И. Н. Молчанов, Л. Д. Николенко

ОСНОВЫ МЕТОДА КОНЕЧНЫХ ЭЛЕМЕНТОВ

АКАДЕМИЯ НАУК
УКРАИНСКОЙ ССР
ИНСТИТУТ КИБЕРНЕТИКИ
ИМЕНИ В. М. ГЛУШКОВА

И. Н. Молчанов,
Л. Д. Николенко

ОСНОВЫ МЕТОДА КОНЕЧНЫХ ЭЛЕМЕНТОВ

КИЕВ НАУКОВА ДУМКА 1989

УДК 519.3

Основы метода конечных элементов / Молчанов И. Н., Николенко Л. Д. Отв. ред. Галба Е. Ф.; АН УССР. Ин-т кибернетики имени В. М. Глушкова. — Киев : Наук. думка, 1989.— 272 с.— ISBN 5-12-000531-4.

В монографии рассмотрены алгоритмы решения некоторых задач математической физики методом конечных элементов (МКЭ). Приведено теоретическое обоснование МКЭ для рассматриваемых классов задач, доказана сходимость построенных приближенных решений и даны оценки их точности. Обсуждены проблемы машинной реализации алгоритмов, а также вопросы достоверности получаемых машинных решений. Применение алгоритмов проиллюстрировано решением ряда модельных задач.

Построение алгоритмов, их теоретическое исследование, вопросы машинной реализации, оценки достоверности получаемых машинных решений являются методологической основой для решения прикладных задач МКЭ. Использование изложенной методики продемонстрировано на примере решения некоторых прикладных задач.

Для научных работников и инженеров, занимающихся численным решением задач термоупругопластичности, задач на колебания и устойчивость, а также для специалистов по прикладной математике. Будет полезна аспирантам и студентам старших курсов соответствующей специальности.

Ил. 49. Табл. 27. Библиогр.: с. 261—267 (160 назв.).

Ответственный редактор *Е. Ф. Галба*

Утверждено к печати ученым советом
Института кибернетики имени В. М. Глушкова АН УССР

Редакция физико-математической литературы

Редактор *В. П. Егорова*

$\text{М}\frac{1602110000\text{-}063}{\text{М221(04)-89}}$ 186-89

ISBN 5-12-000531-4

ПРЕДИСЛОВИЕ

Метод конечных элементов в последние десятилетия стал одним из широко распространенных численных методов. Он является эффективным средством дискретизации (построения дискретного аналога) разнообразных дифференциальных и вариационных задач математической физики. Метод конечных элементов составляет алгоритмическую основу многих пакетов прикладных программ, разрабатываемых и используемых в различных сферах человеческой деятельности.

Следует отметить, что развитие и эффективность МКЭ, как и любого численного метода, обусловлены тесной взаимосвязью трех факторов: современной вычислительной техники, высококачественных математических моделей исследуемых процессов и объектов и, наконец, свойств и характеристик самого численного метода, т. е. МКЭ. Эта триада превращает численный эксперимент в мощное орудие познания окружающего мира. Под численным экспериментом понимается исследование свойств объекта или явления посредством решения на ЭВМ задачи, представляющей собой математическую модель этого явления или объекта. Задавая исходные данные и решая на их основе соответствующую задачу, можно понять значение различных факторов в исследуемом процессе или объекте. Выполнение численного эксперимента (наряду с натурным моделированием) существенно сокращает сроки проектно-конструкторских разработок, снижает расходы материалов и энергозатрат.

Использование математических моделей существенно расширяет возможности познания и прогнозирования, сокращает время проведения исследования при получении фундаментальных научных результатов.

Таким образом, численный эксперимент можно рассматривать как основу при создании систем автоматизации проектирования и автоматизации научных исследований, а МКЭ — как один из эффективных численных методов машинной реализации численного эксперимента.

Основные математические аспекты МКЭ впервые были представлены в работе Р. Куранта [129], где для решения задачи об изгибе мембраны применялся следующий вариант метода Ритца. Исходная область рассматривалась как совокупность квадратов, каждый из которых разбивался диагоналями на треугольники. В качестве базисных функций использовались кусочно-линейные полиномы, непрерывные во всей области и отличные от нуля только на нескольких треугольниках.

В то же время развитие ранее известных инженерных приемов исследования привело в технике к идее представления конструкций в виде дискретных элементов и появлению достаточно общей процедуры изучения широкого класса прикладных задач. Первоначально связь этой процедуры с упомянутым выше вариантом метода Ритца не была замечена. Однако в дальнейшем работами многих исследователей эта связь была установлена и инженерная процедура получила математическое

обоснование. Теоретически обоснованный численный метод, удобный для реализации на ЭВМ, назвали методом конечных элементов.

В настоящей монографии рассматривается математическая трактовка МКЭ. На примере решения краевых задач для различных линейных и нелинейных обыкновенных дифференциальных уравнений, начально-краевых задач для дифференциальных уравнений в частных производных второго порядка параболического типа, нахождения собственных чисел и собственных функций обыкновенных дифференциальных операторов второго и четвертого порядков показаны алгоритм реализации этого метода и его сходимость для классических примеров краевых задач математической физики; приведен порядок точности получаемых приближенных решений. Теоретические положения проиллюстрированы рядом примеров решения различных модельных задач.

Многообразие встречающихся на практике задач далеко не исчерпывается известными классическими примерами. Поэтому для решения прикладных задач на основе общей методологии МКЭ (изложенной в гл. I—V) приходится строить специальные алгоритмы и проводить дополнительные теоретические исследования. Из ряда решенных авторами и их сотрудниками задач были отобраны четыре, характеризующие, с нашей точки зрения, как целесообразность рассмотрения подобных математических моделей, так и особенность возникающих задач. Использование МКЭ при решении этих четырех задач рассмотрено в гл. VI.

При решении задач на ЭВМ любым методом, в том числе и МКЭ, возникает проблема достоверности получаемого машинного решения задачи. Эта проблема многогранна и содержит в себе достоверность математической модели, корректность (правильность) использования МКЭ и аппарата численного анализа при формировании систем уравнений на ЭВМ и машинную реализацию этого метода.

Корректному применению МКЭ и вопросам его машинной реализации посвящены гл. I—VI, вопросам достоверности решений систем уравнений МКЭ, «подводным камням», встречающимся при машинной реализации алгоритмов решения, а также характеристике методов и программ решения — гл. VII монографии.

Отметим, что теоретически обоснованное применение МКЭ для решения современных научно-технических задач требует использования современного математического аппарата и довольно тонких теоретических исследований. Однако только на этом пути можно гарантировать получение достоверных решений, не искажающих физического смысла задач.

Для облегчения чтения монографии основные идеи и теоретические обоснования МКЭ показаны на примерах простейших (одномерных) задач, а затем реализация этих идей иллюстрируется на решении двумерных прикладных задач.

Предлагаемая монография написана на основе лекций, прочитанных одним из авторов в качестве спецкурса студентам Киевского государственного университета им. Т. Г. Шевченко, специализирующимся по вычислительной математике, и студентам-математикам Высшей технической школы им. Отто фон Герике в Магдебурге (ГДР).

Авторы признательны академику А. А. Дородницину за поддержку идеи написания этой монографии и советы, определившие круг освещаемых в ней вопросов. Авторы благодарны коллегам и сотрудникам по работе: М. Ф. Яковлеву, И. С. Левченко, А. В. Попову, А. Ю. Незлиной, Н. А. Бик, В. С. Дейнеке, А. Н. Химичу, Е. Ф. Галбе за ценные замечания, а также Т. А. Герасимовой, А. Н. Нестеренко, И. П. Винокуровой, Д. Н. Назаровой, Н. В. Лапс, Т. А. Лиходзиевской за техническую подготовку рукописи к набору.

Глава I

НЕКОТОРЫЕ ПРЕДВАРИТЕЛЬНЫЕ СВЕДЕНИЯ И ПОНЯТИЕ О МЕТОДЕ КОНЕЧНЫХ ЭЛЕМЕНТОВ (МКЭ)

Глава посвящена общей характеристике метода конечных элементов, являющегося эффективным орудием численного эксперимента в научно-технических исследованиях. Сформулированы математические задачи теории упругости, необходимые для реализации достоверного численного эксперимента. Для удобства чтения в этой главе приводятся также некоторые вспомогательные сведения и понятия, используемые в дальнейшем изложении.

I.1. Постановка задач и метод конечных элементов как средство описания дискретных задач

1. Понятие о численном эксперименте. Внедрение вычислительной техники во все области человеческой деятельности является характерной чертой научно-технической революции. Вычислительные машины находят широкое применение в фундаментальных научных исследованиях, в том числе и в автоматизации научных исследований, в системах автоматизации проектирования, в планировании и управлении народным хозяйством, в управлении дискретным производством (и в реальном масштабе времени), в информационном обеспечении общества.

В ряде областей применения численный эксперимент становится одним из средств научно-технического исследования и прогнозирования. Под численным экспериментом понимают исследование свойств объекта посредством решения на ЭВМ задач, представляющих собой математическую модель объекта. Многократное проигрывание моделей для различных исходных данных позволяет понять роль и значение различных факторов для течения того или иного процесса или поведения объекта и дает возможность правильно планировать и проводить натурный (физический) эксперимент.

Использование численного эксперимента позволяет существенно повысить технический уровень и качество проектируемой и выпускаемой продукции, снизить расходы материалов на изготовление объектов, уменьшить энергетические затраты при эксплуатации проектируемых объектов, сократить сроки и объем натурных испытаний

исследуемых объектов и выявить новые теоретико-технические качества исследуемого объекта.

Для численного эксперимента необходимы математические модели исследуемых объектов или явлений, машинные методы решения соответствующих математических задач и вычислительные машины, на которых численный эксперимент реализуется.

2. Математические задачи теории упругости. Рассмотрим математические модели, с помощью которых описываются задачи теории упругости, где по заданным, действующим на твердое тело, внешним силам требуется найти те изменения формы, которые тело претерпевает, и те внутренние силы упругости, которые возникают между частями тела при этих изменениях формы.

При постановке задач теории упругости обычно задают форму тела, его упругие характеристики, выражающие свойство материала, объемные (массовые) силы, условия нагружения или закрепления тела.

Анализ прикладных задач, описываемых математическими моделями теории упругости, позволяет сделать следующие выводы: наряду с изучением отдельных элементов, узлов конструкций возникает потребность в интегрированном изучении объектов в целом, например, статически напряженного состояния летательного аппарата, причем исследуемые объекты обладают большими размерами и сложной геометрией; существует настоятельная необходимость в решении пространственных задач; наряду со статической прочностью объектом исследования становится динамическая прочность; бóльшую значимость приобретает изучение нелинейных моделей, линеаризация которых может приводить к искажению физического смысла задачи; необходимо рассматривать уравнения с переменными коэффициентами в связи с использованием композиционных материалов, обладающих неоднородными свойствами; возникает потребность в создании объектов, не только прочных и рассчитанных на определенную нагрузку, но и минимальных по весу или экономных по затраченным материалам.

Линейная классическая теория упругости базируется на ряде гипотез, основными из которых являются предположения: о сведении системы сил, действующих на элементарную площадку, только к равнодействующей (отсутствие моментов); о малости градиентов перемещений; о линейной связи между деформациями и перемещениями; об идеальной упругости материала (линейная связь между напряжениями и деформациями).

В зависимости от формы исследуемого тела выбирают ту или иную систему координат. Для определенности нами будут рассматриваться в дальнейшем декартовы координаты. Каждая точка упругого тела, отнесенного к декартовой системе координат $Ox_1x_2x_3$, характеризуется вектором перемещений u с компонентами u_1, u_2, u_3, а также тензором напряжений T_σ и деформаций T_ε:

$$T_\sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}, \quad T_\varepsilon = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{bmatrix}.$$

В классической теории упругости $\sigma_{ik} = \sigma_{ki}$, $\varepsilon_{ki} = \varepsilon_{ik}$. Если известен тензор T_σ, то можно определить компоненты вектора напряжений p_n на любой произвольно ориентированной площадке в данной точке (n — нормаль к площадке):

$$p_{n_i} = \sigma_{i_1} \cos(n, x_1) + \sigma_{i_2} \cos(n, x_2) + \sigma_{i_3} \cos(n, x_3), \quad i = 1, 2, 3.$$

Компоненты тензора деформаций связаны с перемещениями формулами Коши (геометрическими соотношениями):

$$\varepsilon_{ik} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right), \quad i, k = 1, 2, 3. \tag{I.1}$$

Компоненты ε_{ii} характеризуют относительные удлинения в направлении соответствующих осей, а ε_{ik} ($i \neq k$) — относительные сдвиги (изменение углов между осями Ox_i и Ox_k).

Для упругих тел в случае малых деформаций связь между компонентами тензора напряжений и деформаций выражается обобщенным законом Гука

$$\sigma_{ik} = \sum_{l,m=1}^{3} c_{iklm}\varepsilon_{lm}, \quad i, k = 1, 2, 3, \tag{I.2}$$

$$\varepsilon_{ik} = \sum_{l,m=1}^{3} d_{iklm}\sigma_{lm}, \quad i, k = 1, 2, 3, \tag{I.3}$$

где

$$c_{iklm} = c_{kilm} = c_{lmik},$$

$$d_{iklm} = d_{kilm} = d_{lmik}.$$

Потенциальная энергия деформации тела может быть вычислена по формуле

$$\Pi = \frac{1}{2}\iiint_{\Omega} \sum_{i,k=1}^{3} \sigma_{ik}\varepsilon_{ik}\, d\Omega,$$

Здесь $d\Omega = dx_1 dx_2 dx_3$; Ω — объем тела.

Учитывая соотношения (I.2), (I.3), потенциальную энергию можно выразить либо только через компоненты тензора напряжений, либо только через компоненты тензора деформаций:

$$\Pi = \frac{1}{2}\iiint_{\Omega} \sum_{i,k,l,m=1}^{3} d_{iklm}\sigma_{ik}\sigma_{lm}\, d\Omega,$$

$$\Pi = \frac{1}{2}\iiint_{\Omega} \sum_{i,k,l,m=1}^{3} c_{iklm}\varepsilon_{ik}\varepsilon_{lm}\, d\Omega.$$

Выражение

$$\Phi = \Pi + \iiint_{\Omega} \sum_{i=1}^{3} f_i u_i\, d\Omega - \iint_{\Gamma_2} \sum_{i=1}^{3} u_i q_i\, d\Gamma \tag{I.4}$$

определяет полную энергию деформации упругого тела, где Γ_2 — часть поверхности Γ исследуемого тела, f_i — объемные, q_i — поверхностные силы ($i = 1, 2, 3$).

Считая тело идеально упругим и учитывая, что в (1.4) пределы интегрирования и силы f_i, q_i не зависят от перемещений, можно записать начало возможных перемещений в виде

$$\delta\left[\Pi + \iiint\limits_{\Omega} \sum_{i=1} f_i u_i d\Omega - \iint\limits_{\Gamma_2} \sum_{i=1} u_i q_i d\Gamma\right] = 0. \qquad (I.5)$$

Равенство (I.5) показывает, что из всех возможных перемещений при заданных внешних силах, имеющих потенциал, равновесию тела будут соответствовать те перемещения, при которых полная энергия Φ принимает минимальное значение. Поэтому возможна следующая вариационная постановка задачи о равновесии упругого тела: найти вектор-функцию $u = (u_1, u_2, u_3)^T$, доставляющую минимум функционалу (I.4) в классе функций, удовлетворяющих на границе Γ_1 области Ω (в местах закрепления тела) условиям

$$u = g(x_1, x_2, x_3), \ (x_1, x_2, x_3) \in \Gamma_1. \qquad (I.6)$$

Отметим, что воздействие внешних сил, приложенных к исследуемому телу, учитывается в (I.4) интегралом, содержащим q_i, $i = 1, 2, 3$.

Начало возможных перемещений является самым общим началом статики, поэтому из соотношения (I.5) могут быть получены дифференциальные уравнения равновесия упругого тела

$$\sum_{k=1}^{3} \frac{\partial \sigma_{ik}}{\partial x_k} + f_i = 0, \ i = 1, 2, 3, \qquad (I.7)$$

а на границе нагружения тела Γ_2 — условия

$$\sum_{k=1}^{3} \sigma_{ik} \cos(n, x_k) + q_i = 0, \ i = 1, 2, 3, (x_1, x_2, x_3) \in \Gamma_2, \qquad (I.8)$$

соответствующие заданным на поверхности силам; на оставшейся части Γ_1 границы Γ в местах закрепления задают условие (I.6).

Отметим, что дифференциальные уравнения равновесия упругого тела (I.7) могут быть получены и из известных принципов статики в предположении, что составляющие тензора напряжений имеют непрерывные частные производные первого порядка во всей области, занятой телом.

Так как по обобщенному закону Гука напряжения можно выразить через деформации, а следовательно, через перемещения u_1, u_2, u_3, и деформации можно выразить через напряжения, то в теории упругости одну и ту же задачу можно решать либо в перемещениях, либо в напряжениях. В дальнейшем будем рассматривать задачи, сформулированные относительно перемещений.

Итак, ставится следующая математическая задача: найти решение $u = (u_1, u_2, u_3)^T$ системы дифференциальных уравнений (I.7) внутри замкнутой области Ω, удовлетворяющее на границе Γ_2 условиям (I.8), а на Γ_1 условиям (I.6).

В физически линейных и геометрически нелинейных задачах вместо формул Коши применяют соотношения

$$\varepsilon_{ik} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i} + \sum_{\alpha=1}^{3} \frac{\partial u_\alpha}{\partial x_k}\frac{\partial u_\alpha}{\partial x_i}\right),$$
$$i,\ k = 1,\ 2,\ 3.$$

Используя начало возможных перемещений, для задач нелинейной теории упругости можно получить дифференциальные уравнения равновесия упругого тела

$$\sum_{n,k=1}^{3} \frac{\partial}{\partial x_n}\left[\sigma_{nk}\left(\delta_{ik} + \frac{\partial u_i}{\partial x_k}\right)\right] + f_i = 0,\ i = 1,\ 2,\ 3,\ (x_1, x_2, x_3) \in \Omega \quad \text{(I.9)}$$

и граничные условия, где заданы поверхностные силы

$$\sum_{n,k=1}^{3} \sigma_{nk}\left(\delta_{ik} + \frac{\partial u_i}{\partial x_k}\right)\cos(n, x_n) = q_i,\ i = 1,\ 2,\ 3, \quad \text{(I.10)}$$
$$(x_1, x_2, x_3) \in \Gamma_2.$$

Здесь δ_{ik} — символ Кронекера, n — внешняя нормаль. На границе закрепления тела Γ_1 задают перемещения

$$u_i = g_i,\ (x_1, x_2, x_3) \in \Gamma_1, \quad \text{(I.11)}$$

причем $\Gamma = \Gamma_1 \cup \Gamma_2$.

Таким образом, формулируется математическая задача: найти внутри замкнутой области Ω решение системы дифференциальных уравнений (I.9), удовлетворяющее на границе Γ_2 условиям (I.10), а на Γ_1 условиям (I.11).

Используя начало Д'Аламбера, уравнения, например, (I.7) можно обобщить на случай, когда точки твердого тела находятся в движении. Тогда справедливы уравнения динамики

$$\sum_{k=1}^{3} \frac{\partial \sigma_{ik}}{\partial x_k} + f_i - \rho\frac{\partial^2 u_i}{\partial t^2} = 0,\ i = 1,\ 2,\ 3, \quad \text{(I.12)}$$

где ρ — плотность тела, t — время.

Решение уравнений (I.12) находят при краевых условиях (I.8), (I.6), которые также становятся зависящими от времени, и при начальных условиях

$$u_i(x_1, x_2, x_3, 0) = \mu_{1,i},$$
$$\frac{\partial u_i}{\partial t}(x_1, x_2, x_3, 0) = \mu_{2,i},\ i = 1,\ 2,\ 3.$$

При исследовании ряда прикладных задач возникает необходимость в изучении статики или динамики напряженно-деформированного состояния неравномерно напряженных тел, подверженных воздействию источников тепла. Температурное поле $T = T(x_1, x_2, x_3, t)$ получим, решив для твердого тела дифференциальные уравнения параболического типа второго порядка

$$c\frac{\partial T}{\partial t} = \sum_{\alpha=1}^{3} \frac{\partial}{\partial x_\alpha}\left(k\frac{\partial T}{\partial x_\alpha}\right) + \varphi$$

с условиями на границе тела одного из видов

$$T = \mu_1(x_1, x_2, x_3, t),$$

$$k\frac{\partial T}{\partial n} = \mu_2(x_1, x_2, x_3, t),$$

$$k\frac{\partial T}{\partial n} - \beta T = \mu_3(x_1, x_2, x_3, t)$$

(или смешанными граничными условиями) и с начальным условием

$$T(x_1, x_2, x_3, 0) = T_0(x_1, x_2, x_3), \quad (x_1, x_2, x_3) \in \Omega.$$

Здесь $c = c(x_1, x_2, x_3, t) > 0$ — теплоемкость единичного объема, $k = k(x_1, x_2, x_3) > 0$ — коэффициент теплопроводности, $\varphi = \varphi(x_1, x_2, x_3, t)$ — плотность тепловых источников (стоков) тепла внутри области, β — заданный коэффициент. После нахождения распределения температуры в теле необходимо решать уравнения термоупругости.

Если нагреть элемент тела до температуры T и если не будет препятствий для его расширения, то этот элемент расширяется во всех направлениях, причем тепловые деформации его выразятся формулой

$$\varepsilon_{ij} = \begin{cases} \alpha T, & i = j, \\ 0, & i \neq j, \quad i, \quad j = 1, 2, 3, \end{cases}$$

где α — коэффициент линейного теплового расширения.

Если тело нагрето неравномерно или какие-либо участки его поверхности связаны с другими телами, то элементы тела не могут свободно расширяться. В этом случае соотношения (I.3) принимают вид

$$\varepsilon_{ik} = \sum_{l,m=1}^{3} d_{iklm}\sigma_{lm} + \delta_{ik}\alpha T, \quad i, k = 1, 2, 3, \tag{I.13}$$

где δ_{ik} — символ Кронекера.

Решая систему (I.13) относительно σ_{ik}, можно найти

$$\sigma_{ik} = \sum_{l,m=1}^{3} c_{iklm}(\varepsilon_{lm} - \delta_{lm}\alpha T), \; i, k = 1, 2, 3. \tag{I.14}$$

Из принципа возможных перемещений с учетом (I.13), (I.14) можно получить уравнения равновесия термоупругого тела

$$\sum_{k=1}^{3}\frac{\partial}{\partial x_k}\sum_{l,m=1}^{3} c_{iklm}\left\{\frac{1}{2}\left(\frac{\partial u_l}{\partial x_m} + \frac{\partial u_m}{\partial x_l}\right) - \delta_{lm}\alpha T\right\} + f_i = 0, \; i = 1, 2, 3,$$

которые справедливы внутри исследуемого тела, а на поверхности тела можно сформулировать краевые условия одного из видов

$$\frac{1}{2}\sum_{l,m=1}^{3} c_{iklm}\left(\frac{\partial u_l}{\partial x_m} + \frac{\partial u_m}{\partial x_l} - 2\delta_{lm}\alpha T\right)\cos(n, x_k) = q_i,$$

$$u_i = g_i, \; i = 1, 2, 3,$$

или смешанные краевые условия.

Аналогично задачи термоупругости могут быть сформулированы для нелинейных и динамических задач теории упругости. Введением уп-

рощающих гипотез, учитывающих особенности геометрии рассматриваемых тел, свойства материалов этих тел и т. д., был создан ряд прикладных теорий, позволяющих приближенно (аналитически или численно) решать ряд задач. Одной из таких важных задач является задача на колебания или устойчивость, которая сводится к нахождению нескольких минимальных собственных чисел и соответствующих им собственных функций уравнения

$$Lu = \lambda u, \text{ или } Lu = \lambda Mu,$$

где L, M — некоторые дифференциальные операторы, описывающие те или иные математические модели в прикладных теориях.

В качестве примера можно привести задачу о колебании прямоугольной в плоскости Oxy пластины, испытывающей сжимающие и сдвигающие усилия. В этом случае опраторы L и M приобретают следующий вид:

$$Lu = \frac{\partial^2}{\partial x^2}\left(k_1 \frac{\partial^2 u}{\partial x^2} + k_0 \frac{\partial^2 u}{\partial y^2}\right) + 2\frac{\partial^2}{\partial x \partial y}\left(k_3 \frac{\partial^2 u}{\partial x \partial y}\right) + \\ + \frac{\partial^2}{\partial y^2}\left(k_2 \frac{\partial^2 u}{\partial y^2} + k_0 \frac{\partial^2 u}{\partial x^2}\right),$$

$$Mu = \frac{\partial}{\partial x}\left(m_1 \frac{\partial u}{\partial x} + m_3 \frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial y}\left(m_2 \frac{\partial u}{\partial y} + m_3 \frac{\partial u}{\partial x}\right),$$

где k_0, k_1, k_2, k_3, m_1, m_2, m_3 — коэффициенты, представляющие собой жесткости, сжимающие и сдвигающие усилия.

Характер закрепления краев пластины дает различные типы краевых условий. Например, жесткое закрепление характеризуется на границе области условиями $u = 0$, $\frac{\partial u}{\partial n} = 0$ (n — внешняя нормаль к границе).

Вычисленные решения математических задач теории упругости должны быть проанализированы как с качественной, так и с количественной стороны. Качественный анализ показывает наиболее нагруженные и наиболее ослабленные с точки зрения напряженного состояния части конструкции. С помощью количественного анализа можно определить необходимый материал, размер и форму конструкции для того, чтобы удовлетворить нужным условиям прочности и жесткости.

Таким образом, нами рассмотрены некоторые математические задачи, возникающие в теории упругости. Необходимо отметить, что описание одной и той же задачи теории упругости многовариантно. Как это было видно на примере задачи линейной статической теории упругости, для описания ее можно использовать уравнения упругого равновесия тел в перемещениях или напряжениях. Эту же задачу можно сформулировать как вариационную. Отметим, что вариационные постановки задач предъявляют минимальные требования к гладкости как самих решений, так и коэффициентов, используемых в задаче.

Математические задачи теории упругости решаются в основном как методом конечных разностей, так и различными вариантами метода конечных элементов.

С математической точки зрения метод конечных элементов (МКЭ) может трактоваться как обобщение методов Рэлея, Ритца, Бубнова — Галеркина, основанное на специальном выборе координатных (базисных) функций. Вначале он развивался как машинно-ориентированный метод описания и решения прикладных задач. Благодаря усилиям математиков и специалистов в области вычислительной математики было дано теоретическое обоснование МКЭ как средства приближенного решения математических задач.

3. Метод конечных элементов как средство описания дискретных задач. Традиционно усилия специалистов в линейной теории упругости были направлены на решение краевой задачи (I.6) — (I.8). Введением упрощающих гипотез, учитывающих особенности геометрии рассматриваемых тел, свойства материалов и т. д., был создан ряд прикладных теорий, позволяющих приближенно или аналитически решать некоторые частные задачи. С появлением современных ЭВМ широко стали разрабатываться машинные методы решения общих уравнений теории упругости, в том числе и метод конечных элементов. Ряд специалистов-прикладников используют МКЭ как средство получения непосредственно дискретной модели исследуемого физического явления. В данном случае рассматриваемая область разбивается на простые части (элементы) и на основе физики явления составляются уравнения равновесия и неразрывности этих частей, устанавливается взаимосвязь между элементами. В рамках такой дискретной аппроксимации искомые параметры всегда имеют определенный физический смысл. Рассмотрим использование МКЭ в подобном плане на примере решения задач о напряженно-деформированном состоянии тел в нелинейной теории упругости. Предположим, что замкнутая область Ω, в которой ищется решение, разбивается на некоторое число подобластей — элементов Ω_i, $i = 1, 2, \ldots, N$. В каждой из подобластей неизвестные поля могут быть аппроксимированы сравнительно простыми аналитическими выражениями, зависящими от нескольких параметров. Напряженное состояние или соответствующая деформация внутри аппроксимируется, как правило, полиномами. Если используется метод конечных элементов в варианте метода перемещений, то за основные неизвестные принимаются компоненты перемещений в узловых точках, расположение которых зависит от формы подобласти, вида используемого полинома. При этом напряженно-деформированное состояние элемента однозначно определяется через его узловые перемещения. В пространственных задачах перемещения в подобласти аппроксимируются трехмерной вектор-функцией

$$v(x_1, x_2, x_3) = G(x_1, x_2, x_3)\, u,$$

где $G(x_1, x_2, x_3)$ — матрица, u — вектор, состоящий из фиксированных значений искомых перемещений в выбранных узловых точках элемента, по которым позже устанавливается связь с соседними элементами. Используя соотношение между растяжением и сдвигом, представленное в матричном виде

$$\varepsilon(x_1, x_2, x_3) = D_G(x_1, x_2, x_3)\, u, \qquad (I.15)$$

и закон Гука, записанный следующим образом:

$$\sigma(x_1, x_2, x_3) = H\varepsilon,$$

можно выразить напряжения σ через перемещения u в фиксированных узлах. Векторы σ, ε составлены из ненулевых компонент соответствующих тензоров. В равенстве (I.15) D_G можно построить, используя матрицу G и формулы Коши (I.1). Принцип минимума потенциальной энергии позволяет получить условия равновесия

$$f^{(i)} = K_i u^{(i)}$$

на каждом элементе Ω_i, $i = 1, 2, \ldots, N$. Здесь $u^{(i)}$ — вектор искомых перемещений в узловых точках Ω_i, а $f^{(i)}$ — вектор сил, приложенных в узлах Ω_i. Выполнение условий равновесия в узлах всех элементов области приводит к системе линейных алгебраических уравнений относительно перемещений узловых точек

$$Ku = f$$

с симметричной разреженной матрицей K, называемой глобальной матрицей жесткости. На основе физических соображений (заданы условия, сдерживающие перемещения тела как целого) эта матрица — положительно определенная. Решение системы линейных алгебраических уравнений дает нам приближения к перемещениям в узловых точках исследуемой задачи. Имея перемещения, можно легко вычислить компоненты тензора напряжений и деформаций в любой точке тела.

В параграфе I.3 мы продолжим описание метода конечных элементов, рассмотрим некоторые общие вопросы, связанные с реализацией МКЭ как численного метода. Но вначале в параграфе I.2 приведем ряд необходимых вспомогательных сведений, которые будут полезны во всем дальнейшем изложении.

I.2. Необходимые вспомогательные сведения

1. Обозначения и определения. Основную роль в последующем изложении (см. гл. II—V) будут играть функциональные пространства, элементами которых являются вещественные функции $u(x)$ одной переменной x, принадлежащей ограниченному интервалу (a, b), и вещественные функции $u(x, t)$, областью определения которых служит прямоугольник $Q_T = (a, b) \times (0, T)$.

Укажем главные из функциональных пространств, встречающихся в монографии.

Банахово пространство $\mathfrak{B}$, т. е. полное линейное нормированное пространство (норму элемента u будем обозначать как $\|u\|_{\mathfrak{B}}$ или просто $\|u\|$).

Нормированное пространство E называется полным, если для любой последовательности $\{u_n\}$ элементов этого пространства из условия $\|u_p - u_q\| \to 0$ при $p, q \to \infty$ следует существование предельного элемента, принадлежащего E.

Множество $M \subset \mathfrak{B}$ называется плотным в множестве $M_0 \subset \mathfrak{B}$, если M_0 содержится в замыкании M, т. е. $M_0 \subset \overline{M}$. Если $\overline{M} = \mathfrak{B}$,

то M называют всюду плотным множеством (в $\mathfrak{B}$). Очевидно, что любой элемент банахова пространства $\mathfrak{B}$ является пределом последовательности элементов из множества M, плотного в $\mathfrak{B}$.

Если в $\mathfrak{B}$ имеется счетное всюду плотное в $\mathfrak{B}$ множество элементов, то пространство $\mathfrak{B}$ называется сепарабельным.

Множество $M \subset \mathfrak{B}$ называется компактным в банаховом пространстве $\mathfrak{B}$, если любая бесконечная последовательность $\{u_n\}$ элементов из M содержит сходящуюся в себе подпоследовательность, т. е. $\|u_p - u_q\| \to 0$ при $p, q \to \infty$.

Прежде чем переходить к примерам конкретных банаховых пространств, приведем еще несколько определений и понятий.

Функция $u(x)$ называется суммируемой на (a, b), если ее интеграл в смысле Лебега $\int_a^b u(x)\,dx$ конечен.

Приведем определение обобщенной производной (типа функции), принадлежащее С. Л. Соболеву. Ограничимся случаем функции одной переменной (подробнее см. в [98]).

Пусть функции $u(x)$ и $v(x)$ суммируемы на любом строго внутреннем подынтервале (a', b') интервала (a, b). Если для любой бесконечно дифференцируемой на $[a, b]$ функции $\psi(x)$, равной нулю в окрестности концов отрезка $[a, b]$, справедливо тождество

$$\int_a^b u(x)\frac{d\psi}{dx}\,dx = -\int_a^b v(x)\,\psi(x)\,dx,$$

то функция $v(x)$ называется обобщенной производной первого порядка от функции $u(x)$ на (a, b). Обозначается эта производная так, как обычная: $v(x) = \frac{du}{dx}$. Аналогично определяются обобщенные производные высших порядков.

Эквивалентным приведенному является следующее определение обобщенной производной. Функция $v(x)$ называется обобщенной производной первого порядка функции $u(x)$ на (a, b), если существует последовательность непрерывно дифференцируемых на (a, b) функций $u_m(x)$ таких, что

$$\int_{a'}^{b'} |u_m - u|\,dx \to 0, \quad \int_{a'}^{b'} \left|\frac{du_m}{dx} - v\right| dx \to 0$$

при $m \to \infty$, где (a', b') — строго внутренний подынтервал (a, b).

Заметим, что существование обобщенной производной на (a, b) равнозначно абсолютной непрерывности $u(x)$ на (a, b).

Перечислим конкретные банаховы сепарабельные пространства, используемые в дальнейшем.

Пространство $L_r(\Omega)$, $r \geqslant 1$, состоящее из всех функций $u(x)$, суммируемых на Ω со степенью r. Норма в этом пространстве определяется выражением

$$\|u\|_r = \left(\int_a^b |u|^r\,dx\right)^{1/r}.$$

Заметим, что под элементом пространства $L_p(\Omega)$ понимается не какая-либо одна функция $u(x)$, а весь класс функций, эквивалентных ей на Ω. Две функции $u_1(x)$, $u_2(x)$ называются эквивалентными на Ω, если $u_1(x) = u_2(x)$ для почти всех $x \in \Omega$.

Пространство $W_r^m(\Omega)$ состоит из элементов $L_r(\Omega)$, имеющих обобщенные производные до порядка m (включительно), суммируемые на Ω со степенью r. Норма определяется равенством

$$\|u\|_{r,m} = \left(\int_a^b \sum_{k=0}^{m} |u^{(k)}|^r \, dx \right)^{1/r}.$$

Пространство $C^m(\bar{\Omega})$ состоит из непрерывных в $\bar{\Omega} \equiv [a, b]$ функций $u(x)$, имеющих на $[a, b]$ непрерывные производные до порядка m включительно. Норма в $C^m(\bar{\Omega})$ определяется формулой

$$\|u\|_{C^m(\bar{\Omega})} = \max_{\substack{x \in [a,b] \\ 0 \leqslant k \leqslant m}} \left| \frac{d^k u}{dx^k} \right|.$$

Согласно теоремам вложения С. Л. Соболева всякая функция $u(x) \in W_r^m(\Omega)$, $\Omega = (a, b)$, оказывается и функцией из $C^{m-1}(\Omega)$.

Подпространство $\overset{0}{W}{}_r^m(\Omega)$ пространства $W_r^m(\Omega)$ состоит из функций, обращающихся в нуль на концах отрезка $\bar{\Omega}$ вместе со своими производными до порядка $m - 1$ включительно.

Частным случаем банахова пространства является пространство Гильберта. Абстрактное гильбертово пространство будем обозначать через H, а скалярное произведение любой пары его элементов u, v — через $(u, v)_H$; норма в H определяется формулой

$$\|u\|_H = \sqrt{(u, u)_H}.$$

Конкретными примерами гильбертовых пространств являются $L_2(\Omega)$, $W_2^m(\Omega)$, $\overset{0}{W}{}_2^m(\Omega)$, которые описаны выше как частные случаи соответствующих банаховых пространств. В дальнейшем, если это не будет вызывать путаницы, скалярное произведение в $L_2(\Omega)$ будем обозначать через $(\cdot, \cdot)$ без всяких индексов:

$$(u, v) \equiv (u, v)_{L_2(\Omega)} = \int_a^b u(x) v(x) \, dx$$

и соответственно

$$\|u\| \equiv \|u\|_{L_2(\Omega)} = \left(\int_a^b u^2 dx \right)^{1/2}.$$

Скалярное произведение в $W_2^m(\Omega)$ определяется формулой

$$(u, v)_{2,m} = \int_a^b \sum_{k=0}^{m} u^{(k)} v^{(k)} dx,$$

а норма —

$$\|u\|_{2,m} = \left(\int_a^b \sum_{k=0}^{m} (u^{(k)})^2 \, dx \right)^{1/2}.$$

Для прямоугольника Q_T, кроме гильбертовых пространств $L_2(Q_T)$ и $W_2^1(Q_T)$ со скалярными произведениями соответственно

$$(u, v)_{Q_T} = \int_{Q_T} uv dx dt \equiv \int_0^T \int_a^b uv dx dt$$

(норма $\|\cdot\|_{Q_T}$) и

$$(u, v)_{Q_T}^{1,1} = \int_{Q_T} \left(uv + \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial t} \frac{\partial v}{\partial t} \right) dx dt$$

(норма $\|\cdot\|_{Q_T}^{1,1}$), рассматриваются еще и следующие.

Подпространство $W_{2,0}^1(Q_T)$ пространства $W_2^1(Q_T)$, плотным множеством в котором являются гладкие функции, равные нулю вблизи сторон $x = a$, $x = b$. Иными словами, $W_{2,0}^1(Q_T)$ состоит из элементов $u \in W_2^1(Q_T)$, для которых $u(a, t) = u(b, t) = 0$ при $t \in [0, T]$. Гильбертово пространство $W_2^{1,0}(Q_T)$, состоящее из функций $u(x, t) \in \in (L_2(Q_T)$, имеющих обобщенные производные $\frac{\partial u}{\partial x} \in L_2(Q_T)$. Скалярное произведение в этом пространстве определяется равенством

$$(u, v)_{Q_T}^{1,0} = \int_{Q_T} \left(uv + \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} \right) dx dt,$$

а норма обозначается так: $\|\cdot\|_{Q_T}^{1,0}$.

Через $\overset{0}{W}{}_2^{1,0}$ будет обозначено подпространство пространства $W_2^{1,0}(Q_T)$, состоящее из функций $u(x, t) \in W_2^{1,0}(Q_T)$, равных нулю на сторонах $x = a$, $x = b$ прямоугольника Q_T.

Приведем ряд функциональных неравенств, которые будут неоднократно использоваться в дальнейшем.

Неравенство треугольника для любой пары u, v элементов пространства $\mathfrak{B}$

$$\|u + v\|_{\mathfrak{B}} \leqslant \|u\|_{\mathfrak{B}} + \|v\|_{\mathfrak{B}}.$$

В дальнейшем везде для обозначения понятия «любой» будем употреблять символ $\forall$.

Неравество Коши

$$|(u, v)_H| \leqslant \|u\|_H \|v\|_H \text{ при } \forall u, v \in H.$$

Неравенство Гельдера

$$\left| \int_\Omega uv dx \right| \leqslant \left(\int_\Omega |u|^p dx \right)^{1/p} \left(\int_\Omega |v|^q dx \right)^{1/q}$$

справедливо для $\forall u \in L_p(\Omega)$, $v \in L_q(\Omega)$ при $\frac{1}{p} + \frac{1}{q} = 1$, $p, q > 1$.

Пусть $\mathfrak{V}$ и $\mathfrak{V}_1$ — два пространства и M — множество элементов из $\mathfrak{V}$. Если каждому элементу $u \in M \subset \mathfrak{V}$ поставлен в соответствие элемент $w = Au \in \mathfrak{V}_1$, то принято считать, что на M определен оператор A, отображающий M в $\mathfrak{V}_1$. Множество M называют областью определения оператора A, и мы будем обозначать ее через $D(A)$. Множество $\Delta(A)$ всех элементов вида $w = Au$ называется областью значений оператора. Будем считать, что A действует в $\mathfrak{V}$, если $\mathfrak{V} = \mathfrak{V}_1$, иначе оператор A действует из $\mathfrak{V}$ в $\mathfrak{V}_1$.

В дальнейшем всегда будет предполагаться, что $D(A)$ — линейное множество и, как правило, оно плотно в $\mathfrak{V}$. Оператор A называется линейным, если для любых элементов u, $v \in D(A)$ и любых вещественных чисел λ, μ выполняется равенство

$$A(\lambda u + \mu v) = \lambda Au + \mu Av.$$

Отметим, что оператор, для которого элементы области значений $\Delta(A) \in \mathfrak{V}_1$ — вещественные числа, называют функционалом.

Приведем некоторые определения важнейших классов операторов (не только линейных).

1. Оператор A называют непрерывным в точке $u_0 \in D(A)$, если из $\|u_n - u_0\| \to 0$ ($u_n \in D(A)$) следует, что $\|Au_n - Au_0\| \to 0$. Иногда это записывается следующим образом:

$$\lim_{u_n \to u_0} Au_n = Au_0, \quad u_0 \in D(A), \quad u_n \in D(A).$$

Если оператор A непрерывен в каждой точке множества $M \subseteq D(A)$, то считается, что он непрерывен на M.

2. Оператор A называется ограниченным, если он преобразует каждое ограниченное множество из $D(A)$ в ограниченное множество из $\Delta(A)$.

Линейный непрерывный оператор A, отображающий $\mathfrak{V}$ в $\mathfrak{V}_1$, является ограниченным, и для него справедливо неравенство

$$\|Au\|_{\mathfrak{V}_1} \leqslant \|A\| \|u\|_{\mathfrak{V}}, \quad \forall u \in \mathfrak{V},$$

где число $\|A\|$ определяется соотношением $\|A\| = \sup_{\|u\|_{\mathfrak{V}}=1} \|Au\|_{\mathfrak{V}_1}$ и называется нормой оператора A. Верно и обратное утверждение: линейный ограниченный оператор — непрерывен.

3. Оператор A удовлетворяет условию Липшица при ограниченных аргументах, если для любого заданного $c > 0$ существует константа $K = K(c)$ такая, что

$$\|Au - Av\|_{\mathfrak{V}_1} \leqslant K(c) \|u - v\|_{\mathfrak{V}}$$

для любых u, $v \in D(A)$ таких, что $\|u\|_{\mathfrak{V}} \leqslant c$, $\|v\|_{\mathfrak{V}} \leqslant c$.

Напомним определение пространства, сопряженного к банахову пространству.

Пространство всех непрерывных линейных (и, следовательно, ограниченных) функционалов, определенных на банаховом пространстве $\mathfrak{V}$, называется сопряженным с $\mathfrak{V}$ банаховым пространством и обозначается $\mathfrak{V}^*$. За норму элемента $\varphi \in \mathfrak{V}^*$ принимают норму $\|\varphi\|$

функционала $\varphi(u)$, определенного на $D(\varphi) \subset \mathfrak{B}$, т. е.

$$\|\varphi\|_{\mathfrak{B}^*} \equiv \|\varphi\|_* = \sup_{\|u\|_{\mathfrak{B}}=1} |\varphi(u)| \equiv \|\varphi\|.$$

При этом, как указывалось ранее, для линейного ограниченного функционала $\varphi(u)$, $u \in D(\varphi) \subset \mathfrak{B}$, справедливо неравенство

$$|\varphi(u)| \leqslant \|\varphi\| \|u\|_{\mathfrak{B}}.$$

Значение линейного функционала $\varphi \in \mathfrak{B}^*$ в точке $u \in \mathfrak{B}$ будем записывать в виде как $\varphi(u)$, так и $\langle \varphi, u \rangle$. Очевидно,

$$|\langle \varphi, u \rangle| \leqslant \|\varphi\|_* \|u\|_{\mathfrak{B}}.$$

Поскольку $\mathfrak{B}^*$ является банаховым пространством, можно говорить о пространстве $(\mathfrak{B}^*)^* = \mathfrak{B}^{**}$, сопряженном с $\mathfrak{B}^*$. Для каждого элемента $u \in \mathfrak{B}$ существует единственный элемент $u^{**} \in \mathfrak{B}^{**}$, удовлетворяющий соотношению

$$\langle u^{**}, u^* \rangle = \langle u^*, u \rangle, \quad \forall\, u^* \in \mathfrak{B}^*,$$

при этом $\|u\|_{\mathfrak{B}} = \|u^{**}\|_{\mathfrak{B}^{**}}$.

Таким образом, каждый элемент $u \in \mathfrak{B}$ можно отождествить с указанным элементом $u^{**} \in \mathfrak{B}^{**}$ и рассматривать $\mathfrak{B}$ как подпространство пространства $\mathfrak{B}^{**}$, т. е. $\mathfrak{B} \subset \mathfrak{B}^{**}$. Если $\mathfrak{B} = \mathfrak{B}^{**}$, то банахово пространство $\mathfrak{B}$ называется рефлексивным.

Примером рефлексивного пространства является гильбертово пространство. Напомним в этой связи известную теорему Ф. Риса.

Теорема. *Любой линейный непрерывный функционал $l(u)$ в гильбертовом пространстве H имеет вид*

$$l(u) = (u, v)_H,$$

где v — некоторый элемент из пространства H, однозначно определяемый функционалом $l(u)$; при этом $\|l(u)\| = \|v\|$.

Приведем еще несколько определений различных классов операторов.

4. Оператор A, действующий из $\mathfrak{B}$ в $\mathfrak{B}^*$, называется монотонным на множестве $M \subset \mathfrak{B}$, если

$$\langle Au - Av, u - v \rangle \geqslant 0 \text{ для } \forall\, u,\ v \in M,$$

и строго монотонным, если выполняется строгое неравенство при всех $u \neq v$.

5. Оператор A называется сильно монотонным (с постоянной монотонности γ), если

$$\langle Au - Av,\ u - v \rangle \geqslant \gamma \|u - v\|^2_{\mathfrak{B}},\ \gamma > 0,\ u, v \in M.$$

Коснемся еще некоторых сведений о дифференцируемости абстрактных функций, в частности дифференцируемости операторов.

Пусть $f(t)$ — абстрактная функция вещественного аргумента $t \in [a, b]$, значения которой при каждом значении t являются элементами пространства $\mathfrak{B}$: $f(t) \in \mathfrak{B}$, $a \leqslant t \leqslant b$.

Функция $f(t)$ называется дифференцируемой в точке $t_0 \in [a, b]$, если существует такой элемент $\psi \in \mathfrak{B}$, что

$$\left\| \frac{1}{\Delta t} [f(t_0 + \Delta t) - f(t_0)] - \psi \right\| \to 0$$

при $\Delta t \to 0$. Элемент ψ называется производной функции $f(t)$ в точке t_0 и обозначается $\psi = f'(t_0)$.

Функция дифференцируема на отрезке $[a, b]$, если она дифференцируема в каждой его точке. Если при этом производная $f'(t)$ непрерывна, т. е. $\| f'(t + \Delta t) - f'(t) \| \to 0$ при $\Delta t \to 0$, то функция $f(t)$ называется непрерывно дифференцируемой.

Пусть имеется оператор A с областью определения $D(A)$, плотной в банаховом пространстве $\mathfrak{B}$, и пусть $u_0 \in D(A)$ — некоторая фиксированная точка. Если существует такой линейный оператор A_{u_0}, что при всех $v \in \mathfrak{B}$, для которых $u_0 + tv \in D(A)$, справедливо равенство

$$\lim_{t \to 0} \frac{1}{t} [A(u_0 + tv) - A(u_0)] = A_{u_0} v,$$

то A_{u_0} называется производной Гато оператора A в точке u_0.

Производная Гато аналогично определяется для функционала $F(u)$. Здесь важным для дальнейшего изложения (см. гл. V) оказывается понятие градиента функционала. Определим его.

Пусть в банаховом пространстве $\mathfrak{B}$ определен нелинейный функционал $F(u)$, область $D(F)$ определения которого линейна и плотна в $\mathfrak{B}$.

Пусть для элементов u из линейного множества $M \subset D(F)$ предел

$$\lim_{t \to 0} \frac{F(u + tv) - F(u)}{t} = F'_u(v)$$

есть линейный непрерывный (ограниченный) функционал над v. Тогда для $\forall u \in M$ производная $F'_u(v)$ является элементом сопряженного пространства $\mathfrak{B}^*$. Следовательно, можно написать

$$F_u(v) \equiv \langle F'_u, v \rangle = \langle Au, v \rangle, \quad D(A) = M,$$

где A — оператор, который ставит в соответствие каждому элементу $u \in M \subset \mathfrak{B}$ элемент $Au = F'_u(v) \in \mathfrak{B}^*$.

Оператор A, определенный указанным образом, называется градиентом функционала $F(u)$, а $F(u)$ — потенциалом оператора A.

Между ними существует соотношение [9]

$$F(u) = F(u_0) + \int_0^1 \langle A(u_0 + t(u - u_0)), u - u_0 \rangle \, dt,$$

справедливое, если $u, u_0 \in D(A)$.

Если нулевой элемент принадлежит $D(A)$, то, положив $u_0 = 0$, получим

$$F(u) = \int_0^1 \langle A(tu), u \rangle \, dt + \text{const}.$$

Приведем несколько определений, относящихся к свойствам функционала.

Функционал $F(u)$, заданный в нормированном пространстве, называется возрастающим, если $F(u) \to +\infty$, тогда и только тогда, когда $\|u\| \to \infty$.

Функционал $F(u)$ называется полунепрерывным снизу (сверху) в точке u_0, если по данному $\varepsilon > 0$ можно найти такое $\delta > 0$, что из $\|u - u_0\| < \delta$ следует $F(u) - F(u_0) > -\varepsilon$ ($F(u_0) - F(u) > -\varepsilon$). Это свойство можно сформулировать и так: функционал $F(u)$ в точке u_0 полунепрерывен снизу (сверху), если $\varliminf\limits_{u_n \to u_0} F(u_n) \geqslant F(u_0)$ ($\varlimsup\limits_{u_n \to u_0} F(u_n) \leqslant F(u_0)$). Функционал F называется слабо полунепрерывным снизу в точке u_0, если соотношение $\varliminf F(u_n) \geqslant F(u_0)$ справедливо при условии, что u_n слабо сходится к u_0. Функционал полунепрерывен на некотором множестве, если он полунепрерывен в любой точке этого множества.

Функционал $F(u)$ непрерывен в точке u_0, если по заданному $\varepsilon > 0$ можно найти такое $\delta > 0$, что при $\|u - u_0\| < \delta$ обязательно $|F(u) - F(u_0)| < \varepsilon$, и функционал непрерывен на некотором множестве, если он непрерывен в каждой точке этого множества. Функционал непрерывен тогда и только тогда, когда он одновременно полунепрерывен сверху и снизу.

Приведем теорему В. И. Казимирова [64], весьма полезную для установления полунепрерывности функционалов широкого класса. Здесь она формулируется для частного случая функционала

$$F(u, p) = \int_0^1 f(x, u, p)\, dx, \quad u = u(x), \quad p = \frac{du}{dx}.$$

Теорема 1.1. *Пусть*

$$F(u, p) = \int_0^1 f(x, u, p)\, dx,$$

где функция $f(x, u, p)$ определена при $x \in (0, 1)$ и при любых значениях переменных u, p. И пусть во всей области своего определения функция f обладает следующими свойствами:

1) *функция $f(x, u, p)$ — непрерывна вместе с производной $\frac{\partial f}{\partial p}$;*
2) *функция $f(x, u, p)$ — неотрицательна;*
3) *справедливо неравенство*

$$f(x, u, p_1) - f(x, u, p_2) - (p_1 - p_2)\frac{\partial f}{\partial p}(x, u, p_1) \geqslant 0$$

при любых $x \in (0, 1)$, u, p_1, p_2.

Если $u_n(x)$ сильно сходится к $u_0(x)$ в норме некоторого пространства $L_m(0, 1)$, $1 < m < \infty$, а $p_n(x)$ в том же пространстве слабо сходится к $p_0(x)$, то

$$\varliminf_{n \to \infty} F(u_n, p_n) \geqslant F(u_0, p_0).$$

Заметим, что неравенство п. 3) теоремы I.1 выполняется, если $\frac{\partial^2 f}{\partial p^2}(x, u, p) \geqslant 0$ при всех значениях аргументов $x \in (0, 1)$, u, p. Это очевидно из разложения функции $f(x, u, p)$ в ряд Тейлора по аргументу p.

Функционал F называется выпуклым на линейном множестве $M \subset D(F)$, если

$$F(u) + F(v) - 2F\left(\frac{u+v}{2}\right) \geqslant 0, \quad u, v \in M.$$

Функционал называется существенно выпуклым на M, если равенство выполняется лишь при $u = v$.

При исследовании свойств конкретного функционала часто оказывается полезной следующая теорема.

Теорема I.2. *Если градиент функционала F, т. е. $A = \operatorname{grad} F$, имеет производную A'_u, положительную при любом $u \in D(A)$, причем $D(A') \supset D(A)$ для любого $u \in D(A)$, то функционал F существенно выпуклый на множестве $D(A)$. Если, кроме того, $F(u)$ непрерывен, то этот функционал выпуклый на $D(F)$. Далее, если $F(u)$ непрерывен, а производная A'_u его градиента равномерно положительно ограничена снизу, т. е.*

$$\langle A'_u v, v\rangle \geqslant \gamma^2 \|v\|^2, \quad \gamma = \text{const} > 0,$$

где $\|\cdot\|$ — норма пространства, на котором определен функционал F, и $D(A') \supset D(A)$, то функционал $F(u)$ существенно выпуклый на множестве $D(F)$.

Доказательство данной теоремы см. в [64].

Наконец, напомним определение минимизирующей последовательности для функционала. Пусть d — точная нижняя граница ограниченного снизу функционала $F(u)$:

$$d = \inf_{u \in D(F)} F(u).$$

Последовательность $\{u_n\}$ функций, принадлежащих $D(F)$, называется минимизирующей для этого функционала, если

$$\lim_{n\to\infty} F(u_n) = d.$$

2. Положительно определенные операторы и энергетический метод. Линейный оператор A, действующий в вещественном гильбертовом пространстве H, называется симметричным, если область его определения $D(A)$ плотна в H и для любых элементов $u, v \in D(A)$ справедливо равенство

$$(Au, v) = (u, Av).$$

Симметричный оператор A называется положительно определенным, если для любой функции $u \in D(A)$ справедливо неравенство

$$(Au, u) \geqslant \gamma^2 \|u\|^2,$$

где γ — положительная постоянная. Если для $\forall\, u \in D(A)$ выполняется неравенство $(Au, u) \geqslant 0$, причем $(Au, u) = 0$ только тогда,

когда $u = 0$, то симметричный оператор A называется положительным. В дальнейшем при изложении материала параграфа I.2 мы следуем в основном работе [66].

С каждым положительно определенным оператором A можно связать некоторое гильбертово пространство, которое называют энергетическим пространством данного оператора. Это пространство будем обозначать в дальнейшем так: H_A. Строится H_A следующим образом. Каждой паре элементов u, v из множества $D(A)$ поставим в соответствие число $[u, v]_A$:

$$[u, v]_A = (Au, v), \quad \forall u, \ v \in D(A). \tag{I.16}$$

Нетрудно убедиться, что выражение (I.16) удовлетворяет всем аксиомам скалярного произведения. Приняв $[u, v]_A$ за скалярное произведение, множество $D(A)$ можно обычным способом, т. е. введя предельные элементы, пополнить до полного гильбертова пространства. Это пополненное пространство и есть энергетическое пространство H_A. Норма в нем определяется по общему правилу:

$$\|u\|_A^2 = [u, u]_A.$$

Величины $[u, v]_A$ и $\|u\|_A$ называют энергетическим скалярным произведением элементов u, v и энергетической нормой элемента u соответственно.

Доказано, что все элементы пространства H_A принадлежат также исходному пространству H. Множество элементов, образующих энергетическое пространство положительно определенного оператора, плотно в исходном пространстве.

Для всех элементов $u \in H_A$ справедливо соотношение

$$\|u\| \leqslant \frac{1}{\gamma} \|u\|_A.$$

В качестве иллюстрации приведем простой пример положительно определенного оператора A, действующего в $H = L_2(0, 1)$. Область определения его $D(A)$ состоит из функций $u(x) \in C^2[0, 1]$, удовлетворяющих условию $u(0) = u(1) = 0$, а действует оператор по формуле

$$Au = -\frac{d^2u}{dx^2}.$$

Нетрудно убедиться, что данный оператор действительно удовлетворяет всем требованиям, предъявляемым к положительно определенным операторам. Кроме того, можно показать, что энергетическое пространство этого оператора состоит из абсолютно непрерывных на $[0, 1]$ функций $u(x)$, первые производные которых суммируемы с квадратом на $[0, 1]$ и которые удовлетворяют условию $u(0) = u(1) = 0$. Иными словами, H_A состоит из тех же функций, что и пространство $\overset{0}{W}{}_2^1(0,1)$. Для данного H_A энергетическое скалярное произведение и энергетическая норма определяются по формулам

$$[u, v]_A = \int_0^1 \frac{du}{dx} \frac{dv}{dx} dx, \quad \|u\|_A^2 = \int_0^1 \left(\frac{du}{dx}\right)^2 dx.$$

Здесь необходимо подчеркнуть, что в данном случае функции пространства H_A удовлетворяют тем же краевым условиям $u(0) = u(1) = 0$, которым подчинялись функции из области определения $D(A)$ оператора A. Однако если в $L_2(0, 1)$ рассмотреть оператор $Bu = -\frac{d^2u}{dx^2}$, у которого $D(B)$ состоит из функций $u(x) \in C^2[0, 1]$, подчиненных краевым условиям

$$\frac{du}{dx} - \alpha u(x)\Big|_{x=0} = 0, \quad \frac{du}{dx} + \beta u(x)\Big|_{x=1} = 0, \quad \alpha > 0, \quad \beta > 0,$$

то окажется, что энергетическое пространство H_B этого положительно определенного оператора состоит из тех же функций, что и пространство $\overset{\circ}{W}{}_2^1(0, 1)$, причем функции данного энергетического пространства не обязаны удовлетворять никаким краевым условиям.

Краевые условия, которым обязательно удовлетворяют функции из области определения оператора и которым не обязательно должны удовлетворять функции из энергетического пространства, называются естественными для дифференциального оператора. Краевые условия, которым обязательно удовлетворяют функции из энергетического пространства, называют главными. (Подробнее о главных и естественных краевых условиях см. в [66].)

Для положительно определенного оператора A, действующего в гильбертовом пространстве H, квадратичный функционал

$$F(u) = (Au, u) - 2(u, f), \quad f \in H \tag{I.17}$$

называют функционалом энергии оператора A. Очевидно, $D(A) = D(F)$. Доказано, что задача о минимуме функционала энергии на множестве $D(A)$ эквивалентна задаче о решении операторного уравнения

$$Au = f. \tag{I.18}$$

Это устанавливает следующая теорема.

Теорема I.3. *Пусть A — положительно определенный оператор в гильбертовом пространстве H. Если уравнение (I.18) имеет решение, то это решение сообщает функционалу энергии (I.17) наименьшее значение. Обратно, элемент гильбертова пространства u_0, реализующий минимум функционала (I.17), удовлетворяет уравнению (I.18):*

$$Au_0 = f.$$

Указанный элемент u_0 может быть только один.

Таким образом, если одна из этих задач разрешима, то разрешима и другая, и элемент, удовлетворяющий одной из задач, удовлетворяет и другой. Однако из этого утверждения еще не следует существование решения этих задач. Более того, задача о минимуме функционала (I.17) может вообще не иметь решения, если $D(F) = D(A)$.

Однако область определения функционала (I.17) легко расширить на все энергетическое пространство H_A и представить функционал $F(u)$ в виде

$$F(u) = [u, u]_A - 2(u, f), \tag{I.19}$$

где $u \in H_A \subset H$.

Теперь можно искать минимум функционала $F(u)$ не в $D(A)$, а в H_A.

Оказывается, что в энергетическом пространстве существует один и только один элемент $u_0 \in H_A$, на котором функционал (I.19) достигает минимума. Элемент $u_0 \in H_A$, реализующий минимум этого функционала, называют обобщенным решением уравнения (I.18). При этом функционал (I.19) можно записать в виде

$$F(u) = \|u - u_0\|_A^2 - \|u_0\|_A^2.$$

Становится очевидным, что

$$\min_{u \in H_A} F(u) = -\|u_0\|_A^2.$$

Если обобщенное решение u_0 принадлежит области определения оператора A, то u_0 является классическим решением уравнения (I.18), т. е. решением в обычном смысле. Метод решения операторных уравнений (I.18), состоящий, в переходе к вариационной задаче о минимуме функционала (I.19), называют энергетическим методом.

3. Процесс Ритца. Процесс Ритца является одним из методов построения последовательности приближений к элементу, реализующему минимум функционала (I.19) в энергетическом пространстве H_A. Таким образом, этот метод позволяет построить приближение к обобщенному решению уравнения (I.18).

Для осуществления процесса Ритца выбирают последовательность координатных (базисных) элементов

$$\varphi_1, \varphi_2, \ldots, \varphi_n, \ldots, \tag{I.20}$$

удовлетворяющих следующим требованиям:

1) все элементы $\varphi_n \in H_A$;

2) при любом n элементы $\varphi_1, \varphi_2, \ldots, \varphi_n$ линейно независимы;

3) последовательность (I.20) полна в H_A.

Из первых N координатных элементов строят линейную комбинацию

$$u^N = \sum_{k=1}^{N} a_k \varphi_k \tag{I.21}$$

с произвольными числовыми коэффициентами a_k.

Для получения приближенного обобщенного решения в функционале (I.19) полагают $u = u^N$ и неизвестные коэффициенты a_k, $k = 1, 2, \ldots, N$, определяют из условий минимума функции N переменных

$$F(u^N) = F(a_1, a_2, \ldots, a_N).$$

Так как оператор A положительно определенный, можно показать, что условия

$$\frac{\partial F(u^N)}{\partial a_k} = 0, \quad k = 1, 2, \ldots, N, \tag{I.22}$$

являются необходимыми и достаточными для определения минимума функции $F(a_1, a_2, \ldots, a_N)$.

Систему линейных алгебраических уравнений (I.22) можно записать в виде

$$\sum_{j=1}^{N} [\varphi_j, \varphi_k]_A a_j = (f, \varphi_k), \quad k = 1, 2, \ldots, N, \tag{I.23}$$

а если координатные элементы принадлежат $D(A)$, то в виде

$$\sum_{j=1}^{N} (A\varphi_k, \varphi_j) a_j = (f, \varphi_k), \quad k = 1, 2, \ldots, N.$$

Таким образом, построение приближенного обобщенного решения задачи (I.18) сводится к решению системы линейных алгебраических уравнений. Так как элементы $\{\varphi_k\}_1^N$ линейно независимы, то определитель системы линейных алгебраических уравнений (определитель Грама) отличен от нуля и, следовательно, система однозначно разрешима.

Найдя коэффициенты a_k, $k = 1, 2, \ldots, N$, из системы (I.23) и подставив найденные значения в (I.21), получают элемент $(u^N)^*$, который называют приближенным решением уравнения (I.18) по Ритцу. Относительно этого решения справедлива следующая теорема [66].

Теорема I.4. *Если A — положительно определенный оператор, то приближенные по Ритцу решения уравнения (I.18) сходятся к точному обобщенному решению этого уравнения как в энергетической норме, так и в метрике исходного пространства.*

Таким образом, метод Ритца состоит в замене пространства H_A в вариационной задаче (I.19) последовательностью конечномерных подпространств, содержащихся в H_A и имеющих размерность N. Элемент u^N называют допустимой (пробной) функцией. На каждом подпространстве размерности N минимизация функционала приводит к решению системы линейных алгебраических уравнений. Число уравнений совпадает с размерностью подпространства.

Описанный процесс Ритца допускает следующую модификацию. Вместо последовательности координатных элементов (I.20) можно ввести последовательность наборов элементов

$$\begin{array}{l} \varphi_{11}, \varphi_{12}, \ldots, \varphi_{1k_1}; \\ \varphi_{21}, \varphi_{22}, \ldots, \varphi_{2k_2}; \\ \quad \ldots \ldots \ldots \\ \varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nk_n}, \\ \quad \ldots \end{array} \tag{I.24}$$

которые должны подчиняться условиям:

1) все элементы (I.24) принадлежат H_A;

2) для любого n элементы n-го набора $\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nk_n}$ линейно независимы;

3) для произвольного элемента u энергетического пространства H_A и любого числа $\varepsilon > 0$ существует такой номер $N(n, \varepsilon)$, что при любом $n > N$ можно отыскать постоянные $\alpha_1^{(n)}, \alpha_2^{(n)}, \ldots, \alpha_{k_n}^{(n)}$, удовлетворяю-

щие неравенству

$$\left\| u - \sum_{j=1}^{k_n} \alpha_j^{(n)} \varphi_{nj} \right\|_A < \varepsilon.$$

Приближение к решению в этом случае можно записать в виде

$$u^n = \sum_{j=1}^{k_n} a_j \varphi_{nj},$$

а коэффициенты a_j определяются, как и раньше, из условий минимума функционала (I.19). Результаты, касающиеся сходимости приближенных решений к точным, остаются справедливыми и в случае набора координатных элементов вида (I.24).

4. Основные понятия и теоремы о собственном спектре операторов. Пусть A — линейный оператор в гильбертовом пространстве H. Значения числового параметра λ, при которых существуют нетривиальные (отличные от нулевого) решения операторного уравнения

$$Au - \lambda u = 0, \tag{I.25}$$

называются собственными числами уравнения, а соответствующие им нетривиальные решения — собственными элементами. Собственные числа и собственные элементы уравнения (I.25) называют также собственными числами и собственными элементами оператора A. Задачу отыскания λ и нетривиальных решений уравнения (I.25) называют проблемой собственных значений.

Проблема собственных значений при определенных условиях может быть сформулирована как вариационная задача. Это возможно, например, когда оператор A — симметричный и полуограниченный снизу, т. е. удовлетворяет неравенству

$$(Au, u) \geqslant \gamma \| u \|^2,$$

где γ — вещественное, не обязательно положительное число. (Как известно, собственные числа симметричного оператора — вещественны).

Возможность вариационной постановки задачи на собственные значения для данного оператора определяется следующими двумя теоремами.

Теорема 1.5. *Пусть A — полуограниченный снизу симметричный оператор и пусть γ_0 — точная нижняя граница значений функционала*

$$\Phi(u) = \frac{(Au, u)}{(u, u)}. \tag{I.26}$$

Если существует элемент $u_0 \in D(A)$ такой, что

$$\Phi(u_0) = \frac{(Au_0, u_0)}{(u_0, u_0)} = \gamma_0,$$

то γ_0 есть наименьшее собственное число оператора A, а u_0 — соответствующий собственный элемент.

Теорема 1.6. *Пусть $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$ — непосредственно следующие друг за другом первые n собственных чисел симметричного полуограниченного снизу оператора A, а $u_1, u_2, \ldots, u_n$ — соответствующие*

им ортонормированные собственные элементы. Пусть также существует элемент $u = u_{n+1} \neq 0$, *реализующий минимум функционала (1.26) при дополнительных условиях*

$$(u, u_1) = 0, \ (u, u_2) = 0, \ \ldots, \ (u, u_n) = 0. \qquad (1.27)$$

Тогда u_{n+1} *есть собственный элемент оператора* A, *отвечающий собственному числу*

$$\lambda_{n+1} = \frac{(Au_{n+1}, u_{n+1})}{(u_{n+1}, u_{n+1})} = \Phi(u_{n+1}).$$

Это собственное число — ближайшее, следующее за λ_n.

Таким образом, теорема I.5 сводит задачу о нахождении наименьшего собственного значения λ_1 и соответствующего собственного элемента u_1 симметричного полуограниченного снизу оператора A (уравнения (I.25)) к вариационной задаче об отыскании минимума функционала (I.26):

$$\lambda_1 = \min_{u \in D(A)} \Phi(u) = \Phi(u_1),$$

а теорема I.6 сводит отыскание собственных значений λ_{n+1} и u_{n+1}, $n \geqslant 1$, к вариационной задаче определения элементов, реализующих минимум функционала (I.26) при дополнительных условиях (I.27). Положительно определенный оператор является симметричным и полуограниченным снизу, поэтому для него справедливы все результаты, сформулированные выше. Более того, для положительно определенных операторов, возможно, целесообразно введение понятия обобщенных собственных чисел и соответствующих им обобщенных собственных элементов. Это понятие вводится по аналогии с понятием обобщенного решения операторного уравнения $Au = f$ из энергетического пространства H_A.

Элемент $u \in H_A$, $u \neq 0$, и число λ называются обобщенным собственным элементом и обобщенным собственным числом положительно определенного оператора A, если они удовлетворяют тождеству

$$[u, \eta]_A = \lambda(u, \eta), \ \forall \ \eta \in H_A.$$

Вариационная формулировка задачи отыскания наименьшего обобщенного собственного числа λ_1 и соответствующего собственного элемента u_1 (если они существуют) положительно определенного оператора A может быть выражена соотношением

$$\lambda_1 = \min_{u \in H_A} \frac{[u, u]_A}{(u, u)} = \frac{[u_1, u_1]_A}{(u_1, u_1)}.$$

Аналогично если известны n первых обобщенных собственных чисел положительно определенного оператора A

$$\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$$

и соответствующие им попарно ортогональные собственные элементы

$$u_1, u_2, \ldots, u_n$$

и если существует элемент u_{n+1}, реализующий минимум функционала $R(u) = \frac{[u, u]_A}{(u, u)}$,

$$\lambda_{n+1} = \min_{u \in H_A^{(n)}} \frac{[u, u]_A}{(u, u)} = \frac{[u_{n+1}, u_{n+1}]_A}{(u_{n+1}, u_{n+1})} = R(u_{n+1}), \qquad (I.28)$$

то u_{n+1} является обобщенным собственным элементом оператора A, отвечающим собственному числу λ_{n+1}. Это собственное число непосредственно следует за λ_n. В (I.28) через $H_A^{(n)}$ обозначено подпространство пространства H_A, ортогональное в метрике H_A к $u_1, u_2, \ldots, u_n$. Нетрудно показать, что $H_A^{(n)} = H_A \cap H^{(n)}$, где $H^{(n)}$ — подпространство пространства H, ортогональное в метрике H к собственным элементам $u_1, u_2, \ldots, u_n$. (Заметим, что отношение $R(u) = \frac{[u, u]_A}{(u, u)}$ часто называют отношением Рэлея.)

Приведенные результаты подсказывают только способ построения собственных чисел, если существование их уже установлено. Формулируемая ниже теорема определяет достаточные условия существования обобщенных собственных значений положительно определенного оператора [68].

Теорема I.7. *Пусть положительно определенный оператор A, действующий в гильбертовом пространстве H, таков, что любое множество элементов, ограниченное в энергетической норме, компактно в H. Тогда оператор A:*

1) *имеет бесконечную последовательность обобщенных собственных чисел*

$$0 < \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n \leqslant \cdots$$

с единственной предельной точкой на бесконечности;

2) *соответствующие собственные элементы образуют систему, полную как в H, так и в H_A.*

Отметим, что условие данной теоремы можно сформулировать и так: энергетическое пространство H_A вкладывается в исходное пространство H вполне непрерывно.

Условимся в дальнейшем слово «обобщенные» для краткости опускать. Будем считать, что собственные элементы положительно определенного оператора A ортонормированы в исходном пространстве H. Тогда можно показать [68], что система собственных элементов ортогональна и в энергетическом пространстве H_A:

$$(u_i, u_j) = \delta_{ij}; \quad [u_i, u_j]_A = 0, \text{ если } i \neq j,$$

причем $\| u_i \|_A = \sqrt{\lambda_i}$, где λ_i — собственное число, отвечающее элементу u_i.

Важную роль в проблеме собственных значений играет минимаксимальный принцип (или принцип минимакса).

Пусть A — положительно определенный оператор, удовлетворяющий условию теоремы I.7. Если S_n есть n-мерное подпространство

пространства H_A, то собственное число λ_n определяется соотношением

$$\lambda_n = \min_{S_n} \max_{v \in S_n} R(v), \tag{I.29}$$

т. е. λ_n — наименьшее из максимальных значений $R(v)$ на S_n при всевозможных наборах n-мерных подпространств $S_n \subset H_A$. (Доказательство справедливости принципа минимакса можно найти в [101].)

Остановимся весьма кратко и на рассмотрении более общей проблемы собственных значений

$$Au - \lambda Bu = 0, \tag{I.30}$$

но только для случая, когда операторы A и B — положительно определенные и $D(A) \subset D(B) \subset H$. Эти операторы порождают соответствующие энергетические пространства H_A и H_B, так что для любого $u \subset D(A) \subset D(B)$ можно определить как $\|u\|_A$, так и $\|u\|_B$. Относительно собственных чисел и собственных функций операторного уравнения (I.30) имеется ряд теорем, аналогичных теоремам для уравнения $Au = \lambda u$. Сформулируем (без доказательства) некоторые из них [66].

Теорема I.8. *Собственные числа уравнения (I.30) вещественные (и положительные). Собственные элементы, отвечающие различным собственным числам, ортогональны в метрике H_B.*

Можно считать, что совокупность всех собственных элементов уравнения (I.30) ортонормирована в H_B. Нетрудно убедиться, что система собственных элементов ортогональна и в H_A.

Теорема I.9. *Пусть операторы A и B таковы, что всякое множество, ограниченное в H_A, компактно в H_B. Тогда*

1) *уравнение* (I.30) *имеет бесконечное множество собственных чисел*

$$0 < \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n \leqslant \cdots,$$

причем $\lambda_n \to \infty$ при $n \to \infty$;

2) *соответствующие собственные элементы образуют систему, ортонормированную в H_B, ортогональную в H_A и полную в обоих пространствах.*

Теорема I.10. *Пусть d есть точная нижняя грань функционала*

$$\Phi(u \qquad \frac{(Au, u)}{(Bu, u)}. \tag{I.31}$$

Если существует такой элемент u_0, что

$$\Phi(u_0) = \frac{(Au_0, u_0)}{(Bu_0, u_0)} = d,$$

то d есть наименьшее собственное число уравнения (I.30), а u_0 — отвечающий ему собственный элемент этого уравнения.

Теорема I.11. *Пусть $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$ есть n первых собственных чисел уравнения (I.30), а $u_1, u_2, \ldots, u_n$, — соответствующие им собственные элементы, ортонормированные в H_B. Пусть существует элемент u_{n+1}, реализующий минимум функционала (I.31) при дополнительных условиях*

$$(Bu, u_1) = 0, \ (Bu, u_2) = 0, \ \ldots, \ (Bu, u_n) = 0.$$

Тогда u_{n+1} есть собственный элемент уравнения (I.30), соответствующий собственному числу

$$\lambda_{n+1} = \Phi(u_{n+1}) = \frac{(Au_{n+1}, u_{n+1})}{(Bu_{n+1}, u_{n+1})},$$

это собственное число — ближайшее, следующее за λ_n.

Таким образом, теорема I.9 устанавливает существование собственных чисел и собственных элементов уравнения (I.30) при положительно определенных операторах A, B, а теоремы I.10 и I.11 указывают на способ их построения путем решения некоторых вариационных задач.

Для уравнения (I.30) также возможно введение понятия обобщенных собственных чисел и соответствующих им обобщенных собственных векторов. Построение этих обобщенных собственных значений осуществляется посредством решения следующих вариационных задач.

Найти минимум функционала $R(u) = \frac{[u, u]_A}{[u, u]_B}$, $D(R) = H_A$, т. е. найти

$$\lambda_1 = \min_{u \in H_A} R(u) = \frac{[u_1, u_1]_A}{[u_1, u_1]_B},$$

если требуется построить наименьшее собственное число λ_1 и собственный элемент u_1.

Если известны собственные элементы $u_1, u_2, \ldots, u_n$, ортонормированные в H_B, а требуется построить $(n+1)$-е обобщенное собственное число λ_{n+1} и собственный элемент u_{n+1}, то задача формулируется так: найти минимум функционала

$$R(u) = \frac{[u, u]_A}{[u, u]_B}$$

при дополнительных условиях

$$[u, u_1]_B = 0, \ [u, u_2]_B = 0, \ \ldots, \ [u, u_n]_B = 0, \qquad \text{(I.32)}$$

т. е.

$$\lambda_{n+1} = \min_{u \in H_{AB}^{(n)}} \frac{[u, u]_A}{[u, u]_B} = \frac{[u_{n+1}, u_{n+1}]_A}{[u_{n+1}, u_{n+1}]_B},$$

где $H_{AB}^{(n)}$ — подпространство пространства H_A, элементы которого удовлетворяют условиям (I.32).

5. Процесс Рэлея — Ритца в проблеме собственных значений. Для численного решения проблемы собственных значений, сформулированной в вариационной форме, с успехом применяется процесс Ритца. (Свой метод решения вариационных задач Ритц опубликовал в 1908 г. Этот метод является обобщением метода Рэлея, используемого ранее для решения некоторых задач на собственные значения. Поэтому, если речь идет о решении проблемы собственных значений, данный метод (процесс) часто называют процессом Рэлея — Ритца.)

Рассмотрим вначале применение процесса Ритца к решению проблемы собственных значений для уравнения

$$Au - \lambda u = 0,$$

где A — положительно определенный оператор, имеющий бесконечное множество собственных чисел и собственных элементов. В этом случае задача об отыскании наименьшего обобщенного собственного значения λ_1 сводится к отысканию минимума функционала

$$R(u) = \frac{[u, u]_A}{(u, u)}, \quad D(R) = H_A.$$

Для построения процесса Ритца в данной ситуации выбирается последовательность координатных (базисных) функций φ_n, $n = 1, 2, \ldots$, подчиненных тем же условиям, что и система (I.20):

1) $\varphi_n \in H_A$, $n = 1, 2, \ldots$;

2) при любом n элементы $\varphi_1, \varphi_2, \ldots, \varphi_n$ — линейно независимы;

3) система $\{\varphi_n\}$ полна в H_A.

Приближенное решение u^N вариационной задачи ищут в виде

$$u^N = \sum_{k=1}^{N} a_k \varphi_k,$$

где a_k — неизвестные числовые коэффициенты, которые выбираются так, чтобы

$$(u^N, u^N) = \sum_{k,m=1}^{N} a_k a_m (\varphi_k, \varphi_m) = 1, \tag{I.33}$$

а функционал

$$[u^N, u^N]_A = \sum_{k,m=1}^{N} a_k a_m [\varphi_k, \varphi_m]_A \tag{I.34}$$

принимал минимальное значение. Иными словами, дело сводится к отысканию минимума функции $[u^N, u^N]_A$, зависящей от N переменных a_i, $i = 1, 2, \ldots, N$, связанных дополнительным условием (I.33). Для решения этой задачи на условный минимум применяется метод неопределенных множителей Лагранжа (см. [112]) и составляется функция

$$F(a_1, a_2, \ldots, a_N) = [u^N, u^N]_A - \lambda (u^N, u^N), \tag{I.35}$$

где λ — неопределенный числовой множитель.

Приравнивая нулю частные производные функции F по коэффициентам a_m, $m = 1, 2, \ldots, N$, получаем систему уравнений

$$\sum_{k=1}^{N} a_k ([\varphi_k, \varphi_m]_A - \lambda (\varphi_k, \varphi_m)) = 0, \quad m = 1, 2, \ldots, N,$$

или в матричной форме

$$Ka = \lambda Ma, \tag{I.36}$$

где

$$K = \begin{pmatrix} [\varphi_1, \varphi_1] & [\varphi_2, \varphi_1] & \dots & [\varphi_N, \varphi_1] \\ [\varphi_1, \varphi_2] & [\varphi_2, \varphi_2] & \dots & [\varphi_N, \varphi_2] \\ \dots & \dots & \dots & \dots \\ [\varphi_1, \varphi_N] & [\varphi_2, \varphi_N] & \dots & [\varphi_N, \varphi_N] \end{pmatrix},$$

$$M = \begin{pmatrix} (\varphi_1, \varphi_1), & (\varphi_2, \varphi_1), & \dots, & (\varphi_N, \varphi_1) \\ (\varphi_1, \varphi_2), & (\varphi_2, \varphi_2), & \dots, & (\varphi_N, \varphi_2) \\ \dots & \dots & \dots & \dots \\ (\varphi_1, \varphi_N), & (\varphi_2, \varphi_N), & \dots, & (\varphi_N, \varphi_N) \end{pmatrix},$$

$$a = [a_1, a_2, \dots, a_N]^T.$$

Здесь для удобства опущен индекс «A» при обозначении энергетического скалярного произведения $[\cdot, \cdot]_A$. Однородная система (I.36) в силу условия (I.33) должна иметь нетривиальные решения, поэтому ее определитель должен равняться нулю, что и дает уравнение для вычисления значений параметра λ:

$$\det(K - \lambda M) = 0. \tag{I.37}$$

Характеристическое уравнение (I.37) является уравнением N-й степени, так как коэффициент при $(-1)^N\lambda^N$ есть определитель Грама функций $\varphi_1, \varphi_2, \dots, \varphi_N$, а они согласно условию 2) линейно независимы при любом N. Отсюда следует, что уравнение (I.37) имеет N корней. Для каждого корня λ_i, $i = 1, 2, \dots, N$, из системы (I.36) можно вычислить решение

$$a^{(i)} = [a_1^{(i)}, a_2^{(i)}, \dots, a_N^{(i)}]^T,$$

удовлетворяющее условию (I.33), или, точнее, по этому решению можно построить функцию $u_i^N = \sum_{k=1}^{N} a_k^{(i)}\varphi_k$, удовлетворяющую условию (I.33):

$$(u_i^N, u_i^N) = \sum_{k,m=1}^{N} a_k^{(i)} a_m^{(i)} (\varphi_k, \varphi_m) \equiv (a^{(i)})^T M a^{(i)} = 1. \tag{I.38}$$

Покажем, что все корни λ_i уравнения (I.37) положительные. Для этого подставим в (I.36) $\lambda = \lambda_i$, $a = a^{(i)}$ и получим равенство

$$Ka^{(i)} = \lambda_i M a^{(i)},$$

которое с учетом (I.38) преобразуем к виду

$$(a^{(i)})^T K a^{(i)} = \lambda_i (a^{(i)})^T M a^{(i)} = \lambda_i.$$

А так как

$$(a^{(i)})^T K a^{(i)} \equiv \sum_{k,m=1}^{N} [\varphi_k, \varphi_m]_A a_k^{(i)} a_m^{(i)} = [u_i^N, u_i^N]_A,$$

то

$$[u_i^N, u_i^N]_A = \lambda_i, \quad i = 1, 2, \dots, N. \tag{I.39}$$

Из выражения (I.39) следует, что все корни λ_i уравнения (I.37) действительно положительны.

Среди всех функций u_i^N, построенных на основе решений системы (I.36), выберем ту, которая доставляет минимум выражению (I.34). Согласно (I.39) этот минимум является наименьшим корнем уравнения (I.37). Обозначим этот корень через $\lambda_1^{(N)}$, а соответствующую ему функцию — через u_1^N. С возрастанием N значение $\lambda_1^{(N)}$ не возрастает, но и не может стать меньше

$$\lambda_1 = \min_{u \in H_A} \frac{[u, u]_A}{(u, u)}.$$

В работе [66] доказано, что $\lim\limits_{N \to \infty} \lambda_1^{(N)} = \lambda_1$, а это показывает правомерность использования процесса Ритца для построения приближений к наименьшему собственному числу уравнения $Au - \lambda u = 0$, или, что то же, к наименьшему собственному числу оператора A.

Остановимся кратко на определении последующих собственных чисел оператора A. Для получения приближенного значения второго собственного числа λ_2 находят минимум энергетического скалярного произведения (I.34) при дополнительных условиях (I.33) и

$$(u_1^N, u^N) = \sum_{k,m=1}^{N} a_k^{(1)} a_m (\varphi_k, \varphi_m) \equiv (a^{(1)})^T Ma = 0, \qquad \text{(I.40)}$$

где $u_1^N = \sum\limits_{k=1}^{N} a_k^{(1)} \varphi_k$ — приближенное значение первой нормированной по (I.33) собственной функции оператора A.

В этом случае по методу неопределенных множителей Лагранжа составляется вспомогательная функция

$$\Phi(a_1, a_2, \ldots, a_N) = [u^N, u^N]_A - \lambda (u^N, u^N) - 2\mu (u^N, u_1^N) \equiv$$
$$\equiv a^T K a - \lambda a^T M a - 2\mu a^T M a^{(1)},$$
$$a = [a_1, a_2, \ldots, a_N]^T,$$

все частные производные по a_k, $k = 1, 2, \ldots, N$, которой приравниваются к нулю. В результате получается система

$$(K - \lambda M) a = \mu M a^{(1)}, \qquad \text{(I.41)}$$

из которой должны определяться λ_2^N и коэффициенты

$$a = [a_1, a_2, \ldots, a_N]^T$$

для второй приближенной собственной функции u_2^N.

Умножив обе части матричного уравнения (I.41) на вектор $a^{(1)} = [a_1^{(1)}, a_2^{(1)}, \ldots, a_N^{(1)}]^T$, с учетом условия (I.38) при $i = 1$ получим

$$(a^{(1)})^T (K - \lambda M) a = \mu.$$

А так как $(a^{(1)})^T K = \lambda_1 (a^{(1)})^T M$, в силу условия (I.40)

$$(a^{(1)})^T (K - \lambda M) a = (\lambda_1 - \lambda)(a^{(1)})^T Ma = 0,$$

следовательно, множитель μ равен нулю и система (I.41) совпадает с системой (I.36).

Таким образом, как и в случае первого собственного числа, приходим к выводу, что искомый минимум равен некоторому λ, являющемуся корнем уравнения (I.37). Однако здесь нужно взять уже второй по величине корень. Аналогично приближения к последующим (до N-го) собственным числам оператора A можно найти, определяя корни уравнения (I.37). Следует однако подчеркнуть, что достаточно хорошие приближения на практике получаются только для нескольких первых чисел. (Вопрос о точности получаемых приближений будет рассмотрен подробнее дальше в применении к МКЭ.)

В заключение отметим, что процесс Рэлея — Ритца можно применять и для вычисления собственных чисел и собственных элементов уравнения

$$Au - \lambda Bu = 0, \tag{I.42}$$

где A и B — положительно определенные операторы, удовлетворяющие условиям теоремы I.9. В результате искомые приближения к собственным числам и собственным элементам уравнения (I.42) вычисляются из системы

$$\sum_{k=1}^{N} a_k^N ([\varphi_k, \varphi_m]_A - \lambda [\varphi_k, \varphi_m]_B) = 0, \ m = 1, 2, \ldots, N, \ \varphi_m \in H_A,$$

и характеристического уравнения

$$\begin{vmatrix} [\varphi_1, \varphi_1]_A - \lambda [\varphi_1, \varphi_1]_B & [\varphi_2, \varphi_1]_A - \lambda [\varphi_2, \varphi_1]_B & \ldots & [\varphi_N, \varphi_1]_A - \lambda [\varphi_N, \varphi_1]_B \\ [\varphi_1, \varphi_2]_A - \lambda [\varphi_1, \varphi_2]_B & [\varphi_2, \varphi_2]_A - \lambda [\varphi_2, \varphi_2]_B & \ldots & [\varphi_N, \varphi_2]_A - \lambda [\varphi_N, \varphi_2]_B \\ \ldots & \ldots & \ldots & \ldots \\ [\varphi_1, \varphi_N]_A - \lambda [\varphi_1, \varphi_N]_B & [\varphi_2, \varphi_N]_A - \lambda [\varphi_2, \varphi_N]_B & \ldots & [\varphi_N, \varphi_N]_A - \lambda [\varphi_N, \varphi_N]_B \end{vmatrix} = 0.$$

Для приближенного нахождения собственных значений можно использовать и модифицированный процесс Ритца (см. п. 3 параграфа I.2).

6. Метод Бубнова — Галеркина. Этот метод не является вариационным и применяется для нахождения приближенного решения задач, оператор которых не обязательно положительный.

Пусть требуется решить уравнение

$$Au = f,$$

где линейный оператор A определен на множестве $D(A)$, плотном в некотором гильбертовом пространстве H.

Для этого выбирают последовательность базисных элементов $\{\varphi_k\}$, обладающих такими свойствами:

1) $\varphi_k \in D(A)$;

2) элементы $\varphi_1, \varphi_2, \ldots, \varphi_n$ линейно независимы при любом n;

3) система базисных элементов полна в H.

Приближенное решение уравнения (I.18) строят в виде линейной комбинации координатных элементов $u^N = \sum_{k=1}^{N} a_k \varphi_k$ с постоянными коэффициентами a_k.

Коэффициенты a_k определяются из условия, чтобы выражение $Au^N - f$ было ортогонально в H ко всем элементам $\varphi_1, \varphi_2, \ldots, \varphi_N$. Таким образом получают систему линейных алгебраических уравнений относительно неизвестных a_k:

$$\sum_{k=1}^{N} (A\varphi_k, \varphi_j)_H a_k = (f, \varphi_j)_H, \quad j = 1, 2, \ldots, N.$$

Точно так же задачи о собственных значениях для уравнения $Au - \lambda Bu = 0$ с помощью процесса Галеркина приводятся к системе уравнений

$$\sum_{k=1}^{N} a_k [(A\varphi_k, \varphi_j) - \lambda (B\varphi_k, \varphi_j)] = 0, \quad j = 1, 2, \ldots, N,$$

характеристическое уравнение которой

$$\begin{vmatrix} (A\varphi_1, \varphi_1) - \lambda (B\varphi_1, \varphi_1) & (A\varphi_1, \varphi_2) - \lambda (B\varphi_1, \varphi_2) & \ldots & (A\varphi_1, \varphi_N) - \lambda (B\varphi_1, \varphi_N) \\ (A\varphi_2, \varphi_1) - \lambda (B\varphi_2, \varphi_1) & (A\varphi_2, \varphi_2) - \lambda (B\varphi_2, \varphi_2) & \ldots & (A\varphi_2, \varphi_N) - \lambda (B\varphi_2, \varphi_N) \\ \ldots & \ldots & \ldots & \ldots \\ (A\varphi_N, \varphi_1) - \lambda (B\varphi_N, \varphi_1) & (A\varphi_N, \varphi_2) - \lambda (B\varphi_N, \varphi_2) & \ldots & (A\varphi_N, \varphi_N) - \lambda (B\varphi_N, \varphi_N) \end{vmatrix} = 0$$

позволяет определить приближенные значения собственных чисел.

Отметим, что процессы Ритца и Бубнова — Галеркина совпадают в случае положительно определенного оператора, однако метод Бубнова — Галеркина имеет более широкую область применения как для уравнений с неположительными операторами, так и для уравнений, описывающих нестационарные задачи. Применим этот метод и в случае нелинейных дифференциальных уравнений. Сходимость этого метода рассматривается, например, в работах [16, 59, 66].

7. Некоторые трудности численной реализации. В работе [64] рассматривались вопросы численной реализации метода Ритца. Для уменьшения погрешности приближенного решения задач приходится увеличивать число координатных элементов, что приводит к увеличению порядка соответствующих систем линейных алгебраических уравнений, причем матрицы в этих системах плотные.

В случае неудачного выбора координатных функций (близких к линейно зависимым) обусловленность матрицы системы может оказаться плохой. В результате найденное приближенное решение будет искажено за счет как неточного вычисления коэффициентов системы (наследственная погрешность), так и ошибок округления, допущенных в ходе решения систем линейных алгебраических уравнений на ЭВМ. Соответствующие трудности имеются и при машинной реализации задачи методом Бубнова — Галеркина.

1.3. Некоторые общие вопросы метода конечных элементов

1. Метод конечных элементов как средство дискретизации математических задач. С математической точки зрения МКЭ представляет собой обобщение методов Рэлея, Ритца, Бубнова — Галеркина. Это обобщение заключается в специальном выборе базисных (координатных) функций $\varphi_j(x)$, каждая из которых имеет конечный носитель, т. е. отлична от нуля только на небольшой части всей области определения Ω решаемой задачи. Кроме того, каждая $\varphi_j(x)$ является полиномиальной функцией, поэтому все вычисления становятся относительно простыми. Дискретизация вариационной задачи методом конечных элементов начинается с разбиения исходной области Ω на подобласти — элементы. Для удобства задания информации об этих подобластях и обеспечения требуемой гладкости допустимых функций используются сравнительно простые с геометрической точки зрения подобласти: треугольники, прямоугольники — в двумерной области, тетраэдры, параллелепипеды — в трехмерной, отрезки — в одномерной. В результате область Ω представляется в виде объединения отдельных элементов (отрезков, многоугольников, многогранников) Ω_i, $i = 1, 2, \ldots, N$, соседние из которых имеют общие точки, стороны или грани (общие стороны или грани смежных элементов должны быть одинаковыми).

Одно из возможных разбиений заданной трапециевидной области Ω на треугольные элементы (подобласти) показано на рис. 1.

Наряду с разбиением области Ω на элементы строится подходящий для заданной вариационной задачи класс допустимых функций $v^N(x)$, $x \in \Omega$, метода конечных элементов. В качестве таких допустимых функций выбираются кусочные полиномы, которые на каждом элементе подобласти являются полиномами заданной степени с неизвестными коэффициентами, а на всей области Ω обладают гладкостью, предписанной исходной вариационной задачей.

Однозначность определения полинома на каждой подобласти обеспечивается тем, что в заданных узловых точках подобласти фиксируются в качестве неизвестных параметров значения полинома или значения полинома и некоторых его частных производных. Требуемая гладкость допустимой функции на всей области Ω обеспечивается тем, что значения соответствующих параметров в общих узловых точках смежных подобластей совпадают.

Отметим, что для полного описания конечного элемента, используемого при решении задачи, необходимо указывать как геометрическую форму подобластей (элементов), на которые разбивается область, так и вид допустимой функции на элементе.

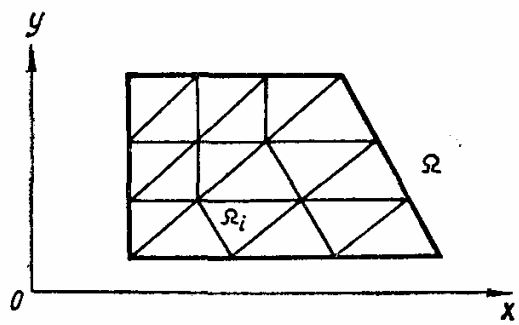

Рис. 1.

В результате дискретизации области и выбора типа конечного элемента вся область, включая границу, оказывается покрытой сеткой, в узлах которой будут определяться значения искомого приближенного ре-

шения и, возможно, его производных. Множество допустимых функций МКЭ является конечномерным, и размерность его определяется общим количеством неизвестных коэффициентов кусочного полинома (или, что то же, общим количеством неизвестных параметров, фиксированных во всех узлах области Ω). Суть МКЭ как приближенного метода решения математических задач состоит в замене бесконечномерного функционального пространства, которому принадлежит решение исходной задачи, конечномерным подпространством допустимых функций МКЭ, содержащимся в исходном функциональном пространстве. При этом в качестве приближенного решения МКЭ принимается та допустимая функция МКЭ, параметры которой находятся из вариационного принципа или некоторого интегрального тождества, лежащего в основе постановки исходной задачи (вариационной или дифференциальной). В результате исходная задача сводится к системе дискретных алгебраических уравнений, решением которой служат искомые параметры (коэффициенты) приближенного решения. Поскольку для построения и решения соответствующих дискретных систем вследствие большого объема перерабатываемой информации приходится применять ЭВМ, весьма важным явилось условие удобства и простоты вычислений, что и определило соответствующий выбор пространства допустимых функций МКЭ, а именно кусочно-полиномиальных. Еще более важен вопрос о точности, с которой допустимые функции могут аппроксимировать искомое решение исходной задачи. Математическое исследование метода показало, что кусочно-полиномиальные функции МКЭ при достаточной гладкости искомого решения могут обеспечить построение приближенного решения любой точности, если ввести достаточное число подобластей-элементов или при заданном разбиении использовать полиномы повышенной степени.

По теоретическому обоснованию МКЭ имеется множество работ (см., например, [48, 67, 83, 84, 101, 102], где представлена обширная библиография). Основные результаты по данному вопросу будут освещены в последующих главах.

Необходимо еще кратко остановится на связях и сравнении МКЭ с методом конечных разностей, этих наиболее распространенных и эффективных численных методов. Как известно, построение конечно-разностных схем обычно требует небольшого объема вычислений, как правило, меньшего, чем в МКЭ. Однако достоинствами МКЭ являются гибкость и разнообразие сеток, стандартные приемы построения дискретных задач для произвольных областей, простота учета естественных краевых условий и т. д. Кроме того, математический анализ МКЭ является более простым, его методы применимы к более широкому классу исходных задач, а оценки погрешностей приближенных решений, как правило, получаются при менее жестких ограничениях, чем в методе конечных разностей. Вместе с тем необходимо подчеркнуть, что основу для исследования МКЭ создали фундаментальные результаты, связанные с исследованием сходимости и устойчивости конечно-разностных схем, проекционных методов, обобщенных решений.

Следует отметить и использование для решения конечно-элементных систем алгебраических уравнений широкого класса эффективных

итерационных методов, разработанных для конечно-разностных уравнений [26, 27, 50, 59, 94—96]. Такая возможность обусловлена тем, что в ряде случаев схемы МКЭ можно строить так, что они будут эквивалентны по спектру определенным схемам метода конечных разностей (см., например, [48, 83] и представленную там библиографию, а также [27—32, 34]).

Подытоживая изложенное, можно выделить следующие этапы решения задач методом конечных элементов:

1. Постановка задачи, т. е. определение понятия разыскиваемого решения (обобщенное или классическое), и выбор варианта МКЭ, основанного на процессе Ритца при вариационной формулировке задачи или на методе Галеркина при определении обобщенного решения через интегральные тождества.

2. Разбиение на элементы области, в которой строится решение.

3. Построение пространства кусочно-полиномиальных допустимых функций МКЭ.

4. Формирование и решение системы дискретных (алгебраических) уравнений МКЭ.

5. Оценка точности получаемого машинного решения задачи.

При практической реализации МКЭ некоторые этапы могут быть совмещены или опущены. Изложим некоторые особенности реализации этих этапов.

2. Дискретизация области, пространства допустимых функций МКЭ, алгебраические системы МКЭ. Разбиение на элементы области Ω, в которой строится решение задачи, является весьма важным этапом, определяющим эффективность реализации МКЭ. Успех дискретизации области зависит от знания физики исследуемого явления, общих особенностей поведения искомого решения, а также от учета влияния типа сетки на эффективность численного решения получаемой дискретной задачи.

Дискретизация области включает задание числа, размеров и формы элементов, которые используются для построения дискретной модели исследуемого объекта. Для обеспечения желаемой точности искомого решения элементы должны быть достаточно малыми, чтобы аппроксимируемые поля в них достаточно хорошо приближались полиномами. Однако малость элементов, а следовательно, и возрастание их количества приводят к дискретным задачам больших размерностей, что в любом случае увеличивает объем вычислительной работы. Кроме того, чрезмерное измельчение сетки ухудшает обусловленность дискретной задачи и может помешать получению машинного решения с требуемой точностью. В связи с этим целесообразно на основе общих соображений о поведении решения задачи измельчать сетку лишь в тех подобластях, где искомое решение сильно меняется (большие значения градиентов), и использовать достаточно крупные элементы там, где решение почти постоянно.

Необходимо подчеркнуть, что от вида разбиения области на элементы существенно зависит возможность использования эффективных численных методов для решения дискретных задач, а следовательно, и минимизация вычислительной работы. В работах [27, 28, 30] рассмот-

рены специальные триангуляции Ω, топологически эквивалентные простейшей триангуляции идеальных областей Q (прямоугольник, прямоугольный треугольник, параллелепипед и т. п.). Эти триангуляции, с одной стороны, обеспечивают обычные оценки точности для приближенных решений МКЭ, а с другой — обеспечивают нахождение операторов, позволяющих строить эффективные итерационные процессы решения соответствующих дискретных задач. Применение таких специальных триангуляций в сочетании с идеей использования последовательностей сгущающихся сеток ведет к асимптотической минимизации вычислительной работы в МКЭ. Например, решение $u \equiv \equiv (u_1, \ldots, u_s)^T \in (W_2^{1+m}(\Omega))^s$, $m > 0$, краевой задачи для сильно-эллиптической системы второго порядка в многоугольнике Ω можно найти с точностью ε в метрике $(W_2^1(\Omega))^s$ при $O(\varepsilon^{-\frac{2}{m}} |\ln \varepsilon|)$ арифметических операциях. (Подробнее о триангуляциях, обеспечивающих асимптотическую минимизацию вычислений, и о применении эффективных вычислительных методов см. в работах [27—32, 34].)

Как уже упоминалось, при решении задач методом конечных элементов используются элементы различных типов. Простейшим среди элементов является одномерный элемент. Наиболее часто такой элемент применяется в одномерных задачах распространения тепла, в задачах строительной механики при расчетах стержневых элементов конструкций (типа ферм) и др. (Подробное описание одномерных элементов дается в гл. II.)

Для построения дискретной модели двумерной области наиболее часто используются треугольные элементы с кусочно-линейной допустимой функцией, которая на каждом треугольнике имеет вид

$$v^N = \alpha_1 + \alpha_2 x + \alpha_3 y,$$

и прямоугольники с кусочно-билинейной допустимой функцией вида

$$v^N = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 xy$$

на каждом прямоугольнике. В качестве неизвестных фиксированных параметров выступают значения $v_i = v^N(x_i, y_i)$ соответствующего

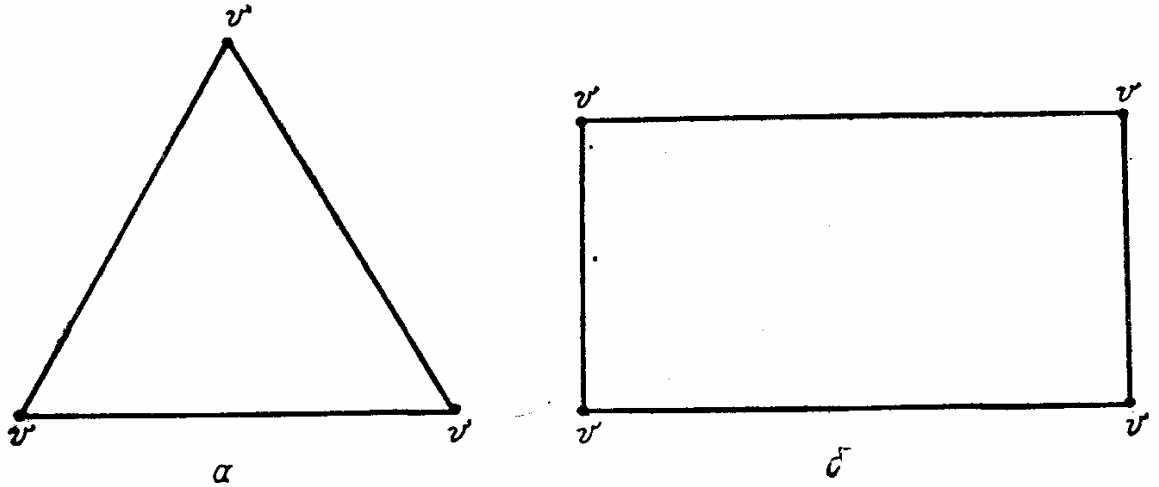

Рис. 2.

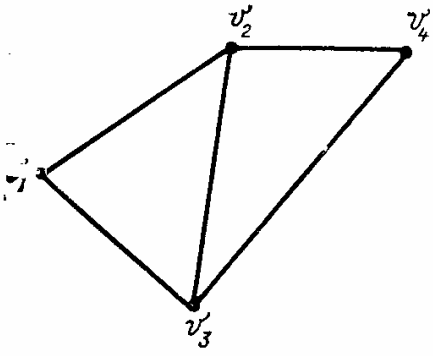

Рис. 3.

полинома $v^N(x, y)$ в узловых точках (x_i, y_i), расположенных в вершинах указанных элементов. Схематически такие конечные элементы можно изображать так, как это показано на рис. 2, *а*, *б*.

Нетрудно убедиться, что при указанном расположении узлов и выборе фиксированных параметров будут обеспечены однозначное определение каждой допустимой функции и непрерывность $v^N(x, y)$ на всех границах смежных элементов, т. е. на всей области Ω. Например, в случае двух смежных треугольников (рис. 3) функция $v^N(x, y)$ будет непрерывной при переходе через общую границу Γ, ибо она на этой границе может быть представлена как полином первой степени от одного параметра s, а следовательно, будет однозначно определяться двумя фиксированными значениями $v_2 = v^N(x_2, y_2)$, $v_3 = v^N(x_3, y_3)$ в конечных (узловых) точках общей стороны.

Таким образом, в данном случае допустимые функции $v^N(x, y)$ принадлежат пространству $W_2^1(\Omega)$.

Если в качестве допустимых функций используют квадратичные или биквадратичные полиномы, то схематически такой элемент с фиксированными узловыми параметрами можно представить в виде, изображенном на рис. 4, *а*, *б*.

Рассмотрим еще случай, когда в качестве допустимых функций МКЭ выбраны кусочно-кубические полиномы, которые для каждого треугольника записываются в виде

$$v(x, y) = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2 + \\ + \alpha_7 x^3 + \alpha_8 x^2 y + \alpha_9 xy^2 + \alpha_{10} y^3.$$

Неизвестные коэффициенты α_i этого полинома можно однозначно определить на треугольнике одним из следующих двух условий:

1) в качестве узловых точек выбираются вершины, центр тяжести треугольника и точки, которые делят каждую сторону на три равные

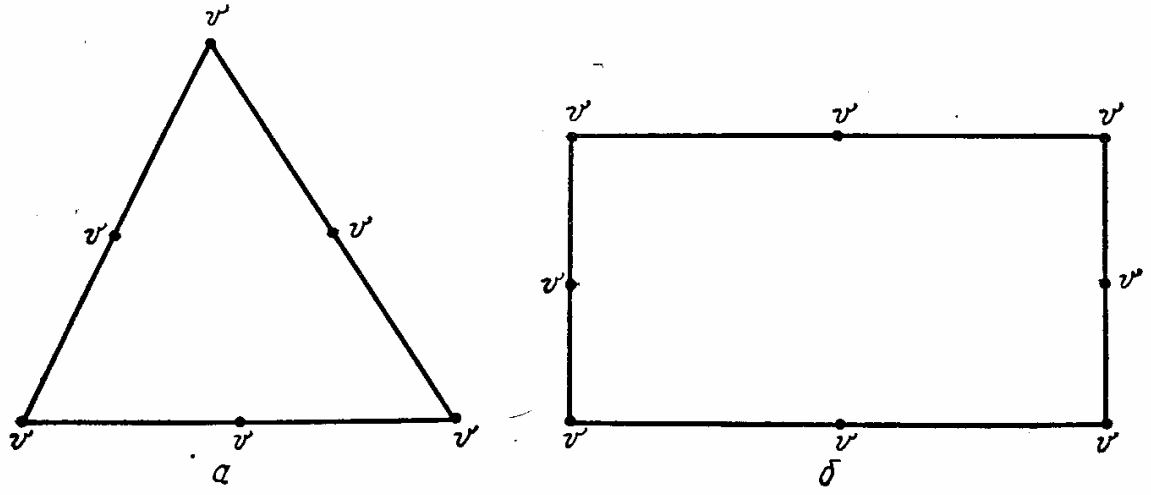

Рис. 4.

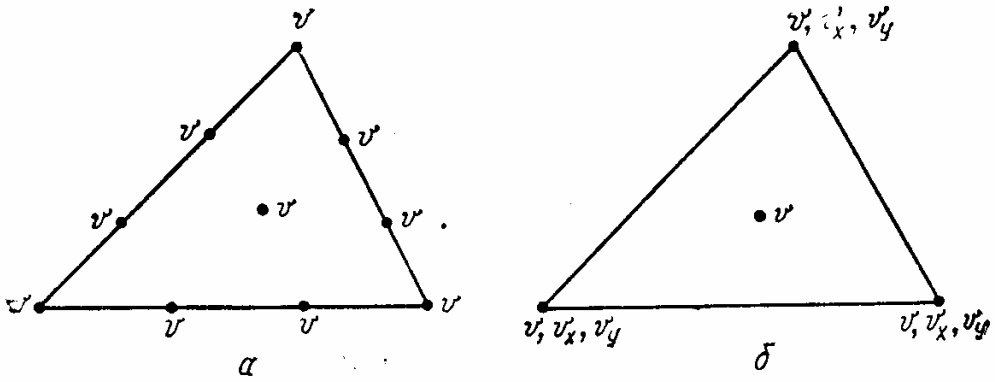

Рис. 5.

части; в каждом узле фиксируется значение допустимой функции v (рис. 5, *а*);

2) узловыми точками служат вершины и центр тяжести треугольника, а узловыми параметрами являются значения функций v во всех узлах и значения первых частных производных v_x, v_y в вершинах треугольника (рис. 5, *б*).

В обоих случаях, как легко убедиться, гарантируется только непрерывность допустимых функций на всей области Ω, а их частные производные претерпевают разрывы первого рода при переходе через сторону в соседний треугольник.

Отметим, что в случае решения задачи в двумерной области, ограниченной ломаной линией, достаточно эффективным может быть разбиение области $\overline{\Omega}$ на сочетание прямоугольников и треугольников (рис. 6) с соответствующими допустимыми функциями: билинейные — линейные, биквадратичные — квадратичные и др. Такое комбинирование треугольников и прямоугольников в случае многоугольной области позволяет точно аппроксимировать границу области $\overline{\Omega}$ и подчас (например, при решении задач с переменными коэффициентами) снизить общие затраты на вычисление решения методом конечных элементов относительно дискретизации с использованием только треугольных элементов.

В ряде случаев необходимо решать прикладные задачи в областях с криволинейной границей. Для определенности рассмотрим двумерную

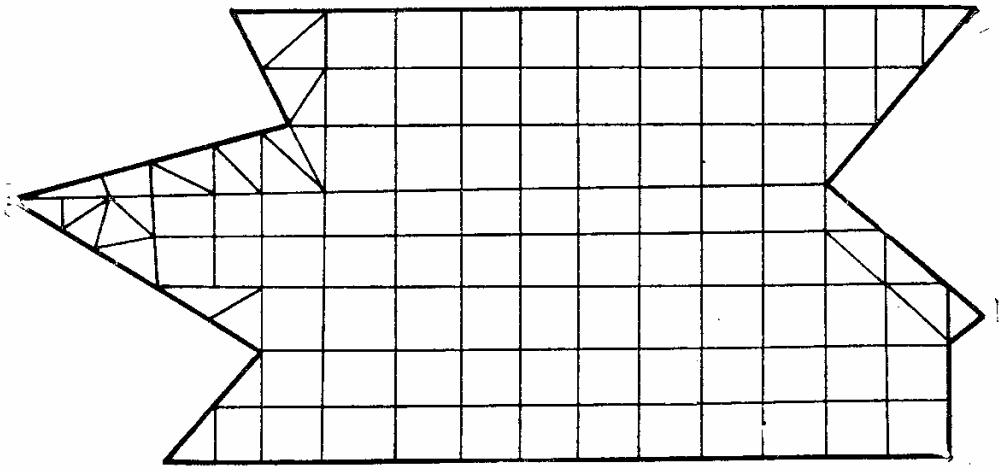

Рис. 6.

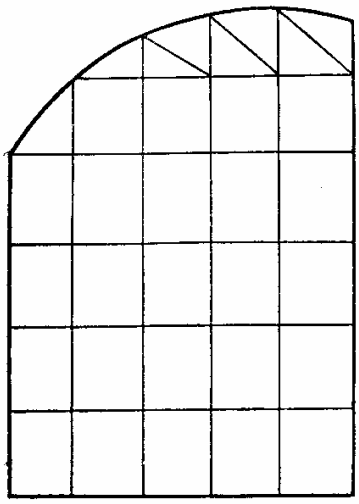

Рис. 7.

область, изображенную на рис. 7. Не всегда такую область можно разбить на комбинации четырехугольных и треугольных элементов или только треугольных элементов, которые обеспечили бы достаточную точность полученного решения (в результате аппроксимации дуг криволинейной области хордами). В этих случаях вместо треугольных элементов, вершины которых лежат на границе области Ω, можно рассматривать треугольные элементы с криволинейной стороной, принадлежащей соответствующему участку границы (рис. 7).

Если исходная область $\bar{\Omega}$ рассматривалась в системе координат x, y, то переходом к новой системе координат ξ, η элемент e_i, изображенный на рис. 8, может быть приведен к стандартной форме (к треугольнику с прямолинейными сторонами в системе координат ξ, η (рис. 9)). Отметим, что соответствующее преобразование координат должно быть взаимно-однозначным и не должно сильно искажать исходный элемент.

При решении прикладных задач в двумерных областях с криволинейной границей достаточно часто и успешно используются изопараметрические элементы. Термин «изопараметрический» означает, что для упомянутого выше преобразования координат выбираются такие же полиномы, как и для самих допустимых функций. Если же степень полиномов, применяемых для преобразования координат, ниже степени полиномов, используемых для допустимых функций, то такой конечный элемент называется субпараметрическим.

При решении пространственных задач наиболее часто используются тетраэдр (рис. 10, *а*) с допустимой функцией вида

$$v^N = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 z$$

и прямоугольные призмы (рис. 10, *б*) с допустимой функцией вида

$$v^N = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 z + \alpha_5 xy + \alpha_6 yz + \alpha_7 xz + \alpha_8 xyz.$$

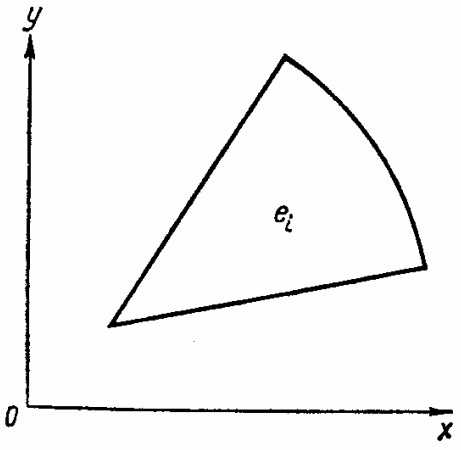

Рис. 8.

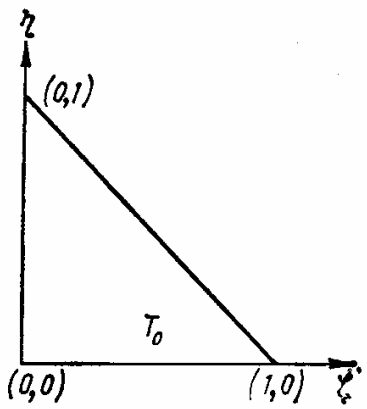

Рис. 9.

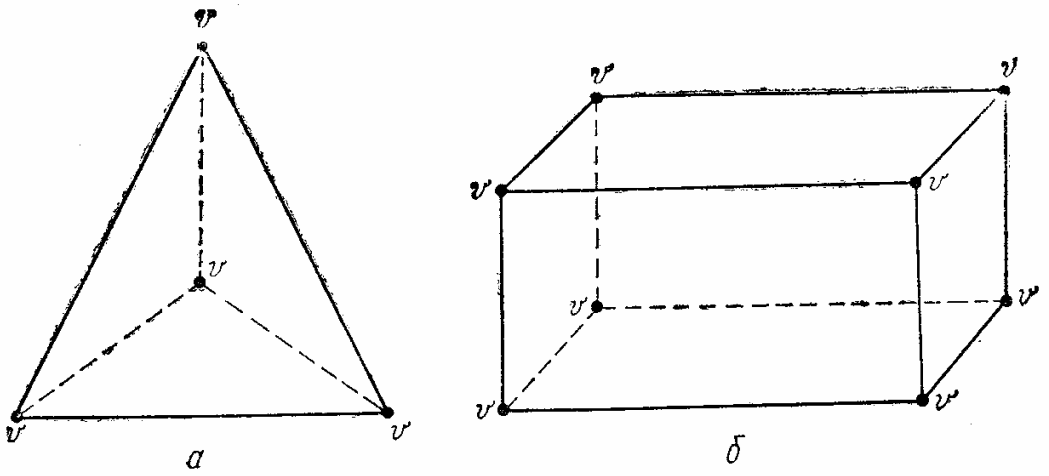

Рис. 10.

Можно использовать на таких элементах в качестве допустимых функций и полиномы более высоких степеней.

Применяются при решении пространственных задач и изопараметрические элементы. На рис. 11 приведен набор различных изопараметрических элементов с полиномиальными функциями (в узлах фиксируются значения v_i), используемых в некоторых пакетах прикладных программ.

Укажем один из возможных способов разбиения двумерной области на треугольные элементы. Сначала двумерная область делится на укрупненные четырехугольные и треугольные подобласти (или зоны), а затем они подразделяются на треугольники. Границы между подобластями должны проходить там, где изменяется приложенная нагрузка или свойства материала.

Если (выделенная) подобласть имеет криволинейную границу, то в этом случае можно использовать два подхода к дискретизации. В первом случае криволинейные границы элементов заменяются прямыми отрезками, во втором — наряду с прямолинейными треугольниками внутри области можно использовать криволинейные, прилегающие к криволинейной границе.

При разбиении треугольной подобласти на более мелкие треугольники целесообразно поступать следующим образом. Каждая сторона исходного треугольника делится на k равных отрезков. Концы отрезков, лежащих на каждой паре сторон (рис. 12, где $k = 3$), соединяются прямыми, которые, пересекаясь, образуют k^2 новых конгруэнтных треугольников.

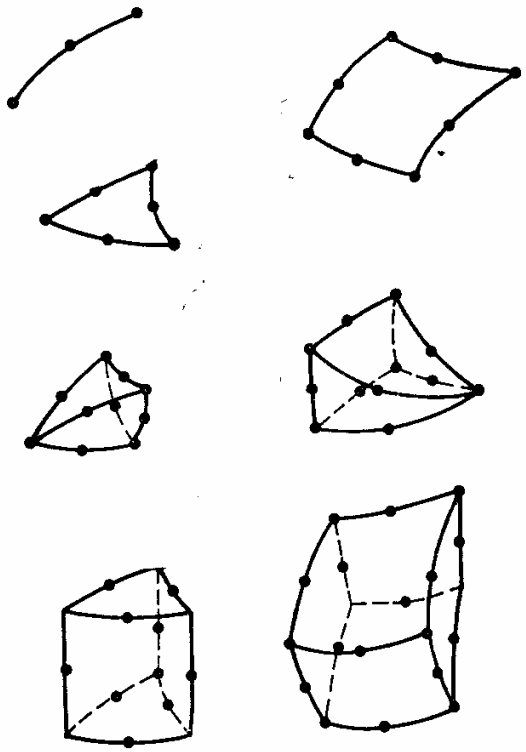

Рис. 11.

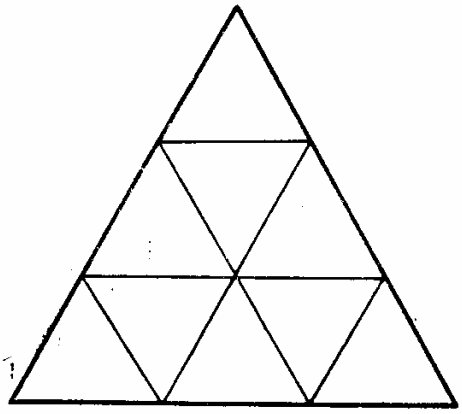

Рис. 12.

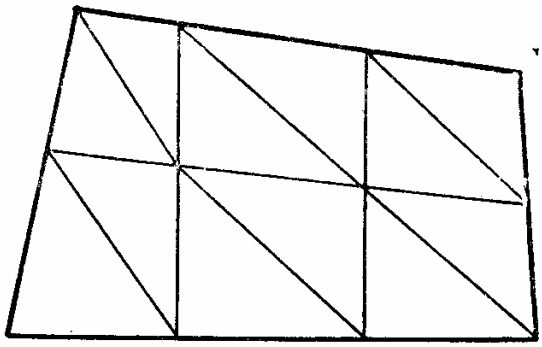

Рис. 13.

ников, заключенных внутри исходного. При таком разбиении не образуется слишком острых углов: самый малый угол исходной (укрупненной) триангуляции равен самому малому углу заключительной триангуляции. Для треугольников с одной искривленной стороной алгоритм аналогичен.

Четырехугольные зоны, как и треугольные, обычно разбивают на элементы соединением концов отрезков, лежащих на противоположных сторонах. Полученные внутренние четырехугольники могут быть, в свою очередь, разбиты на треугольные элементы проведением соответствующей диагонали. При разбиении четырехугольников на треугольники нужно избегать треугольников со слишком острыми углами. Такие элементы обеспечивают более точные результаты.

Стороны четырехугольника могут делиться на отрезки разной длины, чтобы получать элементы разных размеров. В исходном четырехугольнике в результате разбиения будет получено $N = 2kl$ треугольников, если смежные стороны его разбить на k и l отрезков (рис. 13, где $k = 3$, $l = 2$). Кроме того, выбор вида допустимой функции и фиксированных параметров на каждом элементе должен обеспечить однозначность ее определения на общей стороне, что необходимо для удовлетворения условий гладкости допустимых функций на всей области Ω.

Треугольные и четырехугольные подобласти могут иметь общую границу. Число узлов на этой границе для обеих подобластей должно быть одинаковым, и относительное положение узлов должно совпадать.

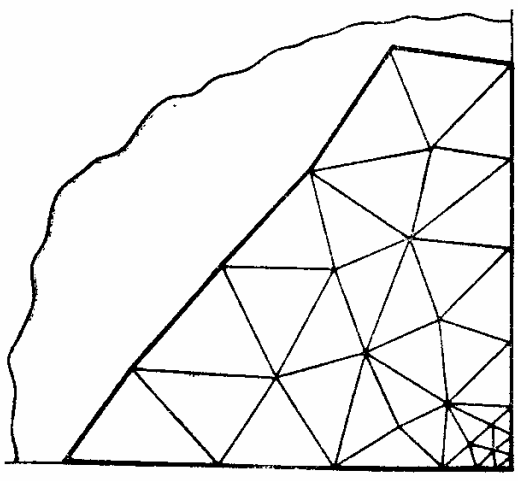

Рис. 14.

Равномерное разбиение, когда все элементы имеют одинаковые форму и размеры, в реальных задачах обычно не проводится, потому что существуют зоны концентрации напряжений, участки с большими температурными градиентами и т. д. В этих зонах, обычно прилегающих к углам, решение необходимо вычислять с более высокой точностью, что может быть достигнуто за счет измельчения элемен-

тов. Возможность варьировать размеры элементов — важное преимущество МКЭ. Пример неравномерной сетки МКЭ дан на рис. 14.

В настоящее время известен ряд алгоритмов с различной степенью автоматизации процесса разбиения области (см., например, [41, 44, 60, 62]). Здесь мы упомянем лишь один — полуавтоматический. Он состоит в том, что в ЭВМ вводят (с учетом геометрии области, характеристик материала, предполагаемого поведения искомого решения) грубую триангуляцию области и величину k, указывающую, на сколько равных отрезков делятся стороны треугольников исходной триангуляции. Далее в автоматическом режиме выполняется разбиение каждого введенного треугольника по алгоритму, описанному выше.

Режим полуавтоматического построения сеток удобно использовать при работе с дисплеем в интерактивном режиме. Достоинством при таком построении конечно-элементной сетки является возможность контроля формирования сетки и небольшой объем входной информации для ЭВМ.

Итак, разбиение исходной области Ω на конечное число элементов Ω_i, $i = 1, 2, \ldots, N$, и определение допустимой функции МКЭ в виде полинома заданной степени n на каждом элементе Ω_i порождают некоторое конечномерное пространство P_n^h кусочно-полиномиальных функций $v^N(x)$. (Здесь h характеризует некоторым образом размер элементов, а через x обозначена произвольная точка области Ω.) Каждая $v^N(x)$ определяется своим набором узловых параметров q_j, $j = 1, 2, \ldots, r$, фиксированных (по одному или по нескольку) в узлах $x = X_k$, $k = 1, 2, \ldots, s$, покрывающей область Ω сетки. Каждый параметр q_j служит значением либо самой функции $v^N(x)$, либо одной из ее производных в конкретном узле X_k. Размерность пространства P_n^h

$$r = \sum_{k=1}^{s} p_k,$$

где p_k — число параметров, фиксированных в узле X_k, $k = 1, 2, \ldots, s$.

В качестве базисных функций МКЭ пространства P_n^h выбираются $\varphi_j(x)$, $j = 1, 2, \ldots, r$,— кусочные полиномы n-й степени, однозначно определяемые следующими условиями.

Узловой параметр q_j функции $\varphi_j(x)$, $j = 1, 2, \ldots, r$, равен единице, а все остальные ее узловые параметры q_k, $k \neq j$, равны нулю. Иными словами, в заданном узле X_i либо значение самой функции $\varphi_j(x)$, либо значение какой-то одной конкретной ее производной (обозначаемое через q_j) равно единице, а все остальные фиксируемые в данном узле X_i параметры $\varphi_j(x)$ полагаются равными нулю. Равны нулю значения функции $\varphi_j(x)$ и ее производных и во всех остальных узлах $x = X_k$, $k \neq i$, области Ω. В этом случае принято считать, что базисная функция $\varphi_j(x)$ соответствует узловому параметру q_j.

При таком определении базисных функций оказывается, что функция $\varphi_j(x)$ отлична от нуля только на совокупности тех подобластей (элементов), которые содержат узел X_i. Область, где базисная функция $\varphi_j(x)$ отлична от нуля, называют носителем $\varphi_j(x)$. Несколько различных форм носителей показано на рис. 15. (Подробнее о построении

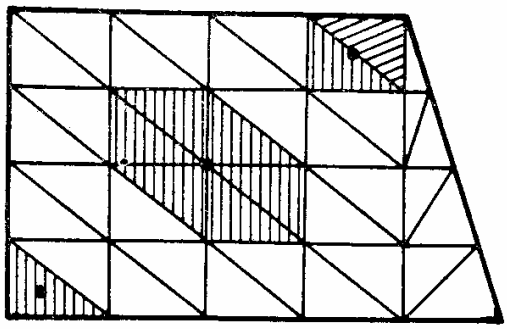

Рис. 15.

и свойствах базисных функций см. в гл. II.) Используя указанный базис, любую допустимую функцию $v^N(x) \in P_n^h$ можно представить в виде

$$v^N(x) = \sum_{j=1}^{r} q_j \varphi_j(x),$$

где q_j — узловые параметры функции $v^N(x)$. Заметим, что при решении вариационных задач непосредственно знать базисные функции МКЭ не обязательно. В дальнейшем будут описаны процедуры, позволяющие выполнить дискретизацию задачи и найти значения q_j из вариационного принципа, не прибегая к явному построению базиса. Для применения варианта МКЭ, основанного на методе Бубнова — Галеркина, необходимо иметь явный вид базисных функций.

В связи с дискретизацией задачи коснемся еще вопросов, связанных с поиском путей сокращения вычислительных трудностей реализации МКЭ на ЭВМ. Здесь можно упомянуть возможность снижения размерности пространства допустимых функций при сохранении порядка точности приближенного решения. Добиться этого можно, в частности, выбором конечного элемента с меньшим числом узлов на элементе, но с большим количеством параметров, фиксированных в одном узле при одинаковой степени полиномов n и одинаковом числе N элементов. Например, нетрудно подсчитать, что общее число фиксированных параметров, т. е. размерность пространства P_3^h, при разбиении Ω, указанном на рис. 1, будет: $r = 100$, если использовать кубический лемент, изображенный на рис. 5, *а*, и $r = 66$, если использовать кубический элемент, изображенный на рис. 5, *б*. (Отметим, что в первом случае имеются в виду кусочно-кубические функции Лагранжа, а во втором — кусочно-кубические полиномы Эрмита.)

Далее, дискретизация МКЭ исходных линейных задач приводит к системе линейных алгебраических уравнений с разреженными матрицами. При этом оказывается, что выбор соответствующей нумерации всех узлов области Ω (и фиксированны в них параметров) позволяет построить такую систему, информацию о которой удобнее и экономнее хранить в памяти ЭВМ и решать прямыми методами. (Использование прямых методов часто является предпочтительным по точности получаемого решения и времени счета.) Рассмотрим один частный случай зависимости структуры симметричной ленточной матрицы от способа нумерации узлов введенной сетки. Напомним при этом, что для систем с симметричной положительно определенной ленточной матрицей $A = (a_{ij})$, у которой $a_{ij} = 0$ при $|i - j| > \alpha > 0$, существуют специальные алгоритмы прямых методов, исключающие действия с нулевыми элементами вне ленты и требующие запоминания только элементов ленты, т. е. хранения массива с размерами $r \times (\alpha + 1)$, где r — порядок системы (размерность пространства P_n^h).

Предположим, что область Ω представлена как совокупность треугольных элементов, узлы которых расположены только в вершинах

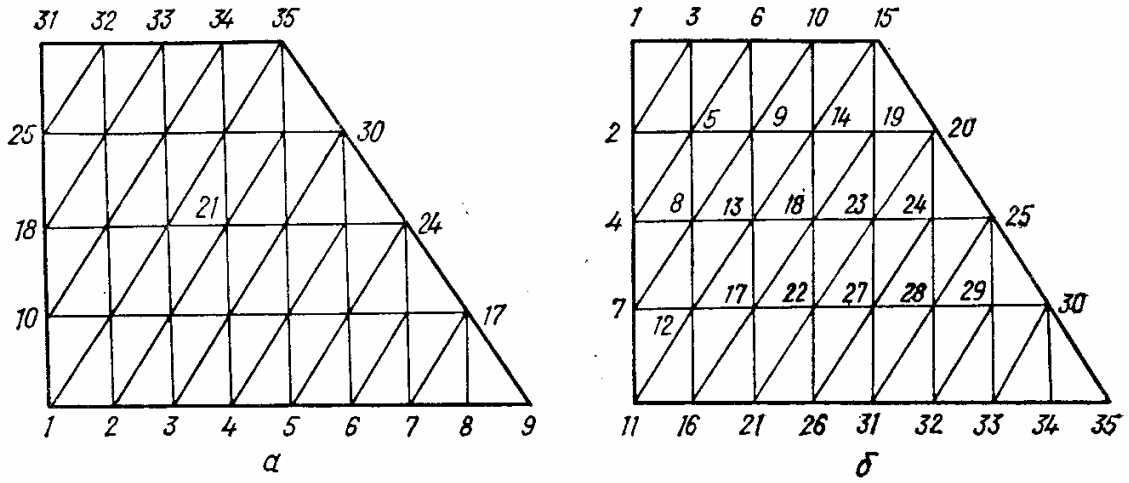

Рис. 16.

треугольников, причем в каждом узле зафиксировано равное количество параметров. При заданной нумерации всех узлов сетки МКЭ каждый элемент имеет свою наибольшую разность между номерами его узлов. Если обозначить через R максимальное (по всей области Ω) значение этой разности, через p — число фиксированных в каждом узле параметров, то

$$\alpha + 1 = (R + 1)\, p.$$

Отсюда ясно, что минимизация α связана с минимизацией R и осуществить ее можно, в частности, соответствующим выбором нумерации узлов имеющегося разбиения области Ω.

На рис. 16, *а* и *б* показаны соответственно первый и второй способы нумерации узлов одной и той же конечно-элементной сетки, покрывающей трапециевидную область Ω. Предположим, что используется конечный элемент вида, изображенного на рис. 2, *а*. Очевидно, что ширина ленты, т. е. $q = 2\alpha + 1$, симметричной матрицы A дискретной системы, полученной при нумерации первым способом, будет равна 21, а вторым — только 11.

Достаточно подробное освещение вопросов, связанных с упорядочением симметричной положительно определенной матрицы A с целью приближенной минимизации количества памяти, необходимой для решения прямым методом системы МКЭ с матрицей A, можно найти в работе [23].

3. Некоторые другие варианты МКЭ. В п. 2 параграфа I.1 описаны постановки задач о равновесии упругих тел в перемещениях. При решении таких задач МКЭ используется для построения приближенной вектор-функции u, характеризующей перемещения тела, а затем с помощью соотношений (I.1), (I.2) по найденному приближенному решению определяются компоненты тензора деформаций и тензора напряжений. Поскольку исследователей в ряде задач интересуют именно напряжения, возникающие в теле, естественно поставить вопрос о непосредственном определении этих величин. В результате исходная постановка задачи может быть сформулирована не в форме, описанной в п. 2 параграфа I.1, а в виде некоторых эквивалентных вариацион-

ных задач [58, 92]. Эти эквивалентные задачи также описывают состояние исследуемого объекта посредством других величин, например компонент тензора напряжений или компонент тензора напряжений и вектора перемещений u и т. п.

Для построения эквивалентных вариационных задач имеется несколько способов. Один из них — использование теории двойственности [118].

Если исходная задача описывает исследуемый процесс через компоненты тензора напряжений и проводится конечноэлементная аппроксимация непосредственно компонент тензора напряжений, то данный вариант метода называют вариантом МКЭ в напряжениях. Если осуществляется дискретизация вариационной задачи, зависящей от компонент как тензора напряжений, так и вектора перемещений, то такой вариант называют смешанным методом конечных элементов. При дискретизации вариационной задачи, определенной на Ω посредством компонент тензора напряжений, а на границе Γ или ее части — компонент вектора перемещений, речь идет о гибридном методе конечных элементов.

Отметим, что для краевой задачи, описываемой системой (I.7) с граничными условиями (I.6), (I.8), эквивалентной вариационной задачей является следующая.

Требуется найти такие σ_{ik}, i, $k = 1, 2, 3$, которые доставляют минимум функционалу

$$R(\sigma) = -\frac{1}{2}\iiint\limits_{\Omega} \sum_{i,k,l,m=1}^{3} d_{iklm}\sigma_{ik}\sigma_{lm}d\Omega + \iint\limits_{\Gamma_1} \sum_{i,k=1}^{3} \sigma_{ik}\cos(n, x_k)\, g_i d\Gamma \tag{I.43}$$

на множестве статически допустимых полей

$$M = \{\sigma : \sigma_{ij} \in L_2(\Omega),\ \forall\, i,\ j,\ \sigma_{ij} = \sigma_{ji},$$

$$\sum_{j=1}^{3} \frac{\partial \sigma_{ij}}{\partial x_j} = -f_i,\ (x_1, x_2, x_3) \in \Omega, \tag{I.44}$$

$$\sum_{j=1}^{3} \sigma_{ij}\cos(n, x_j) = -q_i,\ i = 1,\ 2,\ 3,\ (x_1, x_2, x_3) \in \Gamma_2\Big\}.$$

Если σ — решение вариационной задачи (I.43), (I.44), а u — задачи (I.6) — (I.8), то $R(\sigma) + \Phi(u) = 0$, и, кроме того,

$$\sigma_{ik} = \sum_{l,m=1}^{3} c_{iklm}\frac{1}{2}\left(\frac{\partial u_m}{\partial x_l} + \frac{\partial u_l}{\partial x_m}\right).$$

Переход к эквивалентным вариационным задачам представляет интерес в случае возможности использования допустимых функций с меньшей гладкостью, чем требуется для метода конечных элементов в перемещениях. Кроме того, при дискретизации эквивалентной вариационной задачи можно непосредственно получить приближенные значения наиболее интересных величин, например компонент тензора напряжений. Заметим, что в данном пункте речь шла о разных вариантах МКЭ в зависимости от формы постановки (формулировки) исход-

ной задачи теории упругости. Иногда понятие о вариантах МКЭ связывают с численными методами (Бубнова — Галеркина, Ритца, наименьших квадратов, коллокаций и т. п.), которые используются для решения математических задач и которые легли в основу конкретного варианта МКЭ.

4. Понятие о методе суперэлементов. Метод конечных элементов используется не только для определения напряженно-деформированного состояния отдельных элементов и узлов конструкций. Его также можно применять для определения напряженно-деформированного состояния конструкций в целом. Подчас для получения искомого решения требуемой точности необходимо произвести разбиение рассматриваемой области на большое число элементов (треугольников, прямоугольников и т. д.). Разбиение исследуемого объекта на большое число элементов порождает ряд проблем при машинной реализации метода. Одна из них, как уже упоминалось,— решение систем линейных алгебраических уравнений с большим числом неизвестных (симметричные ленточные матрицы таких систем могут иметь порядок до нескольких десятков тысяч и ширину ленты — свыше тысячи). Кроме того, ситуация еще осложняется и тем, что при исследованиях необходим многовариантный счет задач. Реализация таких задач даже на современных ЭВМ требует сотни часов счета. Поэтому возникла потребность в создании различных модификаций МКЭ, позволяющих сократить затраты времени на решение задачи.

Одной из модификаций является метод суперэлементов [87]. Суть его заключается вот в чем. Исследуемая область $\bar{\Omega}$ разбивается на конечное число элементов $\omega_\varkappa$ ($\bar{\Omega} = \bigcup\limits_{\varkappa=1}^{N} \bar{\omega}_\varkappa$, $\omega_\varkappa \cap \omega_j = 0$ при $\varkappa \neq j$, $\varkappa$, $j = 1, 2, \ldots, N$). Элементы $\omega_\varkappa$ могут быть разнообразной формы, в том числе треугольной, прямоугольной и т. д. В соответствии с этим разбиением можно построить систему линейных алгебраических уравнений МКЭ

$$Kv = f \tag{I.45}$$

с глобальной матрицей жесткости K и вектором нагрузки f. Однако, как отмечалось выше, такая система может иметь очень большой порядок. Для понижения порядка решаемой системы линейных алгебраических уравнений объединяют рядом лежащие элементы $\omega_\varkappa$ в группы $\bar{\Omega}^j$, называемые подструктурами:

$$\bar{\Omega} = \bigcup_{j=1}^{k} \bar{\Omega}^j, \quad \Omega^\nu \cap \Omega^j = 0 \text{ при } \nu \neq j; \quad \nu, j = 1, 2, \ldots, k \leqslant N.$$

Такую подструктуру с соответствующими допустимыми функциями и узловыми точками, которые были определены на ω_j, называют суперэлементом [87]. Узловые точки подобластей $\bar{\Omega}^j$, лежащие на границе $\partial\Omega^j$ области Ω^j, называют граничными, а лежащие внутри — внутренними узловыми точками суперэлемента.

Теперь в системе линейных алгебраических уравнений (I.45) исключим неизвестные, являющиеся внутренними фиксированными парамет-

рами суперэлементов $\overline{\Omega}^j$. Такое исключение неизвестных фиксированных параметров можно произвести на промежуточном этапе, не формируя непосредственно систему (I.45). Для этого в каждой из подструктур Ω^j (рассматривая их как самостоятельные) на основе элементарных матриц жесткости и элементов вектора нагрузки соответствующих конечных элементов $\omega_\varkappa$ строят систему уравнений

$$K^j v^j = f^j. \tag{I.46}$$

Систему (I.46) называют уравнениями равновесия подструктуры $\overline{\Omega}^j$, а K^j — матрицей жесткости подструктуры, v^j — вектором узловых перемещений, f^j — вектором нагрузок, приложенных к узлам подструктуры $\overline{\Omega}^j$.

В соответствии с разделением узлов суперэлемента $\overline{\Omega}^j$ на внутренние (в) и граничные (г) систему уравнений (I.46) можно записать в виде

$$\begin{bmatrix} K^j_{\text{вв}} & K^j_{\text{вг}} \\ (K^j_{\text{вг}})^T & K^j_{\text{гг}} \end{bmatrix} \begin{bmatrix} v^j_{\text{в}} \\ v^j_{\text{г}} \end{bmatrix} = \begin{bmatrix} f^j_{\text{в}} \\ f^j_{\text{г}} \end{bmatrix}, \tag{I.47}$$

где $K^j_{\text{вв}}$ — симметричная положительно определенная матрица с размерами $p \times p$, p — количество неизвестных параметров во внутренних узлах суперэлемента $\overline{\Omega}^j$, $K^j_{\text{вг}}$ — матрица с размерами $p \times r$, r — число параметров в граничных узлах, $K^j_{\text{гг}}$ — симметричная матрица с размерами $r \times r$.

Из (I.47) можно исключить перемещение $v^j_{\text{в}}$ внутренних узлов. В результате получим уравнение равновесия суперэлемента

$$\overline{K}^j v^j_{\text{г}} = \overline{f}^j,$$

где $\overline{K}^j$ — симметричная матрица с размерами $r \times r$, называемая матрицей жесткости суперэлемента $\overline{\Omega}^j$,

$$\overline{K}^j = K^j_{\text{гг}} - Z^T_{\text{вг}} Z_{\text{вг}},$$

$\overline{f}^j$ — вектор узловых усилий суперэлемента $\overline{\Omega}^j$, определяемый соотношением

$$\overline{f}^j = f^j_{\text{г}} - Z^T_{\text{вг}} \overline{f}^j_{\text{в}}.$$

Матрицу $Z_{\text{вг}}$ и вектор $\overline{f}^j_{\text{в}}$ находят из систем уравнений

$$L^T Z_{\text{вг}} = K_{\text{вг}}, \quad L^T \overline{f}^j_{\text{в}} = f^j_{\text{в}},$$

где L — треугольная матрица разложения симметричной и положительно определенной матрицы

$$K^j_{\text{вв}} = L^T L.$$

Реализовать это треугольное разложение можно методом квадратных корней [107].

Из вычисленных матриц жесткости $\overline{K}^j$ суперэлементов и векторов $\overline{f}^j$ строятся соответственно глобальная матрица A и вектор правой части b

системы линейных алгебраических уравнений МКЭ

$$Ax = b. \tag{I.48}$$

Полученная таким образом система линейных алгебраических уравнений (I.48) тождественно совпадает с той системой уравнений, которая была бы получена из (I.45) после исключения в последней всех $v_в^j$, $j = 1, 2, \ldots, k$.

В результате решения системы уравнений (I.48) получают все неисключенные неизвестные $v_г^j$, $j = 1, 2, \ldots, k$, а в случае необходимости неизвестные перемещения v_b^j во внутренних исключенных узловых точках вычисляются с помощью соотношений

$$v_в^j = L^{-1} f_в^j,$$

где

$$\tilde{f}_в^j = \bar{f}_в^j - Z_{вг} v_г^j.$$

Рассмотренная процедура построения суперэлементов может быть проведена в несколько этапов, т. е. подструктуры $\bar{\Omega}^j$ могут быть сами объединены в группы и т. д.

На основании изложенного очевидно, что метод суперэлементов позволяет значительно понизить порядок системы линейных алгебраических уравнений без непосредственного построения полной системы линейных алгебраических уравнений и без снижения точности МКЭ. Вместе с тем в методе суперэлементов могут возникать и новые проблемы, связанные с увеличением ширины ленты матрицы системы уравнений, с алгоритмизацией и др.

Глава II

МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ В КРАЕВЫХ ЗАДАЧАХ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

В данной главе на примере некоторых краевых задач показывается весь ход постановки и решения соответствующей вариационной задачи методом конечных элементов. Рассматриваются варианты МКЭ, основанные на модифицированном процессе Ритца и процессе Бубнова — Галеркина. Исследуются сходимость метода, вопросы его численной реализации, а также оценки точности полученных результатов.

II.1. Постановка задач

1. Обыкновенные дифференциальные уравнения второго порядка. Одна из краевых задач для дифференциальных уравнений второго порядка состоит в отыскании функции $u(x)$, которая на отрезке $[0, l]$ удовлетворяет уравнению

$$-\frac{d}{dx}\left(k(x)\frac{du}{dx}\right)+q(x)u=f(x) \qquad \text{(II.1)}$$

и краевым условиям

$$k(x)\frac{du}{dx}-\beta u(x)|_{x=0}=g_1,\quad u(l)=g_2. \qquad \text{(II.2)}$$

Предположим, что функции $k(x)\geqslant k_0>0$, $\frac{dk}{dx}$, $q(x)\geqslant 0$ непрерывны на $[0, l]$, $f(x)\in L_2(0, l)$, $\beta\geqslant 0$.

К задаче (II.1), (II.2) сводится, например, описание процесса стационарного распределения тепла в неоднородном стержне длины l в условиях интенсивного теплообмена с окружающей средой, когда на конце $x=0$ теплообмен подчиняется закону Ньютона, а на конце $x=l$ поддерживается температура g_2. Существуют и другие практические задачи, в которых приходится решать уравнение вида (II.1) при соответствующих краевых условиях и ограничениях на исходные данные.

В частности, при расчете на прочность вращающихся дисков [22] (рис. 17) основное уравнение растяжения диска, полученное в предположении, что на элемент диска действуют распределенные по граням

окружные и радиальные напряжения, а также объемные силы q_z, можно представить в виде

$$-\frac{d}{dr}\left(\varphi(r)\frac{du}{dr}\right)+\varphi(r)\left[\frac{1}{r^2}-\right.$$
$$\left.-\frac{d}{dr}\left(\frac{\mu}{r}\right)-\frac{\mu}{r}\frac{d(\ln\varphi)}{dr}\right]u=$$
$$=\varphi(r)f(r),\quad a<r<b,$$

где

$$\varphi(r)=\frac{rhE}{1-\mu^2},$$

$$f(r)=(1+\mu)\alpha T\frac{d(\ln\varphi)}{dr}+$$
$$+r\frac{d}{dr}\left[\frac{(1+\mu)\alpha T}{r}\right]-q_r\frac{1-\mu^2}{E},$$

$u=u(r)$ — перемещение, $E=E(r)$ — модуль упругости, $\mu=\mu(r)$ — коэффициент Пуассона, $h=h(r)$ — толщина диска, $\alpha=\alpha(r)$ — коэффициент линейного расширения материала, $T=T(r)$ — температура диска.

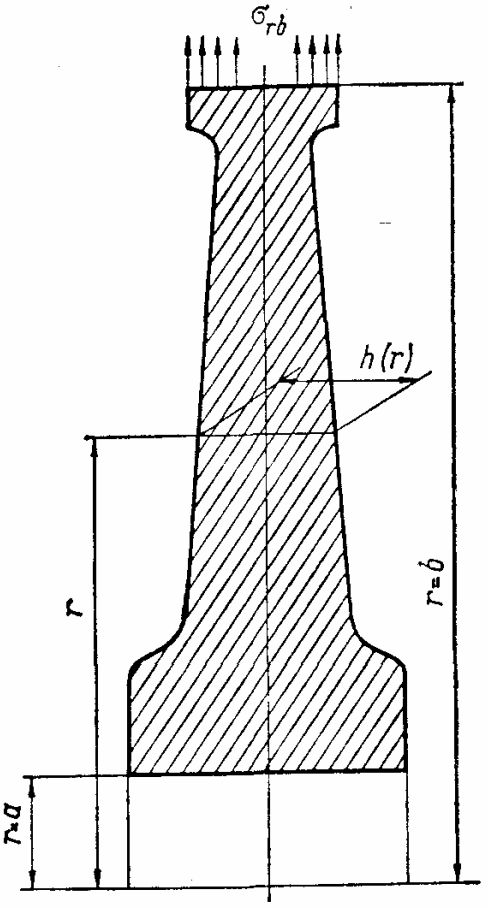

Рис. 17.

На нагруженном контуре диска ($r=b$) могут быть заданы радиальные напряжения σ_{rb}, которые можно представить в виде

$$\frac{E}{1-\mu^2}\left(\frac{du}{dr}+\mu\frac{u}{r}-(1+\mu)\alpha T\right)\Big|_{r=b}=\sigma_{rb}.$$

На внутреннем контуре ($r=a$) граничные условия зависят от условий закрепления диска. Во многих случаях удобно считать заданными силу или напряжения, так что выполняется краевое условие типа

$$\frac{E}{1-\mu^2}\left[\frac{du}{dr}+\mu\frac{u}{r}-(1+\mu)\alpha T\right]\Big|_{r=a}=\sigma_{ra}.$$

Задача (II.1), (II.2), как известно, может быть сведена к задаче с однородными краевыми условиями. Для этого достаточно найти произвольную дважды непрерывно дифференцируемую на $[0, l]$ функцию $u_1(x)$, удовлетворяющую условиям (II.2), и рассмотреть новую неизвестную функцию $w(x)=u(x)-u_1(x)$.

Легко видеть, что относительно функции $w(x)$ задача (II.1), (II.2) перепишется в виде

$$-\frac{d}{dx}\left(k\frac{dw}{dx}\right)+qw=f(x)+\frac{d}{dx}\left(k\frac{du_1}{dx}\right)-qu_1, \qquad \text{(II.3)}$$

$$k\frac{dw}{dx}-\beta w(x)|_{x=0}=0,\quad w(l)=0. \qquad \text{(II.4)}$$

Определим оператор A формулой

$$Av = -\frac{d}{dx}\left(k\frac{dv}{dx}\right) + qv, \quad 0 < x < l, \tag{II.5}$$

и будем рассматривать его в гильбертовом пространстве $H = L_2(0, l)$. За область определения $D(A)$ этого оператора примем множество функций, удовлетворяющих требованиям

$$v(x) \in C^2[0, l], \quad k\frac{dv}{dx} - \beta v(x)|_{x=0} = 0, \quad v(l) = 0. \tag{II.6}$$

Тогда задачу (II.3), (II.4) в операторной форме можно представить в виде

$$Aw = \bar{f}, \tag{II.7}$$

где $\bar{f} = f(x) + \frac{d}{dx}\left(k\frac{du_1}{dx}\right) - qu_1$.

Классическим решением задачи (II.7) (или, что то же, задачи (II.3), (II.4)) назовем функцию $w^* \in D(A)$ и удовлетворяющую уравнению (II.7).

Нетрудно показать, что оператор A уравнения (II.7) — положительно определенный. Действительно, область его определения $D(A)$ плотна в пространстве $L_2(0, l)$, так как она содержит множество всех финитных в $(0, l)$ функций, которое плотно в $L_2(0, l)$.

Симметричность оператора A следует из явно симметричного выражения

$$(Av, w) = \beta w(0) v(0) + \int_0^l \left[k\frac{dv}{dx}\frac{dw}{dx} + qvw\right] dx, \quad v, w \in D(A), \tag{II.8}$$

которое получим, интегрируя по частям первое слагаемое в соотношении

$$(Av, w) = -\int_0^l \frac{d}{dx}\left(k\frac{dv}{dx}\right) w dx + \int_0^l qvw dx$$

и учитывая, что функции v и w удовлетворяют краевым условиям (II.4).

Положив в формуле (II.8) $w = v$, найдем

$$(Av, v) = \beta v^2(0) + \int_0^l \left[k\left(\frac{dv}{dx}\right)^2 + qv^2\right] dx,$$

откуда согласно ограничениям на исходные данные следует неравенство

$$(Av, v) \geqslant k_0 \int_0^l \left(\frac{dv}{dx}\right)^2 dx. \tag{II.9}$$

Так как $v \in C^2[0, l]$ и $v(l) = 0$, то, используя представление

$$v(x) = -\int_x^l \frac{dv}{dx}\, dx$$

и неравенство Коши — Буняковского, легко убедиться, что

$$\int_0^l \left(\frac{dv}{dx}\right)^2 dx \geqslant \frac{2}{l^2}\int_0^l v^2 dx.$$

Таким образом, неравенство (II.9) принимает вид

$$(Av, v) \geqslant \frac{2k_0}{l^2}\int_0^l v^2 dx = \gamma^2 (v, v), \quad \gamma^2 = \frac{2k_0}{l^2},$$

а следовательно, оператор A положительно определенный в $L_2(0, l)$.

Согласно результатам, изложенным в п. 1 параграфа I.2, в энергетическом пространстве H_A этого оператора A существует и единственно обобщенное решение $w_0(x)$ задачи (II.7), или, что то же, задачи (II.3), (II.4). Это решение является функцией, доставляющей минимум функционалу энергии

$$\Phi(w) = [w, w]_A - 2(w, \bar{f}) \tag{II.10}$$

в энергетическом пространстве H_A. На функции $w_0(x)$ функционал $\Phi(w)$ принимает значение

$$\Phi(w_0) \equiv \min_{w \in H} \Phi(w) = -[w_0, w_0]_A.$$

Аналогично тому, как это сделано в работе [68] для оператора краевой задачи с уравнением (II.3) и краевыми условиями $w(0) = w(l) = 0$, можно показать, что в нашем случае H_A совпадает со множеством функций пространства $W_2^1(0, l)$, обращающихся в нуль в точке $x = l$. Для функций $u, v \in H_A$ энергетическое произведение $[u, v]_A$ имеет вид (см. (II.3))

$$[u, v]_A = \int_0^l \left(k\frac{du}{dx}\frac{dv}{dx} + quv\right) dx + \beta v(0) w(0),$$

а энергетическая норма $\|u\|_A^2 = \int_0^l \left(k\left(\frac{du}{dx}\right)^2 + qu^2\right) dx + \beta u^2(0)$.

Итак, функционал (II.10) в данном конкретном случае записывается как

$$\Phi(w) = \int_0^l \left[k\left(\frac{dw}{dx}\right)^2 + qw^2 - 2w\bar{f}\right] dx + \beta w^2(0), \tag{II.11}$$

где $w \in H_A$.

Обобщенное решение $w_0(x)$ задачи (II.3), (II.4) можно найти, минимизируя функционал (II.11) на множестве функций, имеющих

суммируемые с квадратом обобщенные производные первого порядка и удовлетворяющих условию

$$w(l) = 0.$$

Если окажется, что обобщенное решение $w_0(x) \in D(A)$, то оно будет и классическим решением задачи (II.3), (II.4).

Вернемся теперь к краевой задаче (II.1), (II.2) с неоднородными краевыми условиями. Чтобы получить для нее функционал, минимизация которого позволит найти обобщенное решение данной краевой задачи, положим в функционале (II.11) $w(x) = u(x) - u_1(x)$, где $u_1(x)$ — введенная ранее известная функция. Учитывая выражение $\bar{f}(x)$ (см. (II.7)) и то, что $u_1(x)$ удовлетворяет условиям (II.2), получаем

$$\Phi(u - u_1) = F(u) - F(u_1) = F(u) - \text{const},$$

где

$$F(u) = \int_0^l \left[k(x)\left(\frac{du}{dx}\right)^2 + q(x)u^2 - 2uf(x)\right]dx + \beta u^2(0) + 2g_1 u(0). \quad \text{(II.12)}$$

Поскольку вид функции, реализующей минимум, не зависит от постоянных слагаемых функционала, то $u(x)$ — обобщенное решение задачи (II.1), (II.2) — можно найти, минимизируя функционал (II.12) на множестве функций из пространства $W_2^1(0, l)$, удовлетворяющих условию $u(l) = g_2$.

2. Обыкновенные дифференциальные уравнения второго порядка с разрывными коэффициентами. Если описание исследуемого процесса приводит к решению уравнения вида (II.1) с разрывными коэффициентами, претерпевающими разрыв первого рода в конечном числе точек интервала $(0, l)$, то постановка краевой задачи требует доопределения искомого решения: введения дополнительных условий в точках разрывов коэффициентов.

Пусть в уравнении

$$-\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) + q(x)u = f(x), \quad 0 < x < l, \quad \text{(II.13)}$$

функция $k(x) \geqslant k_0 > 0$ — кусочно-непрерывно дифференцируема, а $q(x) \geqslant 0$ — кусочно-непрерывна на отрезке $[0, l]$ и обе функции претерпевают разрыв первого рода в точке $\xi \in (0, l)$; $f(x) \in L_2(0, l)$.

Если согласно характеру исследуемого процесса искомое решение $u(x)$ уравнения (II.13) непрерывно, то наряду с обычными краевыми условиями, например

$$u(0) = u(l) = 0, \quad \text{(II.14)}$$

вводят еще и следующие дополнительные условия сопряжения в точке разрыва:

$$[u]_{x=\xi} = 0, \quad \left[k(x)\frac{du}{dx}\right]_{x=\xi} = 0, \quad \text{(II.15)}$$

где

$$[\psi]_{x=\xi} = \psi(\xi + 0) - \psi(\xi - 0).$$

Однако известны практические задачи, решения которых, как и коэффициенты уравнения, имеют разрывы первого рода. Например, таковой является задача о стационарном распределении температуры в стержне, имеющем при $x = \xi$ разрез с теплоизоляционными свойствами. В этом случае речь может идти, например, об отыскании решения уравнения (II.13) при краевых условиях (II.14) и дополнительных условиях вида

$$\left[k\frac{du}{dx}\right]_{x=\xi} = 0, \quad [u]_{x=\xi} = r\left(k\frac{du}{dx}\right)\Big|_{x=\xi} \equiv rg, \tag{II.16}$$

где r — положительная константа, а через g обозначено значение функции $k(x)\frac{du}{dx}$ в точке $x = \xi$.

Точнее, задачу (II.13), (II.14), (II.16) можно представить в виде

$$-\frac{d}{dx}\left(k_1(x)\frac{du_1}{dx}\right) + q_1(x)u_1 = f_1(x), \quad 0 < x < \xi, \tag{II.17}$$

$$-\frac{d}{dx}\left(k_2(x)\frac{du_2}{dx}\right) + q_2(x)u_2 = f_2(x), \quad \xi < x < l, \tag{II.18}$$

$$u_1(0) = u_2(l) = 0, \tag{II.19}$$

$$k_1(x)\frac{du_1}{dx}\Big|_{x=\xi} = k_2(x)\frac{du_2}{dx}\Big|_{x=\xi} = g, \tag{II.20}$$

$$u_2(\xi) - u_1(\xi) = rg, \tag{II.21}$$

где $k_1(x)$ — функция, непрерывно дифференцируемая на $[0, \xi]$, $k_2(x)$ — на $[\xi, l]$, $k_i(x) \geqslant k_0 > 0$; функции $q_i(x) \geqslant 0$, $i = 1, 2$ и $f_i(x)$, $i = 1, 2$, непрерывны на $[0, \xi]$ и $[\xi, l]$ соответственно.

Решение краевых задач с разрывными коэффициентами, подобных (II.13) — (II.15) или (II.13), (II.14), (II.16), тоже можно свести к решению некоторых вариационных задач. Это можно осуществить тем же методом, что и в предыдущем пункте, т. е. энергетическим.

Остановимся вкратце вначале на задаче (II.13) — (II.15). Можно показать, что в пространстве $L_2(0, l)$ оператор A данной задачи, определенный на множестве $D(A)$ непрерывных и дважды кусочно-дифференцируемых на $[0, l]$ функций, удовлетворяющих условиям (II.14) и (II.15), является положительно определенным оператором.

Действительно, область определения этого оператора $D(A)$ — плотна в $L_2(0, l)$, так как содержит множество $M_0 \subset D(A)$ функций $\varphi(x)$, образованных «склейкой» (сопряжением) финитных в $(0, \xi)$ и в (ξ, l) функций:

$$\varphi(x) = \begin{cases} \varphi_1(x), & 0 \leqslant x \leqslant \xi, \\ \varphi_2(x), & \xi \leqslant x \leqslant l, \end{cases} \tag{II.22}$$

где функция $\varphi_1(x)$ — финитна в $(0, \xi)$, а $\varphi_2(x)$ — в (ξ, l). Данное множество M_0 плотно в пространстве $L_2(0, l)$.

Поэтому симметричность A непосредственно следует из соотношений

$$(Au, v) = \int_0^l \left[-\frac{d}{dx}\left(k\frac{du}{dx}\right)v + quv \right] dx =$$

$$= -\int_0^{\xi} \frac{d}{dx}\left(k\frac{du}{dx}\right)vdx - \int_{\xi}^{l} \frac{d}{dx}\left(k\frac{du}{dx}\right)vdx +$$

$$+ \int_0^l quvdx = \int_0^l \left[k\frac{du}{dx}\frac{dv}{dx} + quv \right] dx, \quad u, v \in D(A),$$

полученных посредством интегрирования по частям с учетом свойств функций, принадлежащих $D(A)$, а именно: функции $u(x)$ и $v(x)$ удовлетворяют условиям (II.14), (II.15). Доказательство о положительной определенности оператора A, т. е. проверка справедливости соотношений

$$(Au, u) = \int_0^l \left[k\left(\frac{du}{dx}\right)^2 + qu^2 \right] dx \geqslant k_0 \int_0^l \left(\frac{du}{dx}\right)^2 dx =$$

$$= k_0 \left[\int_0^{\xi} \left(\frac{du}{dx}\right)^2 dx + \int_{\xi}^{l} \left(\frac{du}{dx}\right)^2 dx \right] \geqslant \gamma^2 \int_0^l u^2 dx = \gamma^2 (u, u),$$

где $\gamma^2 = \min\left(\frac{2k_0}{\xi^2}, \frac{2k_0}{(l-\xi)^2}\right)$, осуществляется так же, как в п. 1 параграфа II.1.

В силу положительной определенности оператора A обобщенное решение краевой задачи (II.13) — (II.15) можно получить, отыскав функцию, которая минимизирует в энергетическом пространстве H_A оператора A функционал

$$F(u) = [u, u]_A - 2(f, u) \equiv \int_0^l \left(k\left(\frac{du}{dx}\right)^2 + qu^2 - 2fu \right) dx. \quad \text{(II.23)}$$

Такое обобщенное решение $u_0(x)$ существует и единственно. Если $u_0(x) \in D(A)$, то это будет решением и краевой задачи (II.13) — (II.15). Можно показать, что в рассматриваемом множестве элементов H_A совпадает с $\overset{0}{W}{}_2^1(0, l)$, т. е. H_A состоит из тех и только тех функций, которые абсолютно непрерывны, имеют обобщенные первые производные, суммируемые с квадратом, и в точках $x = 0$, $x = l$ эти функции обращаются в нуль [68]. Отметим, что функции из данного пространства H_A не обязательно удовлетворяют второму из условий (II.15), т. е. условию

$$k(x)\frac{du}{dx}\Big|_{x=\xi-0} = k(x)\frac{du}{dx}\Big|_{x=\xi+0}.$$

Это условие в данном случае является естественным для оператора A. Чтобы убедиться в этом, рассмотрим необходимое условие существования минимума функционала (II.23). Пусть $u(x) = u_0(x)$ реализует минимум данного функционала, $v(x)$ — произвольная функция из H_A, η — произвольное вещественное число. Тогда $F(u_0 + \eta v) \geqslant F(u_0)$. При фиксированной функции $v(x)$ функционал $F(u_0 + v\eta)$ является функцией от η, достигающей минимума при $\eta = 0$, следовательно,

$$\frac{dF(u_0 + \eta v)}{d\eta}\Big|_{\eta=0} = 0.$$

Согласно (II.23) имеем

$$\frac{dF(u_0 + \eta v)}{d\eta}\Big|_{\eta=0} = 2\int_0^l \left[k\frac{du_0}{dx}\frac{dv}{dx} + qu_0 v - fv\right]dx = 0. \quad \text{(II.24)}$$

Если функция $u_0(x)$ такова, что имеет по две производные в интервалах $(0, \xi)$ и (ξ, l), то в результате интегрирования по частям представим последнее равенство в виде

$$0 = \int_0^{\xi}\left[-\frac{d}{dx}\left(k\frac{du_0}{dx}\right) + qu_0 - f\right]vdx + \int_{\xi}^{l}\left[-\frac{d}{dx}\left(k\frac{du_0}{dx}\right) + \right.$$

$$\left. + qu_0 - f\right]vdx + k\frac{du_0}{dx}v\Big|_{x=\xi-0} - k\frac{du_0}{dx}v\Big|_{x=\xi+0}. \quad \text{(II.24')}$$

В силу произвольности $v(x) \in H_A$ и непрерывности всех функций из H_A, т. е. $v(\xi + 0) = v(\xi - 0)$, из полученного равенства следует, что функция $u_0(x)$, минимизирующая в H_A функционал (II.23) и удовлетворяющая уравнению (II.13), обязательно удовлетворяет и условиям (II.15).

Несколько слов о вариационной постановке задач с разрывным решением типа (II 13), (II.14), (II.16), или, что то же, (II.17) — (II.21). Оператор A, порождаемый в пространстве $L_2(0, l)$ данной задачей, тоже является положительно определенным. Его область определения $D(A)$ образуют функции $u(x) \in L_2(0, l)$ вида

$$u(x) = \begin{cases} u_1(x), & 0 \leqslant x < \xi, \\ u_2(x), & \xi < x \leqslant l, \end{cases} \quad \text{(II.25)}$$

удовлетворяющие требованиям $u_1(x) \in C^2[0, \xi]$, $u_2(x) \in C^2[\xi, l]$, и условиям (II.19) — (II.21).

Отметим, что в этом случае $D(A)$ содержит множество M_0 функций $\varphi(x)$ вида (II.22), т. е. образованных сопряжением двух множеств функций, финитных в интервалах $(0, \xi)$ и (ξ, l).

Доказательство симметричности и положительной определенности рассматриваемого оператора A выполняется так же, как в предыдущем случае. В результате вместо краевой задачи (II.13), (II.14), (II.16) можно решать задачу о минимизации функционала

$$F(u) = \int_0^l \left(k(x)\left(\frac{du}{dx}\right)^2 + q(x)u^2 - 2fu\right)dx + \frac{1}{r}[u]^2_{x=\xi} \quad \text{(II.26)}$$

в энергетическом пространстве H_A оператора A задачи (II.13), (II.14), (II.16).

В функционале (II.26) под $\frac{du}{dx}$ понимается обычная или обобщенная производная, определенная на интервалах $(0, \xi)$ и (ξ, l).

Можно показать [69], что энергетическое пространство H_A в данном случае образуют функции вида (II.25), где $u_1(x) \in W_2^1(0, \xi)$ и $u_1(0) = 0$, а $u_2(x) \in W_2^1(\xi, l)$, $u_2(l) = 0$, т. е. $u_1(x)$ и $u_2(x)$ — абсолютно непрерывные функции соответственно на $[0, \xi]$ и $[\xi, l]$, имеющие там суммируемые с квадратом обобщенные производные и удовлетворяющие условию (II.19).

Заметим, что функции из данного пространства H_A не обязательно должны удовлетворять условиям (II.20), (II.21), или, что то же, (II.16), т. е. эти условия являются естественными для оператора A. Чтобы убедиться в этом, используем известную схему. Пусть $u_0(x)$ реализует минимум функционала (II.26) в H_A исследуемого оператора, $v(x)$ — произвольная функция из H_A, а η — вещественная переменная. Тогда

$$F(u_0 + \eta v) \geqslant F(u_0), \quad \frac{dF(u_0 + \eta v)}{d\eta}\Big|_{\eta=0} = 0.$$

В рассматриваемом случае последнее равенство имеет вид

$$0 = \int_0^{\xi} k \frac{du_0}{dx} \frac{dv}{dx} dx + \int_{\xi}^{l} k \frac{du_0}{dx} \frac{dv}{dx} dx + \cdot$$

$$+ \int_0^{l} q u_0 v dx - \int_0^{l} f v dx + \frac{1}{r} [u_0]_{x=\xi} [v]_{x=\xi}. \qquad \text{(II.27)}$$

Если $u_0(x)$ имеет в интервалах $(0, \xi)$ и (ξ, l) по две производные, то, используя интегрирование по частям двух первых интегралов равенства (II.27), получаем

$$0 = \int_0^{\xi} \left(-\frac{d}{dx}\left(k \frac{du_0}{dx}\right) + q u_0 - f\right) v dx + \int_{\xi}^{l} \left(-\frac{d}{dx}\left(k \frac{du_0}{dx}\right) + \right.$$

$$\left. + q u_0 - f\right) v dx + \left(\frac{1}{r} [u_0]_{x=\xi} - k \frac{du_0}{dx}\Big|_{x=\xi+0}\right) v(\xi + 0) -$$

$$- \left(\frac{1}{r} [u_0]_{x=\xi} - k \frac{du_0}{dx}\Big|_{x=\xi-0}\right) v(\xi - 0).$$

Вследствие произвольности функции $v(x) \in H_A$ можно утверждать, что при сделанном предположении минимизирующая функция $u_0(x)$ удовлетворяет уравнениям (II.17), (II.18) и условиям

$$\frac{1}{r} [u_0]_{x=\xi} - k \frac{du_0}{dx}\Big|_{x=\xi+0} = 0, \quad \frac{1}{r} [u_0]_{x=\xi} - k \frac{du_0}{dx}\Big|_{x=\xi+0} = 0,$$

следовательно,

$$k \frac{du_0}{dx}\Big|_{x=\xi+0} = k \frac{du_0}{dx}\Big|_{x=\xi-0} = \frac{1}{r}(u_0(\xi + 0) - u_0(\xi - 0));$$

иными словами, $u_0(x)$ удовлетворяет требованиям (II.20), (II.21) или (II.16).

Завершая обсуждение постановок задач, связанных с решением обыкновенных дифференциальных уравнений второго порядка, покажем, как можно определить обобщенное решение независимо от задачи о минимизации функционала энергии. Для этого ограничимся рассмотрением задачи (II.13) — (II.15).

Пусть данная задача имеет решение $u(x) \in D(A)$, т. е. $u(x)$ — непрерывная и дважды кусочно-дифференцируемая функция, удовлетворяющая условиям (II.14), (II.15). Умножим скалярно обе части уравнения (II.13) на произвольную функцию $v(x) \in \overset{0}{W}{}_2^1(0, l)$:

$$\int_0^l \left[-\frac{d}{dx}\left(k\frac{du}{dx}\right) + qu - f\right] v dx = -\int_0^{\xi} \frac{d}{dx}\left(k\frac{du}{dx}\right) v dx -$$

$$-\int_{\xi}^{l} \frac{d}{dx}\left(k\frac{du}{dx}\right) v dx + \int_0^l (qu - f) v dx = -k\frac{du}{dx} v\Big|_{x=\xi-0} +$$

$$+ k\frac{du}{dx} v\Big|_{x=\xi+0} + \int_0^l k\frac{du}{dx}\frac{dv}{dx} dx + \int_0^l (qu - f) v dx =$$

$$= \int \left(k\frac{du}{dx}\frac{dv}{dx} + quv - fv\right) dx = 0.$$

Из полученного соотношения видно, что решение $u(x)$ задачи (II.13) — (II.15) удовлетворяет тождеству

$$\int_0^l \left(k\frac{du}{dx}\frac{dv}{dx} + quv\right) dx = \int_0^l fv dx, \quad \forall v \in \overset{0}{W}{}_2^1(0, l). \tag{II.28}$$

Вместе с тем если функция $u(x)$ удовлетворяет тождеству (II.28) при $\forall v \in \overset{0}{W}{}_2^1(0, l)$, условию (II.14) и достаточное число раз кусочно-дифференцируема, то можно показать (см. (II.24) — (II.24′)), что $u(x)$ удовлетворяет уравнению (II.13) и условиям (II.15).

Таким образом, оказывается, что нахождение решения краевой задачи (II.13) — (II.15) эквивалентно нахождению функции, удовлетворяющей интегральному тождеству (II.28) и краевому условию (II.14). Очевидно, что тождество (II.28) имеет смысл для любой функции $u(x) \in \overset{0}{W}{}_2^1(0, l)$ при ограниченных функциях $k(x)$, $q(x)$ и $f(x) \in L_2(0, l)$: все интегралы, входящие в (II.28), конечны.

Приведенные рассуждения позволяют ввести следующее определение обобщенного решения задачи (II.13) — (II.15). Функция $u(x) \in \overset{0}{W}{}_2^1(0, l)$, удовлетворяющая интегральному тождеству (II.28) при произвольной функции $v(x) \in \overset{0}{W}{}_2^1(0, l)$, называется обобщенным решением задачи (II.13) — (II.15).

Аналогичное определение обобщенного решения с помощью интегрального тождества можно вводить и для других краевых задач, выбирая в каждом конкретном случае соответствующее пространство V функций $v(x) \in V$, на которых рассматривается интегральное тождество. Определение обобщенного решения таким способом особенно важно для задач, оператор которых не является положительно определенным, а также в случае особенностей в исходных данных задачи.

Нетрудно убедиться, что для краевых задач с положительно определенными операторами оба определения обобщенных решений совпадают (см., например, [83]).

3. Обыкновенные дифференциальные уравнения четвертого порядка. К решению дифференциальных уравнений четвертого порядка приводит, например, рассмотрение изгиба непризматических балок, лежащих на упругом основании. Исследование поведения балочных элементов позволяет изучить сколь угодно сложную балочную систему, например простые и сложные рамы, судовые перекрытия, изгиб судового корпуса; выполнить расчет общей прочности при спуске судна или постановки его в док и т. д.

Уравнение изгиба балки длины l, лежащей на упругом основании, имеет вид

$$\frac{d^2}{dx^2}\left(EJ(x)\frac{d^2w}{dx^2}\right) - T\frac{d^2w}{dx^2} + q(x)w(x) = f(x), \quad 0 < x < l.$$

Здесь $w(x)$ — прогиб балки в сечении с абсциссой x, $EJ(x)$ — переменная жесткость на изгиб, $q(x)$ — переменная жесткость упругого основания, T — осевые силы, $f(x)$ — интенсивность нормальной нагрузки.

Если концы балки жестко закреплены, то выполняются краевые условия

$$w(0) = w(l) = 0, \quad \left.\frac{dw}{dx}\right|_{x=0} = \left.\frac{dw}{dx}\right|_{x=l} = 0.$$

В зависимости от других видов закрепления концов возникают и другие краевые условия. Например, если конец $x = l$ свободный, а другой жестко закреплен, то краевые условия имеют вид

$$w(0) = 0, \quad \left.\frac{dw}{dx}\right|_{x=0} = 0, \quad \left.\frac{d^2w}{dx^2}\right|_{x=l} = 0,$$

$$\left.\frac{d}{dx}\left(EJ(x)\frac{d^2w}{dx^2}\right)\right|_{x=l} = 0.$$

Пусть в общем случае требуется найти решение уравнения четвертого порядка

$$\frac{d^2}{dx^2}\left(k(x)\frac{d^2u}{dx^2}\right) - \frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u = f(x), \quad 0 < x < l, \tag{II.29}$$

удовлетворяющее краевым условиям

$$u(0) = u(l) = 0, \quad \left.\frac{du}{dx}\right|_{x=0} = \left.\frac{du}{dx}\right|_{x=l} = 0, \tag{II.30}$$

где $k(x) \geqslant k_0 > 0$, $p(x) \geqslant 0$, $q(x) \geqslant 0$, $f(x) \in L_2(0, l)$.

Как и в случае уравнения второго порядка, данная краевая задача равносильна задаче о нахождении функции, доставляющей минимум функционалу

$$F(v) = \int_0^l \left[k(x)\left(\frac{d^2v}{dx^2}\right)^2 + p(x)\left(\frac{dv}{dx}\right)^2 + qv^2 - 2fv \right] dx \qquad \text{(II.31)}$$

на множестве функций, имеющих непрерывные производные до четвертого порядка и удовлетворяющих условиям (II.30).

Обобщенное решение задачи (II.29), (II.30) есть функция, минимизирующая функционал (II.31) в пространстве функций $\overset{0}{W}{}_2^2(0, l)$, т. е. в пространстве функций, имеющих суммируемые с квадратом обобщенные производные второго порядка и удовлетворяющих условиям (II.30).

II.2. Дискретизация обыкновенных дифференциальных уравнений второго порядка

На примере задачи (II.1), (II.2) рассмотрим применение метода конечных элементов для построения системы сеточных уравнений, решение которой обеспечивает соответствующее приближение к искомому решению исходной задачи. Так как обобщенное решение задачи (II.1), (II.2) можно найти, минимизируя функционал (см. (II.12))

$$F(v) = \int_0^l \left[k\left(\frac{dv}{dx}\right)^2 + qv^2 - 2vf \right] dx + \beta v^2(0) + 2g_1 v(0) \qquad \text{(II.32)}$$

на множестве функций, имеющих суммируемые с квадратом обобщенные производные и удовлетворяющих условию

$$v(l) = g_2, \qquad \text{(II.33)}$$

то в данном случае целесообразно использовать вариант МКЭ, основанный на модифицированном процессе Ритца. Для построения конечноэлементной сетки разобьем отрезок $[0, l]$ на N элементарных отрезков $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, N$:

$$0 = x_0 < x_1 < x_2 < \cdots < x_{N-1} < x_N = l. \qquad \text{(II.34)}$$

Точки x_i будем называть узловыми.

В соответствии с выполненным разбиением необходимо определить конечный набор базисных функций $\{\varphi_i^N(x)\}$, обеспечивающих следующие свойства допустимых функций:

$$v^N(x) = \sum_i \omega_i^N \varphi_i^N(x),$$

среди которых ищется приближенное решение $u^N(x)$:

— функции $v^N(x)$ принадлежат множеству, на котором функционал (II.32) достигает минимума;

— полученное приближенное решение $u^N(x)$ позволяет удобно вычислять величины, представляющие физический интерес, например перемещения, напряжения, моменты и т. п.

В данном конкретном случае достаточно, чтобы функции $v^N(x)$ были непрерывными на $[0, l]$, обладали интегрируемыми с квадратом производными и удовлетворяли условиям (II.33). Такие допустимые функции будут принадлежать некоторому конечномерному множеству $P \subset W_2^1(0, l)$. Иногда целесообразно потребовать большей гладкости $v^N(x)$, например непрерывности функции и ее первой производной на $[0, l]$. При этом соответственно изменяется и вид (свойства) базисных функций.

Так как при использовании метода Ритца важно обеспечить только требуемые свойства допустимых функций, а не конкретный вид базиса, то можно и целесообразно строить непосредственно именно допустимые функции. Итак, определим на каждом элементарном отрезке $[x_{i-1}, x_i]$ некоторый полином с неизвестными коэффициентами, подчинив его надлежащим условиям гладкости в граничных точках отрезков. Такая кусочно-полиномиальная на $[0, l]$ функция будет обладать всеми необходимыми свойствами допустимых функций.

Условимся обозначать множество допустимых кусочно-полиномиальных функций через P_n^h, где n — степень используемых полиномов $h = \max\limits_i h_i$, $h_i = x_i - x_{i-1}$.

Опишем теперь применение различных кусочно-полиномиальных функций для дискретизации задачи (II.1), (II.2).

1. Кусочно-линейные полиномы. Пусть приближенное решение $u^N(x)$ задачи (II.1), (II.2) ищется среди функций $v^N(x) \in P_1^N \subset W_2^1(0, l)$, удовлетворяющих условию (II.33). Иными словами, пусть допустимые функции $v^N(x)$ на каждом элементарном отрезке имеют вид линейного полинома с неизвестными коэффициентами и удовлетворяют условию (II.33). Чтобы обеспечить однозначность и непрерывность допустимой функции $v^N(x)$ на всем отрезке $[0, l]$, определим ее коэффициенты на каждом элементарном отрезке $[x_{i-1}, x_i]$ через фиксированные параметры, а именно через значения допустимой функции в узловых точках

$$v_{i-1} = v^N(x_{i-1}), \quad v_i = v^N(x_i). \tag{II.35}$$

Элементарный отрезок $[x_{i-1}, x_i]$, в узлах которого зафиксированы значения допустимой кусочно-линейной функции, будем называть одномерным линейным элементом и обозначать так: «l 1—2» (полином первой степени, параметры фиксируются в двух узлах). Схематически этот элемент можно изображать в виде v ●—● v, или просто ●—●.

Если описанную допустимую функцию $v^N(x)$ подставить в функционал (II.32), то $F(v^N)$ окажется квадратичной функцией всех фиксированных, но неизвестных параметров v_i, $i = 0 \div N$. Найдя значения параметров u_i^N, доставляющие минимум функции $F(v^N)$, получим тем самым значения искомого приближенного решения $u^N(x)$ в узлах x_i:

$$u_i^N = u^N(x_i), \quad i = 0, 1, \ldots, N-1;$$

значение $u^N(x_N)$ уже известно из условия (II.33): $u^N(x_N) = g_2$.

Изложим некоторый алгоритм вычисления u_i^N, допускающий удобную машинную реализацию [158].

Представим функционал (II.32) согласно разбиению (II.34) в виде

$$F(v) = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \left[k \left(\frac{dv}{dx} \right)^2 + qv^2 - 2vf \right] dx + \beta v^2(0) + 2g_1 v(0) \quad \text{(II.36)}$$

и рассмотрим одно из слагаемых этой суммы

$$F_i(v) = \int_{x_{i-1}}^{x_i} \left[k \left(\frac{dv}{dx} \right)^2 + qv^2 - 2vf \right] dx. \quad \text{(II.37)}$$

При $v = v_i^N$ данный функционал превращается в функцию неизвестных параметров $v_{i-1}, v_i : F_i(v^N) = F_i(v_{i-1}, v_i)$. Представим эту зависимость в явном виде.

С целью стандартизации всего алгоритма отобразим посредством преобразования

$$x = x_{i-1} + h_i \xi, \quad h_i = x_i - x_{i-1}, \quad \text{(II.38)}$$

элементарный отрезок $[x_{i-1}, x_i]$ на «канонический отрезок» $[0, 1]$. Тогда получим

$$F_i(v^N) = \int_0^1 \left[\tilde{k} \left(\frac{dr}{d\xi} \right)^2 + \tilde{q} r^2 - 2\tilde{f} r \right] d\xi, \quad \text{(II.39)}$$

где

$$\tilde{k} = \frac{1}{h_i} k(x_{i-1} + h_i \xi), \quad \tilde{q} = h_i q(x_{i-1} + h_i \xi),$$

$$\tilde{f} = h_i f(x_{i-1} + h_i \xi), \quad r(\xi) = v^N(x_{i-1} + h_i \xi) = \alpha_1 + \alpha_2 \xi.$$

Неизвестные коэффициенты α_i определяются соотношениями (II.35):

$$r(0) = \alpha_1 = v_{i-1}, \quad r(1) = \alpha_1 + \alpha_2 = v_i,$$

которые в матричном виде можно записать так:

$$S\alpha = \omega_i, \quad \text{(II.40)}$$

где

$$S = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad \alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \quad \omega_i = \begin{bmatrix} v_{i-1} \\ v_i \end{bmatrix}.$$

Таким образом,

$$\alpha = S^{-1}\omega_i, \quad \text{(II.41)}$$

где

$$S^{-1} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}. \quad \text{(II.42)}$$

Отметим, что матрица S^{-1} одинакова для всех элементов $[x_i, x_{i-1}]$.

Учитывая вид функции $r(\xi)$ и соотношение (II.41), представим функционал (II.39) в виде

$$F_i = \alpha_2^2 \int_0^1 \tilde{k} d\xi + \alpha_1^2 \int_0^1 \tilde{q} d\xi + 2\alpha_1\alpha_2 \int_0^1 \xi\tilde{q} d\xi + \alpha_2^2 \int_0^1 \xi^2\tilde{q} d\xi - 2\alpha_1 \int_0^1 \tilde{f} d\xi -$$

$$- 2\alpha_2 \int_0^1 \xi\tilde{f} d\xi = \alpha^T R_i^1 \alpha + \alpha^T R_i^0 \alpha - 2\alpha^T \delta_i \equiv \omega_i^T K_i^1 \omega_i + \omega_i^T K_i^0 \omega_i - 2\omega_i^T b_i, \tag{II.43}$$

где

$$R_i^1 = \begin{bmatrix} 0 & 0 \\ 0 & \int_0^1 \tilde{k} d\xi \end{bmatrix}, \quad R_i^0 = \begin{bmatrix} \int_0^1 \tilde{q} d\xi & \int_0^1 \xi\tilde{q} d\xi \\ \int_0^1 \xi\tilde{q} d\xi & \int_0^1 \xi^2\tilde{q} d\xi \end{bmatrix}, \quad \delta_i = \begin{bmatrix} \int_0^1 \tilde{f} d\xi \\ \int_0^1 \xi\tilde{f} d\xi \end{bmatrix},$$

$$K_i^j = S^{-T} R_i^j S^{-1}, \quad j = 0, 1; \quad S^{-T} \equiv (S^{-1})^T, \quad b_i = S^{-T}\delta_i, \quad i = 1 \div N. \tag{II.44}$$

Матрицу K_i^1 обычно называют матрицей жесткости i-го элемента, $K_i^0 \equiv M_i$ — матрицей массы элемента, вектор b_i — вектором нагрузки. Матрицы K_i^1 и $K_i^0 \equiv M_i$ симметричны и в случае постоянных коэффициентов уравнения (II.1):

$$k(x) = k_0 = \text{const}, \quad q(x) = q_0 = \text{const},$$

легко вычисляются для каждого элемента:

$$K_i^1 = \frac{k_0}{h_i}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad K_i^0 = \frac{q_0 h_i}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad i = 1 \div N,$$

$$h_i = x_i - x_{i-1}.$$

В случае переменных $k(x)$, $q(x)$ матрица S^{-1} остается постоянной для всех элементов, R_i^1 и R_i^0 вычисляются посредством квадратурных формул для каждого элемента $[x_{i-1}, x_i]$. Аналогично вычисляется вектор b_i.

Согласно (II.37) — (II.43) функционал (II.36) при $v = v^N$ можно записать как функцию параметров v^i ($i = 0 \div N$):

$$F(v^N) \equiv F_1^h(v^N) = \sum_{i=1}^{N} (\omega_i^T K_i \omega_i - 2\omega_i^T b_i) + \beta v_0^2 + 2v_0 g_1, \tag{II.45}$$

где $\omega_i^T = [v_{i-1}, v_i]$, $K_i = K_i^1 + K_i^0$.

Символ $F_1^h(v^N)$ здесь (в дальнейшем F_n^h) будет определять значение функционала, полученное на классе допустимых кусочно-полиномиальных функций первой (n-й) степени при максимальной длине h элементарных отрезков $[x_{i-1}, x_i]$:

$$h = \max_i h_i, \quad h_i = x_i - x_{i-1}.$$

Введение общего вектора

$$\omega^T = [v_0, v_1, \ldots, v_{N-1}, v_N]$$

позволяет представить выражение (II.45) в виде

$$F_1^h(v^N) = \omega^T \bar{K} \omega - 2\omega^T \bar{b},$$

где симметричная матрица $\bar{K}$ порядка $N+1$ построена из элементарных матриц $K_i = K_i^1 + K_i^0$ (II.44) с учетом слагаемого βv_0^2 очевидным образом.

Действительно, пусть $N = 3$ и

$$K_i = \begin{bmatrix} k_{11}^{(i)} & k_{12}^{(i)} \\ k_{12}^{(i)} & k_{22}^{(i)} \end{bmatrix}, \quad i = 1, 2, 3.$$

Тогда

$$F_1^h(v^3) = [v_0, v_1] \begin{bmatrix} k_{11}^{(1)} & k_{12}^{(1)} \\ k_{12}^{(1)} & k_{22}^{(1)} \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} + [v_1, v_2] \begin{bmatrix} k_{11}^{(2)} & k_{12}^{(2)} \\ k_{12}^{(2)} & k_{22}^{(2)} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} +$$

$$+ [v_2, v_3] \begin{bmatrix} k_{11}^{(3)} & k_{12}^{(3)} \\ k_{12}^{(3)} & k_{22}^{(3)} \end{bmatrix} \begin{bmatrix} v_2 \\ v_3 \end{bmatrix} - 2[v_0, v_1] \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \end{bmatrix} - 2[v_1, v_2] \begin{bmatrix} b_1^{(2)} \\ b_2^{(2)} \end{bmatrix} -$$

$$- 2[v_2, v_3] \begin{bmatrix} b_1^{(3)} \\ b_2^{(3)} \end{bmatrix} + \beta v_0^2 + 2v_0 g_1,$$

или

$$F_1^h(v^3) = [v_0, v_1, v_2, v_3] \begin{bmatrix} k_{11}^{(1)} + \beta & k_{12}^{(1)} & & \\ k_{12}^{(1)} & k_{22}^{(1)} + k_{11}^{(2)} & k_{12}^{(2)} & \\ & k_{12}^{(2)} & k_{22}^{(2)} + k_{11}^{(3)} & k_{12}^{(3)} \\ & & k_{12}^{(3)} & k_{22}^{(3)} \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \end{bmatrix} -$$

$$- 2[v_0, v_1, v_2, v_3] \begin{bmatrix} b_1^{(1)} - g_1 \\ b_2^{(1)} + b_1^{(2)} \\ b_2^{(2)} + b_1^{(3)} \\ b_2^{(3)} \end{bmatrix}.$$

Приведенный пример иллюстрирует и построение вектора b из элементарных векторов нагрузки $b_i = [b_1^i, b_2^i]^T$.

Схематически вид ленточной матрицы $\bar{K}$ показан на рис. 18 (заштрихованные части изображают суммирование соответствующих элементов матриц K_i, $i = 1 \div N$).

Иногда функцию F_1^h записывают в виде

$$F_1^h(v^N) = \omega^T \bar{K}^1 \omega + \omega^T \bar{K}^0 \omega - 2\omega^T \tilde{b} + \beta v_0^2 + 2v_0 g_1,$$

где матрицы $\bar{K}^1$ и $\bar{K}^0 \equiv \bar{M}$ формируются, как описано выше, из элементарных матриц жесткости K_i^1 и матриц масс элементов $K_i^0 \equiv M_i$

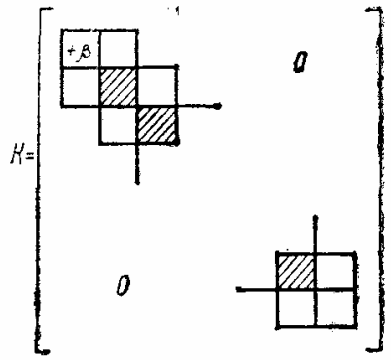

Рис. 18.

и называются матрицей жесткости и матрицей масс соответственно; матрицу $\bar{K} = \bar{K}^1 + \bar{K}^0$ называют глобальной (общей) матрицей жесткости.

Для определения значений параметров v_i, доставляющих минимум функции $F_1^h(v^N)$, ее дифференцируют по всем неизвестным v_i, $i = 0 \div (N-1)$, и полученные производные приравнивают нулю. (Напомним, что значение параметра v_N известно из краевого условия (II.33): $v_N = g_2$.) В результате получается система линейных алгебраических уравнений N-го порядка

$$Kz = b, \tag{II.46}$$

где

$$K = \begin{bmatrix} k_{11}^{(1)} + \beta & k_{12}^{(1)} & & & \\ k_{12}^{(1)} & k_{22}^{(1)} + k_{11}^{(1)} & k_{12}^{(2)} & 0 & \\ & \ddots & \ddots & \ddots & \\ & & k_{12}^{(N-2)} & k_{22}^{(N-2)} + k_{11}^{(N-1)} & k_{12}^{(N-1)} \\ & 0 & & k_{12}^{(N-1)} & k_{22}^{(N-1)} + k_{11}^{(N)} \end{bmatrix},$$

$$b = \begin{bmatrix} b_1^{(1)} - g_1 \\ b_2^{(1)} + b_1^{(2)} \\ \vdots \\ b_2^{(N-2)} + b_1^{(N-1)} \\ b_2^{(N-1)} + b_1^{(N)} - k_{12}^{(N)} g_2 \end{bmatrix}, \quad z = \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{N-1} \end{bmatrix}.$$

Формально переход от общей матрицы $\bar{K}$ и вектора $\bar{b}$ к матрице K и вектору b окончательной сеточной системы (II.46) состоял в том, что последний столбец матрицы $\bar{K}$, являющийся столбцом коэффициентов при известном параметре $v_N = g_2$, был умножен на g_2 и перенесен в правую часть системы (вычтен из вектора $\bar{b}$); кроме того, были отброшены последняя строка матрицы $\bar{K}$ и последний элемент $\bar{b}$, отвечающие в определенном смысле этому, уже известному, параметру v_N.

Систему уравнений (II.46) обычно называют системой уравнений метода конечных элементов. Проиллюстрируем описанную методику решением следующего примера.

Пример 1. Найти численное решение задачи

$$-\frac{d}{dx}\left((1+x^2)\frac{du}{dx}\right) + 12u = -e^x(1+x)^2 + 12e^x - 24, \quad 0 < x < 1,$$

$$\frac{du}{dx}(0) - \frac{1}{2}u(0) = \frac{3}{2}, \quad u(1) = e - 2.$$

Данная задача равнозначна задаче отыскания минимума функционала

$$F(v) = \int_0^1 \left((1+x^2)\left(\frac{dv}{dx}\right)^2 + 12v^2 - 2(12e^x - e^x(1+x)^2 - 24)v\right)dx + \\ + \frac{1}{2}v^2(0) + 3v(0)$$

на множестве непрерывных функций, имеющих интегрируемые с квадратом первые производные и удовлетворяющих условию $v(1) = e - 2$. Для получения приближенного решения область $[0, 1]$ разбивалась на четыре равных отрезка ($h = 0{,}25$) и использовался описанный выше элемент «$l1$—2». Построение соответствующей системы МКЭ и решение ее методом квадратных корней осуществлялось на ЭВМ МИР-2 при разрядности $R = 8$. Для вычисления коэффициентов системы использовались квадратурные формулы Гаусса. Полученные результаты приведены в табл. 1, где приняты следующие обозначения: $u_0(x_i)$ — значение точного решения в точке $x = x_i$; u_i — значение приближенного решения в узле x_i. Точное решение задачи — $u_0(x) = e^x - 2$.

2. Кусочно-квадратичные полиномы. Пусть теперь $v^N(x) \in P_2^h$, где $P_2^h \subset W_2^1(0, l)$ множество непрерывных кусочно-квадратичных функций, удовлетворяющих условию (II.33). Для однозначного определения полинома второй степени на элементарном отрезке $[x_{i-1}, x_i]$ и обеспечения непрерывности допустимой функции $v^N(x)$ на всем отрезке $[0, l]$ достаточно зафиксировать значение $v^N(x)$ на $[x_{i-1}, x_i]$ в трех узловых точках:

$$v_{i-1} = v^N(x_{i-1}),\ v_{i-1/2} = v^N(x_{i-1/2}),\ v_i = v^N(x_i),$$

где

$$x_{i-1/2} = \frac{x_{i-1} + x_i}{2}.$$

Описанный элемент обозначим как «l 2—3» и схематически представим в виде ●—●—●. Алгоритм получения соответствующей системы уравнений МКЭ

$$Kz = b,$$

решением которой являются значения искомого приближенного решения в узловых точках отрезка $[0, l]$

$$z = [u_0^N, u_{1/2}^N, u_1^N, \ldots, u_{N-1/2}^N]^T,$$

здесь совершенно такой же, как в предыдущем пункте. При этом

$$r(\xi) = \alpha_1 + \alpha_2\xi + \alpha_3\xi^2;\quad \alpha = S^{-1}\omega_i,$$

где

$$S^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 4 & -1 \\ 2 & -4 & 2 \end{bmatrix},\quad \omega_i = \begin{bmatrix} v_{i-1} \\ v_{i-1/2} \\ v_i \end{bmatrix}.$$

Таблица 1

x_i	$u_0(x_i)$	u_i
0	—1	—1,00655220
0,25	—0,71597458	—0,72331423
0,5	—0,35127873	—0,35871457
0,75	0,11700002	0,11148608

Матрица жесткости i-го элемента $K_i^1 = S^{-T}R_i^1S^{-1}$ в случае постоянных коэффициентов дифференциального уравнения теперь имеет вид

$$K_i^1 = \frac{k_0}{3h_i}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix},$$

а матрица массы элемента —

$$M_i \equiv K_i^0 = S^{-T}R_i^0S^{-1} = \frac{q_0h_i}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix}.$$

Построение матрицы K и вектора b общей системы уравнений МКЭ из элементарных матриц жесткости масс и элементарных векторов нагрузки выполняется так же, как в п. 1 параграфа II.2. В данном случае при разбиении отрезка $[0, l]$ на N элементарных отрезков получаем систему алгебраических уравнений порядка $2N$.

Пример 2. Найти численное решение краевой задачи примера 1, используя в качестве допустимых функций кусочно-квадратичные полиномы.

Область $[0, 1]$ в данном случае разбивалась на четыре равных элементарных отрезка, т. е. по-прежнему $h = 0{,}25$, но узловых точек теперь почти вдвое больше. Расчет выполнялся так же, как в примере 1, на ЭВМ МИР-2, $R = 8$. Полученные значения приближенного решения приведены в табл. 2.

3. Кусочно-кубические допустимые функции. Остановимся кратко и на использовании кусочно-кубических полиномов в качестве допустимых функций при минимизации функционала (II.32). Здесь можно различать элементы двух видов: «$l3$—4» и «$l3$—2».

В первом случае на элементарном отрезке $[x_{i-1}, x_i]$ кубический полином однозначно определяется фиксированием его значений в четырех узлах:

$$v_{i-1} = v^N(x_{i-1}), \quad v_{i-2/3} = v^N(x_{i-2/3}), \quad v_{i-1/3} = v^N(x_{i-1/3}), \quad v_i = v^N(x_i),$$

где

$$x_{i-2/3} = \frac{2x_{i-1} + x_i}{3}, \quad x_{i-1/3} = \frac{x_{i-1} + 2x_i}{3}.$$

При этом допустимая функция $v^N(x)$ на всем отрезке $[0, l]$ будет непрерывна и кусочно-дифференцируема, т. е.

$$v^N(x) \in P_3^h \subset W_2^1(0, l).$$

Схематический элемент «l 3 —4» будем изображать в виде ●—●—●—● и называть кубическим элементом Лагранжа. Во втором случае функция $v^N(x)$ на каждом отрезке $[x_{i-1}, x_i]$ однозначно определяется условиями

$$v_{i-1} = v^N(x_{i-1}), \quad v'_{i-1} = \frac{dv^N}{dx}\Big|_{x=x_{i-1}}, \quad v_i = v^N(x_i), \quad v'_i = \frac{dv^N}{dx}\Big|_{x=x_i}.$$

Таким образом, допустимые функции в этом случае оказываются не только непрерывными, но и непрерывно дифференцируемыми на всем отрезке $[0, l]$, т. е. $v^N(x) \in P_3^h \subset W_2^2(0, l) \subset W_2^1(0, l)$. Схематически данный элемент будем изображать в виде v, v' ●—● v, v' и называть кубическим элементом Эрмита.

Таблица 2

x_i	$u_0(x_i)$	u_i
0	—1	—0,99997980
0,125	—0,86685155	—0,86684281
0,25	—0,71597458	—0,71596147
0,375	—0,54500859	—0,54501700
0,5	—0,35127873	—0,35126943
0,625	—0,13175404	—0,13177631
0,75	0,11700002	0,11701080
0,875	0,39887529	0,39883981

Дальнейшие рассуждения по построению элементарных матриц жесткости и масс, векторов нагрузки, а также формированию из них соответствующих систем уравнений МКЭ проводятся аналогично предыдущим.

Приведем здесь лишь некоторые результаты, относящиеся к элементу «l3—2». На «каноническом отрезке», полученном преобразованием (II.38), допустимая функция имеет вид

$$v^N(x_{i-1} + h_i\xi) \equiv r(\xi) = \alpha_1 + \alpha_2\xi + \alpha_3\xi^2 + \alpha_4\xi^3.$$

При этом

$$v_{i-1} = r(0) = \alpha_1,$$

$$v'_{i-1} = r_\xi(0)\frac{1}{h_i} = \alpha_2\frac{1}{h_i},$$

$$v_i = r(1) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4,$$

$$v'_i = r_\xi(1)\frac{1}{h_i} = (\alpha_2 + 2\alpha_3 + 3\alpha_4)\frac{1}{h_i},$$

где

$$r_\xi = \frac{dr}{d\xi}, \quad v' \equiv \frac{dv^N}{dx} = \frac{dr}{d\xi}\frac{d\xi}{dx}.$$

Таким образом,

$$\omega_i = \Pi_i S\alpha.$$

Здесь

$$\Pi_i = \begin{bmatrix} 1 & & & \\ & \frac{1}{h_i} & & \\ & & 1 & \\ & & & \frac{1}{h_i} \end{bmatrix}, \quad S = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix}, \quad \omega_i = \begin{bmatrix} v_{i-1} \\ v_{i-1} \\ v_i \\ v_i \end{bmatrix}. \tag{II.47}$$

Следовательно,

$$\alpha = S^{-1}\Pi_i^{-1}\omega_i,$$

где

$$S^{-1}=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -3 & -2 & 3 & -1 \\ 2 & 1 & -2 & 1 \end{bmatrix},\quad \Pi_i^{-1}=\begin{bmatrix} 1 & & & \\ & h_i & & \\ & & 1 & \\ & & & h_i \end{bmatrix}. \tag{II.48}$$

Элементарные матрицы жесткости K_i^1 и масс $K_i^0 \equiv M_i$ строятся способом, подробно описанным в п. 1 параграфа II.2, и имеют вид

$$K_i^1=\Pi_i^{-1}S^{-T}R_i^1S^{-1}\Pi_i^{-1},\quad K_i^0=\Pi_i^{-1}S^{-T}R_i^0S^{-1}\Pi_i^{-1}\equiv M_i.$$

Здесь

$$R_i^1=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \int_0^1 \tilde{k}d\xi & 2\int_0^1 \xi\tilde{k}d\xi & 3\int_0^1 \xi^2\tilde{k}d\xi \\ 0 & 2\int_0^1 \xi\tilde{k}d\xi & 4\int_0^1 \xi^2\tilde{k}d\xi & 6\int_0^1 \xi^3\tilde{k}d\xi \\ 0 & 3\int_0^1 \xi^2\tilde{k}d\xi & 6\int_0^1 \xi^3\tilde{k}d\xi & 9\int_0^1 \xi^4\tilde{k}d\xi \end{bmatrix},$$

$$R_i^0=\begin{bmatrix} \int_0^1 \tilde{q}d\xi & \int_0^1 \xi\tilde{q}d\xi & \int_0^1 \xi^2\tilde{q}d\xi & \int_0^1 \xi^3\tilde{q}d\xi \\ \int_0^1 \xi\tilde{q}d\xi & \int_0^1 \xi^2\tilde{q}d\xi & \int_0^1 \xi^3\tilde{q}d\xi & \int_0^1 \xi^4\tilde{q}d\xi \\ \int_0^1 \xi^2\tilde{q}d\xi & \int_0^1 \xi^3\tilde{q}d\xi & \int_0^1 \xi^4\tilde{q}d\xi & \int_0^1 \xi^5\tilde{q}d\xi \\ \int_0^1 \xi^3\tilde{q}d\xi & \int_0^1 \xi^4\tilde{q}d\xi & \int_0^1 \xi^5\tilde{q}d\xi & \int_0^1 \xi^6\tilde{q}d\xi \end{bmatrix}.$$

В случае постоянных коэффициентов $k(x)\equiv k_0$, $q(x)\equiv q_0$

$$K_i^0=\frac{q_0h_i}{420}\begin{bmatrix} 156 & 22h_i & 54 & -13h_i \\ 22h_i & 4h_i^2 & 13h_i & -3h_i^2 \\ 54 & 13h_i & 156 & -22h_i \\ -13h_i & -3h_i^2 & -22h_i & 4h_i^2 \end{bmatrix}\equiv M_i,$$

$$K_i^1=\frac{k_0}{30h_i}\begin{bmatrix} 36 & 3h_i & -36 & 3h_i \\ 3h_i & 4h_i^2 & -3h_i & -h_i^2 \\ -36 & -3h_i & 36 & -3h_i \\ 3h_i & -h_i^2 & -3h_i & 4h_i^2 \end{bmatrix}.$$

Таблица 3

x_0	$u_0(x_i)$	u_i	u_i'	$u_0'(x_i)$
0	—1	—0,99999760	1,0002970	1
0,25	—0,71597458	—0,71596577	1,2841615	1,28302542
0,5	—0,35127873	—0,35126443	1,6486437	1,64872127
0,75	0,11700002	0,11674656	2,1115582	2,11700002
1	0,71828183	—	2,7173658	2,71828183

Система уравнений МКЭ, сформированная из элементарных матриц жесткости, масс и векторов нагрузки,

$$Kz = b,$$

при разбиении отрезка $[0, l]$ на N элементов вида «l 3—2» имеет порядок $2N + 1$. Решение ее определяет следующие значения приближенного решения $u^N(x)$:

$$z = [u_0^N, (u_0^N)', u_1^N, (u_1^N)', \ldots, u_{N-1}^N, (u_{N-1}^N)', (u_N^N)']^T,$$

где $(u_i^N)' = \left.\frac{du^N}{dx}\right|_{x=x_i}$.

Пример 3. Решить краевую задачу примера 1 с использованием элементов вида «l 3—2».

Для вычисления приближенного решения использовалась равномерная сетка с $h = 0{,}25$. Полученные значения приближенного решения u_i и соответствующие значения производной u_i' в узлах сетки представлены в табл. 3. Расчет выполнялся, как и прежде, на ЭВМ МИР-2, $R = 8$.

4. Дискретизация задач с разрывными решениями. Построение методом конечных элементов приближенного решения краевой задачи с разрывными коэффициентами и решениями (см. п. 2 параграфа II.1) выполняется аналогично описанному в трех предыдущих пунктах. Отметим только некоторые особенности (в предположении наличия одной точки разрыва $x = \xi$).

Разбиение исходного отрезка $[0, l]$ на N элементарных отрезков $[x_{i-1}, x_i]$, $i = 1 \div N$, необходимо выполнять так, чтобы точка $x = \xi$ разрыва решения (см. условие (II.16)) совпадала с некоторым граничным узлом x_p, $0 < p < N$, двух соседних элементов $[x_{p-1}, x_p]$ и $[x_p, x_{p+1}]$. Кроме того, определяя допустимую функцию $v^N(x)$ на этих соседних элементах, в узле $x_p = \xi$ нужно фиксировать два значения приближенного решения:

$$v_p^+ = v^N(\xi + 0), \quad v_p^- = v^N(\xi - 0).$$

При этом, конечно, следует позаботиться о том, чтобы построенная на всем отрезке $[0, l]$ кусочно-полиномиальная функция $v^N(x)$ принадлежала пространству функций, на котором достигается минимум соответствующего функционала (II.26).

В остальном построение элементарных матриц жесткости масс и векторов нагрузки, а также формирование системы уравнений МКЭ не имеет принципиальных отличий. Заметим, что матрица сеточной системы в данном случае, сохраняя ленточный вид, будет состоять из двух отдельных блоков, отвечающих отрезкам $[0, \xi]$ и $[\xi, l]$.

Пример 4. Найти численное решение краевой задачи

$$-\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) = f(x), \quad 0 < x < 1,$$

$$u(0) = 0, \quad u(1) = 0,$$

$$\left[k\frac{du}{dx}\right]_{x=\frac{1}{3}} = 0, \quad [u]_{x=\frac{1}{3}} = \frac{1}{9}\left(k\frac{du}{dx}\right)_{x=\frac{1}{3}},$$

где

$$k(x) = \begin{cases} 0{,}6, & 0 \leqslant x < 1/3, \\ 4, & 1/3 < x \leqslant 1, \end{cases} \quad f(x) = \begin{cases} -76{,}8\pi\sqrt{3}\cos 16\pi x, & 0 \leqslant x < 1/3, \\ 24\pi \sin 2\pi x, & 1/3 < x \leqslant 1. \end{cases}$$

Точное решение задачи следующее:

$$u_0(x) = \begin{cases} \frac{\sqrt{3}}{\pi}\sin^2 8\pi x + 2x, & 0 \leqslant x < 1/3, \\ \frac{1{,}5}{\pi}\sin 2\pi x, & 1/3 < x \leqslant 1. \end{cases}$$

Решение данной дифференциальной задачи эквивалентно отысканию функции, минимизирующей функционал

$$F(v) = \int_0^1 \left[k(x)\left(\frac{dv}{dx}\right)^2 - 2fv\right] dx + 9\,[v]^2_{x=1/3}$$

на множестве функций

$$v(x) = \begin{cases} v_1(x), & 0 \leqslant x < 1/3, \\ v_2(x), & 1/3 < x \leqslant 1, \end{cases}$$

где

$$v_1(x) \in W_2^1(0, 1/3), \quad v_1(0) = 0, \quad v_2(x) \in W_2^1(1/3, 1), \quad v_2(1) = 0.$$

Для построения приближенного решения $v^N(x)$ разобьем отрезок $[0, 1]$ следующим образом:

$$x_i = \frac{1}{24}i, \quad i = 0 \div 8; \quad x_i = \frac{1}{6}(i-6), \quad i = 8 \div 12.$$

Таким образом, отрезок $[0, 1]$ представлен как объединение 12 элементов, причем на участке $[0,1/3]$ длина каждого элемента $h_1 = 1/24$, а на участке $[1/3, 1]$ — $h_2 = 1/6$. Такой выбор сетки диктуется требованием выявить особенности поведения решения, быстро осциллирующего на отрезке $[0,1/3]$.

В качестве допустимых функций в данном примере использовались кусочно-линейные полиномы. Как свидетельствуют результаты, пред-

Таблица 4

x_i	$u_0(x_i)$	u_i
1/24	0,4968300035	0,4968299988
1/12	0,5801633395	0,5801633237
1/8	0,25	0,2499999806
1/6	0,7468300004	0,7468299875
5/24	0,8301633424	0,8301633118
1/4	0,5	0,4999999751
7/24	0,9968299970	0,9968299889
1/3	1,0801633440	1,0801633260
1/3	0,4134966721	0,4134966664
1/2	0	$-0{,}331283\cdot10^{-8}$
2/3	—0,4134966721	—0,4134966738
5/6	—0,4134966721	—0,4134966719

ставленные в табл. 4, линейные полиномы в этой задаче обеспечили получение практически точного решения. Расчет выполнялся на ЭВМ МИР-2 при разрядности $R = 10$.

II.3. Обоснование метода конечных элементов

В данном параграфе будет исследована сходимость приближенного решения, полученного методом конечных элементов, к точному решению соответствующей задачи. Кроме того, будут изложены некоторые практические оценки точности вычисленного на ЭВМ решения.

1. Сходимость МКЭ. Для простоты и ясности изложения остановимся на задаче о нахождении решения уравнения

$$-\frac{d}{dx}\left(k(x)\frac{du}{dx}\right)+q(x)u=f(x),\quad 0<x<l, \tag{II.49}$$

удовлетворяющего краевым условиям

$$u(0)=u(l)=0. \tag{II.50}$$

Ограничения на коэффициенты и правую часть такие же, как в п. 1 параграфа II.1. Эта задача эквивалентна задаче об отыскании функции, доставляющей минимум функционалу

$$F(v)=[v,\,v]_A-2(f,\,v)\equiv\int_0^l\left(k\left(\frac{dv}{dx}\right)^2+qv^2-2fv\right)dx \tag{II.51}$$

в энергетическом пространстве H_A оператора задачи (II.49), (II.50). В данном случае H_A совпадает с пространством функций $\overset{0}{W}{}_2^1(0,\,l)\subset W_2^1(0,\,l)$.

Обозначим через $u(x)$ функцию, доставляющую минимум функционалу $F(v)$. (Заметим, что $u(x)$ является обобщенным (или классическим) решением задачи (II.49), (II.50).) Пусть $u^N(x)$ — приближенное решение, полученное методом конечных элементов по методике,

описанной в параграфе II.2 (вариант метода Ритца). Для конкретности будем предполагать, что $u^N(x)$ доставляет минимум функционалу (II.51) на множестве функций из конечномерного подпространства $\overset{0}{P}{}_1^h \subset \overset{0}{W}{}_2^1(0, l)$, т. е. подпространства кусочно-линейных полиномов, соответствующих элементу «l1—2» и принимающих нулевые значения в точках $x = 0$ и $x = l$.

Оценим близость $u^N(x)$ к точному решению $u(x)$ в метрике пространства $W_2^1(0, l)$, которому принадлежат обе функции. Для этого рассмотрим погрешность $u(x) - u^N(x)$ вначале в энергетической норме

$$\|u - u^N\|_A^2 = [u - u^N, u - u^N]_A = \int_0^l \left(k\left(\frac{du}{dx} - \frac{du^N}{dx}\right)^2 + q(u - u^N)^2\right) dx,$$

а поскольку энергетическое пространство в данном случае состоит из функций, принадлежащих $\overset{0}{W}{}_2^1(0, l)$, и нормы в этих пространствах эквивалентны [68], то из сходимости в энергетической норме сразу будет следовать сходимость в норме пространства $W_2^1(0, l)$. (Напомним, что нормы $\|\ \|_1$ и $\|\ \|_2$, введенные для элементов пространства $\mathfrak{B}$, называются эквивалентными, если для всех $u \in \mathfrak{B}$

$$c_2\|u\|_2 \leqslant \|u\|_1 \leqslant c_1\|u\|_2, \quad c_1,\ c_2 > 0 \text{ — постоянные.)}$$

Как показано в [101],

$$\|u - u^N\|_A^2 = \min_{v^N \in \overset{0}{P}{}_1^h} \|u - v^N\|_A^2. \tag{II.52}$$

Действительно, если u^N минимизирует функционал (II.51) на множестве функций $v^N \in \overset{0}{P}{}_1^h$, то для всех v^N и произвольного числа ε будем иметь

$$F(u^N) \leqslant F(u^N + \varepsilon v^N) = [u^N + \varepsilon v^N, u^N + \varepsilon v^N]_A - 2(f, u^N + \varepsilon v^N) =$$
$$= [u^N, u^N]_A - 2(f, u^N) + 2\varepsilon([u^N, v^N]_A - (f, v^N)) + \varepsilon^2[v^N, v^N]_A =$$
$$= F(u^N) + 2\varepsilon([u^N, v^N]_A - (f, v^N)) + \varepsilon^2[v^N, v^N]_A,$$

откуда следует

$$0 \leqslant 2\varepsilon([u^N, v^N]_A - (f, v^N)) + \varepsilon^2[v^N, v^N]_A.$$

Так как ε — произвольное число любого знака, то

$$[u^N, v^N]_A = (f, v^N), \quad \forall v^N \in \overset{0}{P}{}_1^h. \tag{II.53}$$

В случае минимизации функционала $F(v)$ на всем энергетическом пространстве $\overset{0}{W}{}_2^1(0, l)$ условие (II.53) принимает вид

$$[u, v]_A = (f, v) \text{ для всех } v \in \overset{0}{W}{}_2^1(0, l). \tag{II.54}$$

(Ср. с другой формой определения обобщенного решения, упомянутой в п. 2 параграфа II.1, а именно через интегральное тождество

(II.28).) Далее, поскольку равенство (II.54) справедливо при $\forall v \in \overset{c}{W}{}_2^1(0, l)$, а $v^N \in \overset{0}{P}{}_1^h \subset \overset{0}{W}{}_2^1(0, l)$, имеем

$$[u, v^N]_A = (f, v^N), \quad \forall v^N \in \overset{0}{P}{}_1^h. \tag{II.55}$$

В результате, вычитая (II.55) из (II.53), получаем

$$[u - u^N, v^N]_A = 0, \quad \forall v^N \in \overset{0}{P}{}_1^h. \tag{II.56}$$

Наконец, чтобы убедиться в справедливости (II.52), рассмотрим энергетическое произведение

$$[u - u^N - v^N, u - u^N - v^N]_A = \| u - u^N \|_A^2 - 2[u - u^N, v^N]_A + \| v^N \|_A^2.$$

Согласно (II.56) $\| u - u^N - v^N \|_A^2 \geqslant \| u - u^N \|_A^2$ для любой функции $v^N \in \overset{0}{P}{}_1^h$, причем равенство достигается только в случае $\| v^N \|_A^2 = 0$, т. е. при $v^N = 0$. Поэтому можно утверждать, что минимальное значение выражения $\| u - v^N \|_A^2$ при любой функции $v^N \in \overset{0}{P}{}_1^h$ достигается лишь в случае $v^N = u^N$, т. е. справедливость (II.52) установлена. Таким образом, имеем

$$\| u - u^N \|_A^2 = \min_{v^N \in \overset{0}{P}{}_1^h} \| u - v^N \|_A^2 \leqslant \| u - u_I^N \|_A^2, \tag{II.57}$$

где u_I^N — интерполянт функции $u(x)$ из подпространства $\overset{0}{P}{}_1^h$, а не приближенное решение МКЭ.

Иными словами, $u_I^N(x)$ — кусочно-линейный полином, принимающий в узлах сетки x_i, $i = 0, 1, \ldots, N$, одинаковые с $u(x)$ значения: $u_I^N(x_i) = u(x_i)$, т. е. $u_I^N(x)$ на каждом элементе $[x_{i-1}, x_i]$ является лагранжевым линейным интерполянтом функции $u(x)$.

В дальнейшем нам понадобится следующая элементарная лемма.

Лемма II.1. *Пусть непрерывно дифференцируемая на $[a, b]$ функция $\varphi(x)$ имеет в (a, b) ограниченную вторую производную*

$$\left| \frac{d^2\varphi}{dx^2} \right| \equiv |\varphi''(x)| \leqslant C_2 \quad и \quad \varphi(a) = \eta_1, \ \varphi(b) = \eta_2.$$

Тогда

$$|\varphi(x)| \leqslant \max_i |\eta_i| + \frac{1}{8} C_2 (b - a)^2, \tag{II.58}$$

$$\left| \frac{d\varphi}{dx}(x) \right| \leqslant \frac{|\eta_1| + |\eta_2|}{b - a} + C_2 (b - a), \quad \frac{d\varphi}{dx} \equiv \varphi'(x).$$

Доказательство. Пусть $p_1(x) = \eta_1 \frac{x - b}{a - b} + \eta_2 \frac{x - a}{b - a}$ — интерполяционный полином Лагранжа функции $\varphi(x)$, $x \in [a, b]$.

Тогда можно записать

$$\varphi(x) = p_1(x) + \psi(x), \quad x \in [a, b],$$

где $\psi(x)$ — остаточный член формулы Лагранжа, т. е.

$$\psi(x) = \frac{\varphi''(\xi)}{2} (x - a)(x - b), \quad a \leqslant \xi \leqslant b,$$

для которого в данном случае справедлива оценка

$$|\psi(x)| \leqslant \frac{1}{8} C_2 (b-a)^2.$$

Так как $|p_1(x)| \leqslant \max\limits_i |\eta_i|$, $a \leqslant x \leqslant b$, оценка (II.58) очевидна.

Далее, $\psi(a) = \psi(b) = 0$ и $|\psi''(x)| = |\varphi''(\xi)| \leqslant C_2$, поэтому согласно теореме Ролля на $[a, b]$ существует точка ξ_1, где $\psi'(\xi_1) = 0$, а согласно формуле конечных приращений можно записать

$$\psi'(x) = \psi'(\xi_1) + \psi''(\xi_2)(x-\xi_1), \quad \xi_2 = \xi_1 + \theta(x-\xi_1), \quad 0 < \theta < 1.$$

Таким образом, имеем

$$|\varphi'(x)| \leqslant |p_1'(x)| + |\psi_1'(x)| \leqslant \frac{|\eta_1| + |\eta_2|}{b-a} + C_2(b-a),$$

что и требовалось доказать.

Вернемся теперь к неравенству (II.57) и оценим величину $\|u - u_l^N\|_A^2$. Пусть $u - u_l^N \equiv \eta(x)$. Тогда

$$\|u - u_l^N\|_A^2 \equiv \|\eta\|_A^2 = \int_0^l \left(k\left(\frac{d\eta}{dx}\right)^2 + q\eta^2\right) dx.$$

В соответствии с введенной сеткой, т. е. с разбиением отрезка $[0, l]$ на N элементов вида «l1—2», последний интеграл представим как

$$\|\eta\|_A^2 = \int_0^l \left(k\left(\frac{d\eta}{dx}\right)^2 + q\eta^2\right) dx = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \left(k\left(\frac{d\eta}{dx}\right)^2 + q\eta^2\right) dx$$

и исследуем поведение функции $\eta(x)$ на сегменте $[x_{i-1}, x_i]$ в предположении, что точное решение $u(x)$ имеет ограниченную в $(0, l)$ вторую производную

$$\left|\frac{d^2u}{dx^2}\right| \leqslant M_2.$$

При этом оказывается, что функция $\eta(x)$ на любом отрезке $[x_i, x_{i-1}]$ удовлетворяет всем условиям леммы II.1, т. е. $\eta(x) = u(x) - u_l^N(x)$ есть функция, непрерывно дифференцируемая на $[0, l]$, $\eta(x_{i-1}) = \eta(x_i) = 0$, $\left|\frac{d^2\eta}{dx^2}\right| = \left|\frac{d^2u}{dx^2}\right| \leqslant M_2$, следовательно, справедлива оценка

$$\|\eta\|_A^2 \leqslant \sum_{i=1}^{N} \left(M_2^2 h_i^2 \int_{x_{i-1}}^{x_i} k\,dx + \frac{M_2^2}{64} h_i^4 \int_{x_{i-1}}^{x_i} q\,dx\right) \leqslant M_2^2 C^2 h^2,$$

где $h = \max h_i$, $h_i = x_i - x_{i-1}$, константа C^2 не зависит от h.

Так как $\|u - u^N\|_A^2 \leqslant \|\eta\|_A^2$, полученная оценка показывает, что при измельчении сетки ($h \to 0$) приближенное решение $u^h(x)$ сходится к точному $u(x)$ со скоростью $O(h)$ в энергетической норме, а следова-

тельно, и в норме $\|\ \|_{2,1}$ пространства $W_2^1(0, l)$. Действительно,

$$\|u-u^N\|_A^2=\int_0^l\left(k\left(\frac{du}{dx}-\frac{du^N}{dx}\right)^2+q(u-u^N)^2\right)dx\geqslant$$

$$\geqslant k_0\int_0^l\left(\frac{du}{dx}-\frac{du^N}{dx}\right)^2dx+q_0\int_0^l(u-u^N)^2\,dx.$$

Здесь $q_0=\min\limits_{0\leqslant x\leqslant l} q\ (x)$.

Если $q_0\neq 0$, то

$$\|u-u^N\|_A^2\geqslant m_0\int_0^l\left(\left(\frac{d}{dx}(u-u^N)\right)^2+(u-u^N)^2\right)dx=m_0\|u-u^N\|_{2,1}^2,$$

где $m_0=\min(k_0, q_0)$. Если $q_0=0$, то, использовав неравенство

$$\int_0^l\left(\frac{d\zeta}{dx}\right)^2dx\geqslant\frac{\pi^2}{l^2}\int_0^l\zeta^2dx,$$

справедливое для любой функции $\zeta(x)\in\overset{0}{W}{}_2^1(0, l)$, получим

$$\|u-u^N\|_A^2\geqslant m_1\|u-u^N\|_{2,1}^2,$$

где $m_1=\min\left(\frac{k_0}{2}, \frac{\pi^2k_0}{2l^2}\right)$.

Таким образом, суммируя результаты, можно утверждать справедливость следующей теоремы.

Теорема II.1. *Пусть функция $u(x)$, доставляющая минимум функционалу (II.51) в пространстве $\overset{0}{W}{}_2^1(0, l)$, имеет в $(0, l)$ ограниченную вторую производную*

$$\left|\frac{d^2u}{dx^2}\right|\leqslant M_2. \tag{II.59}$$

Тогда для приближенного решения $u^N(x)$, полученного методом конечных элементов с использованием кусочно-линейных допустимых функций, справедлива оценка

$$\|u-u^N\|_{2,1}\leqslant C_1M_2h,$$

где C_1 — постоянная, не зависящая от максимального шага h введенной сетки.

Нетрудно проверить, что в наших прежних обозначениях $C_1=C\sqrt{G}$, где

$$G=\begin{cases}\frac{1}{m_0}, & m_0=\min(k_0, q_0), \quad \text{при } q_0\neq 0,\\ \frac{1}{m_1}, & m_1=\min\left(\frac{k_0}{2}, \frac{k_0\pi^2}{2l^2}\right) \quad \text{при } q_0=0,\end{cases}$$

$$C\leqslant\left(\int_0^l(k+q)\,dx\right)^{1/2} \quad \text{при } h\leqslant 8.$$

Аналогичным образом можно исследовать сходимость МКЭ при использовании других видов кусочно-полиномиальных допустимых функций. Полученные результаты обобщаются одной теоремой.

Теорема II.2. *Пусть функция* $u(x)$ *минимизирует функционал (II.51) в пространстве* $\overset{0}{W}{}_2^1(0, l)$, *а* $u^N(x)$ — *приближенное решение, минимизирующее функционал (II.51) на конечномерном подпространстве* $\overset{0}{P}{}_n^h \subset \overset{0}{W}{}_2^1(0, l)$ *кусочно-полиномиальных допустимых функций степени* n. *Если* $u(x)$ *имеет в* $(0, l)$ *ограниченную производную порядка* $n+1$:

$$\left|\frac{d^{n+1}u}{dx^{n+1}}\right| \leqslant M_{n+1},$$

то справедлива оценка

$$\|u - u^N\|_{2,1} \leqslant C_n M_{n+1} h^n,$$

где C_n — *постоянная, не зависящая от* h.

Результаты теоремы II.2 при более слабых ограничениях, а именно в предположении, что искомое обобщенное решение $u(x) \in \bar{\in} W_2^{n+1}(0, l)$, непосредственно следуют при учете соотношения (II.57) из теоремы об аппроксимации функций подпространствами P_n^h [101]. В параграфе II.6 мы подробнее остановимся на этих более общих результатах.

Утверждения теоремы II.2 остаются справедливыми и в случае одномерных краевых задач с неоднородными условиями (см., например, [70]). В указанной работе исследуется сходимость приближенного решения краевой задачи с неоднородными граничными условиями, полученного посредством элементов вида «*l*3—2». Случай краевых задач с разрывными коэффициентами и решениями рассмотрен в работе [69]. Однако при этом решение $u(x)$ предполагается кусочно дифференцируемым и условие (II.59) выполняется на каждом отрезке непрерывности. Кроме того, для построения приближенного решения сетка вводится так, чтобы точки разрыва являлись концевыми узлами элементов.

Замечание. При выполнении условия $\frac{d^{n+1}u}{dx^{n+1}} \in L_2(0, l)$ доказано, что скорость сходимости u^N к $u(x)$ в норме $L_2(0, l)$ ($\|u - u^N\|_{L_2}$) имеет порядок $O(h^{n+1})$. Подробное изложение этого факта дано в работе [101] (см. также параграф II.6).

2. Учет ошибок численного интегрирования в МКЭ. Указанная в теореме II.2 скорость сходимости МКЭ получена в предположении, что все вычисления при построении системы уравнений МКЭ и при решении системы выполнялись точно. Однако на практике, как правило, в полученном решении присутствуют ошибки, возникающие как от приближенного вычисления соответствующих интегралов (особенно при переменных коэффициентах), так и от вычисления на ЭВМ решения сеточной системы уравнений. В настоящем пункте рассмотрим влияние на скорость сходимости МКЭ ошибок в интегрировании.

В случае одномерных задач, особенно для уравнений второго порядка, вычисление встречающихся в МКЭ интегралов можно осуществлять двумя способами: либо заменяя все переменные коэффициенты и правую часть интерполирующими полиномами, а потом точно интегрируя полиномы, либо применяя с самого начала какую-нибудь стандартную квадратурную формулу, например Гаусса или Ромберга.

Разберем вначале первую возможность: замена всех переменных коэффициентов и правой части соответствующими интерполяционными полиномами. Возникает вопрос, к каким искажениям приведет такая замена, если предположить, что все последующие вычисления приближенного решения $u^N(x)$ будут выполняться точно.

Рассмотрим этот вопрос на примере минимизации функционала (II.51). Если при построении приближенного решения в функционале (II.51) заменить функции $k(x)$, $q(x)$ и $f(x)$ соответствующими интерполянтами $\tilde{k}(x)$, $\tilde{q}(x)$ и $\tilde{f}(x)$, то в результате придем, по существу, к задаче минимизации нового функционала

$$F(v) = [v, v]_A^* - 2(\tilde{f}, v) = \int_0^l \left(\tilde{k}\left(\frac{dv}{dx}\right)^2 + \tilde{q}v^2 - 2\tilde{f}v\right) dx \qquad \text{(II.60)}$$

на множестве (прежних) допустимых функций из конечномерного подпространства $\overset{0}{P}{}_n^h \subset \overset{0}{W}{}_2^1(0, l)$.

Предположим, что функционалу (II.60) доставляет минимум функция $\tilde{u}^N(x)$, так что аналогично (II.53) справедливо соотношение

$$[\tilde{u}^N, v^N]_A^* = (\tilde{f}, v^N),$$

т. е.

$$\int_0^l \left(\tilde{k}\frac{d\tilde{u}^N}{dx}\frac{dv^N}{dx} + \tilde{q}\tilde{u}^N v^N\right) dx = \int_0^l \tilde{f}v^N dx, \quad \forall\, v^N \in \overset{0}{P}{}_n^h. \qquad \text{(II.61)}$$

Напомним, что для функции $u^N(x)$, доставляющей минимум функционалу (II.51) на $\overset{0}{P}{}_n^h$, соотношение (II.53) имеет вид

$$\int_0^l \left(k\frac{du^N}{dx}\frac{dv^N}{dx} + qu^N v^N\right) dx = \int_0^l fv^N dx, \quad \forall\, v^N \in \overset{0}{P}{}_n^h. \qquad \text{(II.62)}$$

Учитывая (II.61) и (II.62), нетрудно убедиться (непосредственной проверкой) в справедливости тождества

$$\int_0^l \left(\tilde{k}\left(\frac{du^N}{dx} - \frac{d\tilde{u}^N}{dx}\right)^2 + \tilde{q}(u^N - \tilde{u}^N)\right)^2 dx =$$

$$= \int_0^l \left((\tilde{k} - k)\frac{du^N}{dx}\left(\frac{du^N}{dx} - \frac{d\tilde{u}^N}{dx}\right) + (\tilde{q} - q)u^N(u^N - \tilde{u}^N) - \right.$$

$$\left. - (\tilde{f} - f)(u^N - \tilde{u}^N)\right) dx. \qquad \text{(II.63)}$$

Если $\tilde{k}(x) > 0$ и $\tilde{q} \geqslant 0$, то, повторяя рассуждения из доказательства теоремы II.1, легко получить следующую оценку снизу для левой части ($I_л$) данного тождества:

$$\omega \| u^N - \tilde{u}^N \|_{2,1}^2 \leqslant I_л, \tag{II.64}$$

где ω — некоторая положительная константа.

Согласно неравенству Коши — Буняковского правая часть тождества (II.63) оценивается сверху так:

$$I_л \leqslant \left(\int_0^l (\tilde{k} - k)^2 \left(\frac{du^N}{dx} \right)^2 dx \right)^{1/2} \left(\int_0^l \left(\frac{du^N}{dx} - \frac{d\tilde{u}^N}{dx} \right)^2 dx \right)^{1/2} +$$

$$+ \left(\int_0^l (\tilde{q} - q)^2 (u^N)^2 dx \right)^{1/2} \left(\int_0^l (u^N - \tilde{u}^N)^2 dx \right)^{1/2} +$$

$$+ \left(\int_0^l (\tilde{f} - f)^2 dx \right)^{1/2} \left(\int_0^l (u^N - \tilde{u}^N)^2 dx \right)^{1/2} \leqslant \left(\left(\int_0^l (\tilde{k} - k)^2 \left(\frac{du^N}{dx} \right)^2 dx \right)^{1/2} + \right.$$

$$\left. + \left(\int_0^l (\tilde{q} - q)^2 (u^N)^2 dx \right)^{1/2} + \left(\int_0^l (\tilde{f} - f)^2 dx \right)^{1/2} \right) \| u^N - \tilde{u}^N \|_{2,1}. \tag{II.65}$$

Если погрешность замены на каждом элементе $[x_{i-1}, x_i]$ функций $k(x), q(x), f(x)$ их интерполянтами $\tilde{k}(x)$, $\tilde{q}(x)$, $\tilde{f}(x)$ имеет порядок h^p, $h = \max\limits_i (x_i - x_{i-1})$, то из (II.63) — (II.65) следует

$$\| u^N - \tilde{u}^N \|_{2,1} \leqslant \frac{C}{\omega} h^p,$$

где C — некоторая положительная константа. Таким образом, можно утверждать, что в «искаженном» решении $\tilde{u}^N(x)$ погрешность, возникающая от замены переменных коэффициентов и правой части уравнения их интерполяционными полиномами, имеет в норме W_2^1 тот же порядок, что и погрешность интерполяционной формулы. В частности, если при использовании линейных элементов вида «l 1—2» функции $k(x)$, $q(x)$, $f(x)$ на каждом элементе заменяются линейными интерполянтами, для которых погрешность интерполяции имеет порядок h^2, то будет справедлива оценка

$$\| u^N - \tilde{u}^N \|_{2,1} \leqslant Ch^2.$$

Иными словами, в данном случае скорость убывания погрешности, вызванной неточным интегрированием соответствующих величин, будет выше, чем скорость сходимости МКЭ для линейных элементов (см. (II.57)). К аналогичному заключению приходим при использовании других элементов и соответствующих интерполянтов для переменных коэффициентов и функции $f(x)$.

Теперь остановимся кратко на влиянии ошибок интегрирования в случае использования квадратурных формул. Отметим, что в многомерных задачах этот способ вычисления интегралов является наиболее

приемлемым, так как с возрастанием размерности пространства вычислительные трудности, связанные с заменой произвольных функций их интерполянтами и точным интегрированием произведений нескольких полиномов (например, трех), значительно возрастают.

Возникает вопрос: какая точность квадратурной формулы требуется для сохранения «теоретической» скорости сходимости МКЭ, указанной в теореме II.2? Условимся, что рассматриваются только интерполяционные квадратурные формулы:

$$\int_a^b g(x)\,dx \approx \sum_{k=1}^{n} A_k g(x_k), \tag{II.66}$$

т. е. такие, в которых числовые коэффициенты A_k получены интегрированием лагранжевых коэффициентов [52]:

$$A_k = \int_a^b \omega_k(x)\,dx = \int_a^b \frac{\omega(x)\,dx}{(x - x_k)\,\omega'(x_k)}, \tag{II.67}$$

где
$$\omega(x) = (x - x_1)(x - x_2) \ldots (x - x_n),$$
$$\omega'(x) = \frac{d\omega}{dx}.$$

Здесь предполагается, что функция $g(x)$ интерполирована по ее значениям в n произвольных точках $(x_1, x_2, \ldots, x_n)$ отрезка $[a, b]$:

$$g(x) = p_n(x) + r(x),$$

где $p_n(x) = \sum_{k=1}^{n} \frac{\omega(x)}{(x - x_k)\,\omega'(x_k)} g(x_k)$, $r(x)$ — остаток интерполирования.

Принято считать, что квадратурная формула (II.66) имеет алгебраическую степень точности $n - 1$, если она верна для всевозможных многочленов степени $n - 1$ и не верна для многочленов степени n.

Как известно, интерполяционные квадратурные формулы (II.66) точны для всех многочленов степени $n - 1$ (это непосредственно следует из рассмотрения (II.66), (II.67)). Отметим, что данное утверждение справедливо при любом расположении узлов интерполяции x_k. При специальном выборе узлов x_k, $k = 1, \ldots, n$, формула (II.66) может стать верной для всех многочленов степени не выше $2n - 1$.

Примером тому является хорошо известная квадратурная формула Гаусса [52]. Имеются специальные таблицы (см., например, [52]), где указаны для различных n значения x_k и A_k, $k = 1, 2, \ldots, n$. Заметим, что все коэффициенты A_k квадратуры Гаусса — положительные.

Если функция $g(x)$ достаточное число раз дифференцируема, то остаточный член

$$R(g) = \int_a^b g(x)\,dx - \sum_{k=1}^{n} A_k g(x_k)$$

квадратурной формулы степени точности $n-1$ имеет вид

$$R(g)=\frac{1}{n!}\int_a^b \omega(x)\frac{d^n g}{dx^n}(\xi)\,dx,$$

а степени точности $2n-1$ —

$$R(g)=\frac{1}{(2n)!}\int_a^b \omega^2(x)\frac{d^{2n} g}{dx^n}(\xi)\,dx,$$

где $\xi\in(a, b)$.

Итак, для квадратурной формулы степени точности $\alpha-1$ погрешность $R(g)$ численного интегрирования некоторой функции $g(x)$ на элементе $[x_{i-1}, x_i]$ можно оценить следующим образом:

$$|R(g)|\leqslant Ch_i^{\alpha}\int_{x_{i-1}}^{x_i}\left|\frac{d^{\alpha}g}{dx^{\alpha}}\right|dx,\quad h_i=x_i-x_{i-1}, \tag{II.68}$$

где C — положительная постоянная.

После изложенных предварительных сведений наметим пути оценки погрешности $u^N(x)-\tilde{u}^N(x)$, где $\tilde{u}^N(x)$, как и ранее, обозначает функцию, полученную минимизацией «возмущенного» функционала

$$F^*(v^N)=[v^N, v^N]_A^*-2(f, v^N)^*=\sum_{i=1}^{N}F_i^*(v^N)=$$

$$=\sum_{i=1}^{N}\left(\sum_{j=1}^{r}A_j^i\left(k(\xi_j^i)\left(\frac{dv^N}{dx}(\xi_j^i)\right)^2+q(\xi_j^i)(v^N(\xi_j^i))^2-2f(\xi_j^i)v^N(\xi_j^i)\right)\right. \tag{II.69}$$

на множестве функций $v^N\in \overset{0}{P}{}_n^h\subset \overset{0}{W}{}_2^1(0, l)$, а $u^N(x)$ доставляет минимум функционалу (II.69) на этом же множестве $\overset{0}{P}{}_n^h$.

В записи функционала (II.69) предполагается, что отрезок $[0, l]$ разбит на N элементарных отрезков $[x_{i-1}, x_i]$ и на каждом элементе выбрано r квадратурных узлов $\xi_j^i\in[x_{i-1}, x_i]$, $j=1\div r$.

Как следует из общих рассмотрений [101], для функционала (II.69) справедливо тождество, аналогичное (II.63). Если использовать обозначение

$$\eta(x)=u^N(x)-\tilde{u}^N(x),$$

$$R_i(g)=\int_{x_{i-1}}^{x_i}g(x)\,dx-\sum_{j=1}^{r}A_j^i g(\xi_j^i),\quad g(\xi_j^i)=g_j^i,$$

то данное тождество можно записать в виде

$$\sum_{i=1}^{N}\sum_{j=1}^{r}A_j^i\left(k_j^i\left(\frac{d\eta_j^i}{dx}\right)^2+q_j^i(\eta_j^i)^2\right)=2\sum_{i=1}^{N}R_i(f\eta)-$$

$$-\sum_{i=1}^{N}R_i\left(k\frac{du^N}{dx}\frac{d\eta}{dx}+qu^N\eta\right). \tag{II.70}$$

На основании тождества (II.70) легко доказывается следующее утверждение.

Теорема II.3. *Если*

$$\sum_{i=1}^{N}\sum_{j=1}^{r} A_j^i\left(k_j^i\left(\frac{dv^N}{dx}(\xi_j^i)\right)^2+q_j^i(v^N(\xi_j^i))^2\right)\geqslant\theta\|v^N\|_{2,1}^2,\quad\forall v^N\in\overset{0}{P}{}_n^h, \tag{II.71}$$

$$\theta>0-\text{const},$$

и

$$\left|\sum_{i=1}^{N}R_i\left(k\frac{du^N}{dx}\frac{dv^N}{dx}+qu^Nv^N\right)\right|+2\left|\sum_{i=1}^{N}R_i(fv^N)\right|\leqslant Ch^p\|v^N\|_{2,1}, \tag{II.72}$$

$$h=\max_i h_i,$$

то для ошибки приближенного интегрирования справедлива оценка

$$\|u^N-\tilde{u}^N\|_{2,1}\leqslant\frac{1}{\theta}Ch^p. \tag{II.73}$$

При использовании результатов теоремы II.3 и оценки (II.68) в работе [101] показано, что для выполнения условия (II.71) при положительных квадратурных коэффициентах A_j^i должно быть на каждом элементе $[x_{i-1}, x_i]$ по крайней мере n точек интегрирования ξ_j^i, $j=1\div n$, если $v^N(x)$ — полином степени n, т. е. условие (II.71) выполняется при $r=n$.

Далее, установлено, что при использовании квадратуры Гаусса с n узлами интегрирования на каждом элементе $[x_{i-1}, x_i]$ показатель p в оценке (II.72) принимает значение

$$p=(2n-1)-(n-1)+1=n+1, \tag{II.74}$$

если допустимые функции $v^N(x)$ являются полиномами n-й степени.

Таким образом, согласно (II.73)

$$\|u^N-\tilde{u}^N\|_{2,1}\leqslant C_1h^{n+1}, \tag{II.75}$$

что опять-таки не снижает теоретической скорости сходимости МКЭ, обеспечиваемой теоремой II.2.

Отметим, что оценка (II.75) (как и вид показателя p в (II.74)), полученная для частного случая функционала (II.51), следует и из общих результатов, касающихся связи между точностью квадратурной формулы и порядком ошибки $u^N(x)-\tilde{u}^N(x)$. В многомерном пространстве для функционала, зависящего от m-х старших производных допустимых функций v^N, эти общие результаты формулируются следующим образом.

Если с помощью квадратурной формулы точно вычисляется интеграл от любого полинома степени s, умноженного на m-ю производную любой допустимой функции, то справедлива оценка

$$\|u^N-\tilde{u}^N\|_{2,m}\leqslant Ch^{s+1}.$$

Итак, для сведения ошибок численного интегрирования к уровню ошибок МКЭ в общем случае достаточно выбирать квадратурные формулы степени точности 2 $(n - m)$, если допустимые функции были полиномами n-й степени. Тогда и $\| u^N - u \|_{2,m}$, и $\| u^N - \tilde{u}^N \|_{2,m}$ будут иметь одинаковый порядок h^{n-m+1}. Напоминаем, что при этом необходимо добиться выполнения условия $[v^N, v^N]_A^* \geqslant \theta \| v^N \|_{2,m}^2$ (аналог условия (II.71)), которое определяет достаточное количество точек интегрирования на элементе.

3. Погрешности, возникающие при решении на ЭВМ системы уравнений МКЭ. Как известно, подавляющее большинство вычислений на ЭВМ сопровождается ошибками округления, обусловленными как спецификой представления любого числа в ЭВМ в виде конечной дроби с ограниченным количеством разрядов машинной системы счисления, так и спецификой выполнения арифметических операций на ЭВМ.

Рассмотрим влияние этих ошибок на точность вычисленного решения системы линейных алгебраических уравнений метода конечных элементов

$$\tilde{K} z = \tilde{b}. \tag{II.76}$$

В (II.76) через $\tilde{K}$ обозначена глобальная матрица жесткости, построенная из элементарных матриц жесткости и матриц масс, а через b — вектор правых частей уравнений, построенный из элементарных векторов нагрузки с учетом краевых условий задачи.

При этом предполагается, что элементарные матрицы жесткости и масс, а также векторы нагрузки найдены численным, а не точным интегрированием.

Для конкретности будем считать, что система (II.76) получена дискретизацией методом конечных элементов краевой задачи

$$-\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) + q(x)u = f(x), \quad 0 < x < l,$$

$$k\frac{du}{dx} - \beta u\,|_{x=0} = 0, \quad u(l) = 0,$$

или, что то же, равнозначной ей вариационной задачи минимизации функционала

$$F(v) = \int_0^l \left(k\left(\frac{dv}{dx}\right)^2 + qv^2 - 2fv\right)dx + \beta v^2(0) \equiv [v, v]_A - 2(f, v) \tag{II.77}$$

на множестве функций пространства $W_2^1(0, l)$, удовлетворяющих условию

$$v(l) = 0. \tag{II.78}$$

Условия на $k(x)$, $q(x)$ и $f(x)$ такие же, как и раньше. Отрезок $[0, l]$ разбивался на N элементов.

Исследуя вопрос о влиянии ошибок округления, необходимо с самого начала отметить, что вычисленные на ЭВМ элементы матриц $\tilde{K}$

и $\bar{b}$ будут отличаться от «истинных», получаемых при точном выполнении всех арифметических операций. Даже если бы эти элементы были вычислены абсолютно точно где-то вне ЭВМ, то при вводе в ЭВМ они претерпели бы искажения за счет ошибок округления, связанных с переводом десятичного представления чисел в числа машинной арифметики.

Таким образом, приступая к решению на ЭВМ уравнений МКЭ, мы в действительности будем иметь дело не с системой (II.76), а с системой

$$\bar{K}\bar{z} = \bar{b}, \tag{II.79}$$

«возмущенной» ошибками округления входных данных.

Как известно, для решения системы линейных алгебраических уравнений можно использовать прямые и итерационные методы. Для построения эффективного быстросходящегося итерационного процесса необходимо располагать дополнительной информацией о свойствах матрицы системы, что, как правило, является нетривиальной задачей. Однако, как уже отмечалось в гл. I (параграф I.3), в ряде случаев возможно получение систем алгебраических уравнений МКЭ, эквивалентных по спектру хорошо изученным схемам метода конечных разностей. Это позволяет использовать разработанные для конечно-разностных уравнений эффективные итерационные методы и в случае МКЭ. Следует еще заметить, что итерационные методы — приближенные методы, так как дают решение в виде предела сходящейся бесконечной последовательности некоторых векторов. Полученное с их помощью приближенное решение обязательно содержит погрешность метода, существенно зависящую от критерия окончания процесса вычислений. Наряду с этим на точность найденного решения влияют и ошибки округления, неизбежно возникающие при выполнении любых арифметических операций на ЭВМ (см., например, [105]).

В данной работе мы ограничимся рассмотрением вычисления решения системы (II.79) только посредством прямых методов, которые реализуются за конечное число арифметических операций и при точном выполнении всех вычислений обеспечивают получение точного решения системы. Отметим, кроме того, что прямые методы позволяют с меньшими затратами реализовать решение системы со многими правыми частями, а это важно в практике инженерных и исследовательских расчетов.

В последнее время начали использоваться методы решения алгебраических систем, которые представляют собой комбинацию прямых и итерационных методов. К ним относятся метод преобусловливания [121, 136, 139, 146, 152] и методы вычисления решения, реализуемые на последовательности сеток [31, 110, 140—142].

Рассмотрим теперь суммарный эффект влияния ошибок округления, допускаемых при вычислении на ЭВМ решения системы (II.79) каким-нибудь прямым методом. Как показал обратный анализ [14, 105], это влияние равносильно некоторому возмущению исходных данных решаемой системы. Иными словами, реально вычисленное на ЭВМ решение системы (II.79) является точным решением некоторой

возмущенной системы

$$(\bar{K} + dK)\,\bar{z}_в = \bar{b} + db$$

с относительно малыми возмущениями dK и db, называемыми эквивалентными возмущениями. Относительные эквивалентные возмущения $\|dK\|/\|\bar{K}\|$ и $\|db\|/\|\bar{b}\|$ зависят от порядка решаемой системы n, вычислительного алгоритма и значения ε — единичной ошибки округления конкретной ЭВМ в относительных единицах. Сравнение по точности многих прямых методов показало, что «при правильной реализации эквивалентное возмущение оказывается соизмеримым по величине с ошибками округления входных данных» [14]. Такую устойчивость вычислений решения удается организовать по ряду схем метода Гаусса, метода квадратных корней, метода отражений и др. (Оценки величин соответствующих эквивалентных возмущений можно найти в работах [14, 105].)

Таким образом, суммируя все погрешности машинной реализации вычисления решения системы МКЭ (II.76), можно записать, что полученное решение $\bar{z}_в \equiv \tilde{z}$ в действительности точно удовлетворяет некоторой близкой системе

$$(\tilde{K} + d\tilde{K})\,\tilde{z} = \tilde{b} + d\tilde{b}.$$

Относительную погрешность решения $\tilde{z}$ можно оценить по формуле

$$\frac{\|z - \tilde{z}\|}{\|z\|} \leqslant \frac{\|\tilde{K}\|\|\tilde{K}^{-1}\|}{1 - \|\tilde{K}^{-1}\|\|d\tilde{K}\|}\left(\frac{\|d\tilde{K}\|}{\|\tilde{K}\|} + \frac{\|d\tilde{b}\|}{\|\tilde{b}\|}\right), \qquad (II.80)$$

справедливой в любой из согласованных норм при выполнении условия

$$\|\tilde{K}^{-1}\|\|d\tilde{K}\| < 1.$$

Оценка (II.80) является мажорантной, но неулучшаемой (достижимой) на классе всех невырожденных матриц.

Как видно из (II.80), точность вычисленного значения $\tilde{z}$ существенно зависит от числа обусловленности матрицы системы, а именно от числа $H = \|\tilde{K}\|\|\tilde{K}^{-1}\|$. (Заметим, что иногда для оценки точности вычисленного решения используют следующее сугубо приближенное практическое правило [150]: если система с числом обусловленности $H = O(10^r)$ решается на ЭВМ, выполняющей операции с p десятичными цифрами, то вычисленное решение может иметь только $p - r$ верных значащих цифр.)

Таким образом, для оценки точности вычисленного решения системы МКЭ необходимо знать число обусловленности матрицы этой системы или хотя бы порядок его величины.

Значение $H = \|\tilde{K}\|\|\tilde{K}^{-1}\|$ зависит от выбранной нормы (векторной и согласованной с ней матричной), но порядок величины H при этом, как правило, изменяется мало. В дальнейшем мы будем иметь в виду

специальную матричную норму, которая для симметричной положительно определенной матрицы A равна максимальному собственному числу этой матрицы $\|A\| = \lambda_{\max}(A)$. Очевидно, что в этом случае

$$\|A^{-1}\| = 1/\lambda_{\min}(A).$$

Теперь наша цель — оценить $H = \|\tilde{K}\|\|\tilde{K}^{-1}\| = \dfrac{\lambda_{\max}(\tilde{K})}{\lambda_{\min}(\tilde{K})}$ для матрицы системы уравнений МКЭ задачи (II.77), (II.78). При этом мы будем следовать работе [101].

Напомним, что матрица $\tilde{K}$ образуется в результате дискретизации методом конечных элементов выражения $\int_0^l \left(k\left(\frac{dv}{dx}\right)^2 + qv^2\right)dx + \beta v^2(0) \equiv [v, v]_A$ функционала (II.77) на некотором классе допустимых функций $v^N \in P_n^h$.

В силу предположений о коэффициентах $k(x)$ и $q(x)$ справедливо неравенство

$$[v, v]_A \leqslant C\int_0^l \left(\left(\frac{dv}{dx}\right)^2 + v^2\right)dx + \beta v^2(0), \qquad \text{(II.81)}$$

где

$$C = \max\left(k_1 = \max_{0\leqslant x\leqslant l} k(x),\ q_1 = \max_{0\leqslant x\leqslant l} q(x)\right).$$

Используя для дискретизации обеих частей неравенства (II.81) одну и ту же конечно-элементную процедуру, с учетом граничного условия (II.78) получаем

$$\omega^T\tilde{K}\omega \equiv \sum_{i=1}^{N}\omega_i^T\tilde{K}_i\omega_i \leqslant C\sum_{i=1}^{N}\omega_i^T K_i\omega_i, \qquad \text{(II.82)}$$

где $K_i = K_i^1 + K_i^0$, причем элементарная матрица жесткости K_i^1 связана с интегралом $\int_{x_{i-1}}^{x_i}\left(\frac{dv^N}{dx}\right)^2 dx$, а элементарная матрица масс $K_i^0 \equiv$

$\equiv M_i$ — с $\int_{x_{i-1}}^{x_i}(v^N)^2\,dx$.

Так как для любой положительно определенной квадратичной формы y^TAy справедлива оценка $y^TAy \leqslant \lambda_{\max}(A)\,y^Ty$, то из (II.82) следует

$$\omega^T\tilde{K}\omega \leqslant C\sum_{i=1}^{N}\omega_i^T K_i\omega_i \leqslant C\sum_{i=1}^{N}\lambda_{\max}(K_i)\,\omega_i^T\omega_i \leqslant C\Lambda\sum_{i=1}^{N}\omega_i^T\omega_i \leqslant 2C\Lambda\omega^T\omega, \qquad \text{(II.83)}$$

где

$$\Lambda = \max_i(\lambda_{\max}(K_i)).$$

В цепочке неравенств (II.83) было учтено, что любая компонента вектора ω (т. е. любой фиксированный параметр: v_i, v_i и т. п.) встречается не более чем в двух соседних элементах.

Из соотношений (II.83) следует, что

$$\lambda_{\max}(\tilde{K}) \leqslant 2C\Lambda.$$

Так как наименьшее собственное число положительно определенной матрицы A, согласно принципу Рэлея, определяется соотношением

$$\lambda_{\min}(A) = \min \frac{y^T A y}{y^T y},$$

то $\lambda_{\min}(\tilde{K})$ можно оценить следующим образом:

$$\lambda_{\min}(\tilde{K}) = \min \frac{\omega^T \tilde{K} \omega}{\omega^T \omega} = \min\left(\frac{\omega^T \tilde{K} \omega}{\omega^T M \omega} \frac{\omega^T M \omega}{\omega^T \omega}\right) \geqslant$$

$$\geqslant \min \frac{\omega^T \tilde{K} \omega}{\omega^T M \omega} \min \frac{\omega^T M \omega}{\omega^T \omega} = \lambda_1 \lambda_{\min}(M), \qquad \text{(II.84)}$$

где через $M \equiv K^0$ обозначена матрица масс, связанная с $\int_0^l (v^N)^2\, dx$ с учетом краевого условия (II.78), а через λ_1 — минимальное собственное число задачи

$$\tilde{K} y = \lambda M y. \qquad \text{(II.85)}$$

Как будет показано в гл. IV, задача (II.85) является дискретным аналогом соответствующей дифференциальной задачи на собственные значения, в данном случае задачи

$$Lu \equiv -\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) + q(x)u = \lambda u, \quad 0 < x < l,$$

$$k\frac{du}{dx} - \beta u\,|_{x=0} = 0, \; u(l) = 0,$$

решаемой методом конечных элементов. Там же будет показано, что минимальное собственное число λ_1 дискретной задачи всегда не меньше минимального собственного числа $\lambda_1(L)$ исходной задачи:

$$\lambda_1 \geqslant \lambda_1(L).$$

(Данное соотношение, вообще, справедливо для приближенных собственных чисел, найденных процессом Ритца.)

Отметим, что $\lambda_1(L)$, являясь нижней границей для λ_1, никак не зависит от $h_i = x_i - x_{i-1}$.

Далее, чтобы получить оценку $\lambda_{\min}(\tilde{K})$, необходимо оценить $\lambda_{\min}(M)$ (см. (II.84)). Рассуждая так же, как раньше, получим

$$\omega^T M \omega = \sum_{i=1}^{N} \omega_i^T M_i \omega_i \geqslant \sum_{i=1}^{N} \lambda_{\min}(M_i)\, \omega_i^T \omega_i \geqslant \theta \sum_{i=1}^{N} \omega_i^T \omega_i \geqslant \theta \omega^T \omega,$$

где θ — наименьшее из минимальных собственных чисел элементарных матриц масс.

Таким образом,

$$\lambda_{\min}(M) = \min \frac{\omega^T M \omega}{\omega^T \omega} \geqslant \theta$$

и согласно (II.84)

$$\lambda_{\min}(\tilde{K}) \geqslant \lambda_1(L)\,\theta.$$

Суммируя результаты, можно привести следующую оценку числа обусловленности матрицы $\tilde{K}$ системы уравнений МКЭ для любого типа элементов:

$$H = \frac{\lambda_{\max}(\tilde{K})}{\lambda_{\min}(\tilde{K})} \leqslant \frac{2C\Lambda}{\lambda_1(L)\,\theta},$$

где C — некоторая положительная константа, зависящая от коэффициентов $k(x)$ и $q(x)$, Λ — максимальное собственное число из всех собственных чисел элементарных матриц $K_i = K_i^1 + K_i^0$, $i = 1 \div N$, связанных с $\int\limits_{x_{i-1}}^{x_i} \left(\left(\frac{dv^N}{dx}\right)^2 + (v^N)^2\right) dx$ (с учетом граничных условий), $\lambda_1(L)$ — минимальное (первое) собственное число дифференциальной задачи, не зависящее от дискретизации, θ — минимальное собственное число из всех собственных чисел N элементарных матриц масс $K_i^0 \equiv M_i$, связанных с $\int\limits_{x_{i-1}}^{x_i} (v^N)^2\,dx$ (то же с учетом граничных условий).

Получим теперь оценки числа обусловленности для некоторых конкретных типов элементов. Рассмотрим вначале случай использования элемента «l 1—2». Как показано в параграфе II.2,

$$K_i^1 = \frac{1}{h_i}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad M_i = \frac{h_i}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},$$

следовательно,

$$K_i = \begin{bmatrix} \frac{1}{h_i} + \frac{h_i}{3} & \frac{h_i}{6} - \frac{1}{h_i} \\ \frac{h_i}{6} - \frac{1}{h_i} & \frac{1}{h_i} + \frac{h_i}{3} \end{bmatrix}, \quad i = 2 \div (N-1),$$

$$K_1 = \begin{bmatrix} \frac{1}{h_1} + \frac{h_1}{3} + \beta & \frac{h_1}{6} - \frac{1}{h_1} \\ \frac{h_1}{6} - \frac{1}{h_1} & \frac{1}{h_1} + \frac{h_1}{3} \end{bmatrix}, \quad K_N = \frac{1}{h_N} + \frac{h_N}{3}.$$

Непосредственными вычислениями нетрудно убедиться, что здесь при достаточно малом значении $h = \min\limits_i h_i$

$$\Lambda \leqslant \frac{2}{h} + \frac{1}{6} + \beta, \quad \theta = \frac{h}{6}.$$

т. е. $H = O\left(\frac{1}{h^2}\right)$.

Если для дискретизации использовать элемент вида «*l*2—3», то

$$K_i = \begin{bmatrix} \frac{7}{3h_i} + \frac{2h_i}{15} + \delta_{i1}\beta & -\frac{3}{3h_i} + \frac{h_i}{15} & \frac{1}{3h_i} - \frac{h_i}{30} \\ & \frac{16}{3h_i} + \frac{8h_i}{15} & -\frac{8}{3h_i} + \frac{h_i}{15} \\ \text{Симметрично} & & \frac{7}{3h_i} + \frac{2h_i}{15} \end{bmatrix}, \quad \text{(II.86)}$$

$$i = 1 \div (N-1),$$

где δ_{i1} — символ Кронекера ($\delta_{11} = 1$; $\delta_{i1} = 0$, $i \neq 1$),

$$K_N = \begin{bmatrix} \frac{7}{3h_N} + \frac{2h_N}{15} & -\frac{8}{3h_N} + \frac{h_N}{15} \\ -\frac{8}{3h_N} + \frac{h_N}{15} & \frac{16}{3h_N} + \frac{8h_N}{15} \end{bmatrix}, \quad \text{(II.87)}$$

а

$$M_i = \frac{h_i}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix}.$$

Поскольку абсолютные значения собственных чисел матрицы не превосходят любую из норм этой матрицы, то

$$\lambda_{\max}(K_i) \leqslant \| K_i \|_\infty,$$

где норма $\| \|_\infty$ для произвольной матрицы $A = (a_{ij})$, $i, j = 1, \ldots$ ..., n, определяется соотношением $\| A \|_\infty = \max\limits_i \sum\limits_{j=1}^{n} |a_{ij}|$.

Следовательно, согласно (II.86), (II.87) при достаточно малых h_i имеем

$$\lambda_{\max}(K_i) \leqslant \frac{32}{3h_i} - \frac{2h_i}{5},$$

или

$$\Lambda \leqslant \frac{32}{3h}, \quad h = \min_i h_i.$$

Так как матрица M_i — положительно определенная при любом значении $h_i \neq 0$, то ее минимальное собственное число

$$\lambda_{\min}(M_i) = \frac{\alpha h_i}{30},$$

где $\alpha > 0$ — минимальное собственное число матрицы

$$\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix}.$$

Следовательно, $\theta = \frac{\alpha h}{30}$ и $H = O\left(\frac{1}{h^2}\right)$.

Рассмотрим еще случай использования элемента Эрмита «$l3$—2». С помощью аналогичных рассуждений, учитывая вид матриц $K_i = K_i^1 + M_i = \Pi_i^{-1} S^{-T} (R_i^1 + R_i^0) S^{-1} \Pi_i^{-1}$ (см. п. 3 параграфа II.2), нетрудно убедиться, что в данном случае при достаточно малом значении h

$$\Lambda = O\left(\frac{12}{5h}\right), \quad \theta = O\left(\frac{\alpha_1 h^3}{420}\right), \quad \alpha_1 > 0,$$

т. е.

$$H = O\left(\frac{1}{h^4}\right).$$

Однако необходимо отметить, что и при использовании элемента Эрмита «$l3$—2» решаемую систему линейных алгебраических уравнений легко преобразовать таким образом, чтобы число обусловленности матрицы преобразованной системы по-прежнему имело порядок $1/h^2$.

Для этого достаточно «масштабировать» матрицу системы (II.76) следующим способом:

$$\tilde{\tilde{K}} = \Pi \tilde{K} \Pi,$$

где диагональная матрица $\Pi = \operatorname{diag}\{\Pi_1, \Pi_2, \ldots, \Pi_N\}$ (см. (II.47)), и решать систему

$$\tilde{\tilde{K}} y = \Pi \tilde{b}, \quad y = \Pi^{-1} z. \tag{II.88}$$

Такое «масштабирование» равносильно тому, что на каждом элементе вектор фиксируемых параметров имеет вид (ср. с (II.47), (II.48))

$$\tilde{\omega}_i = \Pi_i^{-1} \omega_i = \begin{bmatrix} 1 & & & \\ & h_i & & \\ & & 1 & \\ & & & h_i \end{bmatrix} \begin{bmatrix} v_{i-1} \\ v'_{i-1} \\ v_i \\ v'_i \end{bmatrix} = \begin{bmatrix} v_{i-1} \\ h_i v_{i-1} \\ v_i \\ h_i v_i \end{bmatrix}, \quad i = 1, \ldots, N,$$

а элементарные матрицы жесткости и масс принимают вид

$$\tilde{\tilde{K}}_i^1 = S^{-T} R_i^1 S^{-1}, \quad \tilde{\tilde{M}}_i = S^{-T} R_i^0 S^{-1}.$$

В случае постоянных коэффициентов (ср. с п. 3 параграфа II.2):

$$\tilde{\tilde{K}}_i^1 = \frac{k_0}{30 h_i} \begin{bmatrix} 36 & 3 & -36 & 3 \\ 3 & 4 & -3 & -1 \\ -36 & -3 & 36 & -3 \\ 3 & -1 & -3 & 4 \end{bmatrix},$$

$$\tilde{\tilde{M}}_i = \frac{q_0 h_i}{420}\begin{bmatrix} 156 & 22 & 54 & -13 \\ 22 & 4 & 13 & -3 \\ 54 & 13 & 156 & -22 \\ -13 & -3 & -22 & 4 \end{bmatrix}.$$

Исследуя, как прежде, квадратичную форму $\omega^T \tilde{\tilde{K}} \omega = \sum_{i=1}^{N} \omega_i^T \times$ $\times \tilde{\tilde{K}}_i \omega_i$, $\tilde{\tilde{K}}_i = \tilde{\tilde{K}}_i^1 + \tilde{\tilde{M}}_i$, нетрудно убедиться, что в данном случае $\Lambda = O\ (1/h)$, $\theta = O\ (h)$, следовательно, $H = O\left(\frac{1}{h^2}\right)$. Таким образом, чтобы избежать значительного искажения вычисленного решения ошибками округления, целесообразно строить и решать (в случае использования элементов Эрмита) систему (II.88), а не (II.76)

Аналогичное «масштабирование» можно и нужно применять и при использовании других видов элементов Эрмита, когда в качестве фиксируемых параметров используются значения искомой функции и ее производных. Это позволит «удержать» число обусловленности матрицы соответствующей системы уравнений на уровне величины $O\left(\frac{1}{h^2}\right)$.

4. Практическая оценка точности вычисленного на ЭВМ решения. Как было показано в предыдущих пунктах, теоретически установлена скорость сходимости приближенного решения МКЭ к точному решению краевой или вариационной задачи (теорема II.2). При соблюдении определенной осторожности эту скорость не нарушает и замена точного интегрирования квадратурными формулами. Однако все эти результаты носят асимптотический характер, и их трудно использовать для оценки точности конкретного вычисленного на ЭВМ решения, тем более что полученные результаты искажены и ошибками округления. Отметим, что, зная по теоретическим оценкам порядок числа обусловленности матрицы решаемой системы и учитывая длину машинного слова (точность вычислений), можно по упомянутому ранее практическому правилу приближенно оценить число верных значащих цифр полученного решения z, т. е. верных по отношению к точному решению z дискретной системы МКЭ линейных алгебраических уравнений (II.76). Ориентируясь на эти верные цифры решения z, можно применить следующее практическое правило для оценки погрешности в точке x_i вычисленного приближенного значения $u^N(x_i) = u_i^N$ по отношению к точному решению дифференциальной задачи $u(x_i)$. (Напомним, что компонентами вектора решения z являются фиксированные в узлах значения искомого приближенного решения $u^N(x)$ и некоторых его производных.)

Пусть согласно теореме II.2 приближенное решение сходится в норме пространства W_2^1 к решению дифференциальной задачи со скоростью порядка n. Тогда в силу теоремы вложения Соболева [97] бу-

дет справедлива и оценка

$$\max_{0\leqslant x\leqslant l} |u^N(x) - u(x)| = O(h^n).$$

Предположим, далее, что существует следующее представление:

$$u(x_i) = \tilde{u}_i^N + \alpha(x_i) h^n + O(h^{n+1}), \quad h = \max_i (x_i - x_{i-1}). \quad \text{(II.89)}$$

Рассматривая приближенные решения в общей точке x_i для двух сеток с максимальными размерами элементов h и $\bar{h}$, т. е. выражения (II.89) и

$$u(x_i) = \tilde{\tilde{u}}_i^N + \alpha(x_i) \bar{h}^n + O(\bar{h}^{n+1}), \quad \text{(II.90)}$$

можно с точностью до величин порядка h^{n+1} найти

$$\alpha(x_i) \approx \frac{\tilde{\tilde{u}}_i^N - \tilde{u}_i^N}{h^n - \bar{h}^n}.$$

Это позволяет получить следующие оценки погрешности вычисленного решения в точке x_i:

$$\Delta\tilde{u}_i^N = |u(x_i) - \tilde{u}_i^N| \approx |\alpha(x_i)| h^n,$$

$$\delta\tilde{u}_i^N = \frac{\Delta\tilde{u}_i^N}{|\tilde{u}_i^N|}, \quad \tilde{u}_i^N \neq 0.$$

Используя вычисленные решения систем МКЭ уравнений для двух разбиений области на элементы, можно уточнить получаемые для узлов значения приближенного решения по формуле

$$\tilde{\tilde{u}}_i^N = \sigma\tilde{u}_i^N + (1 - \sigma)\tilde{\tilde{u}}_i^N, \quad \text{(II.91)}$$

где

$$\sigma = \frac{\bar{h}^n}{\bar{h}^n - h^n}. \quad \text{(II.92)}$$

Действительно, учитывая в (II.91) представления (II.89), (II.90), имеем

$$\tilde{\tilde{u}}_i^N = u(x_i) - \alpha(x_i)[\sigma h^n + (1 - \sigma)\bar{h}^n] - \sigma O(h^{n+1}) - (1 - \sigma) O(\bar{h}^{n+1}) = u(x_i) + O(h^{n+1} + \bar{h}^{n+1}),$$

так как согласно (II.92) $\sigma h^n + (1 - \sigma)\bar{h}^n = 0$.

Таким образом,

$$|u(x_i) - \tilde{\tilde{u}}_i^N| = O(h^{n+1} + \bar{h}^{n+1}).$$

Если доказана сходимость решения дискретной задачи к решению дифференциальной, но скорость сходимости неизвестна, то, предпо-

ложив выполнение соотношения

$$u(x_i) = \tilde{u}_i^N + \alpha(x_i) h^\nu + O(h^{\nu+1})$$

и использовав три варианта сетки с максимальными размерами элементов h, $\bar{h}$, $\bar{\bar{h}}$, можно установить экспериментально значение γ из следующего соотношения (справедливого с точностью до $O(h^{\nu+1})$):

$$\frac{|\tilde{u}_i^N - \bar{\tilde{u}}_i^N|}{|\bar{\tilde{u}}_i^N - \bar{\bar{\tilde{u}}}_i^N|} \approx \frac{|\bar{h}^\nu - h^\nu|}{|\bar{\bar{h}}^\nu - \bar{h}^\nu|},$$

где $\tilde{u}_i^N$, $\bar{\tilde{u}}_i^N$, $\bar{\bar{\tilde{u}}}_i^N$ — значения вычисленного решения в узле x_i, общем для всех сеток.

При наличии численных решений, отвечающих трем сеткам, можно вычислить решение в узле x_i с повышенным порядком точности по формуле

$$(u_i^N)^* = \sigma_1 \bar{\tilde{u}}_i^N + \sigma_2 \bar{\bar{\tilde{u}}}_i^N + (1 - \sigma_1 - \sigma_2) \tilde{u}_i^N,$$

где

$$\sigma_1 = \frac{h^{n+1}(\bar{h}^n - h^n) - h^n(\bar{\bar{h}}^{n+1} - h^{n+1})}{(\bar{h}^n - h^n)(\bar{\bar{h}}^{n+1} - \bar{h}^{n+1})},$$

$$\sigma_2 = \frac{h^n(\bar{h}^{n+1} - h^{n+1}) - h^{n+1}(\bar{h}^n - h^n)}{(\bar{h}^n - h^n)(\bar{\bar{h}}^{n+1} - \bar{h}^{n+1})}.$$

Отметим, что при этом имеем $|u(x_i) - (u_i^N)^*| = O(h^{n+2})$.

5. **Численные результаты.** Обратимся теперь к более подробному рассмотрению численных результатов, полученных МКЭ при решении модельного примера пунктов 1—3 параграфа II.2. Остановимся в основном на случае использования кусочно-линейных полиномов, т. е. применения элементов вида «l 1—2».

Вначале рассмотрим поведение приближенного решения при измельчении сетки, когда длина элементарного отрезка принимает значения h, равные 0,25; 0,125; 0,0625. При этом все остальные особенности построения сеточной системы уравнений (в частности, число узлов

Таблица 5

x_i	$u_0(x_i)$	u_i		
		$h = 0{,}25$	$h = 0{,}125$	$h = 0{,}0625$
0	—1	—1,00655220	—1,00161180	—1,00039950
0,25	—0,71597458	—0,72331423	—0,71778243	—0,71642228
0,5	—0,35127873	—0,35871457	—0,35310427	—0,35172841
0,75	0,11700002	0,11148608	0,11565202	0,11667361

Примечание. При $h = 0{,}25$ значение $t = 1$ мин 45 с; при $h = 0{,}125$ значение $t = 3$ мин 20 с; при $h = 0{,}0625$ значение $t = 5$ мин 40 с.

Таблица 6

x_i	$u_0(x_i)$	u_i		
		$h = 0{,}25$	$h = 0{,}125$	$h = 0{,}0625$
0	—1	—1,01675170	—1,00415710	—1,00103560
0,25	—0,71597458	—0,73115442	—0,71973503	—0,71691008
0,5	—0,35127873	—0,36564878	—0,35479609	—0,35214912
0,75	0,11700002	0,10649864	0,11445482	0,11637685

Примечание. При $h = 0{,}25$ значение $t = 45$ с, $H^{(4)} = 6{,}12$; при $h = 0{,}125$ значение $t =$ $= 1$ мин 30 с, $H^{(8)} = 24{,}5$; при $h = 0{,}0625$ значение $t = 2$ мин 55 с, $H^{(16)} = 103{,}8$.

в квадратурных формулах Гаусса и разрядность ЭВМ МИР-2) остаются неизменными: четыре квадратурных узла на каждом элементе $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, N$; $N = 4, 8, 16$; разрядность $R = 8$. Соответствующие значения полученных приближенных решений представлены в табл. 5. Обозначения такие же, как в табл. 1, и, кроме того, n_Γ — число квадратурных узлов на каждом элементе, $n_\Gamma = 4$, t — время получения приближенного решения на ЭВМ МИР-2. Приведенные результаты свидетельствуют о повышении точности приближенного решения с измельчением сетки, причем асимптотический порядок сходимости согласуется с предсказанным теоретическим — $O(h^2)$ (см. замечание в п. 1 параграфа II.3). В частности, для трех сеток ($N = 4, 8, 16$) в каждом узле выполняется соотношение

$$\frac{|u_i^4 - u_i^8|}{|u_i^8 - u_i^{16}|} \approx 4,$$

что обеспечивает поточечную сходимость h^2.

В табл. 6 приведены значения приближенных решений, полученные на тех же сетках, что и в табл. 5, но при использовании квадратурных формул Гаусса с одним узлом на каждом элементе $R = 8$. В этой же таблице даны значения числа обусловленности $H^{(N)}$ матрицы системы уравнений МКЭ. (Время t указано без вычисления $H^{(N)}$). Нетрудно убедиться, что и в данном случае ($n_\Gamma = 1$) асимптотический порядок сходимости приближенных решений остался прежним — $O(h^2)$: в частности, в узловых точках сетки выполняется соотношение

$$\frac{|u_i^4 - u_i^8|}{|u_i^8 - u_i^{16}|} \approx 4.$$

Сравнение временных затрат на получение приближенных решений при использовании различных квадратурных формул (см. данные табл. 5 и 6) показывает, что можно обеспечить требуемую точность искомого решения за меньшее время ЭВМ, если применять квадратурные формулы с минимально необходимым числом узлов, но на более мелких сетках. Однако это возможно лишь при условии достаточного объема оперативной памяти ЭВМ.

Таблица 7

x_l	$u_*(x_l)$	u_l		
		$h = 0{,}25$	$h = 0{,}125$	$h = 0{,}0625$
0	—1	—0,99990220	—1,00002260	—0,99978870
0,25	—0,71597458	—0,71587449	—0,71602041	—0,71566348
0,5	—0,35127873	—0,35118390	—0,35137074	—0,35075443
0,75	0,11700002	0,11707040	0,11689623	0,11769728

Примечание. При $h = 0{,}25$ значение $t = 2$ мин 5 с, $H = 25{,}2$; при $h = 0{,}125$ значение $t = 4$ мин 5 с, $H = 74{,}7$; при $h = 0{,}0625$ значение $H = 475{,}7$.

Обратим внимание еще и на вычисленные значения числа обусловленности $H^{(N)}$ системы уравнений МКЭ (см. табл. 6). Теоретические оценки гарантируют $H^{(N)} = O\left(\frac{1}{h^2}\right)$; полученные значения $H^{(N)}$ хорошо согласуются с этой оценкой. Действительно, $H^{(8)}/H^{(4)} \approx 4$, $H^{(16)}/H^{(8)} \approx \approx 4{,}2$.

Таким образом, все приведенные численные результаты убедительно иллюстрируют изложенные выше теоретические исследования.

Рассмотрим еще решение того же модельного примера с использованием кусочно-квадратичных полиномов, т. е. элемента вида «$l2$—3», на разных сетках. Численные результаты при разрядности 8 ЭВМ МИР-2 и при $n_\Gamma = 2$ представлены в табл. 7.

Нетрудно видеть, что приближенное решение, полученное на сетке $h = 0{,}25$ посредством элемента вида «$l2$—3», $n_\Gamma = 2$, несколько точнее, чем для линейных полиномов даже при $h = 0{,}0625$ и $n_\Gamma = 4$. При этом в случае использования квадратичных полиномов потребовалось 2 мин 5 с машинного времени, а в случае линейных — 5 мин 40 с. Однако, как свидетельствуют результаты табл. 7, дальнейшее измельчение сетки при сохранении разрядности $R = 8$ не ведет к повышению точности получаемого решения: здесь суммарные ошибки округления при построении системы МКЭ и ее решении превышают (перекрывают) погрешность метода конечных элементов. Для повышения точности получаемого решения надо все вычисления выполнять с большей разрядностью. К аналогичным выводам приходим, рассматривая и результаты решения задачи посредством элемента вида «$l3$—2» на различных сетках.

II.4. Базисные функции метода конечных элементов

Как показано в предшествующих параграфах, реализация варианта МКЭ, основанного на процессе Ритца, не требует непосредственного построения базисных функций $\{\varphi_l^N(x)\}$. Однако при других вариантах МКЭ, в частности используемых для решения несамосопряженных задач, знание базисных функций оказывается необходимым. В связи

с этим рассмотрим сейчас построение некоторых видов базисных функций и установим их связь с допустимыми функциями, используемыми в предыдущих параграфах при описании варианта метода Ритца.

1. Кусочно-линейные базисные функции. Предположим, что искомое решение некоторой краевой задачи принадлежит пространству $W_2^1(0, l)$, т. е. является непрерывной функцией на интервале $(0, l)$ и имеет суммируемые с квадратом первые производные. Пусть приближенное решение $v^N(x)$ этой задачи представлено в виде

$$v^N(x) = \sum_i c_i \varphi_i^N(x), \tag{II.93}$$

где $\{\varphi_i^N\}$ — система выбранных базисных функций, c_i — искомые числовые коэффициенты. Чтобы функция $v^N(x) \in W_2^1(0, l)$, достаточно построить систему базисных функций $\{\varphi_i^N(x)\}$ следующим образом.

Разобьем отрезок $[0, l]$ определения искомого решения на N элементарных отрезков $[x_{i-1}, x_i]$, $i = 1 \div N$, $x_0 = 0$, $x_N = l$. Определим базисную функцию $\varphi_i^N(x)$ на всем отрезке $[0, l]$ как кусочно-линейный полином, удовлетворяющий следующим условиям:

$$\varphi_i^N(x_i) = 1, \quad \varphi_i^N(x_{i-1}) = \varphi_i^N(x_{i+1}) = 0,$$

$$\varphi_i^N(x) \equiv 0, \quad x \in [x_0, x_{i-1}] \cup [x_{i+1}, l].$$

Очевидно, что $\varphi_i^N(x)$ на элементарных отрезках $[x_{i-1}, x_i]$ и $[x_i, x_{i+1}]$ представляет собой отрезки линейных интерполяционных полиномов Лагранжа, построенных по заданным значениям, т. е.

$$\varphi_i^N(x) = \begin{cases} \dfrac{x - x_{i-1}}{h_i}, & x \in [x_{i-1}, x_i], \\ \dfrac{x_{i+1} - x}{h_{i+1}}, & x \in [x_i, x_{i+1}], \end{cases} \quad h_i = x_i - x_{i-1}, \; i = 1, 2, \ldots, N-1. \tag{II.94}$$

Вне отрезка $[x_{i-1}, x_{i+1}]$ базисная функция $\varphi_i^N(x)$ тождественно равна нулю. Этот отрезок для функции $\varphi_i^N(x)$ иногда называют носителем, он состоит из двух элементов с общим узлом в точке x_i. Заметим, что $\varphi_0^N(x) = \left(1 - \dfrac{x}{h_1}\right)$, $\varphi_N^N(x) = \dfrac{x - x_{N-1}}{h_N}$.

Очевидно, что построенные базисные функции $\varphi_i^N(x)$ (рис. 19) непрерывны на $[0, l]$ и имеют разрывные первые производные, интегрируемые с квадратом. Иными словами, использование их обеспечит для приближенного решения $v^N(x) = \sum_{i=1}^{n} c_i \varphi_i^N(x)$ принадлежность пространству $W_2^1(0, l)$.

Построенная система базисных функций обладает свойством, которое можно считать аналогом свойства полноты. Эта система полна в том смысле, что любую непрерывную на $[0, l]$ кусочно-линейную

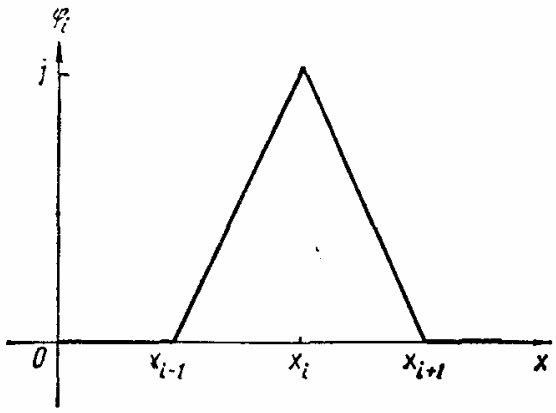

Рис. 19.

функцию $f(x)$ можно представить в виде линейной комбинации данных базисных функций:

$$I(x) = \sum_{i=0}^{N} f_i \varphi_i^N(x), \quad 0 \leqslant x \leqslant l,$$

где $f_i = f(x_i)$.

Данную систему базисных функций можно назвать квазиортогональной, так как в $L_2(0, l)$ функция $\varphi_i^N(x)$ будет ортогональна ко всем $\varphi_k^N(x)$, для которых $|k - i| > 1$.

В силу построения базисных функций (II.94) числовые коэффициенты c_i в разложении приближенного решения (II.93) совпадают со значением этого решения в узловой точке $x = x_i$, поэтому можно записать

$$v(x) \equiv v^N(x) = \sum_{i=1}^{N} v_i \varphi_i^N(x), \quad x \in [0, l],$$

где $v_i = v^N(x_i)$.

Если теперь вернуться к варианту МКЭ, основанному на процессе Ритца, то кусочно-линейную допустимую функцию на элементе $[x_{i-1}, x_i]$ можно представить в виде

$$v^N(x) = v_{i-1} \frac{x_i - x}{h_i} + v_i \frac{x - x_{i-1}}{h_i},$$

или

$$v^N(x) = v_{i-1} \varphi_{i-1}^N(x) + v_i \varphi_i^N(x), \quad x \in [x_{i-1}, x_i]. \qquad \text{(II.95)}$$

Действительно, на основании соглашения об определении допустимой кусочно-линейной функции $v^N = \beta_1 + \beta_2 x$ на элементарном отрезке $[x_{i-1}, x_i]$ через ее значения в узлах $v_i = v^N(x_i)$ имеем

$$v_{i-1} = \beta_1 + \beta_2 x_{i-1}, \quad v_i = \beta_1 + \beta_2 x_i,$$

откуда следует

$$\beta \equiv [\beta_1, \beta_2]^T = S_i^{-1} \omega_i,$$

где

$$\omega_i = [v_{i-1}, v_i]^T, \quad S_i^{-1} = \frac{1}{h_i} \begin{bmatrix} x_i & -x_{i-1} \\ -1 & 1 \end{bmatrix}.$$

Поэтому можно записать

$$v^N(x) = \beta_1 + \beta_2 x = [\beta_1, \beta_2] \begin{bmatrix} 1 \\ x \end{bmatrix} =$$

$$= \omega_i^T S_i^{-T} \begin{bmatrix} 1 \\ x \end{bmatrix} = \omega_i^T \begin{bmatrix} \dfrac{x_{i-1} - x}{h_i} \\ \dfrac{x - x_{i-1}}{h_i} \end{bmatrix} = [v_{i-1}, v_i] \begin{bmatrix} \varphi_{i-1}^N(x) \\ \varphi_i^N(x) \end{bmatrix}, \quad x \in [x_{i-1}, x_i],$$

что подтверждает представление (II.95).

2. Кусочно-квадратичные базисные функции. Согласно введенному в предыдущем пункте разбиению области $[0, l]$ на N элементарных отрезков $[x_{i-1}, x_i]$ можно построить систему базисных функций, каждая из которых будет кусочно-квадратичной и непрерывной на отрезке $[0, l]$. Для этого вводится на элементе $[x_{i-1}, x_i]$ дополнительный узел $x_{i-1/2} = \frac{x_i + x_{i-1}}{2}$ и определяются два вида базисных функций. Функции $\varphi_i^N(x)$, $i = 0, 1, 2, \ldots, N$, соответствующие граничным узлам x_i элементов, задаются соотношениями

$$\varphi_i^N(x_j) = \delta_{ij},$$

где $\delta_{ii} = 1$, $\delta_{ij} = 0$ при $i \neq j$; $i = 0, 1, 2, \ldots, N$, $j = 0, 1/2, 1, 3/2 \ldots, \ldots, N$, а функции $\varphi_{i-1/2}^N(x)$, $i = 1, 2, \ldots, N$, соответствующие внутренним узлам $x_{i-1/2}$,— соотношениями

$$\varphi_{i-1/2}^N(x_{i-1/2}) = 1, \quad \varphi_{i-1/2}^N(x_j) = 0 \text{ при } j \neq (i - 1/2),$$
$$i = 1, 2, \ldots, N, \quad j = 0, 1/2, 1, 1\,1/2, \ldots, N.$$

Для построения указанных базисных функций достаточно использовать отрезки квадратичных интерполяционных полиномов Лагранжа, положив

$$\varphi_i^N(x) = \begin{cases} \dfrac{2(x - x_{i-1})(x - x_{i-1/2})}{h_i^2}, & x \in [x_{i-1}, x_i], \\ \dfrac{2(x - x_{i+1/2})(x - x_{i+1})}{h_{i+1}^2}, & x \in [x_i, x_{i+1}], \end{cases}$$

$$\varphi_i^N(x) \equiv 0, \quad x \in [0, x_{i-1}] \cup [x_{i+1}, l],$$

$$\varphi_{i-1/2}^N(x) = -\frac{4(x - x_{i-1})(x - x_i)}{h_i^2}, \quad x \in [x_{i-1}, x_i],$$

$$\varphi_{i-1/2}^N(x) \equiv 0, \quad \text{если} \quad x \bar{\in} [x_{i-1}, x_i].$$

Отметим, что носитель функции $\varphi_i^N(x)$ (рис. 20, *а*) состоит из двух соседних элементарных отрезков с общим узлом в точке x_i, а носитель функции $\varphi_{i-1/2}^N(x)$ — единственный элемент, содержащий узел

$$x_{i-1/2} = \frac{x_i + x_{i-1}}{2} \text{ (рис. 20, } б).$$

Функция вида

$$v^N(x) = \sum_{i=0}^{N} v_i \varphi_i^N(x) + \sum_{i=1}^{N} v_{i-1/2} \varphi_{i-1/2}^N(x)$$

непрерывна и принадлежит пространству $W_2^1(0, l)$.

Замечания относительно свойств полноты и квазиортогональности остаются справедливыми и для этой системы базисных функций. При этом функция $\varphi_i^N(x)$ ортогональна ко всем функциям системы, кроме $\varphi_{i\pm1}^N(x)$ и $\varphi_{i\pm1/2}^N(x)$, а функция $\varphi_{i-1/2}^N(x)$ — ко всем, кроме $\varphi_i^N(x)$ и $\varphi_{i-1}^N(x)$.

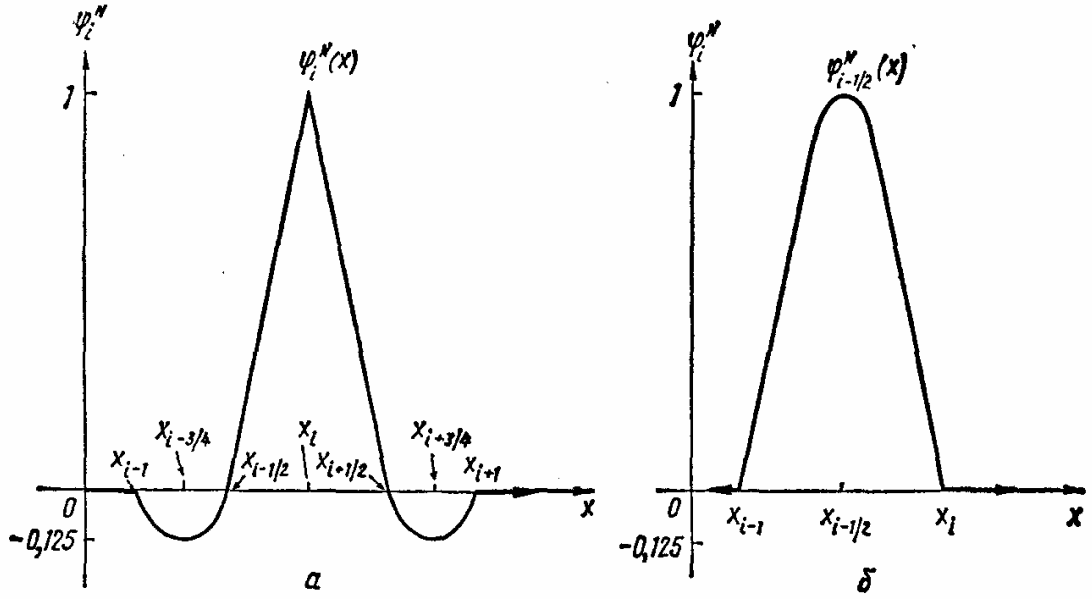

Рис. 20.

В варианте МКЭ, основанном на процессе Ритца, допустимая функция $v^N(x)$ на элементарном отрезке $[x_{i-1}, x_i]$ представляется через базисные функции в виде

$$v^N(x) = v_{i-1}\varphi_{i-1}^N(x) + v_{i-1/3}\varphi_{i-1/3}^N(x) + v_i\varphi_i^N(x).$$

3. **Кусочно-кубические функции.** Остановимся кратко на базисных функциях, отвечающих использованию элементов видов «*l*3—4» и «*l*3—2».

В случае элемента вида («*l*3—4») на каждом элементарном отрезке $[x_{i-1}, x_i]$ фиксируется четыре узла:

$$x_{i-1},\ x_{i-2/3} = \frac{2x_{i-1}+x_i}{3},\quad x_{i-1/3} = \frac{x_{i-1}+2x_i}{3},\quad x_i,$$

и каждому из них соответствует одна базисная функция, являющаяся кусочным полиномом Лагранжа третьей степени. Эти базисные функции определяются следующими соотношениями:

$$\varphi_i^N(x) = \begin{cases} \dfrac{9}{2}\dfrac{(x-x_{i-1})(x-x_{i-2/3})(x-x_{i-1/3})}{h_i^3}, & x\in[x_{i-1}, x_i], \\ -\dfrac{9}{2}\dfrac{(x-x_{i+1/3})(x-x_{i+2/3})(x-x_{i+1})}{h_{i+1}^3}, & x\in[x_i, x_{i+1}], \end{cases} \tag{II.96}$$

$$\varphi_i^N(x) = 0,\quad x\in[0, x_{i-1}]\cup[x_{i+1}, l],$$

$$\varphi_{i-1/3}^N(x) = -\frac{27}{2}\frac{(x-x_{i-1})(x-x_{i-2/3})(x-x_i)}{h_i^3},\quad x\in[x_{i-1}, x_i],$$

$$\varphi_{i-1/3}^N(x) \equiv 0,\quad \text{если}\quad x\bar{\in}[x_{i-1}, x_i], \tag{II.97}$$

$$\varphi_{i-2/3}^N(x) = \frac{27}{2}\frac{(x-x_{i-1})(x-x_{i-1/3})(x-x_i)}{h_i^3},\quad x\in[x_{i-1}, x_i],$$

$$\varphi_{i-2/3}^N(x) \equiv 0,\quad \text{если}\quad x\bar{\in}[x_{i-1}, x_i]. \tag{II.98}$$

Характерные особенности поведения базисных функций (II.96) — (II.98) видны из рис. 21—23.

В случае элемента вида «*l*3—2» с каждым узлом x_i связаны по две базисные функции: $\varphi_i^N(x)$ — соответствующая фиксированному в этом узле параметру $v_i = v^N(x_i)$ и $\psi_i^N(x)$ — соответствующая параметру $v_i' = \frac{dv^N}{dx}(x_i)$.

Каждая из базисных функций $\varphi_i^N(x)$ и $\psi_i^N(x)$ представляет собой кусочный полином Эрмита третьей степени, построенный по следующим данным:

$$\varphi_i^N(x_j) = \delta_{ij}, \quad \frac{d\varphi_i^N}{dx}(x_j) = 0, \quad i, j = 0, 1, 2, \ldots, N, \qquad \text{(II.99)}$$

$$\psi_i^N(x_j) = 0, \quad \frac{d\psi_i^N}{dx}(x_j) = \delta_{ij}, \quad i, j = 0 \div N.$$

Используя интерполяционную формулу Эрмита [7], функции $\varphi_i^N(x)$ и $\psi_i^N(x)$ (рис. 24, *б*) можно определить на $[0, l]$ следующим образом:

$$\varphi_i^N(x) = \begin{cases} \left(1 + \frac{2(x_i - x)}{h_i^2}\right)\frac{(x - x_{i-1})^2}{h_i^2}, & x \in [x_{i-1}, x_i], \\ \left(1 - \frac{2(x_i - x)}{h_{i+1}}\right)\frac{(x - x_{i+1})^2}{h_{i+1}^2}, & x \in [x_i, x_{i+1}] \end{cases},$$

$$\varphi_i^N(x) \equiv 0, \quad \text{если} \quad x \in [0, x_{i-1}] \cup [x_{i+1}, l];$$

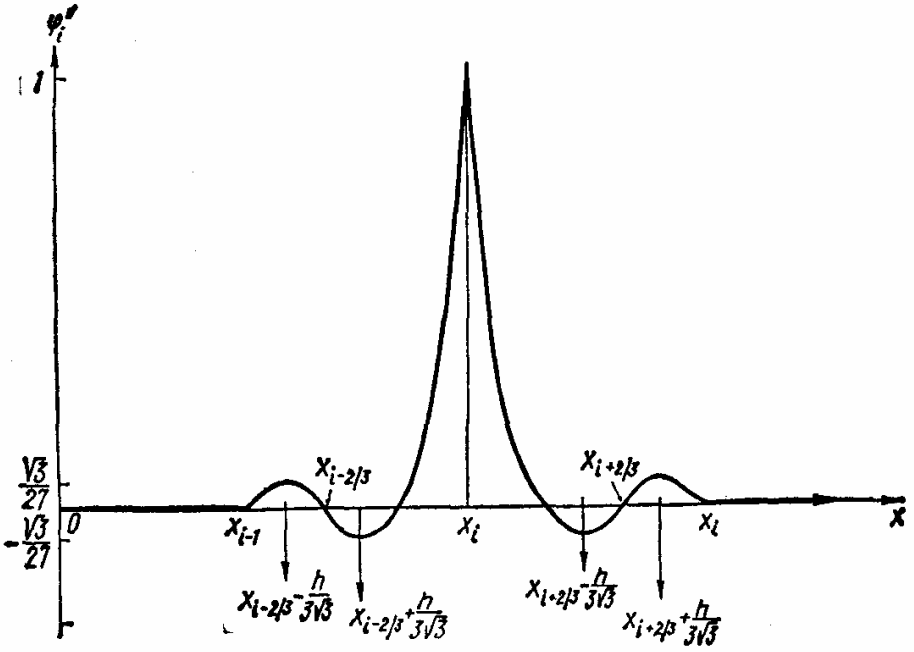

Рис. 21.

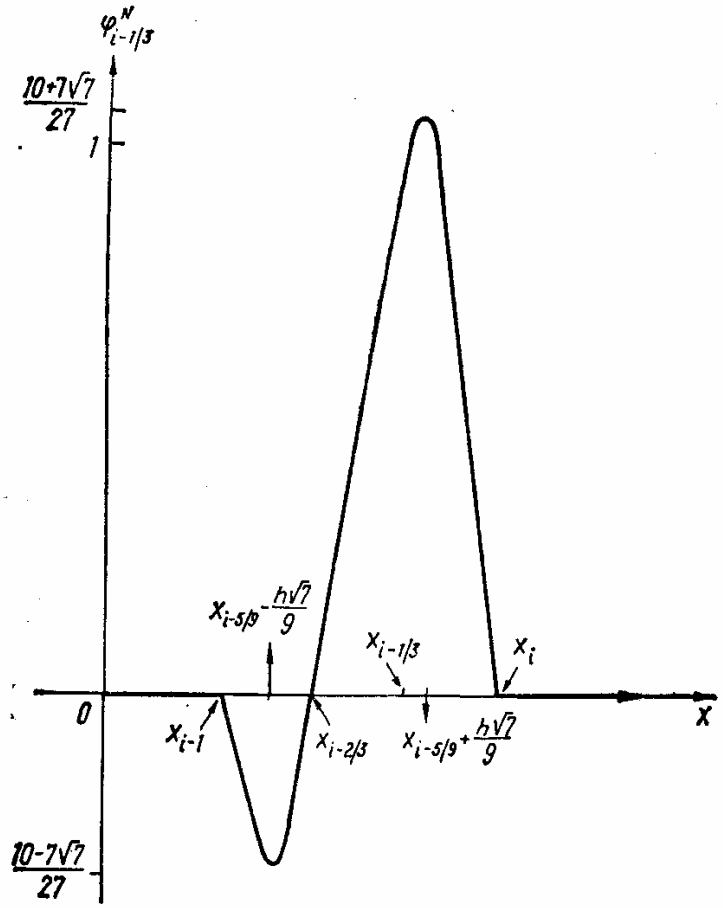
$\varphi^N_{i-1/3}$
$\frac{10+7\sqrt{7}}{27}$
1
$x_{i-5/9}-\frac{h\sqrt{7}}{9}$
$x_{i-1/3}$
x_i
0
x_{i-1}
$x_{i-2/3}$
$x_{i-5/9}+\frac{h\sqrt{7}}{9}$
x
$\frac{10-7\sqrt{7}}{27}$

Рис. 22.

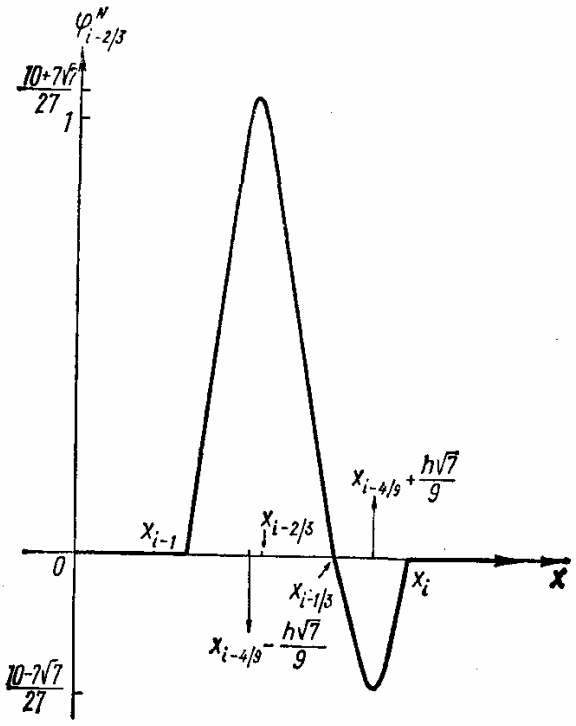
$\varphi^N_{i-2/3}$
$\frac{10+7\sqrt{7}}{27}$
1
$x_{i-4/9}+\frac{h\sqrt{7}}{9}$
x_{i-1}
$x_{i-2/3}$
0
$x_{i-1/3}$
x_i
x
$x_{i-4/9}-\frac{h\sqrt{7}}{9}$
$\frac{10-7\sqrt{7}}{27}$

Рис. 23.

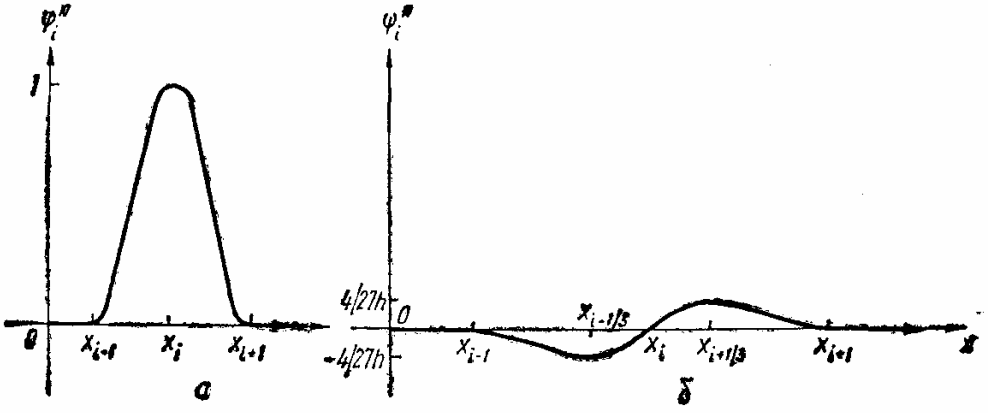

Рис. 24.

$$\psi_i^N(x) = \begin{cases} \frac{1}{h_i^2}(x - x_i)(x - x_{i-1})^2, & x \in [x_{i-1}, x_i], \\ \frac{1}{h_{i+1}^2}(x - x_i)(x - x_{i+1})^2, & x \in [x_i, x_{i+1}], \end{cases}$$

$$\psi_i^N(x) \equiv 0, \quad \text{если} \quad x \in [0, x_{i-1}] \cup [x_{i+1}, l].$$

Каждая из построенных базисных функций (рис. 24) непрерывна и непрерывно дифференцируема на всем отрезке $[0, l]$. Такой же гладкостью будет обладать и приближенное решение, задаваемое выражением

$$v^N(x) = \sum_{i=0}^{N} [v_i \varphi_i^N(x) + v_i' \psi_i^N(x)].$$

Допустимая функция, являющаяся в варианте Ритца кусочно-кубическим полиномом Эрмита, в данном базисе на элементе $[x_{i-1}, x_i]$ записывается в виде

$$v^N(x) = v_{i-1}\varphi_{i-1}^N(x) + v_{i-1}'\psi_{i-1}^N(x) + v_i\varphi_i^N(x) + v_i'\psi_i^N(x), \quad x \in [x_{i-1}, x_i].$$

При таком определении функция $v^N(x)$ непрерывна и непрерывно дифференцируема на $[0, l]$.

Аналогично описанному строятся базисные функции МКЭ и при других свойствах искомого приближенного решения.

II.5. Дискретизация дифференциальных задач посредством варианта метода Галеркина

1. Понятие обобщенного решения. В параграфе II.1 при рассмотрении постановок задач для обыкновенных дифференциальных уравнений было сформулировано понятие обобщенного решения краевой задачи. Оно не связано с вариационной формулировкой задачи, не требует положительной определенности или симметрии оператора краевой задачи, а основывается на использовании некоторого интегрального соотношения.

Напомним и уточним здесь формулировку обобщенного решения. Пусть краевая задача представлена в операторном виде

$$Au = f. \qquad \text{(II.100)}$$

Различные формулировки понятия обобщенного решения этой задачи формально получаются посредством умножения в смысле скалярного произведения пространства L_2 обеих частей уравнения (II.100) на тестовую функцию $v(x)$ из некоторого тестового пространства V:

$$(Au, v) = (f, v). \qquad \text{(II.101)}$$

Под обобщенным решением понимается функция $u(x)$, при которой равенство (II.101) справедливо для каждой функции $v(x)$ из V. В зависимости от выбора тестового пространства V обобщенное решение $u(x)$ будет удовлетворять уравнению (II.100) в том или ином смысле, принадлежать тому или иному функциональному пространству.

Например, если V — пространство бесконечно дифференцируемых функций, равных нулю в приграничной зоне, то, применяя в (II.101) формальное интегрирование по частям, можно все производные перенести с $u(x)$ на $v(x)$ и получить соотношение

$$(u, A^*v) = (f, v), \quad \forall v \in V, \qquad \text{(II.102)}$$

где A^* — формально сопряженный с A оператор.

В этом случае обобщенное решение $u(x)$ удовлетворяет уравнению (II.100) только в смысле тождества (II.102) и поэтому достаточно, чтобы $u(x)$ принадлежало только L_2.

Если $V = W_2^s(0, l)$, то s производных можно перенести с $u(x)$ на $v(x)$, снижая тем самым требования к гладкости искомого решения, т. е. предполагая, что $u(x) \in W^{2m-s}$, где $2m$ — порядок дифференциального уравнения краевой задачи. Наиболее важным является случай, когда $s = m$. При этом искомое решение $u(x)$ и тестовые функции $v(x)$ принадлежат одному пространству $W_2^m(0, l)$.

В определении обобщенного решения важную роль играют краевые условия. Если $s = 0$, то на $u(x)$ налагается полное множество краевых условий задачи.

При $s > 0$, когда $u(x) \in W_2^{2m-s}$, для определения искомого решения понадобятся лишь $2m - s$ производных от $u(x)$, так что будут иметь смысл лишь краевые условия порядка, меньшего $2m - s$. Число условий, налагаемых на тестовые функции $v(x)$, при этом возрастает: они определяются производными до s-го порядка [101].

2. Построение системы уравнений МКЭ при явном использовании базисных функций. Для приближенного решения уравнения (II.100), оператор которого не является положительно определенным, используется вариант МКЭ, основанный на методе Бубнова — Галеркина. Покажем это на следующем примере.

Найти решение уравнения

$$-\frac{d}{dx}\left(k\frac{du}{dx}\right) + p\frac{du}{dx} + qu = f(x), \quad x \in (0, l), \qquad \text{(II.103)}$$

удовлетворяющее условиям

$$u(0) = u(l) = 0. \qquad \text{(II.104)}$$

Будем предполагать, что функции $k(x) \geqslant k_0 > 0$, $\frac{dk}{dx}$, $p(x)$, q ограничены на $[0, l]$, $f \in L_2$. Назовем обобщенным решением из $W_2^1(0, l)$ задачи (II.103), (II.104) функцию $u(x)$ из $\overset{0}{W}{}_2^1(0, l)$, удовлетворяющую тождеству

$$\int_0^l \left(k \frac{du}{dx} \frac{dv}{dx} + p \frac{du}{dx} v + quv\right) dx = \int_0^l fv dx$$

при любой функции $v(x) \in W_2^1(0, l)$.

Таким образом, в данном случае в качестве тестового выступает пространство $V = \overset{0}{W}{}_2^1(0, l)$, а искомое обобщенное решение тоже принадлежит пространству $\overset{0}{W}{}_2^1(0, l)$.

Разрешимость данной задачи (и более сложных в многомерных пространствах) рассматривается в работе [55].

Для построения по МКЭ приближенного обобщенного решения $u^N(x)$ вводим в рассмотрение два конечномерных подпространства: подпространство P_n^h из пространства, которому принадлежит обобщенное решение (в нашем случае $P_n^h \subset \overset{0}{W}{}_2^1$), и подпространство V_n^h из тестового пространства V (здесь тоже $V_n^h \in \overset{0}{W}{}_2^1$). В качестве базисов в обоих подпространствах используются описанные в параграфе II.4 базисные функции МКЭ. Тогда приближенным решением $u^N(x)$ называется такой элемент из P_n^h, что соотношение

$$\int_0^l \left(k \frac{du^N}{dx} \frac{dv^N}{dx} + p \frac{du^N}{dx} v^N + qu^N v^N\right) dx = \int_0^l fv^N dx \qquad \text{(II.105)}$$

выполняется при любой функции $v^N(x) \in V_n^h \subset \overset{0}{W}{}_2^1$. Отметим, что базисные функции в подпространствах P_n^h и V_n^h могут быть в общем случае разного типа (полиномы разной степени), но размерность подпространств должна быть одинаковой.

Ограничимся рассмотрением случая, когда базисные функции $\varphi_i^N(x)$ обоих подпространств одинаковы и размерность подпространств равна r.

Приближенное решение $u^N(x)$ будем искать в виде

$$u^N(x) = \sum_{i=0}^{r-1} c_i^N \varphi_i^N(x),$$

где N — количество элементов $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, N$, $x_0 = 0$, $x_N = l$, r — общее количество узловых параметров и отвечающих им базисных функций $\varphi_i^N(x)$.

Поскольку соотношение (II.105) должно выполняться при любой функции $v^N(x) \in V_n^h$, достаточно, чтобы оно выполнялось для всех базисных функций $\varphi_j^N(x)$ подпространства V_n^h ($j = 0 \div (r-1)$):

$$\int_0^l \left(k \sum_{i=0}^{r-1} c_i^N \frac{d\varphi_i^N}{dx} \frac{d\varphi_j^N}{dx} + p \sum_{i=0}^{r-1} c_i^N \frac{d\varphi_i^N}{dx} \varphi_j^N + q \sum_{i=0}^{r-1} c_i^N \varphi_i^N \varphi_j^N \right) dx =$$

$$= \int_0^l f \varphi_j^N dx, \quad j = 0, 1, \ldots, r-1. \qquad \text{(II.106)}$$

Остановимся более подробно лишь на простейшем случае кусочно-линейных базисных функций (II.94).

Так как искомое обобщенное решение принадлежит пространству $\overset{0}{W}{}_2^1(0, l)$ и $V = \overset{0}{W}{}_2^1(0, l)$, то базис подпространств P_1^h и V_1^h в данном случае образуют функции $\varphi_i^N(x)$, для которых $i = 1, 2, \ldots, N-1$. Поэтому можно положить

$$u^N(x) = \sum_{i=1}^{N-1} u_i^N \varphi_i^N(x), \quad u_i^N = u^N(x_i),$$

и соотношения (II.106) определяют систему линейных алгебраических уравнений относительно неизвестных параметров u_i^N ($i = 1, 2, \ldots, N-1$):

$$\sum_{i=1}^{N-1} u_i^N \int_0^l \left(k \frac{d\varphi_i^N}{dx} \frac{d\varphi_j^N}{dx} + p \frac{d\varphi_i^N}{dx} \varphi_j^N + q \varphi_i^N \varphi_j^N \right) dx = \int_0^l f \varphi_j^N dx, \qquad \text{(II.107)}$$

$$j = 1, 2, \ldots, N-1.$$

Вследствие квазиортогональности системы базисных функций коэффициенты этой системы

$$a_{ij} = \int_0^l \left(k \frac{d\varphi_i^N}{dx} \frac{d\varphi_j^N}{dx} + p \frac{d\varphi_i^N}{dx} \varphi_j^N + q \varphi_i^N \varphi_j^N \right) dx, \quad i, j = 1 \div N-1,$$

для которых $|i - j| > 1$, будут равны нулю, т. е. матрица системы (II.107) будет ленточной трехдиагональной, несимметричной. Решение такой системы легко находится по алгоритмам прямых методов, учитывающим ленточный вид матрицы.

Отметим, что в случае симметричности и положительной определенности оператора краевой задачи варианты МКЭ, основанные на процессах Ритца и Бубнова — Галеркина, приводят к совершенно одинаковым системам линейных алгебраических уравнений.

В данном пункте мы не будем более подробно останавливаться на рассмотрении варианта МКЭ, основанного на процессе Бубнова — Галеркина, в частности на вопросах его обоснования, рассмотрении скорости сходимости и других особенностей (эти вопросы освещены, например, в монографии [101]).

II.6. Дискретизация обыкновенных дифференциальных уравнений высших порядков

В данной главе до настоящего параграфа речь шла в основном о применении МКЭ к решению краевых задач для обыкновенных дифференциальных уравнений второго порядка. Однако ни процедура дискретизации дифференциальной задачи, ни вопросы обоснования МКЭ не имеют каких-то принципиальных отличий в случае дифференциальных уравнений более высоких порядков.

Проиллюстрируем это на случае следующей краевой задачи для уравнения четвертого порядка (см. п. 3 параграфа II.1):

$$\frac{d^2}{dx^2}\left(k(x)\frac{d^2u}{dx^2}\right)-\frac{d}{dx}\left(p(x)\frac{du}{dx}\right)+q(x)u=f(x),\quad 0<x<l, \tag{II.108}$$

$$u(0)=\frac{du}{dx}(0)=0,\quad u(l)=\frac{du}{dx}(l)=0, \tag{II.109}$$

где $k(x)\geqslant k_0>0$, $p(x)\geqslant 0$, $q(x)\geqslant 0$, т. е. оператор задачи симметричный и положительно определенный.

Обобщенным решением данной задачи является функция $u(x)$, минимизирующая функционал

$$F(v)=\int_0^l\left[k\left(\frac{d^2v}{dx^2}\right)^2+p\left(\frac{dv}{dx}\right)^2+qv^2-2fv\right]dx$$

в пространстве $\overset{0}{W}{}_2^2(0, l)$. Поэтому при построении приближенного решения $u^N(x)$ вариантом МКЭ, основанным на модифицированном процессе Ритца, в качестве допустимых функций $v^N(x)$ здесь следует использовать функции из конечномерного подпространства $P_n^h\subset\overset{0}{W}{}_2^2(0, l)$. Иными словами, допустимые функции $v^N(x)$ должны быть непрерывны, непрерывно дифференцируемы на всем отрезке $[0, l]$ и иметь суммируемые с квадратом вторые производные. Таким образом, в данном случае приемлемым оказывается элемент «l3—2», а элементы вида «l2 —3» или «l3—4» использовать нельзя. Возможно применение в качестве допустимых также кусочно-полиномиальных функций высших степеней, но при условии обеспечения их непрерывной дифференцируемости на всей области определения $[0, l]$. Заметим, что использование полиномов слишком высоких степеней сопряжено со значительными вычислительными трудностями и поэтому вряд ли целесообразно.

Рассмотрим без особых подробностей дискретизацию задачи (II.108), (II.109) посредством элемента вида «l3—2», предполагая, что область $[0, l]$ разбита на N отрезков $[x_{i-1}, x_i]$, $i=1, 2, \ldots, N$. В п. 3 параграфа II.2 уже использовался этот элемент при решении краевых задач для дифференциальных уравнений второго порядка. Полученное там выражение для матрицы S^{-1}, устанавливающей на «каноническом отрезке» связь между числовыми коэффициентами допустимой функции $v^N(x)$ и фиксируемыми в узлах x_{i-1}, x_i параметрами v_{i-1}, v'_{i-1}, v_i, v'_i, справедливо и в случае уравнения четвертого порядка.

Аналогичны и элементарные матрицы жесткости $K_i^1 = \Pi_i^{-1} S^{-T} R_i^1 S^{-1} \Pi_i^{-1}$, связанные с интегралом $\int_{x_{i-1}}^{x_i} p \left(\frac{dv^N}{dx}\right)^2 dx$, и элементарные матрицы масс $M_i = \Pi_i^{-1} S^{-T} R_i^0 S^{-1} \Pi_i^{-1}$, отвечающие членам $\int_{x_{i-1}}^{x_i} q\,(v^N)^2\,dx$ (см. п.3 параграфа II.2). Здесь добавляется только новая элементарная матрица — матрица изгиба K_i^2, возникающая из интеграла $\int_{x_{i-1}}^{x_i} k \left(\frac{d^2 v^N}{dx^2}\right)^2 dx$. Напомним, что на «канонический отрезок» [0, 1] любой элементарный отрезок $[x_{i-1}, x_i]$ отображается посредством преобразования $x = x_{i-1} + h_i \xi$, $h_i = x_i - x_{i-1}$; следовательно, $\int_{x_{i-1}}^{x_i} k \left(\frac{d^2 v^N}{dx^2}\right)^2 dx$ переходит в интеграл $\int_0^1 \tilde{k} \left(\frac{d^2 r}{d\xi^2}\right)^2 d\xi$, где $k = \frac{1}{h_i^3} k\,(x_{i-1} + h_i \xi)$, $r(\xi) = v^N (x_{i-1} + h_i \xi) = \alpha_1 + \alpha_2 \xi + \alpha_3 \xi^2 + \alpha_4 \xi^3$.

В результате уже знакомых вычислений находим, что интеграл $\int_0^1 \tilde{k} \left(\frac{d^2 r}{d\xi^2}\right)^2 d\xi$ порождает элементарную матрицу изгиба

$$K_i^2 = \Pi_i^{-1} S^{-T} R_i^2 S^{-1} \Pi_i^{-1},$$

где

$$R_i^2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 4\int_0^1 \tilde{k}\,d\xi & 12\int_0^1 \tilde{k}\xi\,d\xi \\ 0 & 0 & 12\int_0^1 \tilde{k}\xi\,d\xi & 36\int_0^1 \tilde{k}\xi^2\,d\xi \end{bmatrix},$$

матрицы Π_i^{-1}, S^{-1} такие же, как в п.3 параграфа II.2. В случае $k(x) = \text{const} = k_0$ матрица изгиба имеет вид

$$K_i^2 = \frac{k_0}{h_i^3} \begin{bmatrix} 12 & 6h_i & -12 & 6h_i \\ 6h_i & 4h_i^2 & -6h_i & 2h_i^2 \\ -12 & -6h_i & 12 & -6h_i \\ 6h_i & 2h_i^2 & -6h_i & 4h_i^2 \end{bmatrix}.$$

Вид и построение вектора нагрузки i-го элемента ничем не отличаются от рассматриваемых ранее. Аналогично строится и результирующая система уравнений МКЭ.

Сходимость МКЭ в применении к уравнениям четвертого (и более высокого) порядка может быть исследована так же, как в параграфе II.3.

Однако мы приведем здесь результаты, следующие из общей теории МКЭ [101].

Пусть методом конечных элементов решается краевая задача для обыкновенного дифференциального уравнения $2m$-го порядка

$$Au = f,$$

оператор A которой положительно определен в $H = L_2(0, l)$ и действует по формуле

$$Au = \sum_{j=0}^{m} (-1)^j \frac{d^j}{dx^j}\left(a_j(x)\frac{d^j u}{dx^j}\right).$$

Обобщенным точным решением данной задачи является функция $u(x) \in H_A$, доставляющая минимум функционалу

$$F(v) = [v, v]_A - 2(f, v), \quad \forall v \in H_A$$

(отметим, что H_A состоит из функций, принадлежащих $W_2^m(0, l)$).

Приближенным решением МКЭ (вариант процесса Ритца) является функция u^N, доставляющая минимум функционалу $F(v)$ на конечномерном подпространстве $P_n^h \subset H_A$.

Приведем теперь результаты, касающиеся оценки погрешности $u - u^N$. Как уже упоминалось, для погрешности $u - u^N$ справедливы следующие соотношения:

$$[u - u^N, u - u^N]_A = \min_{v^N \in P_n^h} [u - v^N, u - v^N]_A,$$

$$[u - u^N, v^N]_A = 0, \quad \forall v^N \in P_n^h. \tag{II.110}$$

Соотношение (II.110) свидетельствует, что по отношению к энергетическому скалярному произведению $[\cdot, \cdot]_A$ погрешность $u - u^N$ ортогональна подпространству P_n^h, т. е. приближенное решение u^N есть проекция $u(x)$ на P_n^h.

В оценке погрешности $u - u^N$ важную роль играет теорема об аппроксимации функций из пространства W_2^k конечно-элементными подпространствами P_n^h (см. теорему 3.3 из [101]). Применительно к функциям одной переменной $u(x)$, $x \in [a, b]$, эту теорему можно сформулировать следующим образом.

Теорема II.4. *Пусть P_n^h — подпространство кусочных полиномов степени n, имеющих на $[a, b]$ непрерывные производные до $(q-1)$-го порядка, а в производных q-го порядка допускаются лишь разрывы первого рода ($P_n^h \subset W_2^q(a, b)$). Тогда для любой функции $u(x) \in W_2^{n+1}(a, b)$ и для любого целого числа s, $0 \leqslant s \leqslant q$, справедлива оценка*

$$\|u - u_I^N\|_{2,s} \leqslant C_s h^{n+1-s} \left\| \frac{d^{n+1}u}{dx^{n+1}} \right\|_{L_2}, \quad q \leqslant n, \tag{II.111}$$

где $u_I^N(x)$ интерполянт $u(x)$ из P_n^h.

На основе теоремы аппроксимации и с использованием приема Нитше в [101] дана оценка погрешности $u - u^N$ метода конечных элементов в общем случае решения многомерной эллиптической краевой задачи для дифференциального уравнения порядка $2m$.

В рассматриваемом нами случае одномерной краевой задачи соответствующие результаты (с учетом неравенства $\| u - u^N \|_A^2 = \min\limits_{v^N \in P_n^h} \| u - v^N \|_A^2 \leqslant \| u - u_I^N \|_A^2$) формулируются следующим образом.

Теорема II.5. *Пусть $P_n^h \subset W_2^q(a, b)$ — конечномерное подпространство кусочных полиномов степени n, $m \leqslant q \leqslant n$.*

Если искомое обобщенное решение $u(x) \in W_2^{n+1}(a, b)$, то для погрешности $u - u^N$ приближенного решения u^N, полученного методом конечных элементов, справедлива оценка

$$\| u - u^N \|_{2,s} \leqslant C h^{n+1-s} \| u \|_{2,n+1}, \quad 0 \leqslant s \leqslant m. \tag{II.112}$$

Отметим, что в случае многомерной краевой задачи для дифференциального уравнения $2m$-го порядка оценка (II.112) получена в предположении выполнения условия

$$C_1 \| v \|_{2,m}^2 \leqslant \| v \|_A^2 \equiv [v, v]_A \leqslant C_2 \| v \|_{2,m}^2, \quad \forall v \in H_A.$$

Применение теоремы II.5 к оценке погрешности МКЭ в случае краевой задачи для обыкновенного дифференциального уравнения $2m$-го порядка (в частности, четвертого) при использовании подпространств $P_n^h \subset W_2^m(0, l)$, $n + 1 > m$, позволяет определить скорость сходимости (в норме пространства $W_2^m(0, l)$) приближенного решения $u^N(x)$ к точному обобщенному решению $u(x) \in W_2^{n+1}(0, l)$ как величину $O(h^{n+1-m})$, т. е. при $m = 2$ скорость сходимости имеет порядок $n - 1$.

Чтобы не снижалась указанная скорость сходимости при замене в МКЭ точного интегрирования численным, достаточно (см. п. 2 параграфа II.3) использовать квадратурные формулы степени точности $2(n - m)$ при $n > m$.

Если это будут, например, квадратуры Гаусса, то в данном случае достаточно на каждом элементе использовать $n + 1 - m$ узлов интегрирования.

Анализ чисел обусловленности матриц систем уравнений МКЭ, аналогичный подробно описанному в п.3 параграфа II.3, показывает, что в случае уравнения $2m$-го порядка при любых элементах справедлива оценка $H = O\left(\frac{1}{\lambda_1 h^{2m}}\right)$, если используется соответствующее масштабирование матрицы. Иными словами, всегда можно достаточно просто построить систему уравнений МКЭ (см. примеры в п.3 параграфа II.3), матрица которой будет иметь число обусловленности не боль-

Таблица 8

x_i	$u_0(x_i)$, $u_0'(x_i)$	u_i, u_i'	
		$h = 0{,}25$	$h = 0{,}125$
0	$\frac{1}{1}$	$\frac{1{,}0007702}{1{,}0006588}$	$\frac{1{,}0007326}{1{,}0057392}$
0,25	$\frac{1{,}3642770}{2{,}0062897}$	$\frac{1{,}3652687}{2{,}0071896}$	$\frac{1{,}3668287}{2{,}0152401}$
0,5	$\frac{2{,}0609016}{3{,}7096228}$	$\frac{2{,}0621807}{3{,}7106479}$	$\frac{2{,}0663754}{3{,}7248079}$
0,75	$\frac{3{,}3078125}{6{,}4833125}$	$\frac{3{,}3093595}{6{,}4839164}$	$\frac{3{,}3178645}{6{,}5028619}$

Примечание. Над чертой $u_0(x_i)$ и u_i, под чертой $u_0'(x_i)$ и u_i' соответственно.

шее, чем $C/\lambda_1 h^{2m}$, где λ_1 — наименьшее собственное число непрерывной задачи, $h = \min_i h_i$, C — положительная константа.

В заключение приведем пример решения методом конечных элементов краевой задачи

$$\frac{d^2}{dx^2}\left(e^{-x}\frac{d^2u}{dx^2}\right) + 5u = 2 + 5e^x(1 + x^2), \quad 0 < x < 1,$$

$$\frac{d^2u}{dx^2} - 2\frac{du}{dx}\Big|_{x=0} = 1,$$

$$\frac{d^3u}{dx^3} - 2\frac{du}{dx} + 4u(x)|_{x=0} = 9,$$

$$u(1) = 2e, \quad \frac{du}{dx}\Big|_{x=1} = 4e.$$

Эта задача эквивалентна задаче об отыскании функции, минимизирующей функционал

$$F(v) = \int_0^1 \left[e^{-x}\left(\frac{d^2v}{dx^2}\right)^2 + 5v^2 - 2fv\right]dx +$$

$$+ 4v^2(0) + 2\left(\frac{dv}{dx}(0)\right)^2 - 16v(0) + 2\frac{dv}{dx}(0)$$

в классе функций из пространства $W_2^2(0, 1)$, удовлетворяющих условию

$$v(1) = 2e, \quad \frac{dv}{dx}(1) = 4e.$$

(Точное искомое решение задачи $u_0(x) = e^x(1 + x^2)$.)

Для дискретизации задачи использовались кубические полиномы

Эрмита. Все интегралы вычислялись по квадратурным формулам Гаусса с двумя квадратурными узлами на каждом элементе $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, N$. Расчет выполнялся на равномерных сетках при $N = 4,8$ (т. е. при h, равном 0,25 и 0,125). Системы МКЭ решались методом квадратных корней. Вычисления велись на ЭВМ МИР-2 при разрядности $R = 8$. Полученные результаты представлены в табл. 8, где в каждом узле для вычисленного решения первым указывается значение искомой функции, а вторым — значение ее производной в этой же точке.

Как видно из табл. 8, и в данном примере приближенное решение, вычисленное при $h = 0,25$, $R = 8$, обладает достаточно хорошей точностью. Однако при $h = 0,125$ этой разрядности оказалось недостаточно для получения решения с большей точностью: ошибки округления перекрыли ошибки МКЭ.

Глава III

МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ В НЕСТАЦИОНАРНЫХ ЗАДАЧАХ

В данной главе на конкретных примерах дается понятие о применении МКЭ для решения задач, связанных с уравнениями параболического типа Проводится сравнение результатов, полученных при различных подходах использования МКЭ в решении нестационарных задач.

III.1. Решение методом конечных элементов начально-краевых задач для линейных параболических уравнений второго порядка

Особенности применения МКЭ для данного класса задач рассмотрим на следующем примере. В области Q_T найти функцию $u(x, t)$, удовлетворяющую уравнению

$$Lu \equiv \frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(k(x, t)\frac{\partial u}{\partial x}\right) + q(x, t)u = f(x, t), \qquad \text{(III.1)}$$

$$0 < x < l,\ 0 < t < T,$$

и начально-краевым условиям

$$u(x, 0) = u_0(x), \quad x \in [0, l], \qquad \text{(III.2)}$$

$$u(0, t) = u(l, t) = 0, \quad t \in [0, T]. \qquad \text{(III.3)}$$

Предполагается, что

$$0 < c_1 \leqslant k(x, t) < c_2, \qquad \text{(III.4а)}$$

$$0 \leqslant q(x, t) < c_3, \quad c_i = \text{const} \qquad \text{(III.4б)}$$

и все известные функции удовлетворяют некоторым условиям гладкости.

1. **Постановка задачи.** При классической постановке задачи требуется, чтобы решение $u(x, t)$ было непрерывным в Q_T, имело непрерывные производные $\frac{\partial u}{\partial t}$, $\frac{\partial u}{\partial x}$, $\frac{\partial^2 u}{\partial x^2}$ в Q_T и удовлетворяло во всех точках уравнению (III.1) и на границе условиям (III.2), (III.3).

Предположение о достаточной гладкости известных функций в (III.1) — (III.3) обеспечивает существование единственного решения рассматриваемой задачи.

Если исходные данные не являются гладкими функциями, то задача (III.1) — (III.3) в общем случае не имеет классического решения. Тогда можно расширить постановку задачи и заменить ее отысканием некоторых обобщенных решений, принадлежащих различным функциональным пространствам [54].

Наиболее широким классом обобщенных решений задачи (III.1) — (III.2) с недифференцируемой разрывной ограниченной функцией $k(x, t)$ и весьма слабыми условиями на остальные известные функции является класс $\overset{0}{W}{}_2^{1,0}(Q_T)$, элементы которого не обладают никакой гладкостью по t.

В данной главе мы ограничимся понятием обобщенного решения из класса $W_2^1(Q_T)$. Под обобщенным решением задачи (III.1) — (III.3) из пространства $W_2^1(Q_T)$ понимается функция $u(x, t)$ из пространства $W_2^1(Q_T)$ (или, точнее, из $W_{2,0}^1(Q_T)$), которая удовлетворяет условию (III.2) и интегральному тождеству

$$L(u, \eta) \equiv \int_{Q_T} \left(\frac{\partial u}{\partial t} \eta + k(x, t) \frac{\partial u}{\partial x} \frac{\partial \eta}{\partial x} + q(x, t) u\eta \right) dxdt = \int_{Q_T} f\eta dxdt \tag{III.5}$$

при произвольной функции $\eta(x, t)$ из $\overset{0}{W}{}_2^{1,0}(Q_T)$.

Заметим, что формально интегральное тождество (III.5) получено из уравнения (III.1) путем скалярного умножения $(\cdot, \cdot)_{Q_T}$ его обеих частей на функцию $\eta(x, t) \in \overset{0}{W}{}_2^{1,0}(Q_T)$ и последующим интегрированием по частям. Очевидно, что классическое решение задачи (III.1) — (III.3) является и обобщенным решением из класса $W_{2,0}^1(Q_T)$. Существование и единственность обобщенного решения $u(x, t) \in W_{2,0}^1(Q_T)$ доказаны в [54] при выполнении условия (III.4) и следующих ограничениях:

$$\int_0^T \operatorname{vrai} \max_{x \in (0,l)} \left| \frac{\partial k}{\partial t} \right| dt < \infty, \tag{III.6}$$

$$\left(\int_0^T \left(\int_0^l |q(x, t)| dx \right)^2 dt \right)^{1/2} < \infty, \quad f(x, t) \in L_2(Q_T), \quad u_0(x) \in \overset{0}{W}{}_2^1(0, l).$$

В ходе доказательства используется метод Бубнова — Галеркина, который в данном случае реализуется следующим образом.

В пространстве $\overset{0}{W}{}_2^1(0, l)$ выбирается какая-нибудь фундаментальная система $\{\psi_k(x)\}$, т. е. система функций со следующими свойствами:

все функции $\psi_k(x)$, $k = 1, 2, \ldots$, принадлежат $\overset{0}{W}{}_2^1(0, l)$;

функции $\psi_k(x)$, $k = 1, 2, \ldots, N$, линейно независимы при любом значении $N < \infty$;

линейные комбинации $\sum_{k=1}^{N} \alpha_k \psi_k(x)$ с произвольными числовыми коэффициентами α_k и N образуют множество, плотное в $\overset{0}{W}{}_2^1(0, l)$. Иными словами, любой элемент из $\overset{0}{W}{}_2^1(0, l)$ с любой степенью точности может быть аппроксимирован элементами указанного множества.

Для удобства в дальнейшем будем считать систему $\{\psi_k(x)\}$ ортонормированной в пространстве $L_2(0, l)$.

Приближения $u^N(x, t)$ к искомому обобщенному решению задачи (III.1) — (III.3) будем искать в виде

$$u^N(x, t) = \sum_{k=1}^{N} c_k^N(t)\,\psi_k(x), \tag{III.7}$$

где функции $c_k^N(t)$, $k = 1, 2, \ldots, N$, удовлетворяют системе N линейных обыкновенных дифференциальных уравнений

$$\left(\frac{\partial u^N}{\partial t}, \psi_m(x)\right) + \left(k(x, t)\frac{\partial u^N}{\partial x}, \frac{d\psi_m}{dx}\right) + (qu^N, \psi_m) = (f, \psi_m),$$

$$m = 1, 2, \ldots, N,$$

и начальному условию

$$c_m^N(0) = (u_0, \psi_m).$$

Здесь через $(\cdot, \cdot)$ обозначено скалярное произведение в пространстве $L_2(0, l)$, т. е. $(u, v) = \int_0^l uv\,dx$.

Согласно (III.7) указанную систему можно переписать в виде

$$\sum_{k=1}^{N}\left[\frac{dc_k^N}{dt}\int_0^l \psi_k\psi_m dx + c_k^N(t)\int_0^l k(x, t)\frac{d\psi_k}{dx}\frac{d\psi_m}{dx}dx + c_k^N(t)\int_0^l q\psi_k\psi_m dx\right] =$$

$$= \int_0^l f(x, t)\,\psi_m(x)\,dx, \quad m = 1, 2, \ldots, N,$$

$$c_m^N(0) = \int_0^l u_0(x)\,\psi_m(x)\,dx,$$

или, что то же, в виде

$$\sum_{k=1}^{N} a_{mk}\frac{dc_k^N}{dt} + \sum_{k=1}^{N} b_{mk}(t)\,c_k^N(t) = f_m(t),$$

$$c_m^N(0) = \alpha_{0m}, \quad m = 1, 2, \ldots, N,$$

где

$$a_{mk} = \int_0^l \psi_k(x)\,\psi_m(x)\,dx, \quad \alpha_{0m} = \int_0^l u_0(x)\,\psi_m(x)\,dx,$$

$$b_{mk}(t) = \int_0^l \left(k(x, t)\frac{d\psi_k}{dx}\frac{d\psi_m}{dx} + q(x, t)\,\psi_k\psi_m\right)dx.$$

В силу условий (III.4) — (III.6) эта система однозначно определяет абсолютно непрерывные на отрезке $[0, T]$ функции $c_k^N(t)$, $k = 1, 2, \ldots, N$, и приближенное решение $u^N(x, t)$ вида (III.7).

Доказано, что последовательность построенных таким образом решений $u^N(x, t)$ при $N \to \infty$ слабо сходится в $L_2(Q_T)$ к функции $u(x, t) \in W_2^1(Q_T)$, являющейся искомым обобщенным решением из пространства $W_{2,0}^1(Q_T)$.

Заметим, что это обобщенное решение удовлетворяет и интегральному тождеству

$$\left(\frac{\partial u}{\partial t}, \psi\right) + \left(k(x, t)\frac{\partial u}{\partial x}, \frac{\partial \psi}{\partial x}\right) + (qu, \psi) = (f, \psi) \qquad \text{(III.8)}$$

при $\forall\, \psi(x) \in \overset{0}{W}{}_2^1(0, l)$, $t > 0$,

$$(u(x, 0), \psi) = (u_0(x), \psi), \quad \forall\, \psi \in \overset{0}{W}{}_2^1(0, l). \qquad \text{(III.9)}$$

Принятые выше условия для коэффициентов и начальной функции задачи обеспечивают свойство $\frac{\partial u}{\partial t}(x, t) \in \overset{0}{W}{}_2^1(0, l)$ при каждом $t > 0$.

2. Вычисление приближенных решений. Для получения методом конечных элементов приближенного обобщенного решения задачи (III.1) — (III.3) существуют различные подходы. Один из наиболее распространенных основан на процессе Галеркина, используемом для дискретизации МКЭ исходной задачи только по пространственным переменным (называют это полудискретизацией по пространству). В результате получается система обыкновенных дифференциальных уравнений от временной переменной t, которую затем решают каким-нибудь численным методом. При осуществлении полудискретизации исходят из интегрального тождества (III.8), (III.9). Остановимся несколько подробнее на этом подходе.

Для построения полудискретного приближения к обобщенному решению задачи (III.1) — (III.3) выберем конечномерное подпространство $\overset{0}{P}{}_n^h \subset \overset{0}{W}{}_2^1(0, l)$ с базисом $\{\varphi_i^N(x)\}_1^s$, отвечающим разбиению области изменения пространственной переменной x, т. е. отрезка $[0, l]$, на N элементарных отрезков $[x_{k-1}, x_k]$ $k = 1, 2, \ldots, N$. Здесь используются те же обозначения и понятия для элементов и базисных функций, что и в гл. II.

Приближенное обобщенное решение будем искать среди допустимых функций вида

$$u^N(x, t) = \sum_{i=1}^{s} c_i^N(t)\,\varphi_i^N(t) \qquad \text{(III.10)}$$

при условии, что удовлетворяются соотношения

$$\left(\frac{\partial u^N}{\partial t}, \varphi_j^N\right) + \left(k\frac{\partial u^N}{\partial x}, \frac{\partial \varphi_j^N}{\partial x}\right) + (qu^N, \varphi_j^N) = (f, \varphi_j^N), \quad t > 0, \qquad \text{(III.11)}$$

$$(u^N(x, 0), \varphi_j^N) = (u_0, \varphi_j^N) \qquad \text{(III.12)}$$

для всех базисных функций $\varphi_j^N(x)$, $j = 1, 2, \ldots, s$, подпространства $\overset{0}{P}{}_n^h \subset \overset{0}{W}{}_2^1(0, l)$. Соотношения (III.11), (III.12) определяют систему s обыкновенных дифференциальных уравнений

$$M\frac{dc^N}{dt} + Kc^N = F(t), \qquad \text{(III.13)}$$

$$c^N(t) = [c_1^N(t), c_2^N(t), \ldots, c_s^N(t)]^T$$

с начальным условием

$$Mc^N(0) = g,$$

где M — матрица масс, такая же, как в гл. II, K — матрица жесткости (симметричная), элементы которой могут зависеть от t (если коэффициенты $k(x, t)$ и $q(x, t)$ являются функциями t), $F = F(t)$ — вектор, определяемый правой частью уравнения (III.1), g — постоянный s-мерный вектор, определяемый функцией $u_0(x)$.

Заметим, что соотношение (III.12) удобнее заменить на

$$u^N(x, 0) = u_0^h(x). \qquad \text{(III.14)}$$

Здесь $u_0^h(x)$ — интерполянт функции $u_0(x)$ из пространства $\overset{0}{P}{}_n^h$.

Тогда начальное условие для системы (III.13) будет иметь вид

$$c^N(0) = g, \qquad \text{(III.15)}$$

где компонентами вектора g являются значения функции $u_0(x)$ (и некоторых ее производных) в узлах интерполяции. Например, в случае $\overset{0}{P}{}_1^h$ имеем

$$u_0^h(x) = \sum_{i=1}^{N-1} u_0(x_i)\varphi_i(x),$$

$$g = [u_0(x_1), u_0(x_2), \ldots, u_0(x_{N-1})]^T.$$

Порядок точности приближенного решения при такой замене не изменится [101].

Для численного решения задачи (III.13), (III.15) при $t \geqslant 0$ можно использовать ту или иную устойчивую разностную схему.

Опишем, например, применение схемы Кранка — Николсона в случае постоянной матрицы K. (Для удобства опустим значок N в искомом векторе $c^N(t)$, так что при дальнейшем изложении $c^N(t) \equiv c(t)$.) Рассматриваемая схема Кранка — Николсона, представленная в симметричной форме,

$$M(c^{n+1} - c^n) + \frac{\tau}{2}K(c^{n+1} + c^n) = \frac{\tau}{2}(F^{n+1} + F^n),$$

$$c^0 = g,$$

имеет по времени второй порядок точности. Здесь

$$c^n = c(n\tau), \quad n = 0, 1, \ldots, m, \quad \tau = T/m.$$

Вычисление c^{n+1} удобно выполнять на основе системы уравнений

$$\left(M + \frac{\tau}{2} K\right) c^{n+1} = \left(M - \frac{\tau}{2} K\right) c^n + \frac{\tau}{2} (F^{n+1} + F^n),$$

$$c^0 = g, \quad n = 0, 1, \ldots, m, \qquad \text{(III.16)}$$

с положительно определенной ленточной матрицей $A = \left(M + \frac{\tau}{2} K\right)$, одинаковой на каждом временном слое $n = 0, 1, 2, \ldots, m$. Благодаря тому что в системе (III.16) меняется только правая часть $f^n = \left(M - \frac{\tau}{2} K\right) c^n + \frac{\tau}{2} (F^{n+1} + F^n)$, $n = 0, 1, \ldots, m$, решение в данном случае целесообразно находить прямым методом, однократно применив метод квадратных корней для разложения $A = S^T S$, где S — верхняя треугольная матрица, а S^T — транспонированная к ней. Затем на каждом шаге искомый вектор c^{n+1}, $n = 0, 1, \ldots, m$, вычисляется при решении двух треугольных систем

$$S^T z = f^n, \quad S c^{n+1} = z.$$

Если коэффициенты задачи (III.1) — (III.3) зависят от времени, то матрица K тоже будет зависеть от времени $K = K(t)$, следовательно, матрица линейной системы (III.16) будет иной на каждом временном слое. При этом для вычисления c^{n+1} можно использовать систему

$$\left(M + \frac{\tau}{2} K^{n+1/2}\right) c^{n+1} = \left(M - \frac{\tau}{2} K^{n+1/2}\right) c^n + \frac{\tau}{2} (F^{n+1} + F^n), \qquad \text{(III.17)}$$

где

$$K^{n+1/2} = K((n + 1/2)\tau), \quad n = 0, 1, \ldots, m.$$

Однако применение прямых методов для численного решения системы (III.17) может оказаться слишком «дорогим», так как на каждом шаге необходимо строить треугольное разложение новой матрицы. Чтобы обойти эту трудность, ряд авторов [124, 135] предлагают на каждом временном слое вычислять не точное значение c^{n+1}, а некоторое итерационное приближение. (Данный подход справедлив не только для схемы Кранка — Николсона, рассматриваемой здесь в качестве примера, но и для многих других, в том числе и схем более высоких порядков (см., например, [124]).) При этом для итерационного процесса на каждом временном шаге используется специальное начальное приближение, полученное экстраполяцией нескольких ранее вычисленных векторов; например, при вычислении c^{n+1} в качестве начального приближения в итерационном процессе можно взять вектор [124]

$$c_0^{n+1} = 3c^n - 3c^{n-1} + c^{n-2}, \quad 2 \leqslant n < m.$$

В качестве итерационного процесса обычно предлагается использовать модифицированный метод сопряженных градиентов как обладающий достаточно высокой скоростью сходимости и не требующий знания (или оценок) максимального и минимального собственных значений решаемой системы.

Схематично модифицированный метод сопряженных градиентов можно описать на примере решения системы (III.17) следующим образом [135].

Для удобства изложения представим систему (III.17) в виде

$$A^{n+1/2}x = \Phi^n, \tag{III.18}$$

где

$$A^{n+1/2} = M + \frac{\tau}{2}K^{n+1/2}, \quad x = c^{n+1},$$

$$\Phi^n = \left(M - \frac{\tau}{2}K^{n+1/2}\right)c^n + \frac{\tau}{2}(F^{n+1} + F^n).$$

Выберем для данной системы некоторую вспомогательную, не зависящую от n матрицу A_0, которая, как и $A^{n+1/2}$, является положительно определенной, легко обратимой (точнее, решение системы вида $A_0 z = B$ не представляет значительных вычислительных трудностей) и которая удовлетворяет соотношению

$$\gamma_0 (A_0 y, y) \leqslant (A^{n+1/2} y, y) \leqslant \gamma_1 (A_0 y, y),$$

где $0 < \gamma_0 \leqslant \gamma_1$ — известные постоянные, а y — произвольный вектор соответствующего конечномерного подпространства. В частности, в случае системы (III.18) можно принять

$$A_0 = M + \frac{\tau}{2}K^{1/2} \quad \text{или} \quad A_0 = M + \frac{\tau}{2}K^0.$$

Пусть по определенному правилу выбрано начальное приближение x_0 к искомому решению системы (III.18) и вычислена невязка

$$r_0 = A^{n+1/2}x_0 - \Phi^n = s_0.$$

Тогда последующие итерации x_k, $k = 1, 2, \ldots, Q$, определяются по формулам

$$x_{k+1} = x_k + \alpha_k s_k, \quad \alpha_k = -\frac{(A_0^{-1}r_k, r_k)}{(s_k, A^{n+1/2}s_k)},$$

$$r_{k+1} = r_k + \alpha_k A^{n+1/2}s_k, \tag{III.19}$$

$$s_{k+1} = A_0^{-1}r_{k+1} + \beta_k s_k, \quad \beta_k = \frac{(A_0^{-1}r_{k+1}, r_{k+1})}{(A_0^{-1}r_k, r_k)}.$$

Таким образом, при реализации модифицированного метода сопряженных градиентов на каждом временном слое $t_n = n\tau$, $n = 0, 1, \ldots, m$, приходится решать системы уравнений с одинаковой матрицей A_0 и разными правыми частями. Для вычисления решения систем вида $A_0\rho_k = r_k$ удобно применить упоминаемый ранее прямой метод квадратных корней, так что каждый раз вектор $\rho_k = A_0^{-1}r_k$ будут находить при решении двух треугольных систем

$$S^T u = r_k, \quad S\rho_k = u, \quad A_0 = S^T S,$$

одинаковых для всех временных слоев. Подчеркнем, что на каждом временном слое $t_n = n\tau$, $n = 0, 1, \ldots, m$, выполняется только

фиксированное количество итераций (III.19) (см. [124]), поэтому общие вычислительные затраты получаются такого же порядка, как и в случае использования соответствующих разностных схем для решения системы (III.13) с постоянными (не зависящими от времени) коэффициентами.

Необходимо отметить [124], что упомянутый здесь итерационный подход к решению разностных систем на каждом временном слое обеспечивает особенно хорошие результаты для таких схем, как обратная итерация, схема Калахана, и ряда других, основанных на аппроксимации экспоненты e^{-x} рациональными функциями $r(x) = \frac{P(x)}{Q(x)}$, где $P(x)$ и $Q(x)$ — взаимно простые полиномы, удовлетворяющие, в частности, условию $-1 + \delta < \frac{P(x)}{Q(x)} < 1$ при некотором $\delta > 0$ и всех $x > 0$. Так как для схемы Кранка — Николсона $P(x)/Q(x) = \left(1 - \frac{x}{2}\right)\Big/\left(1 + \frac{x}{2}\right)$, $\delta = 0$, то при использовании рассматриваемого итерационного подхода хороших результатов (в смысле устойчивости и скорости сходимости) можно добиться лишь при условии $\tau \leqslant Ch^2$, где $C \leqslant 1$ — некоторая константа, $h = \max\limits_i [x_i - x_{i-1}]$. (В случае $\delta > 0$ никаких условий на соотношения τ и h не требуется.)

3. Численный пример. Рассмотрим построение численного решения МКЭ задачи

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left[(1 + 2tx)\frac{\partial u}{\partial x}\right] + 2t[1 + x(1 + \pi^2)]u =$$

$$= [1 + (1 + 4tx)\pi^2]e^{x+2t}\sin \pi x - 2\pi[1 + t(1 + 2x)]e^{x+2t}\cos \pi x,$$

$$0 < x < 1, \quad 0 < t < 1,$$

$$u(x, 0) = e^x \sin \pi x, \quad u(0, t) = 0, \quad u(1, t) = 0,$$

точное решение которой $u(x, t) = e^{x+2t}\sin \pi x$.

Вначале, как описано в п. 2 параграфа III.1, была выполнена полудискретизация задачи по пространственной переменной x, а затем по схеме Кранка — Николсона строилось решение задачи Коши для системы обыкновенных дифференциальных уравнений. В данном примере матрица K системы (III.13) зависит от времени t, но ее можно представить в виде $K = K(t) = K_0 + 2tK_1$, где матрицы K_i, $i = 0, 1$, от t не зависят, что облегчало построение матрицы $K(t)$ на каждом временном слое. Вектор c^{n+1} из системы

$$\left(M + \frac{\tau}{2}K^{n+1}\right)c^{n+1} = \left(M - \frac{\tau}{2}K^n\right)c^n + \frac{\tau}{2}(F^{n+1} + F^n),$$

$$c^0 = g, \quad n = 0, 1, 2, \ldots, m,$$

$$K^n = K_0 + 2n\tau K_1, \quad \tau = 1/m, \quad c^n = c(n\tau),$$

в данном примере вычислялся на каждом слое методом квадратных корней. Вычисленные компоненты вектора $c^n = [c_1^n, c_2^n, \ldots, c_s^n]^T$ являются значениями приближенного решения $u^N(x, n\tau)$ (или его произ-

Таблица 9

t_j	$\tau_m = 0{,}25$	$\tau_m = 0{,}125$	$\tau_m = 0{,}0625$
	$h = 0{,}25$		
0,25	0,0221	0,0207	0,0198
0,5	0,0335	0,0312	0,0307
0,75	0,0392	0,0378	0,0374
1	0,0439	0,0425	0,0421
	$h = 0{,}125$		
0,25	0,00874	0,00624	0,00539
0,5	0,0116	0,00923	0,00861
0,75	0,0132	0,0112	0,0107
1	0,0145	0,0128	0,0124
	$h = 0{,}0625$		
0,25	0,00552	0,00242	0,00160
0,5	0,00566	0,00312	0,00247
0,75	0,00570	0,00363	0,00308
1	0,00596	0,00405	0,00356

Таблица 10

t_j	$\tau_m = 0{,}25$	$\tau_m = 0{,}125$	$\tau_m = 0{,}062\,5$
	$h = 0{,}25$		
0,25	0,00328	0,00114	$0{,}80 \cdot 10^{-3}$
0,50	0,00246	$0{,}98 \cdot 10^{-3}$	$0{,}69 \cdot 10^{-3}$
0,75	0,00200	$0{,}87 \cdot 10^{-3}$	$0{,}85 \cdot 10^{-3}$
1	0,00172	$0{,}81 \cdot 10^{-3}$	$0{,}63 \cdot 10^{-3}$

водной по x) в узловых точках сетки по пространственной переменной x. Например, в случае линейных базисных функций $\varphi_i^N(x)$ и равномерной сетки с шагом h имеем $c_i^n = u^N(x_i, n\tau)$, $x_i = ih$, $i = 1, 2, \ldots, s = N - 1$, $h = 1/N$.

При полудискретизации рассматриваемой задачи использовались как линейные, так и кусочно-кубические (эрмитовы) базисные функции. Расчет выполнялся на ЭВМ МИР-2 при разрядности 10. Интегралы вычислялись по квадратурным формулам Гаусса с тремя узлами.

В случае линейных базисных функций область [0, 1] изменения x разбивалась на N, $N = 4, 8, 16$, равных элементарных отрезков $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, N$, т. е. $h = 0{,}25; 0{,}125; 0{,}0625$. При решении каждой из получаемых после полудискретизации систем обыкновенных дифференциальных уравнений по схеме Кранка — Николсона выбирались следующие шаги τ_m по временной переменной t: $\tau_m = 1/m$, $m = 4, 8, 16$. Полученные численные результаты были оформлены в виде таблиц. Для удобства восприятия и оценки точности большого массива выходных данных в таблицах представлены только максимальные относительные погрешности вычисленных значений приближенного решения $u^N(x_i, t_j)$ на некоторых временных слоях. Относительная погрешность ε_{ni} вычисленных значений приближенного решения на n-м слое $t_n = nt_m$, $n = 1, 2, \ldots, m$, находилась по формуле

$$\varepsilon_{ni} = \frac{|u(x_i, n\tau_m) - u^N(x_i, n\tau_m)|}{|u(x_i, n\tau_m)|},$$

$$u^N(x_i, n\tau_m) \equiv c_i^n,$$

а затем для каждого слоя выбиралось значение $\varepsilon_n = \max\limits_i \varepsilon_{ni}$. В табл. 9 для случая линейных базисных функций представлены значения ε_n только тех слоев, где $n\tau_m \equiv t_j = 0{,}25; 0{,}5; 0{,}75; 1$.

По такому же принципу в табл. 10 представлены результаты, полученные при полудискретизации исходной задачи посредством

кубических эрмитовых базисных функций. Область изменения пространственной переменной здесь разбивалась только на $N = 4$ равных элементарных отрезков $[x_{i-1}, x_i]$, $i = 1, 2, 3, 4$, а $\tau_m = 0{,}25,\ 0{,}125;\ 0{,}0625$.

Полученные результаты свидетельствуют о достаточно хорошей точности вычисления приближенного решения по описанной методике. Особенно высокая точность достигается при использовании для полудискретизации по пространственной переменной базисных функций повышенных степеней (выше первой).

4. Некоторые варианты применения МКЭ для решения параболических уравнений. Упомянем кратко и другие, отличающиеся от описанного в п. 2 данного параграфа подходы к использованию МКЭ для решения нестационарных задач.

Приближение к обобщенному решению, определяемому интегральным тождеством (III.8), (III.9), можно строить, например, следующим образом. Вначале методом Бубнова — Галеркина, как и в п. 2 настоящего параграфа, необходимо выполнить полудискретизацию по пространственной переменной x, а затем полученную систему обыкновенных дифференциальных уравнений (см. (III.13), (III.15))

$$M\frac{dc}{dt} + Kc(t) = F(t), \quad 0 < t < T, \tag{III.20}$$

$$c(t) = [c_1(t),\ c_2(t),\ \ldots,\ c_s(t)]^T,$$

с начальным условием

$$c(0) = g \tag{III.21}$$

решать методом конечных элементов, используя вновь метод Бубнова — Галеркина для дискретизации по временной переменной. Для этого отрезок $[0, T]$ рассматривается как совокупность $\mathfrak{M}$ одномерных элементов $[t_{k-1}, t_k]$, $k = 1, 2, \ldots, \mathfrak{M}$, а соответствующие допустимые кусочно-полиномиальные вектор-функции можно выбрать в виде

$$c^{\mathfrak{M}}(t) = g\psi_0^{\mathfrak{M}}(t) + \sum_{j=1}^{r} c_j^{\mathfrak{M}}\psi_j^{\mathfrak{M}}(t), \tag{III.22}$$

где $\{\psi_j^{\mathfrak{M}}(t)\}_0^r$ — базисные функции МКЭ, подробно описанные в параграфе II.4, $c_j^{\mathfrak{M}}$ — искомые числовые векторы, компоненты которых будем обозначать через c_{ij}, $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, r$ (см. (III.10)), т. е. $c_j^{\mathfrak{M}} = [c_{1j},\ c_{2j},\ \ldots,\ c_{sj}]^T$.

Заметим, что размерность вектора $c_j^{\mathfrak{M}}$ определяется размерностью конечномерного подпространства, используемого при пространственной полудискретизации исходной задачи (III.1) — (III.3), т. е. равна s.

Выражение (III.22) можно записать и в матричной форме, а именно

$$c^{\mathfrak{M}}(t) = g\psi_0^{\mathfrak{M}}(t) + C^{\mathfrak{M}}\psi^{\mathfrak{M}}(t),$$

где $C^{\mathfrak{M}} = (c_{ij})$ — прямоугольная матрица с размерами $s \times r$, элементы которой — действительные числа c_{ij}, $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, r$, $\psi(t) = [\psi_1^{\mathfrak{M}}(t),\ \psi_2^{\mathfrak{M}}(t),\ \ldots,\ \psi_r^{\mathfrak{M}}(t)]^T$.

Для отыскания приближенного решения МКЭ задачи (III.20), (III.21) применяют обычный метод Бубнова — Галеркина, т. е. подставляют в (III.20) вектор $c^{\mathfrak{M}}(t)$ вместо $c(t)$ и умножают скалярно i-е уравнение системы дифференциальных уравнений (III.20) на каждую базисную функцию $\psi_j^{\mathfrak{M}}(t)$, $j = 1, 2, \ldots, r$, которая участвует в разложении компоненты $c_i^{\mathfrak{M}}(t)$ вектор-функции $c^{\mathfrak{M}}(t)$ (III.22). В результате получают систему линейных алгебраических уравнений для вычисления элементов c_{ij}, $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, r$:

$$Az = b.$$

Здесь A — квадратная ленточная матрица порядка sr, элементы которой не зависят от t, искомое решение

$$z = [c_{11}, c_{21}, \ldots, c_{s1}, c_{12}, c_{22}, \ldots, c_{s2}, \ldots, c_{1r}, c_{2r}, \ldots, c_{sr}]^T,$$

b — известный вектор.

Замечание. Компоненты $c_i^{\mathfrak{M}}(t)$, $i = 1, 2, \ldots, s$, приближенного решения $c^{\mathfrak{M}}(t)$ могут принадлежать разным конечномерным подпространствам $P_{n_i}^h$, т. е. $c_i^{\mathfrak{M}}(t)$ строится в виде разложения

$$c_i^{\mathfrak{M}}(t) = g_i\psi_{i0}^{\mathfrak{M}}(t) + \sum_{j=1}^{r_i} c_{ij}\psi_{ij}^{\mathfrak{M}}(t), \quad \psi_{ij}^{\mathfrak{M}} \in P_{n_i}^h, \quad i = 1, 2, \ldots, s.$$

Процедура получения системы алгебраических уравнений для вычисления коэффициентов c_{ij}, $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, r_i$, остается прежней: i-е уравнение системы (III.20) умножается скалярно на каждую функцию $\psi_{ij}^{\mathfrak{M}}(t)$, $j = 1, 2, \ldots, r_i$; полученная система имеет порядок, равный $\sum_{i=1}^{s} r_i$.

Необходимо вновь подчеркнуть, что выбор базисных функций и соответственно гладкость приближенного решения существенно зависят от гладкости искомого решения дифференциальной задачи, а следовательно, от гладкости ее коэффициентов и граничных условий.

Учитывая, что c_{ij} вычисляются при решении системы уравнений порядка $sr\left(\sum_{i=1}^{s} r_i\right)$ и на практике может интересовать значение искомого решения лишь при $t = T$, бывает целесообразно ограничиться в данном подходе небольшим числом разбиений отрезка $[0, T]$, используя конечные элементы высокого порядка точности. Очевидно также, что данный способ может быть действительно эффективным лишь в случае коэффициентов дифференциальной задачи, не зависящих от времени.

Наконец, упомянем еще один подход к построению приближенного обобщенного решения задачи (III.1) — (III.3) исходя из интегрального тождества (III.5). В этом случае искомое решение $u(x, t)$ можно рассматривать как функцию двух переменных, определенную на прямоугольнике Q_T двумерного пространства, и для дискретизации задачи использовать стандартные двумерные конечные элементы, например треугольники или прямоугольники.

Приближенное обобщенное решение $u^N(x, t) \in W^1_{2,0}(Q^T)$ представляется теперь в виде

$$u^N(x, t) = \sum_i c_i^N \varphi_i^N(x, t), \qquad \text{(III.23)}$$

где $\varphi_i^N(x, t)$ — соответствующие базисные функции множества $P_n^h \subset \subset W^1_{2,0}(Q_T)$, а c_i^N — искомые значения приближенного решения $u^N(x, t)$ или некоторых его производных в узлах конечно-элементной сетки, N — количество элементов, на которые разбит прямоугольник Q_T. Неизвестные числовые параметры c_i^N определяются из условия удовлетворения функции $u^N(x, t)$ интегральному тождеству (III.5), где нужно положить $u(x, t) = u^N(x, t)$, а в качестве $\eta(x,t)$ выбрать базисные функции $\psi_i(x, t)$ некоторого конечномерного множества $P_k^h \subset \overset{0}{W}{}_2^{1,0}(Q_T)$, размерность которого совпадает с размерностью P_n^h. В частности, множества P_k^h и P_n^h могут совпадать.

В результате будет получена система линейных алгебраических уравнений относительно c_i^N, порядок которой определяется конечно-элементной сеткой и количеством неизвестных фиксированных параметров в ее узлах. (Некоторые подробности реализации алгоритма см. в [46].) Вычисленное решение c_i^N, $i = 1, 2, \ldots, r$, этой системы определит значения искомого приближенного решения $u^N(x, t)$ (и некоторых его производных) в узлах сетки, а согласно (III.23) можно получить и аналитическое представление $u^N(x, t)$.

III.2. Сходимость метода конечных элементов при решении параболических уравнений

Приведем некоторые результаты оценок погрешности приближенного решения, полученного посредством полудискретизации по пространственным переменным с последующим применением разностных схем по временной переменной.

Указанным оценкам посвящено много исследований различных авторов, в частности работы [40, 101, 124, 134, 135, 138, 159], где приводится большая библиография по данному вопросу.

Наиболее простым для рассмотрения является случай, когда коэффициенты $k(x, t)$, $q(x, t)$ и краевые условия задачи (III.1) — (III.3) не зависят от времени. Тогда весьма успешным оказывается способ отыскания границ ошибок, основанный на исследовании собственных функций дифференциальной и дискретной задач, на разложении искомых решений по этим собственным функциям. (Подробнее см. [101].) В частности, в случае однородного уравнения вида (III.1), $f(x, t) = 0$, в [159] для схемы Кранка — Николсона получен следующий результат.

Пусть при полудискретизации по пространству использовались базисные функции $\varphi_i^N(x)$, $i = 1, 2, \ldots, s$, конечномерного подпространства $\overset{0}{P}{}_p^h \subset \overset{0}{W}{}_2^1(0, l)$, обладающего следующим свойством (см. теорему

II.4): для любой функции $v(x) \in \overset{0}{W}{}_2^1 \cap W_2^{p+1}$ существует функция $v^N(x) \in \overset{0}{P}{}_p^h$ такая, что

$$\|v(x) - v^N(x)\|_{2,j} \leqslant Ch^{p+1-j} \left\| \frac{d^{p+1}v}{dx^{p+1}} \right\|, \quad j = 0, 1,$$

где C — постоянная, не зависящая от h и $v(x)$.

Предположим, что начальная функция $u_0(x)$ (см. (III.2)) удовлетворяет условиям

$$u_0(x) \in W_2^r(0, l),$$
$$u_0(0) = u_0(l) = 0, \tag{III.24}$$
$$Lu_0 = L^2u_0 = \cdots = L^{\left[\frac{r-1}{2}\right]}u_0 = 0 \quad \text{при } x = 0, \ x = l.$$

Здесь

$$Lv \equiv -\frac{\partial}{\partial x}\left(k(x)\frac{\partial v}{\partial x}\right) + q(x)v.$$

Теперь, наконец, приведем упомянутый выше результат из [159].

Теорема III.1. *Пусть $u_0(x)$ удовлетворяет соотношениям (III.24) при $r = \max((p+3), 6)$. Тогда для приближений $u^N(x, n\tau)$, определяемых по методу конечных элементов посредством (III.10), (III.13), (III.15), (III.16), справедлива оценка*

$$\|u(x, n\tau) - u^N(x, n\tau)\|_{2,1} \leqslant C(h^p + \tau^2)\|u_0\|_{2,r}, \quad 0 \leqslant n \leqslant \frac{T}{\tau}.$$

В этой же работе приводится результат, относящийся к оценке погрешности схемы Кранка — Николсона на достаточно большом временном отрезке, $0 \leqslant n < \infty$, а именно

$$\max_{0 \leqslant n < \infty} \|u(x, n\tau) - u^N(x, n\tau)\|_{L_2} \leqslant C(h^{p+1} + \tau^2) \lg \frac{1}{\tau} \|u_0\|_{2,r}, \tag{III.25}$$

если $u_0(x)$ удовлетворяет (III.24) при $r = \max(p + 1, 4)$. И хотя оценка (III.25) позволяет ожидать достаточно хорошие результаты, однако на практике вычисление приближенных решений при умеренном шаге по времени и недостаточной гладкости начальной функции $u_0(x)$ оказывается неудовлетворительным. А объясняется это тем, что схема Кранка — Николсона не обладает сильной устойчивостью на бесконечности и удовлетворительные результаты возможно получить лишь в случае, когда τ стремится к нулю быстрее h, а именно если $\tau = h^\alpha$, $\alpha \geqslant 1$. В связи с этим в [40] предлагается несколько других одношаговых и многошаговых методов, позволяющих получить хорошие численные результаты без ограничений на τ и h.

Остановимся теперь несколько подробнее на другом способе оценки погрешностей приближенного решения. Этот способ оказывается применимым и в случае, когда коэффициенты задачи (III.1) — (III.3) являются достаточно гладкими функциями временной переменной. Однако чтобы избежать чисто технических трудностей (см. [101]), рассуждения и здесь будут вестись в предположении, что функции $k(x,$

$t) \equiv k(x)$, $q(x, t) \equiv q(x)$ и подчинены прежним условиям, т. е. являются достаточно гладкими и удовлетворяют условиям (III.4а) и (III.4б).

Оценим в норме пространства $L_2(0, l)$ погрешность полудискретизации по пространственной переменной для задачи (III.1) — (III.3), т. е. оценим разность $u(x, t) - u^N(x, t)$, где $u^N(x, t)$ определяется соотношениями (III.10) — (III 12). Заметим, что в силу исходных предположений (см. п. 1 параграфа III.1) точное решение $u(x, t) \in \in \overset{0}{W}{}_2^1(0, l)$ при каждом фиксированном $t \geqslant 0$.

Пусть $\tilde{u}^N(x, t)$ для каждого фиксированного $t \geqslant 0$ является проекцией $u(x, t) \in \overset{0}{W}{}_2^1(0, l)$ на $\overset{0}{P}{}_p^h \subset \overset{0}{W}{}_2^1(0, l)$ в смысле энергетического скалярного произведения $[\cdot, \cdot]_A$ оператора A, определяемого формулой

$$Av = -\frac{\partial}{\partial x}\left(k(x)\frac{\partial v}{\partial x}\right) + q(x)v \qquad \text{(III.26)}$$

на множестве функций $v(x) \in D(A)$, удовлетворяющих условиям

$$v \in C^2[0, l], \quad v(0) = v(l) = 0. \qquad \text{(III.27)}$$

Это означает, что для каждого $t \geqslant 0$ справедливо соотношение

$$[u(x, t) - \tilde{u}^N(x, t), v^N(x, t)]_A = 0, \quad \forall v^N(x, t) \in \overset{0}{P}{}_p^h. \qquad \text{(III.28)}$$

(Напомним, что энергетическое пространство H_A описанного оператора A и пространство $\overset{0}{W}{}_2^1(0, l)$ состоят из одинаковых функций.)

Из (III.28) следует, что

$$[u - \tilde{u}^N, u - \tilde{u}^N]_A = \min_{v^N \in \overset{0}{P}{}_p^h} [u - v^N, u - v^N]_A. \qquad \text{(III.29)}$$

Действительно, пусть $v^N = \tilde{u}^N + w^N$ при $\forall w^N \in \overset{0}{P}{}_p^h$. Тогда с учетом (III.28)

$$[u - v^N, u - v^N]_A = [u - \tilde{u}^N, u - \tilde{u}^N]_A + [w^N, w^N]_A \geqslant$$

$$\geqslant [u - \tilde{u}^N, u - \tilde{u}^N]_A$$

и равенство возможно лишь при $w^N \equiv 0$, т. е. справедливо равенство (III.29).

Таким образом, при каждом значении $t \geqslant 0$ проекцию $\tilde{u}^N(x, t)$ можно рассматривать как приближенное решение некоторой краевой задачи с оператором (III.26), (III.27), и в силу теоремы II.5 имеем

$$\|u(x, t) - \tilde{u}^N(x, t)\| \leqslant Ch^{p+1}\|u\|_{2,p+1}. \qquad \text{(III.30)}$$

Искомую погрешность представим теперь в виде

$$u(x, t) - u^N(x, t) =$$

$$= (u(x, t) - \tilde{u}^N(x, t)) + (\tilde{u}^N(x, t) - u^N(x, t)), \qquad \text{(III.31)}$$

где согласно (III.30) оценка в норме $L_2(0, l)$ для $u(x, t) - \tilde{u}^N(x, t)$ известна.

Для оценки второго слагаемого правой части (III.31) используем лемму (см. работу [101], лемму 7.1).

Лемма. *Для погрешности*

$$z(x, t) = \tilde{u}^N(x, t) - u^N(x, t)$$

справедливо тождество

$$\left(\frac{\partial z}{\partial t}, z\right) + [z, z]_A = \left(\left(\frac{\widetilde{\partial u}}{\partial t}\right)^N - \frac{\partial u}{\partial t}, z\right), \qquad \text{(III.32)}$$

где $(\cdot, \cdot)$ *— скалярное произведение в* $L_2(0, l)$, $\left(\frac{\widetilde{\partial u}}{\partial t}\right)^N$ *— проекция в смысле (III.28)* $\frac{\partial u}{\partial t}$ *на* $\overset{0}{P}{}_p^h$ *при каждом фиксированном значении* $t \geqslant 0$.

Согласно известному соотношению между энергетической нормой и нормой гильбертова пространства $L_2(0, l)$, в котором определен оператор A,

$$[z, z]_A \geqslant \gamma \| z \|^2, \quad \gamma > 0,$$

и очевидному равенству

$$\left(\frac{\partial z}{\partial t}, z\right) \equiv \frac{1}{2} \frac{d}{dt} \| z \|^2 = \| z \| \frac{d \| z \|}{dt}$$

имеем

$$\left(\frac{\partial z}{\partial t}, z\right) + [z, z]_A \geqslant \| z \| \frac{d \| z \|}{dt} + \gamma \| z \|^2.$$

А так как по неравенству Коши — Буняковского правая часть тождества (III.32) ограничена величиной $\left\| \left(\frac{\widetilde{\partial u}}{\partial t}\right)^N - \frac{\partial u}{\partial t} \right\| \| z \|$, то, объединяя указанные результаты, получаем

$$\frac{d}{dt} \| z \| + \gamma \| z \| \leqslant \left\| \left(\frac{\widetilde{\partial u}}{\partial t}\right)^N - \frac{\partial u}{\partial t} \right\|.$$

Умножив обе части этого неравенства на $e^{\gamma t}$, а затем проинтегрировав его по t, найдем

$$\| \tilde{u}^N(x, t) - u^N(x, t) \| \leqslant e^{-\gamma t} \| \tilde{u}^N(x, 0) - u^N(x, 0) \| + \int_0^t e^{\gamma(\tau - t)} \left\| \left(\frac{\widetilde{\partial u}}{\partial t}\right)^N - \frac{\partial u}{\partial t} \right\| d\tau. \qquad \text{(III.33)}$$

Независимо от того, выбирается $u^N(x, 0)$ согласно (III.12) или (III.14), справедлива оценка

$$\| \tilde{u}^N(x, 0) - u^N(x, 0) \| \leqslant C_2 h^{p+1} \| u_0 \|_{2, p+1}. \qquad \text{(III.34)}$$

Так как в силу наших предположений $\frac{\partial u}{\partial t} \in \overset{0}{W}{}_2^1(0, l)$ при каждом фиксированном значении $t > 0$, справедлива оценка, аналогичная (III.30):

$$\left\| \left(\frac{\widetilde{\partial u}}{\partial t} \right)^N - \frac{\partial u}{\partial t} \right\| \leqslant C_3 h^{p+1} \left\| \frac{\partial u}{\partial t} \right\|_{2,p+1}. \qquad \text{(III.35)}$$

Объединение результатов (III.30), (III.33) — (III.35) делает очевидным следующее утверждение.

Теорема III.2. *Пусть $u(x, t)$ — обобщенное решение задачи (III.1) — (III.3), удовлетворяющее соотношениям (III.8), (III.9), а $u^N(x, t)$ — полудискретное приближение к $u(x, t)$ в смысле (III.10), (III.11) и (III.12) или (III.14) в подпространстве $\overset{0}{P}{}_p^h \subset \overset{0}{W}{}_2^1(0, l)$ кусочно-полиномиальных функций степени p.*

Тогда если $u(x, t) \in \overset{0}{W}{}_2^1(0, l) \cap W_2^{p+1}(0, l)$ при каждом значении $t \geqslant 0$, то

$$\| u(x, t) - u^N(x, t) \| \leqslant C h^{p+1} \left(\| u \|_{2,p+1} + e^{-\gamma t} \| u_0 \|_{2,p+1} + \right.$$

$$\left. + \int_0^t e^{\gamma(\tau - t)} \left\| \frac{\partial u}{\partial t} \right\|_{2,p+1} d\tau \right).$$

Таким образом, при полудискретизации начально-краевых задач для параболических уравнений вариантом МКЭ, основанным на процессе Бубнова — Галеркина, порядок погрешности метода таков, как и в краевых задачах для эллиптических уравнений.

В ряде работ, в частности перечисленных в начале данного раздела, изучается также погрешность дискретизации по времени t, т. е. погрешность $\tilde{u}^N(x, n\tau) - u^N(x, n\tau)$, где $\tau = T/m$, $n = 0, 1, \ldots, m$.

Для временной дискретизации используются, как отмечалось, различные устойчивые разностные схемы.

Например, для схемы Кранка — Николсона, кроме результата М. Зламала (см. теорему III.1), получены аналогичные оценки в различных нормах для случая зависимости от t коэффициентов неоднородного уравнения (III.1), а также при различных граничных условиях. Ряд интересных результатов, касающихся применения для решения параболических уравнений с зависящими от времени коэффициентами эффективных разностных схем высокого порядка точности, можно найти в работе [124].

Глава IV

ЗАДАЧИ НА СОБСТВЕННЫЕ ЗНАЧЕНИЯ

Большой и важный класс научно-технических проблем связан с вопросами устойчивости или колебаний некоторых систем и их элементов (см., например, [49]). Математически эти проблемы формулируются в виде задач на собственные значения некоторых операторов. В настоящей главе рассматривается решение методом конечных элементов задач на собственные значения для обыкновенных дифференциальных операторов второго и четвертого порядков, а также затрагиваются вопросы применения МКЭ для подобных задач в более общих случаях.

IV.1. Постановка задач

1. Обыкновенные дифференциальные уравнения второго порядка. Простейший пример задачи на собственные значения возникает при исследовании проблемы устойчивости стержня длины l, один конец которого защемлен, а на другой — свободный — действует в центре тяжести концевой площади сжимающая сила P, направленная вдоль оси стержня [49]. Допустим, что все поперечные сечения стержня одинаковы и их главные оси инерции лежат в двух фиксированных направлениях.

Если значение P выше некоторого критического, то, как известно, прямолинейное положение стержня становится нестабильным и в состоянии устойчивого равновесия стержень имеет изогнутую форму. Рассмотрим начало потери стабильности при продольном изгибе, т. е. рассмотрим положение равновесия стержня, незначительно отличающегося от прямолинейной формы. Пусть начало координат расположено в точке приложения силы P, а ориентация осей x, y такая, как на рис. 25. В слабоизогнутом положении равновесия, незначительно отличающемся от прямолинейного, уравнение упругой линии $y(x)$ имеет вид

$$M = Py = -EJ\frac{d^2y}{dx^2} \equiv -\alpha\frac{d^2y}{dx^2},$$

где M — изгибающий момент, E — модуль упругости, J — осевой момент инерции сечения, $\alpha = EJ$ — жесткость на изгиб.

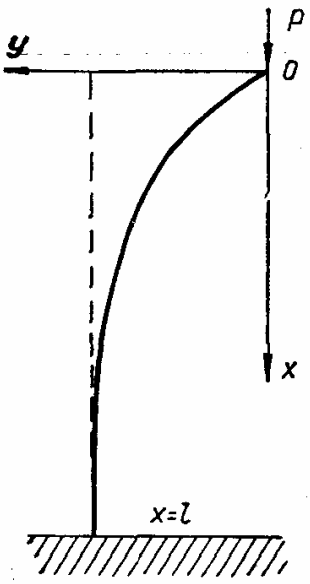

Рис. 25.

Предполагая жесткость постоянной ($\alpha = \text{const}$) и принимая $P/\alpha = \lambda = \omega^2$, записываем уравнение в виде

$$\frac{d^2y}{dx^2} = -\lambda y = -\omega^2 y, \quad 0 < x < l. \tag{IV.1}$$

Из рис. 25 непосредственно вытекают следующие краевые условия для искомого решения:

$$y(0) = 0, \quad \frac{dy}{dx}(l) = 0. \tag{IV.2}$$

Таким образом, задача отыскания критической нагрузки P и соответствующей формы устойчивого равновесия стержня при продольном изгибе свелась к решению математической задачи (IV.1), (IV.2), т. е. к отысканию значений числового параметра $\lambda \equiv \omega^2$, при которых задача имеет нетривиальные решения. Эти значения параметра λ называются собственными числами задачи, а отвечающие им решения $y(x)$ — собственными функциями задачи.

В данном простом примере поставленная математическая задача решается весьма просто, в замкнутой форме. Действительно, общее решение дифференциального уравнения (IV.1) имеет вид

$$y(x) = c_1 \cos \omega x + c_2 \sin \omega x,$$

а учет краевых условий дает

$$c_1 = 0, \quad \omega c_2 \cos \omega l = 0.$$

Из второго равенства следует, что при

$$\omega l = (2k - 1)\pi/2, \quad k = 1, 2, \ldots,$$

задача (IV.1), (IV.2) имеет нетривиальные решения $y_k(x) = c_2 \sin \frac{(2k-1)\pi}{2l} x$, которые соответствуют собственным числам задачи $\lambda_k = \left(\frac{(2k-1)\pi}{2l}\right)^2$, $k = 1, 2, \ldots$ Итак, критические нагрузки при продольном изгибе стержня имеют значения $P_k = \alpha\lambda_k = \alpha\left[\frac{(2k-1)\pi}{2l}\right]^2$, а соответствующие формы положения устойчивого равновесия стержня — $y_k(x) = c \sin \frac{(2k-1)\pi}{2l} x$, где $c = \text{const}$.

К совершенно аналогичной задаче на собственные значения приводит исследование продольных колебаний стержня длины l, один конец которого ($x = 0$) свободен, а второй ($x = l$) жестко закреплен [48] (рис. 26). Смещение $y(x)$ при продольных колебаниях удовлетворяет дифференциальному уравнению

$$-E\frac{d}{dx}\left(s(x)\frac{dy}{dx}\right) = \omega^2 \rho s(x) y, \quad 0 < x < l,$$

где $s(x)$ — площадь поперечного сечения, ρ — плотность и E — модуль упругости материала, из которого сделан стержень. Искомое решение $y(x)$ должно удовлетворять краевым условиям

$$\frac{dy}{dx}(0) = 0, \quad y(l) = 0.$$

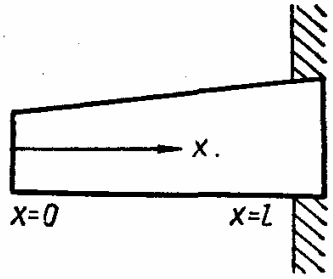

Рис. 26.

Здесь требуется найти круговые частоты ω собственных колебаний и формы $y(x)$ колебаний, отвечающие этим частотам. Эта задача тоже может быть решена в замкнутом виде, если $s(x) = \text{const}$.

Однако в общем случае дифференциальных уравнений с переменными коэффициентами решение задачи на собственные значения сопряжено со значительными трудностями и строится в основном численными методами, в частности, решение может быть найдено методом конечных элементов.

Рассмотрим более общую задачу на собственные значения

$$-\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) + q(x)u = \lambda\rho(x)u, \quad 0 < x < l, \tag{IV.3}$$

$$u(0) = 0, \quad u(l) = 0, \tag{IV.4}$$

где функции $k(x) \geqslant k_0 > 0$, $\frac{dk}{dx}$, $q(x) \geqslant 0$, $\rho(x) \geqslant \rho_0 > 0$ непрерывны на $[0, l]$, λ — числовой параметр. Задачу (IV.3), (IV.4) можно записать в виде операторного уравнения в гильбертовом пространстве $H = L_2(0, l)$. Для этого достаточно ввести оператор A формулой

$$Au = -\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) + q(x)u, \tag{IV.5}$$

оператор B формулой $Bu = \rho(x)u$, и в качестве области определения $D(A) \subset L_2(0, l)$ принять множество функций

$$u(x) \in C^2[0, l], \quad u(0) = 0, \quad u(l) = 0, \tag{IV.6}$$

а в качестве $D(B)$ взять все пространство $L_2(0, l)$,

$$D(A) \subset D(B) = L_2(0, l).$$

Теперь задачу (IV.3), (IV.4) можно представить в виде операторного уравнения

$$Au = \lambda Bu. \tag{IV.7}$$

Если $\rho(x) \equiv 1$, то уравнение (IV.7) принимает вид $Au = \lambda u$, т. е. $B \equiv I$ — тождественный оператор.

Оператор A (см. соотношения (IV.5), (IV.6)) — положительно определенный в силу условий, наложенных на коэффициенты уравнения (IV.3). В этом легко убедиться, повторив дословно рассуждения п. 1 параграфа II.1, относящиеся к оператору, заданному той же формулой (IV.5), но на множестве функций, удовлетворяющих более общим краевым условиям (см. (II.5), (II.6)). Нетрудно проверить и то, что энергетическое пространство H_A рассматриваемого оператора

состоит из тех же функций, что и пространство $\overset{0}{W}{}_2^1(0, l)$; норма в H_A определяется соотношением

$$\|u\|_A^2 = [u, u]_A = \int_0^l \left(k(x)\left(\frac{du}{dx}\right)^2 + q(x)u^2\right)dx,$$

причем в силу условий, наложенных на $k(x)$ и $q(x)$, справедливо неравенство

$$\|u\|_A^2 \geqslant k_0 \int_0^l \left(\frac{du}{dx}\right)^2 dx \equiv k_0 \left\|\frac{du}{dx}\right\|_{L_2}^2. \qquad \text{(IV.8)}$$

Теперь для доказательства существования собственных значений задачи (IV.3), (IV.4) достаточно показать, что оператор A удовлетворяет условиям теоремы I.7, т. е. что любое множество элементов, ограниченное в энергетической норме, компактно в норме исходного пространства $H = L_2(0, l)$.

Итак, пусть множество M функций $u(x)$ ограничено в энергетической норме:

$$\|u\|_A \leqslant C, \quad \forall u \in M \subset H_A. \qquad \text{(IV.9)}$$

Но тогда согласно (IV.8) будет ограничено в $L_2(0, l)$ множество производных $\frac{du}{dx}$:

$$\left\|\frac{du}{dx}\right\|_{L_2} \leqslant C_1 \quad \text{при } \forall u(x) \in M. \qquad \text{(IV.10)}$$

Как известно, для любой функции $u(x) \in \overset{0}{W}{}_2^1(0, l)$ справедливо равенство

$$u(x) = \int_0^x \frac{du}{dt}\,dt, \qquad \text{(IV.11)}$$

которое можно представить в виде

$$u(x) = \int_0^l k(x, t)\frac{du}{dt}\,dt, \qquad \text{(IV.12)}$$

где функция $k(x, t)$ задана соотношением

$$k(x, t) = \begin{cases} 1, & 0 \leqslant t \leqslant x, \\ 0, & x < t \leqslant l. \end{cases}$$

Интегральный оператор

$$Kv \equiv u(x) = \int_0^l k(x, t)v(t)\,dt, \quad v \in L_2(0, l)$$

является оператором Фредгольма, который, как известно (см. [66]), вполне непрерывен в пространстве $L_2(0, l)$. Иными словами, этот оператор преобразует любое ограниченное множество из $L_2(0, l)$ в мно-

жество, компактное в этом же пространстве. Отсюда, полагая $v(x) \equiv \equiv \frac{du}{dx}$ и учитывая (IV.9) — (IV.12), непосредственно убеждаемся, что множество функций $u(x) \in M$ будет компактным в $L_2(0, l) = H$; это и требовалось показать.

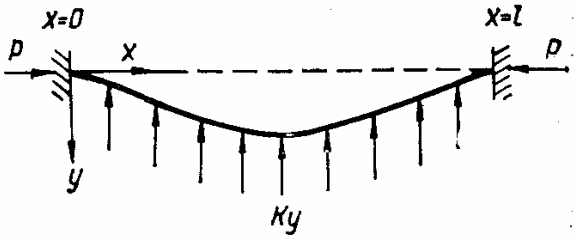

Рис. 27.

Таким образом, установлено, что задача (IV.3), (IV.4) имеет бесконечную последовательность собственных чисел

$$0 < \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n \leqslant \ldots$$

с единственной предельной точкой на бесконечности и соответствующую им ортонормированную в H систему собственных функций

$$u_1(x),\ u_2(x),\ \ldots,\ u_n(x),\ \ldots$$

Вопрос о построении МКЭ приближений к собственным значениям данной задачи будет рассмотрен в параграфе IV.2.

2. Обыкновенные дифференциальные уравнения четвертого порядка. Коснемся кратко постановки задач на собственные значения для операторов четвертого порядка. В качестве примера, приводящего к такому виду задач, рассмотрим устойчивость сжатого стержня на упругом основании [49]. Пусть стержень длины l и переменного сечения, подвергающийся действию осевой сжимающей силы P, лежит на упругом основании. Пусть сила, действующая на единицу длины стержня при прогибе на величину y, пропорциональна прогибу, т. е. Ky. Тогда при малом изгибе (рис. 27) уравнение упругой линии имеет вид

$$M = Py + \int_0^x Ky(\xi)(x - \xi)\,d\xi = -\alpha \frac{d^2y}{dx^2}, \quad \alpha = EJ(x),$$

где, как и в предыдущем примере M — изгибающий момент, E — модуль упругости, $J(x)$ — момент инерции сечения стержня с абсциссой x.

После двукратного дифференцирования получаем дифференциальное уравнение

$$E\frac{d^2}{dx^2}\left(J(x)\frac{d^2y}{dx^2}\right) + Ky = -P\frac{d^2y}{dx^2}. \qquad \text{(IV.13)}$$

При жестком закреплении обоих концов стержня краевые условия следующие:

$$y(0) = y(l) = 0, \quad \frac{dy}{dx}(0) = \frac{dy}{dx}(l) = 0. \qquad \text{(IV.14)}$$

(В случае шарнирного закрепления концов краевые условия таковы:

$$y(0) = y(l) = 0, \quad \frac{d^2y}{dx^2}(0) = \frac{d^2y}{dx^2}(l) = 0.)$$

Задача (IV.13), (IV.14) как задача об устойчивости сжатого стержня состоит в отыскании значений «критических нагрузок» P, при которых уравнение (IV.13) имеет нетривиальные решения, удовлетво-

ряющие (IV.14). Наиболее интересной является наименьшая критическая нагрузка.

В операторной форме в пространстве $H = L_2(0, l)$ задачу (IV.13), (IV.14) можно записать в виде

$$Au = \lambda Bu,$$

где

$$Au = E\frac{d^2}{dx^2}\left(J(x)\frac{d^2u}{dx^2}\right) + Ku, \quad Bu = -\frac{d^2u}{dx^2},$$

$$\lambda = P,$$

$D(A)$ — множество функций $u(x)$, удовлетворяющих условиям

$$u(x) \in C^4[0, l], \quad u(0) = u(l) = 0, \tag{IV.15}$$

$$\frac{du}{dx}(0) = \frac{du}{dx}(l) = 0,$$

$D(B)$ — множество функций $u(x)$, удовлетворяющих условиям

$$u(x) \in C^2[0, l], \quad u(0) = u(l) = 0. \tag{IV.16}$$

Таким образом, рассматриваемый пример из технической механики привел к решению задачи на собственные значения операторного уравнения

$$Au - \lambda Bu = 0.$$

Рассмотрим теперь достаточно общий пример задачи на собственные значения для уравнения четвертого порядка

$$\frac{d^2}{dx^2}\left(k(x)\frac{d^2u}{dx^2}\right) - \frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u =$$

$$= \lambda\left[\rho_0(x)u - \frac{d}{dx}\left(\rho_1(x)\frac{du}{dx}\right)\right], \quad 0 < x < l, \tag{IV.17}$$

$$u(0) = u(l) = 0, \quad \left.\frac{du}{dx}\right|_{x=0} = \left.\frac{du}{dx}\right|_{x=l} = 0, \tag{IV.18}$$

где все коэффициенты ограничены и неотрицательны:

$k(x) \geqslant k_0 > 0$, $p(x) \geqslant p_0 > 0$, $q(x) \geqslant q_0 > 0$, $\rho_s(x) \geqslant \rho > 0$, $s = 1, 0$.

Нетрудно проверить, что в этих предположениях операторы

$$Au = \frac{d^2}{dx^2}\left(k\frac{d^2u}{dx^2}\right) - \frac{d}{dx}\left(p\frac{du}{dx}\right) + qu,$$

$$Bu = -\frac{d}{dx}\left(\rho_1\frac{du}{dx}\right) + \rho_0 u,$$

области определения которых $D(A)$ и $D(B)$ устанавливаются соотношениями (IV.15) и (IV.16), являются положительно определенными, причем

$$[u, u]_A = \int_0^l \left[k\left(\frac{d^2u}{dx^2}\right)^2 + p\left(\frac{du}{dx}\right)^2 + qu^2\right]dx,$$

$$[u, u]_B = \int_0^l \left[\rho_1\left(\frac{du}{dx}\right)^2 + \rho_0 u^2\right]dx. \tag{IV.19}$$

Отметим, что H_A в данном случае составляют функции, принадлежащие $\overset{0}{W}{}_2^2(0, l)$, а H_B — функции из $\overset{0}{W}{}_2^1(0, l)$.

Для доказательства существования бесконечного числа собственных значений задачи (IV.17), (IV.18), т. е. операторного уравнения

$$Au = \lambda Bu,$$

достаточно проверить выполнение условий теоремы I.9.

Итак, пусть $M \subset H_A$ — множество функций $u(x)$, ограниченных в метрике H_A:

$$\|u\|_A^2 \equiv \int_0^l \left[k\left(\frac{d^2u}{dx^2}\right)^2 + p\left(\frac{du}{dx}\right)^2 + qu^2\right] dx \leqslant c, \quad \forall u \in M \subset H_A,$$

откуда следует

$$\int_0^l \left(\frac{d^2u}{dx^2}\right)^2 dx \leqslant \frac{c}{k_0}. \tag{IV.20}$$

Как известно, для любой функции $u(x) \in W_2^2(0, l)$, удовлетворяющей условиям (IV.18), справедливы равенства

$$u(x) = \int_0^x dx \int_0^x \frac{d^2u}{dx^2}\, dx = \int_0^x (x-t)\frac{d^2u}{dt^2}\, dt,$$

$$\frac{du}{dt} = \int_0^x \frac{d^2u}{dt^2}\, dt,$$

которые можно представить в виде

$$u(x) = \int_0^l k_1(x, t)\frac{d^2u}{dt^2}\, dt, \quad k_1(x, t) = \begin{cases} x-t, & 0 \leqslant t \leqslant x, \\ 0, & x < t \leqslant l, \end{cases} \tag{IV.21}$$

$$\frac{du}{dx} = \int_0^l k(x, t)\frac{d^2u}{dt^2}\, dt, \quad k(x, t) = \begin{cases} 1, & 0 \leqslant t \leqslant x, \\ 0, & x < t \leqslant l. \end{cases} \tag{IV.22}$$

Интегральный оператор (IV.21) как оператор Фредгольма преобразует любое множество, ограниченное в $L_2(0, l)$, в множество, компактное в $L_2(0, l)$; следовательно, если $u(x) \in M$, то с учетом (IV.20) можно утверждать, что множество M компактно в $L_2(0, l)$. А это означает, что из любой бесконечной части множества M можно выделить последовательность $\{u_n(x)\}$, сходящуюся в $L_2(0, l)$:

$$u_n(x) = \int_0^l k_1(x, t)\frac{d^2u_n}{dt^2}\, dt, \quad u_n(x) \in M.$$

Если теперь к элементам этой последовательности применим оператор Фредгольма (IV.22)

$$\frac{du_n}{dx} = \int_0^l k(x, t)\frac{d^2u_n}{dt^2}\, dt,$$

то получим множество элементов $\left\{\frac{du_n}{dx}\right\}$ — множество первых производных,— компактное в пространстве $L_2(0, l)$ (см. (IV.20)). В силу компактности множества $\left\{\frac{du_n}{dx}\right\}$ из него можно выделить подпоследовательность $\left\{\frac{du_{n_1}}{dx}\right\}$, сходящуюся в пространстве $L_2(0, l)$. Таким образом, установлено, что из множества функций $u(x) \in M$, ограниченного в метрике H_A, можно выделить подпоследовательность функций $\{u_{n_1}(x)\}$, элементы которой сходятся в $L_2(0, l)$ вместе со своими первыми производными. Иными словами, из M можно выделить последовательность $\{u_{n_1}(x)\}$, сходящуюся в метрике H_B (см. (IV.19)). А это означает, что множество M, ограниченное в метрике H_A, компактно в H_B, что и требовалось показать.

Таким образом, установлено, что задача (IV.17), (IV.18) имеет бесконечную последовательность собственных чисел $0 < \lambda_1 \leqslant \lambda_2 \leqslant \leqslant \ldots \leqslant \lambda_n \leqslant \ldots$ и соответствующую им систему собственных элементов $u_1, u_2, \ldots, u_n, \ldots$, которую можно считать ортонормированной в H_B.

IV.2. Решение задач на собственные значения методом конечных элементов

Для численного решения задач на собственные значения, сформулированных в вариационной форме, с успехом применяется вариант МКЭ, основанный на процессе Ритца. Возможность этого следует из теоремы II.4, устанавливающей полноту кусочно-полиномиальных функций из подпространств P_n^h в пространстве $W_2^m(0, l)$.

Получение соответствующей дискретной задачи в данном случае осуществляется вполне аналогично дискретизации методом конечных элементов дифференциальных краевых задач с положительно определенным оператором (см. гл. II). Поэтому остановимся на данном вопросе весьма бегло, ограничиваясь рассмотрением следующей задачи с постоянными коэффициентами:

$$-k\frac{d^2u}{dx^2} + qu = \lambda u, \quad 0 < x < l, \tag{IV.23}$$

$$u(0) = u(l) = 0, \tag{IV.24}$$

где $k = \text{const} > 0$, $q = \text{const} \geqslant 0$.

Отыскание наименьшего собственного числа λ_1 и соответствующей собственной функции u_1 этой задачи сводится к отысканию минимума функционала

$$R(u) = \frac{\int_0^l \left(k\left(\frac{du}{dx}\right)^2 + qu^2\right)dx}{\int_0^l u^2 dx}, \quad u \in H_A. \tag{IV.25}$$

Напомним, что в данном случае H_A и $\overset{0}{W}{}_2^1(0, l)$ состоят из одинаковых функций, т. е. допустимые функции при минимизации функционала (IV.25) принадлежат $\overset{0}{W}{}_2^1(0, l)$. Для построения методом конечных элементов приближения $u^N(x)$ к функции, доставляющей минимум функционалу (IV.25), достаточно применить процесс Ритца, описанный в п. 5 параграфа I.2, с базисными функциями МКЭ $\{\varphi_i^N(x)\}$. Однако здесь, как и при решении соответствующих краевых задач, можно строить непосредственно допустимые функции $v^N(x)$, являющиеся кусочно-полиномиальными функциями требуемой гладкости из конечномерных подпространств $\overset{0}{P}{}_n^h \subset \overset{0}{W}{}_2^1(0, l)$. Эти функции на каждом элементе $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, N$, имеют вид

$$v^N(x) = \beta_0 + \beta_1 x + \cdots + \beta_n x^n,$$

где β_i — неизвестные числовые коэффициенты, вычисляемые через значения допустимой функции $v_j \equiv v_j^N = v^N(x_j)$ в узлах x_j элементов.

Теперь для получения дискретной задачи МКЭ достаточно применить к вспомогательной функции метода неопределенных множителей Лагранжа

$$F(v^N) = \int_0^l \left(k\left(\frac{dv^N}{dx}\right)^2 + q(v^N)^2\right) dx - \lambda \int_0^l (v^N)^2\, dx \qquad \text{(IV.26)}$$

стандартный алгоритм, подробно описанный в параграфе II.2.

Пусть, например, $v^N(x) \in \overset{0}{P}{}_1^h \subset \overset{0}{W}{}_2^1(0, l)$, т. е. допустимые функции принадлежат подпространству кусочно-линейных полиномов, удовлетворяющих условию

$$v^N(0) = v^N(l) = 0.$$

С помощью тех же рассуждений, что и в п. 1 параграфа II.2, функционал (IV.26) можно записать как функцию параметров $v_i = v^N(x_i)$, $x_i = h_i$, $i = 0, 1, \ldots, N$, $h = l/N$:

$$F(v^N) = \sum_{i=1}^{N} \omega_i^T K_i \omega_i - \lambda \sum_{i=1}^{N} \omega_i^T M_i \omega_i,$$

где $\omega_i^T = [v_{i-1}, v_i]$, $K_i \equiv K_i^1 + K_i^0$, $M_i \equiv \frac{1}{q} K_i^0$,

$$K_i^1 = \frac{k}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad K_i^0 = \frac{qh}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

Введя общий вектор

$$\omega^T = [v_0, v_1, \ldots, v_N],$$

будем иметь

$$F(v^N) = \omega^T \bar{K} \omega - \lambda \omega^T \bar{M} \omega.$$

Здесь трехдиагональные матрицы $\bar{K}$ и $\bar{M}$ строятся обычным образом из элементарных матриц K_i и M_i. Приравняв к нулю частные произ-

водные функции $F(v^N)$ по всем неизвестным параметрам $v_1, v_2, \ldots, v_{N-1}$, получим систему $(N-1)$-го порядка

$$Kv = \lambda Mv, \tag{IV.27}$$

где

$$K = kK^1 + qK^0, \quad M = K^0,$$

$$K^1 = \frac{1}{h}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & 0 & \\ & -1 & 2 & -1 & \\ & \cdots & \cdots & \cdots & \\ & & 0 & -1 & 2 \end{bmatrix}, \quad K^0 = \frac{h}{6}\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & 0 & \\ & 1 & 4 & 1 & \\ & \cdots & \cdots & \cdots & \\ & 0 & & 1 & 4 \end{bmatrix}. \tag{IV.28}$$

Таким образом, дискретизация методом конечных элементов вариационной задачи о минимуме функционала (IV.25) (или, что эквивалентно, исходной задачи на собственные значения (IV.23), (IV.24)) привела к решению алгебраической обобщенной задачи на собственные значения (IV.27). Решив обобщенную задачу (IV.27), найдем приближения λ_i^N к нескольким первым собственным числам λ_i исходной дифференциальной задачи (IV.23), (IV.24), а векторы $v^{(i)} = [v_1^{(i)}, \ldots, v_{N-1}^{(i)}]^T$, отвечающие λ_i^N, дадут значения соответствующих приближенных собственных функций $u_i^N(x)$ во внутренних узлах x_k, $k = 1, 2, \ldots, N-1$, интервала $(0, l)$. Приближение к собственной функции $u_i(x)$, отвечающей собственному числу λ_i исходной дифференциальной задачи, можно получить по формуле

$$u_i^N(x) = \sum_{k=1}^{N-1} v_k^{(i)} \varphi_k(x).$$

Замечание 1. Аналогично выполняется решение задачи (IV.23), (IV.24) и при других краевых условиях, например при

$$\left.\frac{du}{dx}\right|_{x=0} = 0, \quad u(l) = 0, \tag{IV.29}$$

а также в случае переменных коэффициентов. Отметим только, что функции из энергетического пространства H_A, в отличие от функций из $D(A)$, должны удовлетворять только главным краевым условиям и не обязательно должны удовлетворять естественным. Поэтому и допустимые функции $v^N(x) \in P_1^h \subset H_A$ можно не подчинять естественным условиям, например условию $\left.\frac{du}{dx}\right|_{x=0} = 0$ в случае (IV.29).

Замечание 2. Аналогично, следуя методике параграфа II.2, можно получить дискретные задачи и при кусочно-полиномиальных допустимых функциях более высоких степеней ($n > 1$). Отметим, что алгебраическая задача на собственные значения всегда будет обобщенной, а матрицы K и M — хоть и ленточные, но с повышением степени полиномов n ширина их ленты будет расти.

Для иллюстрации изложенного выше приведем результаты отыскания методом конечных элементов нескольких первых собственных чисел задачи (IV.23), (IV.29) при $k = 1$, $q = 0$, $l = 1$.

Как известно, точные собственные числа этой задачи имеют вид

$$\lambda_i = \frac{\pi^2}{4}(2i-1)^2, \quad i = 1, 2, \ldots$$

Для получения их приближенных значений разобьем отрезок [0, 1] на N равных элементов. Используем вначале в качестве допустимых функций кусочно-квадратичные полиномы, т. е. элемент вида «l 2—3». Алгебраическая задача на собственные значения

$$Ku = \lambda Mu \qquad \text{(IV.30)}$$

в этом случае (см. п. 2 параграфа II.2) имеет порядок $2N$.

В (IV.30) K и M — матрицы следующего вида

$$K = \frac{1}{3h}\begin{bmatrix} 7 & -8 & 1 & & & & & 0 \\ -8 & 16 & -8 & & & & & \\ 1 & -8 & 14 & -8 & 1 & & & \\ & & -8 & 16 & -8 & & & \\ & & 1 & -8 & 14 & -8 & 1 & \\ \multicolumn{8}{c}{\dotfill} \\ & & & & 1 & -8 & 14 & -8 \\ & & 0 & & & & -8 & 16 \end{bmatrix},$$

$$M = \frac{h}{30}\begin{bmatrix} 4 & 2 & -1 & & & & & 0 \\ 2 & 16 & 2 & & & & & \\ -1 & 2 & 8 & 2 & -1 & & & \\ & & 2 & 16 & 2 & & & \\ & & -1 & 2 & 8 & 2 & -1 & \\ \multicolumn{8}{c}{\dotfill} \\ & & & & -1 & 2 & 8 & 2 \\ & & 0 & & & & 2 & 16 \end{bmatrix}, \quad h = \frac{1}{N}.$$

Решение алгебраической обобщенной проблемы собственных значений удобно и целесообразно выполнять следующим образом. Вначале методом квадратных корней строится треугольное разложение положительно определенной матрицы $M = LL^T$, которое затем используется для приведения задачи (IV.30) к виду

$$Cy = \lambda y, \qquad \text{(IV.31)}$$

где

$$C = L^{-1}KL^{-T}, \quad y = L^T u.$$

Собственные числа задачи (IV.31) находят в результате выполнения следующих двух вычислительных процессов: приведения матрицы C

Таблица 11

N	m	λ_1^N	δ_1 (%)	λ_2^N	δ_2 (%)	λ_3^N	δ_3 (%)
2	4	2,468656	0,05100	22,94625	3,33	77,06312	24,90
3	6	2,467659	0,01100	22,37364	0,75	64,59863	4,72
4	8	2,467482	0,00330	22,26209	0,25	62,75267	1,75
6	12	2,467418	0,00079	22,21801	0,05	61,91636	0,38

к трехдиагональному виду посредством подобных элементарных преобразований с помощью матриц отражения [106]; вычисления методом бисекций [106] искомых собственных чисел полученной трехдиагональной матрицы.

Отыскание трех первых приближенных собственных чисел задачи (IV.23), (IV.29) (посредством дискретной задачи (IV.30)) осуществлялось на ЭВМ МИР-2 при разрядности 12. Полученные результаты представлены в табл. 11, где использованы следующие обозначения: N — число элементов, на которое разбит отрезок [0, 1], m — число уравнений алгебраической (дискретной) задачи, λ_i^N — вычисленное значение i-го приближенного собственного числа задачи (IV.23), (IV.29) (дано семь значащих цифр), δ_i (%) — относительная погрешность в процентах приближенного собственного числа, определяемая по формуле

$$\delta_i(\%) = \frac{|\lambda_i - \lambda_i^N|}{\lambda_i} \cdot 100\ \%, \quad \lambda_i \text{ — точное собственное число.}$$

Аналогично выполняется решение поставленной задачи при использовании в качестве допустимых функций кусочно-кубических полиномов Эрмита (элемент «l 3—2»).

Для построения матриц K и M алгебраической обобщенной задачи (IV.30) в этом случае использовались элементарные матрицы жесткости и масс, выписанные в п. 3 параграфа II.2 (для постоянных коэффициентов); при этом для упрощения алгебраической системы естественное краевое условие тоже учитывалось. Полученные результаты представлены в табл. 12, где обозначения такие же, как в табл. 11 (вычисленные значения λ_i^N даны с 9—10 значащими цифрами).

Таблица 12

N	m	λ_1^N	δ_1 (%)	λ_2^N	δ_2 (%)	λ_3^N	δ_3 (%)
2	4	2,46741694	$6{,}4\cdot10^{-4}$	22,2482195	0,1900	63,0464209	2,21
3	6	2,46740259	$6{,}0\cdot10^{-5}$	22,2123870	0,0260	61,8723825	0,30
4	8	2,46740134	$9{,}6\cdot10^{-6}$	22,2079141	0,0060	61,7332082	0,08
6	12	2,46740119	$3{,}6\cdot10^{-6}$	22,2067532	0,0007	61,6913401	0,01

Следует подчеркнуть, что при использовании кубических полиномов Эрмита среди вычисленных собственных чисел обобщенной задачи (IV.30) наряду с действительно приближенными собственными числами исходной дифференциальной задачи существуют и такие, которые не приближают никакое собственное число дифференциальной задачи. Приближенных собственных чисел приемлемой точности будет не больше общего количества узловых параметров, являющихся искомыми значениями допустимой функции (в случае задачи (IV.23), (IV.29) — не больше N). Надежно выделить эти, «истинно» приближенные, собственные числа можно лишь по асимптотике их поведения в результате расчетов на нескольких сетках.

Отметим вместе с тем, что и при использовании лагранжевых полиномов, где среди узловых параметров нет значений производных искомой функции, не все вычисленные собственные числа будут хорошо приближать искомые, а только несколько первых. Измельчение сетки позволяет повышать точность первых и увеличивать количество собственных чисел, имеющих приемлемую точность. Эти замечания хорошо иллюстрируются данными табл. 11, 12.

В заключение на основе рассмотрения конкретного примера сделаем несколько замечаний о дискретизации задач на собственные значения для дифференциальных уравнений четвертого порядка. Для простоты, но без потери общности рассуждений, возьмем одну из простейших задач:

$$\frac{d^4u}{dx^4} + q(x)u = \lambda u, \quad 0 < x < 1, \tag{IV.32}$$

$$u(0) = \frac{du}{dx}(0) = 0, \quad \left.\frac{d^2u}{dx^2}\right|_{x=1} = \left.\frac{d^3u}{dx^3}\right|_{x=1} = 0. \tag{IV.33}$$

Наименьшее собственное число и отвечающая ему собственная функция данной задачи есть минимум функционала

$$R(u) = \frac{\int_0^1 \left(\left(\frac{d^2u}{dx^2}\right)^2 + q(x)u^2\right)dx}{\int_0^1 u^2 dx}, \quad u \in H_A,$$

и соответственно функция, доставляющая этот минимум.

Поскольку в данном случае H_A состоит из функций, принадлежащих $W_2^2(0, 1)$, то для построения приближенных решений методом конечных элементов в качестве допустимых следует выбирать кусочно-полиномиальные функции, непрерывные на [0, 1] вместе со своими первыми производными. Это могут быть, в частности, кусочные полиномы Эрмита не ниже третьей степени.

При использовании полиномов третьей степени (элемент вида «l 2—3») по описанным ранее алгоритмам получим соответствующую дискретную задачу вида (IV.30), где при $q \equiv 0$ и равномерной сетке с шагом

$h = 1/N$

$$K = \frac{2}{h^3}\begin{bmatrix}
12 & 0 & -6 & 3h & & & & & & & & \\
0 & 4h^2 & -3h & h^2 & & & & & & & & \\
-6 & -3h & 12 & 0 & -6 & 3h & & & & & & \\
3h & h^2 & 0 & 4h^2 & -3h & h^2 & & & & & & \\
 & & -6 & -3h & 12 & 0 & -6 & 3h & & & & \\
 & & 3h & h^2 & 0 & 4h^2 & -3h & h^2 & & & & \\
 & & & & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & \\
 & & & & -6 & -3h & 12 & 0 & -6 & 3h & & \\
 & & & & 3h & h^2 & 0 & 4h^2 & -3h & h^2 & & \\
 & & & & & & -6 & -3h & 6 & -3h & & \\
 & & & & & & 3h & h^2 & -3h & 2h^2 & &
\end{bmatrix},$$

$$M = \frac{h}{420}\begin{bmatrix}
312 & 0 & 54 & -13h & & & & & & & & \\
0 & 8h^2 & 13h & -3h^2 & & & & & & & & \\
54 & 13h & 312 & 0 & 54 & -13h & & & & & & \\
-13h & -3h^2 & 0 & 8h^2 & 13h & -3h^2 & & & & & & \\
 & & 54 & 13h & 312 & 0 & 54 & -13h & & & & \\
 & & -13h & -3h^2 & 0 & 8h^2 & 13h & -3h^2 & & & & \\
 & & & & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & \\
 & & & & 54 & 13h & 312 & 0 & 54 & -13h & & \\
 & & & & -13h & -3h^2 & 0 & 8h^2 & 13h & -3h^2 & & \\
 & & & & & & 54 & 13h & 156 & -22h & & \\
 & & & & & & -13h & -3h^2 & -22h & 4h^2 & &
\end{bmatrix}.$$

Здесь при построении матриц K и M учитывались только главные условия $u(0) = \frac{du}{dx}(0) = 0$.

Таблица 13

N	m	λ_1^N	δ_1 (%)	λ_2^N	δ_2 (%)	λ_3^N	δ_3 (%)
				«l 3—2»			
3	6	12,36488431	0,020	488,71193	0,66	3902,00	2,51
4	8	12,36312959	$0{,}62 \cdot 10^{-2}$	486,65185	0,23	3865,72	1,55
6	12	12,36254202	$0{,}14 \cdot 10^{-2}$	485,75865	0,049	3820,49	0,36
8	16	12,36240239	$0{,}31 \cdot 10^{-3}$	485,59716	0,016	3811,18	0,12
				«l 4—3»			
2	6	12,36246862	$0{,}85 \cdot 10^{-3}$	485,82775	0,27	3846,37	1,05
3	9	12,36237286	$0{,}77 \cdot 10^{-4}$	485,62155	0,021	3824,05	0,46
4	12	12,36236497	$0{,}13 \cdot 10^{-4}$	485,53706	$0{,}38 \cdot 10^{-2}$	3809,20	0,070
				«l 5—2»			
2	6	12,36236366	$0{,}24 \cdot 10^{-5}$	485,52918	$0{,}21 \cdot 10^{-2}$	3828,08	0,57

При дискретизации задач на собственные значения для уравнения четвертого порядка можно применять также элементы «l 4—3» вида v, v' ●—●—● v, v', «l 5—2» вида v, v', v'' ●—● $v,\ v', v''$ и другие, обеспечивающие непрерывность первой производной допустимой функции. В табл. 13 представлены приближенные значения трех первых собственных чисел задачи (IV.32), (IV.33) при $q = 0$, полученные посредством элементов «l 3—2», «l 4—3», и «l 5—2» на равномерных сетках.

Вычисления выполнялись на ЭВМ МИР-2 при разрядности 12. Алгебраическая обобщенная задача на собственные значения решалась описанным ранее способом. Точные значения собственных чисел задачи (IV.32), (IV.33) есть $\lambda = k^4$, где k — корни уравнения $\cos k \operatorname{ch} k + 1 = 0$. С точностью до десяти значащих цифр $\lambda_1 = 12{,}362\ 363\ 37$, $\lambda_2 = 485{,}518\ 818\ 5$, $\lambda_3 = 3806{,}546\ 266$. Обозначения в табл. 13 такие же, как в табл. 11, 12. Значения λ_1^N даны с десятью значащими цифрами, λ_2^N — с восемью, а λ_3^N — с шестью.

IV.3. Оценки погрешности для собственных чисел и собственных функций

Изложение данного вопроса ведется в соответствии с работой [101]. Остановимся подробно на случае задачи

$$Au = \lambda u,$$

где положительно определенный в H оператор A удовлетворяет условиям теоремы 1.7, т. е. имеет бесконечную последовательность собственных чисел $0 < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n \leqslant \ldots$ с единственной предельной точкой на бесконечности и соответствующие им ортонормированные в H собственные функции u_k, $k = 1, 2, \ldots$, образуют систему, полную в H и в H_A.

В рамках рассмотрений данной главы будем предполагать, что оператор A действует в пространстве $H = L_2(0, l)$ по формуле

$$Au = \sum_{\alpha=0}^{m} (-1)^\alpha \frac{d^\alpha}{dx^\alpha}\left(a_\alpha(x)\frac{d^\alpha u}{dx^\alpha}\right), \quad m = 1 \text{ или } m = 2.$$

При любых однородных краевых условиях норма в энергетическом пространстве H_A данного оператора оценивается сверху нормой в пространстве $W_2^m(0, l)$:

$$\|u\|_A^2 \leqslant C \int_0^l \sum_{\alpha=0}^{m} \left(\frac{d^\alpha u}{dx^\alpha}\right)^2 dx = C\|u\|_{2,m}^2, \quad C = \text{const}. \qquad \text{(IV.34)}$$

Как указано ранее, собственными функциями u_k являются функции, доставляющие минимум отношению

$$R(u) = \frac{[u, u]_A}{(u, u)}, \qquad \text{(IV.35)}$$

на подпространстве $H_A^{(k-1)}$ пространства H_A, ортогональном в метрике

H_A к $u_1, u_2, \ldots, u_{k-1}$, т. е. $H_A^{(k-1)}$ состоит из элементов $u \in H_A$, удовлетворяющих условиям

$$[u, u_i]_A = 0, \quad i = 1, 2, \ldots, k-1, \text{ если } k \geqslant 2.$$

Собственные числа λ_k — значения этого минимума,

$$\lambda_k = \min_{u \in H_A^{(k-1)}} \frac{[u, u]_A}{(u, u)} \equiv \frac{[u_k, u_k]_A}{(u_k, u_k)}.$$

Приближенные собственные числа λ_k^N и собственные функции u_k^N определяются по методу конечных элементов путем минимизации отношения (IV.35) на конечномерных подпространствах $P_n^h \subset H_A$, т. е.

$$\lambda_k^N \equiv \frac{[u_k^N, u_k^N]_A}{(u_k^N, u_k^N)} = \min_{u^N \in P_n^h} \frac{[u^N, u^N]_A}{(u^N, u^N)}, \quad [u^N, u_i^N]_A = 0,$$

$$i = 1, 2, \ldots, k-1,$$

или, что то же,

$$\lambda_k^N = \min_{u^N \in P_n^h} [u^N, u^N]_A = [u_k^N, u_k^N]_A,$$

при условиях

$$(u^N, u^N) = 1, \quad [u^N, u_i^N]_A = 0, \quad i = 1, 2, \ldots, k-1.$$

Как показано в параграфе IV.2, приближенная собственная функция u_i^N оператора A представляется в виде разложения по базисным функциям МКЭ:

$$u_i^N = \sum_{j=1}^{r} u_{ij}^N \varphi_j, \tag{IV.36}$$

где r — размерность P_n^h,
числовые коэффициенты u_{ij}^N являются компонентами собственного вектора, отвечающего собственному числу λ_i^N обобщенной алгебраической задачи

$$Kv_i^N - \lambda_i^N M v_i^N = 0, \quad v_i^N = [u_{i1}^N, u_{i2}^N, \ldots, u_{ir}^N]^T \tag{IV.37}$$

(r — порядок матриц K, M).

Несколько первых собственных чисел λ_i^N алгебраической задачи (IV.37) являются приближенными собственными числами оператора A (см. параграф IV.2).

Для любых собственных векторов v_i^N, v_l^N задачи (IV.37) справедливы соотношения ортогональности

$$(v_l^N)^T M v_i^N = \delta_{il}, \quad (v_l^N)^T K v_i^N = \lambda_i^N \delta_{il}.$$

А для приближенных собственных функций (IV.36) соответственно имеем

$$(u_i^N, u_l^N) = \delta_{il}, \quad [u_i^N, u_l^N]_A = \lambda_i^N (u_i^N, u_l^N) = \lambda_i^N \delta_{il}. \tag{IV.38}$$

Для приближенных собственных чисел λ_k^N справедлив и принцип минимакса:

$$\lambda_k^N = \min_{P^{(k)} \subset P_n^h} \max_{v \in P^{(k)}} R(v), \qquad \text{(IV.39)}$$

где $P^{(k)}$ — любое k-мерное подпространство конечномерного пространства P_n^h, естественно, при условии, что k не больше размерности пространства P_n^h.

Сравнение (IV.39) с (I.29) позволяет утверждать, что для всех k выполняется неравенство

$$\lambda_k^N \geqslant \lambda_k.$$

Действительно, $P^{(k)}$ является частным случаем подпространств S_k, используемых в принципе минимакса; но в (IV.39) $P^{(k)} \subset P_n^h$, а в соотношении (I.29) S_k — *любое* k-мерное подпространство из H_A, т. е. область изменения S_k шире P_n^h.

Для получения оценок погрешностей $\lambda_k^N - \lambda_k$ собственных чисел задачи $Au - \lambda u = 0$ введем оператор проектирования P такой, что $Pu \in P_n^h$, если $u \in H_A$ (т. е. Pu — это составляющая функции u в подпространстве P_n^h), причем

$$[u - Pu,\ w^N]_A = 0, \quad \forall\, w^N \in P_n^h.$$

Как уже отмечалось (см. (III.28)), это означает, что

$$[u - Pu,\ u - Pu]_A = \min_{u^N \in P_n^h} [u - u^N,\ u - u^N]_A.$$

Таким образом, если $u\,(x) \in H_A$, где A — оператор, описанный выше, то элемент Pu можно рассматривать как приближенное решение МКЭ (вариант Ритца) некоторой краевой задачи с оператором A. Согласно теореме II.5 это гарантирует справедливость неравенства

$$\|u - Pu\|_{2,s} \leqslant C_1 h^{n+1-s} \|u\|_{2,n+1}, \quad 0 \leqslant s \leqslant m. \qquad \text{(IV.40)}$$

Далее, для оценки $\lambda_k^N - \lambda_k$ введем еще подпространство E_k, натянутое на точные собственные функции $u_1, u_2, \ldots, u_k$ оператора A, а также множество e_k единичных векторов E_k и обозначим

$$\sigma_k^N = \max_{u \in e_k} |\,2(u,\ u - Pu) - (u - Pu,\ u - Pu)\,|. \qquad \text{(IV.41)}$$

В работе [101] показано, что при выполнении условия $\sigma_k^N < 1$ приближенные собственные значения λ_k^N ограничены и сверху:

$$\lambda_k \leqslant \lambda_k^N \leqslant \frac{\lambda_k}{1 - \sigma_k^N}.$$

Оценим теперь σ_k^N исходя из соотношения (IV.41). При этом используем тождество

$$(u,\ u-Pu)=\sum_{i=1}^{k}\frac{c_i}{\lambda_i}[u_i-Pu_i,\ u-Pu]_A,$$

справедливость которого в случае $u=\sum_{i=1}^{k}c_iu_i\in e_k$ доказана в [101].

Для первого слагаемого σ_k^N согласно (IV.41) и с учетом (IV.34), (IV.40) имеем

$$|2(u,\ u-Pu)|=2|[w_k-Pw_k,\ u-Pu]_A|\leqslant$$
$$\leqslant 2\|w_k-Pw_k\|_A\|u-Pu\|_A\leqslant 2C\|w_k-Pw_k\|_{2,m}\|u-Pu\|_{2,m}\leqslant$$
$$\leqslant 2CC_1^2h^{2(n+1-m)}\|w_k\|_{2,n+1}\|u\|_{2,n+1}=Rh^{2(n+1-m)}\|u\|_{2,n+1},\qquad\text{(IV.42)}$$

где $w_k=\sum_{i=1}^{k}\frac{c_i}{\lambda_i}u_i$, а R — постоянная, не зависящая от h.

Второй член в σ_k^N (см. (IV.41)) оценивается так:

$$(u-Pu,\ u-Pu)=\|u-Pu\|^2\leqslant C_1^2h^{2(n+1)}\|u\|_{2,n+1},$$

т. е. он имеет более высокий порядок относительно h, чем (IV.42). Поэтому окончательную оценку σ_k^N можно представить в виде

$$\sigma_k^N<k_1h^{2(n+1-m)}\|u\|_{2,n+1},\qquad\text{(IV.43)}$$

где k_1 — постоянная, не зависящая от h.

Если взять h настолько малым, чтобы обеспечить $\sigma_k^N\leqslant\frac{1}{2}$, то согласно (IV.42) и (IV.43) будем иметь

$$\lambda_k\leqslant\lambda_k^N\leqslant\frac{\lambda_k}{1-\sigma_k^N}\leqslant\lambda_k(1+2\sigma_k^N)\leqslant\lambda_k+2k_1h^{2(n+1-m)}\|u\|_{2,n+1}.$$

Итак, приведенные результаты можно сформулировать в виде следующей теоремы.

Теорема IV.1. *Если $P_n^h\subset W_2^m(0,\ l)$ образовано кусочными полиномами степени n, а собственные функции оператора A принадлежат $W_2^{n+1}(0,\ l)$, то существует такая постоянная k_1, что при достаточно малых значениях h будут справедливы для приближенных собственных чисел λ_k^N оценки*

$$\lambda_k\leqslant\lambda_k^N\leqslant\lambda_k+2k_1h^{2(n+1-m)}\|u\|_{2,n+1},\quad 1\leqslant k<r,\qquad\text{(IV.44)}$$

или

$$\lambda_k^N-\lambda_k=O(h^{2(n+1-m)}).$$

Отметим, что в работе [101] выполнено более подробное исследование погрешности приближенных собственных чисел, чем приведенное здесь, и получена оценка

$$\lambda_k\leqslant\lambda_k^N\leqslant 2k_2\lambda_k^{\frac{n+1}{m}}h^{2(n+1-m)}.$$

Эта оценка показывает, что точность приближенных собственных чисел с большим номером существенно ниже, чем первых чисел. И это положение хорошо подтверждается практическими вычислениями.

Приведем теперь результаты о точности приближенных собственных функций u_k^N. При этом будут использоваться следующие два тождества [101]:

$$[u_k - u_k^N,\ u_k - u_k^N]_A = \lambda_k \| u_k - u_k^N \|^2 + \lambda_k^N - \lambda_k, \qquad \text{(IV.45)}$$

при нормировке

$$(u_k, u_k) = (u_k^N, u_k^N) = 1;$$

$$(\lambda_j^N - \lambda_k)(Pu_k, u_j^N) = \lambda_k (u_k - Pu_k, u_j^N) \quad \text{при } \forall\, j,\ k. \qquad \text{(IV.46)}$$

Если λ_k — изолированное собственное число (не кратное), то согласно (IV.44) найдется такая постоянная ρ, что для малых значений h будем иметь

$$\frac{\lambda_k}{|\lambda_j^N - \lambda_k|} \leqslant \rho < \infty \quad \text{для любого } j. \qquad \text{(IV.47)}$$

Множество приближенных собственных функций $u_1^N,\ u_2^N,\ \ldots,\ u_s^N$ образует ортонормированный базис в пространстве P_n^h (см. (IV.38)), поэтому

$$Pu_k = \sum_{j=1}^{r} (Pu_k, u_j^N)\, u_j^N. \qquad \text{(IV.48)}$$

Обозначим через β коэффициент (Pu_k, u_k^N) и оценим остальные члены в (IV.48).

Учитывая тождество (IV.46) и неравенство (IV.47), получаем

$$\| Pu_k - \beta u_k^N \|^2 = \sum_{j=1}^{r}{}' (Pu_k, u_j^N)^2 = \sum_{j=1}^{r}{}' \left(\frac{\lambda_k}{\lambda_j^N - \lambda_k} \right)^2 (u_k - Pu_k, u_j^N)^2 \leqslant$$

$$\leqslant \rho^2 \sum_{j=1}^{r}{}' (u_k - Pu_k, u_j^N)^2 \leqslant \rho_1^2 \| u_k - Pu_k \|^2,$$

где $\sum_{j=1}^{r}{}'$ — сумма, в которой отсутствует слагаемое с $j = k$. Теперь можно написать

$$\| u_k - \beta u_k^N \| \leqslant \| u_k - Pu_k \| + \| Pu_k - \beta u_k^N \| \leqslant (1 + \rho_1) \| u_k - Pu_k \|,$$

а с учетом (IV.40)

$$\| u_k - \beta u_k^N \| \leqslant C_2 h^{n+1} \| u_k \|_{2,n+1}. \qquad \text{(IV.49)}$$

Применив неравенство треугольника, получим два соотношения:

$$\| u_k - u_k^N \| \leqslant \| u_k - \beta u_k^N \| + \| (\beta - 1) u_k^N \|, \qquad \text{(IV.50)}$$

$$\| u_k \| - \| u_k - \beta u_k^N \| \leqslant \| \beta u_k^N \| \leqslant \| u_k \| + \| u_k - \beta u_k^N \|. \qquad \text{(IV.51)}$$

Если учесть нормированность векторов u_k и u_k^N и выбрать их знаки так, чтобы коэффициент β был неотрицательным, $\beta \geqslant 0$, то соотношение (IV.51) можно переписать в виде

$$1 - \| u_k - \beta u_k^N \| \leqslant \beta \leqslant 1 + \| u_k - \beta u_k^N \|,$$

что равносильно неравенству

$$|\beta - 1| \leqslant \| u_k - \beta u_k^N \|.$$

Поэтому согласно (IV.49) и (IV.50)

$$\| u_k - u_k^N \| \leqslant 2 \| u_k - \beta u_k^N \| \leqslant C_3 h^{n+1} \| u_k \|_{2,n+1}. \qquad \text{(IV.52)}$$

Если использовать тождество (IV.45), то с учетом (IV.46) и (IV.52) можно получить

$$\| u_k - u_k^N \|_A^2 \leqslant \lambda_k C_3^2 h^{2(n+1)} \| u_k \|_{2,n+1}^2 + 2k_1 h^{2(n+1-m)} \| u \|_{2,n+1} \leqslant$$
$$\leqslant k h^{2(n+1-m)} \| u \|_{2,n+1},$$

где постоянная k не зависит от h.

Таким образом, приведенные выше рассмотрения доказывают справедливость следующего утверждения.

Теорема IV.2. *Если выполняются условия теоремы IV.1 и λ_k — изолированное не кратное собственное число, то при малых значениях h для приближенных по Ритцу собственных функций u_k^N положительно определенного оператора A, порождаемого дифференциальным уравнением $2m$-го порядка, оценки погрешности имеют вид*

$$\| u_k - u_k^N \| \leqslant C h^{n+1}, \quad \| u_k - u_k^N \|_A^2 \leqslant C_1 h^{2(n+1-m)},$$

где C, C_1 — постоянные, не зависящие от h.

Вновь отметим, что оценки, полученные в работе [101], отражают зависимость точности приближенных собственных функций от номера k, а именно

$$\| u_k - u_k^N \| \leqslant C h^{n+1} \lambda_k^{\frac{n+1}{2m}}, \quad \| u_k - u_k^N \|_A^2 \leqslant C_1 h^{2(n+1-m)} \lambda_k^{\frac{n+1}{m}}.$$

Как утверждается в работе [101], для случая кратных собственных чисел справедливы аналогичные оценки. Этот случай подробно рассматривается и в работе [67]. Аналогично исследуются приближенные решения и обобщенной задачи на собственные значения $Au - \lambda Bu = 0$.

РЕШЕНИЕ НЕКОТОРЫХ КЛАССОВ НЕЛИНЕЙНЫХ ЗАДАЧ МЕТОДОМ КОНЕЧНЫХ ЭЛЕМЕНТОВ

В данной главе рассматривается использование МКЭ при решении нелинейных краевых задач для эллиптических уравнений второго порядка и некоторых нелинейных вариационных задач (без перехода к дифференциальным уравнениям). Термин «нелинейные вариационные задачи» означает, что дифференциальное уравнение Эйлера, соответствующее функционалу данной задачи, будет нелинейным; иными словами, рассматриваются вариационные задачи с функциональными, отличными от квадратичных. Как и в предыдущих главах, здесь в основном изложение будет касаться функций одной переменной $u(x)$, $x \in [a, b]$. Однако представленные результаты аналогичны и для случая функций многих переменных.

V.1. Нелинейные краевые задачи

Будем рассматривать решение краевых задач для уравнений вида

$$L(u) \equiv \frac{d}{dx} a_1\left(x, u, \frac{du}{dx}\right) + a_0\left(x, u, \frac{du}{dx}\right) = 0, \quad a < x < b, \tag{V.1}$$

где

$$\nu(|u|) \leqslant \frac{\partial a_1(x, u, p)}{\partial p} \leqslant \mu(|u|), \quad p = \frac{du}{dx},$$

$\nu(t)$ — положительная невозрастающая непрерывная функция, определенная при $t \geqslant 0$, $\mu(t)$ — положительная неубывающая непрерывная функция, определенная при $t \geqslant 0$.

Ограничимся задачей Дирихле, т. е. решением уравнения (V.1) при краевых условиях

$$u(a) = u(b) = 0. \tag{V.2}$$

1. Обобщенное решение задачи. Определим понятие искомого решения задачи (V.1), (V.2). Функцию $u(x) \in \overset{0}{W}{}_2^1(l)$ называют обобщенным решением задачи (V.1), (V.2), если она удовлетворяет интегральному тождеству

$$B(u, \eta) \equiv \int_a^b \left[a_1\left(x, u, \frac{du}{dx}\right)\frac{d\eta}{dx} - a_0\left(x, u, \frac{du}{dx}\right)\eta\right] dx = 0 \tag{V.3}$$

при всех $\eta(x) \in \overset{0}{W}{}_2^1(l)$, $l \equiv (a, b)$.

Укажем для функций $a_i(x, u, p)$, $i = 1, 0$, ограничения, гарантирующие существование и единственность данного обобщенного решения $u(x)$. При этом не будем стремиться к максимальному ослаблению требований (соответствующие ограничения см. в работе [56]).

Пусть выполняются следующие условия:

1) функции $a_i(x, u, p)$, $i = 1, 0$, непрерывны по $x \in [a, b]$ и по (u, p) при любых действительных значениях;

2) функции $a_i(x, u, p)$, $i = 1, 0$, удовлетворяют неравенствам

$$|a_1(x, u, p)| \leqslant \mu_1(u)(|p| + \varphi_1(x)), \quad \varphi_1 \in L_2(l), \tag{V.4}$$
$$|a_0(x, u, p)| \leqslant \mu_2(u)(|p|^2 + \varphi_2(x)), \quad \varphi_2 \in L_1(l),$$

где $\mu_i(u)$ — непрерывные функции u;

3) выполнено условие коэрцитивности, т. е. для любых $u(x) \in \overset{0}{W}{}_2^1(l)$ справедливо неравенство

$$B(u, u) \geqslant f(\|u\|_{2,1}) - c_1, \tag{V.5}$$

где $c_1 > 0$, $f(\tau)$ — непрерывная положительная функция, стремящаяся к бесконечности при $\tau \to \infty$;

4) для любых не равных тождественно друг другу элементов $u, v \in \overset{0}{W}{}_2^1(l)$ справедливо неравенство

$$B(u, u - v) - B(v, u - v) > 0, \tag{V.6}$$

т. е. выполняется условие строгой монотонности по u квазибилинейной формы $B(u, \eta)$.

Отметим, что при выполнении условия (V.4) тождество (V.3) имеет смысл, т. е. образующие его интегралы конечны при любых функциях $u(x)$ и $\eta(x)$ из пространства $\overset{0}{W}{}_2^1(l)$.

Из неравенства (V.5) следует (см. [56]), что для всех возможных решений задачи (V.1), (V.2) будет справедлива оценка $\|u\|_{2,1} \leqslant c_2$.

Если вместо условия 1) предположить, что функции $a_i(x, u, p)$, $i = 1, 0$, дифференцируемы по u и p, то условие (V.6), которое в развернутой записи имеет вид

$$\int_a^b \left\{\left[a_1\left(x, u, \frac{du}{dx}\right) - a_1\left(x, v, \frac{dv}{dx}\right)\right]\left(\frac{du}{dx} - \frac{dv}{dx}\right) - \right.$$
$$\left. - \left[a_0\left(x, u, \frac{du}{dx}\right) - a_0\left(x, v, \frac{dv}{dx}\right)\right](u - v)\right\} dx > 0,$$

будет следовать из выполнения для произвольных u, p, ξ_1, ξ_0 неравенства

$$\frac{\partial a_1(x, u, p)}{\partial p}\xi_1^2 + \left(\frac{\partial a_1}{\partial u} - \frac{\partial a_0}{\partial p}\right)\xi_1\xi_0 - \frac{\partial a_0}{\partial u}\xi_0^2 \geqslant$$
$$\geqslant \nu_1(x, u, p)\xi_1^2 + \nu_2(x, u, p)\xi_0^2,$$

где $\nu_i(x, u, p)$ — непрерывные неотрицательные функции, причем $\nu_1(x, u, p) > 0$ при значениях x, принадлежащих некоторым «подотрезкам» $\bar{l}_1$ отрезка $\bar{l} \equiv [a, b]$, а $\nu_2(x, u, p) > 0$ при $x \in \overline{l - l_1}$.

Как показано в работе [56], условия 1) — 4) гарантируют существование и единственность обобщенного решения $u(x) \in \overset{0}{W}{}_2^1(l)$ задачи (V.1), (V.2). Сделано это с использованием теории монотонных операторов и метода Бубнова — Галеркина. Схема этого метода в данном случае следующая.

Пусть $\{\psi_k(x)\}$ — фундаментальная система линейно независимых функций в пространстве $\overset{0}{W}{}_2^1(l)$, P_N — конечномерное подпространство пространства $\overset{0}{W}{}_2^1(l)$ с базисом $\{\psi_k(x)\}_1^N$.

Приближением по методу Бубнова — Галеркина обобщенного решения $u(x)$ называют функцию $u^N(x) \in P_N$ вида

$$u^N = \sum_{k=1}^{N} c_k \psi_k(x),$$

если она удовлетворяет соотношению

$$B(u^N, \psi_k) = 0, \quad k = 1, 2, \ldots, N. \tag{V.7}$$

Соотношение (V.7) представляет собой систему N нелинейных уравнений относительно N неизвестных c_k.

На основании определения подпространства $P_N \subset \overset{0}{W}{}_2^1(l)$ можно показать, что из системы (V.7) следует справедливость тождества

$$B(u^N, \eta) = 0 \text{ при } \forall\, \eta \in P_N. \tag{V.8}$$

Выражение $B(u^N, \eta)$ является ограниченным линейным функционалом в P_N над η при любом фиксированном элементе $u^N \in P_N$, так как для него из предположений 1), 2) следует оценка

$$|B(u^N, \eta)| \leqslant k \|\eta\|_{2,1},$$

где k — константа, зависящая от u^N.

Согласно теореме Рисса об общем виде ограниченного линейного функционала в гильбертовом пространстве (см. п. 1 параграфа I.2), справедливо представление

$$B(u^N, \eta) = (g, \eta)_{P_N}, \tag{V.9}$$

где $(\cdot, \cdot)_{P_N}$ — скалярное произведение в гильбертовом пространстве $P_N \subset \overset{0}{W}{}_2^1$, $g \equiv A(u^N)$ — некоторый элемент пространства P_N, однозначно определяемый по u^N.

Таким образом, соотношение (V.9) определяет некоторый оператор A, действующий в пространстве P_N. Доказано (см. [56]), что оператор A, определенный на всем пространстве P_N, непрерывен и система (V.7) эквивалентна (см. (V.8), (V.9)) тождеству

$$(Au^N, \eta) = 0 \text{ при } \forall\, \eta \in P_N \tag{V.10}$$

или операторному уравнению $Au^N = 0$ в гильбертовом пространстве $P_N \in \overset{0}{W}{}_2^1(l)$.

Предположения 3), 4), обеспечивающие ограниченность всех возможных обобщенных решений (и их приближений) и строгую монотонность оператора A, позволяют доказать, что система (V.7), определяющая галеркинское приближение $u^N(x)$, однозначно разрешима при любом значении N, а также, что существует единственный элемент $u(x) \in \overset{0}{W}{}_2^1(l)$, для которого справедливо тождество

$$B(u, \eta) = 0 \text{ при } \forall \eta \in \overset{0}{W}{}_2^1(l). \tag{V.11}$$

Иными словами, в рамках предположений 1) — 4) для задачи Дирихле (V.1), (V.2) доказано существование и единственность обобщенного решения $u(x)$ из пространства $\overset{0}{W}{}_2^1(l)$. Поскольку соотношение

$$B(u, \eta) = (Au, \eta)_{2,1}, \quad \forall \eta \in \overset{0}{W}{}_2^1(l), \tag{V.12}$$

определяет оператор A на всем пространстве $\overset{0}{W}{}_2^1(l)$, тождество (V.11) эквивалентно операторному уравнению

$$Au = 0$$

в пространстве $\overset{0}{W}{}_2^1(l)$.

Замечание. Как уже упоминалось, подробное исследование разрешимости n-мерных краевых задач в различных функциональных пространствах (а не только в $W_2^1(\Omega)$) для уравнений вида (V.1), где $x = (x_1, \ldots, x_n)$, $\frac{du}{dx}(x) = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \ldots, \frac{\partial u_n}{\partial x_n}\right)$, и изучение дифференциальных свойств этих решений даны в [56].

В частности, для одномерной задачи Дирихле (V.1), (V.2) можно определить обобщенное решение $u(x)$ из $\overset{0}{W}{}_r^1(l)$, $r > 1$, как элемент пространства $\overset{0}{W}{}_r^1(l)$, удовлетворяющий интегральному тождеству (V.3) $B(u, \eta) = 0$ при любой функции $\eta(x) \in \overset{0}{W}{}_r^1(l)$.

Если при этом функции $a_i(x, u, p)$, $i = 1, 0$, удовлетворяют условиям 1), 3), 4) (с соответствующей поправкой: функции $u(x)$, $v(x)$ принадлежат $\overset{0}{W}{}_r^1(l)$, $r > 1$), а условие 2) заменено условием 2′);

2′) функции $a_i(x, u, p)$, $i = 1, 0$, удовлетворяют неравенствам (ср. с (V.4))

$$|a_1(x, u, p)| \leqslant \mu_1(u)(|p|^{r-1} + \varphi_1(x)), \quad \varphi_1 \in L_{\frac{r}{r-1}}(l),$$

$$|a_0(x, u, p)| \leqslant \mu_2(u)(|p|^r + \varphi_2(x)), \quad \varphi_2 \in L_1(l),$$

где $\mu_i(u)$ — непрерывные функции u, то утверждения об однозначной разрешимости задачи Дирихле (V.1), (V.2) в $\overset{0}{W}{}_r^1(l)$ и однозначной разрешимости систем для соответствующих галеркинских приближений

остаются в силе. (Доказательство основано на использовании свойств монотонных операторов в рефлексивных банаховых пространствах.)

2. Оценка погрешности метода Бубнова — Галеркина. Для оценки близости приближенного по методу Бубнова — Галеркина решения $u^N(x)$ к искомому обобщенному решению $u(x) \in \overset{0}{W}{}_2^1(l)$ задачи (V.1), (V.2) сформулируем представленные в работе [11] результаты в виде следующей теоремы.

Теорема V.1. *Если выполняются условия 1), 2) и существует положительная константа k такая, что*

$$B(u, u-v) - B(v, u-v) \geqslant k \| u-v \|_{2,1}^2, \quad \forall u, v \in \overset{0}{W}{}_2^1(l), \quad \text{(V.13)}$$

то для обобщенного решения $u(x) \in \overset{0}{W}{}_2^1(l)$ задачи (V.1), (V.2) и единственного приближенного по Галеркину решения $u^N(x) \in P_N \subset \overset{0}{W}{}_2^1(l)$ справедлива оценка

$$\| u-u^N \|_{2,1}^2 \leqslant c \inf_{v \in P_N} \| u-v \|_{2,1}, \quad \text{(V.14)}$$

где c — некоторая положительная константа.

Если же оператор A, определяемый соотношением (V.12), удовлетворяет условию Липшица при ограниченных аргументах, то справедливо неравенство

$$\| u-u^N \|_{2,1} \leqslant c \inf_{v \in P_N} \| u-v \|_{2,1}. \quad \text{(V.15)}$$

Отметим, что в данной теореме условие (V.13), гарантирующее сильную монотонность оператору A, обеспечивает тем самым выполнение условий 3), 4).

Замечание. В случае обобщенного решения $u(x) \in \overset{0}{W}{}_r^1(l)$ аналогичный результат вытекает из следующей теоремы [16].

Теорема V.2. *Пусть оператор A, действующий из рефлексивного банахова пространства $\mathfrak{B}$ в сопряженное пространство $\mathfrak{B}^*$, сильно монотонен и удовлетворяет условию Липшица, т. е.*

$$\| Au - Av \|_* \leqslant M \| u-v \|_{\mathfrak{B}}, \quad \forall u, v \in \mathfrak{B},$$

$$M = \text{const}.$$

Тогда имеет место оценка $\| u^N - u \|_{\mathfrak{B}} \leqslant k \inf_{v \in P^N} \| v-u \|_{\mathfrak{B}}$, *где* $k = \frac{M}{\gamma} > 0$, *γ — постоянная монотонности.*

Неравенства (V.14), (V.15) позволяют оценить для некоторых нелинейных краевых задач погрешность приближенного решения, полученного методом конечных элементов (вариант метода Галеркина). В этом случае в качестве пространства P_N выбираются описанные ранее конечномерные пространства P_n^h метода конечных элементов, базисными функциями $\varphi_i^N(x)$ которых являются кусочные полиномы с конечными носителями. Искомое приближение и в данном случае

однозначно представляется в виде разложения по базисным функциям $\varphi_i^N(x)$, $i = 1, 2, \ldots, s$, МКЭ

$$u^N(x) = \sum_{i=1}^{s} \omega_i \varphi_i^N(x),$$

где N — число элементов на отрезке $\bar{l} = [a, b]$, ω_i — узловые параметры функции $u^N(x)$, которые являются искомыми неизвестными дискретной задачи (V.7), s — размерность пространства P_n^h.

Теперь для получения на основе неравенства (V.14), (V.15) оценки погрешности приближенного обобщенного решения задачи (V.1), (V.2), построенного методом конечных элементов, достаточно использовать аппроксимационную теорему II.4. Объединяя приведенные выше результаты, легко увидеть справедливость следующей теоремы.

Теорема V.3. *Пусть обобщенное решение* $u(x) \in \overset{0}{W}{}_2^1(l)$ *задачи (V.1), (V.2) имеет суммируемую с квадратом обобщенную производную* $u^{(n+1)}(x)$ $(n+1)$*-го порядка, а определяемый соотношением (V.12) оператор* A *удовлетворяет условию Липшица для ограниченных аргументов. Если степень пространства* P_n^h *метода конечных элементов равна* n*, то для соответствующего приближенного решения* $u^N \in P_n^h \subset \overset{0}{W}{}_2^1(l)$ *задачи (V.1), (V.2) справедлива оценка*

$$\|u - u^N\|_{2,1} \leqslant c_1 h^n \|u^{(n+1)}\|, \quad h = \max_i |x_i - x_{i-1}|.$$

Если же оператор A *не удовлетворяет условию Липшица, но для квазибилинейной формы задачи (V.1), (V.2) выполняется условие (V.13), то имеет место оценка*

$$\|u - u^N\|_{2,1} \leqslant c_2 h^{n/2} \|u^{(n+1)}\|.$$

Справедливость приведенных оценок непосредственно следует из (V.14), (V.15), (II.111) с учетом неравенства $\|u - u_I\|_{2,1} \geqslant \inf\limits_{v \in P_n^h} \|u - v\|_{2,1}$, где u_I — интерполянт $u(x)$ из пространства P_n^h.

Из приведенных результатов видно, что для некоторых классов нелинейных краевых задач применение МКЭ обеспечивает такую же точность приближенных решений, как и для линейных.

3. Численные примеры. Рассмотрим несколько примеров решения нелинейных краевых задач методом конечных элементов, основанным на процессе Галеркина.

Пример 1. Найти приближенное решение краевой задачи

$$\frac{d}{dx}\left(\frac{du}{dx} - 3u\right) - \frac{1}{2}\left[(u+x)^3 - \frac{6}{2-x} + 7\right] = 0, \quad 0 < x < 1, \tag{V.16}$$

$$u(0) = u(1) = 0, \tag{V.17}$$

точное решение которой $u(x) = \dfrac{x(x-1)}{2-x}$.

Рассмотрим подробно данную задачу в свете теоретических положений, изложенных в данном параграфе. Под обобщенным решением

задачи (V.16), (V.17) будем понимать функцию $u(x) \in \overset{0}{W}{}_2^1(l)$, $l = (0, 1)$, удовлетворяющую интегральному тождеству

$$B(u, \eta) \equiv \int_0^1 \left[\left(\frac{du}{dx} - 3u\right)\frac{d\eta}{dx} + \frac{1}{2}\left((u+x)^3 + 7 - \frac{6}{2-x}\right)\eta\right] dx = 0 \tag{V.18}$$

при любой функции $\eta(x) \in \overset{0}{W}{}_2^1(l)$.

Очевидно, что условия 1), 2) п.1 параграфа V.1 здесь выполнены. Покажем, что выполняется и условие (V.13), гарантирующее сильную монотонность оператора A, определяемого соотношением

$$B(u, \eta) = (Au, \eta)_{2,1}, \quad \forall\, \eta \in \overset{0}{W}{}_2^1(l). \tag{V.19}$$

Действительно, согласно (V.18) имеем

$$B(u, \eta) - B(v, \eta) = \int_0^1 \left[\left(\frac{du}{dx} - \frac{dv}{dx}\right)\frac{d\eta}{dx} - 3(u-v)\frac{d\eta}{dx} + \right.$$

$$\left. + \frac{1}{2}(u-v)\left((u+x)^2 + (u+x)(v+x) + (v+x)^2\right)\eta\right] dx. \tag{V.20}$$

Полагая $\eta = u - v$, где u и v — произвольные функции из $\overset{0}{W}{}_2^1(l)$, $u \neq v$, находим

$$B(u, u-v) - B(v, u-v) = \int_0^1 \left(\frac{du}{dx} - \frac{dv}{dx}\right)^2 dx +$$

$$+ \frac{1}{2}\int_0^1 (u-v)^2\left[(u+x)^2 + (u+x)(v+x) + (v+x)^2\right] dx, \tag{V.21}$$

так как $\int_0^1 3(u-v)\frac{d(u-v)}{dx}dx = 0$ при $\forall\, u, v \in \overset{0}{W}{}_2^1(l)$. Поскольку подынтегральное выражение второго интеграла в соотношении (V.21) неотрицательно при любых $u, v \in \overset{0}{W}{}_2^1(l)$, имеем

$$B(u, u-v) - B(v, u-v) \geqslant \int_0^1 \left(\frac{du}{dx} - \frac{dv}{dx}\right)^2 dx,$$

откуда по известному неравенству Фридрихса

$$\int_0^1 \varphi^2 dx \leqslant \frac{1}{2}\int_0^1 \left(\frac{d\varphi}{dx}\right)^2 dx, \quad \forall\, \varphi \in \overset{0}{W}{}_2^1(l),$$

следует справедливость условия (V.13):

$$B(u, u-v) - B(v, u-v) \geqslant \int_0^1 \left(\frac{du}{dx} - \frac{dv}{dx}\right)^2 dx =$$

$$= \frac{1}{3}\int_0^1 \left(\frac{du}{dx} - \frac{dv}{dx}\right)^2 dx + \frac{2}{3}\int_0^1 \left(\frac{du}{dx} - \frac{dv}{dx}\right)^2 dx \geqslant$$

$$\geqslant \frac{2}{3}\int_0^1 \left[\left(\frac{du}{dx} - \frac{dv}{dx}\right)^2 + (u-v)^2\right] dx = \frac{2}{3}\|u-v\|_{2,1}^2.$$

Таким образом, для рассматриваемой краевой задачи гарантируется существование и единственность определенного выше обобщенного решения, а также возможность построения приближенного решения по варианту МКЭ, основанному на методе Бубнова — Галеркина.

Докажем теперь, что оператор A, определяемый равенством (V.19), удовлетворяет условию Липшица для ограниченных аргументов, т. е. для любого заданного $c > 0$ существует такая константа $K = K(c)$, что $\|Au - Av\|_{2,1} \leqslant K(c)\|u - v\|_{2,1}$ при любых $u, v \in \overset{0}{W}{}_2^1(l)$, удовлетворяющих условию $\|u\|_{2,1} \leqslant c$, $\|v\|_{2,1} \leqslant c$.
Так как

$$\|Au - Av\|_{2,1} \equiv \sup_{\eta \neq 0} \frac{|(Au - Av, \eta)_{2,1}|}{\|\eta\|_{2,1}}, \qquad \text{(V.22)}$$

рассмотрим выражение (см. (V.19))

$$|(Au - Av, \eta)_{2,1}| = |B(u, \eta) - B(v, \eta)|,$$

которое обозначим через M.
Согласно (V.20) имеем

$$M \equiv |B(u, \eta) - B(v, \eta)| \leqslant \left|\int_0^1 \left(\frac{du}{dx} - \frac{dv}{dx}\right)\frac{d\eta}{dx}dx\right| +$$

$$+ 3\left|\int_0^1 (u-v)\frac{d\eta}{dx}dx\right| +$$

$$+ \frac{1}{2}\left|\int_0^1 (u-v)[(u+x)^2 + (u+x)(v+x) + (v+x)^2]\,\eta dx\right|.$$

Учитывая условие $\|u\|_{2,1} \leqslant c$, $\|v\|_{2,1} \leqslant c$, можно написать

$$M \leqslant \left[\int_0^1 \left(\frac{d}{dx}(u-v)\right)^2 dx\right]^{1/2}\left[\int_0^1 \left(\frac{d\eta}{dx}\right)^2 dx\right]^{1/2} +$$

$$+ 3\left(\int_0^1 (u-v)^2 dx\right)^{1/2}\left(\int_0^1 \left(\frac{d\eta}{dx}\right)^2 dx\right)^{1/2} +$$

$$+ \frac{3(c+1)^2}{2}\left(\int_0^1 (u-v)^2 dx\right)^{1/2}\left(\int_0^1 \eta^2 dx\right)^{1/2},$$

что определяет справедливость оценки

$$|(Au - Av, \eta)_{2,1}| \equiv M \leqslant K(c) \|u - v\|_{2,1} \|\eta\|_{2,1},$$

где $K(c) = \left[1 + \sqrt{3} + \frac{(c+1)^2}{2}\right]$.

Полученная оценка с учетом (V.22) обеспечивает выполнение неравенства

$$\|Au - Av\|_{2,1} \leqslant K(c) \|u - v\|_{2,1},$$

следовательно, оператор A удовлетворяет условию Липшица для ограниченных аргументов. А это гарантирует сходимость приближенного МКЭ решения к искомому обобщенному решению со скоростью, определенной теоремой V.3.

Для дискретизации исследуемой задачи и вычисления ее приближенного решения разобьем отрезок [0, 1] на N отрезков $[x_{i-1}, x_i]$, длина которых $h_i = x_i - x_{i-1}$, $i = 1, 2, \ldots, N$, $x_0 = 0$, $x_N = 1$. Поскольку искомое приближенное решение $u^N(x)$ должно принадлежать пространству $W_2^1(l)$, достаточно предположить, что $u^N(x)$ является непрерывной кусочно-линейной функцией, однозначно определяемой на каждом отрезке $[x_{i-1}, x_i]$ своими значениями в двух узловых точках: x_{i-1} и x_i. Согласно теореме V.3 в этом случае гарантируется сходимость решения МКЭ в норме $\|\ \|_{2,1}$ со скоростью порядка h.

В качестве базиса конечномерного подпространства $P_1^h \subset \overset{0}{W}{}_2^1(l)$, содержащего все непрерывные кусочно-линейные полиномы, обращающиеся в нуль на концах отрезка [0, 1], выберем функции $\varphi_i^N(x)$, $i = 1, 2, \ldots, N-1$, вида

$$\varphi_i^N(x) = \begin{cases} \dfrac{x - x_{i-1}}{h_i}, & x_{i-1} \leqslant x \leqslant x_i, \\ \dfrac{x_{i+1} - x}{h_{i+1}}, & x_i \leqslant x \leqslant x_{i+1}, \\ 0, & x \in [0, x_{i-1}] \cup [x_{i+1}, 1], \end{cases} \tag{V.23}$$

т. е. используем элемент «l 1 — 2».

Искомое приближенное решение представим в виде

$$u^N(x) = \sum_{i=1}^{N-1} u_i \varphi_i^N(x)$$

и для определения $(N-1)$ неизвестных $u_i = u^N(x_i)$ получим систему $(N-1)$ нелинейных алгебраических уравнений (см. (V.18))

$$B(u^N, \varphi_i^N) = 0, \quad i = 1, 2, \ldots, N-1. \tag{V.24}$$

Для рассматриваемой задачи (V.16), (V.17) в случае равномерной сетки $h_i \equiv h$, $i = 1, 2, \ldots, N-1$, система (V.24) имеет вид (интегралы вычислялись точно)

$$\frac{1}{h}(u_{i-1} - 2u_i + u_{i+1}) - 1{,}5(u_{i+1} - u_{i-1}) - 0{,}2hu_i^3 -$$

$$-0{,}075hu_i^2(u_{i-1}+u_{i+1})-0{,}05hu_i(u_{i-1}^2+u_{i+1}^2)-$$
$$-0{,}025h(u_{i-1}^3+u_{i+1}^3)-0{,}75h^2iu_i^2-0{,}25h^2i(u_{i-1}+u_{i+1})u_i-$$
$$-0{,}1h^2u_i(u_{i+1}-u_{i-1})-0{,}125h^2i(u_{i-1}^2+u_{i+1}^2)-$$
$$-0{,}075h^2(u_{i+1}^2-u_{i-1}^2)-(i^2+0{,}1)h^3u_i-(i^2+0{,}075)h^3\times$$
$$\times(u_{i-1}+u_{i+1})-0{,}025h^3i(u_{i+1}-u_{i-1})-0{,}25h^4(2i^3+i)-$$
$$-3{,}5h+\frac{3}{h}(2-x_{i-1})\ln(2-x_{i-1})-\frac{6}{h}(2-x_i)\ln(2-x_i)+$$
$$+\frac{3}{h}(2-x_{i+1})\ln(2-x_{i+1})=0,\quad i=1,\ 2,\ \ldots,\ N-1,\qquad (V.25)$$
$$h=\frac{1}{N},\ u_0=u_N=0.$$

Решение данной системы $u=[u_1,\ u_2,\ \ldots,\ u_{N-1}]^T$ вычислялось квазиньютоновским методом, идею которого можно представить следующим образом. Как известно, решение системы нелинейных уравнений

$$F(y)=0:F_i(y_1,\ y_2,\ \ldots,\ y_n)=0,\quad i=1\div n,$$

при достаточно хорошем начальном приближении $y^0=[y_1^0,\ y_2^0,\ \ldots,\ y_n^0]^T$ эффективно вычисляется методом Ньютона:

$$y^{k+1}=y^k-M_k^{-1}F(y^k),\quad k=0,\ 1,\ \ldots,$$

где M_k — матрица Якоби:

$$M_k\equiv M(y^k)=\left[\begin{array}{ccc}\frac{\partial F_1}{\partial y_1} & \frac{\partial F_1}{\partial y_2} & \cdots\ \frac{\partial F_1}{\partial y_n}\\ \cdots & \cdots & \cdots\\ \frac{\partial F_n}{\partial y_1} & \frac{\partial F_n}{\partial y_2} & \cdots\ \frac{\partial F_n}{\partial y_n}\end{array}\right]_{y=y^k}$$

Обращение M_k на каждом шаге итерационного процесса — весьма трудоемкая операция. Поэтому в квазиньютоновском методе предлагается обращать (прямым методом) только $\tilde{M}_0$ — некоторое начальное приближение к якобиану системы, а на последующих шагах $\tilde{M}_k^{-1}$, $k=1,\ 2,\ \ldots$, определяется лишь путем некоторой модификации $\tilde{M}_{k-1}^{-1}$. Матрица $\tilde{M}_k^{-1}$ отличается от $\tilde{M}_{k-1}^{-1}$ только матрицей ранга 2; вид формул модификации зависит от того, является ли матрица Якоби положительно определенной или нет. Заметим, что $\tilde{M}_0$ не обязательно совпадает с $M_0(y^0)$. Матрица $\tilde{M}_0$ может быть матрицей Якоби, вычисленной для некоторого $\tilde{y}$, отличного от начального приближения y^0. (Подробнее о квазиньютоновском методе см. в [148].)

Решение системы (V.25) для различных h вычислялось на ЭВМ ЕС 1060 с двойным машинным словом (64 бита). Критерием окончания итерационного процесса служило условие:

$$\|u^k-u^{k-1}\|_E\leqslant\varepsilon,\quad \|F(u^k)\|_E\leqslant\varepsilon_1,$$

Таблица 14

x_i	$u_T(x_i)$	$h = 0{,}25$		$h = 0{,}125$		$h = 0{,}0625$	
		u^N	δ (%)	u^N	δ (%)	u^N	δ (%)
0,125	—0,0583 (3)			—0,05856	0,39	—0,05839	0,09
0,25	—0,107142857	—0,1092	1,88	—0,1076	0,45	—0,10726	0,11
0,375	—0,14423077			—0,1450	0,54	—0,144420	0,13
0,5	—0,166 (6)	—0,1711	2,64	—0,1677	0,63	—0,16692	0,16
0,625	—0,1704 (45)			—0,1717	0,74	—0,17077	0,18
0,75	—0,15	—0,1557	3,81	—0,1513	0,89	—0,15033	0,22
0,875	—0,0972 (2)			—0,09827	1,79	—0,09748	0,26

Примечание. При $h = 0{,}25$ значение $\tilde{u}^0 = [-0{,}075, -0{,}15, -0{,}225]^T$, количество выполненных итераций $s = 4$, время (процессорное) решения алгебраической системы $t = 1{,}09$ с; при $h = 0{,}125$ значение $\tilde{u}^0 = [-0{,}0375, -0{,}075, -0{,}1125, -0{,}15, -0{,}1875, -0{,}225, -0{,}2625]^T$, $s = 5$, $t = 1{,}81$ с; при $h = 0{,}0625$ значение $\tilde{u}^0 = [-0{,}01875, -0{,}0375, -0{,}05625, -0{,}075, -0{,}09375, -0{,}1125, -0{,}13125, -0{,}15, -0{,}1687, -0{,}1875, -0{,}2062, -0{,}225, -0{,}2437, -0{,}2625, -0{,}2807]^T$, $s = 5$, $t = 4{,}73$ с.

где $u = [u_1, u_2, \ldots, u_{N-1}]^T$ — вектор решения системы (V.25), k — номер итерации квазиньютоновского процесса, $\| \|_E$ — евклидова норма.

Полученные результаты для разных значений приведены в табл. 14. В качестве начального приближения к решению здесь везде выбирался нулевой вектор $u^0 = [0, 0, \ldots, 0]^T$, итерационный процесс оканчивался при $\varepsilon = 5 \cdot 10^{-5}$ и $\varepsilon_1 = 10^{-5}$. В табл. 14 (и последующих) используются такие обозначения: $u_T(x_i)$ — значения точного решения краевой задачи в узловых точках; $u^N = [u_1, u_2, \ldots, u_{N-1}]^T$ — значения вычисленного МКЭ приближенного решения задачи, т. е. $u_i \equiv u^N(x_i)$; $\delta(\%) = \dfrac{|u_T(x_i) - u^N(x_i)|}{|u_T(x_i)|} \cdot 100\ \%$ — относительная погрешность приближенного решения $\tilde{u}^0 = [\tilde{u}_1^0, \tilde{u}_2^0, \ldots, \tilde{u}_{N-1}^0]^T$ — значения неизвестных для вычисления $\tilde{M}_0$.

Пример 2. Найти приближенное решение краевой задачи

$$-\frac{d}{dx}\left(u\frac{du}{dx} + f(x)\frac{du}{dx}\right) + 2\left(\frac{du}{dx}\right)^2 + 3c\frac{du}{dx} + c^2 = 0,$$

$$f(x) = cx + 1, \quad c = e - 1, \quad 0 < x < 1, \tag{V.26}$$

$$u(0) = u(1) = 0, \tag{V.27}$$

единственное точное решение которой есть $u(x) = e^x - f(x)$.

Вариант 1. Для получения приближенного решения использовались кусочно-линейные базисные функции (V.23), интегралы вычислялись точно. В этом случае на равномерной сетке соответствующая система нелинейных алгебраических уравнений имела вид

$$\frac{1}{2h}u_{i-1}^2 - \frac{2}{h}u_{i-1}u_i + \frac{3}{h}u_i^2 - \frac{2}{h}u_iu_{i+1} + \frac{1}{2h}u_{i+1}^2 -$$

$$-\frac{ch(i+1)+1}{h}u_{i-1} + \frac{2(chi+1)}{h}u_i - \frac{ch(i-1)+1}{h} + ch^2 = 0,$$

$$i = 1, 2, \ldots, N-1, \quad u_0 = u_N = 0, \quad h = \frac{1}{N}. \tag{V.28}$$

Решалась система (V.28) квазиньютоновским методом. Полученные результаты представлены в табл. 15 для разных значений h. Обозначения в табл. 15 такие же, как и в табл. 14; итерации прекращались при $\varepsilon = 5 \cdot 10^{-5}$, $\varepsilon_1 = 10^{-5}$.

В а р и а н т 2. Приближенное решение задачи (V.26), (V.27) определялось с использованием в качестве базисных функций МКЭ кусочно-кубических полиномов Эрмита:

$$\varphi_i^N(x) = \begin{cases} \dfrac{(x-x_{i-1})^2}{h_i^2}\left(2\dfrac{x_i-x}{h_i}+1\right), & x_{i-1} \leqslant x \leqslant x_i, \\ \dfrac{(x_{i+1}-x)^2}{h_{i+1}^2}\left(2\dfrac{x-x_i}{h_{i+1}}+1\right), & x_i \leqslant x \leqslant x_{i+1}, \\ 0, \quad x \in [0, x_{i-1}] \cup [x_{i+1}, 1], \end{cases}$$

$$\psi_i^N(x) = \begin{cases} \dfrac{(x-x_i)(x-x_{i-1})^2}{h_i^2}, & x_{i-1} \leqslant x \leqslant x_i, \\ \dfrac{(x-x_i)(x-x_{i+1})^2}{h_{i+1}^2}, & x_i \leqslant x \leqslant x_{i+1}, \\ 0, \quad x \in [0, x_{i-1}] \cup [x_{i+1}, 1], \end{cases} \qquad \text{(V.29)}$$

т. е. использовался элемент вида «*l*3—2».

Расчет выполнялся на сетке с равномерным шагом $h_i \equiv h = 1/N$ для $N = 2, 4, 8$. При построении системы нелинейных алгебраических уравнений интегралы вычислялись точно. Вид системы не приводится вследствие громоздкости. Системы решались квазиньютоновским методом, итерации прекращались при $\varepsilon = 5 \cdot 10^{-5}$, $\varepsilon_1 = 10^{-5}$. Полученные результаты представлены в табл. 16, 17. Заметим, что здесь значения $u_T(x_i)$ не выписаны, так как они имеются в табл. 15. Табл. 16 содержит значения производной точного решения $u_T'(x_i)$ задачи (V.26), (V.27) и значения приближенного решения $u^N(x_i)$ и его производной $(u^N)'|_{x=x_i}$, полученные при h равном 0,5 и 0,25. В обоих случаях использовалось нулевое начальное приближение к решению системы алгебраических уравнений, а соответствующие значения вектора $\tilde{u}^0$ указаны в примечании к таблице. Результаты, полученные на сетке с шагом $h = 0{,}125$ при двух разных начальных приближениях u^0, даны в табл. 17; матрица Якоби здесь вычислялась при одном и том же значении $\tilde{u}^0$ (см. примечание к табл. 17).

В а р и а н т 3. Приближенное решение задачи (V.26), (V.27) вычислялось с использованием кусочно-кубических базисных функций (V.29), но интегралы при построении системы нелинейных алгебраических уравнений вычислялись по квадратурным формулам Гаусса с тремя квадратурными узлами. Вид полученной системы не приводится вследствие громоздкости. Сетка всюду равномерная. Система, как и прежде, решалась квазиньютоновским методом, всюду итерационный процесс оканчивался при $\varepsilon = 5 \cdot 10^{-5}$, $\varepsilon_1 = 10^{-5}$. Полученные результаты представлены в табл. 18, где используются прежние

Таблица 15

x_i	$u_T(x_i)$	$h = 0{,}25$		$h = 0{,}125$		$h = 0{,}0625$	
		u^N	δ (%)	u^N	δ (%)	u^N	δ (%)
0,125	—0,081636775	—	—	—0,0821	0,6	—0,0817	0,15
0,25	—0,1455450404	—0,1495	2,7	—0,1465	0,7	—0,1457	0,17
0,375	—0,1893642710			—0,1907	0,7	—0,1896	0,18
0,5	—0,210419643	—0,2173	3,3	—0,2120	0,9	—0,2108	0,19
0,625	—0,205680185			—0,2074	0,8	—0,2061	0,2
0,750	—0,171711354	—0,1784	4,0	—0,1732	0,9	—0,1721	0,23
0,875	—0,10462130	—	—	—0,1056	0,9	—0,1048	0,25

Примечание. При $h = 0{,}25$ значение $\tilde{u}^0 = [-0{,}1, -0{,}2, -0{,}1]^T$, количество выполненных итераций $s = 7$; при $h = 0{,}125$ значение $\tilde{u}^0 = [-0{,}1, -0{,}1, -0{,}2, -0{,}2, -0{,}2, -0{,}1, -0{,}1]^T$, $s = 8$; при $h = 0{,}0625$ значение $\tilde{u}^0 = [-0{,}1, -0{,}1, -0{,}1, -0{,}1, -0{,}2, -0{,}2, -0{,}2, -0{,}2, -0{,}2, -0{,}2, -0{,}2, -0{,}2, -0{,}1, -0{,}1, -0{,}1, -0{,}1]^T$, $s = 10$.

Таблица 16

x_i	u_T	$h = 0{,}5$				$h = 0{,}25$			
		u^N	δ (%)	$(u^N)'$	δ (%)	u^N	δ (%)	$(u^N)'$	δ (%)
0	—0,718281828	—	—	—0,7156	0,37	—	—	—0,7180	0,033
0,25	—0,434256442	—	—	—	—	—0,1455	0,0046	—0,4341	0,029
0,5	—0,069560558	—0,2103	0,051	—0,0684	1,55	—0,2104	0,0043	—0,0694	0,160
0,75	0,398718188	—	—	—	—	—0,1717	0,0053	0,3987	0,011
1	1	—	—	0,9977	0,23	—	—	0,9996	0,037

Примечание. При $h = 0{,}5$ значение $\tilde{u}^0 = [-0{,}8, -0{,}2, -0{,}06, 1{,}0]^T$, количество выполненных итераций $s = 5$, время (процессорное) решения системы $t = 1{,}92$ с; при $h = 0{,}25$ значение $\tilde{u}^0 = [-0{,}8, -0{,}15, -0{,}43, -0{,}2, -0{,}06, -0{,}15, 0{,}31, 1{,}0]^T$, $s = 5$, $t = 3{,}28$ с.

Таблица 17

x_i	$u^0 = [0, 0, \ldots, 0]^T$				$u^0 = \tilde{u}^0$			
	u^N	δ (%)	$(u^N)'$	δ (%)	u^N	δ (%)	$(u^N)'$	δ (%)
0	—	—	—0,7182509	0,0043	—	—	—0,7182535	0,0039
0,25	—0,1455439	$0{,}7 \cdot 10^{-3}$	—0,4342473	0,0021	—0,145544578	$0{,}3 \cdot 10^{-3}$	—0,4342492	0,0016
0,5	—0,21041806	$0{,}7 \cdot 10^{-3}$	—0,06955238	0,0118	—0,2104191272	$0{,}3 \cdot 10^{-3}$	—0,06955277	0,0112
0,75	—0,1717098	$0{,}8 \cdot 10^{-3}$	0,3987241	0,0015	—0,1717107748	$0{,}3 \cdot 10^{-3}$	0,3987264	0,0026
1	—	—	0,9999412	0,0059	—	—	0,99994626	0,0054

Примечание. В обоих случаях $\tilde{u}^0 = [-0{,}8, -0{,}1, -0{,}615, -0{,}15, -0{,}43, -0{,}2, -0{,}245, -0{,}2, -0{,}06, -0{,}2, 0{,}125, -0{,}15, 0{,}31, -0{,}1, 0{,}495, 1{,}0]^T$; при $u^0 = 0$ количество выполненных итераций $s = 5$, время (процессорное) решения системы $t = 9{,}25$ с.; при $u^0 = \tilde{u}^0$ $s = 9$, $t = 11{,}89$ с.

Таблица 18

x_i	$h = 0{,}5$				$h = 0{,}25$				$h = 0{,}125$			
	u^N	δ (%	$(u^N)'$	δ (%)	u^N	δ (%)	$(u^N)'$	δ (%)	u^N	δ (%)	$(u^N)'$	δ (%)
0	—	—	—0,717340	0,13	—	—	—0,7203975	0,29	—	—	—0,719692	0,19
0,25	—	—	—	—	—0,145610	0,04	—0,434392	0,03	—0,145571	0,018	—0,434435	0,04
0,5	—0,210477	0,028	—0,066277	4,72	—0,210490	0,03	—0,069185	0,53	—0,210450	0,01	—0,0694731	0,12
0,75	—	—	—	—	—0,171761	0,02	0,399322	0,15	—0,171734	0,01	0,398913	0,048
1	—	—	—1,002847	0,28	—	—	1,001594	0,15	—	—	1,000853	0,08

Примечание. Для каждого h соответствующее значение $\tilde{u}^0$ такое же, как в табл. 16, 17, при $h = 0{,}5$ количество выполненных итераций $s = 6$, время решения системы $t = 1{,}51$ с; при $h = 0{,}25$ $s = 5$, $t = 1{,}95$ с.; при $h = 0{,}125$ $s = 9$, $t = 7{,}53$ с.

обозначения. Заметим, что в данном варианте в качестве $\tilde{M}_0$ всюду бралась матрица Якоби, вычисленная в точке начального приближения к решению алгебраической системы, т. е. для значений $\tilde{u}^0 = [(u_0')^0, u_1^0, (u_1')^0, u_2^0, \ldots, (u_n')^0]^T$.

V.2. Решение нелинейных вариационных задач

Как известно, исследование многих научно-технических проблем сводится к решению вариационных задач о минимизации заданного функционала. В настоящее время широко известны численные методы, позволяющие находить приближенные решения таких задач, минуя переход к дифференциальным уравнениям. Таким, в частности, является метод Ритца. Основываясь на этом методе и используя базисные функции МКЭ, можно в ряде случаев эффективно находить решения достаточно широкого класса нелинейных задач.

1. О существовании решения вариационной задачи. Попытаемся рассмотреть особенности применения МКЭ к решению нелинейных вариационных задач на примере задачи об отыскании функции $u(x)$, доставляющей наименьшее значение простейшему функционалу

$$F(u) = \int_0^1 f\left(x, u, \frac{du}{dx}\right) dx \qquad \text{(V.30)}$$

на множестве $\mathfrak{M}$, состоящем из всех функций, на которых функционал $F(u)$ конечен и которые удовлетворяют условию

$$u(0) = u(1) = 0. \qquad \text{(V.31)}$$

Эта задача, как известно, тесно связана с задачей решения уравнения Эйлера

$$L(u) \equiv \frac{d}{dx}\left(\frac{\partial f}{\partial p}\right) - \frac{\partial f}{\partial u} = 0, \quad p = \frac{du}{dx} \qquad \text{(V.32)}$$

при граничных условиях (V.31) и с задачей отыскания стационарных точек рассматриваемого функционала, т. е. отыскания таких функций $u(x)$, на которых первая вариация $\delta F(u, \eta)$ функционала (V.30), (V.31) обращается в нуль при всех допустимых вариациях $\eta(x)$:

$$\delta F(u, \eta) \equiv \int_0^1 \left(\frac{\partial f}{\partial p}\frac{d\eta}{dx} + \frac{\partial f}{\partial u}\eta\right) dx = 0. \qquad \text{(V.33)}$$

Отметим, что (V.33) является интегральным тождеством, определяющим обобщенные решения уравнения (V.32) (см. п.1 параграфа V.1), т. е. стационарные точки функционала $F(u)$ являются обобщенными решениями краевой задачи (V.32), (V.31) из соответствующего функционального пространства.

В простейшем случае, когда функция $f(x, u, p)$ достаточно гладкая по переменным u, p и в области ее задания выполняется условие

$$\frac{\partial^2 f}{\partial p^2}\xi^2 + 2\frac{\partial^2 f}{\partial p \partial u}\xi\eta + \frac{\partial^2 f}{\partial u^2}\eta^2 > 0 \qquad \text{(V.34)}$$

при любых вещественных параметрах ξ, η, не равных одновременно нулю, то все три упомянутые задачи эквивалентны друг другу. Единственное общее решение $u(x)$ этих задач реализует абсолютный минимум функционала (V.30), (V.31) [56].

При невыполнении условия (V.34) нельзя гарантировать эквивалентность задачи об отыскании наименьшего значения функционала (V.30), (V.31) и краевой задачи (V.31), (V.32).

Действительно, пусть

$$F(u) = \int_0^1 \left(\frac{du}{dx} - \frac{1}{4}\right)^2 \left(\frac{du}{dx} + \frac{3}{4}\right)^2 dx, \quad u(0) = u(1) = 0.$$

Соответствующее уравнение Эйлера

$$4\frac{d}{dx}\left[\left(\left(\frac{du}{dx}\right)^2 - \frac{1}{16}\right)\left(\frac{du}{dx} + \frac{3}{4}\right)\right] = 0$$

имеет единственное решение $u_0(x) \equiv 0$, удовлетворяющее краевым условиям $u(0) = u(1) = 0$.

Однако $F(u_0) = \frac{9}{256}$, а непрерывная функция

$$u_1(x) = \begin{cases} \frac{x}{4} & \text{при } x \in \left[0, \frac{3}{4}\right], \\ \frac{3}{4}(1-x) & \text{при } x \in \left[\frac{3}{4}, 1\right], \end{cases}$$

принадлежащая области определения $\mathfrak{M}$ рассматриваемого функционала F, доставляет ему наименьшее значение: $F(u_1) = 0$. Более того, для функции

$$u_2(x) = \begin{cases} -\frac{3}{4}x, & x \in \left[0, \frac{1}{4}\right], \\ \frac{1}{4}(x-1), & x \in \left[\frac{1}{4}, 1\right] \end{cases}$$

тоже имеем $F(u_2) = 0$.

Нетрудно убедиться, что неравенство (V.34) в данном случае не выполняется, так как $\frac{\partial^2 f}{\partial p^2}(x, u, p) = 2\left(6p^2 + 3p - \frac{1}{8}\right)$ и при $p = 0$ имеем $\frac{\partial^2 f}{\partial p^2} < 0$ (производные $\frac{\partial^2 f}{\partial p \partial u}$ и $\frac{\partial^2 f}{\partial u^2}$ здесь тождественно равны нулю).

Итак, рассмотрим вопрос о существовании решения вариационной задачи типа (V.30), (V.31) без перехода к дифференциальному уравнению. При этом возможны различные подходы, но всюду будем предполагать, что функционал $F(u)$ определен в рефлексивном банаховом пространстве $\mathfrak{B}$, а области определения функционала и его градиента A, т. е. $D(F)$ и $D(A)$, всегда линейны и плотны в $\mathfrak{B}$. В ряде случаев

ответ на вопрос о существовании решения задачи может дать следующая теорема.

Теорема V.4. *Пусть функционал $F(u)$, определенный на некотором рефлексивном банаховом пространстве, возрастающий, существенно выпуклый и слабо полунепрерывный снизу. Тогда этот функционал ограничен снизу и его нижняя грань достигается в единственной точке u_0, к которой слабо сходится любая минимизирующая последовательность.*

Для многих задач, в частности задач теории пластичности, эффективным может оказаться подход к постановке и решению вариационных задач, основанный на следующей теореме.

Теорема V.5. *Пусть функционал $F(u)$ непрерывен и его градиент $A = \operatorname{grad} F$ имеет при любом элементе $u \in D(A) \subset D(F)$ производную A'_u, равномерно положительно ограниченную снизу, т. е.*

$$\langle A'_u z, z\rangle \geqslant \gamma^2 \|z\|^2_{\mathfrak{B}}, \quad \gamma = \text{const} > 0, \quad z \in D(A'_u). \tag{V.35}$$

Тогда любая минимизирующая последовательность для функционала $F(u)$ сходится в метрике пространства $\mathfrak{B}$ к некоторому пределу, не зависящему от выбора минимизирующей последовательности.

Доказательство. Согласно теореме I.2 функционал $F(u)$ — существенно выпуклый на $D(F)$. Положим

$$\rho(u, v) = \left[F(u) + F(v) - 2F\left(\frac{u+v}{2}\right)\right]^{1/2}. \tag{V.36}$$

Как показано в [64], при $u, v \in D(A)$ справедливо соотношение

$$\rho(u, v) = \left[\int_0^1 \int_0^1 \left\langle A'_\xi \frac{z}{2}, \frac{z}{2}\right\rangle dt\,d\tau\right]^{1/2},$$

где $z = u - v$, $\xi = u + (t+\tau)\frac{z}{2}$, а вследствие (V.35) — и неравенство

$$\rho(u, v) \geqslant \frac{\gamma}{2}\|u - v\|_{\mathfrak{B}}, \quad u, v \in D_A. \tag{V.37}$$

Нетрудно убедиться, что соотношение

$$\rho(u, v) \geqslant K\|u - v\|_{\mathfrak{B}}, \quad K = \text{const} > 0, \tag{V.38}$$

справедливо при любых $u, v \in D(F)$.

Действительно, пусть $\bar{u} \neq \bar{v}$ — произвольные элементы из $D(F)$ и $\|\bar{u} - \bar{v}\|_{\mathfrak{B}} = \alpha$. Положим $\varepsilon = \frac{\gamma^2}{288}\alpha^2$. Так как F непрерывен, то для любого ε можно выбрать такое $\delta_0 > 0$, что при $\|u - \bar{u}\| < \delta_0$, $\|v - \bar{v}\| < \delta_0$ будут справедливы неравенства

$$F(\bar{u}) > F(u) - \varepsilon, \quad F(\bar{v}) > F(v) - \varepsilon,$$

$$F\left(\frac{\bar{u}+\bar{v}}{2}\right) < F\left(\frac{u+v}{2}\right) + \varepsilon,$$

а следовательно, и неравенство

$$\rho^2(\bar{u}, \bar{v}) \equiv F(\bar{u}) + F(\bar{v}) - 2F\left(\frac{\bar{u}+\bar{v}}{2}\right) \geqslant \rho^2(u, v) - 4\varepsilon. \quad \text{(V.39)}$$

Поскольку области $D(A) \subset D(F)$ плотны в $\mathfrak{B}$, любой элемент из $D(F)$ можно сколь угодно близко аппроксимировать элементами из $D(A)$. Поэтому можно предполагать, что в (V.39) элементы $u, v \in D(A)$. А так как при любом $\delta \leqslant \delta_0$ все написанные выше соотношения остаются в силе, то можно взять $\delta \leqslant \frac{1}{3}\|\bar{u} - \bar{v}\|_{\mathfrak{B}} \equiv \frac{\alpha}{3}$ и $\|u - \bar{u}\|_{\mathfrak{B}} < \frac{\alpha}{3}$, $\|u - \bar{v}\|_{\mathfrak{B}} < \frac{\alpha}{3}$. При этом получим

$$\alpha \equiv \|\bar{u} - \bar{v}\|_{\mathfrak{B}} \leqslant \|\bar{u} - u\|_{\mathfrak{B}} + \|u - v\|_{\mathfrak{B}} + \|v - \bar{v}\|_{\mathfrak{B}} < \|u - v\|_{\mathfrak{B}} + \frac{2}{3}\alpha,$$

т. е.

$$\|u - v\|_{\mathfrak{B}} > \frac{\alpha}{3} \equiv \frac{\|\bar{u} - \bar{v}\|_{\mathfrak{B}}}{3}. \quad \text{(V.40)}$$

Неравенства (V.37), (V.40) и условие $\varepsilon = \frac{\gamma^2}{288}\|\bar{u} - \bar{v}\|^2_{\mathfrak{B}}$ позволяют представить (V.39) в виде

$$\rho^2(\bar{u}, \bar{v}) \geqslant \frac{\gamma^2}{72}\|\bar{u} - \bar{v}\|^2_{\mathfrak{B}}, \quad \forall \bar{u}, \bar{v} \in D(F),$$

т. е. убедиться в справедливости соотношения (V.38) для любых элементов из $D(F)$.

Пусть $\{u_n\}$ — минимизирующая последовательность для функционала $F(u)$, так что $F(u_n) \to d = \inf\limits_{u \in D(F)} F(u)$.

Полагая в формуле (V.36) $u = u_n$, $v = u_m$, имеем

$$\rho(u_n, u_m) = \left[F(u_n) + F(u_m) - 2F\left(\frac{u_n + u_m}{2}\right)\right]^{1/2}.$$

Поскольку $F\left(\frac{u_n + u_m}{2}\right) \geqslant \inf\limits_{u \in D(F)} F(u) = d$, при достаточно больших значениях m, n (когда $F(u_n) < d + \varepsilon$ и $F(u_m) < d + \varepsilon$ для произвольно заданной положительной величины ε), получим $\rho(u_n, u_m) < \sqrt{2\varepsilon}$ и, следовательно,

$$\lim_{m,n \to \infty} \rho(u_n, u_m) = 0.$$

Согласно (V.38) можно утверждать, что и $\|u_n - u_m\|_{\mathfrak{B}} \to 0$ при $m, n \to \infty$, т. е. минимизирующая последовательность сходится в метрике пространства $\mathfrak{B}$ к некоторому элементу u_0.

Этот элемент не зависит от выбора минимизирующей последовательности. Если $\{u_n\}$, $\{v_n\}$ — две разные последовательности, то $u_1, v_1, u_2, v_2, \ldots, u_n, v_n, \ldots$ — тоже минимизирующая последовательность и имеет согласно доказанному предел. Но тогда ее подпоследовательности $\{u_n\}$ и $\{v_n\}$ имеют один и тот же предел. Теорема V.5 доказана.

Общий предел u_0 минимизирующих последовательностей будем называть обобщенным решением задачи о минимуме функционала $F(u)$.

Если $u_0 \in D(F)$ (например, при $D(F) = \mathfrak{B}$), то $\min\limits_{u \in D(F)} F(u) = F(u_0)$ и u_0 является классическим решением вариационной задачи.

Отметим, что при выполнении условия (V.35) задача об отыскании минимума функционала (см. п.1 параграфа I. 2)

$$F(u) = \int_0^1 \langle A(tu), u \rangle \, dt + \text{const}, \tag{V.41}$$

определенного на $D(F) = D(A)$, и задача о решении операторного уравнения $Au = 0$ эквивалентны.

Упомянем еще один подход [56] к решению задачи о существовании функции $u(x)$, реализующей минимум функционала

$$F(u) = \int_0^1 f\left(x, u, \frac{du}{dx}\right) dx,$$

$$u(0) = u(1) = 0.$$

В качестве области определения данного функционала примем пространство $\mathfrak{B} = \overset{0}{W}{}_r^1(\Omega)$, $\Omega = (0, 1)$.

Теорема V.6. *Пусть функция $f(x, u, p)$ непрерывна вместе со своими производными $\frac{\partial f}{\partial u}$, $\frac{\partial f}{\partial p}$ и удовлетворяет неравенству $f(x, u, p_1) - f(x, u, p_2) - (p_1 - p_2) \frac{\partial f}{\partial p}(x, u, p_2) \geqslant 0$ при любых $x \in [0, 1]$, u, p_1, p_2. Пусть, кроме того,*

$$|f(x, u, p)| \leqslant \mu(|u|)[|p|^r + \psi_1(x)], \tag{V.42}$$

$$\left|\frac{\partial f}{\partial p}(x, u, p)\right| \leqslant \mu(|u|)[|p|^{r-1} + \psi_2(x)],$$

где $\mu(\tau)$ — неубывающая непрерывная положительная функция $\tau \geqslant 0$, $\psi_1(x) \in L_1$, $\psi_2(x) \in L_{r'}$, $r' = \frac{r}{r-1}$, и

$$F(u) = \int_0^1 f\left(x, u, \frac{du}{dx}\right) dx \geqslant \varphi(\|u\|_{r,1}) - c, \quad c = \text{const},$$

$$\text{при } \forall u(x) \in \overset{0}{W}{}_r^1(\Omega),$$

где $\varphi(\tau)$ — непрерывная функция, стремящаяся к ∞ при $\tau \to \infty$.

Тогда существует хотя бы одна функция $u(x) \in \overset{0}{W}{}_r^1(\Omega)$, доставляющая функционалу $F(u)$ значение, не превосходящее значение $F(u)$ на любой другой функции из $W_r^1(\Omega)$.

Доказательство этой теоремы, подробно изложенное в [56], также основано на исследовании сходимости минимизирующей последовательности для функционала $F(u)$.

Утверждение теоремы V.6 остается в силе, если условие (V.42) заменить требованием

$$f(x, u, p) \geqslant 0$$

и предположением, что существует по крайней мере одна функция $\tilde{u}(x) \in \overset{0}{W}{}^1_r$, на которой $F(\tilde{u}) < \infty$.

Функции, существование которых обеспечивается теоремой V.6, называют обобщенными решениями из класса $\overset{0}{W}{}^1_r$ вариационной задачи о минимуме функционала (V.30), (V.31).

В монографии [56] приводятся и условия, гарантирующие единственность обобщенного решения из $\overset{0}{W}{}^1_r$, реализующего абсолютный минимум функционала $F(u)$. Они касаются поведения функции $f(x, u, p)$ и имеют вид

$$\gamma(|u|)(1+|p|)^{r-2} \leqslant \frac{\partial^2 f}{\partial p^2}(x, u, p) \leqslant \mu(|u|)(1+|p|)^{r-2},$$

$$\left(\left|\frac{\partial f}{\partial p}\right| + \left|\frac{\partial^2 f}{\partial p \partial u}\right|\right)(1+|p|) + \left|\frac{\partial f}{\partial u}\right| + \left|\frac{\partial^2 f}{\partial u^2}\right| \leqslant \mu(|u|)(1+|p|)^r;$$

кроме того, предполагается, что функция $f(x, u, p)$ обладает, например, следующим свойством: существует число $k > 0$ такое, что при $x \in (0, 1)$, $u > k$ справедливы неравенства

$$f(x, u, 0) \geqslant f(x, k, 0),$$
$$f(x, u, p) > f(x, k, 0), \text{ если } |p| > 0,$$

а при $x \in (0, 1)$, $u < -k$ — неравенства

$$f(x, u, 0) \geqslant f(x, -k, 0),$$
$$f(x, u, p) > f(x, -k, 0), \text{ если } |p| > 0.$$

(Последнее требование можно заменить и некоторыми аналогичными, см. [56].)

2. Построение приближенного решения МКЭ. Изложенные в предыдущем пункте результаты показывают, что приближенное решение вариационной задачи в ряде случаев можно получить, построив минимизирующую для данного функционала последовательность. Член этой сходящейся последовательности с достаточно большим номером — искомое приближенное решение.

Построение минимизирующей последовательности можно выполнять посредством процесса Ритца, о чем свидетельствует приведенная ниже теорема V.7 [64].

Пусть, как и прежде, решается задача о минимуме функционала

$$F(u) = \int_0^1 f\left(x, u, \frac{du}{dx}\right) dx, \quad u(0) = u(1) = 0, \tag{V.43}$$

область определения $D(F)$ которого линейна и плотна в сепарабельном банаховом пространстве $\mathfrak{B}$.

Предположим, что функционал $F(u)$ — непрерывно дифференцируем на любом конечномерном линеале из области $D(F)$, т. е. выражения $F\left(\sum_{k=1}^{N} a_k u_k\right)$ для любого N и любых $u_k \in D(F)$ являются дифферен-

цируемыми функциями числовых переменных a_k при всех конечных значениях этих переменных.

Пусть, далее, элементы последовательности $\{\psi_n(x)\}_1^\infty$ удовлетворяют условиям:

1) $\psi_n \in D(F)$;

2) $\psi_1(x)$, $\psi_2(x)$, ..., $\psi_N(x)$ — линейно независимы при любом значении N;

3) последовательность $\{\psi_n\}_1^N$ полна в $\mathfrak{B}$, т. е. множество всевозможных конечных линейных комбинаций ее элементов плотно в $\mathfrak{B}$.

Обозначим, как и прежде, через P_N конечномерное подпространство с базисом $\{\psi_k\}_1^N$ и будем называть приближенным по Ритцу решением задачи о минимуме функционала (V.43) функцию $u^N(x)$ вида

$$u^N(x) = \sum_{k=1}^{N} c_k \psi_k(x),$$

если она доставляет минимум этому функционалу на P_N.

Числовые коэффициенты c_k при этом удовлетворяют системе уравнений

$$\frac{\partial F\left(\sum_{k=1}^{N} c_k \psi_k\right)}{\partial c_j} = 0, \quad j = 1, 2, \ldots, N. \tag{V.44}$$

Теорема V.7. *Пусть выполнены сформулированные выше условия относительно функционала $F(u)$ и последовательности функций $\{\psi_n\}_1^\infty$. Если функционал $F(u)$ в метрике пространства $\mathfrak{B}$ — возрастающий и полунепрерывный сверху, то приближенные по Ритцу решения можно построить при любом значении N и для функционала $F(u)$ эта последовательность — минимизирующая.*

Доказано, что система Ритца (V.44) для рассматриваемого функционала

$$F(u) = \int_0^1 f\left(x, u, \frac{du}{dx}\right) dx, \quad u(0) = u(1) = 0$$

и система Галеркина (см. (V.7) — (V.10))

$$B\left(\sum_{n=1}^{N} c_n \psi_n, \psi_k\right) \equiv \left[\int_0^1 \left[\frac{\partial f}{\partial p}\left(x, u^N, \frac{du^N}{dx}\right)\frac{d\psi_k}{dx} + \right.\right.$$

$$\left.\left. + \frac{\partial f}{\partial u}\left(x, u^N, \frac{du^N}{dx}\right)\psi_k\right] dx \equiv \langle Au^N, \psi_k\rangle = 0\right.$$

$$\left(p = \frac{du}{dx}\right), \quad k = 1, 2, \ldots, N,$$

для определения приближения к обобщенному решению уравнения Эйлера (V.32) при граничных условиях (V.31) равносильны, если $\psi_k \in D(A)$, где $A = \operatorname{grad} F(u)$.

Как и в параграфе V.1, для численного решения нелинейных вариационных задач в качестве подпространства P_N можно использовать соответствующие конечномерные пространства МКЭ. Для более детального рассмотрения этой возможности конкретизируем постановку вариационной задачи, в частности область определения функционала и вид базисных функций конечномерных подпространств данного банахова пространства.

Пусть решается задача о минимизации функционала

$$F(u) = \int_0^1 f\left(x, u, \frac{du}{dx}\right) dx, \qquad \text{(V.45)}$$

$$u(0) = u(1) = 0,$$

областью определения которого служит сепарабельное и рефлексивное банахово пространство $\overset{0}{W}{}_r^1(0, 1)$, $1 < r < \infty$.

Заметим, что большинство результатов будет справедливо (или аналогично) и для более сложных функционалов: содержащих производные высших порядков или функции многих переменных.

В качестве базисных (координатных) функций МКЭ будем использовать функции, подробно исследованные в [67]. Дадим их краткое описание.

Пусть q — некоторое заданное натуральное число, а $\omega_s(t)$, $s = 0, 1, \ldots, q-1$,— совокупность функций одной переменной, удовлетворяющих следующим условиям:

1) $\omega_s(t) \in W_r^q(R_1)$, $1 \leqslant r \leqslant \infty$; R_1 — одномерное евклидово пространство;

2) $\operatorname{supp} \omega_s(t) \subset \{t : 0 \leqslant t \leqslant 2\}$ (следовательно, $\omega_s^{(\alpha)}(0) = \omega_s^{(\alpha)}(2) = 0$, $0 \leqslant \alpha$, $s \leqslant q-1$, $\omega_s^{(\alpha)} \equiv \frac{d^\alpha \omega_s}{dt^\alpha}$);

3) $\omega_s^{(\alpha)}(1) = \delta_{\alpha s}$, $0 \leqslant \alpha$, $s \leqslant q-1$;

4) $$\sum_{s=0}^{q-1} \frac{\omega_s(t)}{(q-s)!} = \frac{t^q}{q!}, \quad 0 \leqslant t \leqslant 1,$$

$$\omega_\gamma(t+1) + \sum_{s=0}^{\gamma} \frac{\omega_s(t)}{(\gamma-s)!} = \frac{t^\gamma}{\gamma!}, \quad 0 \leqslant \gamma \leqslant q-1, \quad 0 \leqslant t \leqslant 1.$$

Функции $\omega_s(t)$ называют исходными, а всю совокупность исходных функций $\omega_s(t)$, $s = 0 \div (q-1)$,—исходной системой. Способ построения функций $\omega_s(t)$, удовлетворяющих условиям 1) — 4) и являющихся полиномами степени не выше $2s-1$, подробно описан в [67] для произвольного значения q. Более того, для $q = 1 \div 6$ указан явный вид $\omega_s(t)$, $s = 0, 1, \ldots, q-1$, причем каждая функция — полином степени $2q-1$. Например, при $q = 1$ ($s = 0$)

$$\omega_0(t) = \begin{cases} t, & 0 \leqslant t \leqslant 1, \\ 2-t, & 1 \leqslant t \leqslant 2; \end{cases}$$

при $q = 2$ $(s = 0,\ 1)$

$$\omega_0(t) = \begin{cases} t^2(3-2t), & 0 \leqslant t \leqslant 1, \\ (2-t)^2(2t-1), & 1 \leqslant t \leqslant 2; \end{cases}$$

$$\omega_1(t) = \begin{cases} t^2(t-1), & 0 \leqslant t \leqslant 1, \\ (2-t)^2(t-1), & 1 \leqslant t \leqslant 2. \end{cases}$$

Вне интервала (0, 2) функции $\omega_s(t)$ согласно условию 2) полагают равными нулю.

Координатная система, т. е. базисные функции МКЭ для решения вариационной задачи (V.45),строятся из исходной системы по формуле

$$\varphi_{sj}^h(x) = \omega_s\left(\frac{x}{h} - j\right), \quad s = 0,\ 1,\ \ldots,\ q-1, \qquad \text{(V.46)}$$

где $h = 1/2k$ — шаг равномерной сетки, покрывающей отрезок [0, 1], а j — целое число в пределах $-1 \leqslant j \leqslant 2k-1$.

Такая координатная система обеспечивает решение задачи эрмитовой интерполяции: построить функцию $\tilde{u}(x)$, которая вместе со своими производными до порядка $q-1$ включительно совпадает с функцией $u(x)$ и ее соответствующими производными в заданных точках $(j+1)h$.

Легко видеть, что координатные функции (V.46) при $q = 1$ $(s = 0)$

$$\varphi_{0j}^h(x) \equiv \varphi_j^h(x) = \begin{cases} \dfrac{x - jh}{h}, & jh \leqslant x \leqslant (j+1)h, \\ \dfrac{(2+j)h - x}{h}, & (j+1)h \leqslant x \leqslant (j+2)h, \\ 0, & x \in [0, jh] \cup [(j+2)h,\ 1], \end{cases}$$

где $-1 \leqslant j \leqslant 2k-1$, совпадают с выписанными ранее кусочно-линейными базисными функциями МКЭ (см. (V.23)), если в (V.23) отрезок [0, 1] разбит на четное число $N = 2k$ равных элементов, длина которых $x_i - x_{i-1} = h = \frac{1}{N}$. Аналогичное заключение справедливо для функций (V.46) при $q = 2$ $(s = 0,\ 1)$

$$\varphi_{0j}^h(x) \equiv \varphi_j^h(x),$$

$$\varphi_{1j}(x) \equiv \psi_j^h(x), \quad j = -1,\ 2,\ \ldots,\ 2k-1,$$

и кусочно-кубических базисных функций (V.29)

$$\varphi_i^N(x), \quad \psi_i^N(x), \quad i = 1,\ 2,\ \ldots,\ N,$$

при $N = 2k$, $h_i \equiv x_i - x_{i-1} = h = 1/N$.

Координатная система (V.46), как доказано в [67], полна в $W_r^q(0, 1)$, т. е. любую функцию $u(x) \in W_r^q(0, 1)$ можно с любой точностью аппроксимировать в метрике пространства $W_r^q(0, 1)$ функцией вида

$$u_l(x) = \sum_{s=0}^{q-1} \sum_{j=-1}^{2k-1} a_{sj}\omega_s\left(\frac{x}{h} - j\right) \equiv \sum_{s=0}^{q-1} \sum_{j=-1}^{2k-1} a_{sj}\varphi_{sj}^h(x),$$

где числовые коэффициенты $a_{sj} = h^s u^{(s)}((j+1)h)$. Порядок аппроксимации в зависимости от степени гладкости аппроксимируемой функции устанавливает следующая теорема.

Теорема V.8. *Пусть* $u(x) \in W_r^{n+1}(0, 1)$. *Если в качестве аппроксимирующей функции выбрать*

$$u_I(x) = \sum_{s=0}^{q-1} \sum_{j=-1}^{2k-1} h^s u^{(s)}((j+1)h)\, \varphi_{sj}^h(x),$$

где $q \leqslant n \leqslant 2q-1$, *то справедлива оценка*

$$\|u - u_I\|_{r,\bar{q}} \leqslant c \|u\|_{r,n+1} h^{n+1-\bar{q}}, \quad c = \text{const}, \quad 0 \leqslant \bar{q} \leqslant q.$$

Отметим, что аналогичные результаты имеют место и для координатных функций многих переменных (см. [67]).

Таким образом, описанные координатные функции (V.46) вполне обосновано можно использовать для получения приближенного по Ритцу решения задачи о минимуме функционала (V.45), где $D(F) = \overset{0}{W}{}_r^1(0, 1)$, $1 < r < \infty$, т. е. для построения приближенного решения МКЭ.

3. Оценка погрешности приближенного решения МКЭ. Предположим, что существует единственная функция $u^*(x) \in W_r^q(0, 1)$, доставляющая минимум функционалу $F(u)$ (V.45). Тогда приближенное решение МКЭ данной вариационной задачи можно искать в виде разложения

$$u^N = \sum_{s=0}^{q-1} \sum_{j=-1}^{N-1} c_{sj} \varphi_{sj}^h(x), \quad N = 2k, \quad h = \frac{1}{N}, \tag{V.47}$$

числовые коэффициенты c_{sj} которого определяются из условия минимума функционала $F(u^N)$ на подпространстве $P^h \subset W_r^q(0, 1)$, где базисом служат функции $\varphi_{sj}^h(x)$ (см. (V.46)); размерность P^h равна $(N+1)q$.

Заметим, что в силу свойств базисных функций (V.46) и условия 3) для $\omega_s(t)$ приближенное решение $u^N(x)$ будет удовлетворять граничным условиям $u^N(0) = u^N(1) = 0$, если в разложении (V.47) $c_{0,-1} = c_{0,N-1} = 0$. Поэтому в базис конечномерного подпространства, на котором минимизируется $F(u^N)$, можно не включать координатные функции $\varphi_{0,-1}^h(x)$ и $\varphi_{0,N-1}^h(x)$. Такое конечномерное подпространство с «укороченным» базисом будем обозначать через $\overset{0}{P}{}^h \subset \subset W_r^q(0, 1) \cap \overset{0}{W}{}_r^1(0, 1)$. Размерность $\overset{0}{P}{}^h$ равна $(N+1)q - 2$. Итак, неизвестные коэффициенты c_{sj} искомого приближенного решения (V.47) должны удовлетворять системе нелинейных алгебраических уравнений

$$\frac{\partial F(u^N)}{\partial c_{sj}} = 0, \quad 0 < s \leqslant q-1, \quad -1 \leqslant j \leqslant N-1; \qquad s = 0, \quad 0 \leqslant j \leqslant N-2. \tag{V.48}$$

В рамках условий теоремы V.5 для возрастающего функционала $F(u)$ система уравнений (V.48) разрешима при любом значении N и дает по формуле (V.47) приближенное решение МКЭ, обеспечивающее абсолютный минимум функционала $F(u)$ на $\overset{0}{P}{}^h$ [10]. Будем обозначать это приближенное решение через $u_*^N(x)$:

$$F(u_*^N) = \min_{u \in \overset{0}{P}{}^h} F(u).$$

Согласно результатам теорем V.5 и V.7 последовательность $\{u_*^N\}$ при $N \to \infty$ является минимизирующей для функционала $F(u)$ и сходится в $\overset{0}{W}{}_r^1(0, 1)$ к некоторому пределу u_*. Поскольку мы предположили, что $D(F) = \overset{0}{W}{}_r^1(0, 1)$, $1 < r < \infty$, то $u_* \in D(F)$ и

$$F(u_*) = \min_{u \in \overset{0}{W}{}_r^1} F(u).$$

Вопрос о точности описанного приближенного решения МКЭ для рассматриваемого класса вариационных задач решается с помощью следующей теоремы.

Теорема V.9. *Пусть функционал $F(u)$, $D(F) = \overset{0}{W}{}_r^1(0, 1)$, $1 < r < \infty$, задачи (V.45) — возрастающий, удовлетворяет условиям теоремы V.5 и, кроме того, неравенству*

$$|F(u) - F(v)| \leqslant K(c)\|u - v\|_{r,1}^{\alpha}, \quad \alpha > 0, \quad K = \text{const}, \tag{V.49}$$

при любых $u, v \in D(F)$ таких, что $\|u\|_{r,1} \leqslant c$, $\|v\|_{r,1} \leqslant c$.

Если при этом доставляющая минимум функционалу $F(u)$ функция $u_*(x) \in W_r^{n+1}(0, 1) \cap \overset{0}{W}{}_r^1(0, 1)$, то для приближенного решения $u_*^N(x) \in \overset{0}{P}{}^h \subset W_r^{n+1} \cap \overset{0}{W}{}_r^1(0, 1)$, полученного при $q = q_0$, $2q_0 - 1 \leqslant n$, справедлива оценка

$$\|u_* - u_*^N\|_{r,1} \leqslant Mh^{\frac{(2q_0-1)\alpha}{2}},$$

где M — постоянная, не зависящая от h.

Доказательство. Положим

$$\rho(u, v) = \left[F(u) + F(v) - 2F\left(\frac{u+v}{2}\right)\right]^{1/2}, \quad \forall u, v \in D(F).$$

Как показано при доказательстве теоремы V.5, выполняется неравенство

$$K_1\|u - v\|_{r,1} \leqslant \rho(u, v), \quad K_1 = \frac{\gamma}{\sqrt{72}} > 0.$$

Пусть $u = u_*$, $v = v_*^N$. Тогда

$$K_1\|u_* - u_*^N\|_{r,1} \leqslant \rho(u_*, u_*^N) = \left[F(u_*) + F(u_*^N) - 2F\left(\frac{u_* + u_*^N}{2}\right)\right]^{1/2}.$$

Поскольку

$$F\left(\frac{u_* + u_*^N}{2}\right) \geqslant F(u_*), \quad F(u_*^N) \leqslant F(u_I),$$

где

$$u_I(x) = \sum_{s=1}^{q_0-1} \sum_{j=-1}^{n-1} h^s u_*^{(s)}((j+1)h)\,\varphi_{sj}^h(x) + \sum_{j=0}^{n-2} u_*((j+1)h)\,\varphi_{0j}^h(x)$$

есть эрмитов интерполянт функции $u_*(x)$ на соответствующем подпространстве $\overset{0}{P}{}^h$, то справедливо неравенство

$$\|u_* - u_*^N\|_{r,1} \leqslant \frac{1}{K_1}[F(u_I) - F(u_*)]^{1/4}. \tag{V.50}$$

Если для построения базиса $\overset{0}{P}{}^h$ натуральное число q_0 выбрано так, что $2q_0 - 1 \leqslant n$, то согласно теореме V.8 для любого такого q_0

$$\|u_* - u_I\|_{r,1} \leqslant K_2 \|u_*\|_{r,2q_0} h^{2q_0-1}. \tag{V.51}$$

Таким образом, из неравенства (V.50) с учетом выполнения условия (V.49) и оценки (V.51) непосредственно следует

$$\|u_* - u_*^N\|_{r,1} \leqslant Mh^{\frac{(2q_0-1)\alpha}{2}},$$

где M — константа, не зависящая от h. Теорема доказана.

Вопрос о точности приближенного решения $u_*^N(x)$ можно решать и на основе следующей теоремы.

Теорема V.10. *Пусть функционал $F(u)$, $D(F) = \overset{0}{W}{}_r^1(0, 1)$, $1 < r < \infty$, задачи (V.45) — возрастающий, удовлетворяет условиям теоремы V.5 и, кроме того, его градиент $A = \operatorname{grad} F$, $D(A) = D(F)$, удовлетворяет для ограниченных аргументов условию Липшица, т. е.*

$$\|Au - Av\|_* \leqslant K(c)\|u - v\|_{r,1} \tag{V.52}$$

при любых $u, v \in D(A)$ таких, что $\|u\|_{r,1} \leqslant c$ и $\|v\|_{r,1} \leqslant c$.

Если $u_*(x) \in W_r^{n+1}(0, 1) \cap \overset{0}{W}{}_r^1(0, 1)$, а $u_*^N(x)$ получено при $q = q_0$, $2q_0 - 1 \leqslant n$, то

$$\|u_* - u_*^N\|_{r,1} \leqslant Mh^{2q_0-1},$$

где постоянная M не зависит от h.

Доказательство. Как и при доказательстве теоремы V.9, легко получаем (V.50)

$$\|u_* - u_*^N\|_{r,1} \leqslant \frac{1}{K_1}[F(u_I) - F(u_*)]^{1/2},$$

где u_I — интерполянт u_* из $\overset{0}{P}{}^h$.

Согласно существующему соотношению (V.41) между функционалом $F(u)$ и его градиентом A в нашем случае справедливо тожде-

ство [64]

$$F(v) - F(u) = \int_0^1 \langle A(u + t(v-u)),\ v-u\rangle\, dt \qquad \text{(V.53)}$$

при любых $u, v \in D(A) = D(F)$. Положим теперь $u = u_*$, $v = u_l$. Поскольку $F(u_*) = \min\limits_{u \in D(F)} F(u)$, имеем $Au_* = 0$ и соотношение (V.53) можно переписать в виде

$$F(u_l) - F(u_*) = \int_0^1 \langle A(u_* + t(u_l - u_*)) - Au_*,\ u_l - u_*\rangle\, dt.$$

Так как

$$|\langle A(u_* + t(u_l - u_*)) - Au_*,\ u_l - u_*\rangle| \leqslant$$
$$\leqslant \| A(u_* + t(u_l - u_*)) - Au_* \|_* \| u_l - u_* \|_{r,1},$$

то с учетом (V.52) имеем

$$|F(u_l) - F(u_*)| \leqslant K(c)/2 \| u_l - u_* \|_{r,1}^2. \qquad \text{(V.54)}$$

Учитывая (V.50), (V.54) и оценку (V.51), окончательно получаем

$$\| u_* - u_*^N \|_{r,1} \leqslant M h^{2q_0 - 1},$$

где M — постоянная, не зависящая от h, что и требовалось доказать. Как следует из формулировки и доказательства теорем V.9 и V.10, их результаты будут справедливы и в случае функционалов, более общих, чем (V.45), но подчиняющихся аналогичным условиям.

4. Численные примеры. Пример 1. Рассмотрим подробно численное решение нелинейной вариационной задачи методом конечных элементов в случае отыскания минимума функционала

$$F(u) = \int_0^1 \left[\left(\frac{du}{dx} \right)^2 + u^2 + u^4 - f(x)\, u \right] dx, \qquad \text{(V.55)}$$

$$f(x) = 2[2 + x(1-x) + 2x^3 (1-x)^3],$$

на множестве функций, удовлетворяющих условию

$$u(0) = u(1) = 0; \qquad \text{(V.56)}$$

точное решение задачи $u_T = x(x-1)$.

Для этого вначале исследуем свойства данного функционала, чтобы убедиться в существовании искомой функции $u_*(x)$, доставляющей минимум функционалу $F(u)$, и в теоретически обоснованном применении МКЭ (см. теоремы V.5, V.7, V.10).

Пусть $D(F) = \overset{0}{W}{}_2^1(0, 1)$, в котором норму определим равенством

$$\| u \|_{2,1}^{(0)} = \left(\int_0^1 \left(\frac{du}{dx} \right)^2 dx \right)^{1/2}.$$

(При описании примера 1 эту норму будем обозначать так: $\| \ \|$.)

Напомним, что функции $u(x) \in \overset{0}{W}{}_2^1(0, 1)$ абсолютно непрерывны на отрезке [0, 1] и для них справедливо неравенство

$$\int_0^1 u^2 dx \leqslant \frac{1}{2} \int_0^1 \left(\frac{du}{dx}\right)^2 dx.$$

1. Докажем, что функционал $F(u)$ непрерывен в пространстве $\overset{0}{W}{}_2^1(0, 1)$. Для этого достаточно убедиться в непрерывности функционала

$$\Phi(u) = \int_0^1 \left[\left(\frac{du}{dx}\right)^2 + u^2 + u^4\right] dx.$$

Пусть $u_n, u \in \overset{0}{W}{}_2^1(0, 1)$ и $\| u_n - u \| \to 0$ при $n \to \infty$. Тогда

$$|\Phi(u_n) - \Phi(u)| \leqslant \left| \int_0^1 \left(\frac{du_n}{dx} - \frac{du}{dx}\right)\left(\frac{du_n}{dx} + \frac{du}{dx}\right) dx \right| +$$

$$+ \left| \int_0^1 (u_n - u)(u_n + u)\, dx \right| + \left| \int_0^1 (u_n - u)(u_n + u)(u_n^2 + u^2)\, dx \right|. \quad \text{(V.57)}$$

Рассмотрим последний интеграл:

$$I_3 = \left| \int_0^1 (u_n - u)(u_n + u)(u_n^2 + u^2)\, dx \right|.$$

Поскольку $\| u_n - u \| \to 0$ при $n \to \infty$, то для достаточно большого значения n, $\| u_n \| \leqslant c_1 \| u \|$, $c_1 = \text{const}$, и согласно теореме вложения Соболева при $x \in [0, 1]$ справедливо неравенство $u_n^2(x) + u^2(x) \leqslant c \| u \|^2$, $c = \text{const}$. (В дальнейшем все постоянные будем обозначать буквой c, хотя они различны по значению.)

Теперь легко получить оценку интеграла

$$I_3 \leqslant c \| u \|^2 \int_0^1 |u_n - u\ \ u_n + u|\, dx \leqslant c \| u \|^2 \left(\int_0^1 (u_n - u)^2\, dx \right)^{1/2} \times$$

$$\times \left(\int_0^1 (u_n + u)^2\, dx \right)^{1/2} \leqslant \frac{c}{2} \| u \|^2 \left(\int_0^1 \left(\frac{du_n}{dx} - \frac{du}{dx}\right)^2 dx \right)^{1/2} \times$$

$$\times \left(\int_0^1 \left(\frac{du_n}{dx} + \frac{du}{dx}\right)^2 dx \right)^{1/2} = c \| u_n - u \| \| u_n + u \|$$

и аналогично — остальных интегралов неравенства (V.57). В результате находим

$$|\Phi(u_n) - \Phi(u)| \leqslant c \| u_n - u \| \| u_n + u \|.$$

Так как норма $\|u_n + u\|$ ограничена при $\|u_n - u\| \to 0$, последнее неравенство доказывает непрерывность $\Phi(u)$, а значит, и $F(u)$ в $\overset{0}{W}{}_2^1(0, 1)$.

2. Рассмотрим теперь вопрос о существовании и свойствах градиента A исследуемого функционала $F(u)$. Для $\forall\, u, v \in D(F) = \overset{0}{W}{}_2^1(0, 1)$ из непосредственных подсчетов получаем

$$F'_u \equiv \frac{dF(u+tv)}{dt}\Big|_{t=0} = 2\int_0^1 \left(\frac{du}{dx}\frac{dv}{dx} + uv + 2u^3v\right)dx - \int_0^1 fv\,dx.$$

Для доказательства того, что $\frac{dF(u+tv)}{dt}\Big|_{t=0}$ при любом фиксированном элементе $u \in \overset{0}{W}{}_2^1(0, 1)$ есть линейный ограниченный функционал от $v \in \overset{0}{W}{}_2^1(0, 1)$, достаточно доказать это для выражения

$$\Phi(u, v) = \int_0^1 \left(\frac{du}{dx}\frac{dv}{dx} + uv + 2u^3v\right)dx.$$

С помощью рассуждений, аналогичных приведенным в п. 1, нетрудно получить оценку

$$|\Phi(u, v)| \leqslant c\|u\|\|v\|,$$

которая свидетельствует, что $\Phi(u, v)$ есть линейный ограниченный функционал от $v \in \overset{0}{W}{}_2^1(0, 1)$ при любом фиксированном элементе $u \in \overset{0}{W}{}_2^1(0, 1)$. Из ограниченности линейного функционала $\frac{d}{dt}F(u + tv)|_{t=}$ следует существование оператора A с $D(A) = D(F) = \overset{0}{W}{}_2^1(0, 1)$, который является градиентом функционала $F(u)$, $A = \operatorname{grad} F(u)$, и определяется равенством

$$\langle F'_u, v\rangle \equiv \langle Au, v\rangle = 2\int_0^1 \left(\frac{du}{dx}\frac{dv}{dx} + uv + 2u^3v - \frac{1}{2}fv\right)dx. \quad \text{(V.58)}$$

Убедимся, что оператор $A = \operatorname{grad} F(u)$ имеет производную A_u. Пусть $w \in D(A) \equiv \overset{0}{W}{}_2^1(0, 1)$. Тогда

$$\frac{d}{dt}\langle A(u + tw), v\rangle|_{t=0} \equiv \langle A'_u w, v\rangle =$$

$$= 2\int_0^1 \left(\frac{dw}{dx}\frac{dv}{dx} + wv + 6u^2wv\right)dx \quad \text{(V.59)}$$

и для любого фиксированного элемента $u \in \overset{0}{W}{}_2^1(0, 1)$ и $\forall\, w, v \in \overset{0}{W}{}_2^1(0, 1)$ легко получить

$$|\langle A'_u w, v\rangle| \leqslant c\|w\|\|v\|,$$

откуда следует, что $\langle A'_u w, v\rangle$ есть билинейный функционал над v и w, ограниченный в $\overset{0}{W}{}_2^1(0, 1)$.

Таким образом, производная A'_u существует при $\forall\, u \in \overset{0}{W}{}_2^1(0, 1)$ и $D(A_u)$ совпадает с $\overset{0}{W}{}_2^1 = D(F) = D(A)$. Оператор A'_u определяется выражением (V.59). Для A'_u при $\forall\, v \in \overset{0}{W}{}_2^1(0, 1)$ справедлива оценка

$$\langle A'_u v, v\rangle \geqslant \gamma^2\|v\|^2, \quad \gamma > 0 - \text{const},$$

что непосредственно следует из (V.59):

$$\langle A'_u v, v\rangle \geqslant 2\int_0^1 \left(\frac{dv}{dx}\right)^2 dx = 2\|v\|^2.$$

3. Покажем, что функционал $F(u)$ — возрастающий в принятой норме пространства $\overset{0}{W}{}_2^1(0, 1)$. Действительно,

$$F(u) = \int_0^1 \left[\left(\frac{du}{dx}\right)^2 + u^2 + u^4 - fu\right] dx \geqslant \int_0^1 \left(\frac{du}{dx}\right)^2 dx -$$

$$- \int_0^1 fu\,dx \geqslant \|u\|^2 - \left(\int_0^1 f^2 dx\right)^{1/2}\left(\int_0^1 u^2 dx\right)^{1/2} \geqslant \|u\|^2 - \left(\int_0^1 f^2 dx\right)^{1/2}\frac{\|u\|}{\sqrt{2}},$$

т. е. $F(u) \to \infty$ при $\|u\| \to \infty$.

В то же время если $\|u\| \leqslant M$, то легко получить оценку

$$|F(u)| \leqslant \left(\frac{3}{2} + c\right)\|u\|^2 + \frac{1}{\sqrt{2}}\left(\int_0^1 f^2 dx\right)^{1/2}\|u\|,$$

откуда следует ограниченность $|F(u)|$ при ограниченности $\|u\|$, т. е. функционал $F(u)$ — возрастающий.

4. В заключение докажем, что оператор $A = \operatorname{grad} F(u)$, определяемый выражением (V.58), удовлетворяет условию Липшица при ограниченных аргументах.

Пусть $u, w \in D(A) = \overset{0}{W}{}_2^1(0, 1)$ и $\|u\| \leqslant M$, $\|w\| \leqslant M$. Тогда согласно (V.58)

$$|\langle Au - Aw, v\rangle| \leqslant 2\left(\left|\int_0^1 \left(\frac{du}{dx} - \frac{dw}{dx}\right)\frac{dv}{dx}\,dx\right| +\right.$$

$$+\left|\int_0^1 (u-w)\,v\,dx\right|+2\left|\int_0^1 (u^3-w^3)\,v\,dx\right|\Bigg).$$

Повторяя рассуждения, аналогичные приводимым выше, легко получить оценку

$$|\langle Au-Aw, v\rangle| \leqslant 2\|u-w\|\|v\|+\|u-w\|\|v\|+ \\ +4c(M)\|u-w\|\|v\|, \qquad \text{(V.60)}$$

Таблица 19

x_i	u_T	$h=0{,}25$		$h=0{,}125$	
		u^N	δ (%)	u^N	δ (%)
0,25	—0,1875	—0,18944	1,0	—0,18798	0,2
0,5	—0,25	—0,25261	1,0	—0,25665	0,2
0,75	—0,1875	—0,18944	1,0	—0,18798	0,2

Примечание. При $h=0{,}25$ значение $t=2{,}98$ с, $s=4$, при $h=0{,}125$ значение $t=3{,}43$ с, $s=4$.

Таблица 20

x_i	u_T	u'_T	u^N	δ (%)	$(u^N)'$	δ (%)
0	0	—1	—	—	—0,9999	0,01
0,25	—0,1875	—0,5	—0,18749	$5\cdot10^{-3}$	—0,4999	0,02
0,5	—0,25	0	—0,24999	$4\cdot10^{-4}$	$0{,}3\cdot10^{-8}$	—
0,75	—0,1875	0,5	—0,18749	$5\cdot10^{-3}$	0,4999	0,02
1	0	1	—	—	0,9999	0,01

Примечание. Здесь $h=0{,}125$, $t=3{,}43$, $\overline{s}=4$.

Таблица 21

x_i	u_T	$h=0{,}25$		$h=0{,}125$		$h=0{,}0625$	
		u^N	δ (%)	u^N	δ (%)	u^N	δ (%)
0,25	—4,5802	—4,6301	1,09	—4,5935	0,29	—4,5832	0,066
0,5	—6,5684	—7,0256	6,93	—6,7051	2,08	—6,6019	0,511
0,75	—4,7635	—4,8158	1,09	—4,7811	0,36	—4,7686	0,065

Примечание. При $h=0{,}25$ значение $\widetilde{u}^0=[-4{,}6,\ -6{,}8,\ -4{,}6]^T$ число итераций $s=7$; при $h=0{,}125$ значение $\widetilde{u}^0=[-2{,}6,\ -4{,}6,\ -6{,}0,\ -6{,}8,\ -6{,}0,\ -4{,}6,\ -2{,}6]^T$, $s=9$; при $h=0{,}0625$ значение $\widetilde{u}^0=[-1{,}4,\ -2{,}6,\ -3{,}7,\ -4{,}6,\ -5{,}4,\ -6{,}0,\ -6{,}5,\ -6{,}8,\ -6{,}5,\ -6{,}0,\ -5{,}4,\ -4{,}6,\ -3{,}7,\ -2{,}6,\ -1{,}4]^T$, $s=12$; время решения всех трех систем $t=12{,}26$ с.

x_j	u'_T	$h = 0{,}25$			
		u^N	δ (%)	$(u^N)'$	δ (%)
0	—20,2585	—	—	—20,2877	0,14
0,25	—14,5546	—4,5807	0,012	—14,5535	0,007
0,5	—0,5037	—6,5725	0,062	—0,4811	4,49
0,75	14,3540	—4,7627	0,017	14,3355	0,13
1	22,0457	—	—	22,0745	0,13

Примечание. При всех h значение $\tilde{u}^0 = u^0 = 0$; при $h = 0{,}25$ число итераций $s = 16$, время $t = 2$ мин 25, 62 с.

где $c(M)$ — константа, значение которой зависит от заданного значения M. Поскольку, как уже упоминалось ранее,

$$\|Au - Aw\|_* = \sup_{v \neq 0} \frac{|\langle Au - Aw, v\rangle|}{\|v\|},$$

из (V.60) непосредственно следует

$$\|Au - Aw\|_* \leqslant (3 + 4c(M))\|u - w\|,$$

что и требовалось доказать.

Проведенные рассмотрения позволяют вполне обоснованно применять МКЭ для вычисления приближенного решения задачи (V.55), (V.56). Для этого использовались кусочно-линейные и кусочно-кубические эрмитовы базисные функции, подробно описанные в п. 2 параграфа V.2.

Согласно теореме V.10 если точное решение $u_*(x) \in W_2^2(0, 1)$, то использование линейных базисных функций ($q_0 = 1$, так как $n = 1$) обеспечит скорость сходимости приближенных решений u_*^N в норме $\|\ \|_{2,1}$ порядка h; а при $u_*(x) \in W_2^4(0, 1)$ и использовании кубических эрмитовых базисных функций ($q_0 = 2$, $n = 3$) скорость сходимости в той же норме $\|\ \|_{2,1}$ будет порядка h^3. Дискретизация задачи (V.55), (V.56) методом конечных элементов приводит к системе нелинейных алгебраических уравнений с симметричной матрицей Якоби.

Решение данной системы выполнялось квазиньютоновским методом, идея которого изложена в п. 3 параграфа V.1.

Система решалась на ЭВМ ЕС 1060 для различных h с двойным машинным словом. Критерием окончания итерационного процесса служило условие

$$\|u^k - u^{k-1}\|_E \leqslant \varepsilon, \quad \|F(u^k)\|_E \leqslant \varepsilon_1.$$

Полученные результаты для некоторых значений h приведены в табл. 19 (линейные базисные функции) и 20 (кубические). В качестве начального приближения к решению везде выбирался нулевой вектор, кроме того, $\tilde{u}^{(0)} = u^0 = [0, \ldots, 0]^T$; итерационный процесс оканчивался при $\varepsilon_1 = 10^{-5}$ и $\varepsilon = 5 \cdot 10^{-5}$.

Таблица 22

$h = 0{,}125$				$h = 0{,}0625$			
u^N	δ (%)	$(u^N)'$	δ (%)	u^N	δ (%)	$(u^N)'$	δ (%)
—	—	—20,2609	0,012	—	—	—20,2597	0,006
—4,5802	$6 \cdot 10^{-4}$	—14,5536	0,006	—4,5805	0,006	—14,5561	0,013
—6,5685	0,001	—0,4975	1,24	—6,5695	0,017	—0,5099	1,23
—4,7634	0,001	14,3512	0,019	—4,7641	0,013	14,3564	0,016
—	—	22,0490	0,016	—	—	22,0489	0,014

(процессорное) решения системы $t = 9{,}83$ с; при $h = 0{,}125$ $s = 33$, $t = 24{,}05$ с; при $h = 0{,}0625$ $s = 199$,

В табл. 19 и 20 использовались обозначения, принятые в п. 3 параграфа V.1.

Пример 2. Найти функцию $y(x)$, минимизирующую функционал

$$F(y) = \int_0^1 \left[\frac{1}{48} \left(\frac{dy}{dx} - 1 \right)^4 + \left(\frac{dy}{dx} \right)^2 + y^2 - fy \right] dx \qquad (V.61)$$

при условии, что

$$y(0) = y(1) = 0. \qquad (V.62)$$

Здесь

$$f(x) = -\frac{81}{4} \frac{\sin 3x}{\sin 3} \left(\frac{\cos^2 3x}{\sin^2 3} + \frac{80}{81} \right) + 2x.$$

Единственное точное решение задачи $y_T = x - \dfrac{\sin 3x}{\sin 3}$.

Приближенное решение задачи (V.61), (V.62) вычислялось процессом Ритца с базисными функциями МКЭ (линейные и кубические полиномы Эрмита). Система нелинейных алгебраических уравнений дискретной задачи решалась квазиньютоновским методом на ЭВМ ЕС 1060 с двойным машинным словом. Начальное приближение u^0 везде нулевое. Условия окончания итерационного процесса такие же, как для предыдущего примера. Результаты счета приведены в табл. 21 (линейные базисные функции) и 22 (кубические). Все обозначения такие же, как в табл. 16—20.

Отметим, что в случае кубических базисных функций на точность вычисленного решения u^N при $h = 0{,}0625$ (см. табл. 22) заметное влияние оказывают уже и ошибки округления, а не только погрешность МКЭ.

Глава VI

ЧИСЛЕННОЕ РЕШЕНИЕ НЕКОТОРЫХ ПРИКЛАДНЫХ ЗАДАЧ

В настоящей главе будет показано, как методы, описанные в предыдущих главах, могут успешно применяться на практике. С этой целью отобрано несколько решенных авторами и их сотрудниками прикладных задач, рассмотрение которых позволяет увидеть теоретические и практические трудности, встречающиеся при решении на ЭВМ реальных задач.

В одной из задач исследуется напряженно-деформированное состояние толстых цилиндрических оболочек вращения, регулярно подкрепленных кольцевыми ребрами жесткости и находящихся под всесторонним давлением. Материал ребер и оболочки может быть ортотропен и различен. Соединение оболочки с ребрами может быть как жестким. так и «скользящим». Такие задачи описываются уравнениями упругого равновесия тел в перемещениях, записанных в цилиндрической системе координат, причем в случае скользящего соединения ребер с оболочкой на участке их сопряжения допускается разрыв в одной из компонент решения. Их называют смешанными краевыми задачами (контактными, с односторонними связями).

В другой задаче определяются частоты и формы собственных колебаний компрессорных лопаток турбомашин. Она относится к классу задач на собственные значения некоторых дифференциальных операторов. Рассматривая лопатку как стержень переменного сечения, для различных видов колебаний (изгибных, крутильных и т. д.) получаем задачи на собственные значения для обыкновенных дифференциальных операторов. Если рассматривать лопатку как пластину или оболочку переменной толщины, то ставится задача на собственные значения для дифференциальных операторов в частных производных в двумерной области.

Одной из важных прикладных задач является расчет упруго-пластического состояния панели летательного аппарата, находящегося под силовым и температурным воздействием. Решение такой задачи может быть найдено путем минимизации функционала полной энергии системы, записанного в перемещениях. Вследствие нелинейности физических соотношений напряжений и деформаций рассматриваемый функционал не квадратичен.

VI.1. Исследование напряженно-деформированного состояния толстой цилиндрической оболочки, подкрепленной ребрами жесткости

1. Постановка задачи [19]. Пусть требуется определить напряженно-деформированное состояние достаточно длинной осесимметрично загруженной однородной внешней нагрузкой толстой ортотропной цилиндрической оболочки вращения, регулярно подкрепленной ортотропными кольцевыми ребрами (на рис. 28 изображена половина сечения конструкции плоскостью, содержащей ось вращения z).

На значительном расстоянии от концов конструкции двумя плоскостями, перпендикулярными оси вращения (одна из которых проходит через середину ребра, а вторая — через середину пролета между ребрами), выделяем элемент (на рис. 28 заштрихованная часть). Уравнения равновесия линейной теории упругости, записанные в цилиндрической системе координат (r, φ, z), для этого элемента имеют вид [57]

$$-\left[a_{11}\frac{\partial}{\partial r}\left(r\frac{\partial u_r}{\partial r}\right)-a_{33}\frac{u_r}{r}+a_{44}\frac{\partial}{\partial z}\left(r\frac{\partial u_r}{\partial z}+r\frac{\partial u_z}{\partial r}\right)+\right.$$
$$\left.+a_{12}\frac{\partial}{\partial r}\left(r\frac{\partial u_z}{\partial z}\right)-a_{23}\frac{\partial u_z}{\partial z}\right]=f_1(r, z), \qquad \text{(VI.1)}$$
$$-\left[a_{12}\frac{\partial}{\partial z}\left(r\frac{\partial u_r}{\partial r}\right)+a_{22}\frac{\partial}{\partial z}\left(r\frac{\partial u_z}{\partial z}\right)+a_{23}\frac{\partial u_r}{\partial z}+\right.$$
$$\left.+a_{44}\frac{\partial}{\partial r}\left(r\frac{\partial u_r}{\partial z}+r\frac{\partial u_z}{\partial r}\right)\right]=f_2(r, z),$$

где u_r и u_z — перемещения по осям Or и Oz, область D лежит в плоскости (r, z), $D = D_1 \cup D_2$ (рис. 29), a_{ij} — упругие постоянные (в общем случае различны для D_1 и D_2, так как ребра и оболочка изготовлены из разных однородных ортотропных материалов).

Краевые условия на $\Gamma\left(\Gamma = \bigcup_{i=1}^{7}\Gamma_i\right)$ задаются следующим образом:

$$u_z = g(r), \quad (r, z)\in\Gamma^{(1)}, \qquad \text{(VI.2)}$$
$$\sigma_{rz} = g_1(r), \quad (r, z)\in\Gamma^{(1)}, \qquad \text{(VI.3)}$$
$$t(u) = q(r, z), \quad (r, z)\in\Gamma^{(2)}, \qquad \text{(VI.4)}$$

где $t(u)$ — вектор напряжений, $u = [u_r, u_z]^T$, $\Gamma = \Gamma^{(1)} \cup \Gamma^{(2)}$, $\Gamma^{(1)} = \Gamma_1 \cup \Gamma_2 \cup \Gamma_4$, $\Gamma^{(2)} = \Gamma_3 \cup \Gamma_5 \cup \Gamma_6 \cup \Gamma_7$; σ_{rr}, σ_{rz}, ...— компоненты тензора напряжений.

Условия на Γ_8 (участке соединения ребра с оболочкой) задаются в зависимости от того, как связаны ребра с оболочкой. Например, в случае жесткого их соединения (задача с разрывными коэффициентами)

$$[u_r]|_{\Gamma_8} = [u_z]|_{\Gamma_8} = 0, \qquad \text{(VI.5)}$$
$$[\sigma_{rr}]|_{\Gamma_8} = [\sigma_{rz}]_{\Gamma_8} = 0, \qquad \text{(VI.6)}$$

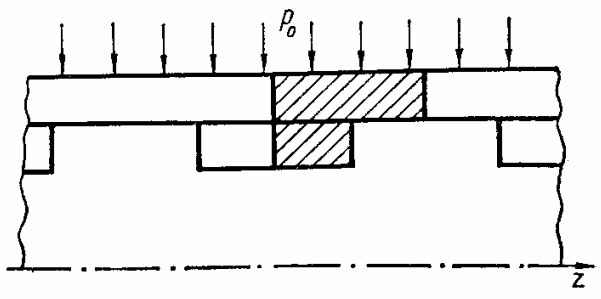

Рис. 28.

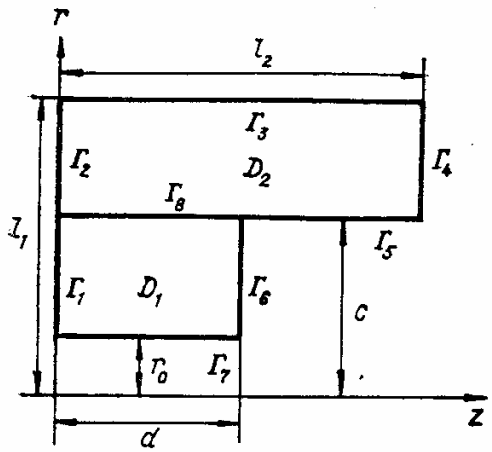

Рис. 29.

где

$$[\psi]|_{\Gamma_8} = \psi(c+0, z) - \psi(c-0, z),$$

$$0 \leqslant z \leqslant d, \quad (\psi(c+0, z) \equiv \psi^+(c, z),$$

$$\psi(c-0, z) \equiv \psi^-(c, z), \quad 0 \leqslant z \leqslant d),$$

а в случае скользящего соединения (задача, где дополнительно допускается разрыв в одной из компонент решения)

$$[u_r]|_{\Gamma_8} = 0, \qquad \text{(VI.7)}$$

$$[\sigma_{rr}]|_{\Gamma_8} = 0, \quad \sigma_{rz}^+|_{\Gamma_8} = \sigma_{rz}^-|_{\Gamma_8} = 0. \qquad \text{(VI.8)}$$

Уравнения (VI.1) запишем в операторном виде

$$LU = f,$$

где

$$f = [f_1(r, z), f_2(r, z)]^T.$$

Для каждой из рассматриваемых задач нетрудно проверить, что оператор L является симметричным и положительно определенным на множестве M_0 дважды непрерывно дифференцируемых в каждой из областей $\bar{D}_1$, $\bar{D}_2$ вектор-функций U, удовлетворяющих однородным условиям (VI.2) — (VI.4) и условиям (VI.5), (VI.6) или (VI.7), (VI.8).

Каждой из задач (VI.1), (VI.2) и (VI.1) — (VI.4), (VI.7), (VI.8) соответствует вариационная задача: определить вектор-функцию $U = [u_r, u_z]^T$, компоненты которой на каждой из областей D_1, D_2 принадлежат пространству $W_2^1(D_k)$, $k = 1, 2$, и которая реализует минимум функционала

$$I(U) = \iint\limits_D \left[a_{11}\left(\frac{\partial u_r}{\partial r}\right)^2 + a_{44}\left(\frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r}\right)^2 + \right.$$

$$+ a_{33}\frac{u_r^2}{r^2} + a_{22}\left(\frac{\partial u_z}{\partial z}\right)^2 + 2a_{12}\frac{\partial u_r}{\partial r}\frac{\partial u_z}{\partial z} +$$

$$\left. + 2a_{13}\frac{u_r}{r}\frac{\partial u_r}{\partial r} + 2a_{23}\frac{u_r}{r}\frac{\partial u_z}{\partial z}\right] r d\Omega -$$

$$- 2\iint\limits_D (f_1 u_r + f_2 u_z)\, d\Omega - 2\int\limits_{\Gamma^{(1)}} r u_r g_1 \cos(n, z)\, d\Gamma -$$

$$- 2\int\limits_{\Gamma^{(2)}} r(q_1 u_r + q_2 u_z)\, d\Gamma \quad (d\Omega = dr dz) \qquad \text{(VI.9)}$$

с учетом того, что выполнены соответствующие главные условия: (VI.2), (VI.5) или (VI.2), (VI.7).

Определив скалярное произведение двух вектор-функций $\varphi = [\varphi_1, \varphi_2]^T$, $\psi = [\psi_1, \psi_2]^T$ по формуле

$$(\varphi, \psi) = \iint\limits_D (\varphi_1\psi_1 + \varphi_2\psi_2)\, d\Omega,$$

перепишем (VI.9) в виде

$$l(U) = F(U) - 2(f, U) - 2l(U), \tag{VI.10}$$

где $F(U)$ — квадратичный функционал из (VI.9), $l(U)$ — линейный функционал из (VI.9), учитывающий естественные краевые условия.

Заметим, что все проведенные рассуждения справедливы также для случая, когда $d = l_2$, т. е. $\Gamma_5 = \varnothing$ (рис. 30).

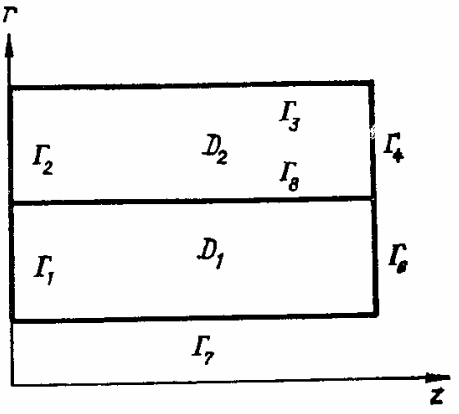

Рис. 30.

2. Дискретизация задачи. Область $\bar{D}$ (см. рис. 29, 30) триангулируется конечным числом прямоугольных треугольников, а именно: каждая из областей $\bar{D}_1$, $\bar{D}_2$ линиями, параллельными осям координат, разбивается на прямоугольники (вершины прямоугольников на Γ_8 общие для $\bar{D}_1$ и $\bar{D}_2$) и далее каждый из них диагональю, образующей с осью Oz угол, больший $\pi/2$, разбивается на два прямоугольных треугольника. В общем случае сетка триангуляции не равномерная.

Вершины и центры тяжести треугольников T_i будем называть узловыми точками. На Γ_8 в каждой вершине два узла.

Пусть на каждом треугольнике $\bar{T}_i \left(\bar{D} = \bigcup_{i=1}^{N} \bar{T}_i\right)$ приближенное решение, т. е. вектор-функция $U^N = [u_r^N, u_z^N]$ является полным кубическим полиномом Эрмита (см. п. 2 параграфа I.3; рис. 5, *б*).

Для однозначного определения на T_i вектор-функции $U^N(r, z)$, т. е. ее десяти (для каждой компоненты) числовых коэффициентов, фиксируются ее значения во всех узловых точках и значения ее первых частных производных в вершинах треугольника.

Непрерывность функции $U^N(r, z)$ на $\bar{D}_1$, $\bar{D}_2$ достигается за счет приравнивания одноименных фиксированных параметров в общих вершинах треугольников триангуляции, а выполнение условия (VI.5) задачи (VI.1) — (VI.6) обеспечивается требованием

$$(U^N)^+ = (U^N)^-, \quad \left(\frac{\partial U^N}{\partial z}\right)^+ = \left(\frac{\partial U^N}{\partial z}\right)^-. \tag{VI.11}$$

Если решается задача (VI.1) — (VI.4), (VI.7), (VI.8), то на Γ_8 требуется выполнение условий (VI.11) только для первой компоненты U_r^N вектор-функции $U^N(r, z)$.

Для каждой из рассмотренных задач в узловых точках, лежащих на $\Gamma^{(1)}$, требуется, чтобы вторая компонента u_z^N допустимой вектор-функции U^N удовлетворяла условиям

$$u_z^N = g(r), \quad \frac{\partial u_z^N}{\partial r} = \frac{\partial g}{\partial r}. \tag{VI.12}$$

Система алгебраических уравнений МКЭ строится из элементарных матриц жесткости (в рассматриваемом случае с размерами

20 × 20). Коэффициенты элементарных матриц жесткости можно получать путем вычисления соответствующих интегралов и последующего матричного умножения. В большинстве случаев затраты времени можно существенно уменьшить, если элементарные матрицы жесткости формировать из блоков, например, с размерами 10 × 10. Причем эти блоки могут быть одинаковыми для всей области $\bar{D}$.

Рассмотрим следующую процедуру построения элементарной матрицы жесткости. Каждый треугольник T_i с помощью преобразования

$$z = z_1(i) + h_1(i)\,\xi, \quad r = r_1(i) + h_2(i)\,\eta, \tag{VI.13}$$
$$h_1(i) = z_2(i) - z_1(i), \quad h_2(i) = r_3(i) - r_1(i),$$

отобразим на канонический треугольник T_0 в плоскости (ξ, η) с вершинами в точках (0, 0), (1, 0), (0, 1). Отметим, что нумерация вершин $(r_k(i), z_k(i))$, $k = 1, 2, 3$, каждого треугольника T_i в плоскости (r, z) выполняется против часовой стрелки, начиная с вершины прямого угла. Тогда согласно триангуляции области D и с учетом преобразования (VI.13) функционал (VI.9) можно представить в виде

$$I(V^N) = \sum_{i=1}^{N} [I_i(V^N) - 2\delta_i(V^N)], \tag{VI.14}$$

где

$$V^N(\xi, \eta) = [w(\xi, \eta), v(\xi\ \eta)]^T \equiv [u_r^N(r(\eta), z(\xi)),$$
$$u_z^N(r(\eta), z(\xi))]^T, \tag{VI.15}$$

$$I_i(V^N) = \iint_{T_0} \left[a_{11} \frac{r_1}{h_2^2} \left(\frac{\partial w}{\partial \eta}\right)^2 + \frac{a_{11}}{h_2} \eta \left(\frac{\partial w}{\partial \eta}\right)^2 + a_{44} \frac{r_1}{h_1^2} \left(\frac{\partial w}{\partial \xi}\right)^2 + \right.$$
$$+ a_{44} \frac{h_2}{h_1^2} \eta \left(\frac{\partial w}{\partial \xi}\right)^2 + 2 \frac{a_{44} r_1}{h_1 h_2} \frac{\partial w}{\partial \xi} \frac{\partial v}{\partial \eta} + 2 \frac{a_{44}}{h_1} \eta \frac{\partial w}{\partial \xi} \frac{\partial v}{\partial \eta} +$$
$$+ \frac{a_{44} r_1}{h_2^2} \left(\frac{\partial v}{\partial \eta}\right)^2 + \frac{a_{44}}{h_2} \eta \left(\frac{\partial v}{\partial \eta}\right)^2 + a_{33} \frac{w^2}{r_1 + h_2 \eta} + a_{22} \frac{r_1}{h_1^2} \left(\frac{\partial v}{\partial \xi}\right)^2 +$$
$$+ a_{22} \frac{h_2}{h_1^2} \eta \left(\frac{\partial v}{\partial \xi}\right)^2 + 2 a_{12} \frac{r_1}{h_1 h_2} \frac{\partial w}{\partial \eta} \frac{\partial v}{\partial \xi} + 2 \frac{a_{12}}{h_1} \eta \frac{\partial w}{\partial \eta} \frac{\partial v}{\partial \xi} +$$
$$\left. + 2 \frac{a_{13}}{h_2} w \frac{\partial w}{\partial \eta} + 2 \frac{a_{23}}{h_1} w \frac{\partial v}{\partial \xi} \right] h_1 h_2 d\xi d\eta,$$

$$\delta_i(V^N) = \iint_{T_0} (w f_1 + v f_2)\, h_1 h_2 d\xi d\eta + l_i,$$

$$h_k = h_k(i), \quad k = 1, 2, \quad r_1 = r_1(i).$$

Значение $l_i = 0$, если $\text{mes}\,(\bar{T}_i \cap \Gamma) = 0$, в противном случае l_i получаем из (VI.10).

Громоздкое на вид выражение (VI.15) достаточно удобно используется при реализации МКЭ. Действительно, пусть

$$I_i(V^N) = I_i^1 + I_i^2 + I_i^3,$$

где

$$I_i^1 = \frac{h_1}{h_2} a_{11} r_1 \iint\limits_{T_0} \left(\frac{\partial w}{\partial \eta}\right)^2 d\Omega + a_{11} h_1 \iint\limits_{T_0} \eta \left(\frac{\partial w}{\partial \eta}\right)^2 d\Omega +$$

$$+ a_{41} r_1 \frac{h_2}{h_1} \iint\limits_{T_0} \left(\frac{\partial w}{\partial \xi}\right)^2 d\Omega + a_{44} \frac{h_2^2}{h_1} \iint\limits_{T_0} \eta \left(\frac{\partial w}{\partial \xi}\right)^2 d\Omega +$$

$$+ a_{33} h_1 h_2 \iint\limits_{T_0} \frac{w^2}{r_1 + h_2 \eta} d\Omega + 2a_{13} h_1 \iint\limits_{T_0} w \frac{\partial w}{\partial \eta} d\Omega,$$

$$I_i^2 = 2a_{44} r_1 \iint\limits_{T_0} \frac{dw}{d\xi} \frac{\partial v}{\partial \eta} d\Omega + 2a_{44} h_2 \iint\limits_{T_0} \eta \frac{\partial w}{\partial \xi} \frac{\partial v}{\partial \eta} d\Omega +$$

$$+ 2a_{12} r_1 \iint\limits_{T_0} \frac{\partial w}{\partial \eta} \frac{\partial v}{\partial \xi} d\Omega + 2a_{12} h_2 \iint\limits_{T_0} \eta \frac{\partial w}{\partial \eta} \frac{\partial v}{\partial \xi} d\Omega +$$

$$+ 2a_{23} h_2 \iint\limits_{T_0} w \frac{\partial v}{\partial \xi} d\Omega, \qquad \text{(VI.16)}$$

$$I_i^3 = a_{44} r_1 \frac{h_1}{h_2} \iint\limits_{T_0} \left(\frac{\partial v}{\partial \eta}\right)^2 d\Omega + a_{44} h_1 \iint\limits_{T_0} \eta \left(\frac{\partial v}{\partial \eta}\right)^2 d\Omega +$$

$$+ a_{22} r_1 \frac{h_2}{h_1} \iint\limits_{T_0} \left(\frac{\partial v}{\partial \xi}\right)^2 d\Omega + a_{22} \frac{h_2^2}{h_1} \iint\limits_{T_0} \eta \left(\frac{\partial v}{\partial \xi}\right)^2 d\Omega.$$

Здесь и далее $d\Omega = d\xi d\eta$.

Поскольку

$$w(\xi, \eta) = \alpha_1 + \alpha_2 \xi + \alpha_3 \eta + \alpha_4 \xi^2 + \alpha_5 \xi\eta + \alpha_6 \eta^2 + \alpha_7 \xi^3 + \alpha_8 \xi^2 \eta + \alpha_9 \xi \eta^2 + \alpha_{10} \eta^3, \qquad \text{(VI.17)}$$

$$v(\xi, \eta) = \beta_1 + \beta_2 \xi + \beta_3 \eta + \beta_4 \xi^2 + \cdots + \beta_{10} \eta^3,$$

нетрудно заметить, что каждый интеграл из I_i^1 имеет вид

$$\iint\limits_{T_0} (\ldots)\, d\Omega = \omega^T \Lambda S^{-T} \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} S^{-1} \Lambda \omega, \qquad \text{(VI.18)}$$

из I_i^2 —

$$\iint\limits_{T_0} (\ldots)\, d\Omega = \omega^T \Lambda S^{-T} \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} S^{-1} \Lambda \omega, \qquad \text{(VI.19)}$$

из I_i^3 —

$$\iint\limits_{T_0} (\ldots)\, d\Omega = \omega^T \Lambda S^{-T} \begin{bmatrix} 0 & 0 \\ 0 & C \end{bmatrix} S^{-1} \Lambda \omega. \qquad \text{(VI.20)}$$

Здесь

$$S^{-1} = \begin{bmatrix} S_1^{-1} & 0 \\ 0 & S_1^{-1} \end{bmatrix}$$

есть матрица двадцатого порядка, устанавливающая для каждого треугольника T_i связь между некоторыми «масштабированными» узловыми параметрами $\bar{\omega} \equiv \Lambda\omega$ вектор-функции $U^N(r, z)$ и неизвестными числовыми коэффициентами α_j, β_j ($j = 1, 2, \ldots, 10$) полиномов (VI.17), а S_1^{-1} — соответствующая (10×10) матрица связи для каждой из компонент вектор-функции U^N; матрица S_1 — постоянна для всей области $\bar{D}$.

Вектор узловых параметров ω функции $U^N(r, z)$, фиксированных на треугольнике T_i, имеет вид

$$\omega \equiv \omega^i = [u_{r1}, p_1, q_1, u_{r2}, p_2, q_2, u_{r3}, p_3, q_3, u_{r0}, u_{z1}, \bar{p}_2, \bar{q}_1, u_{z2}, \bar{p}_2, \bar{q}_2, u_{z3}, \bar{p}_3, \bar{q}_3, u_{z0}]^T, \quad \text{(VI.21)}$$

где

$$u_{rk} = u_r^N(X_k), \; p_k = \frac{\partial u_r^N}{\partial z}(X_k), \; q_k = \frac{\partial u_r^N}{\partial r}(X_k),$$

$$u_{zk} = u_z^N(X_k), \; \bar{p}_k = \frac{\partial u_z^N}{\partial z}(X_k), \; \bar{q}_k = \frac{\partial u_z^N}{\partial r}(X_k),$$

$X_k = X_k(i)$, $k = 1, 2, 3$,— вершины T_i, $X_0 = X_0(i)$ — центр тяжести треугольника T_i, $\Lambda = \operatorname{diag}(1, h_1, h_2, 1, h_1, h_2, 1, h_1, h_2, 1, 1, h_1, h_2, 1, h_1, h_2, 1, h_1, h_2, 1)$.

Отметим, что всевозможные квадратные блоки десятого порядка типов B и C, возникающие при подсчете каждого интеграла соответственно из I_i^2 и I_i^3, не зависят от номера T_i, $i = 1, 2, \ldots, N$ (каждый интеграл порождает свой блок, одинаковый для всей области $\bar{D}$). Блоки десятого порядка типа A, возникающие при вычислении интегралов в I_i^1, кроме интеграла $\iint\limits_{T_0} \frac{\omega^2}{r_1(i) + h_2\eta} d\xi d\eta$, тоже одинаковы для любого треугольника T_i области $\bar{D}$.

Учитывая блочную структуру матрицы S^{-1}, выражения (VI.18) — (VI.20) можно записать в виде

$$\iint\limits_{T_0} (\ldots)\, d\Omega = \omega^T \Lambda \begin{bmatrix} \bar{A} & 0 \\ 0 & 0 \end{bmatrix} \Lambda\omega, \quad \text{(VI.22)}$$

$$\iint\limits_{T_0} (\ldots)\, d\Omega = \omega^T \Lambda \begin{bmatrix} 0 & \bar{B} \\ \bar{B}^T & 0 \end{bmatrix} \Lambda\omega, \quad \text{(VI.23)}$$

$$\iint\limits_{T_0} (\ldots)\, d\Omega = \omega^T \Lambda \begin{bmatrix} 0 & 0 \\ 0 & \bar{C} \end{bmatrix} \Lambda\omega, \quad \text{(VI.24)}$$

где $\bar{A} = S_1^{-T} A S_1^{-1}$, $\bar{B} = S_1^{-T} B S_1^{-1}$, $\bar{C} = S_1^{-T} C S_1^{-1}$.

Кроме того, для слагаемых (VI.16), содержащих $\left(\frac{\partial w}{\partial \eta}\right)^2$ и $\left(\frac{\partial v}{\partial \eta}\right)^2$, $\eta\left(\frac{\partial w}{\partial \eta}\right)^2$ и $\eta\left(\frac{\partial v}{\partial \eta}\right)^2$ и т. д., соответствующие блоки $\bar{A}$ и $\bar{C}$ будут одними и теми же.

Таким образом, один раз формируются необходимые блоки $\bar{A}$, $\bar{B}$, $\bar{C}$, а для получения элементарной матрицы жесткости K_i необходимо лишь умножать их на соответствующие коэффициенты и диагональные элементы матрицы Λ. Заметим, что величины K_i в случае равномерной сетки на каждом слое необходимо вычислять лишь дважды — для четных и нечетных треугольников (нумерация треугольников в области D принята слева направо и снизу вверх).

Формулы (VI.22) — (VI.24) получены в предположении, что нумерация фиксированных параметров произведена по типу (VI.21). Однако для удобства построения общей матрицы системы линейных алгебраических уравнений МКЭ целесообразно изменить упорядочение фиксируемых на T_i параметров, а именно принять

$$\omega \equiv \omega^i = [u_{r1},\ p_1,\ q_1,\ u_{z1},\ \bar{p}_1,\ \bar{q}_1,\ u_{r2},\ p_2,\ q_2,\ u_{z2},$$
$$\bar{p}_2,\ \bar{q}_2,\ u_{r3},\ p_3,\ q_3,\ u_{z3},\ \bar{p}_3,\ \bar{q}_3,\ u_{r0},\ u_{z0}]^T.$$

Тогда параметры u_{r0}, u_{z0} легко исключить из рассмотрения, выразив их через оставшиеся (см. параграф 1.5); соответственно преобразованная элементарная матрица жесткости $\bar{K}_i$ в результате будет не 20-го, а 18-го порядка. Это позволит формировать глобальную матрицу жесткости K, оперируя блоками с размерами 6×6. За счет исключения u_{r0}, u_{z0} порядок общей алгебраической системы МКЭ уменьшается на $2N$ без снижения точности искомого приближенного решения.

Для решения систем алгебраических уравнений МКЭ с симметричными положительно определенными матрицами использовался метод квадратных корней и метод LDL^T-разложения [106].

3. **Сходимость приближенных решений.** Пусть $d = l_2$, т. е. $\Gamma_5 = 0$ (см. рис. 30). Через $\{Q^n\}$ обозначим последовательность разбиений области $\bar{D} = \bar{D}_1 \cup \bar{D}_2$ на замкнутые треугольники T_k^n, $k = 1, 2, \ldots, N_n$ (не обязательно прямоугольные), которые имеют следующие свойства:

1) $\bar{D} = \bigcup\limits_{k=1}^{N_n} \bar{T}_k^n$, $Q^n = \{\bar{T}_1^n, \ldots, \bar{T}_{N_n}^n\}$

(N_n — общее число треугольников n-го разбиения области $\bar{D}$);

2) $T_i^n \cap T_j^n = 0$, $i, j = 1, 2, \ldots, N_n$, $i \neq j$;

3) часть $\Gamma^{(1)} = \Gamma_1 \cup \Gamma_2 \cup \Gamma_4$ границы Γ представляет собой объединение некоторых сторон треугольников разбиения Q^n области $\bar{D}$;

4) $\lim\limits_{n\to\infty} h_n = 0$, где h_n — длина максимальной стороны всех треугольников из Q^n;

5) $0 < \alpha \leqslant \vartheta_n$, где ϑ_n — минимальный угол в разбиении Q^n, а α — фиксированный угол.

Узловыми точками в разбиении Q^n назовем вершины треугольников и центр тяжести каждого такого треугольника.

Введем в рассмотрение функционал

$$\Phi(V, U) = \iint\limits_D \left[a_{11} \frac{\partial v_1}{\partial r} \frac{\partial u_1}{\partial r} + a_{44} \left(\frac{\partial v_1}{\partial z} + \frac{\partial v_2}{\partial r} \right) \left(\frac{\partial u_1}{\partial z} + \frac{\partial u_2}{\partial r} \right) + \right.$$

$$+a_{33}\frac{v_1u_1}{r^2}+a_{22}\frac{\partial v_2}{\partial z}\frac{\partial u_2}{\partial z}+a_{12}\left(\frac{\partial v_1}{\partial r}\frac{\partial u_2}{\partial z}+\frac{\partial u_1}{\partial r}\frac{\partial v_2}{\partial z}\right)+$$

$$+a_{13}\left(\frac{v_1}{r}\frac{\partial u_1}{\partial r}+\frac{u_1}{r}\frac{\partial v_1}{\partial r}\right)+a_{23}\left(\frac{v_1}{r}\frac{\partial u_2}{\partial z}+\frac{u_1}{r}\frac{\partial v_2}{\partial z}\right)\Bigg]rdrdz,$$

где $V=[v_1(r, z), v_2(r, z)]^T$, $U=[u_1(r, z), u_2(r, z)]^T$, $V, U\in\mathfrak{M}$, $\mathfrak{M}$ — множество вектор-функций $V(r, z)$, каждая компонента которых на каждой из областей D_1, D_2 принадлежит пространству W_1^2 и которые удовлетворяют условиям сопряжения (VI.5) или (VI.7), т. е. $[V]|_{\Gamma_s}=0$ или $[v_1]|_{\Gamma_s}=0$. Нетрудно проверить, что

$$\begin{gathered}\Phi(V, V)\geqslant 0,\quad \Phi(V, U)=\Phi(U, V),\\ \Phi(\alpha U, V)=\alpha\Phi(U, V),\\ \Phi(V, U+W)=\Phi(V, U)+\Phi(V, W),\\ \Phi(\alpha V+\beta U, \alpha V+\beta U)=\alpha^2\Phi(V, U)+2\alpha\beta\Phi(V, U)+\beta^2\Phi(U, U),\end{gathered}\tag{VI.25}$$

где $V, U, W\in\mathfrak{M}$, α, β — произвольные вещественные числа.

Построение и исследование приближенного решения $U^N(r, z)$ можно выполнять, минимизируя не функционал (VI.9) (или, что то же, (VI.10)), а функционал вида

$$\tilde{I}(V)=\frac{1}{2}\Phi(V, V)-(f, V)-l(V),\quad \Phi(V, V)\equiv F(V),$$

на множестве вектор-функций $V(r, z)\in\overset{0}{\mathfrak{M}}\subset\mathfrak{M}$, удовлетворяющих условию (VI.2), т. е. $v_2|_{\Gamma^{(1)}}=g(r)$.

Далее будут использованы следующие леммы.

Лемма VI.1. *Если вектор-функция* $\tilde{U}(r, z)$, *минимизирующая функционал* $\tilde{I}(V)$ *на множестве* $\overset{0}{\mathfrak{M}}$, *непрерывно дифференцируема на* $\bar{D}_1$, $\bar{D}_2$ *и имеет в* D_1, D_2 *ограниченные и почти всюду непрерывные частные производные второго порядка, то для произвольной вектор-функции* $V(r, z)\in\mathfrak{M}$ *справедливо соотношение*

$$\Phi(\tilde{U}, V)=(f, V)+l(V)+R(\tilde{U}, V),$$

где

$$R(\tilde{U}, V)\equiv R(V)=\int\limits_{\Gamma^{(1)}}v_2\left(a_{12}\frac{\partial\tilde{u}_1}{\partial r}+a_{22}\frac{\partial\tilde{u}_2}{\partial z}+a_{23}\frac{\partial\tilde{u}_1}{r}\right)r\cos(n, z)\,d\Gamma.$$

Доказательство проводится аналогично тому, как это сделано в работе [156]. Нетрудно доказать следующую лемму.

Лемма VI.2. *Если выполнены предположения леммы VI.1, то для каждой функции* $V\in\mathfrak{M}$ *верно соотношение*

$$\tilde{I}(V)-\tilde{I}(\tilde{U})=\frac{1}{2}\Phi(U-V, \tilde{U}-V)+R(V-U).\tag{VI.26}$$

Доказательство проводится на основании свойств (VI.25).

В соответствии с разбиением Q'' введем в рассмотрение последовательность линейных множеств $\{H^n\}$ вектор-функций $V^n \in \mathfrak{M}$, являющихся кубическими полиномами Эрмита на каждом $\overline{T}_k^n$.

Лемма VI.3. *Если $V^n \in H^n$ и для этой функции в центрах тяжести X_0 и вершинах X_j треугольников разбиения Q^n справедливы соотношения*

$$V^n(X_0) = \tilde{U}(X_0),\ V^n(X_j) = \tilde{U}(X_j),$$

$$\frac{\partial V^n}{\partial r}(X_j) = \frac{\partial \tilde{U}}{\partial r}(X_j),\ \frac{\partial V^n}{\partial z}(X_j) = \frac{\partial \tilde{U}}{\partial z}(X_j),$$

функция $\tilde{U}$ непрерывно дифференцируема на $\overline{D}_1$, $\overline{D}_2$ и имеет ограниченные числом M_4 частные производные четвертого порядка в каждой из областей D_1, D_2, то

$$\lim_{n\to\infty} \tilde{I}(V^n) = \tilde{I}(\tilde{U}) \tag{VI.27}$$

и скорость сходимости в (VI.27) будет $O(h_n^4)$; если же $g(r)$ в (VI.2) — полином не выше третьей степени на $\Gamma^{(1)}$, то скорость сходимости в (VI.27) — $O(h_n^6)$.

Доказательство проводится с учетом результатов теоремы из работы [157] об аппроксимации функций кубическими полиномами Эрмита на треугольниках, соотношения (VI.26) и того факта, что

$$\Phi(V, V) \equiv F(V) = 2\iint_D \Pi(V)\, r\,dr\,dz, \tag{VI.28}$$

где $\Pi(V)$ — упругий потенциал.

Теорема VI.1. *Для последовательности приближенных решений $\{\tilde{U}^n\}$, соответствующей последовательности разбиений $\{Q^n\}$, имеют место соотношения*

$$\lim_{n\to\infty} \tilde{I}(\tilde{U}^n) = \tilde{I}(\tilde{U}), \tag{VI.29}$$

$$\lim_{n\to\infty} F(\tilde{U} - \tilde{U}^n) = 0. \tag{VI.30}$$

Если $g(r)$ из (VI.2) является полиномом не выше третьей степени на $\Gamma^{(1)}$, то скорость сходимости в (VI.29), (VI.30) имеет порядок h_n^6, в противном случае — $O(h^4)$.

Доказательство. В силу (VI.26) имеем

$$\tilde{I}(\tilde{U}^n) - \tilde{I}(\tilde{U}) = \frac{1}{2} F(\tilde{U} - \tilde{U}^n) - R(\tilde{U} - \tilde{U}^n). \tag{VI.31}$$

С учетом результатов работы [157] на $\Gamma^{(1)}$ справедливо неравенство

$$|\tilde{u}_2 - \tilde{u}_2^n| \leqslant \frac{3}{\sin\vartheta_n} M_4 h_n^4.$$

Если $g(r)$ в условии (VI.2) на $\Gamma^{(1)}$ является полиномом не выше третьей степени, то $R(\tilde{U}-\tilde{U}^n)=0$, в противном случае

$$|R(\tilde{U}-\tilde{U}^n)|=O(h_n^4). \tag{VI.32}$$

Так как $F(V)\geqslant 0$ для $\forall V\in\mathfrak{M}$, из (VI.31) следует, что $\tilde{J}(\tilde{U}^n)+R(\tilde{U}-\tilde{U}^n)\geqslant\tilde{I}(\tilde{U})$. Однако для произвольной вектор-функции $V^n\in\overset{0}{H}{}^n$ (множество $\overset{0}{H}{}^n\subset H^n$ состоит из тех функций, которые на $\Gamma^{(1)}$ удовлетворяют условию (VI.12), т. е. $v_2^n=g(r)$, $\frac{\partial v_2^n}{\partial r}=\frac{\partial g}{\partial r}$) справедливо неравенство

$$\tilde{I}(V^n)\geqslant\tilde{I}(\tilde{U}^n). \tag{VI.33}$$

Тогда на основании (VI.32), (VI.33) имеем

$$\tilde{I}(V^n)+R(\tilde{U}-\tilde{U}^n)\geqslant\tilde{I}(\tilde{U}^n)+R(\tilde{U}-\tilde{U}^n)\geqslant\tilde{I}(\tilde{U}). \tag{VI.34}$$

Если вместо произвольной функции V^n, удовлетворяющей (VI.34), взять функцию $V_I^n\in\overset{0}{H}{}^n$, являющуюся кубическим интерполянтом Эрмита функции $\tilde{U}$ на каждом треугольнике разбиения Q^n, то из (VI.34) на основании (VI.27), (VI.32) следует (VI.29), а из (VI.29), (VI.31), (VI.32) получаем (VI.30).

С учетом результатов леммы VI.3 и соотношений (VI.31), (VI.32), (VI.34) получаем, что скорость сходимости в (VI.29), (VI.30) будет $O(h_n^6)$, если $g(r)$ на $\Gamma^{(1)}$ — полином не выше третьей степени, в противном случае — $O(h_n^4)$.

4. Обусловленность матрицы системы алгебраических уравнений МКЭ. Так как упругий потенциал $\Pi(U)$ строго больше нуля в деформированном состоянии, он является положительно определенной квадратичной формой относительно компонент тензора деформаций ε_{rr}, ε_{rz}, ε_{zz}, $\varepsilon_{\varphi\varphi}$:

$$2\Pi(U)=a_{11}\varepsilon_{rr}^2+2a_{12}\varepsilon_{rr}\varepsilon_{zz}+2a_{13}\varepsilon_{rr}\varepsilon_{\varphi\varphi}+a_{22}\varepsilon_{zz}^2+2a_{23}\varepsilon_{zz}\varepsilon_{\varphi\varphi}+a_{33}\varepsilon_{\varphi\varphi}^2+$$
$$+a_{44}\varepsilon_{rz}^2>0. \tag{VI.35}$$

Следовательно, для произвольной функции $U\in\mathfrak{M}$, $U=[u_1,\ u_2]^T$,

$$\gamma_1(\varepsilon_{rr}^2+\varepsilon_{zz}^2+\varepsilon_{rz}^2+\varepsilon_{\varphi\varphi}^2)\leqslant 2\Pi(U)\leqslant\gamma_4(\varepsilon_{rr}^2+\varepsilon_{zz}^2+\varepsilon_{rz}^2+\varepsilon_{\varphi\varphi}^2), \tag{VI.36}$$

где

$$\varepsilon_{rr}=\frac{\partial u_1}{\partial r},\quad \varepsilon_{zz}=\frac{\partial u_2}{\partial z},\quad \varepsilon_{rz}=\frac{\partial u_1}{\partial z}+\frac{\partial u_2}{\partial r},\quad \varepsilon_{\varphi\varphi}=\frac{u_1}{r},$$

$$\gamma_1=\min(\gamma_1^-,\ \gamma_1^+),\quad \gamma_4=\max(\gamma_4^-,\ \gamma_4^+),$$

$$\gamma_1^{\pm}=\min(\lambda_1,\ \lambda_2,\ \lambda_3)^{\pm},$$

$$\gamma_4^{\pm}=\max(\lambda_1,\ \lambda_2,\ \lambda_3,\ a_{44})^{\pm},$$

$(\lambda_1, \lambda_2, \lambda_3)^{\pm}$ — собственные числа матрицы $A^{\pm} = (a_{ij}^{\pm})$, $i, j = 1, 2, 3$.

В дальнейшем будем предполагать, что значение $\hbar = \max(|h_1(i)|, |h_2(i)|)$ достаточно мало, что для всех рассматриваемых разбиений области триангуляция имеет вид, представленный на рис. 31, и что

$$c_1 \leqslant h_1(i)/h_2(i) \leqslant c_2,$$
$$c_3 \leqslant h_k^+(i)/h_k^-(i) \leqslant c_4, \quad k = 1, 2. \tag{VI.37}$$

(здесь $h_1(i)$ — шаги по оси z, $h_2(i)$ — по оси r).

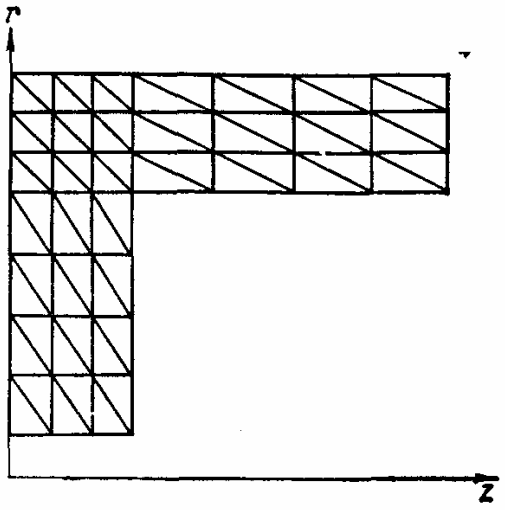

Рис. 31.

Введем обозначение

$$h = \min(|h_1(i)|, |h_2(i)|).$$

Для оценки числа обусловленности матрицы общей системы алгебраических уравнений в рассматриваемой задаче предположим еще, что $g(r) \equiv 0$ (см. (VI.2)). Тогда для $U^n \in \overset{0}{H}{}^n$, $U^n = [u_1^n, u_2^n]^T$, учитывая соотношения (VI.28), (VI.36) и неравенство

$$\iint_D (u_2^n)^2 \, d\Omega \leqslant l_2^2 \iint_D \left(\frac{\partial u_2^n}{\partial z}\right)^2 d\Omega,$$

нетрудно получить

$$\gamma^2 (U^n, U^n) \leqslant F(U^n) \leqslant \mu_2 \left[\left(\frac{\partial U^n}{\partial z}, \frac{\partial U^n}{\partial z}\right) + \left(\frac{\partial U^n}{\partial r}, \frac{\partial U^n}{\partial r}\right) + (U^n, U^n)\right], \tag{VI.38}$$

где

$$\gamma^2 = \gamma_1 r_0 \min(l_1^{-2}, l_0^{-2}), \quad \mu_2 = \gamma_4 l_1 \max(2, r_0^{-2}).$$

В соответствии с триангуляцией области D, как и прежде, можно записать

$$(U^n, U^n) = \sum_{i=1}^{N_n} \iint_{T_0} (w^2(\xi, \eta) + v^2(\xi, \eta)) h_1 h_2 d\xi d\eta \geqslant$$
$$\geqslant h^2 \sum_{i=1}^{N_n} \iint_{T} (w^2 + v^2) \, d\xi d\eta,$$

где $w(\xi, \eta) = u_1^n(r(\eta), z(\xi))$, $v(\xi, \eta) = u_2^n(r(\eta), z(\xi))$. Для каждого T_k^n справедливо неравенство

$$I^k = \iint_{T_0} (w^2 + v^2) \, d\xi d\eta \geqslant 0. \tag{VI.39}$$

Если $I^k = 0$, то $w(\xi, \eta) \equiv 0$, $v(\xi, \eta) \equiv 0$, а следовательно, согласно (VI.18) $\alpha_i = \beta_i \equiv 0$, $i = 1, 2, \ldots, 10$. Зафиксируем на каждом T_k^n, $k =$

$= 1, 2, \ldots, N_n$, «масштабированные» параметры (см. п. 3 параграфа II.3): $\tilde{\omega}^k = [u_{r1}, p_1h_1, q_1h_2, u_{r2}, p_2h_1, q_2h_2, u_{r3}, p_3h_1, q_3h_2, u_{r0}, u_{z1}, \bar{p}_1h_1, \bar{q}_1h_2, u_{z2}, \bar{p}_2h_1, \bar{q}_2h_2, u_{z3}, \bar{p}_3h_1, \bar{q}_3h_2, u_{z0}]^T$.

Тогда на основании связи $[\alpha, \beta]^T = S^{-1}\tilde{\omega}^k$, $[\alpha, \beta]^T = [\alpha_1, \alpha_2, \ldots, \alpha_{10}, \beta_1, \ldots, \beta_{10}]^T$ из $I^k = 0$ следует $\tilde{\omega}^k \equiv 0$, т. е. I^k — положительно определенная квадратичная форма параметров $\tilde{\omega}^k$:

$$I^k = \iint\limits_{T_0} (w^2 + v^2)\, d\xi d\eta = (\tilde{\omega}^k)^T M_k \tilde{\omega}^k \geqslant \nu_0 (\tilde{\omega}^k)^T \tilde{\omega}^k, \qquad \text{(VI.40)}$$

где ν_0 — минимальное собственное число матрицы M_k, которое может быть легко вычислено. Просуммируем (VI.40) по всем значениям $k = 1, 2, \ldots, N_n$. С учетом (VI.38) получим

$$F(U^n) \geqslant Ch^2 \tilde{\omega}^T \tilde{\omega}, \quad U^n \in \mathring{H}^n, \qquad \text{(VI.41)}$$

где параметры в векторе $\tilde{\omega}$ упорядочены в соответствии с нумерацией, $k = 1, 2, \ldots, N_n$, треугольников разбиения, $C = \nu_0 \gamma^2 \min(1, c_3^2, c_4^{-2})$.

Оценку сверху для $F(U^n)$ запишем в виде

$$F(U^n) \leqslant \mu_2 \sum_{k=1}^{N_n} (\bar{I}_k + \underline{I}_k).$$

Здесь

$$\bar{I}_k = \left(\frac{\partial U^n}{\partial z}, \frac{\partial U^n}{\partial z}\right)_{T_k} + \left(\frac{\partial U^n}{\partial r}, \frac{\partial U^n}{\partial r}\right)_{T_k} \equiv$$

$$\equiv \iint\limits_{T_k^n} \left[\left(\frac{\partial u_1^n}{\partial z}\right)^2 + \left(\frac{\partial u_2^n}{\partial z}\right)^2 + \left(\frac{\partial u_1^n}{\partial r}\right)^2 + \left(\frac{\partial u_2^n}{\partial r}\right)^2\right] dr dz,$$

$$\underline{I}_k = (U^n, U^n)_{T_k} \equiv \iint\limits_{T_k} [(u_1^n)^2 + (u_2^n)^2]\, dr dz.$$

Или с учетом (VI.14)

$$\bar{I}_k = \iint\limits_{T_0} \left[\left(\frac{\partial w}{\partial \xi}\right)^2 \frac{h_2}{h_1} + \left(\frac{\partial w}{\partial \eta}\right)^2 \frac{h_1}{h_2} + \left(\frac{\partial v}{\partial \xi}\right) \frac{h_2}{h_1} + \left(\frac{\partial v}{\partial \eta}\right)^2 \frac{h_1}{h_2}\right] d\xi d\eta,$$

$$\underline{I}_k = h_1 h_2 \iint\limits_{T_0} (w^2 + v^2)\, d\xi d\eta.$$

Каждую компоненту вектор-функции $U^n = [w, u]^T$ можно представить в виде

$$w(\xi, \eta) = \sum_{j=1}^{10} \tilde{\omega}_j^k \varphi_j(\xi, \eta), \quad v(\xi, \eta) = \sum_{j=1}^{10} \tilde{\omega}_{j+10}^k \varphi_j, \qquad \text{(VI.42)}$$

где φ_j — соответствующие кубические полиномы.

Легко проверить, что

$$\max_j \left(\left| \frac{\partial \varphi_j}{\partial \xi} \right|, \left| \frac{\partial \varphi_j}{\partial \eta} \right| \right) \leqslant c_5, \quad j = 1, 2, \ldots, 10, \quad \xi, \eta \in \bar{T}_0. \tag{VI.43}$$

Из (VI.37) следует:

$$\max_i \left(\frac{h_1}{h_2}, \frac{h_2}{h_1} \right) \leqslant \max \left(c_2, \frac{1}{c_1} \right) = c_0. \tag{VI.44}$$

Согласно (VI.42) — (VI.44) имеем

$$\bar{I}_k \leqslant \bar{c}_0 (\tilde{\omega}^k)^T \tilde{\omega}^k \text{ и } \underline{I}_k \leqslant \bar{\hbar}^2 \nu_1 (\tilde{\omega}^k)^T \tilde{\omega}^k,$$

где $\bar{c}_0 = 10 c_0 c_5^2$, ν_1 — максимальное собственное число матрицы M_k (см. (VI.40)).

Таким образом,

$$F(U^n) \leqslant \mu_2 (\bar{c}_0 + \nu_1 \hbar^2) \sum_{k=1}^{N_n} (\tilde{\omega}^k)^T \tilde{\omega}^k,$$

или

$$F(U^n) \leqslant c_6 \tilde{\omega}^T \tilde{\omega}, \tag{VI.45}$$

где $c_6 = 6\mu_2 (\bar{c}_0 + \nu_1 \hbar^2) \max (1, c_4^2, c_3^{-2})$, $\tilde{\omega}$ — те же, что и в (VI.41). Так как $\hbar \to 0$ при $n \to \infty$, то

$$c_6 \leqslant 6\mu_2 (\bar{c}_0 + \nu_1) \max (1, c_4^2, c_3^{-2}) = \bar{C}.$$

Итак, для случая $g(r) \equiv 0$ установлено (см. (VI.41), (VI.45)), что при $\forall\, U^n \in \overset{0}{H}{}^n$ функционал $F(U^n)$ является положительно определенной квадратичной формой параметров $\tilde{\omega}$, т. е. $F(U^n) = \tilde{\omega}^T \tilde{K} \tilde{\omega}$, и число обусловленности «масштабированной» матрицы $K = \Lambda^{-1} K \Lambda^{-1}$ результирующей системы алгебраических уравнений МКЭ будет $O(h^{-2})$:

$$C h^2 \tilde{\omega}^T \tilde{\omega} \leqslant F(U^n) \leqslant \bar{C} \tilde{\omega}^T \tilde{\omega}, \quad \text{т. е. } \frac{\lambda_{\max}}{\lambda_{\min}} \leqslant \frac{\bar{C}}{C} h^{-2},$$

где

$$\bar{C} = 6\mu_2 (\bar{c}_0 + \nu_1) \max (1, c_4^2, c_3^{-2}),$$

$$C = \nu_0 \gamma^2 \min (1, c_3^2, c_4^{-2}).$$

Поскольку вид матрицы результирующей системы МКЭ для функционала (VI.10) определяется только квадратичным слагаемым $F(U)$, независимо от граничной функции $g(r)$ число обусловленности «масштабированной» матрицы системы остается $O(h^{-2})$.

5. **Численный пример.** Искомое решение V_0 рассматриваемых прикладных задач строилось в виде

$$V_0 = V_{\text{I}} + m V_{\text{II}},$$

где числовой коэффициент m определялся из некоторого специального условия, V_{I}, V_{II} — решения системы дифференциальных уравнений

$$a_{11}\frac{\partial}{\partial r}\left(r\frac{\partial u_r}{\partial r}\right)-a_{33}\frac{u_r}{r}+a_{44}\frac{\partial}{\partial z}\left(r\frac{\partial u_r}{\partial z}+r\frac{\partial u_z}{\partial r}\right)+$$

$$+a_{12}\frac{\partial}{\partial r}\left(r\frac{\partial u_z}{\partial z}\right)-a_{23}\frac{\partial u_z}{\partial z}=0,$$

$$a_{12}\frac{\partial}{\partial z}\left(r\frac{\partial u_r}{\partial r}\right)+a_{22}\frac{\partial}{\partial z}\left(r\frac{\partial u_z}{\partial z}\right)+a_{23}\frac{\partial u_r}{\partial z}+$$

$$+a_{44}\frac{\partial}{\partial r}\left(r\frac{\partial u_r}{\partial z}+\frac{\partial u_z}{\partial r}\right)=0,\quad (r,\ z)\in D,$$

при соответствующих краевых условиях:
для $V_{\mathrm{I}}(r, z)$

$$u_z=0,\ (r,\ z)\in\Gamma^{(1)};$$

$$\sigma_{rz}=0,\ (r,\ z)\in\Gamma;\quad \sigma_{zz}=0,\ (r,\ z)\in\Gamma_6;$$

$$\sigma_{rr}=0,\ (r,\ z)\in\Gamma_5\cup\Gamma_7;\quad \sigma_{rr}=-p_0,\ (r,\ z)\in\Gamma_3;$$

для $V_{\mathrm{II}}(r, z)$

$$u_z=0,\ (r,\ z)\in\Gamma_1\cup\Gamma_2,\ u_z=g,\ (r,\ z)\in\Gamma_5;$$

$$\sigma_{rz}=0,\ (r,\ z)\in\Gamma;\quad \sigma_{zz}=0,\ (r,\ z)\in\Gamma_6;$$

$$\sigma_{rr}=0,\ (r,\ z)\in\Gamma_3\cup\Gamma_5\cup\Gamma_7.$$

При жестком соединении ребра с оболочкой условия сопряжения имели вид

$$[u_r]|_{\Gamma_8}=[u_z]|_{\Gamma_8}=0,\ [\sigma_{rr}]|_{\Gamma_8}=[\sigma_{rz}]|_{\Gamma_8}=0,$$

а при скользящем —

$$[u_r]|_{\Gamma_8}=0,\ [\sigma_{rr}]|_{\Gamma_8}=0,\ \sigma_{rz}^{+}|_{\Gamma_8}=\sigma_{rz}^{-}|_{\Gamma_8}=0.$$

Приведем исходные данные в безразмерных величинах: упругие постоянные для ребра

$$a_{11}=2{,}49946,\quad a_{12}=1{,}203\,815,\quad a_{13}=0{,}971\,5716,$$

$$a_{22}=2{,}9144,\quad a_{23}=2{,}061\,307,\quad a_{33}=7{,}965\,14,\quad a_{44}=1;$$

для оболочки

$$a_{11}=2{,}212512,\quad a_{12}=1{,}17133,\quad a_{22}=5{,}325998,$$

$$a_{23}=1{,}325998,\quad a_{44}=0{,}5,\quad a_{13}=a_{12},\quad a_{33}=a_{22}.$$

Параметры области $\bar{D}$ (см. рис. 29) следующие: $c=3{,}8$, $l_1=3{,}92$, $r_0=3{,}6$, $l_2=0{,}3$, $d=0{,}1$, $p_0=10^{-3}$, g — постоянная величина, вычисляемая через p_0, параметры области $\bar{D}$ и соответствующую составляющую модуля Юнга для оболочки.

Решение прикладных задач для случаев жесткого и скользящего соединений ребер с оболочкой проведено на двух разных неравномерных сетках, схематически эти триангуляции изображены на рис. 31, 32.

Анализ результатов определения напряженно-деформированного

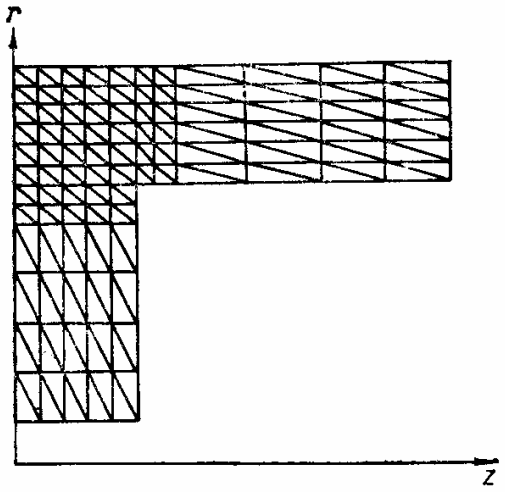

Рис. 32.

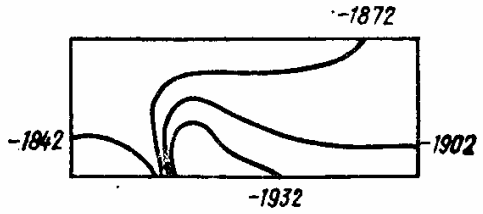

Рис. 33.

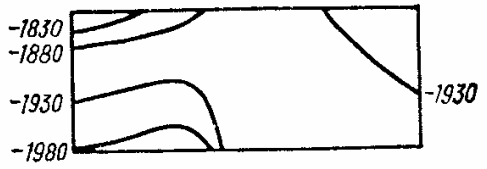

Рис. 34.

состояния ортотропной оболочки вращения, регулярно подкрепленной кольцевыми ребрами жесткости, показывает, что сгущение неравномерной сетки в окрестности точки с координатами (c, d) (см. рис. 32) позволяет улучшить результаты в ее окрестности. Однако вдали от этой точки результаты практически не изменяются, и можно считать, что решение практически достигнуто на достаточно грубой сетке (см. рис. 31), для чего хватило 3 мин на ЭВМ БЭСМ-6.

На рис. 33 показано распределение напряжений $\sigma_{\varphi\varphi}$ для случая жесткого соединения ребра с оболочкой, а на рис. 34 — для скользящего.

VI.2. Определение частот и форм собственных колебаний различных моделей компрессорных лопаток [76]

1. Постановка задачи. Пусть требуется определить частоты и формы собственных колебаний компрессорной лопатки турбомашины (рис. 35). Края лопатки свободны, а хвостовик ее вставлен в замок диска турбины, т. е. жестко защемлен. Опишем постановки некоторых математических задач, соответствующих требуемому расчету. Если рассматривать лопатку как стержень переменного сечения, то для различных видов колебаний (изгибных, крутильных и т. д.) ставятся задачи на собственные значения для обыкновенных дифференциальных операторов. Рассматривая лопатку как пластину или оболочку переменной толщины, получаем задачи на собственные значения для дифференциальных операторов в частных производных в двумерной области (обычно прямоугольной). Задача для пластины является частным случаем задачи для оболочки и отдельно нами не исследуется.

Если используется стержневая модель и изучаются изгибные колебания, то задача на собственные значения, поставленная в слабой форме, запишется так:

$$a(U, V) - \lambda b(U, V) = 0, \quad \forall V \in H(\Omega), \tag{VI.46}$$

где $\Omega = (\alpha, \beta)$, $H = H_1 \times H_1$, H_1 — пространство функций $u_k(x)$,

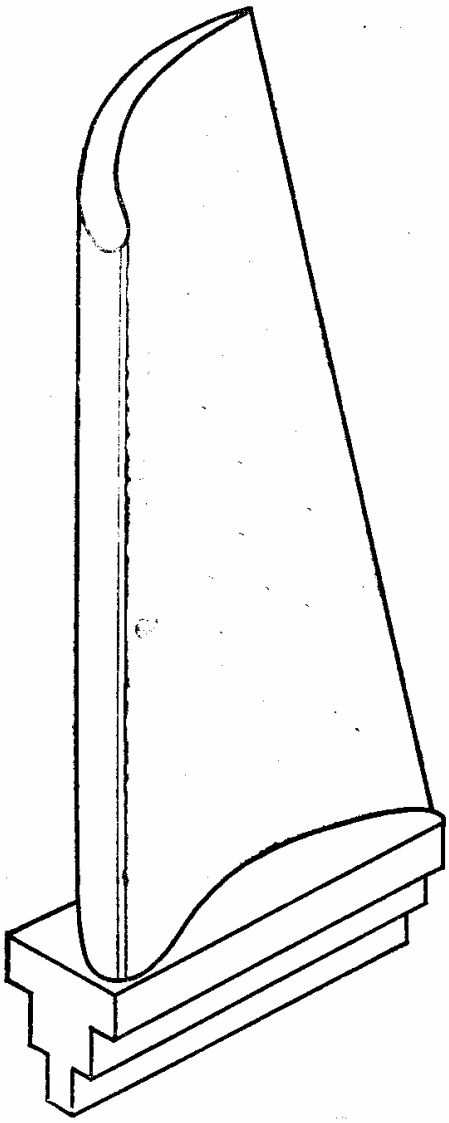

Рис. 35.

принадлежащих $W_2^2(\Omega)$ и удовлетворяющих условиям

$$u_k(\beta) = \frac{du_k}{dx}\Big|_{x=\beta} = 0, \quad \text{(VI.47)}$$

а билинейные функционалы $a(U, V)$, $b(U, V)$ от вектор-функций $U, V \in H(\Omega)$, $U = [u_1, u_2]^T$, $V = [v_1, v_2]^T$ определяются формулами [24]

$$a(U, V) = \int_\alpha^\beta E\left[I_2(x)\frac{d^2u_1}{dx^2}\frac{d^2v_1}{dx^2} + I_{12}(x)\left(\frac{d^2u_1}{dx^2} + \frac{d^2u_2}{dx^2}\frac{d^2v_1}{dx^2}\right) + I_1(x)\frac{d^2u_2}{dx^2}\frac{d^2v_2}{dx^2}\right]dx, \quad \text{(VI.48)}$$

$$b(U, V) = \int_\alpha^\beta \rho(u_1v_1 + u_2v_2)Q(x)\,dx.$$

Здесь $I_1(x)$, $I_2(x)$, $I_{12}(x)$ — моменты инерции относительно некоторых осей, $Q(x)$ — площадь поперечного сечения стержня. Если

$$E > 0,\ \rho > 0,\ 0 < c_1 \leqslant Q(x) \leqslant c_2,$$
$$0 < c_3 \leqslant I_1(x),\ I_2(x) \leqslant c_4,$$
$$I_1(x)I_2(x) - I_{12}^2(x) \geqslant c_5 > 0,$$

то выполняются соотношения

$$a(U, V) = a(V, U),\ b(U, V) = b(V, U),$$
$$a(U) = a(U, U),\ b(U) = b(U, U), \quad \text{(VI.49)}$$
$$k_A\|U\|_H^2 \leqslant a(U) \leqslant K_A\|U\|_H^2\ (k_A, K_A > 0),$$
$$a(U) \geqslant c_A\|U\|_0^2\ (c_A > 0),$$
$$k_B\|U\|_0^2 \leqslant b(U) \leqslant K_B\|U\|_0^2\ (k_B, K_B > 0), \quad \text{(VI.50)}$$
$$a(U) \geqslant \gamma_0^2 b(U),$$

где

$$\|U\|_H^2 = (\|u_1\|_{2,2}^2 + \|u_2\|_{2,2}^2),\ \|u_k\|_{2,2} = \|u_k\|_{W_2^2},$$
$$\|U\|_0^2 = (\|u_1\|^2 + \|u_2\|^2),\ \|u_k\| = \|u_k\|_{L_2}.$$

В случае оболочечной модели

$$\Omega = \{x = (x_1, x_2),\ \alpha_1 < x_1 < \beta_1,\ \alpha_2 < x_2 < \beta_2\}. \quad \text{(VI.51)}$$

$H = H_1 \times H_1 \times H_3$, где H_1 — пространство функций $u_k(x)$, принад-

лежащих $W_2^1(\Omega)$ и удовлетворяющих условию

$$u_k(x) = 0, \quad x \in \Gamma_1 \tag{VI.52}$$

($\Gamma_1 = \{x = (x_1, x_2), \alpha_1 \leqslant x_1 \leqslant \beta_1, x_2 = \beta_2\}$), H_3 — пространство функций $u_3(x)$, принадлежащих $W_2^2(\Omega)$ и удовлетворяющих условиям

$$u_3(x) = \frac{\partial u_3}{\partial x_2} = 0, \quad x \in \Gamma_1. \tag{VI.53}$$

В этом случае билинейные функционалы от вектор-функций $U, V \in H(\Omega)$ ($U = [u_1, u_2, u_3]^T$, $V = [v_1, v_2, v_3]^T$) имеют в (VI.46) вид [17]

$$a(U, V) = \iint_\Omega \frac{E}{1-\nu^2} \Big\{ \varepsilon_1(U)\varepsilon_1(V) + \varepsilon_2(U)\varepsilon_2(V) +$$

$$+ \frac{1-\nu}{2}\omega(U)\omega(V) + \nu[\varepsilon_1(U)\varepsilon_2(V) + \varepsilon_2(U)\varepsilon_1(V)] +$$

$$+ \frac{\delta^2(x)}{12}[\nu(\varkappa_1(U)\varkappa_2(V) + \varkappa_1(U)\varkappa_2(V)) + 2(1-\nu)\tau(U)\tau(V) +$$

$$+ \varkappa_2(U)\varkappa_2(V) + \varkappa_1(U)\varkappa_1(V)]\Big\}\delta(x)A_1(x)A_2(x)\,dx, \tag{VI.54}$$

$$b(U, V) = \iint_\Omega \rho(u_1v_1 + u_2v_2 + u_3v_3)\delta(x)A_1(x)A_2(x)\,dx,$$

где

$$\varepsilon_i(W) = \frac{1}{A_i}\frac{\partial w_i}{\partial x_i} + \frac{1}{A_1A_2}\frac{\partial A_i}{\partial x_j}w_j + k_{ii}(x)w_3,$$

$$\omega(W) = \frac{A_1}{A_2}\frac{\partial}{\partial x_2}\left(\frac{w_1}{A_1}\right) + \frac{A_2}{A_1}\frac{\partial}{\partial x_1}\left(\frac{w_2}{A_2}\right) + k_{12}(x)\cdot 2\cdot w_3,$$

$$\varkappa_i(W) = \frac{1}{A_i}\frac{\partial\gamma_i(W)}{\partial x_i} + \frac{1}{A_1A_2}\frac{\partial A_i}{\partial x_j}\gamma_j(W) +$$

$$+ \frac{k_{12}(x)}{2A_1A_2}\left(\frac{\partial}{\partial x_i}(A_jw_j) - \frac{\partial}{\partial x_j}(A_iw_i)\right), \tag{VI.55}$$

$$\gamma_i(W) = k_{ii}(x)w_i + k_{12}(x)w_j - \frac{1}{A_i}\frac{\partial w_3}{\partial x_i},$$

$$\tau(W) = \frac{1}{A_2}\frac{\partial\gamma_1}{\partial x_2} + \frac{1}{A_1}\frac{\partial\gamma_2}{\partial x_1} + \frac{k_{12}}{A_1}\frac{\partial w_1}{\partial x_1} + \frac{k_{11}}{A_2}\frac{\partial w_1}{\partial x_2} + \frac{k_{22}}{A_1}\frac{\partial w_2}{\partial x_1} +$$

$$+ \frac{k_{12}}{A_2}\frac{\partial w_2}{\partial x_2} - (k_{11}+k_{22})\left[\frac{1}{A_1A_2}\left(\frac{\partial A_1}{\partial x_2}w_1 + \frac{\partial A_2}{\partial x_1}w_2\right) - 2k_{12}w_3\right] +$$

$$+ \frac{1}{A_1^2A_2}\frac{\partial A_1}{\partial x_2}\frac{\partial w_3}{\partial x_1} + \frac{1}{A_1A_2^2}\frac{\partial A_2}{\partial x_1}\frac{\partial w_3}{\partial x_2} \quad (W \in H(\Omega),\ i = 1, 2;\ j = 3-i),$$

$\delta(x)$, $A_1(x)$, $A_2(x)$, $k_{11}(x)$, $k_{12}(x)$, $k_{22}(x)$ — непрерывные в Ω (VI.51) функции (в дальнейшем по ходу изложения предполагается их необходимая гладкость). При условиях

$$\rho > 0,\ E > 0,\ |\nu| < 1,$$

$$0 < c_6 \leqslant \delta(x) \leqslant c_7, \quad 0 < c_8 \leqslant A_i(x) \leqslant c_9,$$

$$|k_{ij}(x)| \leqslant c_{10} < \infty, \quad i = 1, 2, \quad j = 1, 2, \quad k_{21} = k_{12},$$

используя [102] и [117], можно доказать неравенства (VI.49), (VI.50), где теперь $\| U \|_H^2 = \| u_1 \|_{2,1}^2 + \| u_2 \|_{2,1}^2 + \| u_3 \|_{2,2}^2 \|$ и $\| U \|_0^2 = \| u_1 \|^2 + \| u_2 \|^2 + \| u_3 \|^2$.

Введенные в представленных задачах функционалы $a(U, V)$, $b(U, V)$ можно рассматривать как скалярные произведения в некоторых энергетических пространствах H_A, H_B положительно определенных операторов A и B.

Применяя теоремы вложения Соболева [98], можно доказать, что в обеих задачах пространство H_A вкладывается в H_B вполне непрерывно. Следовательно, можно использовать вариационную постановку этих задач, а именно находить экстремальные точки отношения Рэлея

$$R(U) = \frac{a(U)}{b(U)}$$

или решать задачу на условный минимум

$$\lambda_l = \min_{U \in H_A} a(U) = a(V_l), \quad l = 1, 2, \ldots, n, \tag{VI.56}$$

при условиях

$$b(U) = 1, \tag{VI.57}$$

$$b(U, V_k) = 0, \quad k = 1, 2, \ldots, l-1, \quad l \geqslant 2. \tag{VI.58}$$

При $k_{11} = k_{12} = k_{22} = 0$ и $A_1 = A_2 = 1$ вместо оболочки мы имеем задачу для пластинки. Легко проверить, что в этом случае задача (VI.46), (VI.51) — (VI.55) распадается на две задачи — об изгибных и о тангенциальных колебаниях.

2. Дискретизация задач. Рассмотрим вначале дискретизацию задачи (VI.46) — (VI.48), т. е. случай стержневой модели.

Разобьем отрезок $[\alpha, \beta]$ точками

$$\alpha = x_0 < x_1 < x_2 < \cdots < x_{N-1} < x_N = \beta \tag{VI.59}$$

на N элементов $E_i = (x_{i-1}, x_i)$, $i = 1, 2, \ldots, N$. Точки (VI.59) являются узлами полученной сетки, которая характеризуется величиной

$$h = \max_i h_i \equiv \max_i (x_i - x_{i-1}). \tag{VI.60}$$

Будем искать приближенное решение задачи (VI.46) — (VI.48) в конечномерном подпространстве P^3 пространства $H(\alpha, \beta)$. Это подпространство образуем из вектор-функций $U^h = [u_1^h, u_2^h]^T$, компоненты которых на каждом из $\overline{E_i}$ являются кусочными эрмитовыми кубическими полиномами.

За счет приравнивания в узлах одноименных узловых параметров $\left(\text{т. е. значений } u_k^h \text{ и } \frac{du_k^h}{dx}\right)$ обеспечивается требование $u_k^h \in \mathring{W}_2^2(\alpha, \beta)$, $k = 1, 2$, и если $u_k^h(x)$ подчинить условиям (VI.47), то $u_k^h \in P_3 \subset H_1(\alpha, \beta)$. Так как $P^3 = P_3 \times P_3$, то имеем необходимое соотношение $P^3 \subset H(\alpha, \beta)$.

Выражая функцию U^h через узловые параметры и подставляя ее в (VI.48), записываем функционалы $a(U^h, V^h)$, $b(U^h, V^h)$ в виде билинейных форм неизвестных узловых параметров из $[\alpha, \beta]$ (исключаются параметры, определяемые условиями (VI.47)):

$$a(U^h, V^h) = (q(U^h))^T Kq(V^h),$$
$$b(U^h, V^h) = (q(U^h))^T Mq(V^h),$$

где $q(\quad)$ — вектор размера n неизвестных узловых параметров из $[\alpha, \beta)$ для вектор-функции $W^h \in P^3$, K и M — симметричные положительно определенные матрицы порядка n (K — матрица жесткости, M — матрица массы), $n = 4N$ — размерность подпространства P^3.

Приближенная задача на собственные значения записывается так:

$$\lambda_l^h = \min_{U^h \in P^3} a(U^h) \equiv a(V_l^h), \quad l = 1, 2, \ldots, n, \tag{VI.61}$$

при условиях

$$b(U^h) = 1, \tag{VI.62}$$

$$b(U^h, V_k^h) = 0, \quad k = 1, 2, \ldots, l-1, \ l \geqslant 2, \tag{VI.63}$$

где λ_l^h, V_l^h — приближенные собственное число и соответствующая ему собственная вектор-функция. Исходя из (VI.61) — (VI.63) методом неопределенных множителей Лагранжа получаем обобщенную алгебраическую задачу на собственные значения

$$Kq = \lambda^h Mq. \tag{VI.64}$$

В случае задачи для оболочечной модели (VI.46), (VI.51) — (VI.55) прямоугольная область Ω (VI.51) разбивается на прямоугольные элементы

$$E_i = \{x = (x_1, x_2) : x_{1,j_1-1} < x_1 < x_{1,j_1}, \ x_{2,j_2-1} < x_2 < x_{2,j_2}\}.$$

Здесь $x_{k,0} < x_{k,1} < x_{k,2} < \ldots < x_{k,N_{k-1}} < x_{k,N_k}$, $x_{k,0} = \alpha_k$, $x_{k,N_k} = \beta_k$, $i = j_1 + (j_2 - 1) N_1$, $j_k = 1, 2, \ldots, N_k$, $k = 1, 2$. Общее количество таких элементов $N = N_1 \times N_2$. Точки (x_{1,j_1}, x_{2,j_2}) являются вершинами прямоугольников $\bar{E}_i$ и узлами полученной прямоугольной сетки. Она характеризуется величиной

$$h = \max_{j_1, j_2} (h_1^2 + h_2^2)^{1/2}, \tag{VI.65}$$

где

$$h_k = x_{k,j_k} - x_{k,j_k-1}, \quad k = 1, 2.$$

Конечномерное подпространство $P \subset H(\Omega)$ здесь образуем из вектор-функций $U^h = [u_1^h, u_2^h, u_3^h]^T$, компоненты которых на каждом из E_i определяются формулами

$$u_3^h(x) \equiv \psi_{i_3}(x) = \beta_1^{(3)} + \beta_2^{(3)} x_1 + \beta_3^{(3)} x_2 + \beta_4^{(3)} x_1^2 + \beta_5^{(3)} x_1 x_2 + \beta_6^{(3)} x_2^2 +$$
$$+ \beta_7^{(3)} x_1^2 x_2 + \beta_8^{(3)} x_1 x_2^2 + \beta_9^{(3)} x_1^3 + \beta_{10}^{(3)} x_2^3 + \beta_{11}^{(3)} x_1^3 x_2 + \beta_{12}^{(3)} x_1 x_2^3 + \beta_{13}^{(3)} x_1^2 x_2^2 +$$
$$+ \beta_{14}^{(3)} x_1^3 x_2^2 + \beta_{15}^{(3)} x_1^2 x_2^3 + \beta_{16}^{(3)} x_1^3 x_2^3,$$

$$u_k^h(x) \equiv \psi_{i_k}(x) = \beta_1^{(k)} + \beta_2^{(k)} x_1 + \beta_3^{(k)} x_2 +$$

$$+ \beta_4^{(k)} x_1^2 + \beta_5^{(k)} x_1 x_2 + \beta_6^{(k)} x_2^2 + \beta_7^{(k)} x_1^2 x_2 + \beta_8^{(k)} x_1 x_2^2, \quad x \in \bar{E}_i, \quad k = 1, 2.$$

Полиномы $\psi_{i_1}(x)$, $\psi_{i_2}(x)$, $\psi_{i_3}(x)$ определяются однозначно через узловые параметры. Для $\psi_{i_3}(x)$ выбраны значения u_3^h, $\frac{\partial u_3^h}{\partial x_1}$, $\frac{\partial u_3^h}{\partial x_2}$, $\frac{\partial^2 u_3^h}{\partial x_1 \partial x_2}$ в каждом узле сетки. Для $\psi_{i_k}(x)$, $k = 1, 2$, узловыми параметрами служат значения u_k^h и либо $\frac{\partial u_k^h}{\partial x_1}$, либо $\frac{\partial u_k^h}{\partial x_2}$ в каждой вершине элемента $\bar{E}_i$, причем узлы, где фиксируются одноименные производные, расположены в Ω (VI.51) в шахматном порядке (рис. 36). Таким образом, имеем два типа наборов узловых параметров на элементе, а элементы с одинаковыми наборами параметров также располагаются в Ω в шахматном порядке.

Как и в предыдущей задаче, приравниваем одноименные узловые параметры в общих вершинах элементов $\bar{E}_i$, $i = 1, 2, \ldots, N$. Известно (см. например, [101]), что за счет этого обеспечивается условие $u_3^h \in W_2^2(\Omega)$, а для u_k^h, $k = 1, 2$, аналогично можно доказать, что $u_k^h \in W_2^1(\Omega)$. Если потребовать, чтобы на Γ_1 функции u_1^h и u_2^h удовлетворяли условиям (VI.52), а u_3^h — условиям (VI.53), то $u_k^h \in P_k \subset H_k$ ($H_2 \equiv H_1$); здесь P_k — конечномерное множество функций u_k^h, $k = 1, 2, 3$. Так как $P = P_1 \times P_2 \times P_3$, будет выполняться соотношение $P \subset H(\Omega)$.

Аналогично предыдущему можно поставить соответствующую вариационную задачу вида (VI.61) — (VI.63) и получить алгебраическую задачу (VI.64). В этом случае $n = (8N_2 + 1)(N_1 + 1)$.

3. Оценка точности приближенных решений. Используя методику, изложенную в работе [101], можно доказать теоремы о сходимости решений приближенных задач вида (VI.61) — (VI.63) к решениям задач вида (VI.46), (VI.56) — (VI.58).

Теорема VI.2. *Если коэффициенты в выражениях (VI.48) такие, что выполняются соотношения (VI.49), (VI.50) и собственные вектор-функции $V_l = [v_{l_1}, v_{l_2}]^T$ задачи (VI.46), (VI.47), (VI.56) — (VI.58) удовлетворяют условиям*

$$v_{l_k} \in W_2^4(\alpha, \beta), \quad k = 1, 2,$$

то при малых значениях h (VI.60) и $l \leqslant n$ имеют место оценки

$$\lambda_l \leqslant \lambda_l^h \leqslant \lambda_l + c_1 h^4, \quad c_1 > 0,$$
$$\|V_l^h - V_l\|_0 \leqslant c_1' h^4, \quad c_1 > 0. \tag{VI.66}$$

Теорема VI.3. *Если коэффициенты в выражениях (VI.65), (VI.55) такие, что собственные элементы $V_l = [v_{l_1}, v_{l_2}, v_{l_3}]^T$ задачи (VI.46), (VI.51) — (VI.55), (VI.56) — (VI.58) удовлетворяют условиям*

$$v_{l_k} \in W_2^3(\Omega), \quad k = 1, 2, \; v_{l_3} \in W_2^4(\Omega)$$

и выполняются соотношения (VI.49), (VI.50), то при малых значениях h (VI.65) и $l \leqslant n$ справедливы оценки

$$\lambda_l \leqslant \lambda_l^h \leqslant \lambda_l + c_2 h^4, \quad c_2 > 0,$$
$$\| V_l^h - V_l \|_0 \leqslant c_2' h^3, \quad c_2' > 0. \qquad \text{(VI. 67)}$$

Если для дискретизации задачи, описывающей изгибные колебания пластинки, использовать пространство кусочных бикубических полиномов P_3, то при условии $v_{l_3} \in \overset{\circ}{W}{}_2^2(\Omega)$ имеем оценку

$$\| v_{l_3}^h - v_{l_3} \| \leqslant c_3 h^4,$$

а собственные числа оцениваются, как в (VI.67).

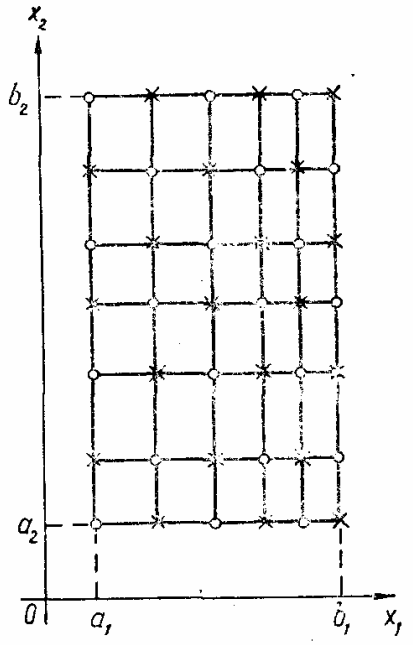

Рис. 36.

Если для дискретизации задачи, описывающей тангенциальные колебания пластинки, использованы указанные выше пространства кусочных полиномов P_1 и P_2, то справедливы оценки (VI.67) при условии $v_{l_1}, v_{l_2} \in \overset{\circ}{W}{}_2^3(\Omega)$, $V_l = [v_{l_1}, v_{l_2}]^T$.

Для решения задачи (VI.64), где матрицы K и M положительно определенные, использованы широко известные методы, упоминаемые в параграфе IV.2. Поскольку погрешность вычисления методом деления отрезка пополам наименьших собственных чисел (представляющих особый интерес для практики) заметно выше, чем наибольших, то целесообразно вместо задачи (VI.64) решать задачу

$$Mq = \mu K q, \quad \lambda^h = \frac{1}{\mu},$$

теми же методами.

4. Численные примеры. Рассмотренные схемы МКЭ применялись для численного решения задач определения частот и форм собственных колебаний различных моделей компрессорных лопаток.

С использованием стержневой модели решались задачи для лопаток со следующими данными:

$$I_1(x) = I_\xi(x)\cos^2\varphi(x) + I_\eta(x)\sin^2\varphi(x),$$
$$I_2(x) = I_\xi(x)\sin^2\varphi(x) + I_\eta(x)\cos^2\varphi(x),$$
$$I_{12}(x) = (I_\eta - I_\xi)\cos\varphi\sin\varphi, \quad Q(x) = \frac{2\sqrt{3}}{5} b\delta(x),$$
$$I_\xi(x) = \frac{16\sqrt{3}}{735} b^3\delta(x),$$
$$I_\eta(x) = \left(\frac{7424}{315\,315} h_0^2 + \frac{9}{385}\delta^2(x)\right)\sqrt{3}\, b\delta(x),$$
$$\delta(x) = (\delta_k - \delta_0)\frac{x}{\beta} + \delta_0, \quad \varphi(x) = \frac{\varphi_0}{\beta} x,$$

Таблица 23

φ_0	h_0	N	ω_1^h	ω_2^h	ω_3^h	ω_4^h
0	0	4	6774,4382	25555,712	32480,314	84217,633
0	0	8	6774,1907	25555,109	32453,363	83782,470
0	0	16	6774,1761	25555,069	32450,848	83737,426
0	1,5	4	6823,3527	25555,712	32792,407	85104,426
0	1,5	8	6823,1081	25555,109	32765,031	84665,612
0	1,5	16	6823,0867	25555,069	32762,497	84619,068
20	1,5	4	6835,1838	24340,619	34474,996	83848,922
20	1,5	8	6834,6892	24336,482	34433,286	83322,263
20	1,5	16	6834,6541	24336,132	34429,934	83271,631
0	5,0	4	7296,0634	25555,712	35757,238	93480,481
0	5,0	8	7295,7954	25555,109	35726,073	92988,601
0	5,0	16	7295,7738	25555,069	35723,360	92937,531

Таблица 24

δ_0	δ_k	N_1	N_2	ω_1^h	ω_2^h	ω_3^h	ω_4^h
10,8	10,8	2	4	4796,1372	23840,485	71715,571	156302,82
10,8	10,8	3	6	4788,5338	23784,728	71506,218	145269,93
10,8	10,8	3	7	4786,9966	23775,523	71470,859	144940,52
10,8	10,8	4	8	4785,9569	23770,126	71450,620	144737,96
7,2	14,4	2	4	6923,5804	28723,949	71094,661	138378,10
7,2	14,4	3	6	6916,1302	28697,160	70989,368	137061,00
7,2	14,4	4	8	6913,6301	28689,524	70956,047	136832,70

Таблица 25

k_{11}	δ_k	δ_0	N_1	N_2	ω_1^h	ω_2^h	ω_4^h	ω_7^h
3,835	11,55	5,774	2	4	2186,558	10371,21	11846,65	47043,93
3,835	11,55	5,774	2	5	2184,461	10351,58	11834,23	46610,53
3,835	11,55	5,774	3	4	2186,219	10364,70	11831,59	46843,72
3,835	11,55	5,774	2	6	2183,282	10342,43	11825,81	46461,44
3,835	11,55	5,774	3	5	2183,921	10345,26	11820,05	46451,91
3,835	11,55	5,774	2	7	2182,822	10339,87	11822,39	46446,12
3,835	10,10	7,217	2	6	1833,777	9814,308	10788,12	46819,14
3,835	8,660	8,660	2	6	1518,695	9091,844	9797,142	45272,65
0	11,55	5,774	2	6	2153,290	10174,30	11635,63	46158,39
1,280	11,55	5,774	2	4	2159,646	10211,98	11698,64	46551,30
1,280	11,55	5,774	2	6	2156,661	10193,39	11671,66	46196,20
7,643	11,55	5,774	2	6	2269,798	10767,50	12089,85	47028,29
12,63	11,55	5,774	2	4	2475,481	10769,49	12489,48	48966,55
12,63	11,55	5,774	2	6	2469,365	10706,89	12466,78	47714,66

$\alpha = 0$, $\beta = 0{,}1$ м, $\rho = 4500$ кг/м^3, $E = 1{,}16 \cdot 10^{11}$ н/м^2, где b — длина хорды профиля лопатки, δ_0 — максимальная толщина сечения $x = \alpha$, δ_k — максимальная толщина сечения $x = \beta$, φ_0 — угол, на который сечение $x = \alpha$ повернуто относительно сечения $x = \beta$, h_0 — максимальное отклонение средней линии профиля лопатки от его хорды. Некоторые результаты решения таких задач при $b = 0{,}05$ м, $\delta_0 = 7{,}2$ мм, $\delta_k = 14{,}4$ мм и различных значениях φ_0 (в градусах) и h_0 (в миллиметрах) приведены в табл. 23. В табл. 23—25 значения $\omega_l^h = (\lambda_l^h)^{1/2}$ даны в радианах в секунду. Все задачи решались при постоянных шагах: $h = \beta/N$ для одномерных задач, $h_k = \beta_k/N_k$, $k = 1, 2$, для двумерных. С использованием пластиночной модели решались задачи для лопаток со следующими характеристиками:

$$\delta(x) = \frac{3\sqrt{3}}{2}\sqrt{\frac{x_1}{\beta_1}}\left(1 - \frac{x_1}{\beta_1}\right)\left[(\delta_k - \delta_0)\frac{x_2}{\beta_2} + \delta_0\right], \quad \text{(VI.68)}$$

$\alpha_1 = \alpha_2 = 0$, $\beta_1 = 0{,}05$ м, $\beta_2 = 0{,}1$ м, $\rho = 4500$ кг/м^3, $E = 1{,}16 \times 10^{11}$ н/м^2, $\nu = 0{,}32$. Здесь δ_0 и δ_k имеют тот же смысл, что и в предыдущем примере. Некоторые результаты решения задачи об изгибных колебаниях пластинки приведены в табл. 24 (δ_0 и δ_k здесь и в табл. 25 даны в миллиметрах).

Оболочечная модель использовалась для лопатки со следующими данными: $\delta(x)$ задается в виде (VI.68), $k_{11} = \text{const}$, $k_{22} = k_{12} = 0$, $A_1 = A_2 = 1$, $\alpha_1 = \alpha_2 = 0$, $\beta_1 = 0{,}0625$ м, $\beta_2 = 0{,}16$ м, $\rho = 4500$ кг/м^3, $E = 1{,}16 \cdot 10^{11}$ н/м^2, $\nu = 0{,}32$. В табл. 25 приведены некоторые результаты решения этой задачи при различных значениях k_{11} (м$^{-1}$), δ_0 и δ_k. На рис. 37 и 38 приведены узловые линии для компоненты $v_{l_3}^h$ некоторых приближенных вектор-функций (при $N_1 = 2$, $N_2 = 6$) в зависимости от k_{11}, δ_0 и δ_k.

На рис. 37 представлены результаты, соответствующие $\delta_0 = 5{,}774$ мм, $\delta_k = 11{,}55$ мм: — — — — при $k_{11} = 1{,}280$ м$^{-1}$; ——— при $k_{11} = 3{,}835$ м$^{-1}$; —.—.— при $k_{11} = 7{,}643$ м$^{-1}$; —.. — .. — — при $k_{11} = 12{,}63$ м$^{-1}$; $v_{l_3}^h \equiv v_{l,3}^h$, $l = 2, 3, 5, 6, 7$.

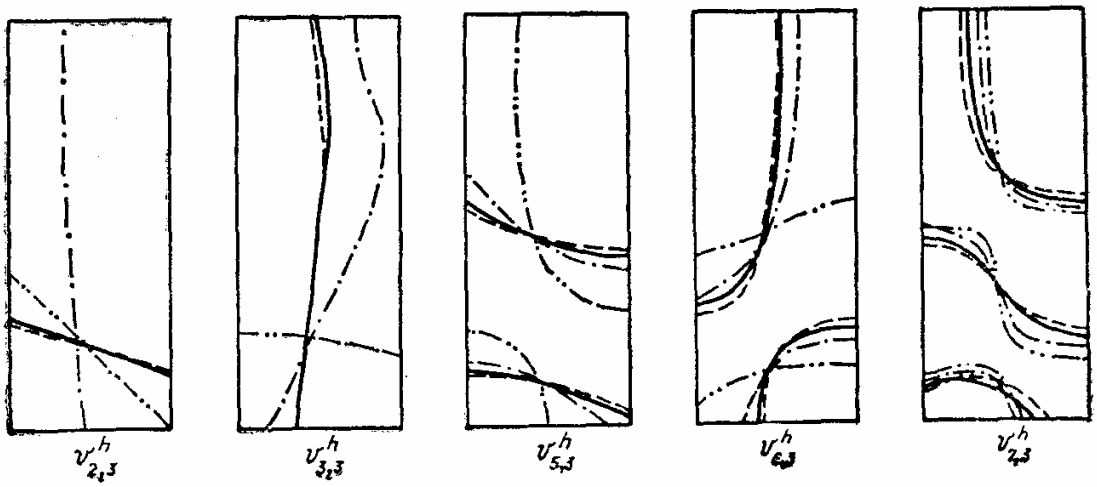

Рис. 37.

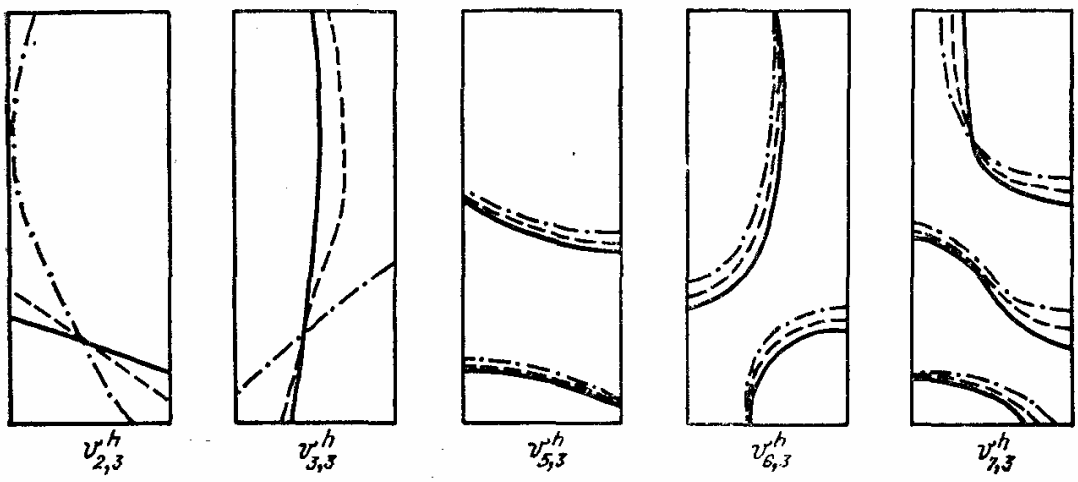

Рис. 38.

На рис. 38 узловые линии представлены для значения $k_{11} = 3{,}835$ м$^{-1}$: —— — при $\delta_0 = 5{,}77$ мм, $\delta_k = 11{,}55$ мм; — — — — при $\delta_0 = 7{,}217$ мм, $\delta_k = 10{,}10$ мм; —·—·— — при $\delta_0 = \delta_k = 8{,}660$ мм.

Анализ полученных результатов показал, что с помощью рассмотренных схем МКЭ с достаточной точностью могут быть вычислены 12—15 минимальных собственных чисел, порядки сходимости для которых вполне удовлетворительно совпадают с теоретическими в оценках (VI.66) и (VI.67). Это устраивает практику, так как при использовании стержневых моделей необходимо 3—4 минимальных собственных числа, а при использовании оболочечной и пластиночных моделей представляют интерес собственные числа, соответствующие частотам до 20 кГц.

Из анализа результатов решения ряда задач при различных моделях лопатки и затраченного машинного времени, учитывая, что рассмотренные здесь схемы имеют четвертый порядок скорости сходимости для собственных чисел, можно сделать вывод о целесообразности применения стержневых моделей как более грубых для предварительных расчетов. Для более точного определения частот и форм собственных колебаний должны использоваться пластиночные (если кривизны k_{11}, k_{22}, k_{12} малы) или оболочечная модели.

VI.3. Расчет упруго-пластического состояния элемента летательного аппарата [82]

1. Постановка задачи. Рассмотрим задачу расчета упруго-пластического состояния обшивки гиперзвукового самолета (рис. 39). Можно выделить из конструкции повторяющийся элемент и все дальнейшие рассуждения проводить относительно этого элемента (рис. 40). Конструкция подвержена силовому воздействию и находится в неравномерном температурном поле. Для построения математической модели исследуемых процессов фигуру рис. 40 мысленно развернем в плоскости xOy (рис. 41).

По границе EE' приложена равномерно распределенная нагрузка N_2, по границе $A'E'$ — нагрузка N_1. Внутренняя граница CC' закреп-

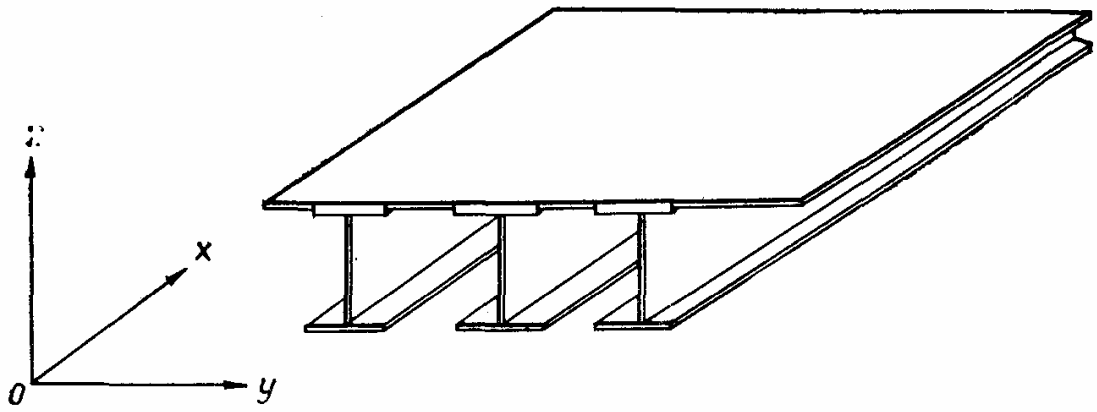

Рис. 39.

лена от перемещений вдоль оси Oy, граница $ABCDE$ от перемещений вдоль оси Ox. Граница AA' свободна от нагрузок и связей. Вследствие того что обшивка представляет собой длинную панель, а фигура на рис. 40 — лишь воображаемый вырез из этой непрерывной конструкции, появляются дополнительные граничные условия: на границе EE' все перемещения вдоль оси Oy должны быть одинаковые, а на границе $A'E'$ — одинаковые перемещения вдоль оси Ox. (Заметим, что толщина составляющих фигуру (рис. 41) элементов I—IV различна и элементы нагреты по-разному.)

Для математической постановки задачи нужно выписать функционал полной энергии системы

$$Э = \int_V \Pi dV + \int_S f dS, \tag{VI.69}$$

где V — область, в которой решается задача, S — граница области V, Π — потенциал деформации, f — внешние нагрузки.

Потенциал деформации выражается формулой

$$\Pi = U + \int_0^{\Gamma} g(\Gamma)\,\Gamma d\Gamma + \frac{E\Gamma^2}{6}. \tag{VI.70}$$

Здесь U — упругая энергия объемного сжатия, в случае пластичности $U = 0$, Γ — интенсивность деформаций, $g(\Gamma)$ — функция связи интенсивности деформаций и интенсивности касательных напряжений σ_i на линейном участке:

$$g(\Gamma) = \frac{\sigma_i}{\Gamma}, \text{ или } \sigma_i = \Gamma g(\Gamma). \tag{VI.71}$$

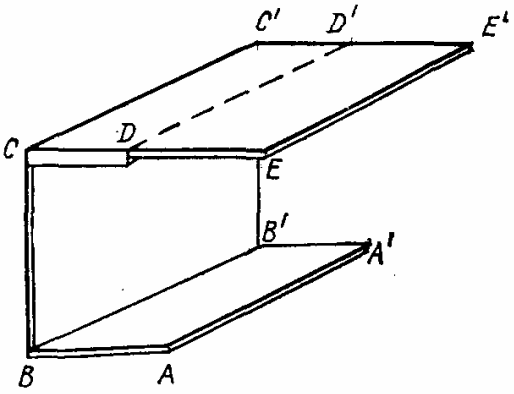

Рис. 40.

Функция (VI.71) в классической теории упруго-пластичности обычно представлена в виде семейства кривых, например таких, как на рис. 42. Данные зависимости получаются при экспериментальном одноосном растяжении цилиндрических образцов различных

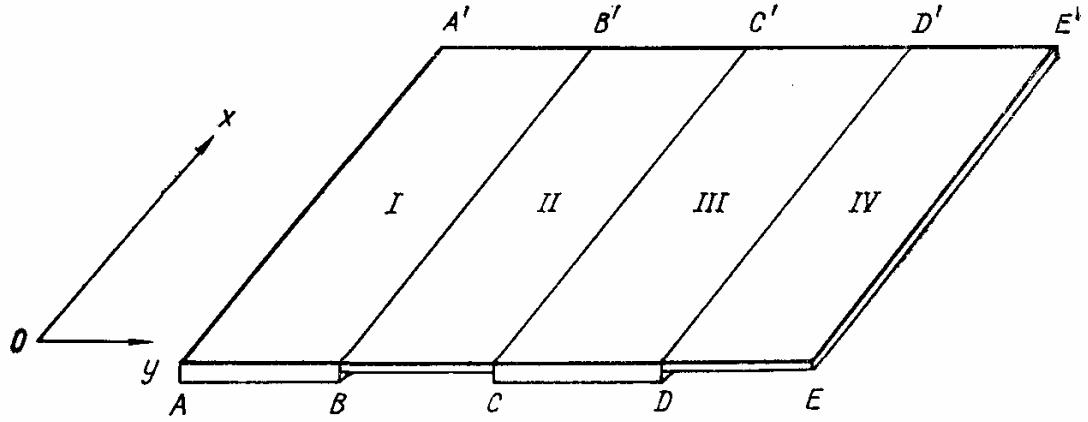

Рис. 41.

несжимаемых материалов в процессе простого нагружения. В этих предположениях считается, что закон одноосного растяжения верен для соотношения интенсивности касательных напряжений σ_i и интенсивности деформаций Γ.

Для математического описания исследуемой задачи требуется аналитический вид закона (VI.71). Получить его можно путем некоторой аппроксимации нелинейного участка семейства кривых (см. рис. 42). Пусть этот закон задан в форме

$$\Gamma g(\Gamma) = 3^{-\frac{m+1}{2}} E \left(\frac{\Gamma}{c}\right)^m, \qquad \text{(VI.72)}$$

где E — модуль упругости материала, c, m — константы, определяемые экспериментально.

Тогда интеграл полной энергии (VI.69) с учетом (VI.70) и (VI.72) будет иметь вид

$$\Э = \int_V E \left(\frac{\Gamma^2}{6} + \frac{\Gamma^{m+1}}{3^{\frac{m+1}{2}} c^m (m+1)} \right) dV + \int_S f dS.$$

Дальнейшие рассуждения по описанию задачи проводятся относительно срединной плоскости (xOy) фигуры, представленной на рис. 41.

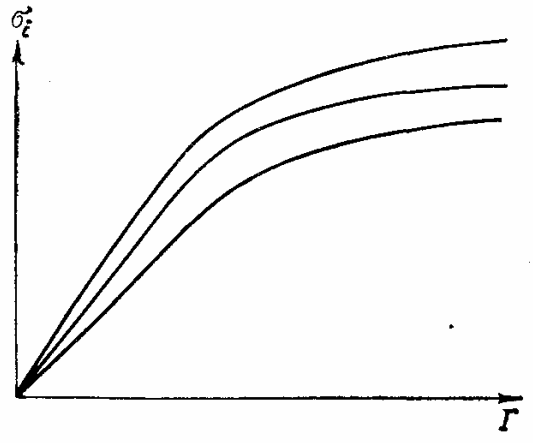

Рис. 42.

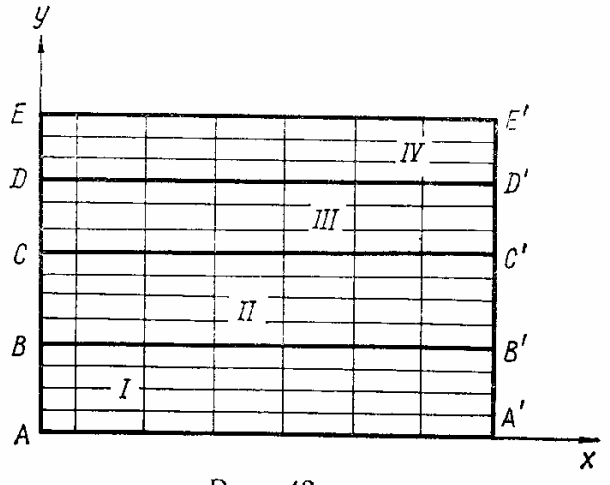

Рис. 43.

Тогда областью Ω определения задачи является прямоугольник $AA'E'E$ (рис. 43).

В случае простого нагружения справедливо соотношение

$$\varepsilon \equiv \varepsilon_{11} + \varepsilon_{22} + \varepsilon_{33} = 3\alpha\theta,$$

где ε_{ij} — компоненты тензора деформаций, α — коэффициент линейного теплового расширения, θ — температура.

Тогда формула для Γ в рамках принятых предположений принимает вид

$$\Gamma = 2\left(\varepsilon_{11}^2 + \varepsilon_{22}^2 + \varepsilon_{11}\varepsilon_{22} + \varepsilon_{12}^2 + 3\alpha^2\theta^2 - 3\alpha\theta\left(\varepsilon_{11} + \varepsilon_{22}\right)\right)^{1/2}. \quad \text{(VI.73)}$$

В свою очередь для компонент деформаций верны соотношения Коши

$$\varepsilon_{11} = \frac{\partial u}{\partial x}, \quad \varepsilon_{22} = \frac{\partial v}{\partial y}, \quad \varepsilon_{12} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right). \quad \text{(VI.74)}$$

Здесь $u(x, y)$ — перемещения вдоль оси Ox, $v(x, y)$ — перемещения вдоль оси Oy. Используя (VI.73) и (VI.74), функционал полной энергии системы (см. рис. 41) в перемещениях можно представить относительно срединной плоскости в форме

$$\Theta(u, v) = \frac{2}{3} E \int_{\Omega} \left[\frac{1}{4}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 + \right.$$

$$\left. + \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} + 3\alpha^2\theta^2 - 3\alpha\theta\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)\right] d\Omega +$$

$$+ \frac{2^{m+1}}{3^{\frac{m+1}{2}}(m+1)c^m} E \int_{\Omega} \left[\frac{1}{4}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 + \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} + \right.$$

$$\left. + 3\alpha^2\theta^2 - 3\alpha\theta\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)\right]^{\frac{m+1}{2}} d\Omega - \int_{A'}^{E'} \frac{N_1}{s_1} u dy - \int_{E}^{E'} \frac{N_2}{s_2} v dx, \quad \text{(VI.75)}$$

где s_1 и s_2 — площадь сечения фигуры вдоль границ $A'E'$ и AE' соответственно.

Отметим, что в описываемой нами конкретной задаче были определены следующие значения констант: $m = 0{,}7$, $c = 50$. Кроме того, предполагалось, что температура θ непрерывно изменялась по линейному закону от 280 °С (участок I) до 450 (участок IV).

Таким образом, задача расчета упруго-пластического состояния элемента конструкции (см. рис. 40) в результате приведенных выше рассуждений и с учетом особенностей физической модели была сведена к решению следующей вариационной задачи.

Найти функции $u(x, y)$, $v(x, y)$, доставляющие минимум функционалу энергии (VI.75) на классах функций из пространства $W_2^1(\Omega)$, удовлетворяющих соответственно условиям

$$u(x, y)|_{AE} = 0, \quad u(x, y)|_{A'E'} = u(AA', 0), \quad \text{(VI.76)}$$

$$v(x, y)|_{CC'} = 0, \quad v(x, y)|_{EE'} = v(0, AE). \quad \text{(VI.77)}$$

2. Дискретизация и численное решение задачи. Построение приближенного решения задачи минимизации функционала (VI.75) при

условиях (VI.76), (VI.77) осуществлялось вариантом МКЭ, основанным на процессе Ритца. Для этого область Ω, изображенная на рис. 43, покрывалась прямоугольной сеткой. Приближенное решение — функции $u^h(x, y)$, $v^h(x, y)$ — строилось в виде кусочно-билинейных полиномов, которые на каждом прямоугольном элементе имели вид

$$u^h(x, y) = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 xy,$$
$$v^h(x, y) = \alpha_5 + \alpha_6 x + \alpha_7 y + \alpha_8 xy. \qquad \text{(VI.78)}$$

Фиксацией значений перемещений u_i, v_i в узлах сетки (в вершинах прямоугольников) было обеспечено однозначное определение билинейного полинома на каждом конечном элементе и непрерывность кусочно-билинейных полиномов на всей области Ω, т. е. принадлежность искомых функций $W_2^1(\Omega)$.

С помощью описанной ранее процедуры искомые коэффициенты α_j, $j = 1, \ldots, 8$, разложения (VI.78) выражаются через узловые перемещения u_i, v_i, которые в свою очередь находятся из системы алгебраических уравнений

$$\frac{\partial \Theta^h}{\partial u_i^h} = 0, \quad \frac{\partial \Theta^h}{\partial v_i^h} = 0, \quad \forall i \in I, \qquad \text{(VI.79)}$$

где $\Theta^h \equiv \Theta(u^h, v^h)$ — функционал (VI.75), представленный в виде суммы интегралов по всем элементам области Ω, I — множество номеров узлов сетки.

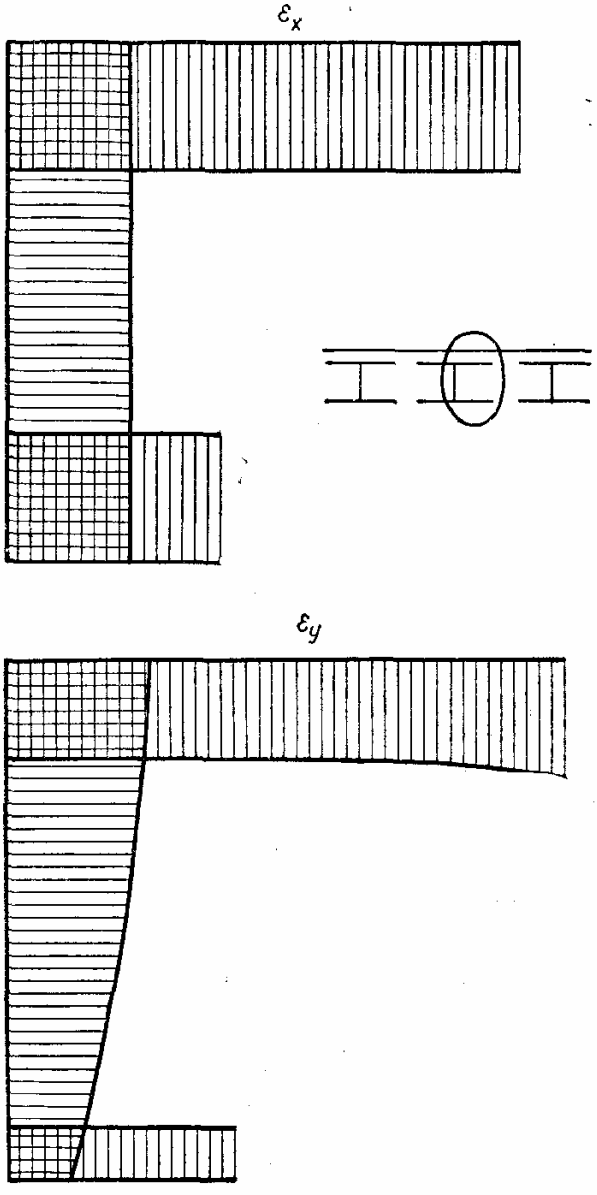

Рис. 44.

Система (VI.79) представляет собой систему нелинейных алгебраических уравнений с симметричной положительно определенной разреженной матрицей Якоби. Симметричность и положительная определенность матрицы обусловлены свойствами самого исходного функционала (а именно его выпуклостью). Удовлетворение условий (VI.76), (VI.94) приводит к тому, что в ленточной матрице Якоби системы (VI.79) появляются симметричные выбросы в строках и столбцах с номером узловой неизвестной u_i^h в точке (AA', O) и с номером неизвестной v_i^h в точке (O, AE). Следует отметить, что при машинной реализации алгоритма ре-

шения система нелинейных алгебраических уравнений содержалась в машине неявно. Фактически же строилась лишь обратная матрица Якоби нелинейной системы, необходимая для решения этой системы по описанному ранее методу квазиньютоновского типа (см. гл. V). Начальным приближением к решению системы нелинейных алгебраических уравнений служило решение линеаризованной исходной задачи. Следует отметить, что квазиньютоновские методы обладают сверхлинейной скоростью сходимости, что позволяет получить решение системы (VI.79) с достаточной точностью за 10—15 итераций.

Таблица 26

Граница	Длина, мм	Толщина, мм
AB	15	2
BC	35	0,6
CD	15	3,2
DE	20	1,2
AA'	300	2

Описанная конкретная задача расчета упруго-пластического состояния обшивки крыла гиперзвукового самолета решалась на машине ЕС 1060 с двойным машинным словом. Геометрические размеры элемента конструкции (см. рис. 40) представлены в табл. 26. Задача решалась относительно срединной плоскости с учетом толщины составляющих конструкцию элементов.

Температура линейно изменялась от 280° С (участок I) до 450 °С (участок IV): $N_1 = 3600$ кГ, $N_2 = 3000$ кГ. При разбиении области Ω

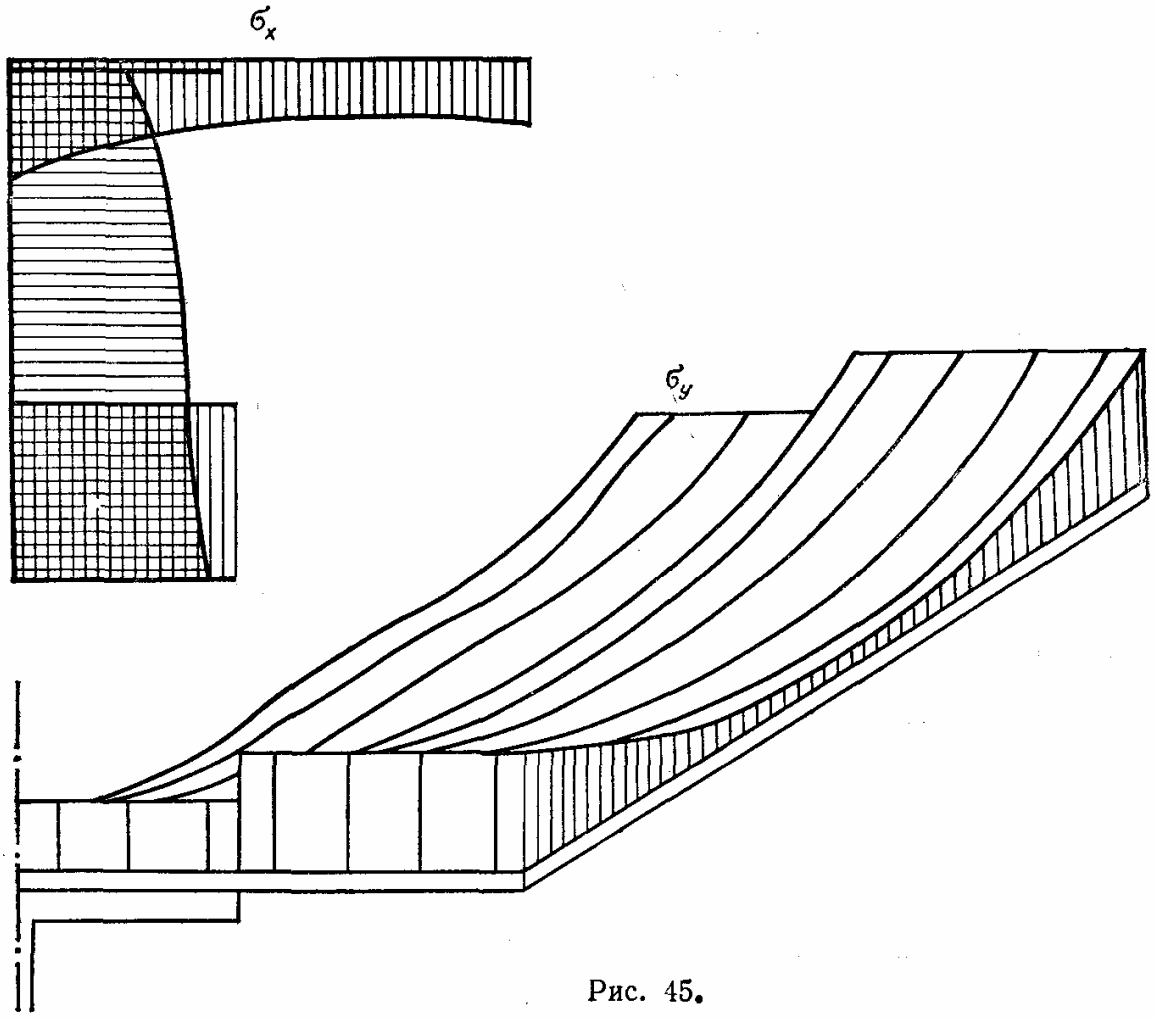

Рис. 45.

(см. рис. 43) на прямоугольники вдоль оси Ox принимался шаг $h_1 = 60$ мм, вдоль оси Oy шаг $h_2 = 5$ мм. Численное интегрирование при дискретизации вариационной задачи проводилось по квадратурным формулам Гаусса с тремя узлами. Время счета задачи 10 мин. Согласно полученным результатам были построены графики распределения полей деформаций (рис. 44) и напряжений (рис. 45) конструкции, иллюстрирующие возникающие зоны пластичности.

Результаты счета хорошо согласуются с практикой эксперимента.

VI.4. Расчет на прочность имитационной модели самолета в целом

Покажем в данном параграфе некоторые характерные особенности решения больших систем линейных алгебраических уравнений МКЭ.

1. Постановка задачи. Пусть требуется выполнить расчет на прочность толстой пластины сложной геометрической формы. Срединная плоскость пластины совпадает с плоскостью xOy декартовой системы координат (x, y, z); геометрические размеры ее в метрах указаны на рис. 46. Толщина пластины — кусочно-постоянная, материал — изотропный.

Пластина находится под действием равных по величине дискретных сил R_i, $i = 1 \div 4$, точки приложения A_i которых отмечены на рис. 46. Направление действия сил и размещение точек их приложения выбраны из условия взаимного уравновешивания сил. Единственность вектора перемещений $U(x, y, z) = [u, v, w]^T$ обеспечивается закреплением пластины в двух точках P и Q (см. рис. 46).

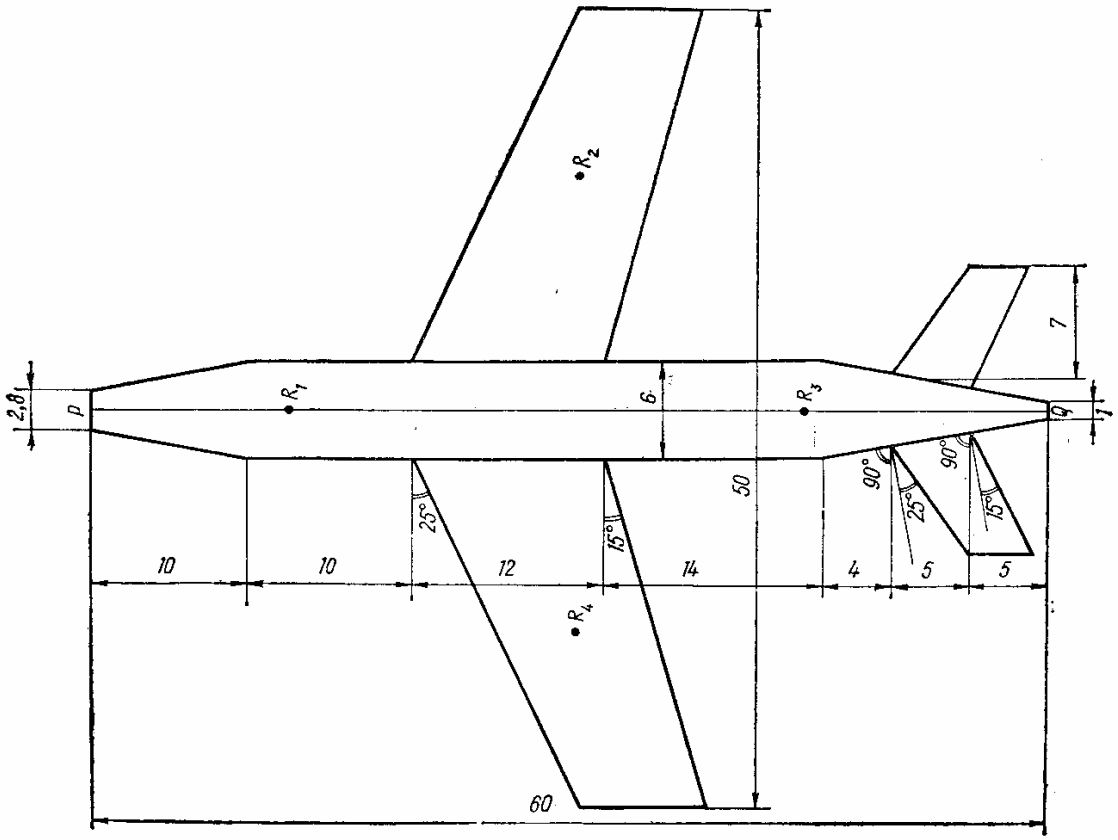

Рис. 46.

В соответствии с теорией толстых пластин математическая постановка задачи сводится к решению вариационной задачи об отыскании вектор-функции $U^* = [u^*, v^*, w^*]^T$, доставляющей минимум функционалу

$$I(U) = \frac{1}{2}\int\limits_{\Omega}\left[\mu'\left(\frac{\partial u}{\partial x}\right)^2 + 2\lambda'\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} + \mu'\left(\frac{\partial v}{\partial y}\right)^2 + \mu\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2 + \right.$$

$$\left. + \mu\left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}\right)^2 + \mu\left(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z}\right)^2\right] d\Omega - \sum_{i=1}^{4} R_i w(A_i) \quad \text{(VI.80)}$$

на множестве вектор-функций $U \in H(\Omega)$.

Здесь Ω — трехмерная область, срединная плоскость которой представлена на рис. 46, пространство $H(\Omega) = H_1 \times H_1 \times H_1$, где $H_1 = H_1(\Omega)$ — множество функций $u(x, y, z)$ (соответственно $v(x, y, z)$ $w(x, y, z)$), принадлежащих пространству Соболева $W_2^1(\Omega)$ и удовлетворяющих условию

$$u(P) = u(Q) = 0,$$

$$(v(P) = v(Q) = 0,\ w(P) = w(Q) = 0), \quad \text{(VI.81)}$$

$$\mu' = \frac{E}{1-\nu^2},\ \lambda' = \frac{E\nu}{1-\nu^2},\ \mu = \frac{E}{2(1+\nu)},$$

где E — модуль Юнга, ν — коэффициент Пуассона.

2. Дискретизация задачи. При дискретизации вариационной задачи (VI.80), (VI.81) применен подход, используемый для расчета толстостенных оболочек (см. [39]). Каждый из элементов, на которые разбивается пластина, образован четырьмя плоскими сечениями, перпендикулярными срединной плоскости, и представляет собой прямую призму. Высота каждого элемента постоянная и равна толщине пластины t. Такой элемент однозначно описывается четырьмя точками $i_{\text{ср}}$ — вершинами соответствующего четырехугольника на срединной плоскости (рис. 47) — и толщиной t_k пластины в данном (k-м) элементе; точки $i_{\text{ср}}$ задаются своими декартовыми координатами (x_i, y_i), $i = 1, 2, 3, 4$, ($z_i = 0$ — $i_{\text{ср}}$ принадлежит срединной плоскости). Согласно [39] вводится для каждого элемента локальная система координат (ξ, η, ζ): плоскость $\xi O\eta$ совпадает со срединной плоскостью элемента, а ось $O\zeta$ совпадает с направлением оси Oz; предполагается, что значения ξ, η, ζ изменяются от -1 до $+1$ на соответствующих поверхностях каждого элемента. Зависимость между координатами (x, y, z) и (ξ, η, ζ) для любой точки k-го элемента имеет вид

$$x = \sum_{i=1}^{4} x_i \varphi_i(\xi, \eta),\ y = \sum_{i=1}^{4} y_i \varphi_i(\xi, \eta),$$

$$z = \frac{\zeta t_k}{2}\sum_{i=1}^{4}\varphi_i(\xi, \eta). \quad \text{(VI.82)}$$

Здесь $\varphi_i(\xi, \eta)$ — кусочно-билинейная базисная функция МКЭ, рассматриваемая на каноническом квадрате и отвечающая его i-му, $i = 1, 2, 3, 4$, узлу (рис. 48).

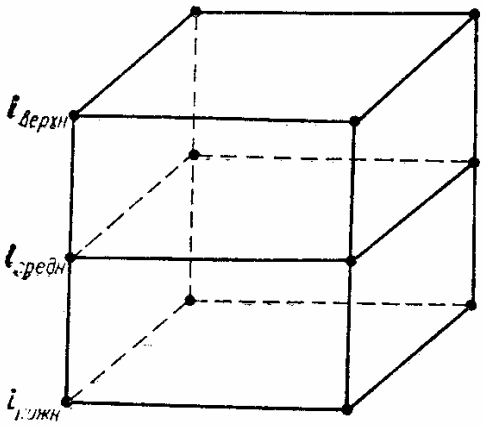

Рис. 47.

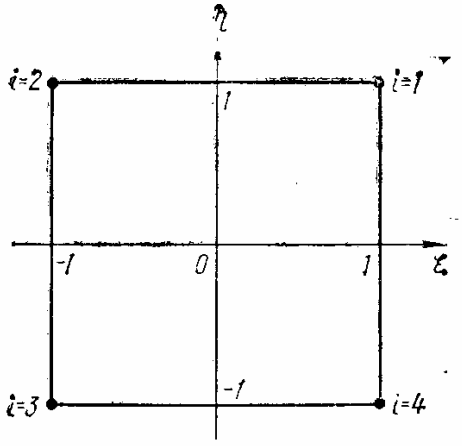

Рис. 48.

Обозначив через (ξ_i, η_i) координаты i-го узла, можно представить общий вид функций $\varphi_i(\xi, \eta)$, $i = 1, 2, 3, 4$, на каноническом квадратном элементе формулой

$$\varphi_i(\xi, \eta) = \frac{1}{4}(1 + \xi\xi_i)(1 + \eta\eta_i). \qquad \text{(VI.83)}$$

Если в i-й, $i = 1, 2, 3, 4$, узловой точке срединной плоскости элемента определить (зафиксировать) по пять параметров, а именно u_i, v_i, w_i, т. е. декартовы компоненты узлового перемещения срединной плоскости в i-м узле, α_i, β_i — углы поворота нормали к срединной плоскости в i-м узле вокруг оси y и оси x соответственно, то согласно [39] перемещения $U = [u, v, w]^T$ в любой точке элемента можно найти по формулам

$$u = \sum_{i=1}^{4} u_i\varphi_i(\xi, \eta) + \frac{t_k\zeta}{2}\sum_{i=1}^{4}\alpha_i\varphi_i(\xi, \eta), \qquad \text{(VI.84)}$$

$$v = \sum_{i=1}^{4} v_i\varphi_i(\xi, \eta) - \frac{t_k\zeta}{2}\sum_{i=1}^{4}\beta_i\varphi_i(\xi, \eta), \quad w = \sum_{i=1}^{4} w_i\varphi_i(\xi, \eta),$$

где t_k — толщина пластины в данном (k-м) элементе.

Таким образом, задача об отыскании вектор-функции $U^*(x, y, z)$ минимизирующей функционал (VI.80) при условии (VI.81), свелась к задаче отыскания обобщенных узловых перемещений (u_i, v_i, w_i, α_i, β_i) в срединной плоскости пластины.

Используя взаимосвязь локальных и глобальных координат, квадратурные формулы Гаусса с восемью узлами на «каноническом» кубе (координаты вершин которого суть $(\pm 1, \pm 1, \pm 1)$) получили согласно (VI.80), (VI.82) — (VI.84) по обычному алгоритму элементарные матрицы жесткости (двадцатого порядка) относительно зафиксированных в срединной плоскости узловых параметров, а затем построили общую систему линейных алгебраических уравнений МКЭ.

При решении данной задачи вся область $\overline{\Omega}$ была разбита на 8600 элементов. На рис. 49 для каждой подобласти Ω указано количество элементов, расположенных по соответствующим направлениям.

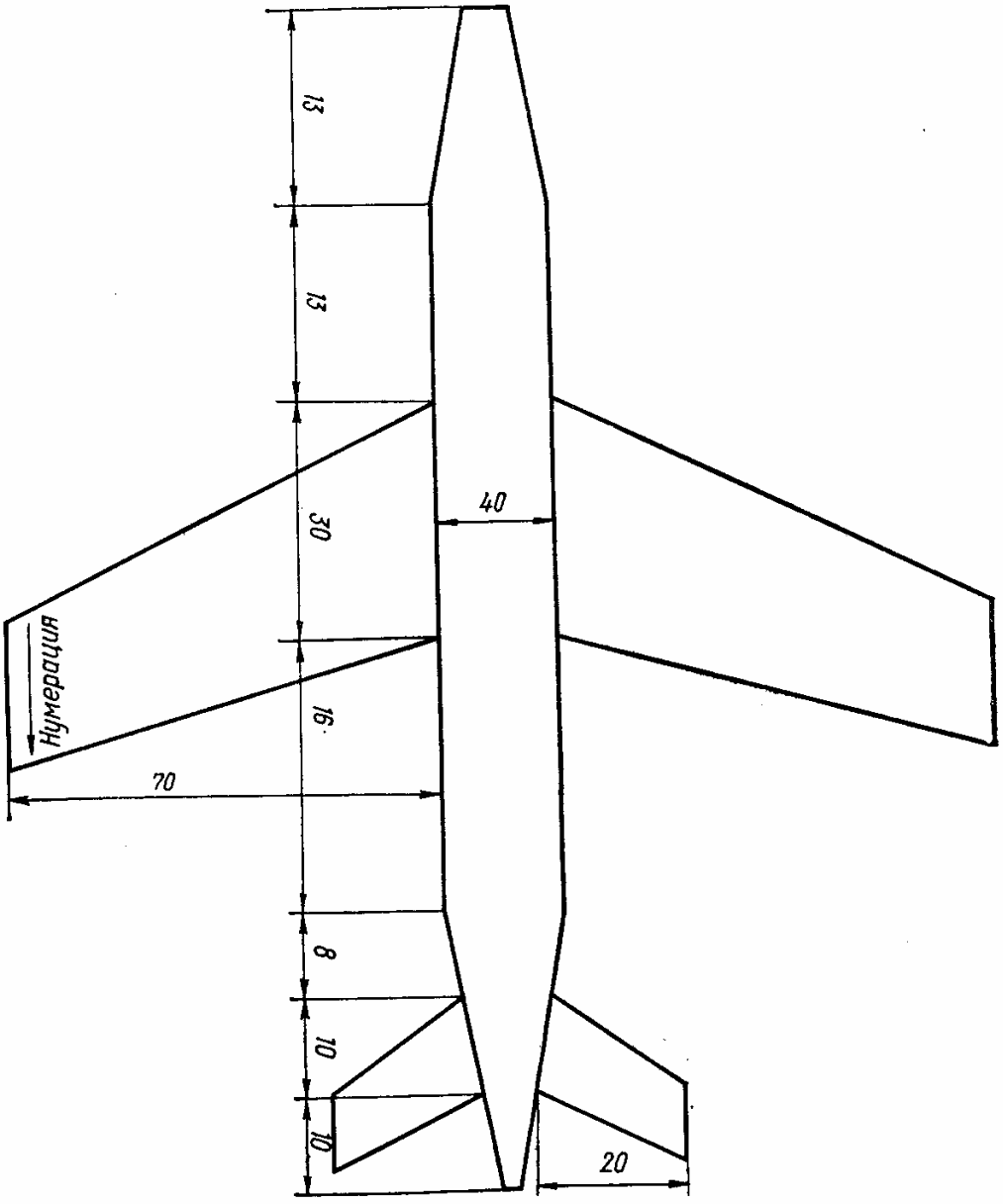

Рис. 49.

Общее количество узлов на срединной плоскости пластины 8921. Согласно соотношению (VI.84) выполнение условия закрепления пластины (VI.81) будет обеспечено, если в узлах P и Q (рис. 46) положить

$$u(P) = u(Q) = 0,\ v(P) = v(Q) = 0,$$
$$w(P) = w(Q) = 0, \qquad \text{(VI.85)}$$
$$\alpha(P) = \alpha(Q) = 0,\ \beta(P) = \beta(Q) = 0,$$

что и учитывалось в дальнейшем расчете.

Исходная глобальная нумерация узлов срединной плоскости проводилась по направлению, параллельному прямой PQ, начиная с места, указанного стрелкой на рис. 49. Это обеспечивало более узкую ширину ленты матрицы системы уравнений МКЭ. В результате с учетом условий (VI.85) была получена система алгебраических уравнений с симметричной положительно определенной ленточной матрицей 44 595-го порядка, лента которой имеет переменную ширину и малоза-

полнена: максимальное количество ненулевых элементов во всей ленточной строке равно 45, а максимальная полуширина ленты — 515 (включая главную диагональ).

3. Решение системы уравнений МКЭ. Рассмотрим теперь особенности вычисления решения большой системы линейных алгебраических уравнений

$$Ax = b, \qquad \text{(VI.86)}$$

описанной в предыдущем пункте. Решалась система на старших моделях ЕС ЭВМ (ЕС 1060, 1066). В связи с тем что теоретические оценки числа обусловленности матрицы A дают величину $O(10^7)$, вычисления выполнялись при удвоенной длине машинного слова (8 байт).

Отметим прежде всего, что в нижнем треугольнике (включая главную диагональ) симметричной и положительно определенной матрицы A имеется несколько менее 10^6 ненулевых элементов, так что только для их хранения необходим объем памяти порядка 8 Мбайт.

Если заданную систему решать одним из наиболее эффективных прямых методов — методом квадратных корней, то в процессе факторизации $A = LL^T$, где L — нижняя треугольная матрица, возникает много новых ненулевых элементов в тех позициях, где у A прежде были нули (произойдет заполнение ленты). При машинной реализации соответствующего вычислительного алгоритма, даже с учетом переменной ширины ленты (профильный алгоритм), для хранения ленты только матрицы L потребуется уже порядка 116 Мбайт, т. е. более одного магнитного диска с объемом хранимой информации 100 Мбайт. Очевидно, что решение системы (VI.86) данным алгоритмом на ЕС 1060 с использованием нескольких таких дисков является нереальным: машинное время будет неприемлемо большим, особенно если учесть среднее время наработки на отказ ЭВМ ЕС и надежность периферийных устройств.

Как известно (см., например, [23]), в настоящее время существуют различные процедуры переупорядочения строк и столбцов разреженной симметричной положительно определенной матрицы, обеспечивающие уменьшение ширины ее ленты (или уменьшение размера профиля). Заметим, что для систем уравнений МКЭ любая процедура упорядочения матрицы равносильна некоторой перенумерации узлов сетки (см. п. 2 параграфа I.3, рис 16). Однако попытка упорядочения матрицы A системы (VI.86) с целью сужения ее ленты (уменьшения размера профиля) не увенчалась успехом: первоначально выбранная нумерация узлов введенной сетки (см. рис. 49) обеспечила минимальную ширину ленты.

Можно было бы попытаться использовать какую-нибудь другую процедуру упорядочения (см. [13, 23, 114]), чтобы обеспечить эффективность применения прямого метода к системе (VI.86).

Однако при их практической реализации с использованием внешней памяти возникает ряд трудностей, снижающих эффективность этих процедур [51, 113]. Поэтому при вычислении решения на ЭВМ ЕС 1060 с дисками объемом 100 Мбайт более целесообразным представлялся все-таки отказ от «чисто» прямого метода в пользу одного из мето-

дов сопряженных направлений, а именно — метода неполного разложения [15].

Суть этого метода в нашем случае сводится к следующему. Симметричная положительно определенная матрица A системы (VI.86) раскладывается так:

$$A = M + N,$$

где $M = M^T$, $N = N^T$, причем матрица M — невырожденная и представлена в форме, обеспечивающей сравнительно «недорогое» вычисление решения системы $Mz = f$. Итерационный процесс решения системы (VI.86) организуется следующим образом. Для подходящего начального вектора x_0 находится невязка $r_0 = Ax_0 - b$ и решается система $Mu_1 = r_0$, а далее вычислительный процесс осуществляется по схеме

$$r_i = r_{i-1} + \alpha_i A u_i,\ Mv_i = r_i,\ u_{i+1} = v_i + \beta_i u_i,\ x_i = x_{i-1} + \alpha_i u_i,$$
$$i = 1, 2, \ldots, \qquad \text{(VI.87)}$$

где

$$\alpha_i = -\frac{(r_{i-1}, u_i)}{(Au_i, u_i)},\ \beta_i = \frac{(v_i, r_i)}{(v_{i-1}, r_{i-1})}.$$

Заметим, что в рассматриваемом случае начальный вектор x_0 называется подходящим, если он обеспечивает выполнение условия $(v_i, r_i) \neq 0$ при всех i, для которых $r_i \neq 0$. Если матрица M не только симметрична, но и положительно определенная, то любой вектор x_0 является подходящим. В противном случае для произвольного x_0 не гарантируется отсутствие вырождений в процессе (VI.87). На практике в качестве x_0 можно брать почти любой вектор; если же все-таки случится обрыв процесса, то его нужно возобновить с другого начального вектора.

Как известно (см. [15]), искомое решение будет получено по процессу (VI.87) особенно быстро, если матрица имеет простую структуру и мало попарно различных собственных чисел. Обеспечить такие свойства матрице $M^{-1}A$ можно, например, в том случае, если в представлении $A = M + N$ матрица M будет «близка» к A, так что большинство собственных чисел матрицы N и, следовательно, $M^{-1}N$ будут равны нулю или близки к нулю: $M^{-1}A = E + M^{-1}N$; в «пределе» $M^{-1} = A^{-1}$.

Замечание. Получение решения системы (VI.86) по схеме (VI.87) с симметричной и положительно определенной матрицей M равносильно решению обычным методом сопряженных градиентов специально «подготовленной» системы

$$\tilde{A}y = \tilde{b}, \qquad \text{(VI.88)}$$

где $\tilde{A} = L^{-1}AL^{-T}$,

$$M = LL^T,\ y = L^T x,\ \tilde{b} = L^{-1}b.$$

Процедуру перехода от системы (VI.86) к системе (VI.88) называют преобусловливанием системы, а матрицу M — преобусловливателем.

Схему (VI.87) с указанной матрицей M называют иногда методом сопряженных градиентов с преобусловливанием.

Если матрица M близка к матрице A в указанном выше смысле, то обычный метод сопряженных градиентов будет работать гораздо быстрее в случае системы (VI.88), чем исходной системы (VI.86), а следовательно, схема (VI.87) более эффективна, чем обычный метод сопряженных градиентов.

Вернемся к описанию процесса решения нашей конкретной системы МКЭ (VI.86). Для вычислительной схемы (VI.87) матрица M строилась с помощью неполной факторизации матрицы A (вариантом метода квадратных корней):

$$A = LDL^T + N, \quad M = LDL^T \approx A,$$

где L — нижняя треугольная матрица с единичной главной диагональю, D — диагональная матрица, N — матрица погрешности неполной факторизации.

Процесс неполной факторизации выполнялся таким образом, что после завершения вычисления всех элементов очередной строки матрицы L (и D) в этой строке L оставлялось только 100 наибольших по модулю элементов. В результате была построена нижняя треугольная матрица L, количество ненулевых элементов которой равнялось 3 894 996. Матрица M хранится в факторизованном виде $M = LDL^T$, что удобно для реализации процесса (VI.87).

Заметим, что в настоящее время для процедуры неполной факторизации заданной матрицы предложены различные подходы и различные критерии отбрасывания элементов в строках L (см. обзор [122]). При решении системы (VI.86) использовалась процедура из [78]. Полученная по ней матрица M не является знакоопределенной.

Коротко об организации вычислений на ЭВМ. Решение системы (VI.86) выполнялось на ЕС 1060 с двумя объемом 100 Мбайт дисками.

Информация о нижнем треугольнике матрицы A (включая главную диагональ) хранилась в трех массивах: массиве значений ненулевых элементов; массиве, указывающем номер столбца для каждого ненулевого элемента; массиве, указывающем общее количество ненулевых элементов во всех предшествующих строках нижнего треугольника, включая данную. Аналогично хранится информация о матрице L. Элементы матриц A и L хранились как двойные машинные слова.

Вся информация, необходимая для решения прикладной задачи, занимала один диск. Другой диск использовался для хранения некоторой вспомогательной информации, необходимой для эффективной организации вычислительного процесса.

Для построения матрицы M и выполнения процесса (VI.87) элементы нижнего треугольника матрицы A вызывались в оперативную память поблочно: матрица была построчно разбита на 14 блоков, доступ к которым — последовательный. Все программы были разработаны с учетом возможности их выполнения с разрывом во времени, что позволяло противостоять машинным сбоям и давало возможность управлять процессом решения.

Таблица 27

Этап расчета	Время выполнения общее (*CPU*)		Используемая оперативная память, кбайт
	ЕС 1060	ЕС 1066	
Построение разбиения	2′ (1′01″)	1′ (12″)	972
Построение алгебраической системы	75′ (69′33″)	17′ (16′05″)	3766
Неполная факторизация	482′ (461′16″)	153′ (142,′10″)	7124
Вычисление по итерационной схеме (25 итераций)	281′ (68′45″)	229′ (15′31″)	6234
Итого	840′ (600′)	400′ (174′)	

Построенная по упомянутому способу матрица M обеспечила получение решения системы (VI.86) по схеме (VI.87) за 25 итераций. Критерием окончания процесса на i-й итерации служило условие $\| r_i \|_E \leqslant \leqslant \varepsilon = 10^{-8}$. (Решение, полученное при $\varepsilon = 10^{-11}$, отличалось от предыдущего в шестой значащей цифре.) Вычисленное решение хорошо согласуется с представлениями инженерной практики.

Как уже упоминалось, основная цель описываемого расчета состояла не в исследовании физических свойств конкретного объекта, а в оценке затрат вычислительных ресурсов на машинную реализацию расчета и в поиске путей сокращения этих затрат. В табл. 27 приводятся данные, характеризующие эти затраты на выполнение всех этапов расчета.

Таким образом, на этапе решения описанным выше способом уже сформированной системы (VI.86) потребовалось иметь максимальный используемый объем оперативной памяти 7124 кбайт, а затраты машинного времени составили: 530 мин (*CPU*) и 763 мин (общее) на ЕС 1060, 157 мин 41 с (*CPU*) и 382 мин (общее) на ЕС 1066.

В заключение отметим, что на ЕС 1066 с двумя дисками объемом 200 Мбайт решение системы (VI.86) было получено и по программе пакета *APAC* [75], реализующей ленточный алгоритм варианта метода квадратных корней (LDL^T-разложение). Матрица при этом разбивалась на 53 блока, которые хранились в файле последовательного доступа одного диска, а факторизованная матрица хранилась на втором диске. Для вычисления решения системы (VI.86) в этом случае потребовалось 6912 кбайт оперативной памяти, 301 мин общего машинного времени и 181 мин 26 с — времени *CPU*.

Глава VII

МАШИННЫЕ МЕТОДЫ РЕШЕНИЯ НЕКОТОРЫХ КЛАССОВ МАТЕМАТИЧЕСКИХ ЗАДАЧ

Ряд прикладных задач может быть сведен методом конечных элементов к решению систем линейных алгебраических уравнений, задаче нахождения собственных чисел и собственных векторов матриц, решению систем нелинейных уравнений, задачам Коши для систем обыкновенных дифференциальных уравнений. В настоящей главе рассмотрим трудности машинного решения этих дискретных математических задач.

Необходимо с самого начала подчеркнуть, что машинное решение математической задачи в ряде случаев может значительно отличаться от точного (математического) решения. Поэтому очень важен вопрос не только вычисления машинного решения, но и оценки его близости к математическому решению.

Следуя [79], для найденного решения рассмотрим оценки погрешностей, зависящих от неточности исходных данных, используемого вычислительного метода и суммарного влияния ошибок округлений, возникающих при реализации вычислений на ЭВМ. (Погрешность метода конечных элементов как средства получения дискретных задач была рассмотрена в предыдущих главах.)

VII.1. Системы линейных алгебраических уравнений с квадратными вещественными матрицами

1. Постановка задач и некоторые определения. В практических задачах крайне редко возникают системы

$$\tilde{A}\tilde{x} = \tilde{b} \tag{VII.1}$$

с точными исходными данными. Наиболее типичным является задание приближенной системы

$$Ax = b \tag{VII.2}$$

с указанием погрешности в исходных данных

$$\|\tilde{A} - A\| = \|\Delta A\| \leqslant \varepsilon_A, \quad \|\tilde{b} - b\| = \|\Delta b\| \leqslant \varepsilon_b. \tag{VII.3}$$

Таким образом, в качестве формального решения задачи (VII.1) — (VII.3) может рассматриваться любой вектор, который обращает в тождество уравнение (VII.2) с матрицей A' и вектором b', удовлетворяющими неравенствам (VII.3). Отметим, что в случае вырожденной ($\det \tilde{A} = 0$) матрицы точной системы (VII.1) приближенная система (VII.2) может оказаться несовместной при любой точности задания исходных данных.

Для удобства точное решение $\tilde{x}$ системы (VII.1) с точными, но неизвестными исходными данными будем условно называть физическим решением, а точное решение x заданной системы (VII.2) — математическим решением задачи.

Погрешность в решении x, вызванную неточным заданием исходных данных, называют наследственной погрешностью. Значение ее зависит как от погрешности исходных данных, так и от свойств матрицы.

Решение системы уравнений (VII.2), получаемое некоторым численным методом на ЭВМ, будем называть машинным решением задачи. Вследствие погрешности перевода числовых исходных данных из десятичной системы в двоичную, погрешности вычислительного алгоритма и погрешности его машинной реализации вычисленное (машинное) решение, как правило, отличается от точного математического решения задачи. Для получения достоверного решения задачи принципиально важны классификация и определение свойств решаемой системы.

Как известно, задача является корректно поставленной, если: решение задачи существует и единственно при любых входных данных из некоторой области их изменения; решение задачи непрерывно зависит от исходных данных.

Отыскание классического решения x системы (VII.2), (VII.3), т. е. вектора x, для которого невязка $r = b - Ax$ тождественно равна нулю, будет корректно поставленной задачей, если $\det A \neq 0$ и

$$\| \Delta A A^{-1} \| < 1, \text{ или } \| \Delta A \| \| A^{-1} \| < 1 \qquad \text{(VII.4)}$$

при произвольном возмущении ΔA из окрестности, определяемой (VII.3).

Например, задача отыскания классического решения системы

$$\begin{aligned} 25x_1 &- 36x_2 = 1, \\ 16x_1 &- 23x_2 = -1 \end{aligned}$$

с приближенными исходными данными является корректно поставленной, если погрешность задания элементов матрицы будет удовлетворять условию $\| \Delta A \|_\infty \leqslant 0{,}015$.

Действительно, поскольку

$$A^{-1} = \begin{bmatrix} -23 & 36 \\ -16 & 25 \end{bmatrix}, \text{ т. е. } \| A^{-1} \|_\infty = 59,$$

условие (VII.4) выполняется, а следовательно, гарантируется невырожденность матрицы при любых допустимых возмущениях исходных данных и непрерывная зависимость решения от этих данных. Точное решение заданной системы следующее: $x_1 = -59$, $x_2 = -41$.

Однако если исходные данные этой системы будут изменяться в пределах $\| \Delta A \|_\infty \leqslant 0{,}02$, $\| \Delta b \| \leqslant 0{,}01$, то задачу следует рассматривать как некорректную: условие (VII.4) не выполняется. Действительно, в пределах указанных возмущений исходных данных можно получить несовместную систему

$$25{,}01x_1 - 35{,}99x_2 = 1,$$
$$15{,}99x_1 - 23{,}01x_2 = -1$$

или совместную

$$25{,}01x_1 - 36x_2 = 1,$$
$$15{,}99x_1 - 23{,}01x_2 = -1,$$

единственное точное решение которой есть $x_1 = -369{,}04315$, $x_2 = -256{,}41025$ (в пределах восьми значащих цифр).

Для корректной постановки задачи об отыскании решения системы линейных алгебраических уравнений (VII.2) с вырожденной матрицей (что может случиться, например, при решении по МКЭ прикладной задачи, математической моделью которой служит соответствующая задача Неймана) необходимо уточнить понятие искомого решения.

Обобщенным решением в смысле наименьших квадратов системы (VII.2) называется любой вектор x_0, для которого евклидова норма невязки достигает своего наименьшего значения

$$\min_x \| b - Ax \|_E^2 = \| b - Ax_0 \|_E^2.$$

Решение $x_н$, имеющее наименьшую евклидову норму:

$$\min_{x_0} \| x_0 \|_E = \| x_н \|_E,$$

называют нормальным обобщенным решением.

Решение $x_н$ можно определить как обобщенное решение, ортогональное всем линейно независимым классическим решениям однородной системы v_j:

$$Av = 0,$$

т. е.

$$(x_н, v_j) \equiv x_н^T v_j = 0, \; j = 1, 2, \ldots, k.$$

Нормальное обобщенное решение всегда существует и единственно. Очевидно, что нормальное обобщенное решение совпадает с классическим в случае $\det A \neq 0$.

Легко убедиться, что обобщенные решения и только они являются классическими решениями системы

$$A^T Ax = A^T b.$$

Для системы с квадратной вырожденной матрицей корректная постановка задачи об отыскании нормального решения требует дальнейшего уточнения. Это объясняется тем, что в данном случае как угодно малые возмущения элементов матрицы могут вызывать изменение ранга матрицы и, следовательно, скачкообразные (не непрерывные) изменения решения.

В работах [103, 104] рассматриваются различные постановки задач о решении вырожденных систем линейных алгебраических уравнений с приближенными исходными данными, а в работах [18, 71, 79, 103, 104, 109] предлагаются специальные методы их решения.

2. **Классификация корректно поставленных задач.** Корректно поставленные задачи о решении систем линейных алгебраических уравнений в зависимости от устойчивости решения к погрешностям в исходных данных можно разделить на хорошо и плохо обусловленные.

Приведем некоторые примеры, характеризующие зависимость наследственной погрешности от свойств матрицы и погрешности в исходных данных. Так, в системах

$$2x_1 - x_2 = 0, \qquad 2x_1 - x_2 = 0,$$
$$-x_1 + 2x_2 = 3, \qquad -x_1 + 2x_2 = 3{,}003$$

правые части различаются между собой в четвертой значащей цифре. Решение первой из них есть $x_1 = 1$, $x_2 = 2$, а второй — $x_1 = 1{,}001$, $x_2 = 2{,}002$, т. е. они отличаются тоже в четвертой значащей цифре.

Правые части систем уравнений

$$100x_1 + 500x_2 = 1700, \qquad 100x_1 + 500x_2 = 1700,$$
$$15x_1 + 75{,}01x_2 = 255, \qquad 15x_1 + 75{,}01x_2 = 255{,}03$$

различаются в пятой значащей цифре, а их решения не имеют ничего общего: у первой из них $x_1 = 17$, $x_2 = 0$, а у второй $x_1 = 2$, $x_2 = 3$. Отметим, что определитель этих двух систем равен единице.

При решении корректно поставленных задач наряду с получением единственного классического решения задачи возникает необходимость в оценке наследственной погрешности или близости математического и физического решений задачи.

Верхнюю границу «относительной» наследственной погрешности точного решения x системы (VII.2) можно выразить через «относительные» погрешности задания матрицы A и вектора b как

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leqslant \frac{\|A\|\|A^{-1}\|}{1 - \|\Delta A\|\|A^{-1}\|}\left[\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right], \qquad \text{(VII.5)}$$

или

$$\frac{\|x - \tilde{x}\|}{\|\tilde{x}\|} \leqslant \frac{\|A\|\|A^{-1}\|}{1 - \dfrac{\|\Delta b\|}{\|b\|}}\left[\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right], \qquad \text{(VII.6)}$$

при условии $\|\Delta A\|\|A^{-1}\| < 1$ и естественном предположении $\dfrac{\|\Delta b\|}{\|b\|} < 1$. Обе оценки являются мажорантными. Однако они **неулучшаемы (достижимы)** на всем классе невырожденных матриц.

Из формул (VII.5), (VII.6) очевидно, что устойчивость решения к изменениям исходных данных в значительной степени зависит от величины $\operatorname{cond} A = \|A\|\,\|A^{-1}\|$, которая называется **числом** обусловленности матрицы. Если значение $\operatorname{cond} A$ невелико, то матрица системы линейных алгебраических уравнений называется хорошо

обусловленной. Если значение $\operatorname{cond} A$ велико, то матрица системы называется плохо обусловленной. В зависимости от способов введения норм матрицы рассматривают несколько видов чисел обусловленности. Например, для симметричных матриц

$$\operatorname{cond} A \equiv \frac{\max\limits_i |\lambda_i|}{\min\limits_i |\lambda_i|},$$

где λ_i — собственные числа ($i = 1, 2, \ldots, n$) матрицы A; для несимметричных матриц

$$\operatorname{cond} A = \sqrt{\frac{\mu_n}{\mu_1}},$$

где μ_n и μ_1 — наибольшее и наименьшее собственные числа матрицы A^TA. Некоторые другие числа обусловленности рассматривались в работе [107—108].

Из оценок (VII.5) и (VII.6) следует, что значение наследственной погрешности математического решения зависит не от абсолютного значения числа обусловленности матрицы $\operatorname{cond} A$ или погрешностей исходных данных системы, а от значения

$$m = \operatorname{cond} A \left[\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right]. \qquad \text{(VII.7)}$$

Эту величину m будем называть числом обусловленности системы линейных алгебраических уравнений. Формула (VII.7) связывает воедино свойства матрицы системы и погрешность в задании исходных данных. В реальных задачах есть смысл рассматривать те системы, для которых оценки m заметно меньше единицы, например $m \leqslant 0{,}01$.

Из формулы (VII.7) следует, что для уменьшения наследственной погрешности в математическом решении необходимо стремиться к уменьшению числа m за счет либо увеличения точности задания исходных данных, либо уменьшения $\operatorname{cond} A$ (путем переформулирования исходной математической задачи).

Итак, рассмотрение устойчивости решения к изменению исходных данных для систем линейных алгебраических уравнений с квадратной невырожденной матрицей A, удовлетворяющей условию (VII.7), позволяет выделить хорошо и плохо обусловленные системы.

Из рассмотренного выше видно, что еще до решения системы может сложиться ситуация, при которой точное решение системы, хранящейся в памяти ЭВМ, может не иметь ничего общего с физическим решением задачи.

Реализация численных методов на ЭВМ вносит свою погрешность, которую необходимо учитывать при анализе машинного решения задачи.

3. Погрешность реализации вычислительных алгоритмов на ЭВМ. Здесь и в дальнейшем при анализе машинной реализации вычислительных алгоритмов будем иметь в виду некоторую абстрактную ЭВМ, которая производит вычисления в двоичной системе счисления в режиме с плавающей запятой. При вычислениях с плавающей запятой

каждое ненулевое число x представимо в виде $x = 2^b a$, где b — целое (положительное или отрицательное) число и $\frac{1}{2} \leqslant |a| < 1$. Число b называют порядком, a — мантиссой числа x. Будем предполагать, что a имеет t цифр после двоичной запятой, т. е. данная ЭВМ имеет «t-разрядное слово», и что нуль имеет нестандартное представление, в котором $a = b = 0$.

При введении почти всех чисел в ЭВМ в числа вносится некоторая ошибка, связанная с их округлением и стандартной формой представления, характерной для каждой ЭВМ. Обозначим через $\mathrm{fl}(x)$ конечную дробь, которая получается после округления мантиссы числа x до t-го разряда после запятой. Тогда

$$\mathrm{fl}(x) \equiv x(1 + \varepsilon).$$

Если $\mathrm{fl}(x) \neq 0$, то $|\varepsilon| \leqslant 2^{-t}$; если $\mathrm{fl}(x) = 0$, но $x \neq 0$, то $\varepsilon = -1$

Равенство вида

$$z = \mathrm{fl}\begin{pmatrix} + \\ - \\ x \times y \\ : \end{pmatrix}$$

будет означать, что x, y и z — стандартные числа с плавающей запятой и что z получено из x и y выполнением соответствующей операции с плавающей запятой. Будем предполагать, что ошибки округления в этих операциях таковы, что

$$z = \mathrm{fl}\begin{pmatrix} + \\ - \\ x \times y \\ : \end{pmatrix} = \begin{pmatrix} + \\ - \\ x \times y \\ : \end{pmatrix}(1 + \varepsilon),$$

где

$$|\varepsilon| \leqslant 2^{-t}, \text{ если } \mathrm{fl}\begin{pmatrix} + \\ - \\ x \times y \\ : \end{pmatrix} \neq 0;$$

$$\varepsilon = -1, \text{ если } \mathrm{fl}\begin{pmatrix} + \\ - \\ x \times y \\ : \end{pmatrix} = 0, \text{ но } \begin{pmatrix} + \\ - \\ x \times y \\ : \end{pmatrix} \neq 0.$$

Суммарное влияние ошибок округления на машинное решение рассмотрим на следующем простом примере. Пусть требуется решить систему линейных алгебраических уравнений с симметричной положитель-

но определенной матрицей

$$\begin{aligned} 0{,}135x_1 + 0{,}188x_2 + 0{,}191x_3 + 0{,}178x_4 &= 0{,}3516,\\ 0{,}188x_1 + 0{,}262x_2 + 0{,}265x_3 + 0{,}247x_4 &= 0{,}4887,\\ 0{,}191x_1 + 0{,}265x_2 + 0{,}281x_3 + 0{,}266x_4 &= 0{,}5105,\\ 0{,}178x_1 + 0{,}247x_2 + 0{,}266x_3 + 0{,}255x_4 &= 0{,}4818. \end{aligned} \qquad \text{(VII.8)}$$

В результате решения этой системы методом LL^T-разложения (методом квадратных корней) на ЭВМ ЕС с одинарной точностью получим машинное решение

$$x^{(1)} = \begin{bmatrix} 0{,}408\,930\,72 \\ 0{,}494\,605\,78 \\ 0{,}597\,760\,26 \\ 0{,}501\,327\,40 \end{bmatrix}. \qquad \text{(VII.9)}$$

Однако, как нетрудно убедиться, математическое решение системы (VII.8)

$$x = \begin{bmatrix} 0{,}4 \\ 0{,}5 \\ 0{,}6 \\ 0{,}5 \end{bmatrix}. \qquad \text{(VII.10)}$$

Рассмотрим некоторые причины такого явления. Одним из источников погрешности в $x^{(1)}$ являются ошибки округления, возникающие при вводе информации в ЭВМ (погрешность перевода из одной системы счисления в другую, использование элементарных функций, которые в ЭВМ вычисляются по приближенным формулам и т. д.). Например, в результате ввода в ЭВМ ЕС с одинарной точностью элементов матрицы A и правой части b системы (VII.8) в памяти машины будут храниться следующие массивы данных в шестнадцатиричной системе счисления:

$$\bar{A} = \begin{bmatrix} 402285C & 403020C4 & 4030E560 & 4029168 \\ 403020C4 & 4043126E & 404370A & 4033B64 \\ 4030E560 & 404370A & 4047E9 & 40441893 \\ 4029168 & 4033B64 & 40441893 & 404147AE \end{bmatrix},$$

$$\bar{b} = [405A0275 \quad 471B71 \quad 4082B020 \quad 407B573E]^T.$$

Заметим, что здесь для представления шестнадцатиричных цифр используются следующие обозначения:

$$A = 10, \quad B = 11, \quad C = 12, \quad D = 13, \quad E = 14, \quad F = 15.$$

Далее, первые две цифры каждого из приведенных шестнадцатиричных чисел обозначают, как принято в ЭВМ ЕС, «смещенный» порядок числа, т. е. $40 = 16 \cdot 4 + 0 = 64$, что соответствует нулевому «несмещенному» порядку числа с плавающей запятой (подробнее о представлении чисел в ЭВМ ЕС см. в работе [43]).

Массивам $\bar{A}$ и $\bar{b}$ шестнадцатиричных чисел соответствуют (с точностью до восьми десятичных значащих цифр) следующие массивы де-

сятичных чисел:

$$\begin{bmatrix} 0{,}13499999 & 0{,}18799996 & 0{,}19099998 & 0{,}17799997 \\ 0{,}18799996 & 0{,}26199996 & 0{,}26499999 & 0{,}24699998 \\ 0{,}19099998 & 0{,}26499999 & 0{,}28099996 & 0{,}26599997 \\ 0{,}17799997 & 0{,}24699998 & 0{,}26599997 & 0{,}25500000 \end{bmatrix},$$

$$\begin{bmatrix} 0{,}35159999 \\ 0{,}48869997 \\ 0{,}51049995 \\ 0{,}48179996 \end{bmatrix}.$$

Таким образом, после ввода матрицы и вектора правых частей системы в ЭВМ нужно, точнее говоря, рассматривать уже не систему (VII.8), а некоторую возмущенную

$$\bar{A}\bar{x} = \bar{b}, \tag{VII.11}$$

которую будем называть машинной задачей. Отметим, что точное решение системы (VII.11) в пределах семи значащих цифр

$$\bar{x} = \begin{bmatrix} 0{,}5174050 \\ 0{,}4290924 \\ 0{,}5705147 \\ 0{,}5174870 \end{bmatrix} \tag{VII.12}$$

и отличается от математического решения (VII.10).

Соответствующее искажение исходных данных систем происходит и в случае, когда коэффициенты и свободные члены машинной задачи вычисляются (как в МКЭ) по определенным, подчас громоздким формулам. В этом случае погрешность исходных данных складывается из ошибок округления, допускаемых при вводе числовой информации в ЭВМ, и из суммарных ошибок округления, сопровождающих каждую арифметическую операцию расчета по формулам. Указанные возмущения исходных данных системы являются причиной наследственной погрешности машинного решения. Вторым источником погрешности машинного решения может стать погрешность используемого вычислительного метода.

Для решения систем линейных алгебраических уравнений применяются прямые и итерационные методы. Прямые методы дают решение системы за конечное число арифметических операций. Если все операции выполняются точно (без ошибок округления), то вычисленное решение заданной системы будет точным, поэтому эти методы иногда называют точными, не имеющими погрешности.

Итерационные методы являются приближенными: они дают решение системы как предел приближений, вычисленных некоторым единообразным процессом. Поскольку на практике итерационный процесс всегда конечен, то полученное решение обязательно содержит некоторую погрешность, даже если все арифметические операции выполнялись точно. Это и есть погрешность машинного решения, зависящая от

метода, используемого для численного решения системы линейных алгебраических уравнений. К вопросу точности машинного решения, полученного итерационным методом, вернемся еще в дальнейшем изложении, а сейчас рассмотрим еще один (третий) источник погрешности машинного решения — ошибки округления, сопровождающие реализацию на ЭВМ любого вычислительного алгоритма.

Начнем с машинной реализации алгоритмов прямых методов. Решая систему (VII.8) (в действительности — машинную систему (VII.11)) методом квадратных корней на ЭВМ ЕС с одинарной точностью, вследствие накопления ошибок округления при арифметических операциях получаем решение (VII.9), которое отличается от точного решения системы (VII.11).

Если вычисление решения системы (VII.11) выполнить с удвоенной точностью, то тем же методом квадратных корней можно получить результат, совпадающий с (VII.12) в пределах семи значащих цифр, но не совпадающий с точным решением (VII.10) системы (VII.8).

Если же и ввод данных системы (VII.8) и вычисление решения методом квадратных корней осуществим с удвоенной точностью ЭВМ ЕС, то получим (в пределах десяти значащих цифр)

$$x^{(1)} = \begin{bmatrix} 0{,}4000000001 \\ 0{,}4999999999 \\ 0{,}6000000000 \\ 0{,}5000000000 \end{bmatrix},$$

что вполне хорошо согласуется с (VII.10).

Суммарное влияние ошибок округления в прямых методах можно рассматривать (см. [105]) как соответствующее эквивалентное возмущение исходных данных. Поэтому вычисленное машинное решение $x^{(1)}$ системы (VII.2) является точным для некоторой возмущенной системы, например

$$(A + dA)\, x^{(1)} = b + db,$$

или

$$(A + F)\, x^{(1)} = b, \qquad \text{(VII.13)}$$

и приближенным, не совпадающим с математическим решением системы (VII.2). Здесь dA, db, F — соответствующие эквивалентные возмущения, зависящие от метода решения, порядка системы, длины мантиссы t машинного слова.

Справедливы следующие оценки:

$$\frac{\|x^{(1)} - x\|}{\|x^{(1)}\|} \leqslant \frac{\|A\|\|A^{-1}\|}{1 - \dfrac{\|db\|}{\|b\|}} \left[\frac{\|dA\|}{\|A\|} + \frac{\|db\|}{\|b\|}\right], \qquad \text{(VII.14)}$$

$$\frac{\|x^{(1)} - x\|}{\|x\|} \leqslant \frac{\|A\|\|A^{-1}\|}{1 - \|dA\|\|A\|} \left[\frac{\|dA\|}{\|A\|} + \frac{\|db\|}{\|b\|}\right], \qquad \text{(VII.15)}$$

$$\frac{\|x^{(1)} - x\|}{\|x^{(1)}\|} \leqslant \|F\|\|A^{-1}\|. \qquad \text{(VII.16)}$$

Оценки (VII.14) — (VII.16) свидетельствуют о том, что погрешности машинной реализации особенно опасны для систем линейных алгебраических уравнений с большим числом обусловленности матрицы.

Рассмотрение этих оценок позволяет говорить о хорошей или плохой обусловленности задачи в связи с численной устойчивостью машинного решения к суммарным ошибкам округления. Но здесь понятие хорошей или плохой «машинной» обусловленности системы тесно связано с возможностями конкретной машины.

В результате одна и та же система может классифицироваться для одной машины или одной длины мантиссы машинного слова как плохо «машинно обусловленная», а для другой — как хорошо «машинно обусловленная». Увеличивая длину машинного слова, всегда можно получить машинное решение, достаточно близкое к математическому решению задачи.

Теперь рассмотрим некоторые вопросы машинной реализации итерационных методов. В итерационных методах отличие машинного от математического решения может определяться как погрешностью округления арифметических операций на каждом шаге итерационного процесса, так и условиями окончания итерационных процессов. Покажем влияние машинной реализации итерационных процессов на точность получаемых результатов на примере решения системы уравнений с симметричной положительно определенной матрицей A (и правой частью b) явным итерационным методом:

$$x^{(k+1)} = x^{(k)} + \tau r^{(k)}, \quad r^{(k)} = b - Ax^{(k)}, \quad \tau = \frac{2}{\delta + \Delta}. \qquad \text{(VII.17)}$$

Здесь $\delta \leqslant \lambda_i \leqslant \Delta$, где λ_i, $i = 1, 2, \ldots, n$ — собственные числа матрицы. С учетом ошибок округления справедлива оценка

$$\| x - \bar{\bar{x}}^{(k)} \| \leqslant \frac{\tau}{1-q} [q^k \| Ax^{(0)} - b \|] + \frac{1-q^k}{1-q} \tau\omega,$$

где $\bar{\bar{x}}^{(k)}$ — машинное решение, полученное на k-й итерации, $q = \| E - \tau A \|$, E — единичная матрица, $x^{(0)} \equiv x_0$ — начальное приближение, $\omega = \max\limits_{1 \leqslant i \leqslant k} \| \varepsilon_i \|$, ε_i — ошибки округления i-го шага итерационного процесса. Отметим, что $\bar{\bar{x}}^{(k)}$ точно удовлетворяет итерационной схеме (VII.17) с возмущенной правой частью, причем это возмущение имеет вид

$$db_k = \sum_{p=0}^{k} ((E - \tau A)^p)^{-1} \beta_k,$$

где β_k есть суммарная ошибка округления k-го шага.

При использовании итерационного процесса всегда предполагается, что процесс останавливается на каком-то k-м приближении. Окончание итерационного процесса (VII.17) при выполнении условия

$$\max_i \frac{| x_i^{(k+1)} - x_i^{(k)} |}{| x_i^{(k)} |} \leqslant \tau \frac{\lambda_{\min} \varepsilon}{1 + \varepsilon}, \quad x_i^{(k)} \neq 0$$

(где ε — некоторое заданное положительное число) гарантирует справедливость неравенства

$$\max_i \frac{|x_i - x_i^{(k+1)}|}{|x_i|} \leqslant \varepsilon, \quad x_i \neq 0.$$

Здесь x_i — компоненты точного решения заданной системы. Другие условия окончания различных итерационных процессов приведены в работе [72]. Отметим, что для рассматриваемого в (VII.17) итерационного процесса справедливо соотношение

$$\operatorname{cond} A \approx \frac{2}{\ln \dfrac{\| x^{(k+1)} - x^{(k)} \|}{\| x^{(k)} - x^{(k-1)} \|}}.$$

Увеличение длины машинного слова при использовании итерационных процессов позволяет, как это было видно выше, снизить погрешность машинного решения.

Необходимо отметить два принципиальных момента, касающихся увеличения длины машинного слова при решении системы линейных алгебраических уравнений. Во-первых, для систем с невырожденными матрицами (для которых выполнено условие (VII.4)) никакое повышение разрядности вычислений не может приблизить математическое решение x к физическому решению $\tilde{x}$. Во-вторых, для некорректно поставленных задач изменение длины машинного слова не может помочь в построении решения обычными методами, даже если исходные данные системы будут точными, а система совместной. В подобных задачах не существует «классического» единственного решения. При рассмотрении таких задач необходимо доопределить понятие искомого решения и использовать специальные алгоритмы его построения.

4. Характеристика некоторых методов и программ решения систем линейных алгебраических уравнений. Для решения систем линейных алгебраических уравнений в настоящее время разработаны машинные методы (см., например, [14, 23, 79, 96, 106]) и программы, удовлетворяющие самым высоким требованиям. Так, для решения достаточно широкого класса систем линейных алгебраических уравнений имеется возможность использования программ из пакетов *LINPACK* [133] и *APAC* [75].

В этих пакетах представлены программные средства для решения систем с квадратными матрицами общего вида, симметричными положительно определенными (плотными, ленточными, некоторыми разреженными), ленточными общего вида и т. п. Имеются средства и для решения систем с вырожденными матрицами.

В пакете программ *APAC* (в отличие от *LINPACK*) предусмотрена возможность решения заданной системы в автоматическом режиме, при котором в процессе решения исследуются свойства системы, находится подходящий алгоритм решения, решается задача и оценивается достоверность полученного решения. Кроме того, в пакете *APAC* имеются программы, использующие в процессе решения дисковую па-

мять ЭВМ ЕС. Это позволяет, например, на ЕС 1060 (аналог *IBM* 370/168) с дисками объемом 100 Мбайт решить систему с плотной матрицей 1200-го порядка за 120 мин 17,23 с (*CPU*) на оперативной памяти (ОП) объемом 1294 кбайт, а систему с ленточной положительно определенной матрицей 28 000-го порядка при половине ширины ленты 200 — за 128 мин 48 с (*CPU*), ОП — 750 кбайт.

Применение МКЭ для решения современных прикладных задач в ряде случаев приводит к возникновению очень больших систем линейных алгебраических уравнений с разреженными матрицами (порядок — до нескольких десятков тысяч). Для эффективного решения таких больших задач в настоящее время широко применяются два подхода, конкретный выбор которых зависит от типа используемого метода решения — прямой или итерационный.

Решение больших разреженных систем прямыми методами с приемлемыми затратами вычислительных ресурсов предполагает максимальный учет разреженности исходной матрицы системы и минимизацию ее заполнения в процессе решения, т. е. ограничение появления новых ненулевых элементов там, где прежде были нули. Оказалось, что для большинства разреженных систем возможна такая перенумерация переменных и перестановка уравнений (т. е. такое упорядочение строк и столбцов исходной матрицы), которая приводит к существенному сокращению заполнения, а следовательно, к экономии памяти и машинного времени решения. Особенно эффективно проблема подходящего упорядочения решается для разреженных систем с симметричными положительно определенными матрицами, так как в данном случае обеспечена численная устойчивость процесса вычисления решения.

Проиллюстрировать эффективность и роль процедуры переупорядочения можно на следующем простейшем примере. Симметричная и положительно определенная матрица системы имеет структуру вида

$$\begin{bmatrix} X & X & X & X & X & X \\ X & X & & & & \\ X & & X & & 0 & \\ X & & & X & & \\ X & & 0 & & X & \\ X & & & & & X \end{bmatrix},$$

где X — ненулевые элементы; порядок матрицы — 3000, количество ненулевых элементов в нижнем треугольнике, включая главную диагональ,— 5999.

Если для решения системы с такой матрицей применить стандартную программу метода квадратных корней (профильный алгоритм), то в процессе факторизации нижняя треугольная матрица превратится в плотную, и для хранения ее элементов потребуется 4 501 500 машинных слов. В результате предпринятого упорядочения (предназначенного для уменьшения профиля) исходная матрица была преобра-

зована к виду

$$\begin{bmatrix}
X & & & & & & & X & 0 \\
 & X & & & & & & X & 0 \\
 & & X & & & & & X & 0 \\
 & & & X & & 0 & & X & 0 \\
 & & & & X & & & X & 0 \\
 & 0 & & & & X & & X & 0 \\
 & & & & & & X & X & 0 \\
X & X & X & X & X & X & X & X & X \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & X & X
\end{bmatrix}.$$

Теперь при ее факторизации тем же профильным алгоритмом метода квадратных корней для хранения необходимой информации о нижней треугольной матрице понадобилось только 5999 слов, т. е. в данном случае не появилось ни одного добавочного ненулевого элемента. Для упорядочения и вычисления решения заданной системы при двойном машинном слове на ЕС 1060 потребовалось 4,97 с (*CPU*) и оперативная память 204 кбайт. Несомненно, что столь «совершенное» упорядочение на практике случается исключительно редко; тем не менее для многих больших разреженных систем можно найти достаточно удачное упорядочение, чтобы при использовании соответствующего вычислительного алгоритма сделать процесс решения эффективным.

Как известно, структура разреженности матрицы конечно-элементной системы линейных алгебраических уравнений существенно зависит от нумерации узлов сетки. Поэтому проблема упорядочения матрицы является проблемой соответствующей (подходящей для данного алгоритма решения) нумерации узлов. Важно отметить, что найти эффективное упорядочение («удачную» нумерацию узлов) можно до реального вычисления элементов матрицы и ее реальной факторизации, работая только с соответствующим (конечно-элементным) графом матрицы, что значительно упрощает весь процесс решения задачи. Подробное описание подхода к решению больших разреженных систем прямыми методами можно найти, например, в работах [23, 114]. В настоящее время имеется также ряд пакетов программ, реализующих этот подход (см., например, описание пакета *SPARSPAK* в [23], а также других пакетов в [37, 38]).

Второй подход к эффективному решению больших разреженных систем связан с использованием итерационных методов и сводится к предварительной подготовке заданной системы для применения итерационного процесса. Процедура подготовки известна под названием преобусловливание, и ее подробное описание можно найти в работах [26, 121, 136, 139, 146, 152]. Преобусловливание состоит в переходе от исходной системы к эквивалентной, но с таким изменением спектра матрицы, что для преобразованной системы становится эффективным применение итерационного метода, скорость сходимости которого зависит от числа обусловленности матрицы системы. Явное или неяв-

ное преобразование исходной системы линейных алгебраических уравнений в эквивалентную ей, но с другим спектром матрицы в разных работах называют по-разному: масштабированием, нормализацией, преобусловливанием, регуляризацией и т. д. Связь между преобусловливанием и построением быстросходящихся итерационных процессов решения систем разностных уравнений, по-видимому, впервые была установлена в работе [26].

Описание результатов решения одной большой разреженной системы с использованием процедуры преобусловливания приведено в параграфе VI.4 данной монографии.

В настоящее время для решения систем линейных алгебраических уравнений, возникающих при использовании МКЭ и метода конечных разностей, разработаны и широко применяются многосеточные итерационные процессы [110, 111, 140, 141]. Характерной особенностью их является одновременное использование нескольких сеток для построения каждого последующего приближения к искомому решению. Самой мелкой сетке соответствует основная дискретная задача, решение которой и является искомым. Для того чтобы избежать непосредственного решения этой большой системы, предлагается в итерационном процессе использовать еще несколько аналогичных вспомогательных дискретных задач, соответствующих более грубым сеткам. Многосеточные итерационные процессы отличаются высокой скоростью сходимости и достаточно экономичны: количество вычислений на одну итерацию, как правило, пропорционально числу узлов сетки; искомое решение с точностью ε может быть получено за $O(|\log \varepsilon| n)$ операций, где n — порядок основной системы.

5. Оценки достоверности решений, полученных прямыми методами. В пунктах 2, 3 рассматривались теоретические аспекты близости математического к физическому и машинного к математическому решениям задачи. Рассмотрим теперь некоторые практические приемы, позволяющие оценить достоверность полученного на ЭВМ решения $x^{(1)}$. Для упрощения будем считать, что погрешность перевода исходных данных из десятичной системы в двоичную равна нулю («машинная» система совпадает с системой (VII.2)). Рассмотрение невязки $r = b - Ax^{(1)}$ для оценки точности вычисленного решения иногда оказывается бесполезным, так как далекие от точного решения векторы могут давать очень малые невязки. Это бывает в системах, у которых норма обратной матрицы велика. Действительно, если точное решение системы (VII.2) представим в виде $x = x^{(1)} + \delta$, то для вычисления поправки δ получим систему

$$A\delta = b - Ax^{(1)} \equiv r^{(1)}.$$

Отсюда следует, что $\delta = A^{-1}r^{(1)}$ и $\|\delta\| \leqslant \|A^{-1}\| \, \|r^{(1)}\|$. Очевидно, даже при малой невязке погрешность δ в решении может оказаться весьма большой, если норма $\|A^{-1}\|$ велика.

Получить информацию о близости машинного решения $x^{(1)}$ к математическому решению x, а в ряде случаев устранить погрешность, вызванную суммарным эффектом ошибок округления, можно с помощью

итерационного процесса [105, 106]

$$r^{(s)} = b - Ax^{(s)}, \quad QP\delta^{(s)} = r^{(s)}, \quad x^{(s+1)} = x^{(s)} + \delta^{(s)}, \quad s = 0, 1, 2, \ldots, \quad x^{(0)} \equiv 0, \quad \delta^{(0)} \equiv x^{(1)}. \tag{VII.18}$$

Здесь поправки $\delta^{(s)}$, $s = 1, 2, \ldots$, находят с использованием уже реально выполненной при вычислении $x^{(1)}$ факторизации матрицы A : $QP = A + F$, где F — соответствующее эквивалентное возмущение (см. (VII.13)). Таким образом, вычисление $\delta^{(s)}$ ($s \geqslant 1$) сводится к соответствующей обработке правых частей и вычислениям по формулам обратного хода. Поэтому процесс итерационного уточнения решения требует небольшого дополнительного времени по сравнению с первоначальным решением системы и может быть организован как решение выбранным прямым методом одной системы с разными последовательно вводимыми правыми частями. В (VII.18) невязки $r^{(s)}$ необходимо вычислять с повышенной точностью за счет либо накопления скалярного произведения, либо удвоения длины машинного слова на этом этапе вычислений по сравнению со всем остальным расчетом. В случае когда не слишком плохо «машинно обусловлена» матрица системы, последовательность $x^{(s)}$, $s = 1, 2, \ldots$, сходится к точному (в пределах длины машинного слова) решению заданной системы. Практически матрица A рассматривается как слишком плохо «машинно обусловленная», если наблюдается неубывание норм двух последовательных поправок $\delta^{(s)}$ или чрезмерно медленное их убывание: за итерацию уточняется менее двух двоичных или половина десятичного разряда. Заметим, что термин «слишком плохая машинная обусловленность» матрицы относится к данной конкретной длине машинного слова.

Сравнение решения $x^{(1)}$ и первой поправки $\delta^{(1)}$ позволяет (при заметном убывании $\delta^{(1)}$) определить число η верных десятичных цифр полученного решения $x^{(1)}$:

$$\eta = \left| -\lg \frac{\| \delta^{(1)} \|_\infty}{\| x^{(1)} \|_\infty} \right|.$$

На основе выполненной при вычислении $x^{(1)}$ факторизации матрицы системы $A \approx QP$ можно получить достаточно хорошую оценку числа обусловленности матрицы A. Для этого достаточно решить системы уравнений

$$(QP)^T y = f \text{ и } QPz = y$$

со специально подобранным вектором f и вычислить оценку по формуле

$$\operatorname{cond} A \approx \| A \|_1 \frac{\| z \|_1}{\| y \|_1}.$$

Этот способ достаточно экономичен (так как сводится только к вычислению двух решений системы уже с факторизованной матрицей) и достаточно надежен. Подробное описание его и обоснование различных методик выбора вектора дано в работе [128].

Используя приближенное значение cond A, можно вычислить число обусловленности системы (VII.2) по формуле (VII.7), а по формулам (VII.5), (VII.6) — относительную наследственную погрешность математического решения задачи. Если окажется, что относительная погрешность велика, то к такому математическому результату следует подходить с осторожностью. В этом случае желательно повысить точность задания исходных данных, а если это невозможно, задачу необходимо сформулировать по-новому относительно других параметров.

VII.2. Задачи на собственные значения матриц

1. Обусловленность в задачах на собственные значения. Как показано в гл. 4, при дискретизации задач на собственные значения дифференциальных операторов возникают системы линейных алгебраических уравнений

$$Kv = \lambda Mv,$$

где K — матрица жесткости, а M — матрица массы (см., например, (IV.27), (IV.28)). Данная обобщенная задача на собственные значения может быть сведена к обычной проблеме на собственные значения. При описании прикладных задач на собственные значения крайне редко возникают индивидуально задаваемые системы

$$\tilde{A}\tilde{v} = \tilde{\lambda}\tilde{v}.$$

Наиболее типичным является задание системы

$$Av = \lambda v \tag{VII.19}$$

и погрешности в исходных данных

$$\|\tilde{A} - A\| = \|\Delta A\| \leqslant \varepsilon. \tag{VII.20}$$

Не всегда близость элементов матриц A и $\tilde{A}$ обеспечивает близость собственных значений матрицы. Так, собственные числа матрицы

$$\begin{bmatrix} 1 & 10\,000 \\ 0 & 1 \end{bmatrix}$$

суть $\lambda_1 = 1$, $\lambda_2 = 1$, а возмущенной матрицы

$$\begin{bmatrix} 1 & 10\,000 \\ 0{,}0001 & 1 \end{bmatrix}$$

$\lambda_1 = 2$, $\lambda_2 = 0$.

При вычислении собственных векторов матриц также возникает задача оценки достоверности получаемых решений. Рассмотрим простой пример. Пусть заданная матрица имеет лишь простые собственные числа. Однако при некотором определенном изменении ее элементов в пределах точности их задания можно получить матрицу, имеющую кратные собственные числа. В этом случае каноническая форма матрицы при изменении ее элементов в пределах точности их задания может

перейти от чисто диагональной формы к недиагональной, канонической форме Жордана. При этом изменяется даже количество линейно независимых собственных векторов.

Из приведенных выше рассуждений видно, что прикладные задачи на собственные значения порождают (с учетом (VII.20)) множество уравнений (VII.19), обладающих достаточно широким классом формально допустимых решений. Решение проблемы собственных чисел заключается в определении одного из допустимых решений, получении этого решения и оценке наследственной погрешности в математическом решении задачи.

Для характеристики устойчивости (чувствительности) решений задачи (VII.19) к возмущениям исходных данных ΔA используют понятие обусловленности задачи. Однако в данном случае ситуация сложнее, чем в задаче решения системы линейных алгебраических уравнений (см. п. 2 параграфа VII.1). Здесь можно говорить об обусловленности матрицы A по отношению к полной проблеме всех собственных чисел, об устойчивости отдельного собственного числа, об обусловленности проблемы собственных векторов.

Рассмотрим вопрос об устойчивости проблемы собственных значений лишь для случая матрицы простой структуры, т. е. когда каждому собственному числу λ_i матрицы A соответствует столько собственных векторов X_i, какова кратность λ_i как корня характеристического уравнения $\det(A - \lambda E) = 0$.

Пусть λ_i, $i = 1, 2, \ldots, n$,— собственные числа матрицы A порядка n, среди которых могут быть и равные, а λ — собственное число возмущенной матрицы $A + \Delta A$. Тогда, как показано в монографии [105], справедлива оценка

$$\min_i |\lambda_i - \lambda| \leqslant \|H^{-1}\|_2 \|H\|_2 \|\Delta A\|_2, \qquad \text{(VII.21)}$$

где H — матрица, составленная из собственных векторов матрицы A; $\|\ \|_2$ — спектральная норма матрицы.

Из приведенного неравенства очевидно, что изменение собственных чисел может быть в $K(H) = \|H^{-1}\|_2 \|H\|_2$ раз большее, чем возмущение $\|\Delta A\|_2$. Величину $K(H)$ называют спектральным числом обусловленности матрицы A по отношению к полной проблеме собственных чисел.

Так как матрица H — не единственна (каждый столбец может быть умножен на произвольный числовой множитель), то считают, что спектральное число обусловленности матрицы A — это наименьшее значение $K(H)$ для всех допустимых H.

Если A — нормальная матрица: $\bar{A}^T A = A\bar{A}^T$ (в частности, если A — вещественная симметричная), то можно взять матрицу H унитарной (ортогональной) и тогда $K(H) = 1$, ибо спектральная норма такой матрицы равна единице. Отсюда следует, что проблема собственных чисел для нормальных (в частности, вещественных симметричных) матриц всегда хорошо обусловлена и

$$|\Delta\lambda_i| \equiv |\lambda - \lambda_i| \leqslant \|\Delta A\|.$$

Отметим, что это не обязательно справедливо для проблемы собственных векторов.

Для характеристики устойчивости отдельного собственного числа λ_i матрицы A простой структуры используется величина

$$\operatorname{cond}\lambda_i \equiv \frac{1}{|X_i^T Y_i|}, \quad i = 1, 2, \ldots, n,$$

где X_i, Y_i — соответствующие нормированные (спектральной нормой, т. е. евклидовой длиной) собственные векторы матриц A и A^T.

Пусть матрица A имеет только простые собственные числа, а возмущенная матрица $A + \Delta A$ близка к ней, т. е. ΔA — мало. Если собственные числа возмущенной матрицы обозначить через $\lambda_i + \Delta\lambda_i$, то с точностью до членов второго порядка малости справедлива оценка (см. [108])

$$|\Delta\lambda_i| \leqslant \operatorname{cond}\lambda_i \, \|\Delta A\|_2.$$

Если собственные векторы возмущенной матрицы $A + \Delta A$ обозначить через $X_i + \Delta X_i$, $i = 1, 2, \ldots, n$, то в данном случае соответственно имеем

$$\|\Delta X_i\|_2 \leqslant \|\Delta A\|_2 \sum_{j=1}^{n} \frac{\operatorname{cond}\lambda_j}{|\lambda_i - \lambda_j|}, \quad j \neq i. \tag{VII.22}$$

Отметим, что с учетом указанной нормировки собственных векторов это соотношение является оценкой относительной погрешности собственных векторов.

Приведенная оценка свидетельствует, что обусловленность проблемы собственных векторов существенно зависит от близости собственных чисел. Поэтому-то, даже в случае нормальной (вещественной симметричной) матрицы A с простыми собственными числами проблема собственных векторов далеко не всегда хорошо обусловлена.

2. Погрешность машинной реализации алгоритмов. Основными вопросами машинной реализации численных методов являются получение на ЭВМ решения, обеспечение оценки близости машинного и математического решений задачи. Машинная реализация методов нахождения собственных чисел и собственных векторов задачи (VII.19), (VII.20) вносит погрешность, определяемую свойствами матрицы A, методами решения проблемы собственных значений и особенностями вычислений на ЭВМ.

Вычисленные на ЭВМ собственные значения являются точными собственными значениями некоторой возмущенной матрицы $A + dA$, причем возмущение dA — не единственно и определяется выбранным алгоритмом, порядком матрицы A и длиной мантиссы машинного слова.

Теоретически при вычислении на ЭВМ всех собственных чисел можно использовать следующую апостериорную оценку близости машинного и математического решений задачи [105]:

$$\min_i |\mu_i - \lambda_i| \leqslant K(H) \frac{\|\theta\|_2}{\|V_i\|_2},$$

где μ_i — машинное приближение к λ_i (собственному числу матрицы A), V_i — машинное приближение к собственному вектору, соответствующему λ_i, $K(H)$ — спектральное число обусловленности, $\theta = AV_i - \mu_i V_i$ — невязка задачи на собственные значения.

Рассматривая суммарную погрешность машинной реализации в задачах на собственные значения как возмущение dA матрицы A, можно получить для соответствующих возмущений собственных чисел и собственных векторов оценки, аналогичные приведенным в предыдущем пункте. Так как здесь dA определяется порядком матрицы, методом решения задачи и длиной машинного слова, то, увеличивая длину машинного слова, можно повысить точность вычисленных собственных значений и получить достаточную близость машинного и математического решений задачи.

3. Характеристика некоторых методов и программ вычисления собственных значений. Для нахождения собственных чисел и собственных векторов матриц разработаны машинные методы (см., например [79, 86, 105, 106]) и создан ряд программных средств достаточно высокого научного уровня [81, 106, 149].

Пакеты программ *EISPACK* [149] и СПАН [81] позволяют проводить идентификацию задачи и обеспечивают автоматический выбор алгоритма решения задачи. Пакет СПАН позволяет получить оценки погрешности вычисленного решения. В этом пакете имеются программы решения «больших» задач на собственные значения, которые активно используют дисковую память ЭВМ ЕС. Имеются программы, реализующие алгоритмы переупорядочения для разреженных матриц.

На ЭВМ ЕС 1060 с помощью программных средств пакета СПАН были решены задачи нахождения всех собственных чисел и собственных векторов плотных симметричных матриц до 1000-го порядка, задачи нахождения нескольких наименьших собственных чисел и собственных векторов ленточных симметричных матриц до 10 000-го порядка и разреженных симметричных матриц до 20 000-го порядка.

Как уже упоминалось в начале данного параграфа, обобщенную задачу на собственные значения

$$Av = \lambda Bv, \quad \det B \neq 0,$$

часто преобразуют к обычной

$$Cx = \lambda x.$$

Имеется несколько способов осуществления такого преобразования, например формированные матрицы $C = B^{-1}A$ ($x \equiv v$). Однако матрица C может оказаться несимметричной, даже если A и B были вещественными симметричными, а это, как известно, ухудшает обусловленность задачи.

Если A и B — вещественные симметричные матрицы и B — положительно определенная, то вместо исходной обобщенной задачи можно решать обычную, где $C = B^{-1/2}AB^{-1/2}$, $x = B^{1/2}v$, причем такая матрица C будет симметричной. Если матрица B — симметричная и положительно определенная, то можно применить способ преобразования, уже рассмотренный в параграфе IV.2, а именно: построить

треугольное разложение $B = LL^T$, сформировать $C = L^{-1}AL^T$ и положить $x = L^Tv$. Матрица C и в этом случае будет симметричной, если симметрична матрица A. Последний способ реализован в [81, 149] для решения обобщенной задачи на собственные значения с плотными симметричными матрицами A, B, где B — положительно определенная матрица. Отметим, что ленточная и профильная структуры матрицы B наследуются матрицей L.

В работе [130] предложен алгоритм, позволяющий обобщенную задачу на собственные значения для ленточных A и B привести к обычной задаче без использования дополнительной памяти. (Возможны также некоторые способы преобразования.)

При дискретизации задач на колебания и устойчивость методом конечных элементов возникает обобщенная проблема собственных значений с профильно-разреженными структурами матриц A и B высокого порядка, в которых требуется вычислить несколько минимальных собственных чисел и принадлежащих им собственных векторов. В последнее время для решения таких задач получили применение метод итерирования подпространств и метод Ланцоша. Программы, реализующие эти алгоритмы, имеются в пакете программ СПАН. Программа метода итерирования подпространств имеется в работе [5], а [131] содержит набор программ решения задач линейной алгебры методом Ланцоша.

Одной из наиболее трудоемких операций как в методе итерирования подпространств, так и в методе Ланцоша является факторизация симметричных разреженных матриц. В ряде случаев для решения этой задачи можно применять процедуру переупорядочения матриц посредством перестановки строк и столбцов с тем, чтобы уменьшить вычислительные затраты при факторизации [91].

Когда треугольное разложение матриц затруднительно (например, в случае очень высоких порядков A и B), то для специального класса задач иногда могут быть полезными метод обратных итераций [86], метод последовательной верхней релаксации [155], градиентные методы [33, 88, 93].

Для ускорения сходимости итерационных методов градиентного типа используются приемы преобусловливания, или, что то же, регуляризации (см. например, [33, 47, 88, 89, 93]).

Представляет определенный интерес решение обобщенной проблемы собственных значений итерационными методами, реализуемыми на последовательности сеток [100, 115].

VII. 3. Нелинейные алгебраические и трансцендентные уравнения

Эффективное и надежное получение решений систем нелинейных алгебраических уравнений с приближенными исходными данными требует, как и в случае линейных систем (параграф VII.1), рассмотрения вопросов о корректной постановке задачи, об обусловленности системы, о способах оценки достоверности вычисленных решений и т. п. В данном случае эти вопросы еще более трудны и недостаточно изучены.

Для иллюстрации некоторых ситуаций, возникающих на практике, рассмотрим следующий простой пример.

Найти принадлежащие отрезку [0, 5] корни уравнения

$$\tilde{f}(x) \equiv x^2 - 2x + \tilde{c} = 0, \quad 0 \leqslant x \leqslant 5, \tag{VII.23}$$

если заданы приближенное значение свободного члена $c = 0{,}84$ и оценка погрешности

$$|c - \tilde{c}| \leqslant \delta. \tag{VII.24}$$

Легко проверить, что приближенное уравнение

$$f(x) \equiv x^2 - 2x + c = 0, \quad c = 0{,}84, \tag{VII.25}$$

имеет на отрезке [0, 5] два корня:

$$x_1 = 0{,}6, \quad x_2 = 1{,}4. \tag{VII.26}$$

Но как эти значения связаны с корнями точного уравнения (VII.23) при условии (VII.24)? Допустим, что $\delta = 0{,}2$. Тогда $\tilde{c}$ может принимать любое значение из отрезка [0,64; 1,04]. При этом в пределах точности исходных данных возможны следующие ситуации: уравнение (VII.23) — имеет на указанном отрезке два корня, если $\tilde{c} < 1$;—имеет один двукратный корень, если $\tilde{c} = 1$; — не имеет ни одного действительного корня, если $\tilde{c} > 1$. Очевидно, поставленную задачу (VII.24), (VII.25) при $\delta = 0{,}2$ нельзя считать корректно поставленной. Предположим теперь, что $\delta = 0{,}09$, т. е. значение $\tilde{c}$ находится в пределах отрезка [0,75; 0,93]. В этом случае при любом значении $\tilde{c}$ уравнение (VII.23) имеет два корня — x_1^*, x_2^*:

$$0{,}5 \leqslant x_1^* \leqslant 1 - \sqrt{0{,}07} \approx 0{,}74, \quad 1{,}26 \leqslant x_2^* \leqslant 1{,}5.$$

Они достаточно хорошо приближаются корнями (VII.26):

$$|x_1 - x_1^*| \leqslant 0{,}14, \quad |x_2 - x_2^*| \leqslant 0{,}14.$$

Таким образом, в данном случае можно говорить о корректной постановке задачи (VII.24), (VII.25) с приближенными исходными данными. Поскольку исследование упомянутых выше вопросов еще далеко от завершения, в данном параграфе будут затронуты лишь некоторые из них.

1. Погрешность машинной реализации вычислительных алгоритмов. Пусть задана система нелинейных уравнений

$$F(x) \equiv x - \Phi(x) = 0, \quad x \in D, \tag{VII.27}$$

где

$$F(x) = [f_1(x), f_2(x), \ldots, f_n(x)]^T, \quad x = [x_1, x_2, \ldots, x_n]^T,$$

D — область n-мерного евклидова пространства. Отображение Φ называется сжимающим на множестве $D_0 \subset D$, если существует такое

значение $0 < \alpha < 1$, что $\| \Phi(x) - \Phi(y) \| < \alpha \| x - y \|$ при любых $x, y \in D_0$.

Как известно (см., например, [6, 85]), если отображение $\Phi(x)$ — сжимающее на замкнутом множестве $D_0 \subset D$ и значение $\Phi(x) \in D_0$ при $x \in D_0$, то система (VII.27) имеет в D_0 единственное решение x^*.

Пусть для вычисления этого решения используется метод простой итерации, реализуемый по формуле

$$x^k = \Phi(x^{k-1}), \quad k = 1, 2, \ldots, \tag{VII.28}$$

где x^0 — некоторое начальное приближение. Доказано (см., например, [6, 51]), что при выполнении указанных выше условий (обеспечивающих существование единственного x^*) последовательность $x^1, x^2, \ldots$ сходится при любом $x^0 \in D_0$ к решению x^*, а скорость сходимости характеризуется неравенствами

$$\| x^k - x^* \| \leqslant \alpha^k \| x^0 - x^* \|, \tag{VII.29}$$

или

$$\| x^k - x^* \| \leqslant \frac{\alpha^k}{1-\alpha} \| x^0 - \Phi(x^0) \|, \quad k = 1, 2, \ldots$$

Однако при построении приближений (VII.28) на ЭВМ неизбежно возникают погрешности и за счет применения вычислений по приближенным формулам, например для значений элементарных функций, и за счет ошибок округления. При этом вместо (VII.28) выполняются равенства

$$x^k = \Phi(x^{k-1}) + \Delta_k, \quad k = 1, 2, \ldots,$$

где Δ_k — погрешность на k-й итерации, для которой известна оценка

$$\| \Delta_k \| \leqslant \delta, \quad k = 1, 2, \ldots$$

В результате, как показано в работе [51], вместо (VII.29) имеем

$$\| x^k - x^* \| \leqslant \alpha^k \| x^0 - x^* \| + \frac{\delta}{1-\alpha},$$

откуда следует, что при $k \to \infty$ приближенные решения x^k уже могут не сходиться к точному решению x^* системы (VII.27).

Проиллюстрируем сказанное решением уравнения

$$f(x) \equiv 0{,}001 - 100e^{-(10+x/5)} = 0, \quad 0 \leqslant x \leqslant 10,$$

корень которого

$$x = 5(5 \ln 10 - 10) \approx 7{,}565.$$

Приближения к искомому корню строим по формуле

$$x_0 = 1,$$

$$x_{k+1} = x_k - 0{,}001 + 100e^{-(10+x_k/5)}, \quad k = 0, 1, \ldots$$

В данном примере функция

$$\varphi(x) \equiv x - 0{,}001 + 100e^{-(10+x/5)}$$

является сжимающей, так как

$$|\varphi'(x)| = |1 - 20e^{-(10+x/5)}| = \alpha < 1, \quad 0 \leqslant x \leqslant 10,$$

и для нее справедливо соотношение

$$|\varphi(x) - \varphi(y)| \leqslant \alpha |x - y|, \quad 0 \leqslant x, y \leqslant 10.$$

Задача решалась на ЭВМ МИР-2 при разрядности 4. После выполнения условий окончания итераций $\left(\text{а именно } \frac{|x_{k+1} - x_k|}{|x_k|} \leqslant \varepsilon\right)$ при значении $\varepsilon = 10^{-2}$ было получено машинное решение $\tilde{x} = 5{,}005$, при $\varepsilon = 10^{-3}$ тоже $\tilde{x} = 5{,}005$. Увеличение длины машинного слова до восьми десятичных разрядов по той же программе на той же ЭВМ позволило получить $x_{\bar{k}} = 7{,}518\,856\,5$ за $\bar{k} = 219\,54$ итераций (при $\varepsilon = 10^{-2}$) и $x_{\bar{k}} = 7{,}559\,634\,1$ за $\bar{k} = 33237$ итераций (при $\varepsilon = 10^{-3}$).

Для решения этой задачи использован метод Ньютона, реализуемый по формуле $x_0 = 1$,

$$x_{k+1} = x_k - 0{,}00005e^{(10+x_k/5)} + 5, \quad k = 0, 1, 2, \ldots,$$

со своим условием окончания итераций (см. [79]). При длине машинного слова в четыре десятичных знака для $\varepsilon = 10^{-2}$ за четыре итерации было получено $x_4 = 7{,}518$, а для $\varepsilon = 10^{-3}$ за пять итераций $x_5 = 7{,}579$. Использование в расчетах восьми десятичных цифр при реализации той же программы метода Ньютона для $\varepsilon = 10^{-2}$ за четыре итерации позволило получить $x_4 = 7{,}564\,40\,56$ и для $\varepsilon = 10^{-3}$ за пять итераций $x_5 = 7{,}564\,628\,4$.

Из этих примеров видно, что для одних методов длина машинного слова в четыре десятичных разряда оказалась недостаточной, а для других ее вполне хватило, чтобы получить хорошее приближение к искомому решению.

Подводя итоги по вопросам машинной реализации алгоритмов решения нелинейных уравнений, важно сделать следующие выводы: решение задачи может быть с достоверностью получено, если выбранный метод решения, длина машинного слова и критерии окончания счета будут соответствовать свойствам задачи.

2. Характеристика некоторых методов и программ решения систем нелинейных уравнений. Численное решение систем нелинейных алгебраических или трансцендентных уравнений представляет собой сложную и до конца не исследованную проблему вычислительной математики. Для решения систем нелинейных уравнений можно использовать метод простой итерации, метод секущих, метод Ньютона, квазиньютоновские методы и другие, описание которых можно найти, например, в работах [6, 79, 85]. Простая итерация обладает линейной скоростью сходимости, метод Ньютона — квадратичной (при выполнении соответствующих условий), а квазиньютоновские методы — сверхлинейной скоростью сходимости. Несмотря на то что квазиньютоновские методы обладают более медленной по сравнению с методом Ньютона

теоретической скоростью сходимости, они требуют при своей реализации меньшего числа машинных операций по сравнению с методом Ньютона. Однако все эти методы обладают локальной сходимостью, т. е. сходимостью лишь при хорошем начальном приближении.

Отметим, что решение системы нелинейных уравнений может быть сведено к задаче минимизации функций.

Для получения хорошего начального приближения при решении систем нелинейных уравнений используют те или иные соображения о начальном приближении, учитывающие физику процесса, а если таковых не имеется, то применяют те или иные методы спуска и комбинируют их с методами, обладающими более высокой скоростью сходимости.

В настоящее время разрабатываются методы решения систем нелинейных уравнений, обладающих глобальной сходимостью (см., например, [8]). К этим же методам принадлежит метод продолжения по параметру (см., например, [120, 143]).

Среди часто используемых программ решения систем нелинейных уравнений с матрицами Якоби произвольного вида можно отметить программы, реализующие методы Брента и Брауна [125, 151].

В настоящее время разработаны и некоторые алгоритмы квазиньютоновского типа, специально ориентированные на решение систем нелинейных алгебраических уравнений, возникающих в рамках использования метода конечных элементов. Среди алгоритмов этой группы надо прежде всего отметить алгоритм *BFGS* [148], предназначенный для решения нелинейных систем как с симметричной положительно определенной матрицей Якоби, так и с несимметричной. Данный алгоритм сохраняет структуру матрицы Якоби и ее свойства.

VII.4. Задачи Коши для систем обыкновенных дифференциальных уравнений

1. Постановка задач, некоторые определения. Правильно поставленная прикладная задача всегда имеет решение. При рассмотрении прикладных задач вводят те или иные упрощающие гипотезы и исследуемую задачу сводят к некоторым физическим моделям. Решение физической модели условно назовем физическим решением задачи. При описании физических моделей с помощью математического аппарата возникают математические модели, в частности задача Коши для обыкновенных дифференциальных уравнений.

Задачи Коши для систем обыкновенных дифференциальных уравнений получают и в результате применения метода Бубнова — Галеркина к задачам с начально-краевыми условиями. Так, если решаются уравнения параболического типа, то получают следующую задачу:

$$M\frac{du}{dt} = -Ku + F, \quad Mu(0) = \Phi.$$

Как следует из рассмотрений гл. III, компоненты вектора правых частей F имеют вид $F_k = \int f(x, t)\,\varphi_k(x)\,dx$. Применяя те или иные формулы численного интегрирования, мы вносим некоторую погрешность.

Кроме того, вносим погрешность и при вычислении вектора начальных условий. Возникает вопрос о влиянии этих погрешностей на точность решения исходной задачи. Для выяснения этого вопроса рассмотрим задачу Коши для одного уравнения:

$$\frac{du}{dx} = f(u, x), \quad x \in [x_0, X], \tag{VII.30}$$

$$u(x_0) = u_0. \tag{VII.31}$$

Пусть в замкнутом прямоугольнике D плоскости (x, u), определяемом неравенствами $|x - x_0| \leqslant a$, $|u - u_0| \leqslant b$, где a, b — некоторые положительные постоянные, функция $f(x, u)$ непрерывна и удовлетворяет условию Липшица по u:

$$|f(x, u_1) - f(x, u_2)| \leqslant L|u_1 - u_2|, \quad L - \text{const}.$$

Как известно, при этих предположениях гарантируются существование и единственность решения задачи (VII.30), (VII.31) в некотором промежутке $x_0 - X_1 \leqslant x \leqslant x_0 + X_1$. В дальнейшем всегда предполагается, что точное решение задачи Коши существует на всем заданном промежутке $[x_0, X]$.

Пусть наряду с задачей (VII.30), (VII.31) имеется возмущенная задача

$$\frac{dy}{dx} = f(x, y) + \delta(x), \quad x \in [x_0, X], \tag{VII.30'}$$

$$y(x_0) = u_0 + \delta_0, \tag{VII.31'}$$

где $\delta(x)$, δ_0 предполагаются достаточно малыми.

Однако даже малые погрешности в задании исходных данных могут существенно исказить решение. Действительно, точное решение задачи Коши

$$\frac{du}{dx} = u - x, \quad 0 \leqslant x \leqslant 100, \quad u(0) = 1$$

имеет вид $u(x) = 1 + x$, следовательно, $u(100) = 101$. Задача Коши с возмущенным начальным условием

$$\frac{dy}{dx} = y - x, \quad 0 \leqslant x \leqslant 100, \quad y(0) = 1 + 10^{-6}$$

имеет точное решение $y(x) = 1 + x + 10^{-6}e^x$ и $y(100) \approx 2{,}7 \cdot 10^{37}$.

Таким образом, небольшое изменение исходных данных сильно изменило решение. В связи с этим весьма существенным является вопрос об отыскании условий, при которых достаточно малые изменения начальных данных вызывают малые изменения в решении задачи Коши. Если x изменяется на конечном отрезке $[x_0, X]$, то ответ на поставленный вопрос дает *теорема о непрерывной зависимости решений от начальных значений* (см., например, [119]). Если же x может принимать сколь угодно большие значения, то эти вопросы относятся к теории устойчивости [21, 58].

Пусть выполняются сформулированные выше условия для функции $f(x, u)$. Тогда для решения $y(x)$ возмущенной задачи (VII.30'), (VII.31')

можно получить следующую оценку:

$$\max_{x_0 \leqslant x \leqslant X} |u(x) - y(x)| \leqslant \frac{\varepsilon}{L}((L+1)e^{LX} - 1),$$

где $\varepsilon = \max(|\delta_0|, \max_{x_0 \leqslant x \leqslant X} |\delta(x)|)$.

Аналогичные оценки могут быть получены и для систем обыкновенных дифференциальных уравнений.

Рассмотрим задачу Коши для линейной системы n обыкновенных дифференциальных уравнений

$$\frac{dU}{dx} = AU, \quad U(0) = U_0, \quad x \in [0, X], \tag{VII.32}$$

где $U(x) = [u_1(x), u_2(x), \ldots, u_n(x)]^T$, A — вещественная матрица с простыми собственными числами. Тогда существует такая невырожденная матрица C, что

$$C^{-1}AC = \Lambda,$$

где

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix},$$

λ_i — i-е собственное число матрицы A. Если ввести замену переменных

$$U(x) = CZ(x), \quad Z(x) = [z_1(x), z_2(x), \ldots, z_n(x)]^T,$$

то система (VII.32) преобразуется в систему

$$\frac{dz_i}{dx} = \lambda_i z_i(x), \quad i = 1, 2, \ldots, n. \tag{VII.33}$$

Таким образом, исследовать поведение решения системы (VII.32) можно с помощью системы (VII.33). Более того, устойчивость решения системы дифференциальных уравнений (VII.33) можно изучать на уравнении

$$\frac{du}{dx} = \lambda u, \tag{VII.34}$$

которое называют тестовым [147]. Будем предполагать, что $\lambda = \alpha + i\beta$ принадлежит полю комплексных чисел. Решение уравнения (VII.34) называется асимптотически устойчивым, если $\alpha = \operatorname{Re}\lambda < 0$, устойчивым, если $\alpha = 0$, и неустойчивым, если $\alpha = \operatorname{Re}\lambda > 0$.

Несколько слов о постановке и исследовании задачи Коши для системы n обыкновенных дифференциальных уравнений

$$\frac{dU}{dx} = F(x, U), \quad U(x_0) = U_0, \quad x \in [x_0, X], \tag{VII.35}$$

где

$$U(x) = [u_1(x), u_2(x), \ldots, u_n(x)]^T,$$

$$F(x, U) = [f_1(x, U), f_2(x, U), \ldots, f_n(x, U)]^T.$$

Пусть в замкнутом параллелепипеде D

$$|x - x_0| \leqslant a, \quad |u_i - u_{i0}| \leqslant b, \quad i = 1, 2, \ldots, n,$$

функции $f_i(x, U)$ непрерывны и удовлетворяют условию Липшица по переменным u_k:

$$|f_i(x, \tilde{u}_1, \ldots, \tilde{u}_n) - f_i(x, u_1, \ldots, u_n)| \leqslant L \sum_{k=1}^{n} |\tilde{u}_k - u_k|,$$

$$i = 1, 2, \ldots, n,$$

L — константа Липшица.

Выполнение этих условий гарантирует существование и единственность решения задачи (VII.35).

Заметим, что условие Липшица часто записывают в виде

$$\|F(x, \tilde{U}) - F(x, U)\| \leqslant L \|\tilde{U} - U\|,$$

где $\|\cdot\|$ — некоторая норма в конечномерном пространстве.

Если функция $F(x, U)$ — непрерывно дифференцируема по U, то качественное исследование системы дифференциальных уравнений (VII.35) вблизи некоторого частного решения $U^*(x)$ можно провести следующим образом. Разложим $F(x, U)$ в окрестности $U^*(x)$ в ряд Тейлора:

$$F(x, U) = F(x, U^*) + J(x, U^*)(U - U^*) + \ldots,$$

где $J(x, U^*)$ — матрица Якоби, $J(x, U^*) = \left\{\frac{\partial f_i}{\partial u_j}\right\}\Big|_{U=U^*}$, $i, j = 1, 2, \ldots, n$. Тогда согласно (VII.35) имеем

$$\frac{dU}{dx} \approx \frac{dU^*}{dx} + J(x, U^*)(U - U^*). \qquad \text{(VII.36)}$$

Если изменение элементов матрицы $J(x, U^*)$ на некотором интервале изменения x достаточно мало, то $J(x, U^*)$ можно заменить локально постоянной матрицей A и свести исследование устойчивости решения системы (VII.35) к исследованию устойчивости решения линейной системы обыкновенных дифференциальных уравнений с постоянными коэффициентами.

Рассмотрим теперь задачу Коши для обыкновенного дифференциального уравнения

$$\frac{du}{dx} = q(u - p(x)) + \frac{dp}{dx}, \quad u(0) = u_0, \qquad \text{(VII.37)}$$

где $q = \text{const}$. Решение этой задачи, как нетрудно убедиться,

$$u(x) = (u_0 - p(0))\, e^{qx} + p(x).$$

Если q — большое положительное число, то решение задачи (VII.37) неустойчиво; если q — малое положительное число, то решение задачи устойчиво на некотором конечном интервале. В случае, когда q большое по модулю отрицательное число, каким бы ни было выбрано значение u_0, через достаточно малый промежуток $[0, x_1]$, называемый

переходным участком (или пограничным слоем), кривая решения $u(x)$ становится как угодно близкой к кривой частного решения $u^*(x) = p(x)$ уравнения (VII.37). Эта сверхустойчивость решения дифференциальной задачи является идеальной в смысле распространения наследственной ошибки в дифференциальном уравнении, но она создала ряд трудностей численного решения задач на ЭВМ. Одна из них состоит в том, что хотя решение за пределами переходного участка ведет себя как $p(x)$ и практически не зависит от q (при $q < 0$), тем не менее из условия устойчивости шаг численного интегрирования приходится выбирать зависящим от q (см. работу [90]). Чем большее значение $|q|$, тем меньший шаг интегрирования (жесткие ограничения на шаг интегрирования). Такие задачи получили название жестких. Аналогично может быть рассмотрена задача Коши для систем обыкновенных дифференциальных уравнений.

Для выяснения вопроса, является ли задача Коши (VII.35) жесткой, необходимо исследовать поведение решений системы уравнений (VII.35) в окрестности некоторого частного решения $U^*(x)$ этой системы. Будем предполагать, что система локально устойчива, т. е. все локальные собственные числа $\lambda_i(x)$ матрицы Якоби $J(x) = J(x, U^*(x))$ различны и $\operatorname{Re}\lambda_i < 0$, $i = 1, 2, \ldots, n$. При этих предположениях в работе [147] находим следующее определение жесткой задачи Коши.

Задача Коши (VII.35) называется жесткой на некотором интервале $J \subset [x_0, X]$, если для всех $x \in J$ выполняются условия

$$\operatorname{Re}\lambda_i < 0, \quad i = 1, 2, \ldots, n, \quad s(x) = \max_i \operatorname{Re}(-\lambda_i)/\min_i \operatorname{Re}(-\lambda_i) \gg 1,$$

где λ_i — собственные числа матрицы Якоби $J(x) = J(x, U^*)$. Величину $s(x)$ называют локальным коэффициентом жесткости задачи. Если $s(x)$ есть величина $O(10)$, то задачу можно считать жесткой. В ряде прикладных задач коэффициент жесткости достигает величины $O(10^6)$.

Отметим, что с нашей точки зрения это определение полезно дополнить еще условием: большие по модулю собственные числа имеют большую по модулю отрицательную вещественную часть.

Приведем еще одно определение жесткой системы [90]. Система обыкновенных дифференциальных уравнений (VII.35) называется жесткой на отрезке $[c, d]$, принадлежащем интервалу существования ее решений, если при любом векторе начальных значений $U_0 = U(x_0)$ и на любом отрезке $[x_0, x_0 + \xi] \subset [c, d]$ найдутся такие числа τ, L, N, удовлетворяющие соотношениям

$$0 < \tau \ll d - c, \quad 0 < L \leqslant \rho(J(x, U)) \leqslant \|J(x, U)\|, \quad (x, U) \in D, \quad N \gg 1,$$

что справедливы неравенства

$$\left|\frac{du_k}{dx}\right|_{x \in \Delta_1} \leqslant \frac{L}{N} \max_{x \in \Delta_2} |u_k(x)|, \quad k = 1, 2, \ldots, n, \qquad \text{(VII.38)}$$

где $\Delta_1 = [x_0 + \tau, x_0 + \xi]$, $\Delta_2 = [x_0, x_0 + \xi]$.

В данном определении использовались следующие обозначения: $\rho(J(x, U))$ — максимальный модуль собственных чисел матрицы Якоби, $\|\cdot\|$ — принятая норма матрицы. Отметим, что для жестких

дифференциальных уравнений «почти всегда» существуют участки решения (переходные участки и стационарные участки) с существенно различным характером его поведения, причем продолжительность переходных участков τ значительно меньше стационарных.

Решение одного жесткого дифференциального уравнения быстро стремится к такому решению, которое не зависит от начальных условий. Однако при малых отклонениях решений производные их резко отличаются. Отметим, что жесткость зависит от самого дифференциального уравнения, а не от поведения решения. Следует учитывать, что дифференциальное уравнение может быть жестким на некоторых участках интервала интегрирования и нежестким на других.

Определенный условиями (VII.38) класс жестких систем обыкновенных дифференциальных уравнений может быть расширен за счет уравнений, у которых на переходных участках и вне их сильно различаются по величине производные не первого, а более высокого порядка. В этом случае к числу жестких относят такие системы, для которых вне пограничного слоя вместо неравенства (VII.38) выполняется условие

$$\left|\frac{d^l u_k}{dx^l}\right|_{x\in\Delta_1} \leqslant \left(\frac{L}{N}\right)^l \max_{x\in\Delta_2} |u_k(x)|, \quad l>1.$$

2. Погрешность и устойчивость машинных алгоритмов численного интегрирования. Численные методы являются основным средством решения возникающих на практике задач с начальными условиями. При построении численных методов решения задачи Коши исходную дифференциальную задачу заменяют (аппроксимируют) дискретной задачей, в которую входит параметр дискретизации; разрабатывают вычислительную схему решения дискретной задачи, удобную для реализации на ЭВМ; рассматривают вопрос численной устойчивости и доказывают сходимость решения дискретной задачи к решению исходной дифференциальной задачи.

Рассмотрим основные проблемы, возникающие при численном решении задачи Коши для одного уравнения:

$$\frac{du}{dx} = f(x, u), \quad 0 \leqslant x \leqslant X, \tag{VII.39}$$

$$u(0) = u_0. \tag{VII.40}$$

Предполагаем, что существует достаточно гладкое решение этой задачи, а функция $f(x, u)$ удовлетворяет по u условию Липшица

$$|f(x, u_1) - f(x, u_2)| \leqslant L|u_1 - u_2|.$$

Для численного решения задачи (VII.39), (VII.40) область непрерывного изменения аргумента $[0, X]$ заменяют сеткой

$$\omega_h = \left\{x_i = ih, i = 0, 1, \ldots, N, h = \frac{X}{N}\right\}.$$

Здесь N — число узлов сетки, h — шаг сетки.

Интегрируя уравнение (VII.39) на отрезке $[x_i, x_{i+1}]$, получаем

$$u(x_{i+1}) - u(x_i) = \int_{x_i}^{x_{i+1}} f(x, u(x))\,dx. \qquad \text{(VII.41)}$$

Заменяя интеграл в правой части (VII.41) различными формулами численного интегрирования, получают различные методы численного решения задач Коши, обладающие теми или иными качествами. Замена интеграла по формуле прямоугольников приводит нас к явной

$$y_{i+1} = y_i + hf(x_i, y_i) \qquad \text{(VII.42)}$$

или неявной

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1})$$

формуле метода Эйлера.

Численное решение задачи (VII.39), (VII.40) явным методом Эйлера реализуется следующим образом. Зная начальное условие

$$y_0 = u_0, \qquad \text{(VII.43)}$$

по формуле (VII.42) можно вычислить y_1, y_2 и т. д. до y_N.

Для численного решения задачи неявным методом Эйлера, зная начальное условие (VII.43), на каждом шаге интегрирования можно использовать итерационную формулу

$$y_{i+1}^{(k+1)} = y_i + hf(x_{i+1}, y_{i+1}^{(k)}), \quad k = 0, 1, 2, \ldots,$$

или итерационный процесс на основе метода Ньютона

$$y_{i+1}^{(k+1)} = y_{i+1}^{(k)} + \left(1 - h\frac{\partial f(x_{i+1}, y_{i+1}^{(k)})}{\partial y_{i+1}^{(k)}}\right)^{-1} (hf(x_{i+1}, y_{i+1}^{(k)}) + y_i - y_{i+1}^{(k)}), \quad k = 0, 1, 2, \ldots,$$

который при достаточно хорошем начальном приближении $y_{i+1}^{(0)}$ сходится к соответствующему решению нелинейного уравнения. Начальное приближение выбирают, используя явную формулу метода Эйлера. Численные методы решения обыкновенных дифференциальных уравнений, которые для вычисления значения y_{i+1} используют значение y_i, называют одношаговыми. Явный и неявный метод Эйлера принадлежит к классу одношаговых численных методов.

В вычислительной машине реализуется не схема (VII.42), (VII.43), а некоторая возмущенная схема

$$y_{i+1} = y_i + hf(x_i, y_i) + \delta_i, \quad i = 1, 2, \ldots, N, \quad y_0 = u_0 + \delta_0,$$

где $|\delta_i| < \delta$, $i = 0, 1, \ldots, N$.

Для вычисленного значения y_N справедлива оценка (см., например, [35, 80])

$$|y_N - u(x_N)| < e^{LX}\left(O(h) + \delta + \frac{\delta}{hL}\right), \qquad \text{(VII.44)}$$

где $O(h)$ — погрешность дискретизации.

Из оценки (VII.44) следует, что при $h \to 0$ ошибка дискретизации стремится к нулю, а погрешность реализации схемы на ЭВМ — к бесконечности. Таким образом, при решении задач Коши на ЭВМ полученное машинное решение всегда будет отличаться от математического решения. Из (VII.44) следует также, что выбор шага интегрирования является принципиальным моментом. Шаг интегрирования определяется соображениями не только близости решения разностного уравнения к решению дифференциального, но и связанными с влиянием возмущений, возникающих в процессе вычислений.

При численном решении задач с начальными условиями применяют вычислительные схемы, которые имеют вид конечно-разностных уравнений, определенных на сетке ω_h. Чаще всего это линейное разностное уравнение с постоянными коэффициентами вида

$$\sum_{s=0}^{m} a_s y_{i+s} = h \sum_{s=0}^{m} b_s f_{i+s}, \quad i = 0, 1, 2, \ldots, \quad a_m \neq 0. \quad \text{(VII.45)}$$

Этому уравнению ставят в соответствие характеристическое уравнение

$$\sum_{s=0}^{m} a_s r^s = 0. \quad \text{(VII.46)}$$

Решение разностного уравнения (VII.45) называют численно устойчивым, если все корни характеристического уравнения (VII.46) по модулю меньше или равны единице и корни, по модулю равные единице, простые. Численная устойчивость разностной схемы решения задачи Коши не означает уменьшения начального возмущения от шага к шагу интегрирования, а только ограничивает рост этого возмущения, если он происходит. Решение разностного уравнения (VII.45) называют асимптотически численно устойчивым, если все корни его характеристического уравнения (VII.46) по модулю меньше единицы. И наконец, решение численно неустойчиво, если по крайней мере модуль одного из корней больше единицы.

Рассмотрим явную схему метода Эйлера (VII.42), (VII.43), примененную к тестовому уравнению (VII.34)

$$y_{i+1} = y_i + \lambda h y_i, \quad i = 0, 1, 2, \ldots$$

Требование асимптотической численной устойчивости в этом случае приводит к условию

$$|1 + h\lambda| < 1,$$

или

$$(1 + h\alpha)^2 + h^2\beta^2 < 1, \quad \text{(VII.47)}$$

где $\alpha = \operatorname{Re} \lambda$, $\beta = \operatorname{Im} \lambda$. Областью асимптотической численной устойчивости в этом случае будет внутренность круга единичного радиуса с центром $h\alpha = -1$, $h\beta = 0$ на плоскости $(h\alpha, h\beta)$. Для чисто мнимого значения λ ($\alpha = 0$) неравенство (VII.47) не выполняется ни при каком значении $h > 0$. Отметим, что явный метод Эйлера численно устойчив. Действительно, если для него составить характеристическое уравнение

$$r - 1 = 0,$$

то оно имеет корень $r = 1$, что показывает численную устойчивость явной схемы.

Рассмотрим теперь численную устойчивость неявного метода Эйлера, который, будучи применен к тестовому уравнению (VII.34), записывается в виде

$$y_{i+1} = y_i + h\lambda y_{i+1}.$$

Нетрудно убедиться, что и этот метод численно устойчив, так как $r = 1$. Условие асимптотической численной устойчивости имеет вид

$$|(1 - \lambda h)^{-1}| < 1,$$

или

$$(1 - h\alpha)^2 + h^2\beta^2 > 1.$$

Областью асимптотической численной устойчивости этого метода будет вся плоскость $(h\alpha, h\beta)$, за исключением единичного круга с центром (1, 0).

Численные методы решения обыкновенных дифференциальных уравнений, у которых область асимптотической устойчивости при решении тестового уравнения (VII.34) содержит всю левую полуплоскость плоскости $(h\alpha, h\beta)$, называют A-устойчивыми [132]. Иными словами, A-устойчивыми называют численные методы, если их область асимптотической устойчивости включает всю полуплоскость $\mathrm{Re}\,(h\lambda) < 0$. Таким образом, неявный метод Эйлера является A-устойчивым.

Отметим, что в жестких дифференциальных уравнениях влияние на решение погрешности в задании исходных данных убывает со временем. Применение для решения таких задач численно устойчивых методов, в которых может ограниченно расти погрешность машинной реализации, нецелесообразно. Уменьшение погрешности от шага к шагу обеспечивается использованием A-устойчивых численных методов.

Рассматриваемые вопросы возникают и для других схем, обладающих более высоким порядком точности численного интегрирования обыкновенных дифференциальных уравнений, а также для систем таких уравнений (см., например, [80, 126]).

Проиллюстрируем важность рассматриваемых нами теоретических вопросов на примере решения следующей модельной задачи Коши для системы обыкновенных дифференциальных уравнений:

$$\frac{du_1}{dx} = -0{,}013u_2 - 1000u_1u_2 - 2500u_1u_3,\ u_1(0) = 0,$$

$$\frac{du_2}{dx} = -0{,}013u_2 - 1000u_1u_2,\ u_2(0) = 1,$$

$$\frac{du_3}{dx} = -2500u_1u_3,\ u_3(0) = 1,$$

$$0 \leqslant x \leqslant 50.$$

Эта задача решалась на ЭВМ МИР-2 с длиной машинного слова 12 десятичных знаков по программе метода Кутта — Мерсона с автоматическим выбором шага интегрирования. Машинное решение $y = [-0{,}389\,244\,498 \cdot 10^{-4},\ 0{,}597\,572\,459,\ 1{,}402\,341\,728]^T$ в точке $x =$

$= 50$ было получено за 102 ч работы машины, его относительная погрешность 2000 %. Шаг интегрирования всегда выбирался автоматически не больше чем 10^{-3}. При исследовании выяснилось, что эта система принадлежит к классу жестких. Применение метода решения жестких систем на этой же машине с прежней длиной машинного слова позволило получить всего за 20 мин значение $y = [-1{,}842\,945\,749 \times \times 10^{-6},\ 0{,}547\,654\,343,\ 1{,}402\,343\,765]^T$ в точке $x = 50$. Относительная погрешность этого машинного решения порядка тысячной доли процента.

Решать все системы обыкновенных дифференциальных уравнений по программам решения жестких систем слишком дорого. Ведь на каждом шаге, как мы видим, приходится решать нелинейные уравнения. Поэтому, с одной стороны, целесообразно различать, с жесткой или нежесткой системой нам приходится иметь дело, а с другой — необходимо учитывать, что в настоящее время разработаны и разрабатываются методы решения жестких систем, снимающие требования A-устойчивости (см., например, [137]) и, следовательно, уменьшающие затраты на получени решения.

3. Характеристика некоторых методов и программ решения. Разработка методов решения задач Коши для систем обыкновенных дифференциальных уравнений всегда была почти такой же популярной, как и разработка методов решения систем линейных алгебраических уравнений. Именно поэтому к настоящему времени имеется большое количество численных методов интегрирования задач Коши [4, 6, 99]. Все численные методы можно подразделить на одношаговые и многошаговые, на явные и неявные. Собственно выбор метода интегрирования обусловливается в основном характером решаемой задачи, поведением ее решения (основную роль в этом играют свойства правых частей системы обыкновенных дифференциальных уравнений), а также желаемой точностью получения решения и опытом специалиста, решающего задачу. До начала 70-х годов предпочтение при выборе численных методов решения рассматриваемых задач отдавалось явным одношаговым методам решения. Однако проведенные впоследствии различными авторами исследования показали, что в случае когда система дифференциальных уравнений относится к классу жестких, ее целесообразно решать неявными методами. В связи с этим началось интенсивное исследование «старых» неявных методов и создание новых методов, позволяющих уменьшить время получения решения на одном шаге интегрирования [1, 90, 126, 137, 154].

При решении задач Коши для систем обыкновенных дифференциальных уравнений возникает еще одна нетривиальная проблема — выбор такого шага интегрирования, который обеспечивал бы желаемую точность решения. Эта проблема может быть решена одним из подходов [154]: 1) использованием экстраполяции по проведению одного шага длиной $2h$ и двух шагов длиной h (экстраполяция Ричардсона); 2) использованием спаренного метода (последний применяется и для оценивания глобальной ошибки аппроксимации одношаговых методов на основе теоремы Штеттера [20, 99, 116]).

К настоящему времени создан ряд пакетов программ интегрирования задач Коши для систем обыкновенных дифференциальных уравнений, как обычных, так и жестких. В основе почти всех существующих пакетов лежит разработанная Гиром программа *DIFSUB* [137]. Обычные системы уравнений здесь интегрируются методом Адамса переменного порядка, жесткие — методом Гира, порядок которого может меняться от первого до шестого. Возможность объединения указанных выше двух методов в одной программе обусловлена тем, что в обоих методах используется один и тот же предиктор.

В качестве контроля для выбора шага и порядка метода используется требование, чтобы ошибка метода на каждом шаге интегрирования не превышала ε. При этом если решение является возрастающей функцией, то контролируется относительная ошибка метода, а при выходе решения на стационарное значение — абсолютная ошибка. Ошибка оценивается через поправку корректора.

В исходной информации для программы *DIFSUB*, кроме ε, подпрограммы вычисления правых частей системы, границ изменения шага интегрирования и некоторых других параметров, задается указатель метода *MF* с возможными значениями 0, 1, 2. Если $MF = 0$, то при интегрировании используется метод Адамса; если $MF = 1$, то используется метод Гира с аналитически заданной матрицей Якоби (при этом необходимо написать подпрограмму ее вычисления); если $MF = 2$, то используется метод Гира, но при этом матрица Якоби вычисляется автоматически с помощью численного дифференцирования правых частей.

Как уже упоминалось, на основе идей, заложенных в программу *DIFSUB*, и самой программы созданы многочисленные пакеты, например *GEAR, GEARBI, GEARS, GEARIB* и др. Отметим, что пакет *GEARBI* ориентирован на решение жестких систем дифференциальных уравнений с матрицей Якоби, имеющей ленточную структуру или с матрицей Якоби, которая хорошо приближается матрицей ленточной структуры. Пакет *GEARS* интегрирует системы с разреженной матрицей Якоби.

Как дальнейшее развитие пакетной проблематики создан *ODEPACK* — систематизированный набор программ для решения обыкновенных дифференциальных уравнений [144]. В этот набор включено пять программ решения задач Коши для систем обыкновенных дифференциальных уравнений. Набор разработан в Национальной лаборатории им. Лоуренса в Ливерморе (штат Калифорния, США). Приведем состав набора с указанием названия программ и их назначения:

LSODE — предназначена для интегрирования жестких и нежестких систем обыкновенных дифференциальных уравнений, объединяет возможности пакетов *GEAR* и *GEARB*;

LSODI — предназначена для интегрирования задач Коши для систем уравнений вида $A(t, y)\frac{dy}{dt} = g(t, y)$, создана на основе пакета *GEARIB*;

LSODES — предназначена для интегрирования разрешенных относительно производных систем дифференциальных уравнений с

разреженной матрицей Якоби, в основе лежит пакет *GEARS*; системы линейных алгебраических уравнений решаются программой *YSMP*;

LSODA — предназначена для автоматического выбора метода интегрирования;

LSODAR — предназначена для решения нелинейных функциональных уравнений.

Опишем еще один тип пакетов (см. работу [127]). Предназначенные для решения задач Коши для систем обыкновенных дифференциальных уравнений пакеты *EPISODE, EPISODEB, DISPL* появились из решения уравнений в частных производных после дискретизации уравнений по пространственным переменным. В пакетах *EPISODE* и *EPISODEB* используются формулы обратных разностей переменного порядка с переменным шагом по времени для решения жестких систем обыкновенных дифференциальных уравнений и формулы Адамса переменного порядка с переменным шагом для уравнений, не разрешенных относительно производных. Пакет *EPISODE* предназначен для решения систем дифференциальных уравнений с полной матрицей Якоби, а *EPISODEB* — с матрицей Якоби ленточной структуры.

В пакете *DISPL* используется метод Бубнова — Галеркина и *B*-сплайны для дискретизации уравнений по пространственным переменным.

ЗАКЛЮЧЕНИЕ

Применение метода конечных элементов для дискретизации и последующее решение дискретных задач дает возможность получить приближенные решения ряда задач теории термоупруго-пластичности, а также задач на колебания и устойчивость, связанных с расчетом на прочность узлов, элементов и конструкций в целом.

Для решения реальных научно-технических задач не существует готовых рецептов. Однако, как показывает опыт, методика, изложенная в монографии, облегчает решение практических задач. Действительно, можно многие научно-технические задачи вполне определенно отнести к тому или иному классу, описанному в данной работе (или исследованному в других работах), а следовательно, и найти путь решения конкретной прикладной задачи.

Несмотря на большие возможности современных ЭВМ, они все же имеют ограничения как по быстродействию, так и по объему запоминающих устройств. Об этих ограничениях не следует забывать при решении прикладных задач. Построение приемлемой математической модели, которая, с одной стороны, достаточно полно отражает физику протекающих процессов, а с другой — может быть решена на имеющейся ЭВМ, осуществляется, как правило, в несколько этапов, путем постепенного уточнения.

Для построения дискретных задач можно использовать как схемы МКЭ, когда в качестве базисных или допустимых функций используются линейные (полилинейные) полиномы, так и схемы высокого порядка точности, где в качестве базисных или допустимых функций применяются полиномы высших степеней.

Разбиение исходной области на элементы и выбор базисных функций производятся с учетом требований к точности искомого приближенного решения и теоретических оценок погрешности метода конечных элементов. Следует заметить, что эти оценки устанавливают только скорость сходимости в некоторой норме приближенного решения к точному (например, $O(h^{\nu})$), без конструктивного и достаточно точного определения константы при множителе h^{ν}, а потому для первого конкретного выбора конечного элемента приближенное решение МКЭ может оказаться недостаточно близким к точному решению исходной задачи.

В ряде случаев расчет приходится выполнять на нескольких сетках (для нескольких разбиений), чтобы обеспечить нужную точность решения.

В связи с рассмотрением погрешностей МКЭ необходимо упомянуть также ошибки, возникающие при аппроксимации криволинейной границы области определения задачи некоторой кусочно-полиномиальной границей. Поскольку в данной монографии подробно представлены исследования только для одномерных задач, интересующихся этим типом ошибок отсылаем к работам [50, 101] и другим литературным источникам.

В результате выполненной МКЭ дискретизации вопрос о построении приближенного решения в зависимости от исходной задачи сводится к решению либо системы алгебраических уравнений (линейных или нелинейных), либо алгебраической задачи на собственные значения, либо задачи Коши для обыкновенных дифференциальных уравнений и т. п. Здесь также важно помнить о следующих источниках погрешностей вычисленного на ЭВМ приближенного решения.

Во-первых, это ошибки, возникающие при вычислении коэффициентов и параметров указанных выше дискретных задач, например за счет численного интегрирования, замены начальной функции ее интерполянтом и т. п. Эти ошибки, которые можно рассматривать как наследственные для дискретных задач, достаточно полно исследованы и указаны способы их корреляции с погрешностями самого МКЭ.

Во-вторых, это погрешности, связанные с накоплением ошибок округления в процессе машинной реализации численного метода решения дискретной задачи. Чтобы уменьшить их влияние (снизить до уровня, не превышающего погрешность МКЭ), необходимо выбирать соответствующую длину машинного слова. В частности, при решении систем линейных алгебраических уравнений МКЭ, для которых априорные оценки числа обусловленности матрицы системы известны, такой выбор, как указано в гл. II, не представляет особого труда.

Выбор численного метода и алгоритма для решения дискретной задачи — далеко не простая проблема, имеющая свои сложности и особенности. Классификация, учитывающая специфические свойства дискретной задачи, помогает обеспечить эффективное вычисление искомого решения. В качестве иллюстрации коснемся лишь задачи решения систем линейных алгебраических уравнений МКЭ и существующих для этого наиболее эффективных численных алгоритмов. Как известно, матрицы таких систем — разреженные: ленточные, профильные или с произвольным расположением ненулевых элементов и ненулевых блоков. Вид их определяется исходной задачей, выбором конечного элемента и нумерацией узлов сетки. С переходом к решению дифференциальных и вариационных задач в трехмерных пространствах порядок и разреженность матриц существенно возросли. Для решения таких систем уравнений используются прямые и итерационные методы, различные алгоритмы которых ориентированы на специальный вид (и свойства) матриц системы. Для повышения эффективности использования алгоритмов прямых методов, в частности для экономии машинного времени и памяти, широко применяются алгоритмы предварительь-

ного упорядочения матрицы системы. (Краткая характеристика этих алгоритмов и соответствующая библиография приведены в гл. VII.)

Повышение эффективности использования итерационных методов для решения больших разреженных систем успешно достигается в ряде случаев применением специальной процедуры предварительной подготовки матрицы системы, а именно преобусловливанием. Эта процедура использовалась в параграфе VI.4 при описании решения прикладной задачи, а краткая библиография по данному вопросу дана в гл. VII.

Для минимизации объема вычислительной работы при решении различных дискретных задач, возникающих в рамках МКЭ, в последнее время широко применяются многосеточные итерационные процессы. Для ознакомления с этими методами можно воспользоваться их краткой характеристикой и библиографией, указанной в гл. VII.

Как уже упоминалось, априорные теоретические оценки погрешности МКЭ практически малопригодны для контроля достоверности вычисленных на ЭВМ решений. Поэтому особый интерес представляют удобные для машинной реализации достаточно надежные апостериорные оценки, не требующие к тому же слишком больших вычислительных затрат. В последние годы разработке именно таких оценок уделяется все больше внимания. Эти оценки особенно важны и потому, что являются основным средством конструирования адаптивных процедур для эффективного вычисления МКЭ решений требуемой точности или для получения решений наибольшей возможной точности в рамках допустимых вычислительных затрат. Сейчас уже общепризнано, что эффективное решение современных прикладных задач практически неосуществимо без использования того или иного адаптивного процесса.

Главные «рычаги» управления адаптивным процессом, предназначенным для уменьшения погрешности метода конечных элементов (т. е. повышения точности аппроксимации искомого решения), следующие: равномерное сгущение первоначально введенной сетки; изменение вида сетки (например, переход к неравномерной); повышение степени кусочно-полиномиальных базисных функций МКЭ или добавление к старому базису функций, описывающих поведение решения вблизи особых точек (т. е. переход к другому виду конечного элемента).

Каждый из этих способов управления имеет свои достоинства и недостатки, поэтому представляется целесообразным их комбинированное использование. Однако теоретически обоснованное построение подобной эффективной комбинации (даже для линейных задач) остается еще нерешенной проблемой. На практике обычно при организации адаптивного процесса ограничиваются пока использованием одного из упомянутых способов, чаще всего это измельчение (адаптация) сетки в зависимости от получаемых в ходе вычислений апостериорных оценок погрешности найденного МКЭ решения.

Весьма важным в рамках данной тематики является и вопрос о создании программного обеспечения (некоторой программной системы) для машинной реализации адаптивного применения МКЭ при решении больших реальных задач. Подробнее ознакомиться с апостериорными оценками погрешностей МКЭ и использованием этих оценок для

адаптивных процедур решения некоторых классов задач можно, например, по работе [153] и имеющейся в ней библиографии.

Данная монография посвящена алгоритмическим аспектам МКЭ и может служить основой для построения дискретных задач и эффективных численных методов их решения на ЭВМ.

Для массовых практических расчетов в прикладных организациях на основе теоретических исследований должны создаваться пакеты прикладных программ. Под пакетом программ понимается комплекс программных средств, включающий как набор функциональных модулей, характерных для данной прикладной области, так и наборы управляющих и вспомогательных модулей. Этот комплекс должен обеспечивать пользователю возможность в интерактивном режиме построить эффективную программу и вычислить по ней решение конкретной прикладной задачи в соответствии со спецификой задачи, а также с учетом особенностей ЭВМ, на которой выполняется вычисление искомого решения.

Пакет программ должен ускорить программирование и получение решения достаточно широкого класса математических и прикладных задач, дать пользователю не только решение задачи, но и некоторую достоверную информацию о точности вычисленного решения.

Пакеты программ, основанные на МКЭ, и создаваемые с их помощью программы должны иметь единый интерфейс с программными средствами систем автоматизированного проектирования и автоматизации научных исследований, а также единую с этими системами информационную базу.

В работе [73] приведена некоторая характеристика зарубежных пакетов программ, которые используются в машиностроении, судостроении, ядерных и аэрокосмических исследованиях, на транспорте, в строительстве и других прикладных областях. Эти пакеты организованы на различных принципах, но в алгоритмической основе их лежит метод конечных элементов. Использование таких пакетов прикладных программ сокращает время подготовки и решения задач на ЭВМ и повышает производительность труда пользователей.

СПИСОК ЛИТЕРАТУРЫ

1. *Артемьев С. С., Демидов Г. В.* A-устойчивый метод типа Розенброка четвертого порядка точности решения задачи Коши для жестких систем обыкновенных дифференциальных уравнений // Некоторые проблемы вычислительной и прикладной математики.— Новосибирск, 1975.— С. 212—219.
2. *Астраханцев Г. П.* Сведение задачи об изгибе пластины к системе уравнений второго порядка // Вариационно-разностные методы решения задач математической физики.— Новосибирск, 1976.— С. 62—72.
3. *Астраханцев Г. П.* О численном решении задачи Дирихле в произвольной области // Разностные и вариационно-разностные методы.— Новосибирск, 1977 — Вып. 2.— С. 63—72.
4. *Бабушка И., Витасек Э., Прагер М.* Численные процессы решения дифференциальных уравнений.— М. : Мир, 1969.— 368 с.
5. *Бате К., Вилсон Е.* Численные методы анализа и метод конечных элементов.— М. : Стройиздат, 1982.— 447 с.
6. *Бахвалов Н. С.* Численные методы (анализ, алгебра, обыкновенные дифференциальные уравнения).— М. : Наука, 1973.— 632 с.
7. *Березин И. С., Жидков Н. П.* Методы вычислений.— М. : Физматгиз, 1962.— Т. 1.
8. *Бурдаков О. П.* Некоторые глобально сходящиеся модификации метода Ньютона для решения систем нелинейных уравнений // Докл. АН СССР.— 1980.— 254, № 3.— С. 521—523.
9. *Вайнберг М. М.* Вариационные методы исследования нелинейных операторов.— М. : Гостехиздат, 1956.— 344 с.
10. *Вайнберг М. М.* Функциональный анализ.— М. : Просвещение, 1979.— 128 с.
11. *Варга Р.* Функциональный анализ и теория аппроксимации в численном анализе.— М. : Мир, 1974.— 126 с.
12. *Винокурова И. П.* Об одном варианте метода блочного исключения Гаусса для решения больших разреженных систем линейных алгебраических уравнений / Ин-т кибернетики АН УССР.— Киев, 1986.— 22 с.— Деп. в ВИНИТИ 16.12.86, № 8630-В Деп.
13. *Винокурова И. П., Черненко А. С.* Анализ логической факторизации для метода блочного исключения Гаусса при решении одной конечно-элементной задачи с большим числом неизвестных // Оптимизация алгоритмов программного обеспечения ЭВМ.— Киев, 1986.— С. 44—48.
14. *Воеводин В. В.* Вычислительные основы линейной алгебры.— М. : Наука, 1977.— 303 с.
15. *Воеводин В. В., Кузнецов Ю. А.* Матрицы и вычисления.— М. : Наука, 1984.— 320 с.
16. *Гаевский Х., Грёгер К., Захариас К.* Нелинейные операторные уравнения и операторные дифференциальные уравнения.— М. : Мир, 1978.— 336 с.
17. *Гольденвейзер А. Л., Лидский В. Г., Товстик П. Е.* Свободные колебания тонких упругих оболочек.— М. : Наука, 1979.— 384 с.

18. *Гордонова В. И., Морозов В. А.* Численные алгоритмы выбора параметров в методе регуляризации // Журн. вычисл. математики и мат. физики.— 1973.— **13**, № 3.— С. 539—545.
19. *Дейнека В. С., Молчанов И. Н.* Схема метода конечных элементов повышенного порядка точности для решения задач теории упругости // Там же.— 1981.— **21**, № 2.— С. 452—469.
20. *Демидов Г. В., Новиков Е. А.* Оценка ошибки одношаговых методов интегрирования обыкновенных дифференциальных уравнений // Численные методы механики сплошной среды.— 1985.— **16**, № 1.— С. 27—42.
21. *Демидович Б. П.* Лекции по математической теории устойчивости.— М. : Наука, 1967.— 472 с.
22. *Демьянушко И. В., Биргер И. А.* Расчет на прочность вращающихся дисков.— М. : Машиностроение, 1978.— 247 с.
23. *Джордж А., Лю Дж.* Численное решение больших разреженных систем уравнений.— М. : Мир, 1984.— 333 с.
24. *Динамика* авиационных газотурбинных двигателей.— М. : Машиностроение, 1981.— 232 с.
25. *Дьяконов Е. Г.* Итерационные методы решения разностных аналогов краевых задач для уравнений эллиптического типа.— Киев, 1970.— 144 с. (Современные численные методы: Материалы междунар. летней шк. по числ. методам, Киев, 1966; Вып. 4).
26. *Дьяконов Е. Г.* Разностные методы решения краевых задач. М. : Изд-во МГУ, 1971.— 242 с.
27. *Дьяконов Е. Г.* Некоторые классы операторов, эквивалентные по спектру, и их применение // Вариационно-разностные методы в математической физике.— Новосибирск, 1976.— С. 49—61.
28. *Дьяконов Е. Г.* О выборе триангуляции в проекционно-разностном методе, связанном с минимизацией вычислительной работы // Докл. АН СССР.— 1977.— **235**, № 4.— С. 757—760.
29. *Дьяконов Е. Г.* О решении систем уравнений проекционно-разностного метода для неотрицательных операторов // Вычислительные методы линейной алгебры.— Новосибирск, 1977.— С. 51—60.
30. *Дьяконов Е. Г.* О некоторых модификациях проекционно-разностных методов // Вестн. Моск. ун-та. Сер. Вычисл. математика и кибернетика.— 1977.— **1**, № 2.— С. 3—19.
31. *Дьяконов Е. Г.* Об использовании последовательностей сеток при решении сильноэллиптических систем // Вычислительные методы линейной алгебры.— Новосибирск, 1977.— С. 146—162.
32. *Дьяконов Е. Г.* Асимптотическая минимизация вычислительной работы при применении проекционно-разностных методов // Вариационно-разностные методы в математической физике.— Новосибирск, 1978.— С. 149—164.
33. *Дьяконов Е. Г.* Модифицированные итерационные методы в задачах на собственные значения // Вычислительные методы линейной алгебры.— Новосибирск, 1978.— С. 39—61.
34. *Дьяконов Е. Г.* Асимптотическая минимизация вычислительной работы при решении сильноэллиптических краевых задач // Теория кубатурных формул и вычислительная математика.— Новосибирск, 1980.— С. 31—37.
35. *Дьяченко В. П.* Основные понятия вычислительной математики.— М. : Наука, 1972.— 119 с.
36. *Дюво Г., Лионс Ж. Л.* Неравенства в механике и физике.— М. : Наука, 1980.— 383 с.
37. *Еремин А. Ю., Марьяшкин Н. Я.* Пакет программ *SPARSE* для решения систем линейных алгебраических уравнений с разреженными матрицами.— М., 1978.— 31 с.— В надзаг: БЦ АН СССР.
38. *Еремин А. Ю., Марьяшкин Н. Я.* Пакет программ *FEMS* для решения эллиптических краевых задач методом конечных элементов.— М., 1981.— 50 с.— В надзаг : ВЦ АН СССР.
39. *Зенкевич О.* Метод конечных элементов в технике.— М. : Мир, 1975.— 541 с.
40. *Зламал М.* Метод конечных элементов для уравнения теплопроводности // Вариационно-разностные методы решения задач математической физики.— Новосибирск, 1976.— С. 21—26.

41. *Ильин В. П., Катешов В. А.* Автоматизация описания двумерных краевых задач.— Препр. № 173 / ВЦ Сиб. отд-ния АН СССР.— Новосибирск, 1979.— 22 с.
42. *Ильюшин А. А., Огибалов Г. И.* Упруго-пластические деформации полых цилиндров.— М : Изд-во МГУ, 1960.— 227 с.
43. *Каган Б. М., Каневский М. М.* Цифровые вычислительные машины и системы.— М. : Энергия, 1973.— 679 с.
44. *Камель Х. А., Эйзенштейн Г. К.* Автоматическое построение сетки в двух- и трехмерных составных областях // Расчет упругих конструкций с использованием ЭВМ.— Л., Судостроение, 1974.— Т. 2.— С. 46—58.
45. *Ким Г. Д.* О статическом исследовании ошибок округления при решении систем линейных алгебраических уравнений итерационными методами // Ошибки округления в алгебраических процессах.— М., 1968.— С. 74—102.
46. *Климашевская Ю. А.* Об одном варианте метода конечных элементов решения первой начально-краевой задачи для уравнения параболического типа // Эффективная организация вычислений и численные методы.— Киев, 1983.— С. 60—66.
47. *Князев А. В.* О методах одновременного вычисления нескольких собственных векторов.— Препр. № 3724/16 / Ин-т атом. энергии им. И. В. Курчатова.— М.; 1983.— 18 с.
48. *Коллатц Л.* Численные методы решения дифференциальных уравнений.— М. : Изд-во иностр. лит., 1953.— 459 с.
49. *Коллатц Л.* Задачи на собственные значения.— М. : Наука, 1968.— 503 с.
50. *Корнеев В. Г.* Схемы метода конечных элементов высоких порядков точности.— Л. : Изд-во ЛГУ, 1977.— 208 с.
51. *Красносельский М. А. и др.* Приближенное решение операторных уравнений / Красносельский М. А., Вайнико Г. М., Забрейко П. П. и др.— М. : Наука, 1969.— 456 с.
52. *Крылов В. И.* Приближенное вычисление интегралов.— М. : Физматгиз, 1959.— 327 с.
53. *Курант Р., Гильберт Д.* Методы математической физики.— М. ; Л. : Гостехтеориздат, 1951.— 476 с.
54. *Ладыженская О. А., Солонников В. А., Уральцева Н. Н.* Линейные и квазилинейные уравнения параболического типа.— М. : Наука, 1967.— 736 с.
55. *Ладыженская О. А.* Краевые задачи математической физики.— М. : Наука, 1973.— 407 с.
56. *Ладыженская О. А., Уральцева Н. Н.* Линейные и квазилинейные уравнения эллиптического типа.— М. : Наука, 1973.— 576 с.
57. *Лехницкий С. Г.* Теория упругости анизотропного тела.— М. ; Л. : Гостехиздат, 1950.— 300 с.
58. *Ляпунов А. М.* Общая задача об устойчивости движения.— М. ; Л. : ОНТИ, 1935.— 382 с.
59. *Марчук Г. И.* Методы вычислительной математики.— М. : Наука, 1977.— 454 с.
60. *Мацокин А. М.* Автоматизация триангуляции областей с гладкой границей при решении уравнений эллиптического типа.— Препр. № 15 / ВЦ Сиб. отд-ния АН СССР.— Новосибирск, 1975.— 93 с.
61. *Мацокин А. М.* Вариационно-разностный метод решения эллиптических уравнений в круге // Численные методы механики сплошной среды.— 1976.— 7, № 7.— С. 51—62.
62. *Методы* и алгоритмы автоматического формирования сетки треугольных элементов : Программы и материалы по мат. обеспечению ЭВМ / Сост. Бабич Ю. Н., Цыбенко А. С.— Киев, 1977.— 93 с.— В надзаг. : Ин-т проблем прочности АН УССР.
63. *Михлин С. Г.* Проблемы минимума квадратичного функционала.— М. ; Л. : Гостехтеориздат, 1952.— 216 с.
64. *Михлин С. Г.* Численная реализация вариационных методов.— М. : Наука, 1966.— 432 с.
65. *Михлин С. Г.* Курс математической физики.— М. : Наука, 1968.— 575 с.
66. *Михлин С. Г.* Вариационные методы в математической физике.— М. : Наука, 1970.— 454 с.
67. *Михлин С. Г.* Вариационно-сеточная аппроксимация // Зап. науч. семинаров. Ленинг. отд-ние мат. ин-та.— 1974.— 48.— С. 32—188.

68. *Михлин С. Г.* Линейные уравнения в частных производных.— М. : Высш. шк., 1977.— 431 с.
69. *Молчанов И. Н., Николенко Л. Д.* Вариационный метод в некоторых краевых задачах с разрывными коэффициентами // Численный анализ.— Киев, 1975.— С. 71—83.
70. *Молчанов И. Н., Николенко Л. Д.* Метод конечных элементов и его применение для решения некоторых одномерных краевых задач.— Препр. № 14 / Ин-т кибернетики АН УССР.— Киев, 1976.— 72 с.
71. *Молчанов И. Н., Николенко Л. Д., Яковлев М. Ф.* О решении одного класса систем линейных алгебраических уравнений с вырожденными матрицами // Вычислительные методы линейной алгебры.— Новосибирск, 1977.— С. 97—109.
72. *Молчанов И. Н., Яковлев М. Ф.* Условия окончания итерационных процессов, гарантирующие заданную точность // Докл. АН УССР. Сер. А.—1980.— № 6.— С. 21—23.
73. *Молчанов И. Н.* О некоторых проблемах использования ЭВМ в прочностных расчетах.— Препр. № 10 / Ин-т кибернетики АН УССР.— Киев : 1981.— 39 с.
74. *Молчанов И. Н., Тарасова Л. Г.* Об одном критерии окончания итерационных процессов решения нелинейных уравнений // Докл. АН УССР. Сер. А.— 1981.— № 10.— С. 13—15.
75. *Молчанов И. Н. и др.* Пакет программ АРАС / Молчанов И. Н., Зубатенко В. С., Николенко Л. Д., Яковлев М. Ф. // Пакеты прикладных программ : Вычислительный эксперимент.— М., 1983.— С. 129—139.
76. *Молчанов И. Н., Попов А. В.* Схема повышенного порядка точности для некоторых задач на собственные значения // Вариационно-разностные методы в математической физике.— М., 1984.— С. 185—195.
77. *Молчанов И. Н., Николенко Л. Д., Незлина А. Ю.* Решение методом конечных элементов некоторых классов нелинейных задач.— Препр. № 35 / Ин-т кибернетики АН УССР.— Киев, 1984.— 50 с.
78. *Молчанов И. Н., Рябцев В. Е.* О реализации методов преобусловливания на многопроцессорных системах // Оптимизация численных методов решения задач на ЭВМ.— Киев, 1986.— С. 44—48.
79. *Молчанов И. Н.* Машинные методы решения прикладных задач. Алгебра и приближение функций.— Киев : Наук. думка, 1987.— 285 с.
80. *Молчанов И. Н.* Машинные методы решения прикладных задач. Дифференциальные уравнения.— Киев : Наук. думка, 1988.— 343 с.
81. *Молчанов И. Н. и др.* Структура и принципы организации ППП СПАН для вычисления собственных значений и собственных векторов матриц / Молчанов И. Н., Зубатенко В. С., Химич А. Н., Решетуха И. В. // Пакеты прикладных программ и численные методы.— Киев, 1988.— С. 92—96.
82. *Незлина А. Ю.* Сходимость метода конечных элементов при решении нелинейных краевых задач // Докл. АН УССР. Сер. А.— 1983.— № 7.— С. 16—19.
83. *Оганесян Л. А., Руховец Л. А.* Вариационно-разностные методы решения эллиптических уравнений.— Ереван : Изд-во АН АрмССР, 1979.— 336 с.
84. *Оден Дж.* Конечные элементы в нелинейной механике сплошных сред.— М. : Мир, 1976.— 464 с.
85. *Ортега Дж, Рейнболдт В. С.* Итерационные методы решения нелинейных систем уравнений со многими неизвестными.— М. : Мир, 1975.— 558 с.
86. *Парлетт Б.* Симметричная проблема собственных значений. Численные методы.— М. : Мир, 1983.— 382 с.
87. *Постнов В. А.* Метод суперэлементов в расчетах инженерных сооружений / Постнов В. А., Дмитриев С. А., Ентышев Б. К., Родионов А. А.— Л. : Судостроение, 1979.— 287 с.
88. *Приказчиков В. Г.* Прототипы итерационных процессов в задачах на собственные значения // Дифференц. уравнения.— 1980.— **16**, № 9.— С. 1688—1697.
89. *Приказчиков В. Г., Химич А. Н.* Итерационный метод решения задач устойчивости и колебания пластин и оболочек // Прикл. механика.— 1984.— **20**, № 1.— С. 88—94.
90. *Ракитский Ю. В., Устинов С. М., Черноруцкий И. Г.* Численные методы решения жестких систем.— М. : Наука,— 1979.— 208 с.
91. *Решетуха И. В., Рудич О. В.* Использование переупорядочения разреженных матриц при решении задач на собственные значения методом итерирования

подпространств // Оптимизация вычислений и численные методы.— Киев, 1987.— С. 17—20.
92. *Розин Л. А.* Вариационные постановки задач для упругих систем.— Л. : Изд-во ЛГУ, 1978.— 223 с.
93. *Савинов Г. В.* Метод сопряженных градиентов для определения собственных значений // Тр. Ленингр. кораблестроит. ин-та.— 1977.— Вып. 120.— С. 55—58.
94. *Самарский А. А.* Введение в теорию разностных схем.— М. : Наука, 1971.— 552 с.
95. *Самарский А. А.* Теория разностных схем.— М. : Наука, 1977.— 653 с.
96. *Самарский А. А., Николаев Е. С.* Методы решения сеточных уравнений.— М. : Наука, 1978.— 559 с.
97. *Смирнов В. И.* Курс высшей математики.— М. : Физматгиз, 1959.— Т. 5.
98. *Соболев С. Л.* Некоторые применения функционального анализа в математической физике.— Л. : Изд-во ЛГУ, 1950.— 225 с.
99. *Современные* численные методы решения обыкновенных дифференциальных уравнений / Под. ред. Дж. Холла, Дж. Уатта — М. : Мир, 1979.— 312 с.
100. *Страховская Л. Г.* Итерационный метод вычисления первой собственной функции эллиптического оператора // Журн. вычисл. математики и мат. физики.— 1977.— **17**, № 3.— С. 649—664.
101. *Стренг Г., Фикс Дж.* Теория метода конечных элементов.— М. : Мир, 1977.— 349 с.
102. *Сьярле Ф.* Метод конечных элементов для эллиптических задач.— М. : Мир, 1980.— 512 с.
103. *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач.— М. : Наука, 1979.— 224 с.
104. *Тихонов А. Н.* О приближенных системах линейных алгебраических уравнений // Журн. вычисл. математики и мат. физики.— 1980.— **20**, № 6.— С. 1373—1383.
105. *Уилкинсон Дж. Х.* Алгебраическая проблема собственных значений.— М. : Наука, 1970.— 563 с.
106. *Уилкинсон Дж. Х., Райнш К.* Справочник алгоритмов на языке АЛГОЛ.— М. : Машиностроение, 1976.— 390 с.
107-108. *Фаддеев Д. К., Фаддеева В. Н.* Вычислительные методы линейной алгебры.— 2-е изд., доп.— М. ; Л. : Физматгиз, 1963.— 734 с.
109. *Фаддеева В. Н.* Сдвиг для систем с плохо обусловленными матрицами // Журн. вычисл. математики и мат. физики.— 1965.— **5**, № 5.— С. 907—911.
110. *Федоренко Р. П.* О скорости сходимости одного итерационного процесса // Там же.— 1964.— **3**, № 8.— С. 559—564.
111. *Федоренко Р. П.* Итерационные методы решения разностных эллиптических уравнений // Успехи мат. наук.— 1973.— **28**, № 2.— С. 121—182.
112. *Фихтенгольц Г. М.* Курс дифференциального и интегрального исчисления.— М. ; Л. : Гостехтеориздат, 1948.— Т. 1.
113. *Черненко А. С.* О двух подходах в формировании общей СЛАУ метода конечных элементов / Ин-т кибернетики АН УССР.— Киев, 1986.— 20 с.— Деп. в ВИНИТИ 16. 12. 86, № 8631-В Деп.
114. *Чубань В. Д.* Об одном эффективном прямом методе решения систем линейных алгебраических уравнений метода конечных элементов // Журн. вычисл. математики и мат. физики.— 1978.— **18**, № 5.— С. 1075—1082.
115. *Шайдуров В. В.* О решении спектральной вариационно-разностной задачи на последовательности сеток // Вариационно-разностные методы в математической физике.— М., 1984.— С. 149—160.
116. *Штеттер Х.* Анализ методов дискретизации для обыкновенных дифференциальных уравнений.— М. : Мир, 1978.— 461 с.
117. *Эйдус Д. М.* О смешанной задаче теории упругости // Докл. АН СССР.— 1951.— **76**, № 2. С. 181—184.
118. *Экланд И., Темам Р.* Выпуклый анализ и вариационные проблемы.— М. : Мир, 1979.— 399 с.
119. *Эльсгольц Л. Э.* Дифференциальные уравнения и вариационное исчисление.— М. : Наука, 1965.— 402 с.
120. *Allgower E. L., Georg K.* Simplicial and continuation methods for approximating fixed points and solutions to systems of equations // SIAM Rev.— 1980.— **22**. N 1.— P. 28—85.

121. *Axelsson O.* A class of interative methods for finite element equations // Comput. Meth. Appl. Mech. and Eng.— 1976.— 9, N 2.— P. 123—137.
122. *Axelsson O.* A survey of preconditioned iterative methods for linear systems of algebraic equations // BIT (Dan.).— 1985.— 25, N 1.— P. 166—187.
123. *Baker G. A.* Error estimates for finite element methods for second order hyperbolic equations // SIAM J. Numer. Anal.— 1976.— 13, N 4.— P. 564—576.
124. *Bramble J., Sammon P.* Efficient higher order single step methods for parabolic problems // Math. Comput.— 1980.— 35, N 151.— P. 655—677.
125. *Brent R. P.* Some efficient algorithms for solving systems of non-linear equations / SIAM J. Numer. Anal.— 1973.— 10, N 2.— P. 327—344.
126. *Bui T. D., Oppenheim A. K., Pratt D. T.* Recent advances in methods for numerical solution of ODE initial value problems // J. Comput. and Appl. Math.— 1984.— 11, N 3.— P. 283—296.
127. *Byrne G. D.* Some software for stiff systems of differential euations // Numerical Methods for Differential Equations and Simulation : Proc. IMACS (AICA) Int. Symp. Simul. Software and Numer. Meth. Differ. Equat., Blacksburg, Va, 1977, March 9—11.— Amsterdam etc., 1978.— P. 45—50.
128. *Cline A. K. et. al.* An estimate for the condition number of a matrix / Cline A. K., Moler C. B., Stewart C. W., Wilkinson J. H. // SIAM J. Numer. Anal.— 1979.— 16, N 2.— P. 368—375.
129. *Courant R.* Variational methods for the solution of problems of equilibrium and vibrations // Bull. Amer. Math. Soc.— 1943.— 49, N 1.— P. 1—23.
130. *Crawford C. R.* Reduction of a band-symmetric generalized eigenvalue problem // Communs ACM.— 1973.— 16, N 1.— P. 41—44.
131. *Cullum J. K., Willoughby R. A.* Lanczos algorithms for large symmetric eigenvalue computation.— Boston ets. : Birk häuset, 1985.— Vol. 2 : Programms.— 497 p.
132. *Dahlquist G.* A special stability problem for linear multistep methods // BIT.— 1963.— 3, N 1.— P. 27—43.
133. *Dongarra J. J. et al.* LINPACK users guide / Dongarra J. J., Bunch J. R., Moler C. B., Stewart G. W.— Philadelphia : SIAM.— 1979.— 366 p.
134. *Douglas J., Dupon T.* Galerkin methods for parabolic equations // SIAM J. Numer. Anal.— 1970.— 7, N 4.— P. 575—626.
135. *Douglas J., Dupont T., Ewing R.* Incomplete iteration for time-stepping Galerkin method for quasi-linear parabolic problem // Ibid — 1979.— 16, N 3.— P. 503—522.
136. *Evans D. J.* The use of pre-conditioning in iterative methods for solving linear equations with symmetric positive definite matrices // J. Inst. Math. and Appl.— 1968.— 4, N 3.— P. 295—314.
137. *Gear C. W.* Numerical initial value problem in ordinary differential equations.— New Jersey : Prentice-Hall, 1971.— 253 p.
138. *Gekeler E.* Linear multistep methods and Galerkin procedures for initial boundary value problems // SIAM J. Numer. Anal.— 1976.— 13, N 4.— P. 536—548.
139. *Gustafsson I.* A class of first order factorization methods // BIT.— 1978.— 18, N 4.— P. 142—156.
140. *Hackbusch W.* A multi-grid method applied to a boundary value problem with variable coefficients in a rectangle.— Köln, 1977.— 48 p.— (Angew. Math. / Math. Inst. Univ. Köln; Rept. 77-17).
141. *Hackbusch W.* On the multi-grid method applied to difference equations // Computing.— 1978.— 20, N 4.— P. 291—306.
142. *Hackbusch W.* Analysis and multi-grid solutions of mixed finite element and mixed difference equations.— Prepr. / Math. Inst. Ruhr-Univ.— Bochum, Oct. 1980.— 29 p.
143. *Hackl J., Wacker H. J., Zulehner W.* An efficient step size control for continuation methods // BIT.— 1980.— 20, N 4.— P. 475—485.
144. *Hindmarsh A. C.* ODEPACK, a systematized collection of ODE solvers // Sci. Comput. : Appl. Math. and Comput. Phys. Sci. 10th IMACS World Congr. Syst. Simul. and Sci. Comput., Montreal 8-13, Aug., 1982.— Amsterdam etc., 1983.— Vol. 1.— P. 55—64.
145. *Kaps P., Ponn S. W. H., Bui T. D.* Rosenbrock methods for stiff ODEs : a comparison of Richardson extrapolation and embedding technique // Computing.— 1985.— 34, N 1.— P. 17—40.

146. *Kershaw D. S.* The incomplete Cholesky — conjugate gradient method for the iterative solution of systems of linear equations // J. Comput. Phys.— 1978.— **26**, N 1.— P. 43—65.
147. *Lambert J. D.* Computational methods in ordinary differential equations.— London : Wiley, 1973.— 278 p.
148. *Matthies H., Strang J.* The solution of nonlinear finite element equations // Int. J. Numer. Meth. Eng.— 1979.— **14**, N 11.— P. 1613—1626.
149. *Matrix* eigensystem routines / Smith B. T., Boyle I. M., Dongarra J. J. et al.— Berlin ; New York : Springer — Verlag, 1976.— 551 p.— (EISPACK guide).— (Lect. Notes Comput. Sci.; vol. 6).— Ibid. / Garbow B. S., Boyle J. M., Dongarra J. J., Moler C. B.— Berlin; New York : Springer — Verlag, 1977.— 343 p. (EISPACK guide extension).— (Lect. Notes Comput. Sci.; vol. 51).
150. *Moler C.* Three research problems in numerical linear algebra // Proc. Symp. Appl. Math.— 1978.— Vol. 22.— P. 1—18.
151. *Moré J. J., Cosnard M. Y.* Numerical solution of non-linear equations // ACM Trans. Math. Software.— 1979.— **5**, N 1.— P. 64—85.
152. *Munksgaard N.* Solving sparse symmetric sets of linear equations by pre-conditioned conjugate gradients // Ibid.— 1980.— **6**, N 2.— P. 206—216.
153. *Rheinboldt W. C.* Error estimates for non-linear finite element computations // Comput. and Struct.— 1985.— **20**, N 1—3.— P. 91—98.
154. *Rosenbrock H. H.* Some general implicit processes for the numerical solution of differential equations // Comput. J. 1963.— **5**, N 4.— P. 329—330.
155. *Ruhe A.* SOR-methods for the eigenvalue problem with large sparse matrices // Math. Comput.— 1974.— **28**, N 127.— P. 695—710.
156. *Zenizek A.* Convergence of a finite element procedure for solving boundary value problems of systems of elliptic equations // Apl. Mat.— 1969.— **14**, N 3.— P. 355—377.
157. *Zlamal M.* On the finite element method // Numer. Math.— 1968.— **12**, N 4.— P. 394—409.
158. *Zlamal M.* On some finite element procedures for solving second order boundary problems // Ibid.— 1969.— **14**, N 1.— P. 42—48.
159. *Zlamal M.* Finite element methods for parabolic equations // Math. Comput.— 1974.— **28**, N 126.— P. 393—404.
160. *Zlamal M.* Finite element multi-step discretization of parabolic boundary value problems // Ibid.— 1975.— **29**, N 130.— P. 350—359.

Научное издание

МОЛЧАНОВ Игорь Николаевич
НИКОЛЕНКО Лариса Даниловна

ОСНОВЫ МЕТОДА КОНЕЧНЫХ ЭЛЕМЕНТОВ

Художественный редактор *И. П. Антонюк*
Технический редактор *И. Н. Лукашенко*
Корректоры *Е. А. Михалец, Л. М. Тищенко*

ИБ № 9888

Сдано в набор 30.05.88. Подп. в печ. 28.11.88. БФ 01673. Формат 60×90/16. Бум. тип. № 1. Лит. гарн. Выс. печ. Усл. печ. л. 17,0. Усл. кр.-отт. 17,0. Уч.-изд. л. 17.3. Тираж 2470 экз. Заказ 8—1728. Цена 3 р. 50 к.

Издательство «Наукова думка». 252601, Киев, ул. Репина, 3.

Отпечатано с матриц Головного предприятия республиканского производственного объединения «Полиграфкнига». 252057 Киев ул. Довженко, 3 в Киевской книжно-журнальной типографии научной книги. 252004 Киев 4, ул. Репина, 4. Зак. 9-32.

ИЗДАТЕЛЬСТВО «НАУКОВА ДУМКА»
В 1989 Г. ВЫПУСТИТ
В СВЕТ КНИГИ:

ДИНАМИКА ОДНОМЕРНЫХ ОТОБРАЖЕНИЙ

/ А. Н. Шарковский, С. Ф. Коляда, А. Г. Сивак, В. В. Федоренко.—
12 л.— 2 р. 70 к.

В монографии дается краткая характеристика свойств одномерных динамических систем, задаваемых произвольными непрерывными отображениями. Рассматриваются наиболее простые нелинейные отображения интервала — унимодальные, содержащие одну критическую точку. Для них рассматриваются периодические точки и периодические интервалы различных типов, их сосуществование, возможные бифуркации и связанные с ними универсальные свойства в семействах отображений. Строится спектральное разложение отображения, анализируются его устойчивость и бифуркации. Изучаются перемешивающие аттракторы, квазиаттракторы, репеллеры, области притяжения аттракторов. Представлены результаты исследования топологической энтропии, инвариантных мер отображений, обсуждается типичность различных свойств динамических систем.

Для научных и инженерно-технических работников, а также преподавателей вузов, интересующихся нелинейной динамикой и ее приложениями.

Шор Н. З., Стеценко С. И.

КВАДРАТИЧНЫЕ ЭКСТРЕМАЛЬНЫЕ ЗАДАЧИ И НЕДИФФЕРЕНЦИРУЕМАЯ ОПТИМИЗАЦИЯ.— 10 л.— 2 р. 10 к.

В монографии исследуются квадратичные экстремальные задачи (КЭЗ). Показано, что целевая функция и ограничения КЭЗ выражаются полиномами со степенями, не превышающими 2, позволяющими строить двойственные оценки с использованием методов недифференцируемой оптимизации. На основе применения этих оценок рассмотрены комбинаторные задачи на графах. Для задачи о максимальном независимом множестве графа предложен оригинальный алгоритм, показано равенство известных оценок Ловаса и двойственных оценок. Путем сведения к КЭЗ получены оценки глобального минимума полиномиальной функции. Рассмотрены методы решения задач минимизации невыпуклых квадратичных функций, в частности задач о линейной дополнительности, играющих большую роль в экономических исследованиях.

Для научных работников и инженеров, интересующихся вопросами математического программирования.

Предварительные заказы на эти книги принимает магазин издательства «Наукова думка» (252001 Киев 1, ул. Кирова, 4), который вышлет их иногородним заказчикам после выхода из печати наложенным платежом.