Глава 1

Метод конечных элементов в одномерном случае

Основная область применения метода конечных элементов это, конечно же, решение уравнений в частных производных для двумерных и трехмерных областей, особенно, когда область имеет сложную геометрическую форму. Однако, для понимания идей метода и большинства его наиболее важных особенностей, удобнее начать с рассмотрения одномерных краевых задач. Традиционно принято излагать основы метода, используя вариационную формулировку задачи. Однако, на взгляд авторов, это существенно облегчает лишь проведение строгих математических доказательств, но не вносит ничего нового в понимание сути метода. Намного проще изначально использовать проекционный подход, трактуя метод конечных элементов как разновидность метода Галеркина.

1.1 Сильное и слабое решение задачи

Рассмотрим первую краевую задачу (задачу Дирихле) для определения функции $u(x)$

$$-\frac{d}{dx}\left(p(x)\frac{du(x)}{dx}\right)+q(x)u(x)=f(x), \qquad 0<x<1, \tag{1.1}$$

$$u(0)=0, \qquad u(1)=0, \tag{1.2}$$

где $p(x)$, $q(x)$, $f(x)$ — известные функции.

Предположим, что функции $q(x)$, $f(x)$ непрерывны, а функция $p(x)$ — непрерывно дифференцируема. Решение задачи (1.1), (1.2), являющееся дважды непрерывно дифференцируемой функцией, будем называть ***сильным решением***.

Можно попытаться переформулировать задачу (1.1), (1.2) с целью снижения требований, накладываемых на непрерывность функций $p(x)$, $q(x)$, $f(x)$ и функции $u(x)$. Для этого умножим уравнение (1.1) на некоторую

функцию $v(x)$ и проинтегрируем на отрезке $[0,1]$

$$-\int_0^1 \frac{d}{dx}\left(p(x)\frac{du(x)}{dx}\right)v(x)\,dx+\int_0^1 q(x)u(x)v(x)\,dx=\int_0^1 f(x)v(x)\,dx.$$

Используя формулу интегрирования по частям, имеем

$$\int_0^1 p\frac{du}{dx}\frac{dv}{dx}\,dx-p\frac{du}{dx}v\Big|_0^1+\int_0^1 quv\,dx=\int_0^1 fv\,dx.$$

Потребуем, чтобы функция $v(x)$ удовлетворяла тем же краевым условиям, что и $u(x)$, т. е.

$$v(0)=0,\quad v(1)=0. \tag{1.3}$$

Тогда

$$\int_0^1\left(p\frac{du}{dx}\frac{dv}{dx}+quv-fv\right)dx=0,\quad v(0)=0,\quad v(1)=0. \tag{1.4}$$

Если соотношение (1.4) выполнено для любых $v(x)$, то оно называется ***слабой формой*** (или вариационной формой) записи задачи (1.1), (1.2). На самом деле, следует дополнительно указать какому классу принадлежат функции $v(x)$, например, $v\in C^1$.

Сам способ получения соотношения (1.4) показывает, что если функция $u(x)$ является решением задачи (1.1), (1.2), то эта же функция будет решением (1.4). Обратное же утверждение в общем случае неверно. Это видно уже из того, что для функций $p(x)$, $q(x)$, $f(x)$ и $u(x)$, входящих в (1.4), можно требовать гораздо меньших ограничений на гладкость, чем для соответствующих функций задачи (1.1), (1.2). Так, например, можно считать функции $p(x)$, $q(x)$, $f(x)$ непрерывными ($p,q,f\in C^0$), а функции $u(x)$, $v(x)$ непрерывно дифференцируемыми ($u,v\in C^1$). На самом деле, можно даже требовать от $p(x)$, $q(x)$, $f(x)$ кусочной непрерывности, а от $u(x)$, $v(x)$ — кусочной гладкости. Заметим также, что для функции $u(x)$, входящей в (1.4), не требуется даже выполнения краевых условий (1.2).

Функцию $u(x)$, являющуюся решением задачи (1.4), будем называть ***слабым решением*** задачи (1.1), (1.2). Из вышесказанного ясно, что сильное и слабое решения задачи (1.1), (1.2) в общем случае не совпадают. Подчеркнем, что сильные и слабые решения могут различаться и в случае, когда для функций $p(x)$, $q(x)$, $f(x)$ сохранены такие же требования на гладкость, как в исходной задаче (1.1), (1.2).

Проблемы связи между собой сильных и слабых решений задачи являются очень важными для метода конечных элементов. Дело в том, что традиционно метод используется для получения приближенных решений именно задач в слабой формулировке (т. е. слабых решений)[1]. Различие

[1] В частности, FreeFem++ предназначен для решения именно таких задач.

между слабым и сильным решениями может быть одной из причин, по которой приближенное решение, полученное методом конечных элементов, может не иметь никакой связи с решением исходной задачи.

Далее вопросы, связанные со сходимостью слабых решений к сильному, не рассматриваются, т. к. основная цель книги — это описание работы с FreeFem++. Ограничимся лишь ссылкой на [3, 5], где эти вопросы исследуются применительно к методу конечных элементов и в которых имеется достаточно обширная библиография.

1.2 Построение приближенного решения задачи

1.2.1 Слабое решение

Будем искать приближенное решение задачи (1.4) в виде

$$u^h(x) = \sum_{k=1}^{n-1} c_k \varphi_k(x), \tag{1.5}$$

где $u^h(x)$ — приближенное решение, функции $\varphi_k(x)$ предполагаются известными, линейно независимыми и называются *базисными функциями* (другие названия — *тестовые* функции, *пробные* функции), c_k — коэффициенты, подлежащие определению.

Формула (1.5) задает аппроксимацию функции $u(x)$ в виде некоторого сужения (проектирования) на *конечномерное* пространство, определяемое базисом $\varphi_k(x)$, $k = 1, \ldots, n-1$. Точность аппроксимации, естественно, будет зависеть от того, насколько «хорошо» выбранное конечномерное пространство приближает исходное пространство, которому должна принадлежать функция $u(x)$. Заметим, что величина u^h при заданных $\varphi_k(x)$ полностью определяется набором чисел $(c_1, c_2, \ldots, c_{n-1})$, т. е. можно считать, что $u^h = (c_1, c_2, \ldots, c_{n-1})$ является вектором, принадлежащим $\mathbb{R}^{n-1}$.

Подставляя (1.5) в (1.4), получим

$$\int_0^1 \left(p \sum_{k=1}^{n-1} c_k \varphi_k' v' + q \sum_{k=1}^{n-1} c_k \varphi_k v - f v \right) dx = 0.$$

Учитывая, что (1.4) должно выполняться для любых $v(x)$, принадлежащих некоторому классу, выберем в качестве $v(x)$ некоторый набор функций, а именно $v(x) = \varphi_i(x)$, $i = 1, \ldots, n-1$. В соответствии с требованиями $v(0) = 0$, $v(1) = 0$ (см. (1.4)) будем считать, что

$$\varphi_k(0) = 0, \quad \varphi_k(1) = 0, \quad k = 1, \ldots, n-1. \tag{1.6}$$

В этом случае имеем

$$\int_0^1 \left(p \sum_{k=1}^{n-1} c_k \varphi_k' \varphi_i' + q \sum_{k=1}^{n-1} c_k \varphi_k \varphi_i - f \varphi_i \right) dx = 0, \quad i = 1, \ldots, n-1.$$

Перепишем это соотношение в виде

$$\sum_{k=1}^{n-1} A_{ik} c_k = b_i, \quad i = 1, \ldots, n-1, \tag{1.7}$$

где введены следующие обозначения

$$A_{ik} = \int_0^1 \left(p\varphi_i'\varphi_k' + q\varphi_i\varphi_k\right) dx, \quad b_i = \int_0^1 f\varphi_i \, dx, \quad i, k = 1, \ldots, n-1. \tag{1.8}$$

Таким образом, для определения c_k имеем систему линейных алгебраических уравнений. Решая эту систему и подставляя c_k в (1.5), получим приближенное решение задачи (1.4). Подчеркнем, что приближенное решение (1.5) в силу условий (1.6) автоматически удовлетворяет краевым условиям (1.2) для исходной задачи.

1.2.2 Метод Галеркина. Сильное решение

Разыскивать решение в виде (1.5) можно и для исходной задачи. Подставляя (1.5) в (1.1), имеем

$$\delta(x) = -\frac{d}{dx}\left(p\sum_{k=1}^{n-1} c_k\varphi_k'\right) + q\sum_{k=1}^{n-1} c_k\varphi_k - f$$

или

$$\delta(x) = -p'\sum_{k=1}^{n-1} c_k\varphi_k' - p\sum_{k=1}^{n-1} c_k\varphi_k'' + q\sum_{k=1}^{n-1} c_k\varphi_k - f. \tag{1.9}$$

Здесь $\delta(x)$ — невязка, возникающая после подстановки приближенного решения $u^h(x)$ в уравнение (1.1).

Умножим соотношение (1.9) на φ_i и, проинтегрировав на $[0, 1]$, потребуем выполнения равенств

$$\int_0^1 \delta(x)\varphi_i(x)\, dx = -\int_0^1 \left(\sum_{k=1}^{n-1} c_k \left(p'\varphi_k' + p\varphi_k'' - q\varphi_k\right) + f\right)\varphi_i \, dx = 0 \tag{1.10}$$

или

$$\sum_{k=1}^{n-1} c_k \int_0^1 \left(-p'\varphi_k'\varphi_i - p\varphi_k''\varphi_i + q\varphi_k\varphi_i\right) dx - \int_0^1 f\varphi_i \, dx = 0, \quad i = 1, \ldots, n-1.$$

Таким образом, вновь получена система линейных алгебраических уравнений для определения c_k

$$\sum_{k=1}^{n-1} \tilde{A}_{ik} c_k = b_i, \quad i = 1, \ldots, n-1,$$

где

$$\tilde{A}_{ik} = \int_0^1 (-p'\varphi_k'\varphi_i - p\varphi_k''\varphi_i + q\varphi_k\varphi_i)\,dx, \quad b_i = \int_0^1 f\varphi_i\,dx. \tag{1.11}$$

Заметим, что функция (1.5) будет являться приближением решения исходной задачи (1.1), (1.2), если на базисные функции $\varphi_k(x)$ будут наложены дополнительные ограничения (1.6) — требования удовлетворения граничным условиям (1.2).

Укажем на основное различие между матрицами A_{ik} и $\tilde{A}_{ik}$. Для вычисления матрицы A_{ik} (см. (1.8)) от функций φ_k требуется, чтобы φ_k' была кусочно-непрерывной. В случае матрицы $\tilde{A}_{ik}$ (см. (1.11)) на функции φ_k следует накладывать более сильные ограничения — требуется, чтобы φ_k'' была кусочно-непрерывной. Конечно, имеется в виду, что матрица $\tilde{A}_{ik}$ вычисляется непосредственно по приведенной формуле и операция интегрирования по частям в (1.11) не используется. Если же в (1.11) произвести операцию интегрирования по частям, то вновь получится слабая формулировка задачи и соотношения (1.11) совпадут с (1.7).

Как уже говорилось, формула (1.5) определяет аппроксимацию функции $u(x)$ в виде некоторого сужения (проектирования) на некоторое *конечномерное* пространство, определяемое базисом $\varphi_k(x)$, $k = 1, \ldots, n-1$. Соотношения (1.10) при условии выполнения (1.6) задают *проекцию* невязки $\delta(x)$ (см. (1.9)) исходной задачи (1.1), (1.2) на *то же самое* пространство. Конечно, это вовсе не обязательно (в случае более сложных задач — просто неправильно) и соотношения (1.10) следует заменить более общими, задающими проектирование невязки $\delta(x)$ на конечномерное пространство, определяемое иным базисом, например, $\psi_k(x)$, $k = 0, \ldots, n-1$

$$\int_0^1 \delta(x)\psi_i(x)\,dx = -\int_0^1 \left(\sum_{k=1}^{n-1} c_k\,(p'\varphi_k' + p\varphi_k'' - q\varphi_k) + f\right)\psi_i dx = 0. \tag{1.12}$$

Подчеркнем, что в этом случае матрица $\tilde{A}_{ik}$ будет определяться соотношениями, отличными от (1.11), и переход от $\tilde{A}_{ik}$ к A_{ik} будет невозможен. Во избежание недоразумений, заметим, что способ проектирования (1.10) или (1.12) при помощи скалярного произведения, определяемого интегрированием, уже подразумевает некоторую слабую формулировку исходной задачи. Действительно, если, например, в формулировку исходной задачи (1.1), (1.2) включено требование принадлежности функций p', q, f, u'' классу непрерывных функций, то при построении приближенного решения с использованием соотношений (1.10) уже можно ограничиваться требованием принадлежности функций p', q, f, φ_k'' (а следовательно и u'') лишь классу кусочно-непрерывных функций.

1.2.3 Конечно-разностный метод

Описанные приемы построения приближенного решения, конечно же, не являются специфическими способами решения исходной задачи. При-

ведем, в частности, конечно-разностный (сеточный) метод, важный для дальнейших целей. Для простоты ограничимся случаем, когда $p(x) = 1$, $q(x) = 0$, т. е. задачей

$$-u''(x) = f(x), \quad 0 < x < 1, \quad u(0) = 0, \quad u(1) = 0. \tag{1.13}$$

Разобьем отрезок $[0, 1]$ на интервалы с одинаковой длиной $[x_k, x_{k+1}]$, $k = 0, \ldots, n-1$, $x_i = ih$, $h = 1/n$. Рассмотрим уравнение (1.13) в точке x_k и аппроксимируем производную $u''(x_k)$ центральной конечной разностью (см. [4]). Тогда, с учетом краевых условий, для определения $u_k = u(x_k)$, $k = 0, \ldots, n$ получим систему линейных алгебраических уравнений

$$-u_{k-1} + 2u_k - u_{k+1} = h^2 f(x_k), \quad k = 1, \ldots, n-1, \quad u_0 = 0, \quad u_n = 0. \tag{1.14}$$

Матрица коэффициентов для этой системы имеет трехдиагональный вид

$$\begin{pmatrix} 2 & -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ \ldots \\ u_{n-2} \\ u_{n-1} \end{pmatrix} = \begin{pmatrix} h^2 f(x_1) \\ h^2 f(x_2) \\ h^2 f(x_3) \\ h^2 f(x_4) \\ \ldots \\ h^2 f(x_{n-2}) \\ h^2 f(x_{n-1}) \end{pmatrix}. \tag{1.15}$$

1.3 Выбор базисных функций

Проведем некоторый сравнительный анализ метода, использованного для построения слабого решения, и конечно-разностного метода (аналогично можно сделать и для метода Галеркина). В дальнейшем, если не оговорено противное, и для метода Галеркина, и для метода построения слабого решения, изложенного в п. 1.2.1, будем использовать общее название — *проекционные методы*, имея ввиду тот факт, что при соответствующих требованиях на гладкость и проведении интегрирования по частям матрицы A_{ik} и $\tilde{A}_{ik}$ будут, конечно же, совпадать.

При построении приближенного решения u^h по формуле (1.5) необходимо задавать набор базисных функций φ_1, φ_2,..., φ_{n-1}, удовлетворящих условиям (1.6). Если какая-либо дополнительная информация о задаче неизвестна, то обычно в качестве такого набора выбирают полиномиальные или тригонометрические функции. Например, для рассматриваемой задачи это могут быть

$$\varphi_1(x) = x(1-x), \quad \varphi_2(x) = x^2(1-x)^2, \ldots$$

или

$$\varphi_k(x) = \sin \pi k x.$$

Вычислительная практика показывает, что при удачном выборе базисных функций бывает достаточно ограничиться их небольшим количеством.

В общем же случае, когда число n велико, для определения коэффициентов c_k придется решать систему линейных уравнений (1.7) с большой размерностью $(n-1)\times(n-1)$ и сильно заполненной матрицей A_{ik}, определенной формулами (1.8). При решении такой системы могут возникнуть значительные трудности — большие погрешности при вычислении, увеличение времени расчетов и т. п.

Напротив, при использовании конечно-разностной схемы (1.14) матрица в (1.15) трехдиагональна и для нахождения решения $u(x_0)$, $u(x_1)$,..., $u(x_n)$ существуют эффективные алгоритмы. Более того, значения функции $u(x)$ непосредственно определяются в точках x_k сразу же после решения системы (1.14), тогда как для нахождения приближенного решения по формуле (1.5) следует сначала, решая (1.7), найти c_k, и лишь затем использовать формулу (1.5).

На самом деле такие преимущества конечно-разностный метод имеет лишь для сравнительно простых уравнений, краевых условий и областей, в которых решается задача. Уже в двумерном случае использование конечно-разностного метода приводит к серьезным проблемам при конструировании краевых условий в областях сложной формы (например, отличных от прямоугольника). Кроме того, часто возникают проблемы сохранения различных свойств исходных задач.

Поясним это на следующем примере. Рассмотрим связанный с задачей (1.1), (1.2) дифференциальный оператор L (строго говоря, в определение следует еще включать требования гладкости)

$$(Lu)(x) \stackrel{def}{=} -\frac{d}{dx}\left(p(x)\frac{du(x)}{dx}\right)+q(x)u(x), \quad 0<x<1, \tag{1.16}$$

$$u(0)=0, \quad u(1)=0.$$

Дифференциальный оператор называется *симметричным*, если для любых $u(x)$, $v(x)$ из допустимого класса функций, в частности, удовлетворяющих краевым условиям (1.2) и требуемым условиям гладкости, выполнены соотношения

$$(Lu,v)=(u,Lv), \tag{1.17}$$

$$(Lu,v) \stackrel{def}{=} \int_0^1 \left\{-\frac{d}{dx}\left(p(x)\frac{du(x)}{dx}\right)+q(x)u(x)\right\} v(x)\,dx. \tag{1.18}$$

Введем также обозначение, которое будем называть ***билинейной формой*** (т. е. *линейной* по u и v)

$$A(u,v) \stackrel{def}{=} \int_0^1 \left(p\frac{du}{dx}\frac{dv}{dx}+quv\right)dx. \tag{1.19}$$

Очевидно, что условие симметричности (1.17) для задачи (1.1), (1.2) выполнены (см. п. 1.1)

$$(Lu,v)=A(u,v), \quad (u,Lv)=A(v,u), \quad A(u,v)=A(v,u).$$

Аналогом свойства симметричности в конечномерном случае, естественно, является симметричность матрицы системы линейных уравнений (1.7). С учетом введенного обозначения (1.19) легко убедиться, что матрица (1.7) симметрична

$$A_{ik} = A(\varphi_i, \varphi_k), \quad A_{ki} = A(\varphi_k, \varphi_i), \quad A_{ik} = A_{ki}.$$

Таким образом, при построении приближенного решения такое важное свойство исходной задачи, как симметричность, сохраняется. Особенно подчеркнем, что *матрица A_{ik} будет симметрична при любом выборе базисных функций.*

По иному обстоит дело в случае конечно-разностного метода. Опуская довольно утомительные выкладки, приведем систему линейных уравнений для приближенного решения задачи (1.13) в случае, когда при аппроксимации не делается предположение об одинаковой длине отрезков $[x_k, x_{k-1}]$, $k = 1, \ldots, n-1$ (ср. с (1.14))

$$-\alpha_k u_{k-1} + 2u_k - \beta_k u_{k+1} = (x_k - x_{k-1})(x_{k+1} - x_k) f(x_k),$$
$$u_0 = 0, \quad u_n = 0, \quad \alpha_k = \frac{2(x_k - x_{k-1})}{(x_{k+1} - x_{k-1})}, \quad \beta_k = \frac{2(x_{k+1} - x_k)}{(x_{k+1} - x_{k-1})}.$$

Теперь соответствующая матрица коэффициентов не является симметричной (ср. с (1.15)), хотя, по-прежнему, остается трехдиагональной

$$\begin{pmatrix} 2 & -\beta_1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ -\alpha_2 & 2 & -\beta_2 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -\alpha_3 & 2 & -\beta_3 & \ldots & 0 & 0 & 0 \\ 0 & 0 & -\alpha_4 & 2 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & -\alpha_{n-2} & 2 & -\beta_{n-2} \\ 0 & 0 & 0 & 0 & \ldots & 0 & -\alpha_{n-1} & 2 \end{pmatrix}.$$

Приведенный пример показывает, что проекционный метод построения слабого решения является предпочтительнее конечно-разностного метода, т. к. он позволяет автоматически сохранять такое важное свойство задачи, как симметричность.

1.3.1 Финитные базисные функции

Оказывается, что существует эффективный способ устранения «главного недостатка» проекционных методов — сильной заполненности матрицы системы линейных уравнений. Специальный выбор базисных функций позволяет сделать матрицу A_{ik} сильно разреженной, а во многих случаях трехдиагональной, блочно-диагональной или ленточной. Более того, это можно сделать не только для одномерных задач, но и в случаях многомерных задач в области с достаточно сложной геометрической формой.

Иными словами, возможно сохранить все преимущества проекционных методов и конечно-разностных методов (разреженность матриц). Алгоритмы, позволяющие это осуществить, уместно называть ***проекционно-сеточными методами*** (проекционно-разностными, вариационно-разностными). Это другое название ***метода конечных элементов***, предложенное, в частности, в [3].

Напомним некоторые определения.

Определение 1.3.1. ***Носителем функции*** *называется замкнутая область, вне которой функция тождественно обращается в нуль. Для носителя функции* $\varphi(x)$ *используется обозначение:* $\operatorname{supp}\varphi$.

Определение 1.3.2. *Функция называется* ***финитной****, если ее носитель является ограниченной областью (более точно, компактной).*

В методе конечных элементов в качестве базисных функций предлагается выбирать финитные функции с размерами носителей, меньшими (как правило, существенно меньшими), чем размеры области, в которой рассматривается решаемая задача.

Для решения задачи (1.4) возможен следующий способ выбора базисных функций. Разобьем отрезок $[0,1]$ на интервалы $[x_{k-1}, x_k]$, $k = 1, \ldots, n$, $x_0 = 0$, $x_n = 1$ (длины интервалов не предполагаются одинаковыми, см. рис. 1.1).

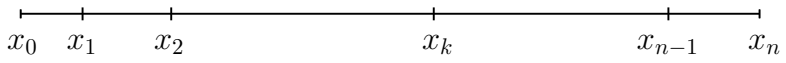

Рис. 1.1. Разбиение отрезка

В качестве $\varphi_k(x)$, $k = 1, \ldots, n-1$ выберем *финитные кусочно-линейные функции*, носителем которых будет отрезок $[x_{k-1}, x_{k+1}]$ (см. рис. 1.2)

$$\varphi_k(x) = \begin{cases} 0, & x < x_{k-1}, \\ \dfrac{x - x_{k-1}}{x_k - x_{k-1}}, & x_{k-1} < x < x_k, \\ \dfrac{x - x_{k+1}}{x_k - x_{k+1}}, & x_k < x < x_{k+1}, \\ 0, & x > x_{k+1}. \end{cases} \tag{1.20}$$

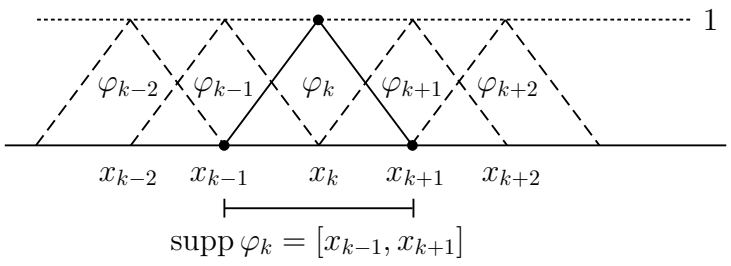

Рис. 1.2. Набор базисных функций $\varphi_k(x)$

Обычно носители базисных функций называют ***конечными элементами***. Часто тот же самый термин используется и для самих базисных функций.

Подчеркнем, что все функции $\varphi_k(x)$ $k = 1, \ldots, n-1$ удовлетворяют условиям (1.6).

Сделаем ряд важных замечаний.

1. Носители базисных функций полностью заполняют отрезок $[0, 1]$, т. е.

$$\bigcup_{k=1}^{n-1} \operatorname{supp} \varphi_k = [0, 1]. \tag{1.21}$$

Иными словами, конечные элементы полностью покрывают область, в которой разыскивается решение задачи. В противном случае, нашлась бы такая точка отрезка $[0, 1]$, в которой приближенное решение $u^h(x)$, задаваемое формулой (1.5), было бы неопределено.

2. Пересечения носителей различных базисных функций не являются пустыми, но таких пересечений достаточно мало. Более точно (очевидно, что при указании области изменения индексов предполагается, что они лежат в интервале от 1 до $n-1$)

$$\operatorname{supp} \varphi_k \bigcap \operatorname{supp} \varphi_i \neq \varnothing, \quad i = k-1, k, k+1, \tag{1.22}$$

$$\operatorname{supp} \varphi_k \bigcap \operatorname{supp} \varphi_i \neq \varnothing, \quad i = k-2, k+2, \tag{1.23}$$

$$\operatorname{mes}(\operatorname{supp} \varphi_k \bigcap \operatorname{supp} \varphi_i) = 0,$$

$$\operatorname{supp} \varphi_k \bigcap \operatorname{supp} \varphi_i = \varnothing, \quad i < k-2, \quad i > k+2. \tag{1.24}$$

Соотношения (1.22)–(1.24) являются наиболее важными для эффективности метода конечных элементов. В конечном итоге, именно они определяют структуру матрицы A_{ik}, задаваемой соотношениями (1.8). Благодаря (1.23), (1.24), большинство интегралов в (1.8) обращается в нуль и матрица A_{ik} становится сильно разреженной.

3. Справедливо соотношение

$$u^h(x_s) = \sum_{k=1}^{n-1} c_k \varphi_k(x_s) = c_s. \tag{1.25}$$

Таким образом, коэффициенты c_s — это значения приближенного решения в точках x_s. В частности, это означает, что при решении системы (1.7) приближенное решение $u^h(x)$ **в узлах сетки** будет получено сразу, без использования формулы (1.5). Напомним, что в случае конечно-разностного метода, приближенное решение в узлах сетки также получалось непосредственно после решения системы линейных уравнений (см. (1.14)).

4. Базисные функции $\varphi_k(x)$ выбраны кусочно-линейными из следующих соображений. Во-первых, кусочно-линейные функции допустимо использовать в задаче (1.4), т. к. производные $\varphi_k'(x)$ в этом случае кусочно-постоянны и интегрирование в (1.4) и последующих соотношениях (1.6)–(1.8) может быть выполнено.

Во-вторых, кусочно-линейные функции являются наиболее простейшими финитными кусочно-полиномиальными функциями, использование которых позволяет учесть все члены в интегралах (1.8). Нельзя выбирать в качестве $\varphi_k(x)$ кусочно-постоянные функции, т. к. в этом случае $\varphi_k'(x)$ обращаются в нуль и члены вида $\varphi_k'(x)\varphi_i'(x)$ будут отсутствовать в интегралах (1.8), что приведет к исчезновению части информации об исходной задаче. По этой же причине, нельзя использовать кусочно-линейные функции непосредственно в выражениях для вычисления матрицы $\tilde{A}_{ik}$, т. к. интегралы (1.11) содержат вторые производные базисных функций.

На самом деле, ситуация более сложная, т. к. при дифференцировании кусочно-постоянных функций будут возникать δ-функции Дирака и при определенных дополнительных предположениях кусочно-постоянные базисные функции все-таки могут быть использованы.

1.3.2 Вычисление элементов матрицы A_{ik} и вектора b_i

Уже на основании свойств носителей базисных функций (конечных элементов) (1.22)–(1.24) можно сделать вывод о том, что матрица A_{ik} будет трехдиагональной. Однако, с учетом явных выражений (1.20) для базисных функций полезно привести конкретные формулы для элементов матрицы A_{ik} и вектора b_i системы (1.7).

С учетом (1.24) ясно, что носитель функции φ_k не пересекается с носителями функций φ_i при $i > k+2$ и $i < k-2$ (см. также рис. 1.2). Это означает, что произведение $\varphi_i\varphi_k = 0$ при $i > k+2$ и $i < k-2$. Это же относится и к производным функции φ_i, заданной формулой (1.20). На самом деле, из (1.20) следует даже бóльшее — $\varphi_i\varphi_k = 0$ при $i \geqslant k+2$ и $i \leqslant k-2$.

Учитывая сказанное, запишем формулы для вычисления интегралов, входящих в (1.8). Заметим, что вместо интегрирования по всей области $[0,1]$ можно ограничиться интегрированием лишь по области, являющейся объединением носителей подинтегральных функций.

Имеем следующие соотношения (пересечение носителей функций либо пусто, либо имеет меру нуль)

$$\int_0^1 p\varphi_k'\varphi_i'\,dx = 0, \quad i \geqslant k+2, \quad i \leqslant k-2. \tag{1.26}$$

Далее, заменяем интегрирование по всей области интегрированием по пересечению носителей (см. рис. 1.2)

$$\int_0^1 p\varphi_k'\varphi_k'\,dx = \int_{x_{k-1}}^{x_{k+1}} p\varphi_k'\varphi_k'\,dx,$$

$$\int_0^1 p\varphi_k'\varphi_{k+1}'\,dx = \int_{x_k}^{x_{k+1}} p\varphi_k'\varphi_{k+1}'\,dx, \quad \int_0^1 p\varphi_k'\varphi_{k-1}'\,dx = \int_{x_{k-1}}^{x_k} p\varphi_k'\varphi_{k-1}'\,dx. \tag{1.27}$$

Аналогично для других членов, входящих в формулу (1.8)

$$\int_0^1 q\varphi_k\varphi_i\,dx = 0, \quad k-2 \leqslant i \leqslant k+2, \tag{1.28}$$

$$\int_0^1 q\varphi_k\varphi_k\,dx = \int_{x_{k-1}}^{x_{k+1}} q\varphi_k\varphi_k\,dx, \quad \int_0^1 q\varphi_k\varphi_{k+1}\,dx = \int_{x_k}^{x_{k+1}} q\varphi_k\varphi_{k+1}\,dx,$$

$$\int_0^1 q\varphi_k\varphi_{k-1}\,dx = \int_{x_{k-1}}^{x_k} q\varphi_k\varphi_{k-1}\,dx, \quad \int_0^1 f\varphi_i\,dx = \int_{x_{i-1}}^{x_{i+1}} f\varphi_i\,dx. \tag{1.29}$$

Таким образом, матрица A_{ik} является трехдиагональной матрицей.

Пример 1.1. В простых случаях интегралы (1.27), (1.29) легко вычисляются. Рассмотрим краевую задачу (см. (1.13))

$$-u''(x) = 1, \quad u(0) = 0, \quad u(1) = 0.$$

Сравнивая с (1.1), запишем

$$p = 1, \quad q = 0, \quad f(x) = 1.$$

Производя вычисление интегралов, получим следующие результаты (в окончательном ответе, для простоты, полагаем $x_k = kh$)

$$\int_{x_{k-1}}^{x_{k+1}} \varphi_k'\varphi_k'dx = \int_{x_{k-1}}^{x_k} \varphi_k'\varphi_k'dx + \int_{x_k}^{x_{k+1}} \varphi_k'\varphi_k'dx = \int_{x_{k-1}}^{x_k} \frac{1}{(x_k - x_{k-1})^2}dx + \int_{x_k}^{x_{k+1}} \frac{1}{(x_k - x_{k+1})^2}dx =$$

$$= \frac{x_k - x_{k-1}}{(x_k - x_{k-1})^2} + \frac{x_{k+1} - x_k}{(x_{k+1} - x_k)^2} = \frac{1}{x_k - x_{k-1}} + \frac{1}{x_{k+1} - x_k} = \frac{2}{h}.$$

$$\int_{x_k}^{x_{k+1}} \varphi_k'\varphi_{k+1}'dx = \int_{x_k}^{x_{k+1}} \frac{1}{x_k - x_{k+1}}\frac{1}{x_{k+1} - x_k}dx = \frac{x_{k+1} - x_k}{(x_k - x_{k+1})(x_{k+1} - x_k)} = \frac{1}{x_k - x_{k+1}} = -\frac{1}{h}.$$

$$\int_{x_{k-1}}^{x_k} \varphi_k'\varphi_{k-1}'dx = \int_{x_{k-1}}^{x_k} \frac{1}{x_k - x_{k-1}}\frac{1}{x_{k-1} - x_k}dx = \frac{x_k - x_{k-1}}{(x_k - x_{k-1})(x_{k-1} - x_k)} = \frac{1}{x_{k-1} - x_k} = -\frac{1}{h}.$$

$$\int_0^1 \varphi_i\,dx = \int_{x_{i-1}}^{x_{i+1}} \varphi_i\,dx = \int_{x_{i-1}}^{x_i} \varphi_i\,dx + \int_{x_i}^{x_{i+1}} \varphi_i\,dx.$$

$$\int_{x_{i-1}}^{x_i} \frac{x - x_{i-1}}{x_i - x_{i-1}}\,dx = \left| \begin{array}{c} x - x_{i-1} = s \\ dx = ds \end{array} \right| = \frac{1}{x_i - x_{i-1}} \int_0^{x_i - x_{i-1}} s\,ds = \frac{1}{x_i - x_{i-1}} \cdot \frac{s^2}{2}\Big|_0^{x_i - x_{i-1}} =$$

$$= \frac{(x_i - x_{i-1})^2}{2(x_i - x_{i-1})} = \frac{x_i - x_{i-1}}{2} = \frac{h}{2}.$$

Аналогично, имеем

$$\int_{x_i}^{x_{i+1}} \varphi_i \, dx = \frac{h}{2}.$$

Используя полученные соотношения, запишем систему (1.7)

$$\underbrace{A_{ii-1}}_{-1/h} c_{i-1} + \underbrace{A_{ii}}_{2/h} c_i + \underbrace{A_{ii+1}}_{-1/h} c_{i+1} = \underbrace{b_i}_{h}$$

или

$$-\frac{c_{i-1} - 2c_i + c_{i+1}}{h^2} = 1.$$

С учетом (1.25) запишем

$$-\frac{u^h(x_{i-1}) - 2u^h(x_i) + u^h(x_{i+1})}{h^2} = 1.$$

Сравнивая полученное выражение с (1.14), видим, что вновь получена обычная конечно-разностная схема (при $f(x) = 1$).

Конечно, такие простые соотношения получаются лишь для простых примеров. В более общем случае для вычисления интегралов (1.27), (1.29) следует использовать те или иные квадратурные формулы. От правильного выбора квадратурной формулы будет зависеть точность вычисления интеграла и, в конечном итоге, точность представления матрицы A_{ik}. Заметим, что FreeFem++ предлагает специальные возможности для использования различных квадратурных формул (см. гл. 18 и [1]).

1.4 Естественные и главные краевые условия

При использовании метода конечных элементов важное значение имеет правильный учет краевых условий рассматриваемой краевой задачи. В случае задачи (1.1), (1.2) при построении приближенного решения в виде (1.5) выполнение краевых условий (1.2) обеспечивалось выбором базисных функций. Действительно, для базисных функций требовалось выполнение краевых условий (1.6), которые и были соблюдены при задании φ_k соотношениями (1.20).

В общем случае для задач с краевыми условиями, отличными от (1.2), не всегда удается удовлетворить краевым условиям, выбирая подходящие базисные функции. Более точно, такой выбор может быть сопряжен с большими техническими трудностями — функции могут оказаться слишком сложными, матрица системы линейных уравнений (1.7) не будет сильно разреженной и пр.

Рассмотрим следующую краевую задачу, которая отличается от задачи (1.1), (1.2) (при $p = 1$, $q = 0$) лишь заданием краевого условия третьего рода при $x = 0$

$$-u''(x) = f(x), \qquad 0 < x < 1, \tag{1.30}$$

$$u'(0) - \alpha u(0) = 0, \qquad u(1) = 0, \tag{1.31}$$

где $f(x)$ — известная функция, α — заданное число.

Следуя процедуре, описанной в п. 1.1, умножим (1.30) на $v(x)$ и проинтегрируем от 0 до 1 с использованием интегрирования по частям

$$\int_0^1 \frac{du}{dx}\frac{dv}{dx}\,dx - \frac{du}{dx}v\Big|_0^1 = \int_0^1 fv\,dx. \tag{1.32}$$

Потребуем, чтобы $v(x)$ при $x = 1$ удовлетворяла тому же краевому условию, что и $u(x)$, т. е.

$$v(1) = 0. \tag{1.33}$$

Заметим, что нельзя требовать выполнения условия $v(0) = 0$, т. к. в этом случае (1.32) примет вид (1.4) (при $p = 1$, $q = 0$), что, очевидно, будет соответствовать решению задачи (1.30), (1.31) с краевым условием $u(0) = 0$ вместо краевого условия $u'(0) - \alpha u(0) = 0$.

Соотношение (1.32) с учетом (1.33) примет вид

$$\int_0^1 \left(\frac{du}{dx}\frac{dv}{dx} - fv\right)dx + u'(0)v(0) = 0, \quad v(1) = 0.$$

Используя краевое условие (1.31) при $x = 0$, получим (ср. с (1.4))

$$\int_0^1 \left(\frac{du}{dx}\frac{dv}{dx} - fv\right)dx + \alpha u(0)v(0) = 0, \quad v(1) = 0. \tag{1.34}$$

Функцию $u(x)$, являющуюся решением задачи (1.34), будем называть ***слабым решением*** задачи (1.30), (1.31).

Приближенное решение задачи (1.34) строим в виде, аналогичном (1.5), ***добавив к базисным функциям еще одну дополнительную функцию*** $\varphi_0(x)$, линейно не зависящую от прежних базисных функций

$$u^h(x) = \sum_{k=0}^{n-1} c_k\varphi_k(x). \tag{1.35}$$

Необходимость введения дополнительной базисной функции достаточно очевидна. Сохранив прежний набор базисных функций, невозможно удовлетворить краевым условиям при $x = 0$, т. к. в силу (1.6) $\varphi_k(0) = 0$, $\varphi_k(1) = 0$, $k = 1, \ldots, n-1$. Понятно также, что, выбирая $\varphi_0(x)$, следует потребовать выполнения условия $\varphi_0(1) = 0$. В противном случае нарушится выполнение краевого условия при $x = 1$ для приближенного решения, т. е. будет $u^h(1) \neq 0$. (Конечно, кроме тривиального случая, когда $c_0 = 0$, что соответствует просто прежнему набору базисных функций.)

Зададим дополнительную базисную функцию $\varphi_0(x)$ в виде *финитной кусочно-линейной функции* с носителем $\operatorname{supp}\varphi_0 = [x_0, x_1]$ (ср. с (1.20) и см. рис. 1.3 и 1.2)

$$\varphi_0(x) = \begin{cases} 1, & x = x_0, \\ \dfrac{x - x_1}{x_0 - x_1}, & x_0 < x < x_1, \\ 0, & x > x_1. \end{cases} \tag{1.36}$$

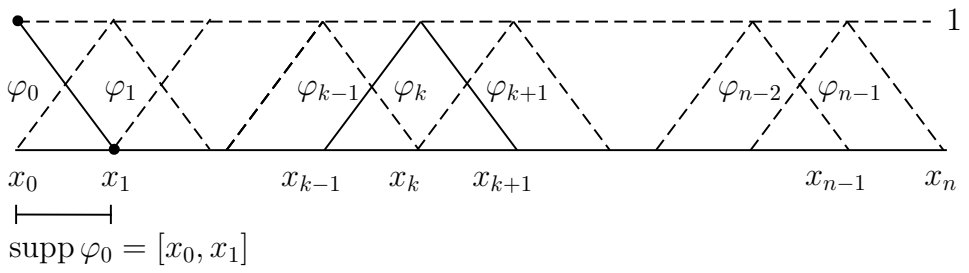

Рис. 1.3. Набор базисных функций $\varphi_k(x)$

Выбор $\varphi_0(x)$ в форме (1.36) продиктован следующими соображениями. Новая базисная функция не должна «сильно отличаться» от прежних, по-крайней мере, принадлежать тому же классу функций (в данном случае, быть кусочно-линейной). Желательно сохранение условия (1.25), т. е. $u^h(x_0) = c_0$, что приводит к условию $\varphi_0(0) = 1$. Желательно также сохранение условий, аналогичных (1.22)–(1.24):

$$\operatorname{supp}\varphi_0 \bigcap \operatorname{supp}\varphi_1 \neq \varnothing, \quad \operatorname{supp}\varphi_0 \bigcap \operatorname{supp}\varphi_i = \varnothing, \quad i > 2, \tag{1.37}$$

$$\operatorname{supp}\varphi_0 \bigcap \operatorname{supp}\varphi_2 \neq \varnothing, \quad \operatorname{mes}(\operatorname{supp}\varphi_0 \bigcap \operatorname{supp}\varphi_2) = 0.$$

Особенно подчеркнем, что функция $\varphi_0(x)$ не удовлетворяет краевому условию (1.31) при $x = 0$. Заманчиво было бы, конечно, выбрать $\varphi_0(x)$ в виде $\varphi_0(x) = 1 + \alpha x$, например, на отрезке $[x_0, x_1]$, и $\varphi_0(x) = 0$ — вне этого отрезка. Это привело бы к $\varphi_0'(0) - \alpha\varphi_0(0) \equiv 0$. Однако, при этом теряется некоторая «универсальность» базисных функций — φ_0 будет зависеть от параметра задачи α. Более того, детальный анализ показывает, что такой выбор существенно ухудшает процесс построения решения.

После того как выбор дополнительной базисной функции осуществлен, схема построения системы линейных уравнений для определения c_k будет такая же, как в п. 1.2.1.

Подставляя (1.35) в (1.34), получим ($k = 0, 1, \ldots, n-1$)

$$\int_0^1 \left(\sum_{k=0}^{n-1} c_k \varphi_k' v' - fv \right) dx + \alpha \sum_{k=0}^{n-1} c_k \varphi_k(0) v(0) = 0.$$

Выбирая в качестве $v(x)$ набор функций $\varphi_i(x)$, $i = 0, \ldots, n-1$, имеем

$$\int_0^1 \left(\sum_{k=0}^{n-1} c_k \varphi_k' \varphi_i' - f\varphi_k \right) dx + \alpha \sum_{k=0}^{n-1} c_k \varphi_k(0) \varphi_i(0) = 0, \quad i = 0, \ldots, n-1.$$

Для определения c_k имеем систему линейных уравнений

$$\sum_{k=0}^{n-1} B_{ik} c_k = b_i, \quad i = 0, \ldots, n-1, \tag{1.38}$$

где введены следующие обозначения

$$\int\limits_0^1 \varphi_i'\varphi_k' dx + \alpha\varphi_i(0)\varphi_k(0) = B_{ik}, \quad \int\limits_0^1 f\varphi_i\, dx = b_i, \quad i,k = 0,\ldots,n-1. \tag{1.39}$$

Нетрудно показать, что элементы матрицы B_{ik} для $i,k = 1,\ldots,n-1$ совпадают с элементами матрицы A_{ik} (при $p = 1$, $q = 0$), определяемыми формулами (1.8). Действительно, базисные функции $\varphi_1, \ldots, \varphi_{n-1}$ остались прежними и $\varphi_k(0) = 0$, $k = 1,\ldots,n-1$.

В силу (1.37) при вычислении интегралов в (1.39) получим (пересечение носителей соответствующих функций имеет нулевую меру)

$$B_{0k} = 0, \quad B_{k0} = 0, \quad k = 2,\ldots,n-1.$$

Используя вид функций φ_0 и φ_1 (см. (1.20) и (1.36)) в случае, когда длины всех отрезков $[x_k, x_{k+1}]$ одинаковы и равны h, получим

$$B_{00} = \frac{1+\alpha h}{h}, \quad B_{10} = B_{01} = -\frac{1}{h}.$$

Таким образом, матрица B_{ik} размерности $n \times n$ отличается от матрицы A_{ik} размерности $(n-1)\times(n-1)$ лишь левым столбцом и верхней строкой. Система уравнений (1.38), для удобства умноженная на h, в матричной форме имеет вид (выделена часть, соответствующая матрице A_{ik})

$$hBc = \left(\begin{array}{c|cccccc} 1+\alpha h & -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ \hline -1 & 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \ldots & \multicolumn{7}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\ 0 & 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{array}\right)\begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ \ldots \\ c_{n-2} \\ c_{n-1} \end{pmatrix} = \begin{pmatrix} hb_0 \\ hb_1 \\ hb_2 \\ hb_3 \\ \ldots \\ hb_{n-2} \\ hb_{n-1} \end{pmatrix}.$$

Интересно посмотреть, чему соответствует первое уравнение системы (1.38). При $i = 0$, с учетом $c_0 = u(x_0) = u(0)$, $c_1 = u(x_1) = u(x_0+h) = u(h)$, имеем

$$(1+\alpha h)u(0) - u'(h) = h\int\limits_0^h f(x)\left(\frac{h-x}{h}\right)dx = \int\limits_0^h f(x)(h-x)\,dx.$$

Раскладывая в ряд при $h \to 0$, выводим

$$(\alpha u(0) - u'(0))h - \frac{1}{2}u''(0)h^2 + \mathcal{O}(h^3) = \frac{1}{2}f(0)h^2 + \mathcal{O}(h^3).$$

Если предполагать выполнение уравнения (1.30) при $x = 0$, то имеем $u''(0) = -f(0)$, и получится, что краевое условие (1.31) при $x = 0$ выполняется с точностью до членов порядка $\mathcal{O}(h^2)$

$$u'(0) - \alpha u(0) = \mathcal{O}(h^2), \quad h \to 0.$$

Этот факт служит косвенным подтверждением правильности выбора базисной функции $\varphi_0(x)$.

Рассмотренные примеры показывают, что имеются различные возможности удовлетворения краевым условиям. В первом случае все краевые условия для задачи (1.1), (1.2) были выполнены *за счет выбора базисных функций* (см. формулу (1.20)). Во втором случае, для задачи (1.30), (1.31) краевое условие (1.31) при $x = 1$ по прежнему выполнено за счет выбора базисных функций. Краевое условие (1.31) при $x = 0$ учитывается «естественным образом». Имеется в виду следующее: в результате преобразования уравнений в (1.34) возникает дополнительный член $\alpha u(0)v(0)$. Затем к базисным функциям добавляется новая функция φ_0 (не удовлетворяющая краевому условию!) и задача приводится к решению системы линейных уравнений (1.38) с матрицей B_{ik}, включающей дополнительные, по сравнению с матрицей A_{ik}, члены $\alpha\varphi_i(0)\varphi_k(0)$.

Определение 1.4.1. *Краевые условия, которым можно удовлетворить за счет выбора базисных функций, называются* ***главными краевыми условиями.***

Определение 1.4.2. *Краевые условия, удовлетворение которых возможно за счет преобразования задачи, а не за счет выбора базисных функций, называются* ***естественными краевыми условиями.***

Таким образом, в задаче (1.1), (1.2) оба краевых условия (1.2) являются ***главными***. В задаче (1.30), (1.31) краевое условие (1.31) при $x = 1$ будет ***главным***, а краевое условие (1.31) при $x = 0$ является ***естественным***.

Понятие о главных и естественных краевых условиях является весьма важным для метода конечных элементов. В случае естественных краевых условий, в некотором смысле, можно не заботиться о выборе базисных функций. В случае же главных краевых условий приходится ставить дополнительные ограничения на выбор базисных функций. Заметим, что большое количество примеров задач с естественными и главными краевыми условиями имеется в [3], где подробно разъясняются всевозможные математические проблемы, возникающие в связи с численной реализацией различных граничных условий.

Скажем несколько слов о случае, когда для задачи (1.30), (1.31) параметр $\alpha = 0$ (краевое условие второго рода). Никаких проблем в данном случае в связи с численной реализацией не возникает — выражения (1.39) остаются справедливыми и при $\alpha = 0$. На первый взгляд может показаться, что задача (1.4) (при $p = 1$, $q = 0$) и задача (1.34) при $\alpha = 0$ одинаковы. На самом деле, это, конечно же, разные задачи. В случае (1.4) требуется выполнение двух условий: $v(0) = 0$ и $v(1) = 0$, тогда как для (1.34) нужно выполнение лишь одного условия $v(1) = 0$ и значение $v(0)$ может быть произвольным.