Федеральное агентство по образованию

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

С.И. Репин М.Е. Фролов

ЧИСЛЕННЫЕ МЕТОДЫ

ОЦЕНКА ПОГРЕШНОСТИ РЕШЕНИЯ

Лабораторный практикум

Санкт-Петербург
Издательство Политехнического университета
2006

Федеральное агентство по образованию

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

С.И. Репин *М.Е. Фролов*

ЧИСЛЕННЫЕ МЕТОДЫ

ОЦЕНКА ПОГРЕШНОСТИ РЕШЕНИЯ

Лабораторный практикум

Санкт-Петербург
Издательство Политехнического университета
2006

УДК 519.6

Репин С.И., Фролов М.Е. **Численные методы. Оценка погрешности решения:** Лаб. практикум. СПб.: Изд-во Политехн. ун-та, 2006. 36 с.

Практикум соответствует государственному образовательному стандарту дисциплины "Численные методы" направления подготовки бакалавров 010500 "Прикладная математика и информатика".

Рассмотрены методы оценки точности приближенных решений различных задач, возникающих при численном моделировании. Практикум включает в себя шесть лабораторных работ, в которых изложена математическая теория, дано конкретное задание, а также алгоритм его выполнения. Для того чтобы выполнить все работы, студент должен знать один из языков программирования высокого уровня и обладать основами работы в пакете MATLAB, являющемся программным продуктом компании The MathWorks, Inc.

Предназначен учащимся СПбГПУ и других университетов, а также преподавателям при составлении курсов лабораторных работ.

Табл. 1. Ил. 2. Библиогр.: 20 назв.

Печатается по решению редакционно-издательского совета Санкт-Петербургского государственного политехнического университета.

ВВЕДЕНИЕ

Проблема контроля точности приближенных решений различных задач вычислительной математики имеет большое прикладное значение. В первую очередь ее исследование связано с необходимостью гарантировать достоверность результатов, получаемых в процессе математического моделирования. Оценить качество той или иной математической модели можно, только сравнивая решение с результатами физических экспериментов. Однако в подавляющем большинстве случаев оказывается невозможно точно решить поставленную задачу. Тогда решение, соответствующее конкретной модели, находится приближенно путем ее дискретизации. Таким образом, помимо ошибки модели, возникает ошибка дискретизации. Более того, решение дискретной задачи также может содержать погрешности, связанные, например, с ошибками численного интегрирования и дифференцирования, использованием итерационных методов, накоплением вычислительной погрешности. В результате влияния всех этих факторов возникает ошибка приближенного решения дискретной задачи. Как следствие, становится невозможным объективно оценить ту или иную модель, сравнивая результат с физическим экспериментом, — ведь существенное различие может возникнуть не из-за неудачно выбранной модели, а из-за неудачной ее дискретизации или высокой неточности при получении приближенного решения. В связи с этим особое значение приобретают знания, дающие возможность оценить составляющие ошибки, связанные с дискретной задачей, и гарантировать их малость по сравнению с погрешностью модели. В результате становится возможным делать достоверные выводы о качестве выбранной математической модели и сравнивать ее с другими.

Лабораторный практикум дает общее представление об источниках возникновения погрешности вычислений и подходах к ее контролю. Содержащиеся в нем работы посвящены как простым, так и более сложным вопросам оценки точности решения различных задач. Издание включает в себя шесть лабораторных работ, в которых изложена математическая теория, дано конкретное задание, а также алгоритм его выполнения. При этом предполагается, что читатель владеет одним из языков программирования высокого уровня и основами работы в пакете MATLAB.

Р а б о т а 1

ОПРЕДЕЛЕНИЕ МАШИННОГО ε

Постановка задачи — исследование наиболее простых эффектов, связанных с особенностями представления вещественных чисел в ЭВМ и ограниченностью разрядной сетки машины. Оно заключается в определении наименьшего числа ε, такого, что

$$1 + \varepsilon > 1.$$

Данное число носит название *машинного* ε и отражает то наименьшее значение, для которого не происходит потерь точности при сложении с единицей (см., например, [1]). В каждом конкретном случае это значение зависит в том числе и от особенностей компилятора, используемых опций и выбранной точности представления чисел с плавающей точкой.

Цель работы. Написать простейшими средствами языка C, Pascal или Fortran программу, вычисляющую значение машинного ε.

Алгоритм выполнения задания.

1. Положим $\varepsilon = 1$, $i = \log_2 \varepsilon = 0$.
2. Если $1 + \varepsilon/2 > 1$, то $\varepsilon = \varepsilon/2$, $i = i - 1$, повторяем шаг 2, иначе, выводим значения ε и соответствующей степени числа два.

Дополнение к заданию. Проиллюстрируем влияние описанного выше эффекта на точность вычисления производных функции при помощи аппроксимации конечной разностью. Как известно, численное дифференцирование функций относится к числу задач, в которых с уменьшением погрешности метода повышается влияние погрешности в исходных данных и вычислительной погрешности (см., например, [2]). Рассмотрим две функции

$$f_1 = \varepsilon \sin \frac{x}{\varepsilon}, \quad f_2 = 1 + \varepsilon \sin \frac{x}{\varepsilon}.$$

Значения производной обеих функций в точке $x = 0$ равны единице. Вычислим эти значения приближенно при помощи конечно-разностной аппроксимации:

$$f'(x) \approx f'_h(x) = \frac{f(x+h) - f(x-h)}{2h}, \tag{1.1}$$

где $x = 0, h = \varepsilon/8$.

Формула (1.1) дает точный результат только для функции f_1, а для f_2 приближенное значение производной $f_h'(0)$ будет равно нулю. Следовательно, даже в простейших ситуациях, далеких от тех, которые реально могут возникнуть на практике, особенности представления чисел в ЭВМ могут критическим образом повлиять на качество полученных результатов.

Р а б о т а 2

ОЦЕНКА ПОГРЕШНОСТИ РЕШЕНИЯ СИСТЕМЫ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Постановка задачи — оценка точности решения систем линейных алгебраических уравнений через норму невязки системы на приближенном решении и через число обусловленности ее матрицы (см., например, [3]). Пусть необходимо решить систему линейных алгебраических уравнений с невырожденной матрицей $A \in \mathbb{M}^{n\times n}$ и правой частью $f \in \mathbb{R}^n$, а именно

$$Au = f,$$

где $u \in \mathbb{R}^n$ — точное решение.

Предположим, что мы получили приближенное решение U данной системы и хотим оценить его точность. Рассмотрев разность

$$Au - AU = f - AU,$$

получаем

$$u - U = A^{-1}(f - AU).$$

Величина $e = u - U$ представляет собой погрешность приближенного решения исследуемой задачи. Величина $r = f - AU$ называется невязкой системы на приближенном решении U. Невязка обращается в ноль тогда и только тогда, когда U совпадает с точным решением u, т.е. погрешность равна нулю. Для ошибки e имеет место неравенство

$$\|e\| \le \|A^{-1}\| \cdot \|r\|, \tag{2.1}$$

где выбор нормы вектора и подчиненной ей нормы матрицы не требует уточнения.

Например, рассмотрим *кубическую* норму вектора и подчиненную ей норму матрицы, которые определяются следующим образом:

$$\|u\| = \max_i |u_i|;$$
$$\|A\| = \max_i \sum_{j=1}^{n} |a_{ij}|.$$

Из оценки (2.1) следует, что норма невязки сама по себе не может контролировать величину погрешности. Хорошо известно, что при решении некоторых систем невязка на приближенном решении мала, а погрешность тем не менее велика.

Чтобы точнее охарактеризовать факторы, влияющие на погрешность приближенного решения, необходимо ввести специальную величину, которая называется *числом обусловленности* матрицы A. Рассмотрим две величины m и M, задаваемые соотношениями

$$m = \inf_{\substack{x\in\mathbb{R}^n,\\ x\neq 0}} \frac{\|Ax\|}{\|x\|}$$

и

$$m = \sup_{\substack{x\in\mathbb{R}^n,\\ x\neq 0}} \frac{\|Ax\|}{\|x\|}.$$

Назовем числом обусловленности матрицы A (обозначается $\operatorname{cond} A$) отношение этих двух величин, а именно

$$\operatorname{cond} A = \frac{M}{m}.$$

Покажем, как введенная характеристика связана с относительной точностью приближенного решения U. По определению величин m и M имеем

$$m \le \frac{\|Ae\|}{\|e\|} \le M, \quad \text{т.е.} \quad m \le \frac{\|r\|}{\|e\|} \le M,$$

и

$$m \le \frac{\|Au\|}{\|u\|} \le M, \quad \text{т.е.} \quad m \le \frac{\|f\|}{\|u\|} \le M.$$

Отсюда получаем двусторонние неравенства

$$m\|e\| \le \|r\| \le M\|e\|$$

и

$$\frac{m}{\|f\|} \le \frac{1}{\|u\|} \le \frac{M}{\|f\|}.$$

Введем величину относительной погрешности $\tilde{e}$, определяемую соотношением

$$\tilde{e} = \frac{\|e\|}{\|u\|}.$$

Из неравенств

$$\|e\| \le \frac{\|r\|}{m}$$

и

$$\frac{1}{\|u\|} \le \frac{M}{\|f\|},$$

полученных ранее, имеем оценку

$$\tilde{e} \le \frac{M}{m} \frac{\|r\|}{\|f\|} = \operatorname{cond} A \frac{\|r\|}{\|f\|},$$

а из неравенств

$$\|e\| \ge \frac{\|r\|}{M}$$

и

$$\frac{1}{\|u\|} \ge \frac{m}{\|f\|}$$

обратную оценку:

$$\tilde{e} \ge \frac{m}{M} \frac{\|r\|}{\|f\|} = (\operatorname{cond} A)^{-1} \frac{\|r\|}{\|f\|}.$$

Таким образом,

$$(\operatorname{cond} A)^{-1} \tilde{r} \le \tilde{e} \le \operatorname{cond} A \, \tilde{r}, \tag{2.2}$$

где

$$\tilde{r} = \frac{\|r\|}{\|f\|}.$$

Из оценки (2.2) следует, что относительная точность приближенного решения контролируется отношением нормы невязки к норме правой части. Однако границы данной двусторонней оценки расходятся в том случае, если мы имеем дело с плохо обусловленной матрицей системы.

Аналогичную оценку можно получить, если предположить наличие погрешности в исходных данных задачи, а именно в правой части системы. При этом вместо исходной системы $Au = f$ решается система $AU = F$. Тогда справедлива оценка

$$\tilde{e} = \frac{\|u - U\|}{\|u\|} \le \operatorname{cond} A \frac{\|f - F\|}{\|f\|}, \tag{2.3}$$

отражающая зависимость отклонения решения от изменения исходных данных.

Цель работы. Решить две системы (с хорошо и плохо обусловленными матрицами) и выяснить влияние обусловленности матрицы системы на зависимость точности приближенных решений от точности определения исходных данных задачи.

Алгоритм выполнения задания. Рассмотрим системы с *известным точным решением* u. Матрицы системы A_1 и A_2 выбираются произвольными, но должны удовлетворять условиям

$$\operatorname{cond} A_1 < 10;$$
$$\frac{\operatorname{cond} A_2}{\operatorname{cond} A_1} > 1000.$$

Соответствующие правые части f_1 и f_2 вычисляют по заданным u, A_1 и A_2.

Лабораторная работа выполняется средствами пакета MATLAB и заключается в проверке оценки погрешности (2.3) для хорошо обусловленной (A_1) и плохо обусловленной (A_2) матрицы системы (в качестве учебного пособия для изучения возможностей пакета можно использовать, например, [4]). Сначала решаются системы

$$A_1 u = f_1$$

и

$$A_1 U = F_1,$$

где вектор F_1 получен незначительным (не более 1 %) возмущением коэффициентов вектора f_1.

Далее вычисляются все величины, входящие в оценку (2.3), и проверяется ее выполнение. Система с плохо обусловленной матрицей исследуется аналогичным образом.

Пример выполнения задания. Покажем последовательность выполнения данной лабораторной работы и кратко опишем необходимые для этого команды пакета MATLAB.

1. Ввод матрицы системы:

$$A_1 = \begin{bmatrix} +1.351 & +0.201 & -0.623 & +0.440 \\ +0.201 & +1.351 & +0.123 & +0.985 \\ -0.623 & +0.211 & +1.351 & -1.002 \\ +0.440 & +0.334 & -0.154 & +1.351 \end{bmatrix}.$$

Команда:
A1 = [1.351, 0.201, –0.623, 0.440; ...
0.201, 1.351, 0.123, 0.985; ...

–0.623, 0.211, 1.351, –1.002; ...
0.440, 0.334, –0.154, 1.351].

2. Ввод точного решения $u = \left[+1, -2, +3, -4\right]^{\mathrm{T}}$.
Команда:
u = [+1; −2; +3; −4]′.
3. Вычисление числа обусловленности матрицы A_1. Команда:
cond(A1)
Ответ: 5.4949.
4. Вычисление правой части f_1. Команда:
f1= A1*u
5. Обращение матрицы системы $B_1 = A_1^{-1}$ и вычисление приближенного решения $u_1 = B_1 f_1$. Команды:
B1= inv(A1);
u1= B1*f1
6. Возмущение правой части системы. Команды:
F1(1) = 1.01*f1(1); F1(3) = 1.01*f1(3);
F1(2) = 0.99*f1(1); F1(4) = 0.99*f1(4);
F1 = F1′
7. Вычисление решения U_1. Вычисление ошибки e_1. Команды:
U1= B1*F1
e1= u1 – U1
8. Вычисление относительной погрешности и относительного возмущения правой части. Команды:
norm(e1)/norm(u1)
norm(f1-F1)/norm(F1)
9. Проверка выполнения неравенства (2.3).
10. Ввод матрицы A_2 и повторение шагов 3–9, где

$$A_2 = \begin{bmatrix} +1.000 & +0.500 & +0.333 & +0.250 \\ +0.500 & +0.333 & +0.250 & +0.200 \\ +0.333 & +0.250 & +0.200 & +0.167 \\ +0.250 & +0.200 & +0.167 & +0.143 \end{bmatrix}.$$

Р а б о т а 3

АПРИОРНАЯ И АПОСТЕРИОРНАЯ ОЦЕНКА ТОЧНОСТИ РЕШЕНИЯ СИСТЕМЫ НЕЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Постановка задачи. Пусть X — полное нормированное (банахово) пространство, а $A : X \to X$ — непрерывный нелинейный оператор.

Рассмотрим задачу на поиск элемента $u \in X$, удовлетворяющего соотношению

$$Au = u. \tag{3.1}$$

Элемент u называется *неподвижной точкой* оператора A. Построим последовательность элементов пространства X, определяемых формулой

$$u_{i+1} = Au_i, \quad i = 0, 1, 2, ..., \tag{3.2}$$

где $u_0 \in X$ — произвольное начальное приближение.

Далее исследуем вопрос об условиях сходимости полученной при помощи данного итерационного процесса последовательности к неподвижной точке оператора.

Пусть (X, d) — метрическое пространство, порожденное метрикой d. Метрикой называется однозначная неотрицательная вещественная функция, подчиненная следующим трем аксиомам, выполняющимся для любых элементов $x_1, x_2 \in X$:

1) $d(x_1, x_2) = 0 \Leftrightarrow x_1 = x_2$;
2) $d(x_1, x_2) = d(x_2, x_1)$;
3) $d(x_1, x_2) \leq d(x_1, z) + d(z, x_2), \quad \forall z \in X$.

Нормированное пространство автоматически является метрическим.

Введем еще одно определение: оператор A называется *сжимающим*, если для любых $x_1, x_2 \in X$ выполняется оценка

$$d(Ax_1, Ax_2) \leq K\, d(x_1, x_2), \quad 0 < K < 1,$$

где число K — *показатель сжатия.*

Примеры применения принципа сжимающих отображений представлены, в частности, в книге [5].

Сформулируем и докажем теорему Банаха о неподвижной точке.

Теорема. *Пусть X — банахово пространство, A является сжимающим оператором с показателем K, тогда неподвижная точка оператора A существует и единственна. Для любого начального приближения $u_0 \in X$, определяемая формулой* (3.2) *последовательность $\{u_i\}_{i=0}^{+\infty}$ сходится к неподвижной точке в пространстве X, а также имеют место следующие оценки погрешности:*

$$d(u_i, u) \leq \frac{K^i}{1-K}\, d(u_0, u_1) \quad \textit{(априорная оценка)} \tag{3.3}$$

и

$$d(u_i, u) \leq \frac{K}{1-K}\, d(u_{i-1}, u_i) \quad \textit{(апостериорная оценка).} \tag{3.4}$$

Доказательство.

Из определяющего соотношения (3.2) получаем

$$d(u_i, u_{i+1}) = d(Au_{i-1}, Au_i) \le Kd(u_{i-1}, u_i) =$$
$$= Kd(Au_{i-2}, Au_{i-1}) \le ... \le K^i d(u_0, u_1).$$

Тогда

$$d(u_i, u_{i+m}) \le d(u_i, u_{i+1}) + ... + d(u_{i+m-1}, u_{i+m}) \le$$
$$\le (K^i + K^{i+1} + ... + K^{i+m-1})d(u_0, u_1) = K^i \tfrac{1-K^m}{1-K} d(u_0, u_1) \le$$
$$\le \tfrac{K^i}{1-K} d(u_0, u_1),$$

или

$$d(u_i, u_{i+m}) \le \frac{K^i}{1-K} d(u_0, u_1). \tag{3.5}$$

Правая часть неравенства (3.5) не зависит от m. Для любого фиксированного m при $i \to +\infty$ получаем $d(u_i, u_{i+m}) \to 0$. Таким образом, последовательность $\{u_i\}_{i=0}^{+\infty}$ фундаментальна, а следовательно, сходится к некоторому пределу $v \in X$.

Покажем, что v — неподвижная точка оператора A. Действительно, в силу непрерывности оператора A и сходимости $\{u_i\}_{i=0}^{+\infty}$ к v

$$d(Au_i, Av) \to 0, \quad d(u_{i+1}, v) \to 0 \quad \text{при} \quad i \to +\infty.$$

Следовательно, по формуле (3.2), поскольку v — фиксированный элемент пространства X, $Av = v$.

Докажем единственность неподвижной точки. Предположим существование двух таких точек — v и u, где $d(u, v) > 0$. По определению неподвижной точки и по свойствам оператора A

$$d(u, v) = d(Au, Av) \le Kd(u, v), \quad \text{где} \quad K < 1.$$

Следовательно, $d(u, v) = 0$.

Устремляя в неравенстве (3.5) $m \to +\infty$, приходим к априорной оценке (3.3). Апостериорная оценка (3.4) получается из следующего неравенства:

$$d(u_i, u_{i+m}) \le d(u_i, u_{i+1}) + ... + d(u_{i+m-1}, u_{i+m}) \le$$
$$(K + K^2 + ... + K^m)d(u_{i-1}, u_i) \le \tfrac{K}{1-K} d(u_{i-1}, u_i),$$

т.е.

$$d(u_i, u_{i+m}) \le \frac{K}{1-K} d(u_{i-1}, u_i).$$

□

Отметим существенное отличие оценки (3.3) от (3.4). Для того чтобы при помощи (3.3) сделать прогноз о числе итерации, которое потребуется для достижения необходимой точности, необходимо знать только два первых приближения u_0 и u_1. Естественно, такой прогноз в большинстве случаев оказывается пессимистичным. С помощью неравенства (3.4) можно уточнять оценку после каждой итерации, что на практике более эффективно. Если показатель сжатия K близок к единице, то (3.3) может дать оценку числа необходимых итераций, в сотни раз превышающую ту, которая получается при использовании неравенства (3.4). Эта односторонняя оценка контролирует величину погрешности только сверху. Двусторонние апостериорные оценки для итерационных методов исследованы, в частности, в монографии [6]. В книге также представлен обширный список литературы.

Цель работы. Сравнить эффективность априорной и апостериорной оценок погрешности решения системы нелинейных алгебраических уравнений, полученного методом простых итераций.

Алгоритм выполнения задания.

1. Рассмотрим систему уравнений (3.1), для которой известно точное решение $u = [0, 0]^{\mathrm{T}}$, а оператор A задается соотношением

$$Ax = \begin{bmatrix} 0.3\left(x_1^2\sqrt{x_1^2+x_2^2} + x_2^2\ln(x_1+\sqrt{x_1^2+x_2^2})\right) \\ 0.1\left(x_2^2\sqrt{x_1^2+x_2^2} + x_1^2\ln(x_2+\sqrt{x_1^2+x_2^2})\right) \end{bmatrix},$$

   где $x = [x_1, x_2]^{\mathrm{T}}$.
2. Выбираем желаемую точность приближенного решения и вектор $u_0 = [1, 1]^{\mathrm{T}}$ в качестве начального приближения.
3. При помощи априорной оценки вычисляем прогнозируемое количество итераций, которое потребуется для достижения точности.
4. Производим дальнейшие вычисления до тех пор, пока апостериорная оценка не покажет, что требуемая точность достигнута.
5. На каждом шаге контролируем реальную величину погрешности и находим тот, на котором на самом деле была достигнута необходимая точность.
6. Сравниваем числа, полученные на этапах 3–5, и делаем соответствующие выводы.

Р а б о т а 4

ДВУСТОРОННИЕ МЕТОДЫ ДЛЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Постановка задачи. Рассмотрим задачу

$$u = Au + f, \tag{4.1}$$

где матрица $A \in \mathbb{M}^{n\times n}$ удовлетворяет условиям, гарантирующим сходимость итерационного процесса:

$$u_{i+1} = Au_i + f. \tag{4.2}$$

Вначале дополнительно потребуем, чтобы все коэффициенты матрицы A были неотрицательны, т.е.

$$a_{jk} \geq 0, \quad \forall j, k = 1, ..., n.$$

Из формул (4.1) и (4.2) находим представление для погрешности

$$u_{i+1} - u = A(u_i - u).$$

Пусть выбраны два вектора u_0^- и u_0^+, для которых покомпонентно выполняется двусторонняя оценка

$$u_0^- \leq u \leq u_0^+.$$

Последующие приближения будем вычислять по формулам

$$u_{i+1}^- = Au_i^- + f;$$
$$u_{i+1}^+ = Au_i^+ + f.$$

Для всех членов последовательностей, построенных таким образом, двусторонняя оценка

$$u_i^- \leq u \leq u_i^+ \tag{4.3}$$

остается справедливой. Действительно, в силу неотрицательности коэффициентов матрицы

$$u_{i+1}^- - u = Au_i^- + f - Au - f = A(u_i^- - u) \leq 0;$$
$$u_{i+1}^+ - u = Au_i^+ + f - Au - f = A(u_i^+ - u) \geq 0.$$

Таким образом, по индукции двусторонняя оценка справедлива для любого i.

Описанная выше итерационная процедура позволяет получить двусторонние оценки точного решения, выполняющиеся покомпонентно. При этом, если $\|A\| = K < 1$, то границы будут сходиться друг к другу, что гарантирует нахождение решения с заданной точностью. Более общее

описание изложенного выше подхода (не только при применении к системам линейных алгебраических уравнений) представлено в книге [7].

Рассмотрим случай, когда коэффициенты матрицы A произвольны. Представим матрицу в виде

$$A = A^+ - A^-,$$

где коэффициенты A^+ и A^- неотрицательны.

Выбирая u_0^- и u_0^+ так, чтобы они покомпонентно оценивали точное решение снизу и сверху, будем вычислять последующие приближения по формулам

$$u_{i+1}^- = A^+ u_i^- - A^- u_i^+ + f; \tag{4.4}$$

$$u_{i+1}^+ = A^+ u_i^+ - A^- u_i^- + f. \tag{4.5}$$

Для последовательностей, построенных таким образом

$$u_{i+1}^- - u = A^+ u_i^- - A^- u_i^+ - Au = A^+(u_i^- - u) - A^-(u_i^+ - u) \le 0;$$
$$u_{i+1}^+ - u = A^+ u_i^+ - A^- u_i^- - Au = A^+(u_i^+ - u) - A^-(u_i^- - u) \ge 0.$$

Следовательно, аналогично предыдущему случаю двусторонняя оценка (4.3) снова справедлива для любого i.

Цель работы. Рассмотрев случаи, когда норма матрицы A намного меньше единицы и когда она близка к единице, изучить сходимость двусторонних оценок точного решения друг к другу.

Алгоритм выполнения задания. Исследуем две системы с *известным точным решением* u. Матрицы A_1 и A_2 выбираем так, что

$$\|A_1\| = 0.5, \quad \|A_2\| = 0.99,$$

где $\|.\|$ — кубическая норма матрицы (см. работу 2).

Правые части соответствующих задач вычисляются по заданному точному решению, а именно

$$f_1 = u - A_1 u$$

и

$$f_2 = u - A_2 u.$$

Далее, представляя матрицы в виде

$$A_1 = A_1^+ - A_1^-; \quad A_2 = A_2^+ - A_2^-$$

и выбирая начальные приближения

$$u_0^- = u - 1; \quad u_0^+ = u + 1,$$

необходимо вычислить несколько последующих приближений по формулам (4.4), (4.5). Условие остановки итерационного процесса на i-м шаге: значение нормы разности $u_i^+ - u_i^-$ становится меньше некоторого ε. Отметим, что при выполнении работы должна наблюдаться следующая закономерность: когда норма матрицы системы близка к единице, достижение необходимой точности требует существенно большего числа итераций.

Пример выполнения задания. Опишем реализацию лабораторной работы при помощи пакета MATLAB (необходимые для этого команды описаны в лабораторной работе 2).

1. Ввод точного решения
$$u = \left[+1, -2, +3, -4\right]^{\mathrm{T}}.$$
2. Определение матриц A_1^- и A_1^+ путем выделения отрицательных и положительных элементов матрицы A_1, где
$$A_1 = \begin{bmatrix} +0.13 & -0.05 & -0.21 & +0.08 \\ -0.35 & -0.11 & +0.01 & +0.03 \\ +0.00 & -0.06 & +0.32 & +0.10 \\ -0.02 & +0.41 & -0.05 & +0.01 \end{bmatrix},$$
$$A_1^- = \begin{bmatrix} 0.00 & 0.05 & 0.21 & 0.00 \\ 0.35 & 0.11 & 0.00 & 0.00 \\ 0.00 & 0.06 & 0.00 & 0.00 \\ 0.02 & 0.00 & 0.05 & 0.00 \end{bmatrix}$$
и
$$A_1^+ = \begin{bmatrix} 0.13 & 0.00 & 0.00 & 0.08 \\ 0.00 & 0.00 & 0.01 & 0.03 \\ 0.00 & 0.00 & 0.32 & 0.10 \\ 0.00 & 0.41 & 0.00 & 0.01 \end{bmatrix}.$$
Команды:
A1 = [...]
A1m = (abs(A) – A)/2
A1p = (abs(A) + A)/2
3. Вычисление кубической нормы матрицы A_1. Команда: norm(A1,inf)
4. Вычисление правой части
$$f_1 = \left[+1.72, -1.78, +2.32, -2.97\right]^{\mathrm{T}}.$$
Команда:
f1 = u – A1*u

5. Ввод требуемой точности ε. Команда:
   eps= 0.001;
6. Ввод начальных приближений $u_0^- = u - 1$ и $u_0^+ = u + 1$. Команды:
   um = u – 1
   up = u + 1
7. Вычисление двусторонних оценок точного решения по формулам (4.4), (4.5) с использованием оператора *while*. Команды:
   i = 0;
   while norm(up – um,inf) > eps
   i = i+1
   tmp = A1p*um – A1m*up + f1;
   up = A1p*up – A1m*um + f1;
   um = tmp;
   um′
   up′
   dist = norm(up – um,inf)
   pause(2)
   end
8. Ввод матрицы A_2 и повторение шагов 2–7, где

$$A_2 = \begin{bmatrix} +0.33 & -0.05 & -0.21 & +0.37 \\ -0.35 & -0.31 & +0.30 & +0.03 \\ +0.00 & -0.35 & +0.52 & +0.10 \\ -0.31 & +0.41 & -0.05 & +0.21 \end{bmatrix}.$$

Р а б о т а 5

ДВУСТОРОННИЕ МЕТОДЫ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Постановка задачи — построение на основе метода Рунге-Кутты двусторонних оценок решения задачи Коши. Эта задача возникает при анализе обыкновенного дифференциального уравнения на отрезке $[a, b]$, дополненного начальным условием в точке $x = a$, и имеет вид: найти функцию $y(x)$, удовлетворяющую соотношениям

$$\begin{cases} y'(x) &= f(x, y); \\ y(a) &= y_0. \end{cases}$$

В большинстве случаев такая задача не может быть решена аналитически. При ее дискретизации возникает проблема вычисления значений

функции y на множестве узлов $\{x_i\}_{i=0}^n$, где $x_0 = a$, $x_n = b$ и $x_i < x_{i+1}$. Заметим, что распределение узлов может быть как равномерным, так и неравномерным, т.е. учитывающим конкретные особенности рассматриваемой задачи. Имея точное значение функции y в точке x_0, для получения вычислительной схемы необходимо указать способ нахождения значений искомой функции в последующих точках сетки по уже имеющимся. В работе рассматриваются схемы, относящиеся к группе методов Рунге–Кутты. Они основаны на использовании для перехода от узла x_i к узлу x_{i+1} определенного количества слагаемых из разложения функции y в ряд Тейлора

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2}y''(x) + \frac{h^3}{6}y'''(x) + ..., \tag{5.1}$$

где $x = x_i$, $h = x_{i+1} - x_i$.

Простейшим методом такого типа можно считать также и хорошо известный метод Эйлера, вычислительная схема которого выглядит следующим образом:

$$y(x+h) \approx y(x) + hf(x,y).$$

Более точное представление (5.1) используется для построения методов третьего порядка, так как мы пренебрегаем величинами более высоких порядков.

Введем обозначение

$$\Delta y = hy'(x) + \frac{h^2}{2}y''(x) + \frac{h^3}{6}y'''(x)$$

и получим вычислительную схему

$$y(x+h) \approx y(x) + \Delta y.$$

Выразим производные функции y через функцию f и ее частные производные:

$$y' = f;$$
$$y'' = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}y' = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}f$$

и

$$y''' = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial x \partial y}f + \frac{\partial^2 f}{\partial x \partial y}f + \frac{\partial^2 f}{\partial y^2}f^2 + \frac{\partial f}{\partial y}\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}f\right) =$$

$$= \frac{\partial^2 f}{\partial x^2} + 2\frac{\partial^2 f}{\partial x \partial y}f + \frac{\partial^2 f}{\partial y^2}f^2 + \left(\frac{\partial f}{\partial y}\right)^2 f + \frac{\partial f}{\partial x}\frac{\partial f}{\partial y}.$$

Таким образом,

$$\Delta y = hf + \frac{h^2}{2}\frac{\partial f}{\partial x} + \frac{h^2}{2}\frac{\partial f}{\partial y}f + \frac{h^3}{6}\frac{\partial^2 f}{\partial x^2} +$$
$$+\frac{h^3}{3}\frac{\partial^2 f}{\partial x \partial y}f + \frac{h^3}{6}\frac{\partial^2 f}{\partial y^2}f^2 + \frac{h^3}{6}\left(\frac{\partial f}{\partial y}\right)^2 f + \frac{h^3}{6}\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}. \qquad (5.2)$$

Дальнейшее построение заключается в аппроксимации функции Δy при помощи линейной комбинации δy значений функции f в некотором наборе точек

$$\delta y = h\left(p_1 K_1 + p_2 K_2 + p_3 K_3\right), \qquad (5.3)$$

где p_1, p_2, p_3 — весовые множители;

$$K_1 = f(x, y);$$
$$K_2 = f(x + h\alpha_2, y + h\beta_{21} K_1);$$
$$K_3 = f(x + h\alpha_3, y + h(\beta_{31} K_1 + \beta_{32} K_2)),$$

$\alpha_2, \alpha_3, \beta_{21}, \beta_{31}, \beta_{32}$ — свободные параметры.

Множители p_1, p_2, p_3 и параметры $\alpha_2, \alpha_3, \beta_{21}, \beta_{31}, \beta_{32}$ подбираются таким образом, чтобы функция δy аппроксимировала функцию Δy с точностью до слагаемых порядка h^3 включительно.

Представим значения K_2 и K_3 в виде разложений в ряд, сохранив только необходимое количество членов. Величины большего порядка малости снова обозначим знаком "...":

$$K_2 = f + h\alpha_2\frac{\partial f}{\partial x} + h\beta_{21}K_1\frac{\partial f}{\partial y} + \frac{h^2}{2}\alpha_2^2\frac{\partial^2 f}{\partial x^2} + h^2\alpha_2\beta_{21}K_1\frac{\partial^2 f}{\partial x \partial y} +$$
$$+\frac{h^2}{2}\beta_{21}^2 K_1^2\frac{\partial^2 f}{\partial y^2} + ... = f + h\alpha_2\frac{\partial f}{\partial x} + h\beta_{21}\frac{\partial f}{\partial y}f +$$
$$+\frac{h^2}{2}\alpha_2^2\frac{\partial^2 f}{\partial x^2} + h^2\alpha_2\beta_{21}\frac{\partial^2 f}{\partial x \partial y}f + \frac{h^2}{2}\beta_{21}^2\frac{\partial^2 f}{\partial y^2}f^2 + ... \quad ; \qquad (5.4)$$

$$K_3 = f + h\alpha_3\frac{\partial f}{\partial x} + h(\beta_{31}K_1 + \beta_{32}K_2)\frac{\partial f}{\partial y} + \frac{h^2}{2}\alpha_3^2\frac{\partial^2 f}{\partial x^2} +$$
$$+h^2\alpha_3(\beta_{31}K_1 + \beta_{32}K_2)\frac{\partial^2 f}{\partial x \partial y} + \frac{h^2}{2}(\beta_{31}K_1 + \beta_{32}K_2)^2\frac{\partial^2 f}{\partial y^2} + ... =$$
$$= f + h\alpha_3\frac{\partial f}{\partial x} + h\beta_{31}\frac{\partial f}{\partial y}f + h\beta_{32}\frac{\partial f}{\partial y}\left(f + h\alpha_2\frac{\partial f}{\partial x} + h\beta_{21}\frac{\partial f}{\partial y}f + ...\right) +$$
$$+\frac{h^2}{2}\alpha_3^2\frac{\partial^2 f}{\partial x^2} + h^2\alpha_3\beta_{31}\frac{\partial^2 f}{\partial x \partial y}f + h^2\alpha_3\beta_{32}\frac{\partial^2 f}{\partial x \partial y}f + \frac{h^2}{2}(\beta_{31} + \beta_{32})^2\frac{\partial^2 f}{\partial y^2}f^2 + ... \qquad (5.5)$$

Заметим, что все значения функций в разложениях (5.4) и (5.5) вычисляются в точке (x, y). Поэтому мы опускаем эти аргументы. С учетом общего множителя h необходимо сохранить в представлениях K_2 и K_3 только те слагаемые, порядок которых по h не превышает второго. Далее, сравнивая представления (5.2) и (5.3), находим условия, связывающие значения весовых множителей и свободных параметров между собой (таблица)

Условие	Слагаемое	Коэффициенты	
		Формула (5.3)	Формула (5.2)
1	hf	$p_1 + p_2 + p_3$	1
2	$h^2 \frac{\partial f}{\partial x}$	$p_2\alpha_2 + p_3\alpha_3$	1/2
3	$h^2 \frac{\partial f}{\partial y} f$	$p_2\beta_{21} + p_3(\beta_{31} + \beta_{32})$	1/2
4	$h^3 \frac{\partial^2 f}{\partial x^2}$	$(p_2\alpha_2^2 + p_3\alpha_3^2)/2$	1/6
5	$h^3 \frac{\partial^2 f}{\partial x \partial y} f$	$p_2\alpha_2\beta_{21} + p_3\alpha_3(\beta_{31} + \beta_{32})$	1/3
6	$h^3 \frac{\partial^2 f}{\partial y^2} f^2$	$(p_2\beta_{21}^2 + p_3(\beta_{31} + \beta_{32})^2)/2$	1/6
7	$h^3 (\frac{\partial f}{\partial y})^2 f$	$p_3\beta_{32}\beta_{21}$	1/6
8	$h^3 \frac{\partial f}{\partial x} \frac{\partial f}{\partial y}$	$p_3\beta_{32}\alpha_2$	1/6

Из условий 7 и 8, 4 и 5 получаем соотношения $\beta_{21} = \alpha_2$ и $\alpha_3 = \beta_{31} + \beta_{32}$ соответственно. Условие 2 оказалось эквивалентно условию 3, а условие 4 — условию 6. Таким образом, можно оставить по одному соотношению из двух и получить следующую систему ограничений:

1a) $p_1 + p_2 + p_3 = 1$;
2a) $p_2\alpha_2 + p_3\alpha_3 = 1/2$;
3a) $\alpha_3 = \beta_{31} + \beta_{32}$;
4a) $\beta_{21} = \alpha_2$;
5a) $p_2\alpha_2^2 + p_3\alpha_3^2 = 1/3$;
6a) $p_3\beta_{32}\alpha_2 = 1/6$.

Из всех возможных вариантов выберем две схемы так, чтобы условия 1a–4a и 6a выполнялись, а условие 5a нарушалось. При этом будут нарушены исходные условия 4 и 6, которые эквивалентны друг другу. Тогда получим представление

$$\Delta y - \delta y = \xi h^3 \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} f^2 \right) + O(h^4);$$

$$\xi = \frac{1}{6} - \frac{p_2\alpha_2^2 + p_3\alpha_3^2}{2},$$

где $O(h^4)$ — величина четвертого порядка малости.

Построение двусторонних оценок точного решения задачи Коши заключается в таком подборе параметров, при котором множитель ξ равняется $-c$ в одном случае и $+c$ — в другом, где c — некоторая постоянная. Тогда для первой схемы

$$\Delta y - \delta y^{(1)} = -ch^3 \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} f^2 \right) + O(h^4), \tag{5.6}$$

а для второй

$$\Delta y - \delta y^{(2)} = +ch^3 \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} f^2 \right) + O(h^4). \tag{5.7}$$

Заметим, что порядок схем (5.6) и (5.7) на единицу меньше, чем у стандартных, так как было нарушено условие 5а и сохранена величина третьего порядка малости относительно h.

Приведем два набора параметров, для которых $c = 1/24$:

схема 1:

$$p_1^{(1)} = 0; \quad p_2^{(1)} = 0; \quad p_3^{(1)} = 1;$$
$$\alpha_2^{(1)} = 1/3; \quad \alpha_3^{(1)} = 1/2;$$
$$\beta_{21}^{(1)} = 1/3; \quad \beta_{31}^{(1)} = 0; \quad \beta_{32}^{(1)} = 1/2;$$

схема 2:

$$p_1^{(2)} = 2/5; \quad p_2^{(2)} = 0; \quad p_3^{(2)} = 3/5;$$
$$\alpha_2^{(2)} = 1/3; \quad \alpha_3^{(2)} = 5/6;$$
$$\beta_{21}^{(2)} = 1/3; \quad \beta_{31}^{(2)} = 0; \quad \beta_{32}^{(2)} = 5/6.$$

При этом получаем двустороннюю оценку значения точного решения y в точке $x_1 = x_0 + h$:

$$\min \left\{ y^{(1)}(x_1), y^{(2)}(x_1) \right\} \le y(x_1) \le \max \left\{ y^{(1)}(x_1), y^{(2)}(x_1) \right\},$$

где

$$y^{(1)}(x_1) = y(x_0) + h \left[p_1^{(1)} K_1^{(1)} + p_2^{(1)} K_2^{(1)} + p_3^{(1)} K_3^{(1)} \right];$$

$$y^{(2)}(x_1) = y(x_0) + h \left[p_1^{(2)} K_1^{(2)} + p_2^{(2)} K_2^{(2)} + p_3^{(2)} K_3^{(2)} \right].$$

Имея начальное условие, определяющее значение $y(x_0)$, и шаг h, вычисляем величины $y^{(1)}(x_1)$ и $y^{(2)}(x_1)$. Одна из них оценивает $y(x_1)$ сверху, а другая — снизу. Далее, используя каждое из этих значений в качестве начального, строим $y^{(1,1)}(x_2)$, $y^{(1,2)}(x_2)$, $y^{(2,1)}(x_2)$ и $y^{(2,2)}(x_2)$, где $x_2 = x_1 + h$. Наибольшее и наименьшее из этих значений обеспечивают

двустороннюю оценку для $y(x_2)$. Таким образом, на i-м шаге необходимо вычислять и сравнивать 2^i приближений, построенных по методу Рунге–Кутты. Следовательно, практическая реализация вычисления двусторонних оценок — это трудоемкая задача.

Цель работы. Исследовать эффективность двустороннего метода и сравнить его с методом Эйлера на примере задачи Коши с известным точным решением.

Алгоритм выполнения задания. Рассмотрим на отрезке $[0, 1]$ следующую задачу:

$$\begin{cases} y'(x) &= y\cos(tx); \\ y(0) &= 1, \end{cases}$$

точным решением которой является функция $y(x) = e^{\sin(tx)/t}$. Параметр t позволяет получить серию результатов. Например, сравнение методов можно провести, выбирая значения $t = 1$ и $t = 10$. При выполнении задания предлагается рассмотреть сетки с шагом $h = 0.01$; 0.005 и 0.0025, оценивая порядок метода Эйлера и построенного двустороннего метода. Для того чтобы сделать вывод и сравнение результатов более наглядными, это следует делать в небольшом, фиксированном для всех случаев наборе точек.

Р а б о т а 6

ВВЕДЕНИЕ В ТЕОРИЮ ПОСТРОЕНИЯ АПОСТЕРИОРНЫХ ОЦЕНОК ТОЧНОСТИ ПРИБЛИЖЕННЫХ РЕШЕНИЙ КРАЕВЫХ ЗАДАЧ

Постановка задачи — анализ одного из бурно развивающихся сегодня разделов вычислительной математики. Для упрощения в теоретической части описаны лишь несколько подходов к построению апостериорных оценок погрешности на примере классической краевой задачи. Более детальный анализ различных аспектов и обширный список литературы, посвященной методам построения апостериорных оценок для краевых задач различных типов, представлены, в частности, в монографиях [6], [8] и [9].

Рассмотрим задачу Дирихле для уравнения Пуассона в плоской ограниченной связной области Ω с непрерывной по Липшицу границей $\partial\Omega$

$$\begin{cases} -\Delta u = f & \text{в } \Omega; \\ u = 0 & \text{на } \partial\Omega, \end{cases} \tag{6.1}$$

где $f \in \mathbf{L}_2(\Omega)$.

Обобщенная формулировка данной задачи имеет следующий вид: найти такой элемент $u \in V_0$, что

$$\int\limits_{\Omega} \nabla u \cdot \nabla w dx = \int\limits_{\Omega} f w dx, \quad \forall w \in V_0, \tag{6.2}$$

где $V_0 = \mathring{\mathbf{W}}_2^1(\Omega)$ — подпространство функций из пространства Соболева $\mathbf{W}_2^1(\Omega)$, обращающихся в ноль на границе области в смысле оператора следа.

Рассмотрим некоторую аппроксимацию $v \in V_0$ точного решения краевой задачи u. Контроль ее точности сводится к оценке некоторой нормы отклонения $e = u - v$. Далее речь будет идти исключительно об апостериорных оценках так называемой энергетической нормы, которая в данном случае определяется как $\mathbf{L}_2$-норма градиента функции

$$\|\nabla e\|_{\Omega}^2 = \int\limits_{\Omega} |\nabla e|^2 dx.$$

Отметим, что для различных $\mathbf{L}_2$-норм вводится единое обозначение — $\|.\|_{\times}$, в котором нижний индекс конкретизирует область.

Для приближенного решения v задачи (6.2) ошибка e является решением аналогичной задачи, а именно

$$\int\limits_{\Omega} \nabla e \cdot \nabla w dx = \int\limits_{\Omega} \left(f w - \nabla v \cdot \nabla w\right) dx, \quad \forall w \in V_0. \tag{6.3}$$

Можно показать (см., например, [9], [10]), что в этом случае энергетическая норма ошибки контролируется через норму невязки $f + \Delta v$ исходного дифференциального уравнения в пространстве $\mathbf{W}_2^{-1}(\Omega)$, которая имеет вид

$$\sup_{w \in V_0, w \neq 0} \frac{\int\limits_{\Omega} \left(f w - \nabla v \cdot \nabla w\right) dx}{\|w\|_{1,\Omega}},$$

где

$$\|w\|_{1,\Omega}^2 = \|w\|_{\Omega}^2 + \|\nabla w\|_{\Omega}^2.$$

В действительности, если использовать другую (эквивалентную стандартной) норму

$$[\![f + \Delta v]\!]_{(-1)} = \sup_{w \in V_0, w \neq 0} \frac{\int\limits_{\Omega} \left(f w - \nabla v \cdot \nabla w\right) dx}{\|\nabla w\|_{\Omega}}, \tag{6.4}$$

в знаменателе которой стоит не полная норма в пространстве V_0, а так называемая старшая норма (вычисленная по старшим производным), то

получим соотношение

$$\|\nabla u - \nabla v\|_\Omega \;=\; [\![f + \Delta v]\!]_{(-1)}, \quad \forall v \in V_0. \tag{6.5}$$

Докажем это утверждение.

Из соотношения (6.3), положив $w = e$, имеем

$$\|\nabla e\|_\Omega \;=\; \frac{\int\limits_\Omega (fe - \nabla v \cdot \nabla e)\, dx}{\|\nabla e\|_\Omega} \;\le\; [\![f + \Delta v]\!]_{(-1)}.$$

С другой стороны, снова используя (6.3), по неравенству Гёльдера получаем обратную оценку:

$$\int\limits_\Omega (fw - \nabla v \cdot \nabla w)\, dx \;\le\; \|\nabla e\|_\Omega \|\nabla w\|_\Omega, \quad \forall w \in V_0,$$

или

$$[\![f + \Delta v]\!]_{(-1)} \;\le\; \|\nabla e\|_\Omega.$$

Отсюда следует, что две нормы совпадают.

На том факте, что так называемая негативная норма невязки позволяет контролировать погрешность приближенных решений рассматриваемой задачи, основано целое направление в теории апостериорного контроля точности, начало которому положили работы [10] и [11].

Явный метод невязок заключается в построении явно вычисляемой верхней оценки для выражения в правой части соотношения (6.5). Важную роль при ее получении играет оператор интерполирования специального вида. Оператор, предложенный Ф. Клеманом в работе [12], позволяет интерполировать функции, не имеющие необходимой гладкости. Известно, что в плоской области Ω классическая схема построения интерполянта не может быть использована для произвольного элемента u, имеющего лишь первые обобщенные производные (значения функции u в отдельных точках могут быть не определены). В работе [12] автор рассматривает способ интерполяции функций из пространств Соболева, аналогичный классическому. Он носит локальный характер и основывается на разбиении области Ω на элементы и привлечении стандартных аппроксимаций метода конечных элементов. Покажем, как это делается, на примере конечномерного подпространства $V_h \subset V_0$, полученного на базе простейших кусочно-линейных аппроксимаций. В данном случае индекс h является стандартным обозначением характерного размера сетки. Предположим, что Ω — область в $\mathbb{R}^2$ с полигональной границей. Рассмотрим разбиение области на треугольные элементы (триангуляцию).

С каждым узлом сетки X свяжем совокупность прилегающих к нему элементов, обозначаемую Ω_X (см. рис. 6.1). Рассмотрим произвольную

функцию $u \in V_0$ и построим ее локальную проекцию на пространство $\mathcal{P}_k(\Omega_X)$ — пространство полиномов степени не выше k. Обозначая полученный полином $\Pi_X u$, имеем

$$\int\limits_{\Omega_X} up\,dx = \int\limits_{\Omega_X} (\Pi_X u)p\,dx, \quad \forall p \in \mathcal{P}_k(\Omega_X).$$

Заметим, что в отличие от самой функции значение полинома $\Pi_X u$ в узле сетки X всегда определено и может быть использовано для построения стандартного кусочно-линейного восполнения на элементах сетки. В результате получаем оператор интерполирования $I_h : V_0 \to V_h$, для которого значение $I_h u$ в узле X определяется следующим образом: $(I_h u)(X) = (\Pi_X u)(X)$. При учете нулевого краевого условия для граничных узлов значение $(I_h u)(X)$ принимается равным нулю. В работе [12] также получены оценки, устанавливающие необходимые аппроксимационные свойства построенного оператора интерполирования, а именно, для каждого элемента разбиения T и каждой его грани E справедливы следующие неравенства:

$$\|u - I_h u\|_T \le C_{\Omega_T}\, h_T\, \|\nabla u\|_{\Omega_T};$$
$$\|u - I_h u\|_E \le C_{\Omega_E}\, \sqrt{h_E}\, \|\nabla u\|_{\Omega_E},$$

где Ω_T и Ω_E — объединения элементов триангуляции, имеющих общие вершины с элементом T и гранью E соответственно; h_T — длина наибольшей стороны элемента; h_E — длина грани E; C_{Ω_T} и C_{Ω_E} — постоянные, которые не зависят от h_T и h_E.

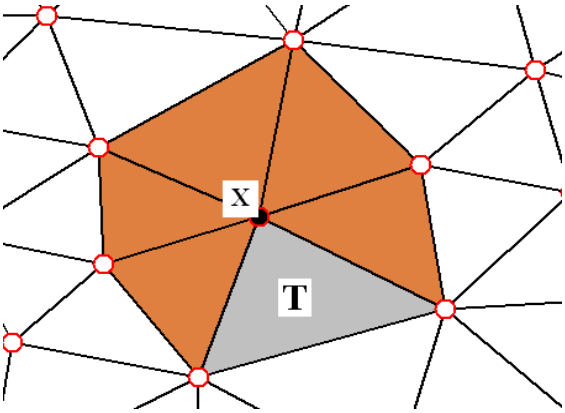

Рис. 6.1. *Совокупность элементов* Ω_X,

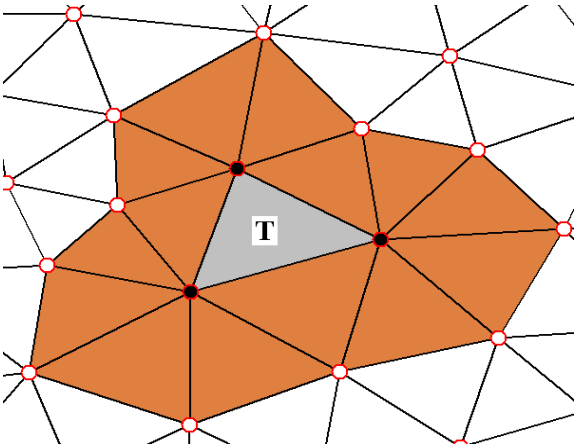

Рис. 6.2. *Совокупность элементов* Ω_T *включающих* X *в качестве вершины*

На рис. 6.2 представлена совокупность элементов Ω_T (Ω_E имеет аналогичную структуру).

С помощью рассмотренного ранее оператора интерполирования оценивают норму (6.4), входящую в соотношение (6.5). Это возможно сделать в случае, когда приближенное решение v является Галеркинской аппроксимацией u_h, т.е. удовлетворяет следующему соотношению:

$$\int_\Omega \nabla u_h \cdot \nabla w_h dx = \int_\Omega f w_h dx, \quad \forall w_h \in V_h.$$

Отсюда следует, что для произвольного элемента $w \in V_0$ имеет место равенство

$$\int_\Omega \nabla u_h \cdot \nabla (I_h w) dx - \int_\Omega f\,(I_h w) dx = 0.$$

Таким образом,

$$\int_\Omega (fw - \nabla u_h \cdot \nabla w) dx = \int_\Omega (f\bar{w} - \nabla u_h \cdot \nabla \bar{w}) dx,$$

где $\bar{w} = w - I_h w$.

Применяя формулу интегрирования по-частям, получаем

$$\int_\Omega (f\bar{w} - \nabla u_h \cdot \nabla \bar{w}) dx = \sum_{T \in \mathcal{T}_h} \left(\int_T (f + \Delta u_h)\bar{w}\, dx - \int_{\partial T \setminus \partial\Omega} (\nabla u_h \cdot n_T)\bar{w}\, ds \right),$$

где $\mathcal{T}_h$ — совокупность всех элементов разбиения; n_T — внешняя нормаль к границе элемента T (в сумму входят только те ее части, которые не попадают на границу области $\partial\Omega$).

Учитывая тот факт, что каждая внутренняя грань является общей для пары элементов и обходится дважды, второе слагаемое можно переписать как сумму по множеству $\mathcal{E}_h$ всех граней (кроме лежащих на границе области Ω), а именно

$$\sum_{T \in \mathcal{T}_h} \left(- \int_{\partial T \setminus \partial\Omega} (\nabla u_h \cdot n_T)\bar{w}\, ds \right) = \sum_{E \in \mathcal{E}_h} \int_E j(\nabla u_h \cdot n_E)\bar{w}\, ds,$$

где n_E — нормаль к соответствующей грани; $j(...)$ — скачок нормальной составляющей вектора ∇u_h.

Для каждой грани E может быть выбрано любое из двух направлений нормали. При этом скачок должен быть определен как разность

$$j(\nabla u_h \cdot n_E) = (\nabla u_h \mid_{T_{(+)}} - \nabla u_h \mid_{T_{(-)}}) \cdot n_E,$$

где элементы выбраны так, чтобы нормаль n_E была направлена в сторону от $T_{(-)}$ к $T_{(+)}$.

Оценим слагаемые сумм по неравенству Гёльдера и используем локальные свойства оператора интерполирования Клемана

$$\sum_{T\in\mathcal{T}_h}\int\limits_T (f+\Delta u_h)\bar{w}\,dx + \sum_{E\in\mathcal{E}_h}\int\limits_E j(\nabla u_h\cdot n_E)\bar{w}\,ds \le$$

$$\le \sum_{T\in\mathcal{T}_h}\|f+\Delta u_h\|_T\,\|\bar{w}\|_T + \sum_{E\in\mathcal{E}_h}\|j(\nabla u_h\cdot n_E)\|_E\,\|\bar{w}\|_E \le$$

$$\le \sum_{T\in\mathcal{T}_h} h_T C_{\Omega_T}\|f+\Delta u_h\|_T\,\|\nabla w\|_{\Omega_T} + \sum_{E\in\mathcal{E}_h}\sqrt{h_E}C_{\Omega_E}\|j(\nabla u_h\cdot n_E)\|_E\,\|\nabla w\|_{\Omega_E}.$$

Применяя неравенство Коши–Шварца и объединяя начало и конец вывода, для $e_h = u - u_h$ получаем следующую оценку:

$$\|\nabla e_h\|_\Omega = \sup_{w\in V_0, w\neq 0}\frac{\int\limits_\Omega (fw - \nabla u_h\cdot\nabla w)dx}{\|\nabla w\|_\Omega} \le$$

$$\le C_1\left(\sum_{T\in\mathcal{T}_h} h_T^2\|f+\Delta u_h\|_T^2\right)^{1/2} + C_2\left(\sum_{E\in\mathcal{E}_h} h_E\|j(\nabla u_h\cdot n_E)\|_E^2\right)^{1/2},$$

где C_1 и C_2 — постоянные, зависящие от локальной структуры всего разбиения. Вообще говоря это значит, что при адаптации сеток значения констант необходимо пересчитывать, что влечет за собой серьезные затруднения при практическом применении явного метода невязок для получения верхних оценок нормы ошибки. Обычно, чтобы избежать проблем с оценкой таких постоянных, говорят лишь об анализе зон, где погрешности велики относительно среднего значения по всей области. При этом значения всех постоянных можно условно принять равными единице. В литературе известны различные индикаторы такого типа. В частности, это индикатор, содержащий исключительно скачки

$$\eta_T^j = \left(\frac{1}{2}\sum_{E\in\mathcal{E}_T} h_E\|j(\nabla u_h\cdot n_E)\|_E^2\right)^{1/2}, \tag{6.6}$$

где $\mathcal{E}_T$ — совокупность граней элемента T, не попадающих на границу области $\partial\Omega$.

Алгоритм его применения для заданного приближенного решения u_h сводится к последовательному вычислению величин η_T^j для всех элементов $T\in\mathcal{T}_h$, к их сравнению между собой и к выбору для дальнейшего измельчения тех элементов, для которых индикатор имеет значение, превосходящее установленный порог.

Метод усреднения градиента приближенного решения основан на принципиально другой идее. Предположим, что имеется некоторая процедура, которая приближает поле градиента аппроксимации u_h к ∇u, т.е. уточняет его. Тогда разница между исходным и уточненным полем может быть использована в качестве индикатора погрешности. Другими словами, если существует оператор G, обеспечивающий выполнение неравенства

$$\|\nabla u - G\nabla u_h\|_\Omega \leq \ \delta \ \|\nabla u - \nabla u_h\|_\Omega \tag{6.7}$$

при $\delta < 1$, то справедлива двусторонняя оценка

$$\frac{1}{1+\delta}\|\nabla u_h - G\nabla u_h\|_\Omega \ \leq \ \|\nabla u - \nabla u_h\|_\Omega \ \leq \ \frac{1}{1-\delta}\|\nabla u_h - G\nabla u_h\|_\Omega.$$

Если неравенство (6.7) выполняется при малом значении δ, то норма разности $\nabla u_h - G\nabla u_h$ дает точное представление об истинном значении энергетической нормы ошибки.

Впервые подход подобного рода был предложен в работе [13]. Дальнейшие исследования показали, что разница между градиентом приближенного решения и его усреднением часто является хорошим индикатором ошибки. Преимущество такого подхода — его простота и малая вычислительная трудоемкость.

Как математическое обоснование метода, так и конструктивное построение некоторого оператора G в большинстве случаев основано на так называемом эффекте суперсходимости. Его исследованию посвящено большое количество работ, среди которых первая [14] (обзор результатов, связанных с эффектом суперсходимости, представлен в монографии [15]). Основываясь на данном эффекте, установлено, что при наличии у точного решения повышенной гладкости (например, $u \in \mathbf{W}_2^3(\Omega)$) процедура усреднения повышает асимптотическую скорость сходимости и имеет место оценка

$$\|\nabla u - G\nabla u_h\|_\Omega \leq Ch^2\|u\|_{3,\Omega},$$

вместо известной асимптотики

$$\|\nabla u - \nabla u_h\|_\Omega \simeq O(h).$$

Тогда при достаточно малых h величина $\|\nabla u_h - G\nabla u_h\|_\Omega$ близка к значению нормы ошибки.

Как отмечалось ранее, первым индикатором, основанным на усреднении градиента, был индикатор, предложенный в работе [13]. В ней рассматривалась процедура усреднения кусочно-постоянного поля напряжений в простом одномерном случае. При этом значение в узле сетки

получалось как полусумма значений на соседних интервалах. В плоском случае аналогичная процедура приводит к индикатору вида

$$\eta_T^{ZZ} = \|\nabla u_h - G\nabla u_h\|_T, \tag{6.8}$$

где $G\nabla u_h$ — кусочно-линейное непрерывное векторное поле, значение которого в узле X определяется по формуле (см. рис. 6.1)

$$G\nabla u_h|_X = \sum_{T\in\Omega_X} \frac{|T|}{|\Omega_X|}\ \nabla u_h|_T. \tag{6.9}$$

Другая эффективная процедура восстановления значений градиента приближенного решения в узлах сетки была предложена в работе [16] и ряде цитируемых в ней работ. Данный метод получил сокращение SPR (superconvergent patch recovery). Он заключается в следующем. Для каждого узла сетки X рассматривается совокупность элементов Ω_X (см. рис. 6.1). На всех элементах вычисляются значения производных приближенного решения в центрах масс. Далее по ним строится полином над Ω_X, значения которого в этих точках приближают заданные значения при помощи метода наименьших квадратов. Значение полученного полинома в центральном узле X дает необходимое усреднение. Получив усредненные производные во всех узлах сетки, можно построить по ним кусочно-линейное поле $G\nabla u_h$. Такая процедура универсальна и легко адаптируется под различные типы конечных элементов. Она локальна и не требует существенных вычислительных затрат. Однако, наряду с преимуществами, метод имеет и ряд недостатков. Прежде всего, явление суперсходимости основано на повышенной гладкости решения и наблюдается далеко не всегда. Даже для простых эллиптических краевых задач решение может не иметь обобщенных производных порядка выше первого (например, для областей с негладкими границами). Решения нелинейных краевых задач (в частности, вариационных неравенств) часто обладают предельной регулярностью, которая не зависит от гладкости внешних данных. Кроме того, данная технология, как и метод невязок, может быть применена только к аппроксимациям специального вида. Однако в литературе описано много примеров, когда рассматриваемый метод позволяет получать качественную индикацию погрешности даже в тех случаях, где его применение не имеет строгого математического обоснования.

Функциональные апостериорные оценки. Рассмотрим метод, который был предложен С.Г. Михлиным в книге [17]. Вначале рассмотрим вариационную формулировку задачи (6.1).

Задача $\mathcal{P}$. Найти такой элемент $u \in V_0$, что

$$J(u) = \inf \mathcal{P} = \inf_{w \in V_0} J(w),$$

где функционал энергии J имеет вид

$$J(w) = \int\limits_\Omega \left(\frac{1}{2} |\nabla w|^2 - fw \right) dx.$$

Преобразуем разность между значением функционала J на произвольном приближенном решении v и на точном решении u

$$J(v) - \inf \mathcal{P} = \int\limits_\Omega \left(\frac{1}{2} |\nabla v|^2 - fv \right) dx - \int\limits_\Omega \left(\frac{1}{2} |\nabla u|^2 - fu \right) dx =$$

$$= \frac{1}{2} \|\nabla v - \nabla u\|_\Omega^2 + \int\limits_\Omega \left(\nabla u \cdot \nabla (v - u) - f(v - u) \right) dx.$$

С учетом соотношения (6.2) для $w = v - u$ имеем следующее представление квадрата нормы ошибки:

$$\|\nabla v - \nabla u\|_\Omega^2 = 2(J(v) - \inf \mathcal{P}).$$

Таким образом, если была бы известна точная нижняя грань рассматриваемой задачи, то ее значение позволило оценить степень близости приближенного решения к точному. К сожалению, эта величина, как правило, неизвестна. Однако в книге [17] предложен способ вычисления сколь угодно точной оценки значения $-\inf \mathcal{P}$ сверху. Она строится при помощи элементов $q^* \in \mathbf{L}_2(\Omega, \mathbb{R}^2)$, удовлетворяющих в обобщенном смысле соотношению $\operatorname{div} q^* + f = 0$, т.е. принадлежащих множеству

$$Q_f^* = \left\{ q^* \in \mathbf{L}_2(\Omega, \mathbb{R}^2) \mid \int\limits_\Omega q^* \cdot \nabla w dx = \int\limits_\Omega fw dx, \quad \forall w \in V_0 \right\}.$$

Оценка имеет вид

$$-2 \inf \mathcal{P} \ \leq \ \|q^*\|_\Omega^2$$

и верна для любого $q^* \in Q_f^*$. Она приводит к неравенству

$$\|\nabla v - \nabla u\|_\Omega^2 \ \leq \ \int\limits_\Omega \left(|\nabla v|^2 - 2fv + |q^*|^2 \right) dx. \tag{6.10}$$

Неравенство (6.10) можно переписать так:

$$\|\nabla v - \nabla u\|_\Omega \ \leq \ \|\nabla v - q^*\|_\Omega, \quad \forall q^* \in Q_f^*. \tag{6.11}$$

Оценка (6.11) содержит свободную переменную, подчиненную сильному ограничению. Чтобы получить даже грубую индикацию погрешности при помощи неравенства (6.11), необходимо построить элемент q^*, который с высокой степенью точности удовлетворяет соотношению $\operatorname{div} q^* + f = 0$. Для правой части f произвольного вида такое построение требует большого объема дополнительных вычислений. Именно это ограничение делает подход не столь эффективным с практической точки зрения.

Преобразуем неравенство (6.11) таким образом, чтобы свободная переменная в нем была подчинена более слабому ограничению. Для этого введем вторую свободную переменную $y^* \in \mathbf{L}_2(\Omega, \mathbb{R}^2)$. Тогда по неравенству треугольника

$$\|\nabla v - \nabla u\|_\Omega \le \|\nabla v - y^*\|_\Omega + \|y^* - q^*\|_\Omega, \qquad \forall q^* \in Q_f^*. \tag{6.12}$$

Возведя обе части (6.12) в квадрат и применив известное алгебраическое неравенство Коши со свободным параметром (см., например, [18])

$$2 \mid ab \mid \le \beta a^2 + \beta^{-1} b^2, \quad \forall \beta > 0,$$

получаем

$$\|\nabla v - \nabla u\|_\Omega^2 \le (1+\beta)\|\nabla v - y^*\|_\Omega^2 + (1+\beta^{-1})\|y^* - q^*\|_\Omega^2. \tag{6.13}$$

Дальнейшие преобразования основаны на работе [19], в которой использовалось ортогональное разложение Гельмгольца для элементов пространства $\mathbf{L}_2(\Omega, \mathbb{R}^2)$, приведенное в книге [17], а именно, для любого элемента $y^* \in \mathbf{L}_2(\Omega, \mathbb{R}^2)$ справедливо следующее представление:

$$y^* = \nabla w_0 + q_0^*, \qquad w_0 \in V_0,\ q_0^* \in Q_0^*, \tag{6.14}$$

где

$$Q_0^* = \left\{ q_0^* \in \mathbf{L}_2(\Omega, \mathbb{R}^2) \mid \int_\Omega q_0^* \cdot \nabla w dx = 0, \quad \forall w \in V_0 \right\}.$$

Вначале получим оценку второго слагаемого в правой части неравенства (6.13) для тривиального случая, когда $f \equiv 0$ и q^* — элемент множества Q_0^*, т.е. $q^* = q_0^*$. В силу представления (6.14) имеем $y^* - q_0^* = \nabla w_0$, следовательно,

$$\|y^* - q_0^*\|_\Omega^2 = \|\nabla w_0\|_\Omega^2.$$

Рассмотрим негативную норму элемента $\operatorname{div} y^*$, определенную аналогично соотношению (6.4):

$$\|\!|\operatorname{div} y^*|\!\|_{(-1)} = \sup_{w\in V_0, w\neq 0} \frac{-\int\limits_\Omega y^*\cdot\nabla w dx}{\|\nabla w\|_\Omega} \geq \frac{\int\limits_\Omega y^*\cdot\nabla w_0 dx}{\|\nabla w_0\|_\Omega}.$$

Используя разложение (6.14) и учитывая, что $q_0^* \in Q_0^*$, имеем

$$\int\limits_\Omega y^*\cdot\nabla w_0 dx = \int\limits_\Omega (\nabla w_0 + q_0^*)\cdot\nabla w_0 dx = \|\nabla w_0\|_\Omega^2.$$

Отсюда получаем неравенство

$$\|y^* - q_0^*\|_\Omega \;=\; \|\nabla w_0\|_\Omega \;\leq\; \|\!|\operatorname{div} y^*|\!\|_{(-1)},$$

которое справедливо для произвольного элемента $y^* \in \mathbf{L}_2(\Omega,\mathbb{R}^2)$.

Теперь предположим, что $f \not\equiv 0$. При этом элемент q^* должен принадлежать не множеству Q_0^*, а множеству Q_f^*. Заметим, что $y^* - \nabla u$ есть элемент пространства $\mathbf{L}_2(\Omega,\mathbb{R}^2)$, для которого верно разложение Гельмгольца

$$y^* - \nabla u = q_0^* + \nabla w_0$$

или

$$y^* = q_0^* + \nabla u + \nabla w_0.$$

В силу интегрального тождества (6.2)

$$\int\limits_\Omega \nabla u\cdot\nabla w dx = \int\limits_\Omega f w dx, \quad \forall w \in V_0,$$

и определения Q_0^* элемент $q^* = q_0^* + \nabla u$ принадлежит множеству Q_f^*. Отсюда получаем оценку

$$\|y^* - q^*\|_\Omega = \|\nabla w_0\|_\Omega \leq \|\!|\operatorname{div}(y^* - \nabla u)|\!\|_{(-1)} = \|\!|\operatorname{div} y^* + f|\!\|_{(-1)},$$

которая приводит к первой форме функциональной апостериорной оценки для задачи (6.1):

$$\|\nabla v - \nabla u\|_\Omega^2 \;\leq\; (1+\beta)\|\nabla v - y^*\|_\Omega^2 + (1+\beta^{-1})\|\!|\operatorname{div} y^* + f|\!\|_{(-1)}^2. \quad (6.15)$$

Ограничим класс допустимых переменных y^* элементами множества

$$Q_{\operatorname{div}}^* = \left\{ y^* \in \mathbf{L}_2(\Omega,\mathbb{R}^2) \mid \operatorname{div} y^* \in \mathbf{L}_2(\Omega) \right\}.$$

Тогда в силу соотношения

$$-\int\limits_\Omega y^*\cdot\nabla w dx = \int\limits_\Omega \operatorname{div} y^* w dx;$$

$$\|\!|\operatorname{div} y^* + f|\!\|_{(-1)} = \sup_{w\in V_0, w\neq 0} \frac{\int\limits_\Omega (\operatorname{div} y^* + f) w dx}{\|\nabla w\|_\Omega} \leq \mathbb{C}_{\mathrm{F}\Omega} \|\operatorname{div} y^* + f\|_\Omega,$$

где $\mathbb{C}_{\mathrm{F}\Omega}$ — константа в неравенстве Фридрихса, которое имеет вид

$$\|w\|_\Omega \leq \mathbb{C}_{\mathrm{F}\Omega} \|\nabla w\|_\Omega, \quad \forall w \in V_0.$$

В итоге получаем оценку точности, которая справедлива для любого приближенного решения $v \in V_0$:

$$\|\nabla(v-u)\|_\Omega^2 \leq \mathcal{M}(v, \beta, y^*), \tag{6.16}$$

где

$$\mathcal{M}(v, \beta, y^*) = (1+\beta)\|\nabla v - y^*\|_\Omega^2 + (1+\beta^{-1})\mathbb{C}_{\mathrm{F}\Omega}^2 \|\operatorname{div} y^* + f\|_\Omega^2, \tag{6.17}$$

y^* — произвольный элемент множества Q^*_{div}, β — произвольный положительный параметр.

Отметим, что в работе [19] оценка (6.16) получена непосредственно, без промежуточной оценки (6.15), для вывода которой первоначально привлекались методы теории двойственности вариационного исчисления (см., например, [20]).

Рассмотрим некоторые вычислительные свойства мажоранты (6.17). Отметим, что неравенство (6.16) может быть записано в виде

$$\|\nabla(v-u)\|_\Omega \leq \|\nabla v - y^*\|_\Omega + \mathbb{C}_{\mathrm{F}\Omega} \|\operatorname{div} y^* + f\|_\Omega, \tag{6.18}$$

который не содержит свободного параметра β. Такое представление мажоранты более удобно при установлении ее свойств. На практике предпочтительнее иметь квадратичную структуру функционала в правой части апостериорной оценки.

Множество Q^*_{div}, которому принадлежит свободный элемент y^*, на самом деле является гильбертовым пространством с нормой

$$\|y^*\|_{\mathrm{div}}^2 = \|y^*\|_\Omega^2 \; + \; \|\operatorname{div} y^*\|_\Omega^2,$$

порожденной соответствующим скалярным произведением. Элемент $p^* = \nabla u$, очевидно, принадлежит данному множеству, поскольку $-\operatorname{div} p^* = f \in \mathbf{L}_2(\Omega)$. При этом значение мажоранты в правой части неравенства (6.18) равно $\|\nabla v - \nabla u\|_\Omega$. Таким образом, рассматриваемая оценка может быть получена сколь угодно точно, и для эффективного контроля погрешности достаточно построить последовательность y_k^*, которая бы сходилась к p^* в пространстве Q^*_{div}.

В заключение более подробно рассмотрим процесс вычисления постоянной $\mathbb{C}_{\mathrm{F}\Omega}$. Как следует из вида мажоранты (6.17), чтобы обеспечить

надежность контроля погрешности, необходимо получить для $\mathbb{C}_{\mathrm{F}\Omega}$ оценку сверху. В общем случае эта константа определяется через наименьшее собственное значение λ_Ω оператора Лапласа для области Ω, а именно

$$\mathbb{C}_{\mathrm{F}\Omega} = 1/\sqrt{\lambda}_\Omega,$$

где

$$\lambda_\Omega = \inf_{w \in V_0} \frac{\int\limits_\Omega |\nabla w|^2 dx}{\int\limits_\Omega |w|^2 dx}.$$

Отметим, что для первой краевой задачи имеет место неравенство

$$\lambda_{\tilde{\Omega}} \le \lambda_\Omega, \quad \forall \tilde{\Omega} \supset \Omega, \tag{6.19}$$

которое позволяет получать оценки сверху для $\mathbb{C}_{\mathrm{F}\Omega}$. Действительно, если $\tilde{\Omega}$ — какая-либо каноническая область (например, прямоугольник или круг), то $\lambda_{\tilde{\Omega}}$ находится аналитически. Из соотношения (6.19) следует, что в качестве оценки снизу для λ_Ω, , можно использовать значение λ для квадрата со стороной l, которое вычисляется по формуле

$$\lambda = 2\pi^2/l^2.$$

Цель работы. На примере краевой задачи (6.1) сравнить качество оценок погрешности приближенных решений, полученных при помощи явного метода невязок, метода усреднения градиента и функционального метода.

Алгоритм выполнения задания. В качестве тестового примера рассмотрим любую задачу (6.1) с известным точным решением. Например, для квадратной области Ω с единичной стороной можно выбрать правую часть f так, чтобы ей соответствовало точное решение вида

$$u = x_1(1-x_1)\ x_2(1-x_2)\ e^{A(x_1-x_{10})^2}\ e^{B(x_2-x_{20})^2}.$$

Варьирование свободных параметров A, B, x_{10}, x_{20} оказывает влияние на форму точного решения и локальное распределение погрешности по области. Приближенно решить такую задачу в постановке (6.2) можно при помощи метода конечных элементов с использованием PDE Toolbox пакета MATLAB. Данное средство позволяет создавать области различной геометрии, разбивать их сетками, состоящими из треугольных элементов, и строить непрерывные кусочно-линейные аппроксимации решения задачи на этих сетках. После построения приближенного решения v, которое будет близко к галеркинской аппроксимации u_h, зная точное решение u, можно вычислить настоящее распределение погрешности по

области Ω, а именно вклады в интеграл ошибки

$$\int\limits_{\Omega} |\nabla(u-v)|^2 dx$$

на каждом элементе триангуляции. Далее предлагается сравнить индикатор (6.6), индикатор (6.8) с усреднением вида (6.9) и мажоранту (6.17). Основным критерием здесь является качество воспроизведения истинного распределения ошибки по области при помощи того или иного метода. При использовании функционального метода необходимо обращать внимание на степень переоценки глобальной величины погрешности, т.е. на близость друг к другу левой и правой частей неравенства (6.16).

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. **Форсайт Д., Малькольм М., Моулер К.** Машинные методы математических вычислений. М.: Мир, 1980. 279 с.
2. **Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.** Численные методы. М.: Лаборатория базовых знаний, 2002. 630 с.
3. **Крылов В.И., Бобков В.В., Монастырный П.И.** Вычислительные методы. Т. 1. М.: Наука, 1976. 304 с.
4. **Ануфриев И.Е., Смирнов А.Б., Смирнова Е.Н.** MATLAB 7: Наиболее полное руководство. СПб.: БХВ-Петербург, 2005. 1082 с.
5. **Колмогоров А.Н., Фомин С.В.** Элементы теории функций и функционального анализа. М.: Наука, 1976. 543 с.
6. **Neittaanmäki P., Repin S.** Reliable methods for computer simulation. Error control and a posteriori estimates. New York: Elsevier, 2004. 305 p.
7. **Коллатц Л.** Функциональный анализ и вычислительная математика. М.: Мир, 1969. 447 с.
8. **Ainsworth M., Oden J.T.** A posteriori error estimation in finite element analysis. New York: John Wiley & Sons, 2000. 240 p.
9. **Verfürth R.** A review of a posteriori error estimation and adaptive mesh-refinement techniques. Chichester; Stuttgart: John Wiley & Sons, B.G. Teubner, 1996. 127 p.
10. **Babuška I., Rheinboldt W.C.** Error estimates for adaptive finite element computations // SIAM J. Numer. Anal. 1978. Vol. 15, N 4. P. 736–754.
11. **Babuška I., Rheinboldt W.C.** A-posteriori error estimates for the finite element method // Int. J. Numer. Methods Eng. 1978. Vol. 12. P. 1597–1615.
12. **Clément Ph.** Approximation by finite element functions using local regularization // Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér. 1975. Vol. 9, NR-2. P. 77–84.

13. **Zienkiewicz O.C., Zhu J.Z.** A simple error estimator and adaptive procedure for practical engineering analysis // Internat. J. Numer. Methods Engrg. 1987. Vol. 24, N 2. P. 337–357.
14. **Оганесян Л.А., Руховец Л.А.** Исследование скорости сходимости вариационно-разностных схем для эллиптических уравнений второго порядка в двумерной области с гладкой границей // Журн. вычислительной математики и матем. физики. 1969. Т. 9. С. 1102–1120.
15. **Wahlbin L.B.** Superconvergence in Galerkin finite element methods. Berlin: Springer-Verlag, 1995. 166 p.
16. **Zienkiewicz O.C., Zhu J.Z.** The superconvergent patch recovery (SPR) and adaptive finite element refinement // Comput. Methods Appl. Mech. Eng. 1992. Vol. 101, N 1–3. P. 207–224.
17. **Михлин С.Г.** Вариационные методы в математической физике. М.: Наука, 1970. 512 с.
18. **Ладыженская О.А.** Краевые задачи математической физики. М.: Наука, 1973. 407 с.
19. **Repin S., Sauter S., Smolianski A.** A posteriori error estimation for the Dirichlet problem with account of the error in the approximation of boundary conditions // Computing. 2003. Vol. 70, N 3. P. 205–233.
20. **Repin S.I.** A posteriori error estimation for variational problems with uniformly convex functionals // Math. Comp. 2000. Vol. 69, N 230. P. 481–500.

Содержание

Введение 3

Р а б о т а 1. Определение машинного ε 4

Р а б о т а 2. Оценка погрешности решения системы линейных алгебраических уравнений 5

Р а б о т а 3. Априорная и апостериорная оценка точности решения системы нелинейных алгебраических уравнений 9

Р а б о т а 4. Двусторонние методы для систем линейных алгебраических уравнений 13

Р а б о т а 5. Двусторонние методы для обыкновенных дифференциальных уравнений 16

Р а б о т а 6. Введение в теорию построения апостериорных оценок точности приближенных решений краевых задач 21

Библиографический список 34

РЕПИН Сергей Игоревич
ФРОЛОВ Максим Евгеньевич

ЧИСЛЕННЫЕ МЕТОДЫ

ОЦЕНКА ПОГРЕШНОСТИ РЕШЕНИЯ

Лабораторный практикум

Редактор О.В. Махрова
Технический редактор А.И. Колодяжная
Оригинал-макет подготовлен авторами

Директор Издательства Политехнического университета *А.В. Иванов*

Свод. темплан 2006 г.

Лицензия ЛР № 020593 от 07.08.97

Налоговая льгота — Общероссийский классификатор продукции
ОК 005-93, т. 2; 95 3005 — учебная литература

Подписано в печать Формат 60x84/16
Усл. печ. л. 2,0 п.л. Уч.-изд. л. 2,0 п.л. Тираж 100. Заказ

Санкт-Петербургский государственный политехнический университет,
Издательство Политехнического университета,
член Издательско-полиграфической ассоциации университетов России.
Адрес университета и издательства:
195251, Санкт-Петербург, Политехническая ул., 29.