоненты вектора G содержат еще ошибки округления, которые ведут себя нерегулярно. Поэтому мы и считаем, что ε имеет порядок ошибок округления, допускаемых компьютером.

четырех уравнений, одно из них есть дифференциальное уравнение, одно представляет начальное условие и два — граничные условия. Расчленение единой задачи на несколько уравнений помимо чисто технических преимуществ позволяет иногда устанавливать более тонкие результаты, касающиеся в отдельности аппроксимации дифференциального уравнения, начальных и граничных условий.

Мы будем рассматривать приближенный метод, который состоит в замене системы уравнений (1) другой системой, зависящей от параметра h:

$$A_h^{(s)} u_h = f_h^{(s)} = \psi_h^{(s)} f^{(s)}, \quad s = 1, 2, ..., m, \tag{2}$$

где $A_h^{(s)} \in L(U_h, F_h^{(s)})$, $\psi_h^{(s)} \in L(F^{(s)}, F_h^{(s)})$. Заметим, что каждое уравнение системы (2) связано лишь с одним из уравнений системы (1), которое имеет тот же номер. Не все применяемые методы укладываются в эту схему. Напомним, что в вариационно-разностных методах *(см. § 1, разд. "Вариационно-разностные схемы")* у нас появлялись уравнения, которые одновременно учитывали как дифференциальное уравнение, так и граничные условия.

Как и раньше, будем считать заданным оператор сноса $\varphi_h \in L(U, U_h)$. Теперь при каждом s мы можем изобразить введенные операторы и пространства схемой, вполне аналогичной той, которая была раньше (рис. 14). Обратим внимание на то, что левая часть этой схемы одна и та же при всех s.

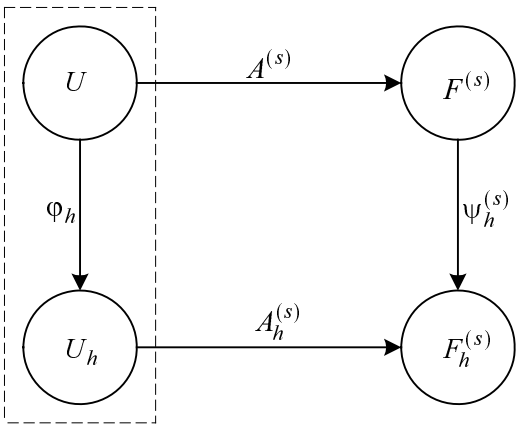


Рис. 14

Разумеется, изучение приближенного метода решения системы можно свести к случаю одного уравнения (так мы и поступали в двух предыдущих параграфах), если ввести пространства $F = F^{(1)} \times F^{(2)} \times ... \times F^{(m)}$, $F_h = F_h^{(1)} \times \times F_h^{(2)} \times ... \times F_h^{(m)}$ и операторы A, A_h, ψ_h, действующие по правилам:

$$Au = (A^{(1)}u, A^{(2)}u, ..., A^{(m)}u) \in F,$$

$$A_h u_h = (A_h^{(1)}u_h, A_h^{(2)}u_h, ..., A_h^{(m)}u_h) \in F_h$$

и для $f = (f^{(1)}, ..., f^{(m)}) \in F$ $\psi_h f = (\psi_h^{(1)} f^{(1)}, ..., \psi_h^{(m)} f^{(m)}) \in F_h$. Тогда система (1) может быть переписана в виде одного уравнения $Au = f$, а система (2) — в виде $A_h u_h = \psi_h f$. Однако мы проведем самостоятельное рассмотрение случая системы уравнений, хотя только что введенные обозначения (пространства F, F_h, операторы A, A_h, ψ_h) нам и потребуются в дальнейшем. При этом для определенности мы будем считать, что нормы в пространствах F и F_h введены по правилу

$$\left\| (f^{(1)}, ..., f^{(m)}) \right\|_F = \max_{1 \le s \le m} \left\| f^{(s)} \right\|_{F^{(s)}},$$

$$\left\| (f_h^{(1)}, ..., f_h^{(m)}) \right\|_{F_h} = \max_{1 \le s \le m} \left\| f_h^{(s)} \right\|_{F_h^{(s)}}.$$

Отметим, что более обще можно было бы положить

$$\left\| (f^{(1)}, ..., f^{(m)}) \right\|_F = \nu\left((\left\| f^{(1)} \right\|, ..., \left\| f^{(m)} \right\|) \right),$$

где ν — некоторая норма в m-мерном пространстве, и аналогично можно было бы определить норму в F_h. Ввиду эквивалентности всех норм в R^m качественный характер результатов, связанных с этими нормами, не изменился бы.

Вполне аналогично тому, как это было раньше, дадим теперь следующее определение.

Определение 1

Пусть $u \in D(A^{(s)})$. *Мерой аппроксимации оператора* $A^{(s)}$ *оператором* $A_h^{(s)}$ на элементе u называется число

$$\gamma_h^{(s)}(u) = \left\| A_h^{(s)} \varphi_h u - \psi_h^{(s)} A^{(s)} u \right\|_{F_h^{(s)}}.$$

Сразу же отметим очевидную связь между $\gamma_h(u) = \| A_h \varphi_h u - \psi_h A u \|_{F_h}$ и "частными" мерами аппроксимации $\gamma_h^{(s)}(u)$: для $u \in D(A) = \bigcap D(A^{(s)})$ выполняется равенство $\gamma_h(u) = \max_s \gamma_h^{(s)}(u)$.

Решение u_h^* системы уравнений (2) мы будем называть *каркасом приближенного решения* и будем говорить, что имеет место сходимость каркасов приближенных решений, если выполняется соотношение $\| \varphi_h u^* - u_h^* \|_{U_h} \to 0$.

Пусть система уравнений (2) однозначно разрешима. Это означает существование обратного оператора A_h^{-1}. Пусть этот оператор ограничен: $A_h^{-1} \in L(F_h, U_h)$. Нам удобно разложить оператор A_h^{-1} на составляющие по следующему правилу. При каждом $s = 1, 2, ..., m$ построим оператор $[A_h^{-1}]^{(s)} \in L(F_h^{(s)}, U_h)$, определяемый формулой

$$[A_h^{-1}]^{(s)} f_h^{(s)} = A_h^{-1}(0, ..., 0, f_h^{(s)}, 0, ..., 0).$$

Тогда ясно, что для любого элемента $f_h = (f_h^{(1)}, ..., f_h^{(m)}) \in F_h$

$$A_h^{-1} f_h = \sum_{s=1}^{m} [A_h^{-1}]^{(s)} f_h^{(s)}. \qquad (3)$$

Норма $\| A_h^{-1} \|$ связана с нормами $\| [A_h^{-1}]^{(s)} \|$ неравенствами

$$\max_s \| [A_h^{-1}]^{(s)} \| \le \| A_h^{-1} \| \le \sum_s \| [A_h^{-1}]^{(s)} \|.$$

Первое из них следует непосредственно из определения операторов $[A_h^{-1}]^{(s)}$, а второе — из равенства (3).

Теперь легко может быть доказана теорема о сходимости каркасов.

Теорема 1

Пусть система уравнений (1) имеет решение u^* и пусть при достаточно малых h существуют ограниченные обратные операторы A_h^{-1}.

Если при всех s выполняются соотношения

$$\left\| [A_h^{-1}]^{(s)} \right\| \cdot \gamma_h^{(s)}(u^*) \to 0,$$

то каркасы приближенных решений сходятся, причем

$$\left\| \varphi_n u^* - u_h^* \right\| \le \sum_{s=1}^{m} \left\| [A_h^{-1}]^{(s)} \right\| \cdot \gamma_h^{(s)}(u^*). \qquad (4)$$

Доказательство

Утверждение о сходимости является непосредственным следствием (4). Само неравенство (4) вытекает из такого равенства:

$$\varphi_h u^* - u_h^* = A_h^{-1} A_h \varphi_h u^* - A_h^{-1} \psi_h A u^* =$$

$$= A_h^{-1}(A_h \varphi_h u^* - \psi_h A u^*) = \sum_{s=1}^{m} [A^{-1}]^{(s)} (A_h^{(s)} \varphi_h u^* - \psi_h^{(s)} A^{(s)} u^*). \quad \blacksquare$$

Отметим, что применение теоремы о сходимости каркасов из *§ 2 главы 1* непосредственно к уравнению $Au = f$, эквивалентному системе (1), привело бы нас к оценке

$$\left\| \varphi_n u^* - u_h^* \right\| \le \left\| A_h^{-1} \right\| \cdot \gamma_h(u^*),$$

правая часть которой удовлетворяет неравенствам

$$\max_s \left\| [A_h^{-1}]^{(s)} \right\| \cdot \gamma_h^{(s)}(u^*) \le \left\| A^{-1} \right\| \cdot \gamma_h(u^*) \le \sum_s \left\| [A_h^{-1}]^{(s)} \right\| \cdot \gamma_h^{(s)}(u^*).$$

Оценка (4) имеет преимущество в том случае, когда для тех s, при которых велика норма $\left\| [A_h^{-1}]^{(s)} \right\|$, мала мера аппроксимации $\gamma_h^{(s)}(u^*)$, и наоборот.

Исследование устойчивости процесса отыскания каркасов приближенных решений уравнения $Au = f$ требует оценки числа обусловленности операторов A_h.

Ввиду очевидного равенства $\|A_h\| = \max \left\| A_h^{(s)} \right\|$ и указанной выше оценки для $\left\| A_h^{-1} \right\|$ и в этой задаче нам достаточно изучить в отдельности операторы $A_h^{(s)}$ и составляющие $[A_h^{-1}]^{(s)}$ обратных операторов. Но следует отметить, что

число обусловленности $\mu(A_h) = \|A_h\| \cdot \|A_h^{-1}\|$ не может быть оценено только через произведения $\|A_h^{(s)}\| \cdot \|[A_h^{-1}]^{(s)}\|$. В то же время для систем уравнений можно естественным образом указать несколько другое понятие устойчивости, к изучению которого мы сейчас переходим.

Введем специальное обозначение

$$\mu^{(s)}(A_h) = \|A_h^{(s)}\| \cdot \|[A_h^{-1}]^{(s)}\|$$

и будем называть это произведение *частным* (s-м) *числом обусловленности* оператора A_h. Ясно, что $\mu^{(s)}(A_h)$ зависит не только от $A_h^{(s)}$, но и от всех остальных $A_h^{(j)}$. Очевидно также, что $1 \le \mu^{(s)}(A_h) \le \mu(A_h)$.

Выявим роль частных чисел обусловленности. Наряду с системой уравнений (2) будем рассматривать возмущенную систему

$$(A_h^{(s)} + \Delta A_h^{(s)})u_h = f_h^{(s)} + \Delta f_h^{(s)}, \tag{5}$$

решение которой обозначим через $u_h^* + \Delta u_h$. Введем относительные погрешности

$$\delta A_h^{(s)} = \frac{\|\Delta A_h^{(s)}\|}{\|A_h^{(s)}\|}, \quad \delta f_h^{(s)} = \frac{\|\Delta f_h^{(s)}\|}{\|f_h^{(s)}\|}, \quad \delta u_h = \frac{\|\Delta u_h\|}{\|u_h^*\|}.$$

Теорема 2

Пусть при достаточно малых h существуют обратные операторы $A_h^{-1} \in L(F_h, U_h)$ и пусть возмущения $\Delta A_h^{(s)} \in L(U_h, F_h^{(s)})$ и $\Delta f_h^{(s)} \in F_h^{(s)}$ уравнений (2) таковы, что при всех s

$$\mu^{(s)}(A_h)\delta A_h^{(s)} \underset{h\to 0}{\to} 0, \quad \mu^{(s)}(A_h)\delta f_h^{(s)} \underset{h\to 0}{\to} 0. \tag{6}$$

Тогда при достаточно малых h существуют обратные операторы $(A_h + \Delta A_h)^{-1} \in L(F_h, U_h)$ и $\delta u_h \underset{h\to 0}{\to} 0$.

Доказательство

Введем "промежуточные" возмущения оператора A_h

$$\Delta^k A_h = (\Delta A_h^{(1)}, ..., \Delta A_h^{(k)}, 0, ..., 0)$$

и методом индукции по k покажем, что при достаточно малых h существуют обратные операторы $(A_h + \Delta^k A_h)^{-1}$ и что при всех $s = 1, 2, ..., m$

$$\mu^{(s)}(A_h + \Delta^k A_h)\delta A_h^{(s)} \underset{h\to 0}{\to} 0. \qquad (7)$$

При $k = 0$ $\Delta^0 A_h = 0$ и соотношения (7) совпадают с (6). Это дает нам базу для индукции. Покажем возможность индуктивного перехода. Допустим, что при некотором $k < m$ оператор $(A_h + \Delta^k A_h)^{-1} \in L(F_h, U_h)$ существует и выполняются соотношения (7). Введем в рассмотрение подпространство $U_k^{(k+1)}$ пространства U_h:

$$U_k^{(k+1)} = \{u_h \in U_h \,\big|\, (A_h^{(s)} + \Delta A_h^{(s)})u_h = 0 \text{ при } s = 1, 2, ..., k,$$

$$A_h^{(s)} u_h = 0 \text{ при } s = k+2, ..., m\}.$$

Ясно, что это подпространство может быть записано также в форме

$$U_k^{(k+1)} = [(A_h + \Delta^k A_h)^{-1}]^{(k+1)} F_h^{(k+1)}.$$

Обозначим через $B_h^{(k+1)}$ сужение оператора $A_h^{(k+1)}$ на подпространство $U_h^{(k+1)}$: $B_h^{(k+1)} = A_h^{(k+1)}\Big|_{U_h^{(k+1)}}$ и соответственно $\Delta B_h^{(k+1)} = \Delta A_h^{(k+1)}\Big|_{U_h^{(k+1)}}$. Ввиду обратимости оператора $A_h + \Delta^k A_h$ оператор $B_h^{(k+1)}$ обратим и

$$(B_h^{(k+1)})^{-1} = [(A_h + \Delta^k A_h)^{-1}]^{(k+1)}.$$

Далее

$$\left\|(B_h^{(k+1)})^{-1}\right\| \cdot \left\|\Delta B_h^{(k+1)}\right\| \le \left\|[(A_h + \Delta^k A_h)^{-1}]^{(k+1)}\right\| \cdot \left\|\Delta A_h^{(k+1)}\right\| =$$

$$= \mu^{(k+1)}(A_h + \Delta^k A_h)\delta A_h^{(k+1)} = \varepsilon_h^{(k+1)} \underset{h\to 0}{\to} 0.$$

Поэтому при достаточно малых h окажется $\varepsilon_h^{(k+1)} < 1/2$ и при таких h существует обратный оператор

$$(B_h^{(k+1)} + \Delta B_h^{(k+1)})^{-1} \in L(F_h^{(k+1)}, U_h^{(k+1)}),$$
$$\left\|(B_h^{(k+1)} + \Delta B_h^{(k+1)})^{-1}\right\| \le 2\left\|(B_h^{(k+1)})^{-1}\right\| = 2\left\|[(A_h + \Delta^k A_h)^{-1}]^{(k+1)}\right\|.$$

Покажем теперь, что при тех же h существует обратный оператор $(A_h + \Delta^{k+1} A_h)^{-1} \in L(F_h, U_h)$. Для этого нам требуется установить однозначную разрешимость системы уравнений

$$(A_h^{(s)} + \Delta A_h^{(s)})u_h = f_h^{(s)}, \quad s = 1, 2, ..., k+1,$$
$$A_h^{(s)} u_h = f_h^{(s)}, \quad s = k+2, ..., m,$$

при произвольных правых частях $f_h^{(s)}$. Решение этой системы будем искать в виде

$$u_h = (B_h^{(k+1)} + \Delta B_h^{(k+1)})^{-1} f_h^{k+1} + \sum_{j \ne k+1} [(A_h + \Delta^k A_h)^{-1}]^{(j)} f_h^{(j)} + z_h.$$

Для нового искомого элемента z_h получается система уравнений

$$(A_h^{(s)} + \Delta A_h^{(s)})z_h = 0 \quad \text{при } s = 1, 2, ..., k,$$

$$(A_h^{(k+1)} + \Delta A_h^{(k+1)})z_h = -\Delta A_h^{(k+1)} \sum_{j \ne k+1} [(A_h + \Delta^k A_h)^{-1}]^{(j)} f_h^{(J)}, \qquad (8)$$
$$A_h^{(s)} z_h = 0 \quad \text{при } s = k+2, ..., m.$$

Здесь мы учли, что в силу определения составляющих обратного оператора выполняются равенства[1]

$$(A_h + \Delta^k A_h)^{(s)} [(A_h + \Delta^k A_h)^{-1}]^{(j)} f_h^{(j)} = \begin{cases} 0 & \text{при } s \ne j, \\ f_h^{(j)} & \text{при } s = j \end{cases}$$

[1] Мы пишем $(A_h + \Delta^k A_h)^{(s)}$ вместо $A_h^{(s)} + \Delta A_h^{(s)}$ при $s = 1, 2, ..., k$ и $A_h^{(s)}$ при $s = k+1, ..., m$.

(в частности, $A_h^{(k+1)}[(A_h+\Delta^k A_h)^{-1}]^{(j)} f_h^{(j)} = 0$ при $j \neq k+1$) и

$$(A_h+\Delta^k A_h)^{(s)}(B_h^{(k+1)}+\Delta B_h^{(k+1)})^{-1} f_h^{(k+1)} = \begin{cases} 0 & \text{при } s \neq k+1, \\ f_h^{(k+1)} & \text{при } s = k+1. \end{cases}$$

Система уравнений (8) однозначно разрешима, ее решение есть

$$z_h = -(B_h^{(k+1)}+\Delta B_h^{(k+1)})^{-1}\Delta A_h^{(k+1)} \sum_{j\neq k+1}[(A_h+\Delta^k A_h)^{-1}]^{(j)} f_h^{(j)}.$$

Этим доказано существование $(A_h+\Delta^{(k+1)}A_h)^{-1}$, причем справедливы формулы

$$[(A_h+\Delta^{(k+1)}A_h)^{-1}]^{(s)} = [(A_h+\Delta^k A_h)^{-1}]^{(s)} -$$

$$-(B_h^{(k+1)}+\Delta B_h^{(k+1)})^{-1}(\Delta A_h^{(k+1)}[(A_h+\Delta^k A_h)^{-1}]^{(s)} \quad \text{при } s \neq k+1$$

$$[(A_h+\Delta^{k+1}A_h)^{-1}]^{(k+1)} = (B_h^{(k+1)}+\Delta B_h^{(k+1)})^{-1}.$$

Поскольку при $s \neq k+1$ $(A_h+\Delta^k A_h)^{(s)} = (A_h+\Delta^{k+1}A_h)^{(s)}$, то при таких s из приведенной формулы следует

$$\mu^{(s)}(A_h+\Delta^{k+1}A_h)\delta A_h^{(s)} \le \mu^{(s)}(A_h+\Delta^k A_h)\delta A_h^{(s)} +$$
$$+2\left\|[(A_h+\Delta^k A_h)^{-1}]^{(k+1)}\right\|\cdot\left\|A_h^{(k+1)}\right\|\cdot\delta A_h^{(k+1)}\cdot\mu^{(s)}(A_h+\Delta^k A_h)\delta A_h^{(s)} =$$
$$=\mu^{(s)}(A_h+\Delta^k A_h)\delta A_h^{(s)}\left(1+2\mu^{(k+1)}(A_h+\Delta^k A_h)\delta A_h^{(k+1)}\right)\underset{h\to 0}{\to} 0,$$

а при $s = k+1$

$$\mu^{(k+1)}(A_h+\Delta^{k+1}A_h)\delta A_h^{(k+1)} \le 2\left\|A_h^{(k+1)}+\Delta A_h^{(k+1)}\right\|\cdot\left\|[(A_h+\Delta^k A_h)^{-1}]^{(k+1)}\right\|\times$$
$$\times\,\delta A_h^{(k+1)} \le 2\mu^{(k+1)}(A_h+\Delta^k A_h)(1+\delta A_h^{(k+1)})\delta A_h^{(k+1)} \underset{h\to 0}{\to} 0.$$

Этим соотношения (7) при $k+1$, а значит, в силу принципа индукции и при всех k доказаны. В частности, при $k = m$ $\Delta^m A_h = \Delta A_h$ и потому

$$\mu^{(s)}(A_h+\Delta A_h)\delta A_h^{(s)} \underset{h\to 0}{\to} 0, \quad s = 1, 2, \ldots, m.$$

Наряду с системами уравнений (2) и (5) рассмотрим еще систему

$$A_h^{(s)} u_h = f_h^{(s)} + \Delta f_h^{(s)}, \quad s = 1, 2, ..., m,$$

решение которой обозначим через $u_h^* + \Delta_1 u_h$. Итак, мы имеем тождества

$$A_h^{(s)} u_h^* = f_h^{(s)}, \tag{9}$$

$$A_h^{(s)} (u_h^* + \Delta_1 u_h) = f_h^{(s)} + \Delta f_h^{(s)}, \tag{10}$$

$$(A_h^{(s)} + \Delta A_h^{(s)})(u_h^* + \Delta u_h) = f_h^{(s)} + \Delta f^{(s)}. \tag{11}$$

Вычитая из (10) тождество (9), получаем $A_h^{(s)} \Delta_1 u_h = \Delta f_h^{(s)}$, откуда

$$\Delta_1 u_1 = \sum_{s=1}^m [A_h^{-1}]^{(s)} \Delta f_h^{(s)},$$

$$\|\Delta_1 u_h\| \le \sum_{s=1}^m \left\|[A_h^{-1}]^{(s)}\right\| \cdot \left\|\Delta f_h^{(s)}\right\| = \sum_{s=1}^m \left\|[A_h^{-1}]^{(s)}\right\| \cdot \left\|f_h^{(s)}\right\| \cdot \delta f_h^{(s)} \le$$

$$\le \left\|u_h^*\right\| \sum_{s=1}^m \mu^{(s)}(A_h) \delta f_h^{(s)}, \quad \frac{\|\Delta_1 u_h\|}{\left\|u_h^*\right\|} \underset{h \to 0}{\to} 0.$$

Вычитая из (11) тождество (10), получаем

$$(A_h^{(s)} + \Delta A_h^{(s)})(\Delta u_h - \Delta_1 u_h) = -\Delta A_h^{(s)} (u_h^* + \Delta_1 u_h),$$

откуда

$$\Delta u_h - \Delta_1 u_h = -\sum_{s=1}^m [(A_h + \Delta A_h)^{-1}]^{(s)} \Delta A_h^{(s)} (u_h^* + \Delta_1 u_h),$$

$$\|\Delta u_h - \Delta_1 u_h\| \le \sum_{s=1}^m \left\|[(A_h + \Delta A_h)^{-1}]^{(s)}\right\| \cdot \left\|A_h^{(s)} + \Delta A_h^{(s)}\right\| \frac{\left\|A_h^{(s)}\right\|}{\left\|A_h^{(s)} + \Delta A_h^{(s)}\right\|} \delta A_h^{(s)} \left\|u_h^* + \Delta_1 u_h\right\| \le$$

$$\le \sum_{s=1}^m \mu^{(s)}(A_h + \Delta A_h) \frac{1}{1 - \left\|\Delta A_h^{(s)}\right\| \Big/ \left\|A_h^{(s)}\right\|} \delta A_h^{(s)} \left(1 + \frac{\|\Delta_1 u_h\|}{\left\|u_h^*\right\|}\right) \left\|u_h^*\right\|.$$

Из этой оценки следует, что $\|\Delta u_h - \Delta_1 u_h\| \Big/ \left\|u_h^*\right\| \to 0$, а значит, и $\delta u_h \to 0$. ■

❑ **Замечание.** Как видно из доказательства, если выполнены условия теоремы и нашлась такая бесконечно малая δ_h, что при всех s

$$\mu^{(s)}(A_h)\delta A_h^{(s)} = O(\delta_h) \quad \text{и} \quad \mu^{(s)}(A_h)\delta f_h^{(s)} = O(\delta_h),$$

то и $\delta u_h = O(\delta_h)$.

Перейдем теперь к определению устойчивости.

Определение 2

Пусть при $h \le h_0$ существуют обратные операторы $A_h^{-1} \in L(F_h, U_h)$ и пусть k — одно из чисел $1, 2, ..., m$. Будем говорить, что процесс отыскания каркасов приближенных решений *устойчив по k-му уравнению*, если, каковы бы ни были возмущения $\Delta A_h^{(k)} \in L(U_h, F_h^{(k)})$ и $\Delta f_h^{(k)} \in F_h^{(k)}$, такие, что

$$\delta A_h^{(k)} \underset{h\to 0}{\to} 0, \quad \delta f_h^{(k)} \underset{h\to 0}{\to} 0,$$

при всех достаточно малых h существуют обратные операторы $(A_h + \Delta A_h)^{-1} \in L(F_h, U_h)$ и выполняется соотношение $\delta u_h \to 0$. Здесь $\Delta A_h = (0, ..., \Delta A^{(k)}, 0, ..., 0)$, u_h^* — решение системы уравнений

$$A_h^{(s)} u_h = 0 \quad \text{при} \quad s \ne k, \quad A_h^{(k)} u_h = f_h^{(k)},$$

а $\Delta u_h^* + \Delta u_h$ — решение системы

$$A_h^{(s)} u_h = 0 \quad \text{при} \quad s \ne k, \quad (A_h^{(k)} + \Delta A_h^{(k)}) u_h = f_h^{(k)} + \Delta f_h^{(k)}.$$

Уравнения системы (1) обычно имеют смысл дифференциального уравнения, граничного или начального условия. Тогда вместо того, чтобы говорить об "устойчивости по k-му уравнению", часто говорят об "устойчивости по дифференциальному уравнению", "устойчивости по граничному условию" и т. п.

Признак устойчивости по k-му уравнению дается следующей теоремой.

Теорема 3

Пусть при достаточно малых $h > 0$ существуют обратные операторы $A_h^{-1} \in L(F_h, U_h)$. Для того чтобы процесс отыскания каркасов приближенных решений был устойчив по *k*-му уравнению, необходимо и достаточно, чтобы частные числа обусловленности $\mu^{(k)}(A_h)$ были равномерно ограничены при всех достаточно малых *h*.

Доказательство

1. *Достаточность*

Пусть при $h \le h_0 \quad \mu^{(k)}(A_h) \le c$ и пусть $\delta A_h^{(k)} \to 0, \quad \delta f_h^{(k)} \to 0$. При $s \ne k$ положим $\Delta A_h^{(s)} = 0$. Тогда при всех $s = 1, 2, ..., m \quad \mu^{(s)}(A_h)\delta A_h^{(s)} \to 0$.

Из теоремы 2 следует, что при достаточно малых *h* существуют обратные операторы $(A_h + \Delta A_h)^{-1} \in L(F_h, U_h)$, и, как видно из доказательства этой теоремы,

$$\mu^{(k)}(A_h + \Delta A_h)\delta A_h^{(k)} \to 0.$$

Доказательство соотношения $\delta u_h \to 0$ проводится теперь точно так же, как в теореме 2. В сущности, мы могли бы просто сослаться на теорему 2, если условиться считать при $f_h^{(s)} = 0$ и $\Delta f_h^{(s)} = 0 \quad \delta f_h^{(s)} = 0$.

2. *Необходимость*

Необходимость будем доказывать от противного. Допустим, что требование равномерной ограниченности при всех достаточно малых *h* чисел $\mu^{(k)}(A_h)$ нарушено. Тогда найдется такая последовательность положительных чисел $h_n \to 0$, что $\mu^{(k)}(A_{h_n}) \to \infty$. Повторим теперь конструкцию, которая уже использовалась при доказательстве *леммы из § 3 главы 1*. При каждом *n* построим такой элемент $f_n^{(k)} \in F_{h_n}^{(k)}$, что для элемента $u_n^{(k)} = [A_{h_n}^{-1}]^{(k)} f_n^{(k)}$ выполняется неравенство

$$\left\| u_n^{(k)} \right\| > \frac{1}{2} \left\| [A_{h_n}^{-1}]^{(k)} \right\| \cdot \left\| f_n^{(k)} \right\|.$$

Далее выберем функционал $G_n \in F^*_{h_n}$ из условий $\|G_n\| = 1, \quad G_n u_n^{(k)} = \left\| u_n^{(k)} \right\|$ и определим оператор $\Delta A_{h_n}^{(k)} \in L(U_{h_n}, F_{h_n}^{(k)})$, положив для произвольного $u_{h_n}^{(k)} \in U_{h_n}$

$$\Delta A_{h_n}^{(k)} u_{h_n} = -G(u_{h_n}) \frac{f_n^{(k)}}{\left\| u_n^{(k)} \right\|}.$$

Тогда $(A_{h_n}^{(k)} + \Delta A_{h_n}^{(k)}) u_n^{(k)} = f_n^{(k)} - f_n^{(k)} = 0$, при $s \neq k \quad A_{h_n}^{(s)} u_n^{(k)} = 0$ ввиду определения $u_n^{(k)}$ и потому $(A_{h_n} + \Delta A_{h_n}) u_n^{(k)} = 0$, так что оператор $A_{h_n} + \Delta A_{h_n}$ необратим. В то же время

$$\left\| \Delta A_{h_n}^{(k)} \right\| = \frac{\left\| f_n^{(k)} \right\|}{\left\| u_n^{(k)} \right\|} < \frac{2}{\left\| \left[A_{h_n}^{(-1)} \right]^{(k)} \right\|}, \qquad \delta A_{h_n}^{(k)} < \frac{2}{\mu^{(k)}(A_{h_n})} \to 0.$$

Таким образом, нашлись такие возмущения $\Delta A_h^{(k)}$ (при $h \neq h_n$, $n = 1, 2, ...$, можно положить $\Delta A_h^{(k)} = 0$), что, несмотря на соотношение $\delta A_h^{(k)} \to 0$, найдутся сколь угодно малые h, при которых оператор $A_h + \Delta A_h$ необратим, т. е. процесс отыскания каркасов неустойчив по k-му уравнению. ■

- **Замечание 1.** Как видно из доказательства теоремы, если процесс отыскания каркасов неустойчив по k-му уравнению, то обязательно нарушено первое условие устойчивости — существуют относительно малые возмущения $\Delta A_h^{(k)} (\delta A_h^{(k)} \to 0)$, при которых обратные операторы $(A_h + \Delta A_h)^{-1}$ не существуют по меньшей мере для бесконечной последовательности значений h, стремящейся к нулю.
- **Замечание 2.** Из доказанной теоремы и теоремы 2 легко следует, что определение устойчивости по k-му уравнению можно дать в другой, эквивалентной определению 2 форме.

Определение 2'

Процесс отыскания каркасов приближенных решений системы уравнений называется *устойчивым по k-му уравнению*, если, каковы бы ни были пра-

вые части $f_h^{(k)}(s=1,2,...,m)$ и возмущения $\Delta A_h^{(k)}$ и $\Delta f_h^{(k)}$, такие, что $\delta A_h^{(k)} \to 0$ и $\delta f_h^{(k)} \to 0$, при всех достаточно малых h существуют обратные операторы $(A_h + \Delta A_h)^{-1}$, где $\Delta A_h = (0,...,\Delta A_h^{(k)},0,...,0)$, и $\delta u_h \to 0$. Здесь u_h^* — решение системы уравнений (2), а $u_h^* + \Delta u_h$ — системы уравнений

$$A_h^{(s)} u_h = f_h^{(s)} \ (s \neq k), \quad (A_h^{(k)} + \Delta A_h^{(k)}) u_h = f_h^{(k)} + \Delta f_h^{(k)}.$$

Дадим теперь определение устойчивости процесса отыскания каркасов для всей системы уравнений в целом. Будем предполагать, что при всех достаточно малых положительных h существуют непрерывные операторы $A_h^{-1} \in L(F_h, U_h)$.

Определение 3

Будем говорить, что процесс отыскания каркасов приближенных решений системы уравнений *устойчив*, если, каковы бы ни были возмущения $\Delta A_h^{(s)} \in L(U_h, F_h^{(s)})$ операторов $A_h^{(s)}$ и возмущения $\Delta f_h^{(s)}$ правых частей $f_h^{(s)}$, такие, что $\delta A_h^{(s)} \underset{h\to 0}{\to} 0$; $\delta f_h^{(s)} \underset{h\to 0}{\to} 0$, при всех достаточно малых h существуют обратные операторы $\left(A_h + \Delta A_h\right)^{-1} \in L(F_h, U_h)$, где $\Delta A_h = (\Delta A_h^{(1)},...,\Delta A_h^{(m)})$, и будет выполняться соотношение

$$\delta u_h \underset{h\to 0}{\to} 0.$$

Здесь u_h^* — решение системы уравнений (2), а $u_h^* + \Delta u_h$ — системы уравнений (5).

Теорема 4

Для того чтобы процесс отыскания каркасов приближенных решений системы уравнений был устойчив, необходимо и достаточно, чтобы он был устойчив по каждому уравнению.

Доказательство

1. *Необходимость* высказанного условия почти очевидна. Действительно, если процесс отыскания каркасов неустойчив по k-му уравнению, то мы мо-

жем при всех $s \neq k$ положить $\Delta A_h^{(s)} = 0$, а возмущения $\Delta A_h^{(s)}$, согласно замечанию 1 к предыдущей теореме, взять такими, что $\delta A_h^{(k)} \to 0$, и тем не менее найдутся сколь угодно малые h, при которых операторы

$$A_h + \Delta A_h = \left(A_h^{(1)}, ..., A_h^{(k)} + \Delta A_h^{(k)}, ..., A_h^{(m)}\right)$$

необратимы, так что процесс отыскания каркасов системы уравнений неустойчив.

2. *Достаточность* является очевидным следствием теорем 2 и 3. ■

Ввиду теоремы 3 теореме 4 можно дать эквивалентную формулировку.

Теорема 4'

Для того чтобы процесс отыскания каркасов приближенных решений системы уравнений был устойчив, необходимо и достаточно, чтобы все частные числа обусловленности $\mu^{(s)}(A)$ были равномерно ограничены при достаточно малых $h > 0$.

Подобно тому, как это делалось в *§ 3 главы 1* для одного уравнения, для систем уравнений также можно ввести понятие степенной и показательной неустойчивости процесса отыскания каркасов приближенных решений. В частности, степенная неустойчивость к k-му уравнению означает существование такого положительного α, что $\mu^{(s)}(A_h) \leq ch^{-\alpha}$. При степенной неустойчивости обычно удается обеспечить такое убывание ошибок $\Delta A_h^{(s)}$ и $\Delta f_h^{(s)}$, чтобы выполнялись соотношения (6) и тем самым согласно теореме 2 относительные погрешности в каркасах приближенных решений, вызванные этими ошибками, стремились к нулю. Подчеркнем, что быстроту убывания относительных погрешностей $\delta A_h^{(s)}$ и $\delta f_h^{(s)}$ в s-м уравнении можно согласовывать лишь с соответствующим s-м частным числом обусловленности $\mu^{(s)}(A_h)$; в частности, если процесс отыскания каркасов устойчив по s-му уравнению, то достаточно, чтобы эти относительные погрешности стремились к нулю. Если каркасы приближенных решений сходятся с порядком $\delta_h : \left\| u_h^* - \varphi u^* \right\| = O(\delta_h)$ и, как это часто бывает, нормы $\left\| u_h^* \right\|$ заключены между двумя положительными числами, то естественно обеспечивать такую

точность построения приближенных уравнений, чтобы выполнялись соотношения

$$\mu^{(s)}(A_h)\delta A_h^{(s)} = O(\delta_h), \quad \mu^{(s)}(A_h)\delta f_h^{(s)} = O(\delta_h).$$

Тогда совершаемые нами ошибки не ухудшат порядок сходимости (δ_h).

Остановимся еще на сравнении понятия устойчивости процесса отыскания каркасов приближенных решений системы уравнений (1) и процесса отыскания каркасов той же системы как единого уравнения $Au = f$. Последнее равносильно равномерной ограниченности (при малых h) чисел обусловленности $\mu(A_h)$. Если процесс отыскания каркасов для этого единого уравнения устойчив, то устойчив и процесс отыскания каркасов для системы уравнений. Это ясно из оценок $\mu^{(s)}(A_h) \le \mu(A_h)$. Но это же нетрудно усмотреть и непосредственно из определений. Действительно, при определении устойчивости процесса отыскания каркасов для системы мы требовали от погрешностей $\Delta A_h^{(s)}$ и $\Delta f_h^{(s)}$ большего — чтобы они стремились к нулю, будучи отнесенными к соответствующим операторам $A_h^{(s)}$ и правым частям $f_h^{(s)}$, в то время как в определении устойчивости для той же системы как единого уравнения требуется лишь стремление к нулю отношений $\left\|A_h^{(s)}\right\| \Big/ \max_k \left\|A_h^{(k)}\right\|$ и $\left\|f_h^{(s)}\right\| \Big/ \max_k \left\|f_h^{(k)}\right\|$. Из устойчивости процесса отыскания каркасов для системы уравнений устойчивость для той же системы как единого уравнения, вообще говоря, не следует.

Рассмотрим теперь тот вариант метода сеток для эллиптического уравнения, который уже изучался в *§ 4*. Уравнения системы линейных алгебраических уравнений этого метода $A_h u_h = f_h$ разобьем на две группы, отнеся к первой те из них, которые аппроксимируют дифференциальное уравнение, а ко второй — граничное условие. Тем самым "приближенное уравнение" запишется в виде системы

$$A_h^{(1)} u_h = f_h^{(1)}, \quad A_h^{(2)} u_h = f_h^{(2)}.$$

По-прежнему будем считать, что $U_h = l_N^\infty$, где $N = N(h)$ — число узлов в замкнутой области $\overline{\Omega}$. В качестве пространств $F_h^{(1)}$ и $F_h^{(2)}$ мы возьмем $F_h^{(1)} = l_{N_1}^{(\infty)}$, $F_h^{(2)} = l_{N_2}^{(\infty)}$, где $N_1 = N_1(h)$ — число внутренних узлов, а

$N_2 = N_2(h)$ — число узлов граничной полоски, так что для всего пространства $F_h = l_N^\infty$ (таким было F_h в *§ 4*) выполняется условие

$$\|f_h\|_{F_h} = \left\|\left(f_h^{(1)}, f_h^{(2)}\right)\right\|_{F_h} = \max\left\{\left\|f_h^{(1)}\right\|_{F_h^{(1)}}, \left\|f_h^{(2)}\right\|_{F_h^{(2)}}\right\}.$$

Поэтому $\left\|[A_h^{-1}]^{(s)}\right\| \le \left\|A_h^{-1}\right\|$ $(s = 1, 2)$ и, как следует из результатов *теоремы 1 из § 4*, при $h \le 1$ $\left\|[A_h^{-1}]^{(s)}\right\| \le (R+1)^2$, т. е эти нормы равномерно ограничены. Совершенно очевидно также, что $\left\|A_h^{(2)}\right\| \le 3$. Итак, при $h \le 1$ $\mu^{(2)}(A_h) \le 3(R+1)^2$ и потому метод сеток для эллиптического уравнения устойчив по граничному условию. Говоря об устойчивости метода здесь и в следующем абзаце, мы имеем в виду устойчивость процесса отыскания каркасов в равномерных нормах.

Аналогично показывается, что явная разностная схема для уравнения теплопроводности при выполнении условия $a_1\tau/h^2 \le 1/2$ устойчива относительно начального и граничных условий.

§ 7. Неявная разностная схема для уравнения теплопроводности

Рассматриваемую граничную задачу для уравнения теплопроводности мы сразу запишем в виде системы уравнений

$$\left.\begin{aligned} A^{(1)}u &= \frac{\partial u}{\partial t} - a\frac{\partial^2 u}{\partial x^2} = f, \\ A^{(2)}u &= u(x, 0) = g, \\ A^{(3)}u &= u(0, t) = \alpha, \\ A^{(4)}u &= u(1, t) = \beta. \end{aligned}\right\} \qquad (1)$$

Решение требуется найти в прямоугольнике $0 \le x \le 1$, $0 \le t \le T$. В этом параграфе коэффициент $a > 0$ дифференциального уравнения мы считаем постоянным.

Мы будем изучать неявную шеститочечную разностную схему, построенную в *§ 1*. Из разностных уравнений найдем таблицу приближенного решения — вектор $u_h = \{u_{jk}\}$ $(j = 0, 1, ..., n, k = 0, 1, ..., m)$ размерности $(n+1)\times(m+1)$, компоненты которого u_{jk} рассматриваются как приближения к значениям точного решения

$$u^*(x_j, t_k), \quad x_j = jh \quad (h = 1/n), \quad t_k = k\tau, \quad m = [T/\tau].$$

Поэтому в аппроксимирующей системе роль пространства U_h будет играть $(n+1)\times(m+1)$-мерное векторное пространство, норму в котором мы определим несколько позже.

Система уравнений, аппроксимирующая систему (1), состоит из четырех "уравнений", каждое из которых есть система линейных алгебраических уравнений. Первое "уравнение" $A_h^{(1)}u_h = f_h$ состоит из уравнений, аппроксимирующих дифференциальное:

$$\frac{u_{jk} - u_{jk-1}}{\tau} - a\left(\frac{u_{j+1k} - 2u_{jk} + u_{j-1k}}{2h^2} + \frac{u_{j+1k-1} - 2u_{jk-1} + u_{j-1k-1}}{2h^2}\right) = \tag{2}$$
$$= f_{jk} \quad \left(f_{jk} = f(x_j, t_k - \tau/2)\right), \quad j = 1, 2, ..., n-1, \quad k = 1, 2, ..., m.$$

Таким образом, $A_h^{(1)}$ есть линейный оператор из $(n+1)\times(m+1)$-мерного векторного пространства U_h в $(n-1)\times m$-мерное векторное пространство $F_h^{(1)}$.

"Уравнение" $A_h^{(2)}u_h = g_h$ состоит из уравнений, аппроксимирующих начальное условие:

$$u_{j0} = g_j \quad \left(g_j = g(x_j)\right), \quad j = 1, 2, ..., n-1,$$

и $A_h^{(2)}$ есть оператор из U_h в $(n-1)$-мерное пространство $F_h^{(2)}$.

Наконец, $A_h^{(3)}u_h = \alpha_h$ и $A_h^{(4)}u_h = \beta_h$ состоят из уравнений, аппроксимирующих первое и второе из граничных условий:

$$u_{0k} = \alpha_k \quad \left(\alpha_k = \alpha(t_k)\right), \quad k = 0, ..., m,$$
$$u_{nk} = \beta_k \quad \left(\beta_k = \beta(t_k)\right), \quad k = 0, ..., m,$$

так что операторы $A_h^{(3)}$ и $A_h^{(4)}$ переводят U_h в $(m+1)$-мерные векторные пространства $F_h^{(3)}$ и $F_h^{(4)}$.

Определим нормы в пространствах $F_h^{(s)}$. Для вектора

$$f_h \in F_h^{(1)},\ f_h = \{f_{jk}\} \quad (j = 1, 2, ..., n-1, \quad k = 1, 2, ..., m)$$

положим

$$\|f_h\|_{F_h^{(1)}} = \max_{1 \le k \le m} \sqrt{h \sum_{j=1}^{n-1} f_{jk}^2}.$$

Поясним смысл этой нормы. Вектор f_h мы можем рассматривать как таблицу некоторой функции $f(x,t)$, заданной в прямоугольнике $0 \le x \le 1$, $0 \le t \le T : f_{jk} = f(x_j, t_k)$. Для функций, заданных в этом прямоугольнике, можно определить норму

$$\|f\| = \max_{t \in [0,T]} \|f(\cdot, t)\|_{L_2(0,1)}$$

— "максимум по t норм в L_2, по x". Если в этой норме интеграл, которым выражается норма в L_2, вычислить по формуле трапеций, пренебрегая слагаемым $(1/2)h\left(f^2(0,t) + f^2(1,t)\right)$, а максимум брать не по всем значениям t из $[0,T]$, а лишь по t_k $(k = 1, 2, ..., m)$, то, как сразу ясно, мы получим $\|f_h\|_{F_h^{(1)}}$.

Таким образом, норма в $F_h^{(1)}$ есть "разностный аналог" только что введенной нормы для функций двух переменных.

Норму в пространстве $F_h^{(2)}$ определим равенством: для $g_h \in F_h^{(2)}$, $g_h = (g_1, ..., g_{n-1})$

$$\|g_h\|_{F_h^{(2)}} = \sqrt{h \sum_{j=1}^{n-1} g_j^2} = \sqrt{h}\,\|g_h\|_2$$

("разностный аналог" нормы в $L_2(0,1)$). Наконец, для $\alpha_h = (\alpha_0, ..., \alpha_m)$ и $\beta_h = (\beta_0, ..., \beta_m)$ положим

$$\|\alpha_h\|_{F_h^{(3)}} = \max_{0 \le k \le m} |\alpha_k| = \|\alpha_h\|_\infty, \quad \|\beta_h\|_{F_h^{(4)}} = \max_{0 \le k \le m} |\beta_k| = \|\beta_h\|_\infty.$$

Так что $F_h^{(3)} = F_h^{(4)} = l_{m+1}^\infty$.

Норму в пространстве $F_h^{(1)} \times F_h^{(2)} \times F_h^{(3)} \times F_h^{(4)}$ определим равенством: для $f_h = (f_h^{(1)}, f_h^{(2)}, f_h^{(3)}, f_h^{(4)})$ положим $\|f_h\|_{F_h} = \max_s \|f_h^{(s)}\|$.

Выберем теперь норму в пространстве U_h так, чтобы она совпадала с нормой в пространстве $F_h = F_h^{(1)} \times F_h^{(2)} \times F_h^{(3)} \times F_h^{(4)}$:

$$\|u_h\|_{U_h} = \max_{0 \le k \le m} \max \left\{ |u_{0k}|, |u_{nk}|, \sqrt{h \sum_{j=1}^{n-1} u_{jk}^2} \right\}.$$

Установим некоторые неравенства для норм в этих пространствах.

Лемма

Для любого вектора $f_h = \{f_{jk}\} \subset R^{N_1} (N_1 = (n-1)m)$ выполняется неравенство

$$\|f_h\|_{F_h^{(1)}} \le \|f_h\|_\infty$$

и для любого вектора $u_h = \{u_{jk}\} \in R^N \left(N = (n+1)(m+1)\right)$ неравенство

$$\|u_h\|_\infty \le \frac{1}{\sqrt{h}} \|u_h\|_{U_h}.$$

Доказательство

Для первого неравенства:

$$\|f_h\|_{F_h^{(1)}} = \max_{1 \le k \le m} \sqrt{h \sum_{j=1}^{n-1} f_{jk}^2} \le \|f_h\|_\infty \sqrt{h\,(n-1)} \le \|f_h\|_\infty.$$

Докажем второе. При $j = 0, n$ очевидно неравенство $|u_{jk}| \le \|u_h\|_{U_h}$, а при $1 \le j \le n-1$

$$|u_{jk}| \le \sqrt{\sum_{l=1}^{n-1} u_{lk}^2} \le \frac{1}{\sqrt{h}} \|u_h\|_{U_h}. \quad \blacksquare$$

Оценим меры аппроксимации $\gamma_h^{(s)}(u^*)$, вычисленные на точном решении u^* нашей задачи.

Теорема 1

Справедливы равенства

$$\gamma_h^{(2)}(u^*) = \gamma_h^{(3)}(u^*) = \gamma_h^{(4)}(u^*) = 0. \tag{3}$$

Если точное решение u^* четырежды непрерывно дифференцируемо в прямоугольнике $\Omega = \{(x,t) \mid 0 \le x \le l,\ 0 \le t \le T\}$, то существуют такие постоянные c_1 и c_2, что $\gamma_h^{(1)}(u^*) \le c_1 h^2 + c_2 \tau^2$.

Доказательство

Начнем со второго утверждения. При составлении разностных уравнений (2) в параграфе 1 отмечалось, что если точное решение u^* четырежды непрерывно дифференцируемо в Ω, то его значения в узлах удовлетворяют этим уравнениям с погрешностью порядка $O(h^2 + \tau^2)$, т. е. все компоненты невязки $\upsilon_h = \{\upsilon_{jk}\}$, получающейся при подстановке этих значений в уравнения (2), допускают оценку $|\upsilon_{jk}| \le c_1 h^2 + c_2 \tau^2$, где постоянные c_1 и c_2 выражаются через максимумы в Ω модулей производных $\dfrac{\partial^3 u^*}{\partial t^3}, \dfrac{\partial^4 u^*}{\partial x^4}, \dfrac{\partial^4 u^*}{\partial x^2 \partial t^2}$ решения u^* и не зависят от h, τ, j и k. Таким образом, $\|\upsilon_h\|_\infty \le c_1 h^2 + c_2 \tau^2$, и остается воспользоваться леммой, учитывая, что $\gamma_h^{(1)}(u^*) = \|\upsilon_h\|_{F_h^{(1)}}$. Второе утверждение доказано.

Первое утверждение очевидно, поскольку уравнениям, аппроксимирующим начальное и граничные условия, решение u^* удовлетворяет точно. ■

❑ **Замечание.** Ввиду формулы (3) и в связи с теоремой 1, доказанной в *§ 6*, при изучении сходимости нашей неявной разностной схемы мы можем ограничиться оценкой лишь одной компоненты оператора A_h^{-1}, а именно $\left[A_h^{-1}\right]^{(1)}$, поскольку в нашем случае

$$\left\|\varphi_h u^* - u_h^*\right\| \le \left\|\left[A_h^{-1}\right]^{(1)}\right\| \cdot \gamma_h^{(1)}(u^*).$$

Однако нас будут интересовать также вопросы устойчивости, поэтому мы оценим $\left\| \left[A_h^{-1} \right]^{(s)} \right\|$ и при $s = 2,3,4$.

Теорема 2

Справедливы оценки

$$\left\| \left[A_h^{-1} \right]^{(1)} \right\| \le T, \quad \left\| \left[A_h^{-1} \right]^{(2)} \right\| \le 1, \quad \left\| \left[A_h^{-1} \right]^{(3)} \right\| \le 3, \quad \left\| \left[A_h^{-1} \right]^{(4)} \right\| \le 3. \tag{4}$$

Доказательство

Доказательство начнем с некоторых обозначений и выкладок, которые будут полезны при оценке всех четырех составляющих оператора A_h^{-1}. Пусть $u_h \in U_h$ — произвольный вектор, удовлетворяющий системе уравнений (2) при некотором наборе правых частей f_{jk} (число уравнений в системе (2) меньше, чем число компонент вектора u_h). Введем в рассмотрение $(n-1)$-мерные векторы $U^{(k)}$, составленные из компонент вектора u_h:

$$U^{(k)} = (u_{1k}, u_{2k}, ..., u_{n-1k}).$$

Умножая равенства (2) на τ и полагая $\kappa = a\tau/(2h^2)$, видим, что выполняются равенства

$$BU^{(k)} = CU^{(k-1)} + \tau V^{(k)}, \quad k = 1, 2, ..., m, \tag{5}$$

где B и C — матрицы порядка $n-1$:

$$B = \begin{pmatrix} 1+2\kappa & -\kappa & & & 0 \\ -\kappa & 1+2\kappa & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 1+2\kappa & -\kappa \\ 0 & & & -\kappa & 1+2\kappa \end{pmatrix}, \quad C = \begin{pmatrix} 1-2\kappa & \kappa & & & 0 \\ \kappa & 1-2\kappa & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 1-2\kappa & \kappa \\ 0 & & & \kappa & 1-2\kappa \end{pmatrix},$$

а $(n-1)$-мерные векторы $V^{(k)} = (\upsilon_{1k}, ..., \upsilon_{n-1k})$ имеют компоненты

$$\upsilon_{1k} = f_{1k} + \frac{a}{2h^2}(u_{0k-1} + u_{0k}), \quad \upsilon_{n-1k} = f_{n-1k} + \frac{a}{2h^2}(u_{nk-1} + u_{nk}),$$

$$\upsilon_{jk} = f_{jk} \text{ при } j = 2, 3, ..., n-2.$$

Поскольку

$$1 + 2\kappa - 2\kappa\cos(k\pi/n) = 1 + 4\kappa\sin^2[k\pi/(2n)] > 0$$

при всех $k = 1, 2, ..., n-1$, то по *теореме 1 из § 2*[1] матрица B обратима, и соотношения (5) можно переписать в виде

$$U^{(k)} = B^{-1}CU^{(k-1)} + \tau B^{-1}V^{(k)}. \qquad (6)$$

Из той же теоремы нам известны все собственные числа μ_l матрицы B^{-1} и λ_l — матрицы $B^{-1}C$:

$$\mu_l = \frac{1}{1 + 2\kappa - 2\kappa\cos(l\pi/n)} = \frac{1}{1 + 4\kappa\sin^2[l\pi/(2n)]},$$

$$\lambda_l = \frac{1 - 2\kappa + 2\kappa\cos(l\pi/n)}{1 + 2\kappa - 2\kappa\cos(l\pi/n)} = \frac{1 - 4\kappa\sin^2[l\pi/(2n)]}{1 + 4\kappa\sin^2[l\pi/(2n)]},$$

причем соответствующие собственные векторы этих матриц таковы:

$$z_l = \sqrt{\frac{2}{n}}\left(\sin\frac{l\pi}{n}, \sin\frac{2l\pi}{n}, ..., \sin\frac{(n-1)l\pi}{n}\right).$$

Очевидным образом для собственных чисел μ_l выполняются оценки $0 < \mu_l < 1$, а для собственных чисел $\lambda_l - |\lambda_l| < 1$. Матрица B^{-1} симметрична. Поэтому

$$\left\|B^{-1}\right\|_2 = \max|\mu_l| < 1.$$

По упоминавшейся *теореме 1 из § 2* матрица $B^{-1}C$ также симметрична, и

$$\left\|B^{-1}C\right\|_2 = \max|\lambda_l| < 1.$$

Итак, доказаны неравенства

$$\left\|B^{-1}\right\|_2 < 1, \quad \left\|B^{-1}C\right\|_2 < 1. \qquad (7)$$

---

[1] Обратим внимание на то, что в этой теореме речь идет о матрицах порядка n, а наши матрицы имеют порядок $n-1$. Этим вызвана замена числа $n+1$ в условиях теоремы на n.

Перейдем теперь к оценке составляющих оператора A_h^{-1}.

1. Вектор $u_h = [A_h^{-1}]^{(1)} f_h$ есть решение системы уравнений

$$A_h^{(1)} u_h = f_h, \quad A_h^{(s)} u_h = 0 \quad \text{при} \quad s = 2, 3, 4,$$

f_h — произвольная правая часть. Нам требуется оценить $\|u_h\|$. Заметим прежде всего, что ввиду равенств $u_{j0} = 0 \quad (j = 1, 2, ..., n-1)$ и $u_{0k} = u_{nk} = 0 \quad (k = 0, ..., m)$

$$\|u_h\| = \max_{1 \le k \le m} \sqrt{h \sum_{j=1}^{n-1} u_{jk}^2} = \sqrt{h} \max_{1 \le k \le m} \left\| U^{(k)} \right\|_2,$$

где $U^{(k)}$ — введенные выше $(n-1)$-мерные векторы. При этом $U^{(0)} = 0$ и векторы $U^{(k)}$ связаны равенствами (6), в которых $V^{(k)} = (f_{1k}, f_{2k}, ..., f_{n-1k})$ и

$$\left\| V^{(k)} \right\|_2 = \sqrt{\sum_{j=1}^{n-1} f_{jk}^2} \le \frac{1}{\sqrt{h}} \max_{1 \le s \le m} \sqrt{h \sum_{j=1}^{n-1} f_{js}} = \frac{1}{\sqrt{h}} \|f_h\|.$$

Равенства (6) и неравенства (7) позволяют нам написать рекуррентную оценку

$$\left\| U^{(k)} \right\|_2 \le \left\| U^{(k-1)} \right\|_2 + \tau \left\| V^{(k)} \right\|_2 \le \left\| U^{(k-1)} \right\|_2 + \frac{\tau}{\sqrt{h}} \|f_h\|.$$

Отсюда, учитывая, что $U^{(0)} = 0,$ методом индукции сразу же получаем

$$\left\| U^{(k)} \right\|_2 \le \frac{k\tau}{\sqrt{h}} \|f_h\| \le \frac{m\tau}{\sqrt{h}} \|f_h\| \le \frac{T}{\sqrt{h}} \|f_h\|$$

и

$$\|u_h\| = \sqrt{h} \max_{1 \le k \le m} \left\| U^{(k)} \right\|_2 \le T \|f_h\|. \tag{8}$$

В частности, если $f_h = 0,$ то и $u_h = 0.$ Этим доказана обратимость оператора A_h. Из неравенства (8) (оно доказано для произвольного вектора $f_h \in F_h^{(1)}$) следует оценка $\left\| \left[A_h^{-1} \right]^{(1)} \right\| \le T.$

2. Оценим $\left\|\left[A_h^{-1}\right]^{(2)}\right\|$. Если u_h есть решение системы уравнений

$$A_h^{(1)}u_h = 0, \quad A_h^{(2)}u_h = g_h, \quad A_h^{(3)}u_h = 0, \quad A_h^{(4)}u_h = 0,$$

то выполняются следующие очевидные утверждения:

а) $\|u_h\| = \sqrt{h}\max\limits_{0\le k\le m}\left\|U^{(k)}\right\|_2$;

б) $U^{(0)} = g_h$;

в) векторы $U^{(k)}$ связаны равенствами (6), в которых $V^{(k)} = 0 \quad (k = 1, 2, ..., m)$, т. е.

$$U^{(k)} = B^{-1}CU^{(k-1)}.$$

Учитывая оценку (7), отсюда сразу же получаем, что при всех k $\left\|U^{(k)}\right\|_2 \le \left\|U^{(0)}\right\|_2$. Итак,

$$\|u_h\| = \sqrt{h}\max_{0\le k\le m}\left\|U^{(k)}\right\|_2 = \sqrt{h}\left\|U^{(0)}\right\|_2 = \sqrt{h}\|g_h\|_2 = \|g_h\|_{F_h^{(2)}},$$

и тем самым $\left\|\left[A_h^{-1}\right]^{(2)}\right\| = 1$.

3. Нормы $\left\|\left[A_h^{-1}\right]^{(3)}\right\|$ и $\left\|\left[A_h^{-1}\right]^{(4)}\right\|$ оцениваются совершенно одинаково, и из соображений симметрии они должны быть равны между собой. Поэтому мы будем оценивать лишь $\left\|\left[A_h^{-1}\right]^{(3)}\right\|$. Итак, пусть u_h есть решение системы уравнений

$$A_h^{(1)}u_h = 0, \quad A_h^{(2)}u_h = 0, \quad A_h^{(3)}u_h = \alpha_h, \quad A_h^{(4)}u_h = 0.$$

Тогда, как легко видеть, для векторов $U^{(k)}$, введенных выше, справедливы утверждения:

а) $U^{(0)} = 0$;

б) $\|u_h\| = \max\left\{\|\alpha_h\|, \sqrt{h}\max\limits_{1\le k\le m}\left\|U^{(k)}\right\|_2\right\}$;

в) векторы $U^{(k)}$ связаны формулой (6), где

$$\upsilon_{1k} = \frac{a}{2h^2}(\alpha_{k-1} + \alpha_k), \quad \upsilon_{jk} = 0 \text{ при } j = 2, 3, ..., n-1.$$

Разложим векторы $V^{(k)}$ и $U^{(k)}$ по собственным векторам z_l матриц B^{-1} и $B^{-1}C$, обозначив соответствующие коэффициенты разложения через b_l^k и c_l^k:

$$V^{(k)} = \sum_{l=1}^{n-1} b_l^k z_l, \quad U^{(k)} = \sum_{l=1}^{n-1} c_l^k z_l.$$

Векторы z_l ортогональны и нормированы. Поэтому

$$b_l^k = (V^{(k)}, z_l) = \sqrt{\frac{2}{n}}\upsilon_{1k} \sin\frac{\pi l}{n} = \sqrt{\frac{2}{n}}\frac{a}{2h^2}\sin\frac{\pi l}{n}(\alpha_{k-1} + \alpha_k).$$

Формула (6) принимает вид

$$\sum_{l=1}^{n-1} c_l^{(k)} z_l = \sum_{l=1}^{n-1} c_l^{(k-1)} \lambda_l z_l + \tau \sum_{l=1}^{n-1} b_l^{(k)} \mu_l z_l.$$

Отсюда, приравнивая коэффициенты при z_l в левой и правой частях, имеем

$$c_l^k = \lambda_l c_l^{(k-1)} + \tau \mu_l c_l^{(k)}.$$

Учитывая, что $c_l^{(0)} = 0$ (т. к. $U^{(0)} = 0$), нетрудно найти явное выражение для c_l^k:

$$c_l^k = \tau \mu_l \sum_{j=1}^{k} \lambda_l^{k-j} b_l^{(j)} = \sqrt{\frac{2}{n}} \kappa \mu_l \sin\frac{\pi l}{n} \sum_{j=1}^{k} \lambda_l^{k-j}(\alpha_{j-1} + \alpha_j).$$

Преобразуем последнюю сумму:

$$\sum_{j=1}^{k} \lambda_l^{k-j}(\alpha_{j-1} + \alpha_j) = \sum_{j=1}^{k} \lambda_l^{k-j}\alpha_j + \sum_{j=0}^{k-1} \lambda_l^{k-j-1}\alpha_j =$$

$$= \alpha_k + \lambda_l^{k-1}\alpha_0 + (1 + \lambda_l)\sum_{j=1}^{k-1} \lambda_l^{k-j-1}\alpha_j.$$

Учитывая это равенство, а также оценку $|\lambda_l| < 1$, теперь получаем

$$|c_l^k| \le \sqrt{\frac{2}{n}} \kappa \mu_l \sin\frac{\pi l}{n} \|\alpha_h\| \left(2 + \frac{1 + \lambda_l}{1 - |\lambda_l|}\right).$$

Если $\lambda_l \le 0$, то $(2+(1+\lambda_l)/(1-|\lambda_l|))=3$ и потому

$$\left|c_l^{(k)}\right| \le 3\sqrt{\frac{2}{n}}\kappa\mu_l \sin\frac{\pi l}{n}\|\alpha_h\| =$$

$$= 3\sqrt{\frac{2}{n}}\frac{\kappa \sin(\pi l/n)}{1+4\kappa\sin^2[\pi l/(2n)]}\|\alpha_h\| \le \frac{3}{4}\sqrt{\frac{2}{n}}\|\alpha_h\|\frac{\sin(\pi l/n)}{\sin^2[\pi l/(2n)]}.$$

Если $\lambda_l > 0$, то

$$2+\frac{1+\lambda_l}{1-\lambda_l} < \frac{3}{1-\lambda_l}, \quad \frac{\mu_l}{1-\lambda_l} = \frac{1}{8\kappa\sin^2[l\pi/(2n)]}$$

и

$$\left|c_l^{(k)}\right| \le \frac{3}{8}\sqrt{\frac{2}{n}}\|\alpha_h\|\frac{\sin(\pi l/n)}{\sin^2[\pi l/(2n)]}.$$

Итак, при всех l

$$\left|c_l^{(k)}\right| \le \frac{3}{4}\sqrt{\frac{2}{n}}\|\alpha_h\|\frac{\sin(\pi l/n)}{\sin^2[\pi l/(2n)]} = \frac{3}{2}\sqrt{\frac{2}{n}}\|\alpha_h\|\frac{\cos[\pi l/(2n)]}{\sin[\pi l/(2n)]},$$

и учитывая, что при $1 \le l \le n-1$ $\sin[\pi l/(2n)] \ge l/n$,

$$\left|c_l^{(k)}\right| \le \frac{3}{2}\sqrt{\frac{2}{n}}\|\alpha_h\|\frac{n}{l},$$

$$\left\|u^{(k)}\right\|_2^2 = \sum_{l=1}^{n-1}(c_l^{(k)})^2 \le \frac{9}{2}n\|\alpha_h\|^2\sum_{l=1}^{n-1}\frac{1}{l^2} \le \frac{3\pi^2}{4}n\|\alpha_h\|^2.$$

В последнем неравенстве учтено, что $\sum_{l=1}^{\infty}1/l^2 = \pi^2/6$. Итак,

$$\|u_h\| = \max\left\{\|\alpha_h\|, \max_{1\le k\le m}\sqrt{h}\left\|U^{(k)}\right\|_2\right\} \le \frac{\pi}{2}\sqrt{3}\sqrt{nh}\|\alpha_h\| = \frac{\pi}{2}\sqrt{3}\|\alpha_h\| \le 3\|\alpha_h\|. \ \blacksquare$$

❑ **Замечание 1.** В отличие от явной разностной схемы для уравнения теплопроводности, рассмотренной в *§ 5*, мы получили оценку нормы обратных операторов вне зависимости от закона изменения шагов $\tau = \tau(h)$, который в теореме 2 может быть произвольным.

❑ **Замечание 2.** Недостатком доказанной теоремы является то, что в правой части оценки $\left\|\left[A_h^{-1}\right]^{(1)}\right\|$ стоит число T, которое велико, если нам требуется найти решение нашей задачи на большом промежутке изменения t. Если наложить на закон зависимости шагов следующее ограничение (которое, впрочем, является вполне естественным): существует такая постоянная c_1, что $\tau \le c_1 h$, то можно получить другую оценку оператора $\left\|\left[A_h^{-1}\right]^{(1)}\right\|$: $\left\|\left[A_h^{-1}\right]^{(1)}\right\| \le c_2$, где постоянная c_2 уже не зависит от T. Для доказательства этой оценки изучим более внимательно собственные числа λ_l матрицы $B^{-1}C$. Для тех l, при которых $4\kappa\sin^2\dfrac{\pi l}{2n} \le 1$, имеем

$$|\lambda_l| = \lambda_l \le 1 - 4\kappa\sin^2\frac{\pi l}{2n} \le 1 - \frac{4\kappa}{\mathrm{n}^2} = 1 - \frac{2a\tau}{n^2h^2} = 1 - 2a\tau = 1 - c_3\tau, \quad c_3 = 2a.$$

Если же $4\kappa\sin^2[\pi l/(2n)] > 1$, то

$$|\lambda_l| = -\lambda_l \le \frac{4\kappa - 1}{4\kappa + 1} \le 1 - \frac{1}{4\kappa} = 1 - \frac{h^2}{2a\tau} \le 1 - \frac{1}{2ac_1^2}\tau = 1 - c_4\tau, \quad c_4 = 1/(2ac_1^2).$$

Итак, в любом случае $|\lambda_l| \le 1 - c_5\tau$, где $c_5 = \min(c_3, c_4)$ и $\left|B^{-1}C\right|_2 \le 1 - c_5\tau$. Теперь из (6) следует

$$\left\|U^{(k)}\right\|_2 \le (1 - c_5\tau)\left\|U^{(k-1)}\right\|_2 + \tau\left\|V^{(k)}\right\|_2 \le \tau\sum_{v=1}^{k}(1 - c_5\tau)^{k-v}\left\|V^{(v)}\right\|_2 \le$$

$$\le \frac{\tau}{\sqrt{h}}\|\upsilon_h\|\sum_{v=1}^{k}(1 - c_5\tau)^{k-v} \le \frac{\tau}{\sqrt{h}}\|\upsilon_h\|\sum_{v=0}^{\infty}(1 - c_5\tau)^{v} = \frac{1}{c_5}\frac{1}{\sqrt{h}}\|\upsilon_h\|.$$

Дальнейшие рассуждения те же, что при доказательстве теоремы 2.

Ввиду того способа, которым была введена норма в пространстве U_h, очевидны равенства $\left\|A_h^{(2)}\right\| = \left\|A_h^{(3)}\right\| = \left\|A_h^{(4)}\right\| = 1$. Поэтому из доказанной теоремы сразу же вытекает такой результат.

Следствие. Рассматриваемая разностная схема для уравнения теплопроводности устойчива по начальному и граничным условиям.

В этом следствии мы кратко сформулировали утверждение, которое более правильно было бы формулировать так: при применении к задаче (1) рассматриваемой разностной схемы процесс отыскания каркасов приближенных решений устойчив по второму, третьему и четвертому уравнениям, если в пространствах U_h и $F_h^{(s)}$ $(s=2,3,4)$ введены нормы, указанные ранее.

Займемся теперь оценкой нормы $\left\|A_h^{(1)}\right\|$, не очень заботясь о том, чтобы получить хорошую по возможности постоянную. Пусть

$$A_h^{(1)}u_h = f_h,\ u_h = \{u_{jk}\}\ (j=0,...,n;\ k=0,...,m),$$
$$f_h = \{f_{jk}\}\ (j=1,...,n-1;\ k=1,...,m).$$

Тогда f_{jk} выражается формулой (2), т. е. f_{jk} есть линейная комбинация шести компонент вектора u_h с коэффициентами, не превосходящими по абсолютной величине $1/\tau + a/h^2$. Учитывая элементарное неравенство

$$\left(\sum\nolimits_{k=1}^{N} a_k\right)^2 \le N\sum\nolimits_{k=1}^{N} a_k^2,$$

получаем

$$f_{jk}^2 \le 6\left(\frac{1}{\tau}+\frac{a}{h^2}\right)^2\left(u_{j-1k-1}^2+u_{jk-1}^2+u_{j+1k-1}^2+u_{j-1k}^2+u_{jk}^2+u_{j+1k}^2\right).$$

Просуммируем такие неравенства по j от 1 до $n-1$:

$$\sum\nolimits_{j=1}^{n-1} f_{jk}^2 \le 6\left(\frac{1}{\tau}+\frac{a}{h^2}\right)^2\left[3\sum\nolimits_{j=1}^{n-1}u_{jk-1}^2+3\sum\nolimits_{j=1}^{n-1}u_{jk}^2+u_{0k-1}^2+u_{nk-1}^2+u_{0k}^2+u_{nk}^2\right].$$

Поскольку

$$h\sum\nolimits_{j=1}^{n-1}u_{jl}^2 \le \left\|u_h\right\|^2,\quad |u_{0l}| \le \left\|u_h\right\|,\quad |u_{nl}| \le \left\|u_h\right\| \quad (l=k-1,k),$$

то

$$h\sum\nolimits_{j=1}^{n-1} f_{jk}^2 \le 60\left(\frac{1}{\tau}+\frac{a}{h^2}\right)^2\left\|u_h\right\|^2$$

при всех k, и потому

$$\|f_h\| \le \sqrt{60}\left(\frac{1}{\tau}+\frac{a}{h^2}\right)\|u_h\|, \quad \left\|A_h^{(1)}\right\| \le \sqrt{60}\left(\frac{1}{\tau}+\frac{a}{h^2}\right).$$

Таким образом,

$$\mu^{(1)}(A_h) \le c\left(\frac{1}{\tau}+\frac{a}{h^2}\right) \qquad (c=\sqrt{60}T),$$

и по дифференциальному уравнению рассматриваемая разностная схема имеет степенную (слабую) неустойчивость, привычную для метода сеток.

Из доказанных теорем 1 и 2 и *теоремы 1 из § 6* непосредственно вытекает следующий результат о сходимости рассматриваемой разностной схемы.

Теорема 3

Если решение u^* рассматриваемой задачи (1) четырежды непрерывно дифференцируемо, то изучаемый вариант метода сеток сходится в том смысле, что

$$\left\|\varphi_h u^* - u_h^*\right\|_{U_h} = O(h^2+\tau^2). \tag{9}$$

- **Замечание 1.** Сходимость метода без оценки быстроты сходимости может быть установлена при меньших предположениях о гладкости точного решения, чем в доказанной теореме. Действительно, если точное решение u^* дважды непрерывно дифференцируемо по x и один раз по t, то, как легко видеть, $\gamma_h^{(1)}(u^*) \to 0$, и из этой же *теоремы 1 (§ 6)* вытекает, что $\left\|\varphi_h u^* - u_h^*\right\| \to 0$.

- **Замечание 2.** Из теоремы 3 и доказанной ранее леммы сразу же следует, что

$$\left\|\varphi_h u^* - u_h^*\right\|_\infty = \max_{jk}\left|u^*(x_j,t_k)-u_{jk}\right| \le \frac{1}{\sqrt{h}}\left\|\varphi_h u^* - u_h^*\right\|_{U_h} = O(h\sqrt{h}+\tau^2/\sqrt{h}), \tag{10}$$

и если закон зависимости шагов таков, что $\tau^2/\sqrt{h} \to 0$, то имеет место и равномерная сходимость приближенных решений к точному.

Отметим еще, что в правую часть оценки (9) шаги h и τ входят в одинаковых степенях, поэтому естественно выбирать закон соответствия шагов таким, что τ есть бесконечно малая того же порядка, что и h: $c'h \le \tau \le c''h \quad (0 < c' \le c'')$. Тогда правую часть (9) можно заменить на $O(h^2)$, а правую часть (10) — на $O(h\sqrt{h})$.

В заключение параграфа проведем некоторое сопоставление явной и неявной разностных схем для уравнения теплопроводности. При одном и том же шаге h явная схема менее трудоемка на каждом шаге по времени t. Но в явной схеме величину шага τ следует выбирать из условия $\tau \le \frac{1}{2} a_1 h^2$, т. е. очень маленькой, а в неявной схеме целесообразно выбирать τ того же порядка, что и h. Поэтому число шагов по времени, которое нам придется сделать по явной схеме для получения решения на интересующем нас промежутке $[0, T]$, обычно несравненно больше, чем по неявной схеме. Если учесть еще, что специфика тех систем уравнений, которые нам придется решать на каждом шаге по времени при применении неявной схемы, такова, что для решения каждой из них требуется число арифметических операций, лишь в несколько раз превышающее число неизвестных в системе (это будет показано в параграфе 9), то можно сделать заключение, что неявная разностная схема обычно оказывается в конечном итоге менее трудоемкой, по крайней мере при достаточно малых h.

§ 8. Метод сеток для уравнения колебаний струны

Уравнение колебаний струны

$$A^{(1)}u = \frac{\partial^2 u}{\partial t^2} - a^2 \frac{\partial^2 u}{\partial x^2} = f \quad (0 \le x \le 1, \quad 0 \le t \le T) \tag{1}$$

мы будем рассматривать при начальных условиях

$$A^{(2)}u = u(x, 0) = g^{(0)}(x), \quad A^{(3)}u = u_t'(x, 0) = g^{(1)}(x) \tag{2}$$

и граничных условиях

$$A^{(4)}u = u(0, t) = \alpha(t), \quad A^{(5)}u = u(1, t) = \beta(t). \tag{3}$$

При изложении алгоритма приближенного решения задачи (1)—(3), которым мы будем заниматься в этом параграфе, можно было бы считать коэффициент a в уравнении (1) переменным: $a = a(x, t)$ — это не вызвало бы никаких дополнительных трудностей. Однако исследовать сходимость и устойчивость метода мы будем в предположении, что этот коэффициент постоянен. Поэтому с самого начала положим $a = \text{const}$.

Для построения разностной схемы решения нашей задачи выберем некоторое натуральное n и положим $h = 1/n, \quad x_j = jh \quad (j = 0, ..., n)$, выберем некоторый шаг τ по переменной t и положим $t_0 = -\tau/2, \quad t_k = t_0 + k\tau = (k - 1/2)\tau$ $(k = 0, ..., m,$ где $m = [T/\tau + 1/2])$. Причина, по которой мы взяли t_0 равным $-\tau/2$, а не нулю, будет ясна из способа аппроксимации начальных условий (2). Приближенное решение мы будем искать в узлах (x_j, t_k) $(j = 0, ..., n, k = 0, ..., m)$, за исключением двух из них — (x_0, t_0) и (x_n, t_0), рассматривая u_{jk} как приближения к $u^*(x_j, t_k)$ (u^*, как обычно, — точное решение задачи (1)—(3)). Расположение узлов, в которых мы ищем решение, изображено на рис. 15. Общее число неизвестных в той системе уравнений, которую мы будем строить, есть $(n + 1)(m + 1) - 2 = N$.

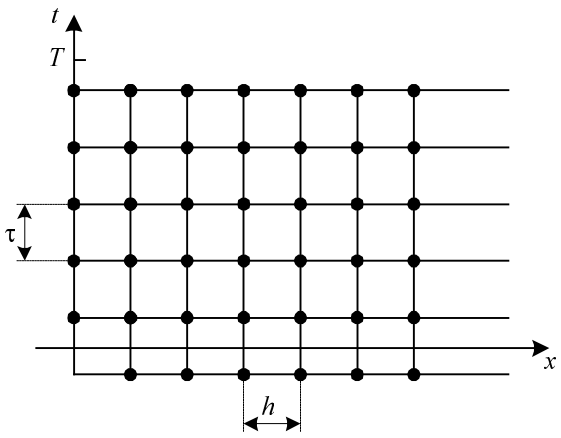


Рис. 15

Уравнение (1) в узле (j,k) $(j=1,...,n-1; k=1,...,m-1)$ мы будем приближать, используя обычные аппроксимации вторых производных. Это даст нам уравнения

$$\frac{u_{jk+1}-2u_{jk}+u_{jk-1}}{\tau^2}-a^2\frac{u_{j+1k}-2u_{jk}+u_{j-1k}}{h^2}=f_{jk} \qquad (4)$$
$$(f_{jk}=f(x_j,t_k)).$$

Шаблон этой разностной схемы изображен на рис. 16. Уравнения (4) мы будем записывать в виде одного уравнения $A_h^{(1)}u_h=f_h^{(1)}$, где $u_h=\{u_{jk}\}$ — искомый N-мерный вектор, $f_h^{(1)}=\{f_{jk}\}-(n-1)(m-1)$-мерный вектор.

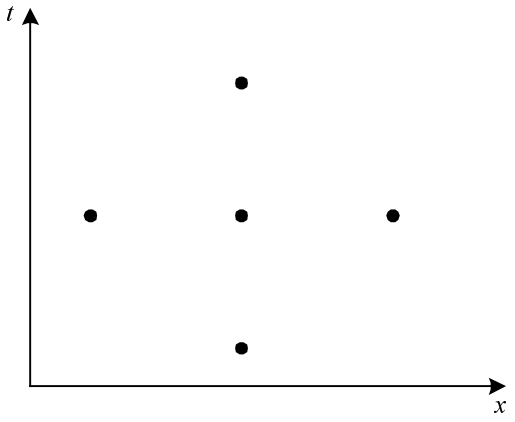


Рис. 16

Если точное решение нашей задачи $u^*(x,t)$ четырежды непрерывно дифференцируемо по x и четырежды непрерывно дифференцируемо по t, то при подстановке в уравнения (4) на место компонент вектора u_h значений этого точного решения в соответствующих узлах все компоненты полученной невязки будут оцениваться через $(1/12)(M_t^{(4)}\tau^2+a^2M_x^{(4)}h^2)$, где

$$M_x^{(4)}=\max_{\Omega}\left|u_{x^4}^{*\mathrm{IV}}\right|,\quad M_t^{(4)}=\max_{\Omega_\delta}\left|u_{t^4}^{*\mathrm{IV}}\right|,$$

Ω — прямоугольник $0\le x\le 1,\ 0\le t\le T$, а Ω_δ — прямоугольник $0\le x\le 1,\ -\delta\le t\le T$, причем предполагается, что $\tau/2\le\delta$. Появление пря-

моугольника Ω_δ связано с тем, что часть узлов нашей сетки (узлы $(j, 0)$) выходит за пределы прямоугольника Ω.

Мы считаем также, что точное решение продолжено в прямоугольник Ω_δ достаточно гладким образом.

В предположении, что решение u^* дважды непрерывно дифференцируемо по t, справедлива формула

$$u^*(x, 0) = \frac{1}{2}\left(u^*(x, \frac{\tau}{2}) + u^*(x - \frac{\tau}{2}) \right) - \frac{\tau^2}{8} u^{*''}_{t^2}(x, \zeta)$$

$(\zeta = \zeta(x) \in (-\tau/2, \tau/2))$. Здесь мы выразили $u^*(x, 0)$ через линейный интерполяционный полином этой функции (как функции от t), построенный по узлам $-\tau/2, \tau/2,$ с остаточным членом в форме Лагранжа. Вышеприведенное равенство дает нам основание написать систему уравнений $A_h^{(2)} u_h = g_h^{(0)}$, аппроксимирующую первое из начальных условий (2), в виде

$$\frac{1}{2}(u_{j1} + u_{j0}) = g_j^0, \quad g_j^0 = g^{(0)}(x_j), \quad j = 1, 2, ..., n-1. \tag{5}$$

Невязки, которые образуются при подстановке в эти уравнения значений точного решения, оцениваются через $(1/8) M_t^{(2)} \tau^2$, где $M_t^{(2)} = \max_{\Omega_\delta} \left| u^{*''}_{t^2} \right|$.

Второе из начальных условий (2) аппроксимируется системой уравнений $A_h^{(3)} u_h = g_h^{(1)}$, при составлении которой применяется обычное представление первой производной через симметричную разделенную разность:

$$\frac{u_{j1} - u_{j0}}{\tau} g_j^1, \quad g_j^1 = g^{(1)}(x_j), \quad j = 1, 2, ..., n-1. \tag{6}$$

Невязки значений точного решения для этих уравнений оцениваются через

$$(1/24) M_t^{(3)} \tau^2, \quad M_t^{(3)} = \max_{\Omega_\delta} \left| u^{*''}_{t^3} \right|.$$

Наконец, аппроксимации $A_h^{(4)} u_h = \alpha_h$, $A_h^{(5)} u_h = \beta_h$ граничных условий (3) получаются тривиальным образом:

$$u_{0k} = \alpha_k = \alpha(t_k), \quad u_{nk} = \beta_k = \beta(t_k), \quad k = 1, 2, ..., m. \tag{7}$$

Общее число уравнений в системе (4)—(7) $(n-1)(m-1)+2(n-1)+2m=N$ совпадает с числом неизвестных.

Построенная разностная схема — явная. Действительно, значения u_{0k} и u_{nk} нам известны из уравнений (7). Значения u_{j0} и u_{j1} легко находятся из уравнений (5) и (6):

$$u_{j0}=g_j^0-\frac{\tau}{2}g_j^1,\quad u_{j1}=g_j^0+\frac{\tau}{2}g_j^1.$$

Все остальные неизвестные u_{jk} можно найти последовательно, придавая j значения $2, 3, ..., m,$ из рекуррентных формул:

$$u_{jk}=2u_{jk-1}-u_{jk-2}+\frac{\alpha^2\tau^2}{h^2}(u_{j+1k-1}-2u_{jk-1}+u_{j-1k-1})+\tau^2 f_{jk-1}, \qquad (8)$$

которые получаются как следствия (4).

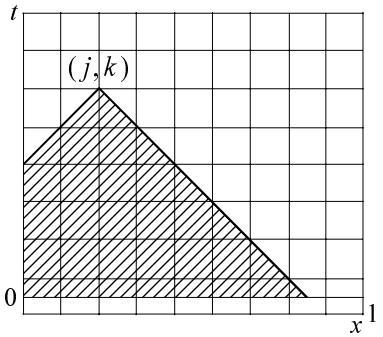


Рис. 17

Из формул (8) видно, что для вычисления значения u_{jk} мы используем в конечном счете значения функций $f, g^{(0)}, g^{(1)}, \alpha$ и β лишь в точках, лежащих ниже прямых $(t-t_k)/\tau=\pm(x-x_j)/h.$ Эта область заштрихована на рис. 17. В то же время значение точного решения u^* в точке (x_j, t_k) зависит от значений этих исходных данных в точках, лежащих ниже характеристик

$t - t_k = \pm(1/a)(x - x_j)$. Если $\tau/h > 1/a$, т. е. $a\tau/h > 1$, то это означает, что для построения u_{jk} мы использовали не все исходные данные, от которых зависит значение $u^*(x_j, t_k)$, и ожидать близости u_{jk} к $u^*(x_j, t_k)$ не приходится. Поэтому на закон соответствия шагов $\tau = \tau(h)$ естественно наложить ограничение $a\tau/h \le 1$. При исследовании нашей разностной схемы мы заменим это условие на несколько более жесткое.

Переходя к исследованию, введем прежде всего нормы в векторных пространствах $F_h^{(s)}$ и U_h. Эти нормы вполне аналогичны тем, которые вводились в предыдущем параграфе, и поэтому пояснять их мы не будем. Для вектора

$$f_h^{(1)} \in F_h^{(1)}, \quad f_h^{(1)} = \{f_{jk}\}, \quad j = 1, ..., n-1; \quad k = 1, ..., m-1,$$

положим

$$\left\| f_h^{(1)} \right\|_{F_h^{(1)}} = \max_{1 \le k \le m-1} \sqrt{h \sum_{j=1}^{n-1} f_{jk}^2}.$$

Для векторов

$$g_h^{(0)} \in F_h^{(2)} \text{ и } g_h^{(1)} \in F_h^{(3)}, \quad g_h^{(0)} = (g_1^0, ..., g_{h-1}^{(0)}), \quad g_h^{(1)} = (g_1^1, ..., g_{h-1}^1)$$

$$\left\| g_h^{(0)} \right\|_{F_h^{(2)}} = \sqrt{h \sum_{j=1}^{n-1} (g_j^0)^2} = \sqrt{h} \left\| g_h^{(0)} \right\|_2,$$

$$\left\| g_h^{(1)} \right\|_{F_h^{(3)}} = \sqrt{h \sum_{j=1}^{n-1} (g_j^1)^2} = \sqrt{h} \left\| g_h^{(1)} \right\|_2.$$

Для векторов $\alpha_h \in F_h^{(4)}$ и $\beta_h \in F_h^{(5)}$, $\alpha_h = (\alpha_1, ..., \alpha_m)$, $\beta_h = (\beta_1, ..., \beta_m)$

$$\left\| \alpha_h \right\|_{F_h^{(4)}} = \max |\alpha_k| = \left\| \alpha_h \right\|_\infty, \quad \left\| \beta_h \right\|_{F_h^{(5)}} = \max |\beta_k| = \left\| \beta_h \right\|_\infty.$$

Наконец, для вектора $u_h \in U_h$, $u_h = \{u_{jk}\}$, $j = 0, ..., n$, $k = 0, ..., m$,

$$\left\| u_h \right\|_{U_h} = \max \left\{ \max_{1 \le k \le m} |u_{0k}|, \max_{1 \le k \le m} |u_{nk}|, \max_{0 \le k \le m} \sqrt{h \sum_{j=1}^{n-1} u_{jk}^2} \right\}.$$

При введенных нормах нетрудно оценить меры аппроксимации на точном решении.

Теорема 1

Если точное решение $u^*(x,t)$ задачи (1)—(3) четырежды непрерывно дифференцируемо в прямоугольнике Ω_δ и $\tau \le 2\delta$, то справедливы оценки

$$\left.\begin{gathered}\gamma_h^{(1)}(u^*) \le c_1(h^2+\tau^2), \quad \gamma_h^{(2)}(u^*) \le c_2\tau^2, \quad \gamma_h^{(3)}(u^*) \le c_3\tau^3, \\ \gamma_h^{(4)}(u^*) = \gamma_h^{(5)}(u^*) = 0.\end{gathered}\right\} \tag{9}$$

Доказательство

Для любого вектора $f_h^{(1)} \in F_h^{(1)}$ очевидным образом выполняется оценка $\left\|f_h^{(1)}\right\|_{F_h^{(1)}} \le \left\|f_h^{(1)}\right\|_\infty = \max\left|f_{jk}\right|$ *(см. также лемму в § 7)*. Мера аппроксимации $\gamma_h^{(1)}(u^*)$ есть норма (в $F_h^{(1)}$) вектора невязок, образующихся при подстановке в уравнения (4) значений точного решения. Как уже отмечалось, все компоненты этого вектора по абсолютной величине не превосходят $(1/12)(M_t^{(4)}\tau^2 + a^2 M_x^{(4)} h^2)$. Поэтому $\gamma_h^{(1)}(u^*) \le c_1(h^2+\tau^2)$, где $c_1 = (1/12)\max\left\{M_t^{(4)}, a^2 M_x^{(4)}\right\}$. Неравенства $\gamma_h^{(2)}(u^*) \le c_2\tau^2$ при $c_2 = (1/8)M_t^{(2)}$ и $\gamma_h^{(3)}(u^*) \le c_3\tau^2$ при $c_3 = (1/24)M_t^{(3)}$ доказываются совершенно аналогично. Равенства $\gamma_h^{(4)}(u^*) = \gamma_h^{(5)}(u^*) = 0$ следуют из того, что значения точного решения u^* уравнениям (7) удовлетворяют точно. ■

Следующий этап исследования — оценка составляющих $[A_h^{-1}]^{(s)}$ обратного оператора. При этом само существование оператора A_h^{-1} сомнений не вызывает — схема явная, и выше были выписаны формулы, по которым можно найти все компоненты вектора $A_h^{-1} f_h$ $(f_h = (f_h^{(1)}, g_h^{(0)}, g_h^{(1)}, \alpha_h, \beta_h))$. Начнем с доказательства некоторых вспомогательных утверждений.

Определение

Многочленами Чебышева второго рода называются функции, задаваемые формулой

$$Q_k(x) = \frac{\sin(k+1)\theta}{\sin\theta}, \quad \theta = \arccos x \quad (|x| \le 1).$$

Ясно, что

$$Q_0(x) \equiv 1, \quad Q_1(x) = 2\cos\theta = 2x. \tag{10}$$

Из формулы

$$\sin(k+1)\theta + \sin(k-1)\theta = 2\cos\theta \cdot \sin k\theta$$

видно, что функции $Q_k(x)$ связаны рекуррентным соотношением

$$Q_k(x) = 2xQ_{k-1}(x) - Q_{k-2}(x) \quad (k = 2, 3, \ldots). \tag{11}$$

Используя (10) и (11), методом индукции легко доказать, что $Q_k(x)$ есть многочлен степени k.

Лемма 1

Для многочленов Чебышева второго рода справедлива оценка

$$|Q_k(x)| \le k+1 \quad (|x| \le 1).$$

Доказательство сразу же следует из элементарного неравенства: при любом натуральном n $|\sin n\theta| \le n|\sin\theta|$. ■

Лемма 2

Пусть $x \le 1$ и числа $c^{(k)}$ связаны рекуррентной формулой $(k = 1, 2, \ldots)$

$$c^{(k+1)} - 2xc^{(k)} + c^{(k-1)} = b^{(k)}. \tag{12}$$

Тогда при $k = 2, 3, \ldots$

$$c^{(k)} = c^{(1)}Q_{k-1}(x) - c^{(0)}Q_{k-2}(x) + \sum_{v=1}^{k-1} b^{(v)}Q_{k-v-1}(x). \tag{13}$$

Доказательство

Доказательство проводится методом индукции. В качестве базы для индукции мы должны доказать формулу (13) при $k = 2$ и $k = 3$. Пусть $k = 2$. Тогда согласно (12)

$$c^{(2)} = 2xc^{(1)} - c^{(0)} + b^{(1)} = c^{(1)}Q_1(x) - c^{(0)}Q_0(x) + b^{(1)}Q_0(x),$$

т. к. $Q_1(x) = 2x, \quad Q_0(x) = 1,$ и при $k = 2$ формула (13) доказана. Пусть $k = 3$. Тогда

$$c^{(3)} = 2xc^{(2)} - c^{(1)} + b^{(2)} = 2x(c^{(1)}Q_1(x) - c^{(0)}Q_0(x) + b^{(1)}Q_0(x)) -$$
$$- c^{(1)} + b^{(2)} = c^{(1)}(2xQ_1(x) - Q_0(x)) - 2xc^{(0)} + b^{(1)}2x + b^{(2)} =$$
$$= c^{(1)}Q_2(x) - c^{(0)}Q_1(x) + \sum_{v=1}^{2} b^{(v)}Q_{2-v}(x).$$

Здесь были использованы формулы (10) и (11). Покажем теперь возможность индуктивного перехода. Пусть для $k-1$ и k формула (13) уже доказана. Тогда

$$c^{(k+1)} = 2xc^{(k)} - c^{(k-1)} + b^{(k)} = 2x[c^{(1)}Q_{k-1}(x) - c^{(0)}Q_{k-2}(x) +$$
$$+ \sum_{v=1}^{k-1} b^{(v)}Q_{k-v-1}(x)] - [c^{(1)}Q_{k-2}(x) - c^{(0)}Q_{k-3}(x) + \sum_{v=1}^{k-2} b^{(v)}Q_{k-v-2}(x)] +$$
$$+ b^{(k)} = c^{(1)}(2xQ_{k-1}(x) - Q_{k-2}(x)) - c^{(0)}(2xQ_{k-2}(x) - Q_{k-3}(x)) +$$
$$+ \sum_{v=1}^{k-2} b^{(v)}(2xQ_{k-v-1}(x) - Q_{k-v-2}(x)) + 2xb^{(k-1)} + b^{(k)} =$$
$$= c^{(1)}Q_k(x) - c^{(0)}Q_{k-1}(x) + \sum_{v=1}^{k} b^{(v)}Q_{k-v}(x). \ \blacksquare$$

Теорема 2

Если закон соответствия шагов $\tau = \tau(h)$ таков, что

$$\kappa = a^2\tau^2/h^2 \le q < 1, \tag{14}$$

то для составляющих обратного оператора A_h^{-1} справедливы оценки

$$\left\| [A_h^{-1}]^{(1)} \right\| \le T^2, \quad \left\| [A_h^{-1}]^{(2)} \right\| \le 1/\sqrt{1-q}, \quad \left\| [A_h^{-1}]^{(3)} \right\| \le T,$$
$$\left\| [A_h^{-1}]^{(4)} \right\| = \left\| [A_h^{-1}]^{(5)} \right\| \le \max\left\{1, \sqrt{2}aT/\sqrt{h}\right\}.$$

Доказательство

Пусть $u_h = \{u_{jk}\}$ — некоторое решение системы уравнений (4) при произвольных правых частях f_{jk}. Введем $(n-1)$-мерные векторы $U^{(k)} = (u_{1k}, u_{2k}, ..., u_{n-1k})$, составленные из компонент вектора u_h. Домножая ра-

венства (4) на τ^2 и используя обозначение (14), видим, что эти векторы связаны рекуррентной формулой

$$U^{(k+1)} = BU^{(k)} - U^{(k-1)} + V^{(k)}, \quad k = 1, 2, ..., m-1, \tag{15}$$

где

$$B = \begin{pmatrix} 2-2\kappa & \kappa & 0 & \cdots & 0 & 0 \\ \kappa & 2-2\kappa & \kappa & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \kappa & 2-2\kappa \end{pmatrix},$$

а вектор $V^{(k)} = (\upsilon_{1k}, ..., \upsilon_{n-1k})$ имеет компоненты

$$\upsilon_{1k} = \tau^2 f_{1k} + \kappa u_{0k}, \quad \upsilon_{n-1k} = \tau^2 f_{n-1k} + \kappa u_{nk}, \quad \upsilon_{jk} = \tau^2 f_{jk}, \quad j \neq 1, n-1.$$

Собственные числа матрицы B известны *(см. теорему 1 из § 1)*:

$$\lambda_l = 2(1-\kappa) + 2\kappa\cos\frac{l\pi}{n} = 2\left(1 - \kappa(1-\cos\frac{l\pi}{n})\right) = 2\left(1 - 2\kappa\sin^2\frac{l\pi}{2n}\right) = 2\sigma_l,$$

где: $\sigma_l = 1 - 2\kappa\sin^2[l\pi/(2n)]$ и ввиду условия (14) $\sigma_l < 1$. Соответствующие собственным числам λ_l собственные векторы матрицы B суть

$$z_l = \sqrt{\frac{2}{n}}\left(\sin\frac{l\pi}{n}, \sin\frac{2l\pi}{n}, ..., \sin\frac{(n-1)l\pi}{n}\right).$$

Эти векторы ортогональны и нормированы: $(z_{\iota_1}, z_{\iota_2}) = \delta_{\iota_1\iota_2}$. Разложим векторы $U^{(k)}$ и $V^{(k)}$ по векторам z_ι:

$$U^{(k)} = \sum_{l=1}^{n-1} c_l^{(k)} z_l, \qquad V^{(k)} = \sum_{l=1}^{n-1} b_l^{(k)} z_l.$$

Подставляя эти разложения в рекуррентную формулу (15) и приравнивая коэффициенты при векторах z_l в левой и правой частях, получаем

$$c_l^{(k+1)} = 2\sigma_l c_l^{(k)} - c_l^{(k-1)} + b_l^{(k)}, \quad k = 1, 2, ..., m-1, \quad l = 1, 2, ..., n-1.$$

Учитывая, что $|\sigma_l| < 1$, и применяя лемму 2, имеем

$$c_l^{(k)} = c_l^{(1)} Q_{k-1}(\sigma_l) - c_l^{(0)} Q_{k-2}(\sigma_l) + \sum_{v=1}^{k-1} b_l^{(v)} Q_{k-v-1}(\sigma_l). \tag{16}$$

После этих предварительных замечаний перейдем непосредственно к оценке составляющих оператора A_h^{-1}.

1. Оценим $[A_h^{-1}]$. Вектор $u_h = \{u_{jk}\} = [A_h^{-1}] f_h^{(1)}$ есть решение системы уравнений

$$A_h^{(1)} u_h = f_h^{(1)}, \quad A_h^{(s)} u_h = 0 \quad \text{при} \quad s = 2, 3, 4, 5.$$

Пусть $U^{(k)}$ — векторы, введенные ранее. Из равенств $A_h^{(2)} u_h = 0, \quad A_h^{(3)} u_h = 0$ следует, что $U^{(0)} = U^{(1)} = 0,$ и, в частности, $c_l^{(0)} = c_l^{(1)} = 0 \quad (l = 1, ..., n-1).$ Ввиду равенств $A_h^{(4)} u_h = A_h^{(5)} u_h = 0$ будет $u_{0k} = u_{nk} = 0,$ и потому при всех $j = 1, 2, ..., n-1 \quad \upsilon_{jk} = \tau^2 f_{jk},$ а также

$$\|u_h\| = \max_{2 \le k \le m} \sqrt{h \sum_{j=1}^{n-1} u_{jk}^2} = \sqrt{h} \max_{2 \le k \le m} \left\| U^{(k)} \right\|_2.$$

Формула (16) принимает в нашем случае вид

$$c_l^{(k)} = \sum_{v=1}^{k-1} b_l^{(v)} Q_{k-v-1}(\sigma_l).$$

По лемме 1 $\left| Q_{k-v-1}(\sigma_l) \right| \le k - v \le k - 1,$ откуда

$$\left| c_l^{(k)} \right| \le (k-1) \sum_{v=1}^{k-1} \left| b_l^{(v)} \right|.$$

Поскольку

$$\left(\sum_{v=1}^{k-1} \left| b_l^{(v)} \right| \right)^2 \le (k-1) \sum_{v=1}^{k-1} (b_l^{(v)})^2,$$

то

$$(c_l^{(k)})^2 \le (k-1)^3 \sum_{v=1}^{k-1} (b_l^{(v)})^2,$$

$$\left\| U^{(k)} \right\|_2^2 = \sum_{l=1}^{n-1} (c_l^{(k)})^2 \le (k-1)^3 \sum_{v=1}^{k-1} \sum_{l=1}^{n-1} (b_l^{(v)})^2 = (k-1)^3 \sum_{v=1}^{k-1} \left\| V^{(v)} \right\|_2^2.$$

Здесь

$$\left\| V^{(v)} \right\|_2^2 = \sum_{j=1}^{k-1} \upsilon_{jv}^2 = \tau^4 \sum_{j=1}^{n-1} f_{jv}^2 \le (\tau^4 / h) \left\| f_h^{(1)} \right\|^2.$$

Поэтому

$$\left\| U^{(k)} \right\|_2^2 \le (1/h)(k-1)^4 \tau^4 \left\| f_h^{(1)} \right\|^2.$$

Учитывая, что $(k-1)\tau \le T$, отсюда имеем

$$\left\| U^{(k)} \right\|_2 \le (1/\sqrt{h}) T^2 \left\| f_h^{(1)} \right\|, \quad \| u_h \| \le T^2 \left\| f_h^{(1)} \right\|.$$

Ввиду произвольности вектора $f_h^{(1)}$ $\left\| [A_h^{-1}]^{(1)} \right\| \le T^2$, и первое из доказываемых неравенств получено.

2. Оценим $\left\| [A_h^{-1}]^{(2)} \right\|$. Пусть u_h — решение системы уравнений

$$A_h^{(2)} u_h = g_h^{(0)}, \quad A_h^{(s)} u_h = 0 \quad \text{при} \quad s = 1, 3, 4, 5.$$

Тогда векторы $U^{(k)}$, построенные для такого u_h, обладают свойствами:

а) $U^{(0)} = U^{(1)} = g_h^{(0)}$;

б) $\| u_h \| = \sqrt{h} \max\limits_{0 \le k \le m} \left\| U^{(k)} \right\|_2$;

в) векторы $U^{(k)}$ связаны рекуррентной формулой (15) при $V^{(k)} = 0$, в частности,

$$c_l^{(k)} = c_l^{(1)} Q_{k-1}(\sigma_l) - c_l^{(0)} Q_{k-2}(\sigma_l).$$

Здесь $c_l^{(1)} = c_l^{(0)} = c_l$ — коэффициенты разложения вектора $g_h^{(0)}$ по векторам z_l:

$$g_h^{(0)} = \sum\nolimits_{l=1}^{n-1} c_l z_l.$$

Итак,

$$c_l^{(k)} = c_l (Q_{k-1}(\sigma_l) - Q_{k-2}(\sigma_l)).$$

Найдем такой угол $\theta_l \in (0, \pi)$, что $\sigma_l = \cos\theta_l$. Тогда

$$\left| Q_{k-1}(\sigma_l) - Q_{k-2}(\sigma_l) \right| = \left| \frac{\sin k\theta_l - \sin(k-1)\theta_l}{\sin\theta_l} \right| =$$

$$= \left| \frac{2\sin\frac{\theta_l}{2} \cdot \cos(k - \frac{1}{2})\theta_l}{2\sin\frac{\theta_l}{2} \cdot \cos\frac{\theta_l}{2}} \right| = \left| \frac{\cos(k - \frac{1}{2})\theta_l}{\cos\frac{\theta_l}{2}} \right| \le \frac{1}{\cos\frac{\theta_l}{2}}.$$

Но

$$\cos\frac{\theta_l}{2}=\sqrt{\frac{1+\cos\theta_l}{2}}=\sqrt{\frac{1+\sigma_l}{2}}=\sqrt{1-\kappa\sin^2\frac{l\pi}{2n}}\ge\sqrt{1-\kappa}\ge\sqrt{1-q},$$

и потому

$$\left|c_l^{(k)}\right|\le\frac{1}{\sqrt{1-q}}|c_l|,$$

$$\left\|U^{(k)}\right\|_2^2=\sum_{l=1}^{n-1}(c_l^{(k)})^2\le\frac{1}{1-q}\sum_{l=1}^{n-1}c_l^2=\frac{1}{1-q}\left\|g_n^{(0)}\right\|_2^2,$$

$$\left\|u_h\right\|=\sqrt{h}\max\left\|U^{(k)}\right\|_2\le\frac{1}{\sqrt{1-q}}\sqrt{h}\left\|g_n^{(0)}\right\|_2=\frac{1}{\sqrt{1-q}}\left\|g_n^{(0)}\right\|,$$

и ввиду произвольности вектора $g_h^{(0)}$ $\left\|[A_h^{-1}]^{(2)}\right\|\le 1/\sqrt{1-q}$.

3. Оценка $\left\|[A_h^{-1}]^3\right\|$. Если u_h есть решение системы

$$A_h^{(3)}u_h=g_h^{(1)},\quad A_h^{(s)}u_h=0\quad\text{при}\quad s=1,2,4,5,$$

то:

а) $U^{(0)}=-(\tau/2)g_h^{(1)},\quad U^{(1)}=(\tau/2)g_h^{(1)}$;

б) $\left\|u_h\right\|=\sqrt{h}\max\limits_{0\le k\le m}\left\|U^{(k)}\right\|_2$;

в) в формуле (15) $V^{(k)}=0$ и

$$c_l^{(k)}=c_l^{(1)}Q_{k-1}(\sigma_l)-c_l^{(0)}Q_{k-2}(\sigma_l)=c_l(Q_{k-1}(\sigma_l)+(Q_{k-2}(\sigma_l)),$$

где $c_l=(\tau/2)d_l$; d_l — коэффициенты разложения вектора $g_h^{(1)}$ по векторам z_l:

$$g_h^{(1)}=\sum_{\iota=1}^{n-1}d_l z_l.$$

Применяя лемму 1, имеем

$$\left|c_\iota^{(k)}\right|\le|c_l|(2k-1)=(\tau/2)(2k-1)|d_l|\le T|d_l|.$$

Здесь мы учли, что $\tau(k-1/2) \le \tau(m-1/2) \le T$. Отсюда

$$\left\| U^{(k)} \right\|_2 \le T \left\| g_h^{(1)} \right\|_2, \quad \| u_h \| \le T \left\| g_h^{(1)} \right\|, \quad \left\| [A_h^{-1}]^{(3)} \right\| \le T.$$

4. Нормы $\left\| [A_h^{-1}]^{(4)} \right\|$ и $\left\| [A_h^{-1}]^{(5)} \right\|$ оцениваются одинаково и из соображений симметрии должны быть равны между собой. Оценим $\left\| [A_h^{-1}]^{(4)} \right\|$. Пусть u_h есть решение системы уравнений

$$A_h^{(4)} u_h = \alpha_h, \quad A_h^{(s)} u_h = 0 \quad \text{при} \quad s = 1, 2, 3, 5.$$

Тогда

а) $U^{(0)} = U^{(1)} = 0$;

б) $\| u_h \| = \max \left\{ \| \alpha_h \|, \sqrt{h} \max_{2 \le k \le m} \left\| U^{(k)} \right\|_2 \right\}$;

в) в рекуррентной формуле (15) $V^{(k)} = (\kappa \alpha_k, 0, 0, ..., 0)$. Тем самым

$$c_l^{(0)} = c_l^{(1)} = 0, \quad b_l^{(k)} = (V^{(k)}, z_l) = \kappa \alpha_k \sqrt{2/n} \sin(l\pi/n)$$

и

$$c_l^{(k)} = \sqrt{\frac{2}{n}} \kappa \sin \frac{l\pi}{n} \sum_{v=1}^{k-1} \alpha_v Q_{k-v-1}(\sigma_l).$$

Полагая, как и при оценке $\left\| [A_h^{-1}]^{(2)} \right\|$, $\sigma_l = \cos\theta_l$ и учитывая, что

$$\left| Q_{k-v-1}(\sigma_l) \right| \le 1/\sin\theta_l, \quad \left| \alpha_k \right| \le \| \alpha_h \| \quad k-1 \le m-1,$$

получаем

$$\left| c_l^{(k)} \right| \le \sqrt{\frac{2}{n}} \kappa \sin \frac{l\pi}{n} (m-1) \frac{1}{\sin\theta_l} \| \alpha_h \|.$$

Отсюда

$$\left\| U^{(k)} \right\|_2^2 \le \frac{2}{n} \kappa (m-1)^2 \sum_{l=1}^{n-1} \frac{\kappa \sin^2(l\pi/n)}{\sin^2\theta_l} \cdot \| \alpha_h \|^2.$$

Оценим последнюю сумму:

$$\sum_{l=1}^{n-1}\frac{\kappa\sin^2(l\pi/n)}{\sin^2\theta_l}=\sum_{l=1}^{n-1}\frac{\kappa\sin^2(l\pi/n)}{1-\sigma_l^2}=$$

$$=\sum_{l=1}^{n-1}\frac{4\kappa\sin^2[l\pi/(2n)]\cdot\cos^2[l\pi/(2n)]}{4\kappa\sin^2[l\pi/(2n)]-4\kappa\sin^4[l\pi/(2n)]}=$$

$$=\sum_{l=1}^{n-1}\frac{1-\sin^2[l\pi/(2n)]}{1-\kappa\sin^2[l\pi/(2n)]}\le n-1<n.$$

Итак,

$$\left\|U^{(k)}\right\|_2\le\sqrt{2}\sqrt{\kappa}(m-1)\left\|\alpha_h\right\|=\sqrt{2}\,\frac{a\tau}{h}(m-1)\left\|\alpha_h\right\|\le\frac{\sqrt{2}Ta}{h}\left\|\alpha_h\right\|,$$

$$\left\|u_h\right\|\le\max\left(1,\frac{\sqrt{2}Ta}{\sqrt{h}}\right)\left\|\alpha_h\right\|,\quad\left\|[A_h^{-1}]^{(4)}\right\|\le\max\left(1,\frac{\sqrt{2}Ta}{\sqrt{h}}\right).\ \blacksquare$$

❑ **Замечание.** Как уже было показано из сравнения областей зависимости от исходных данных точного и приближенного решений, требование $\kappa\le 1$ является существенным для рассматриваемого метода. Если заменить более жесткое условие (14) на условие $\kappa\le 1$ и, в частности, разрешить к принимать значения, равные единице, то, как нетрудно убедиться, просматривая доказательство теоремы, оценки норм $\left\|[A_h^{-1}]^{(s)}\right\|$ при $s=1,3,4,5$ останутся в силе, а для $\left\|[A_h^{-1}]^{(2)}\right\|$ придется дать другую оценку.

Видоизменим п. 2 доказательства следующим образом. Как и раньше,

$$\left|c_l^{(k)}\right|\le\left|c_l\right|/\cos(\theta_l/2),$$

где

$$\cos\frac{\theta_l}{2}=\sqrt{1-\kappa\sin^2\frac{l\pi}{2n}}\ge\sqrt{1-\sin^2\frac{l\pi}{2n}}=\cos\frac{l\pi}{2n}\ge\cos\frac{(n-1)\pi}{2n}=\sin\frac{\pi}{2n}\ge\frac{1}{n}.$$

Отсюда

$$\left|c_l^{(k)}\right|\le\left|c_l\right|\cdot n=\left|c_l\right|/h\ \text{ и }\ \left\|u_h\right\|\le(1/h)\left\|g_h^{(0)}\right\|.$$

Итак, при более слабом, чем (14), условии $\kappa \le 1$ выполняется оценка

$$\left\| [A_h^{-1}]^{(2)} \right\| \le 1/h.$$

Точно такими же выкладками, как в § 7, можно показать, что

$$\left\| A_h^{(1)} \right\| \le c(1/\tau^2 + 1/h^2),$$

где c — некоторая постоянная. Легко видеть, что

$$\left\| A_h^{(2)} \right\| = 1, \quad \left\| A_h^{(3)} \right\| = 2/\tau \text{ и } \left\| A_h^{(4)} \right\| = \left\| A_h^{(5)} \right\| = 1.$$

Из этих оценок и теоремы 2 сразу же вытекает утверждение об устойчивости.

Следствие. Если для закона соответствия шагов выполняется условие (14), то при выбранных нормах в пространствах U_h и $F_h^{(s)}$ процесс отыскания каркасов приближенных решений устойчив по первому начальному условию и имеет степенную неустойчивость по дифференциальному уравнению, второму начальному условию и граничным условиям, причем

$$\mu^{(1)}(A_h) \le c_1(1/\tau^2 + 1/h^2),$$

$$\mu^{(3)}(A_h) \le c_3/\tau,$$

$$\mu^{(4)}(A_h) = \mu^{(5)}(A_h) \le c_4/\sqrt{h}.$$

Если условие (14) заменить более слабым $\kappa \le 1$, то процесс отыскания каркасов и по первому начальному условию имеет степенную неустойчивость, причем

$$\mu^{(2)}(A_h) \le c_2/h.$$

❑ **Замечание.** Поскольку $\|A_h\| = \max \left\| A_h^{(s)} \right\| \le c_5(1/\tau^2 + 1/h^2)$ (как и в оценке для $\mu^{(1)}(A_h)$, можно было бы, впрочем, $1/\tau^2 + 1/h^2$ заменить на $1/\tau^2$, т. к. $1/h^2$ оценивается в силу (14) через $1/\tau^2$) и $\left\| A_h^{-1} \right\| \le \sum_s \left\| [A_h^{-1}]^{(s)} \right\| \le$

$\le c_6/\sqrt{h}$, то для числа обусловленности оператора A_h мы получаем худшую оценку, чем для всех частных чисел обусловленности:

$$\mu(A_h) \le c_7 \frac{1}{\sqrt{h}}\left(\frac{1}{\tau^2}+\frac{1}{h^2}\right) \le c_8 \frac{1}{\sqrt{h}\tau^2}.$$

Теперь нетрудно доказать сходимость рассматриваемой разностной схемы.

Теорема 3

Пусть решение u^* задачи (1)—(3) четырежды непрерывно дифференцируемо по x и по t в прямоугольнике Ω_δ. Пусть для закона соответствия шагов $\tau = \tau(h)$ выполнено условие (14). Тогда каркасы приближенных решений сходятся в отношении выбранных нами в пространствах U_h норм, причем

$$\left\| \varphi_h u^* - u_h^* \right\|_{U_h} = O(h^2 + \tau^2). \qquad (17)$$

Доказательство

Согласно *теореме 1 из § 6*

$$\left\| \varphi_h u^* - u_h^* \right\|_{U_h} \le \sum_{s=1}^{5} \left\| [A_h^{-1}]^{(s)} \right\| \gamma_h^{(s)}(u^*) = \sum_{s=1}^{3} \left\| [A_h^{-1}]^{(s)} \right\| \gamma_h^{(s)}(u^*).$$

В последнем равенстве учтено, что $\gamma_h^{(4)}(u^*) = \gamma_h^{(s)}(u^*)$. Остается оценить оставшиеся слагаемые при помощи теорем 1 и 2. ■

- **Замечание 1.** Ввиду условия (14) в правой части (17) можно вместо $O(h^2+\tau^2)$ писать $O(h^2)$.

- **Замечание 2.** Сам факт сходимости (стремление к нулю $\left\| \varphi_h u^* - u_h^* \right\|_{U_h}$) без оценки быстроты сходимости может быть получен, как обычно, при меньших предположениях о гладкости точного решения. Именно достаточно предполагать, что u^* дважды непрерывно дифференцируемо по x и по t.

❑ **Замечание 3.** Сходимость каркасов приближенных решений в метрике U_h означает, что

$$\max_k \sqrt{h\sum_{j=1}^{n-1}(u^*(x_j,t_k)-u_{jk})^2} \to 0,$$

т. е. эту сходимость можно охарактеризовать как "равномерную по t и в среднем по x". Так же, как и в *§ 7*, легко установить, что в условиях теоремы 3 имеет место и "равномерная" сходимость, причем

$$\max_{j,k}\left|u^*(x_j,t_k)-u_{jk}\right| = O(h\sqrt{h}).$$

§ 9. Метод прогонки

При применении метода сеток к дифференциальным уравнениям приходится решать системы линейных алгебраических уравнений высокого порядка. Эти системы обладают некоторыми специфическими свойствами, и для них естественно разрабатывать свои способы решения. В этом параграфе пойдет речь об одном способе решения сеточных уравнений, которые возникают при применении метода сеток к обыкновенным дифференциальным уравнениям и при применении неявной разностной схемы к уравнению теплопроводности. В последнем случае речь идет о тех системах уравнений, которые приходится решать на каждом шаге по времени. Особенность конечноразностных систем в упомянутых случаях состоит в том, что их матрицы коэффициентов трехдиагональны, т. е. имеют вид

$$\begin{pmatrix} b_1 & c_1 & 0 & \cdots & 0 & 0 \\ a_2 & b_2 & c_2 & \cdots & 0 & 0 \\ 0 & a_3 & b_3 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & a_n & b_n \end{pmatrix}$$

Таким образом, если положить $a_1 = 0,\quad c_n = 0,$ то сама система уравнений выглядит так:

$$a_k x_{k-1} + b_k x_k + c_k x_{k+1} = d_k, \quad k = 1, 2, ..., n. \tag{1}$$

Метод, рассматриваемый в этом параграфе, есть метод последовательного исключения неизвестных. Возможно большое число вариантов метода последовательного исключения неизвестных, отличающихся порядком их исключения. Наиболее естественными для системы (1) представляются два из них, сохраняющие на каждом шаге трехдиагональный характер оставшейся системы. Первый порядок таков — с помощью первого уравнения мы исключаем сначала x_2, затем с помощью второго — x_3 и т. д. В результате все неизвестные оказываются выраженными через x_1, которое находится в конце из последнего уравнения. Однако специфика рассматриваемых систем, возникающих в методе сеток, такова, что эта последовательность исключения приводит часто к существенной потере точности. Поясним это следующим примером.

Пример

Пусть мы решаем методом сеток следующую граничную задачу для обыкновенного дифференциального уравнения:

$$-u''+d^2u=1, \quad u(0)=u(1)=0. \tag{2}$$

Коэффициент d мы считаем постоянным и большим. Точное решение этой задачи есть

$$u^*(x)=\frac{1}{d^2\operatorname{sh}d}\bigl(\operatorname{sh}d-\operatorname{sh}(dx)-\operatorname{sh}(d(1-x))\bigr).$$

Применение к задаче (2) метода сеток[1] с шагом $h=1/n$ приводит к системе уравнений

$$\left.\begin{aligned} \frac{2u_1-u_2}{h^2}+d^2u_1&=1,\\ \frac{-u_{k-1}+2u_k-u_{k+1}}{h^2}+d^2u_k&=1 \quad k=2,3,\ldots,n-2,\\ \frac{-u_{n-2}+2u_{n-1}}{h^2}+d^2u_{n-1}&=1. \end{aligned}\right\}$$

[1] Мы не занимались специально исследованием метода сеток для обыкновенных дифференциальных уравнений. Однако наша задача есть одномерный аналог той, которая рассматривалась для эллиптического дифференциального уравнения в *§ 4*, так что и здесь метод сеток сходится в равномерной метрике с порядком h^2.

Для решения этой системы приме́ним метод последовательного исключения неизвестных, выражая все u_k через u_1: $u_k = \alpha_k u_1 + \beta_k$. Тогда $\alpha_1 = 1, \quad \beta_1 = 0,$ из первого уравнения $\alpha_2 = 2 + d^2h^2, \quad \beta_2 = -h^2,$ из k-го уравнения

$$-(\alpha_{k-1}u_1 + \beta_{k-1}) + 2(\alpha_k u_1 + \beta_k) - (\alpha_{k+1}u_1 + \beta_{k+1}) + d^2h^2(\alpha_k u_1 + \beta_k) = h^2,$$

так что

$$\alpha_{k+1} = (2 + d^2h^2)\alpha_k - \alpha_{k-1},$$
$$\beta_{k+1} = (2 + d^2h^2)\beta_k - \beta_{k-1} - h^2.$$

Если обозначить через λ_1 и λ_2 корни характеристического уравнения $\lambda^2 - (2 + d^2h^2)\lambda + 1 = 0$:

$$\lambda_1 = 1 + \frac{d^2h^2}{2} + \sqrt{d^2h^2 + (1/4)d^4h^4}, \quad \lambda_2 = 1 + \frac{d^2h^2}{2} - \sqrt{d^2h^2 + (1/4)d^4h^4},$$

то для α_k и β_k можно получить явные формулы :

$$\alpha_k = \frac{\lambda_1^k - \lambda_2^k}{\lambda_1 - \lambda_2}, \quad \beta_k = \frac{1}{d^2}\left(1 - \frac{1 + d^2h^2 - \lambda_2}{\lambda_1 - \lambda_2}\lambda_1^{k-1} + \frac{1 + d^2h^2 - \lambda_1}{\lambda_1 - \lambda_2}\lambda_2^{k-1}\right).$$

Пусть x — рациональная точка промежутка (0, 1). Тогда при $n \to \infty$ эта точка бесконечное число раз окажется узлом. Дальше мы будем рассматривать лишь такие n, при которых $nx = k$ — целое, так что x есть узел с номером k. Тогда нетрудно получить асимптотические формулы (при $n \to \infty$, т. е. $h \to \infty$)

$$\alpha_k \approx \frac{1}{2dn}\operatorname{sh}(dx), \quad \beta_k \approx -\frac{\operatorname{sh}(dx) - 1}{d^2}.$$

С увеличением d эти числа быстро возрастают по абсолютной величине. Поскольку само решение при больших d мало, то это означает, что при вычислении u_k по формуле $u_k = \alpha_k u_1 + \beta_k$ мы столкнемся с вычитанием двух больших близких чисел, т. е. произойдет пропадание значащих цифр. Так, при $d = 10$ и достаточно малом h сеточное решение в точке $x = 0{,}8$, близкое к $u^*(0{,}8) \approx 0{,}0086,$ мы получим вычитанием двух чисел, близких к $\beta_{0{,}8n} \approx 14{,}85 \quad (\max\left|u^*(x)\right| = u^*(0{,}5) \approx 0{,}0099)$. Отметим, что коэффициенты

α_k не зависят от правой части дифференциального уравнения, так что пропадание значащих цифр можно ожидать не только при $f \equiv 1$, но и при других правых частях уравнения $-u''+d^2u'=f$.

При применении к уравнению теплопроводности неявной разностной схемы, изучавшейся в § 7, на каждом шаге по времени нам придется решать систему линейных уравнений, матрица которой лишь множителем отличается от матрицы системы уравнений метода сеток для задачи (2) при $d^2 = 2/(a\tau)$. Изучение рассматриваемого сейчас способа последовательного исключения неизвестных для неявной разностной схемы осложняется тем, что и коэффициент $d^2 = 2/(a\tau)$, и правые части сеточных уравнений связаны с шагом h. Однако вывод о катастрофическом пропадании значащих цифр при $\tau \to 0$ остается в силе.

Перейдем к описанию другого естественного порядка исключения неизвестных в системе уравнений (1), который обычно используется и называется *методом прогонки*. Выразим из первого уравнения x_1 через x_2:

$$x_1 = d_1/b_1 - (c_1/b_1)x_2. \tag{3}$$

Подставляя это выражение для x_1 во второе уравнение, мы приходим к уравнению, которое содержит лишь x_2 и x_3, и используем его для выражения x_2 через x_3. Вообще, если мы уже выразили x_{k-1} через x_k:

$$x_{k-1} = A_{k-1}x_k + B_{k-1},$$

то, подставив x_{k-1} в такой форме в уравнение с номером k, мы получим

$$(A_{k-1}a_k + b_k)x_k + c_k x_{k+1} = d_k - B_{k-1}a_k,$$

откуда

$$x_k = A_k x_{k+1} + B_k, \tag{4}$$

где

$$D_k = A_{k-1}a_k + b_k, \quad A_k = -c_k/D_k, \quad B_k = (d_k - B_{k-1}a_k)/D_k. \tag{5}$$

Отметим, что поскольку $a_1 = 0$, то формула (3) также может быть переписана в виде (4), (5) при $k = 1$. Так как $c_n = 0$, то $A_n = 0$, и потому $x_n = B_n$. Зная x_n, из формулы (4) при $k = n-1$ мы можем найти x_{n-1}, затем x_{n-2}, и т. д.,

пока не найдем все неизвестные. Итак, процесс решения системы уравнений (1) методом прогонки состоит в следующем:

1) по формулам (5) при $k = 1, 2, ..., n$ последовательно вычисляем коэффициенты D_k, A_k, B_k. Эта часть вычислений называется *прямой прогонкой*;

2) полагаем $x_n = B_n$ и по формулам (4) при $k = n-1, n-2, ..., 1$ последовательно вычисляем неизвестные $x_{n-1}, x_{n-2}, ..., x_1$. Этот процесс называется *обратной прогонкой.*

Метод прогонки для решения системы уравнений (1) применим, если отличны от нуля все числа D_k, на которые приходится делить (до сих пор мы предполагали это условие выполненным). Покажем, что это так при некоторых условиях, наложенных на коэффициенты системы (1).

Теорема 1

Если при $k = 1, 2, ..., n$ выполняются неравенства

$$|b_k| > |a_k| + |c_k|, \tag{6}$$

то при всех k $D_k \neq 0$.

Доказательство

Покажем методом индукции, что при всех k $|A_k| < 1$ и $D_k \neq 0$. При $k = 1$

$$D_1 = b_1 \neq 0, \quad |A_1| = |c_1| / |b_1| < 1.$$

Если эти соотношения уже доказаны для $k-1$, то имеем

$$|D_k| = |A_{k-1} a_k + b_k| \geq |b_k| - |a_k| > |c_k|, \quad D_k \neq 0,$$

$$|A_k| = |c_k| / |D_k| < |c_k| / |c_k| = 1. \ \blacksquare$$

- **Замечание 1.** Теорема об осуществимости метода прогонки дает, разумеется, и достаточный признак однозначной разрешимости системы уравнений (1). Условие (6) для этой системы есть известный признак Адамара неособенности матрицы.
- **Замечание 2.** Полученная попутно при доказательстве теоремы оценка $|A_k| < 1$ оказывается чрезвычайно полезной при исследовании устойчивости метода прогонки. В частности, если значение x_{k+1} вычислено с некоторой погрешностью δ, то при вычислении x_k по формуле

$x_k = A_k x_{k+1} + B_k$ эта погрешность войдет в x_k с коэффициентом, меньшим единицы, т. е. уменьшится. Если наша система есть система уравнений метода сеток, то x_k и x_{k+1} имеют обычно смысл значений приближенного решения в двух близких точках. Поэтому естественно ожидать, что x_k и x_{k+1} близки между собой, и в формуле $x_k = A_k x_{k+1} + B_k$ первое слагаемое в правой части не может быть по модулю существенно больше суммы, так что пропадания значащих цифр при вычислении по этой формуле не должно быть.

❑ **Замечание 3.** Условие (6) теоремы оказывается выполненным в следующих важных случаях.

1. При применении варианта метода сеток, изложенного в *разд. "Метод сеток для обыкновенного дифференциального уравнения" (§ 1)*, к обыкновенному дифференциальному уравнению, если только коэффициенты p и r этого уравнения удовлетворяют неравенствам $p(x) \ge p_0 > 0, \quad r(x) \le r_0 < 0$ [1], а шаг h достаточно мал. Действительно, в системе конечно-разностных уравнений будет $b_k = -2p_k/h^2 + r_k$, $a_k = p_k/h^2 - q_k/h, \quad c_k = p_k/h^2 + q_k h$ и при условии, что $p_k \ge |q_k| h,$ окажется $|b_k| - |a_k| - |c_k| = |r_k| > 0.$

2. При применении к обыкновенному дифференциальному уравнению метода конечных элементов *(см. разд. "Вариационно-разностные схемы", § 1)*, если $r(x) \ge 0$ и шаг h достаточно мал. Коэффициенты a_k, b_k и c_k системы уравнений метода выражаются здесь через интегралы. При достаточно малых h a_k и c_k имеют знаки, противоположные знакам b_k, так что

$$|b_k| - |a_k| - |c_k| = |b_k + a_k + c_k| > 0.$$

3. При применении к уравнению теплопроводности неявной разностной схемы, изучавшейся в *§ 7* (при этом коэффициент a уравнения не предполагается постоянным). В этом случае

$$b_k = 1/\tau + a_{jk-1/2}/h^2, \quad a_k = c_k = -a_{jk-1/2}/(2h^2).$$

---

[1] Эти неравенства гарантируют и однозначную разрешимость рассматриваемой граничной задачи.

Подсчитаем число арифметических операций для решения системы уравнений (1) методом прогонки. Каждый шаг прямой прогонки требует двух умножений, двух делений, одного сложения и одного вычитания, т. е. шести операций. Каждый шаг обратной прогонки требует одного умножения и одного сложения — двух операций. Поэтому общее число арифметических операций для решения системы есть примерно $8n$ и пропорционально, таким образом, порядку системы n, а не n^3, как это бывает при применении метода последовательных исключений к системам уравнений общего вида.

Рассмотрение метода прогонки позволяет нам остановиться еще на одном чрезвычайно важном круге вопросов, связанных с методом сеток, а отчасти и с другими приближенными методами, который не нашел достаточного отражения в нашей книге. Наличие эффективного способа решения возникающей системы линейных алгебраических уравнений — одна из важнейших характеристик того или иного варианта метода сеток наряду с устойчивостью и порядком аппроксимации. В сущности, в этом направлении перед исследователями встают два класса задач — создание наиболее эффективных методов решения систем уравнений для уже существующих вариантов метода сеток и создание новых вариантов метода сеток, системы линейных алгебраических уравнений которых допускают удобные способы решения.

Введем одно понятие, связанное с первым из двух названных классов задач. Пусть $A_h u_h = f_h$ — системы линейных алгебраических уравнений, аппроксимирующие некоторую задачу $Au = f$. Пусть известен некоторый алгоритм решения этих систем, требующий для своего осуществления N_h арифметических операций. Говорят, что этот алгоритм *оптимален по порядку*, если существует такая постоянная $c > 0$, не зависящая от h, что для любого другого алгоритма решения тех же систем число арифметических операций не менее cN_h. При любом способе решения систем уравнений каждая компонента искомого вектора u_h получается в результате некоторой арифметической операции. Поэтому тривиальной нижней границей для N_h является число компонент вектора u_h, т. е. для метода сеток — число узлов сетки. Если при применении к уравнению теплопроводности неявной разностной схемы на каждом шаге по времени мы применяем метод прогонки, то общее число арифметических операций, потребных для решения всей "большой" системы (с учетом подстановки решения, полученного на предыдущем временном слое, в правые части уравнений для следующего слоя), пропорционально числу узлов. Поэтому такой способ решения системы линейных уравнений неявной разностной схемы для уравнения теплопроводности оптимален по порядку.

В том определении оптимальности по порядку алгоритмов решения систем алгебраических уравнений, которое дано ранее, предполагалось, что мы допускаем к рассмотрению лишь точные методы решения таких систем. Между тем в действительности для решения систем уравнений метода сеток значительно большее распространение имеют итеративные методы. Для включения в рассмотрение и этой группы методов определение оптимальности по порядку можно изменить следующим образом. Будем считать, что при каждом h нам задано число $\varepsilon(h)$, характеризующее ту точность, с которой мы хотим получить решение системы уравнений $A_h u_h = f_h$. Тогда для каждого алгоритма решения этой системы N_h есть число арифметических операций, потребующихся для получения этим алгоритмом вектора, отличающегося от точного решения u_h^* в метрике пространства U_h не более чем на $\varepsilon(h)$. Число $\varepsilon(h)$ обычно естественным образом согласуется с порядком сходимости метода сеток.

Остановимся теперь на одном обобщении метода прогонки. Рассмотрим граничную задачу Дирихле для эллиптического дифференциального уравнения

$$-\frac{\partial}{\partial x}\left(p\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(q\frac{\partial u}{\partial y}\right)+c\frac{\partial u}{\partial x}+d\frac{\partial u}{\partial y}+ru=f, \qquad u\big|_{\partial\Omega}=g. \tag{7}$$

Здесь Ω — прямоугольник $\Omega=\{0\le x\le a,\quad 0\le y\le b\}$. Стороны этого прямоугольника будем считать соизмеримыми, и шаг h при применении к задаче метода сеток выбранным так, что $a=nh, b=mh.$ Можно было бы считать шаги h_1 и h_2 по переменным x и y различными, тогда предположение о соизмеримости сторон Ω нам не потребовалось бы, но суть дела от этого не изменилась бы. Систему сеточных уравнений запишем в привычном виде

$$\begin{aligned}&\frac{-p_{j+\frac12 k}u_{j+1k}+(p_{j+\frac12 k}+p_{j-\frac12 k})u_{jk}-p_{j-\frac12 k}u_{j-1k}}{h^2}+\\&+\frac{-q_{jk+\frac12}u_{jk+1}+(q_{jk+\frac12}+q_{jk-\frac12})u_{jk}-q_{jk-\frac12}u_{jk-1}}{h^2}+\\&+c_{jk}\frac{u_{j+1k}-u_{j-1k}}{2h}+d_{jk}\frac{u_{jk+1}-u_{jk-1}}{2h}+r_{jk}u_{jk}=f_{jk},\\&u_{0k}=g_{0k},\quad u_{nk}=g_{nk},\quad u_{j0}=g_{j0},\quad u_{jm}=g_{jm}.\end{aligned} \tag{8}$$

Здесь u_{jk} — искомое приближение к $u^*(x_j, y_k)$ $(x_j = jh,\ y_k = kh)$, $p_{\mu\nu} = p(\mu h, \nu h)$ (μ и ν — целые или полуцелые), и аналогичный смысл имеют $q_{\mu\nu}, c_{\mu\nu}, d_{\mu\nu}, r_{\mu\nu}, f_{\mu\nu}, g_{\mu\nu}$. Проделаем с нашей системой уравнений следующие операции. Во-первых, исключим из уравнений, аппроксимирующих дифференциальное, неизвестные $u_{0k}, u_{nk}, u_{j0}, u_{jm}$, воспользовавшись граничными условиями. Во-вторых, умножим эти уравнения на h^2. В-третьих, введем в рассмотрение векторы $U^{(j)} = (u_{j1}, ..., u_{jm-1})$ $(j = 1, 2, ..., n-1)$ порядка $m-1$, составленные из компонент искомого вектора u_h. Тогда систему уравнений (8) можно переписать в виде следующих соотношений между векторами $U^{(j)}$:

$$\left.\begin{array}{c} B_1U^{(1)} + C_1U^{(2)} = V^{(1)}, \\ A_2U^{(1)} + B_2U^{(2)} + C_2U^{(3)} = V^{(2)}, \\ \cdots\cdots\ \cdots\cdots\ \cdots\cdots\ \cdots\cdots \\ A_jU^{(j-1)} + B_jU^{(j)} + C_jU^{(j+1)} == V^{(j)}, \\ \cdots\cdots\ \cdots\cdots\ \cdots\cdots\ \cdots\cdots \\ A_{n-1}U^{(n-2)} + B_{n-1}U^{(n-1)} = V^{(n-1)}. \end{array}\right\} \qquad (9)$$

Здесь A_j и C_j — диагональные матрицы порядка $m-1$:

$$A_j = \begin{pmatrix} a_{11}^{(j)} & 0 & \cdots & 0 \\ 0 & a_{22}^{(j)} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{m-1m-1}^{(j)} \end{pmatrix}, \quad C_j = \begin{pmatrix} c_{11}^{(j)} & 0 & \cdots & 0 \\ 0 & c_{22}^{(j)} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & c_{m-1m-1}^{(j)} \end{pmatrix},$$

элементы которых задаются формулами

$$a_{kk}^{(j)} = -p_{j-\frac{1}{2}k} - \frac{1}{2}c_{jk}h, \quad c_{kk}^{(j)} = -p_{j+\frac{1}{2}k} + \frac{1}{2}c_{jk}h,$$

B_j — трехдиагональная матрица

$$B_j - \begin{pmatrix} b_{11}^{(j)} & b_{12}^{(j)} & 0 & \cdots & 0 & 0 \\ b_{21}^{(j)} & b_{22}^{(j)} & b_{23}^{(j)} & \cdots & 0 & 0 \\ 0 & b_{32}^{(j)} & b_{33}^{(j)} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & b_{m-1m-2}^{(j)} & b_{m-1m-1}^{(j)} \end{pmatrix}$$

с элементами

$$b_{k\,k-1}^{(j)} = -q_{\,jk-\frac{1}{2}} - \frac{1}{2} d_{\,jk} h, \quad b_{k\,k+1}^{(j)} = -q_{\,jk+\frac{1}{2}} + \frac{1}{2} d_{\,jk} h,$$

$$b_{k\,k}^{(j)} = p_{\,j+\frac{1}{2}k} + p_{\,j-\frac{1}{2}k} + q_{\,jk+\frac{1}{2}} + q_{\,jk-\frac{1}{2}} + r_{jk} h^2,$$

а $V^{(j)} = (\upsilon_{j1}, ..., \upsilon_{jm-1})$ — $(m-1)$-мерные векторы, причем при $2 \le j \le n-2$ и $2 \le k \le m-2$ $\upsilon_{jk} = h^2 f_{jk}$, а при $j = 1, n-1$ и $k = 1, m-1$ к $h^2 f_{jk}$ добавляются еще слагаемые, выражающиеся через значения функции g.

Систему уравнений (9) мы будем решать теперь точно так же, как и систему (1): будем искать выражение вектора $U^{(j)}$ через $U^{(j+1)}$ в виде[1]

$$U^{(j)} = Q_j U^{(j+1)} + D^{(j)}, \tag{10}$$

где Q_j — матрица порядка $m-1$, а $D^{(j)}$ — $(m-1)$-мерный вектор. Из первого из уравнений (9) имеем $U^{(j)} = Q_1 U^{(2)} + D^{(1)}$, где $Q_1 = -B_1^{-1} C_1$,

---

[1] Другой способ "последовательного исключения векторов $U^{(j)}$", при котором все векторы $U^{(j)}$ выражаются через $U^{(j)}$, приводит к катастрофическому пропаданию значащих цифр. Это видно из тех же соображений, которые были высказаны ранее.

$D^{(1)} = B_1^{-1}V^{(1)}$. Считая, что для $U^{(j-1)}$ формула вида (10) уже получена, и подставляя ее в j-е уравнение (9), получаем

$$A_j(Q_{j-1}U^{(j)} + D^{(j-1)}) + B_jU^{(j)} + C_jU^{(j+1)} = V^{(j)},$$

откуда

$$Q_j = -(A_jQ_{j-1} + B_j)^{-1}C_j, \quad D^{(j)} = (A_jQ_{j-1} + B_j)^{-1}(V^{(j)} - A_jD^{(j-1)}). \qquad (11)$$

Как и в случае системы уравнений (1), можно показать, что $U^{(n-1)} = D^{(n-1)}$. Таким образом, весь процесс решения системы уравнений (9), а значит, и (8), сводится к следующему:

1) по рекуррентным формулам (11) строятся матрицы Q_j и векторы $D^{(j)}$;

2) полагается $U^{(n-1)} = D^{(n-1)}$ и по формулам (10) последовательно находятся векторы $U^{(n-2)}, U^{(n-3)}, ..., U^{(1)}$.

Такой способ решения системы уравнений (9) получил название *метода матричной прогонки*.

Относительно осуществимости метода матричной прогонки может быть получен результат, вполне аналогичный теореме 1. При этом мы не будем предполагать, что система (9) происходит из (8), так что матрицы A_j, B_j и C_j могут быть произвольными, в частности, "полными". Для удобства формулировки введем еще матрицу $A_1 = 0$. Осуществимость метода матричной прогонки означает обратимость всех матриц $A_jQ_{j-1} + B_j$. Ниже $\|\cdot\|$ есть произвольная матричная операторная норма, т. е. норма матрицы как оператора в некотором нормированном $(m-1)$-мерном векторном пространстве.

Теорема 2

Пусть при всех $j = 1, 2, ..., n-1$ матрицы B_j обратимы и выполняются неравенства

$$\left\|B_j^{-1}A_j\right\| + \left\|B_j^{-1}C_j\right\| < 1.$$

Тогда при всех $j = 1, 2, ..., n-1$ матрицы $A_j Q_{j-1} + B_j$ также обратимы и $\|Q_j\| < 1$.

Доказательство

Доказательство проводится методом индукции. При $j = 1$ ввиду $A_1 = 0$ под $A_1 Q_0 + B_1$ следует понимать матрицу B_1, которая обратима. Далее $Q_1 = -B_1^{-1} C_1$, и по предположению теоремы $\|Q_1\| < 1$. Покажем возможность индуктивного перехода. Пусть уже известно, что $\|Q_{j-1}\| < 1$. Тогда имеем

$$A_j Q_{j-1} + B_j = B_j [B_j^{-1} A_j Q_{j-1} + I].$$

Здесь

$$\|B_j^{-1} A_j Q_{j-1}\| \le \|B_j^{-1} A_j\| \cdot \|Q_{j-1}\| \le \|B_j^{-1} A_j\| < 1,$$

и по теореме Банаха существует обратная матрица $[B_j^{-1} A_j Q_{j-1} + I]^{-1}$, причем

$$\|[B_j^{-1} A_j Q_{j-1} + I]^{-1}\| \le 1/(1 - \|B_j^{-1} A_j\|).$$

Тем самым существует и обратная матрица

$$[A_j Q_{j-1} + B_j]^{-1} = [B_j^{-1} A_j Q_{j-1} + I]^{-1} B_j^{-1}.$$

Наконец,

$$\|Q_j\| = \|[A_j Q_{j-1} + B_j]^{-1} C_j\| \le \|[B_j^{-1} A_j Q_{j-1} + I]^{-1}\| \cdot \|B_j^{-1} C_j\| \le$$
$$\le \|B_j^{-1} C_j\| / (1 - \|B_j^{-1} A_j\|) < 1. \blacksquare$$

Покажем применимость этой теоремы к системе уравнений (8) при следующих предположениях об уравнении (7):

а) коэффициент d уравнения тождественно равен нулю;

б) коэффициенты p, q, c и r зависят лишь от переменной x, но не от y;

в) $p(x) \ge p^0 > 0, \quad q(x) \ge q^0 > 0, \quad r(x) \ge 0.$

Кроме того, шаг h будем считать настолько малым, что выполняется неравенство

$$h \cdot \max \left| c(x) \right| < 2p^0. \tag{12}$$

При сделанных предположениях матрицы A_j и C_j принимают вид

$$A_j = \alpha_j I, \quad C_j = \gamma_j I,$$

где[1]

$$\alpha_j = -p_{j-1/2} - 1/2 c_j h, \quad \gamma_j = -p_{j+1/2} + 1/2 c_j h,$$

а элементы трехдиагональной матрицы B_j задаются формулами

$$b^{(j)}_{kk-1} = b^{(j)}_{kk+1} = -q_j, \quad b^{(j)}_{kk} = p_{j+1/2} + p_{j-1/2} + 2q_j + r_j h^2.$$

Все матрицы B_j обратимы по *теореме 1 из § 2*. Собственные числа $\lambda^{(j)}_k$ матриц $B_j^{-1} A_j$ и $\mu^{(j)}_k$ матриц $B_j^{-1} C_j$ даются той же теоремой:

$$\lambda^{(j)}_k = \frac{-p_{j-1/2} - \frac{1}{2} c_j h}{p_{j+1/2} + p_{j-1/2} + 2q_j + r_j h^2 - 2q_j \cos(k\pi/m)},$$

$$\mu^{(j)}_k = \frac{-p_{j+1/2} + \frac{1}{2} c_j h}{p_{j+1/2} + p_{j-1/2} + 2q_j + r_j h^2 - 2q_j \cos(k\pi/m)}.$$

Ввиду условия (12) при всех k выполняются неравенства

$$\left| \lambda^{(j)}_k \right| < \frac{p_{j-1/2} + \frac{1}{2} c_j h}{p_{j+1/2} + p_{j-1/2}}, \quad \left| \mu^{(j)}_k \right| < \frac{p_{j+1/2} - \frac{1}{2} c_j h}{p_{j+1/2} + p_{j-1/2}},$$

[1] Так как коэффициенты уравнения зависят лишь от x, то для обозначения их значений в точке $(vh, \mu h)$ достаточно указывать лишь индекс v.

откуда (учитывая симметричность матриц $B_j^{-1}A_j$ и $B_j^{-1}C_j$)

$$\left\| B_j^{-1}A_j \right\|_2 + \left\| B_j^{-1}C_j \right\|_2 = \max_k \left| \lambda_k^{(j)} \right| + \max_k \left| \mu_k^{(j)} \right| <$$

$$< \frac{p_{j-1/2} + \frac{1}{2}c_j h}{p_{j+1/2} + p_{j-1/2}} + \frac{p_{j+1/2} - \frac{1}{2}c_j h}{p_{j+1/2} + p_{j-1/2}} = 1,$$

и тем самым условия теоремы 2 выполнены и система уравнений (8) разрешима методом матричной прогонки.

Литература

1. Бакушинский А. Б. Один метод численного решения интегрального уравнения. — В кн.: Вычислислительные методы и программирование. — М., Вып. 3. 1965. С. 536—543.
2. Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы. — М., 2001.
3. Вайникко Г. М. Принцип компактной аппроксимации в теории приближенных методов / Журн. вычисл. мат. и мат. физ., 1969. Т. 9, № 4. С. 739—761.
4. Вайникко Г. М. Компактная аппроксимация операторов и приближённое решение уравнений. — Тарту, 1970.
5. Гавурин М. К. Лекции по методам вычислений. — М., 1971.
6. Гавурин М. К., Самокиш Б. А. Замечания по поводу определения устойчивости вычислительных алгоритмов. — В кн.: Методы вычислений. — Л., Вып. 6. 1970. С. 13—21.
7. Годунов С. К., Рябенький В. С. Разностные схемы. Введение в теорию. М., 1977.
8. Даугавет И. К. Об итеративном решении уравнений, возникающих при компактной аппроксимации операторов / Журн. вычисл. мат. и мат. физ., 1980. Т. 20, № 4. С. 1046—1049.
9. Канторович Л. В. Функциональный анализ и прикладная математика / Успехи мат. наук, 1948. Т. 3, № 6. С. 89—185.
10. Канторович Л. В., Акилов Г. П. Функциональный анализ. — М., 1984.
11. Канторович Л. В. Крылов В. И. Приближенные методы высшего анализа. — М.; Л., 1962.
12. Келдыш М. В. О методе Б. Г. Галеркина для решения краевых задач / Изв. АН СССР, сер. мат., 1942, № 6. С. 309—330.
13. Крылов В. И. Приближенное вычисление интегралов. — М., 1967.

14. Марчук Г. И. Методы вычислительной математики. — М., 1977.

15. Михлин С. Г. Вариационные методы в математической физике. — М., 1970.

16. Михлин С. Г. Численная реализация вариационных методов. — М., 1966.

17. Мысовских И. П. О сходимости метода механических кубатур для решения интегральных уравнений. — В кн.: Методы вычислении. — Л., Вып. 4. 1967. С. 63—72.

18. Рихтмайер Р. Д. Разностные методы решения краевых задач. — М., 1960.

19. Самарский А. А. Введение в теорию разностных схем. — М., 1971.

20. Самарский А. Д., Николаев Е.С. Методы решения сеточных уравнений. — М., 1978.

21. Соболев С. Л. Некоторые замечания о численном решении интегральных уравнений / Изв. АН СССР, сер. мат., 1956. Т. 20, № 4. С. 413—436.

22. Харрик И. Ю. О приближении функций, обращающихся в нуль на границе области, функциями особого вида / Мат. сб., 1955. Т. 37, № 2. С. 355—384.

23. Atkinson K. Iterative variants of the Nyström method for the numerical solution of integral equations. — Numer. Math., 1973, vol. 22, № 1. P. 17—31.

24. Taylor J. M. The condition of Gram matrices and related problems. — Proc. Roy. Soc. Edinburgh, 1978, vol. A80, № 1, 2. P. 45—56.