ТБИЛИССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
ИМЕНИ АКАДЕМИКА И.Н.ВЕКУА

С.Г.МИХЛИН

ПОГРЕШНОСТИ ВЫЧИСЛИТЕЛЬНЫХ ПРОЦЕССОВ

ИЗДАТЕЛЬСТВО ТБИЛИССКОГО УНИВЕРСИТЕТА
ТБИЛИСИ - 1983

УДК 519.6

Печатается по постановлению Ученого совета Института прикладной математики имени И.Н.Векуа Тбилисского государственного университета.

В книге рассматривается довольно широкий класс задач вычислительной математики и вычислительные процессы, которые приводят к построению решения данной задачи, анализируются источники погрешностей и оцениваются самые погрешности в предположении, что данная задача поставлена точно. Как кажется автору, в весьма многих случаях существует точно четыре источника погрешностей: погрешность аппроксимации, проистекающая от замены данной задачи другой, упрощенной задачей; погрешность искажения, связанная в тем, что данные упрощенной задачи вычисляются неточно; погрешность алгоритма, возникающая в тех случаях, когда алгоритм, применимый для решения упрощенной задачи, дает после конечного числа шагов лишь приближенное её решение; погрешность округления, возникающая из-за того, что предписания упомянутого алгоритма выполняются неточно. Для широкого круга вычислительных процессов даны оценки всех перечисленных погрешностей.

Книга содержит 9 глав. В главе I дана постановка проблемы и исследованы два более простых примера. В главах II-VI исследуются погрешности линейных вычислительных процессов, в главах VII-IX тот же вопрос исследуется для нелинейных процессов.

Книга содержит расширенное изложение докладов, сделанных автором в Институте прикладной математики имени академика И.Н. Векуа в 1982 году.

Редактор: заслуженный деятель науки ГССР

Л. Г. Магнарадзе

Рецензент: доктор физ.-матем. наук, проф.

В. Г. Мазья

ПРЕДИСЛОВИЕ

В настоящее время ни у кого не вызывает сомнения необходимость исследования и оценки различных погрешностей, возникающих при производстве вычислений, хотя ещё сравнительно недавно - лет 35-40 тому назад - многие видные представители прикладных наук считали, что в таком исследовании нет нужды и что любой вычислительный процесс надёжен, если он кажется таким. Следует признать, что в прошлом эта наивная точка зрения была не совсем безосновательной: до того, как в научный обиход вошли современные вычислительные машины, доступными для реализации были, как правило, только методы вычислений, дававшие грубые приближения; такие методы требовали небольшого объёма вычислений, при котором не могло произойти катастрофических вычислительных погрешностей. Положение изменилось, когда электронные вычислительные машины получили широкое распространение. Стало возможным выполнение вычислений огромного объёма, но при этом довольно скоро обнаружилось, что в ряде случаев результат вычислений становится совершенно недостоверным из-за накопления погрешностей счета. Специалисты по вычислительной математике в своё время обнаружили это явление и исследовали частный, но весьма важный вопрос о накоплении погрешностей округления при решении линейных алгебраических систем /Уилкинсон и др./. В своих работах, относящихся к началу 60-х годов, автор этих строк обратил внимание на погрешности /"погрешности искажения"/ возникающие от неточной замены данной задачи на более простую приближенную задачу. С погрешностью искажения связано введённое тогда же автором понятие устойчивости вычислительного процесса. В последние годы автор занимался выявлением всех возможных источников погрешностей и оценкой этих погрешностей; предлагаемая вниманию читателей книга в основном содержит результаты этих исследований. Наряду с процессами, которые автор называет сьободными /в них результат любого шага не зависит от результатов остальных шагов процесса/, исследованы также рекуррентные процессы. Много внимания уделено нелинейным вычислительным процессам, в частности, тем, которые возникают при решении односторонних вариационных задач.

Ниже следует более подробное изложение содержания книги. Она содержит 9 глав. В главе I описан класс задач вычислительной математики, которые решаются так, что каждая задача заменяется последовательностью более простых задач /точнее, любой задачей из этой последовательности/. Такая замена приводит к тому, что в данной книге названо погрешностью аппроксимации - это некоторый функционал, ограничивающий погрешность от замены данной задачи на более простую. Далее, данные приближенной задачи на самом деле нам не заданы - они нами вычисляются и притом, за редкими исключениями, с некоторой погрешностью. Таким образом, мы получаем приближенную задачу в искаженном виде; если искаженную задачу решить точно, то мы придем не к точному, а к искаженному решению приближенной задачи. Всё это приводит нас ко второму виду погрешности, которую автор называет погрешностью искажения. Как теоретические рассуждения, так и численные примеры показывают, что погрешность искажения для решения может быть очень велика, даже если погрешности данных приближенной задачи малы.

Допустим, что мы умеем решать приближенную задачу, независимо от того, дана она нам в точном или в искаженном виде. Это значит, что нам известен некоторый алгоритм, на каждом шаге которого требуется выполнить конечное число операций, выполнение которых нам доступно, и после некоторого конечного числа шагов этот алгоритм приводит /если упомянутые операции выполнены точно, без погрешности/ либо к точному решению искаженной приближенной задачи, либо к элементу, который в том или ином смысле близок к названному решению, причем эту близость можно оценить. Таким путем мы приходим к третьему виду погрешности - к погрешности алгоритма.

Последняя - четвертая - погрешность есть погрешность округления; она связана с тем, что предписания упомянутого выше алгоритма выполняются не точно, а с ошибками округления.

В той же главе I рассмотрены два примера: квадратурные формулы и системы линейных алгебраических уравнений с квадратной матрицей и с неравным нулю определителем. Первый пример интересен тем, что для него существенна только погрешность аппроксимации; во втором примере эта погрешность отсутствует, хотя, например, по методу Гаусса исходная задача заменяется другой, бо-

лее простой. Вообще, в каждом конкретном случае та или другая из отмеченных выше четырех погрешностей может отсутствовать.

Главы II-VI посвящены погрешностям линейных вычислительных процессов. В главе II для ряда приближенных методов изучается погрешность аппроксимации, в главе III - погрешность искажения, в главе IV - погрешности алгоритма и округления. Глава V содержит анализ погрешностей для метода конечных элементов, глава VI - для различных методов решения интегральных уравнений, фредгольмовских и сингулярных.

В главах VII-IX изучаются погрешности нелинейных вычислительных процессов. В главе VII рассмотрены общие вопросы; полученные результаты применяются к методу Ритца /в частности, к методу конечных элементов/ для нелинейных задач. Подробно исследованы нелинейные рекуррентные процессы, включая метод Ньютона - Канторовича.

В главе VIII рассмотрены приближенные решения одного класса односторонних вариационных задач. Исследованы погрешности аппроксимации и искажения; доказана устойчивость точного решения относительно малых возмущений данных; даны оценки возмущения решения в зависимости от возмущений данных задачи.

Глава IX содержит приложения результатов главы VIII к задачам упруго-пластического состояния, основанные на теории Сен-Венана - Мизеса и вариационном принципе Хаара - Кармана. В § 1 приведены, для удобства ссылок, основные факты теории Сен-Венана - Мизеса и принцип Хаара - Кармана. В § 2 рассмотрена задача упруго-пластического кручения. Отмечены существование решения и эквивалентность принципа Хаара - Кармана системе уравнений упруго-пластического кручения /все эти результаты получены французскими математиками школы Лионса/. Далее, доказана устойчивость точного решения задачи относительно малых возмущений данных задачи - крутящего момента и поперечного сечения стержня; получены оценки погрешностей аппроксимации и искажения. В § 3 аналогичные результаты получены для плоской задачи упруго-пластического состояния; в частности, при некоторых предположениях доказана эквивалентность принципа Хаара - Кармана и уравнений плоской задачи. Аналогичные результаты получены в § 4 для трехмерной задачи.

И.Н.Гончарова провела все вычисления по одному из примеров § 4 главы Ш. В.Б.Тихтин написал добавление к главе УШ, в котором дан обзор ряда современных работ, содержащих оценки погрешности аппроксимации для решений довольно широкого класса вариационных неравенств. Автор рад выразить И.Н.Гончаровой и В.Б. Тихтину свою сердечную благодарность.

Июнь 1982 г. С.Г.Михлин
Ленинград

ГЛАВА I

Постановка проблемы

§ I. Задачи вычислительной математики. Вычислительные процессы

1°. Весьма многие задачи вычислительной математики можно подвести под следующую схему. Дан некоторый объект f ; требуется определить другой объект x, удовлетворяющий двум требованиям: I) x принадлежит некоторому данному классу объектов X ; 2) x находится в некотором данном соотношении τ с данным объектом f . Нашу задачу можно записать так:

$$\{x \in X\,,\ x \tau f\}\,. \tag{I.I.I}$$

Объект x , удовлетворяющий условиям (I.I.I), есть решение нашей задачи.

Примеры. I. Пусть требуется вычислить интеграл

$$\int_0^1 f(t)\,dt\,,$$

где f – заданная функция. В этом случае $f = f(t)$, $X = R_1$, а соотношение τ означает, что x и f связаны уравнением

$$x = \int_0^1 f(t)\,dt\,. \tag{I.I.2}$$

В несколько усложненном примере, когда требуется вычислить интеграл вида

$$x(z) = \int_{\mathcal{D}} f(z,t)\,dt\,, \tag{I.I.3}$$

где $\mathcal{D}$ – множество в R_m, а z – параметр, который может принять любое значение из некоторого множества Ω , f есть функция двух точек z и t , X есть множество всех функций, определенных на Ω , а соотношение τ определяется равенством (I.I.3).

2°. Пусть требуется решить систему линейных алгебраических уравнений

$$Ax = b\,, \tag{I.I.4}$$

где x и b – k-мерные векторы, A – матрица порядка $k \times k$. Пусть для простоты все числа в нашей задаче – вещественные. Тогда $f = b$, $X = R_k$, соотношение τ определяется уравнением (I.I.4).

3. Рассмотрим смешанную задачу для волнового уравнения:

$$\frac{\partial^2 x}{\partial t^2} - \Delta x = f(z,t), \quad t \geq 0, \quad z \in \Omega,$$
$$x(z,0) = \varphi(z), \quad x_t(z,0) = \psi(z), \tag{I.I.5}$$
$$x(z,t)\big|_{z \in \partial\Omega} = \omega(z,t);$$

Δ – оператор Лапласа. Будем, например, искать решение этой задачи, принадлежащее соболевскому классу $W_2^{(2)}([0,\infty) \times \Omega)$. Тогда $X = W_2^{(2)}([0,\infty) \times \Omega)$, f есть совокупность функций $f(z,t)$, $\varphi(z)$, $\psi(z)$, $\omega(z,t)$, а соотношение τ определяется совокупностью уравнений (I.I.5).

4. Пусть $f(x)$ – функционал, определенный на некотором множестве X, и требуется найти элемент $x \in X$, на котором $f(x)$ достигает минимума. В данном случае $f = f(x)$, а соотношение τ можно записать так: если x_* – решение нашей вариационной задачи, то $x_* \tau f$ означает, что $f(x_*) \leq f(x)$, $\forall x \in X$, или, если угодно, что $f(x_*) = \min\limits_{x \in X} f(x)$.

3°. Как правило, задача (I.I.I) решается приближенно. В довольно широком классе случаев это означает следующее. Каждому натуральному n приводятся в соответствие: объект $f^{(n)}$; класс объектов X_n; оператор P_n, отображающий X_n в X; соотношение τ_n. Задача (I.I.I) заменяется любой задачей из последовательности

$$\{ v^{(n)} \in X_n, \; v^{(n)} \tau_n f^{(n)} \}; \tag{I.I.6}$$

мы предполагаем, что решение каждой задачи этой последовательности не зависит от решений последующих задач, но может зависеть от предыдущих. Если $v_*^{(n)}$ есть решение n-й задачи (I.I.6), то элемент $x_*^{(n)} = P_n v_*^{(n)}$ рассматривается как приближенное решение задачи (I.I.I).

Будем говорить, что задачи (I.I.6) аппроксимируют задачу (I.I.I) и допустим, что аппроксимирующие задачи (I.I.6) в том или ином смысле проще, чем аппроксимируемая задача (I.I.I). Процесс решения последовательности задач (I.I.6) назовем вычислительным процессом. Этот процесс будем называть свободным, если при любом n решение $v_*^{(n)}$ определяется независимо от значений элементов $v_*^{(k)}$, $k \neq n$, и рекуррентным, если $v_*^{(k)}$ определяется

через $v_*^{(k)}$, $k<n$.

Пример 5. Будем вычислять интеграл (I.I.2) (пример I), например, по формуле средних прямоугольников с n узлами. В таком случае $f^{(n)}$ есть совокупность n ординат $f(t_k)$ в точках $t_k=\frac{2k-1}{2n}$, $k=1,2,\ldots,n$; $X_n=R_1$, ρ_n есть тождественный оператор в R_1, а соотношение τ_n определяется выбранной квадратурной формулой:

$$v^{(n)}\tau_n f^{(n)} := v^{(n)} = \frac{1}{n}\sum_{k=1}^{n} f\left(\frac{2k-1}{2n}\right). \qquad (I.I.7)$$

Вычислительный процесс (I.I.7) - свободный.

Пример 6. В примере 2 положим $A=I-B$, где I - единичная матрица в R_K, и запишем уравнение (I.I.4) в виде $x=Bx+b$. Воспользуемся методом простой итерации: зададим какой-нибудь вектор $v^{(0)}$ и положим для любого натурального n

$$v^{(n)} = Bv^{(n-1)}+b. \qquad (I.I.8)$$

В данном случае $f^{(n)}$ есть совокупность векторов b и $v^{(n-1)}$, $X_n=X=R_K$, $\rho_n=I$, соотношение τ_n определяется формулой простых итераций

$$v^{(n)}\tau_n f^{(n)} := v^{(n)} = Bv^{(n-1)}+b. \qquad (I.I.9)$$

Вычислительный процесс (I.I.9) - рекуррентный.

Приведем еще некоторые примеры. К свободным вычислительным процессам приводят методы Ритца, Бубнова - Галеркина, конечных элементов, разностные методы, метод прямых. К рекуррентным вычислительным процессам приводят различные итерационные процессы: метод простой итерации, релаксационные методы, градиентные методы, метод наискорейшего спуска, метод Ньютона - Канторовича и пр.

4°. Предположим, что аппроксимирующую задачу (I.I.6) мы "умеем решать" при любом n. Под этим мы понимаем следующее. Существует набор "элементарных" операций, которые мы умеем выполнять, и мы располагаем алгоритмом, который при любом n после выполнения конечного числа "элементарных" операций приводит (если эти операции выполнены точно, без погрешности) либо к точному значению элемента $v_*^{(n)}$, решающего задачу (I.I.6), либо к

некоторому элементу $\tilde{v}^{(n)}$, который в том или ином смысле близок к элементу $v_*^{(n)}$, причем степень этой близости можно оценить. Разумеется, мы предполагаем при этом, что для рассматриваемых n аппроксимирующая задача разрешима.

Поясним сказанное на примере. Пусть задаче (I.I.I) имеет вид

$$Ax = f, \qquad (I.I.10)$$

где A - положительно определенный оператор в некотором гильбертовом пространстве H, x и f - элементы этого пространства, искомый и данный. Если к этой задаче применить какой-нибудь вариант метода Ритца (например, метод конечных элементов), то мы придем к аппроксимирующей линейной алгебраической задаче вида

$$A_n v^{(n)} = f^{(n)}, \qquad (I.I.11)$$

где A_n - числовая матрица порядка $N \times N$, $v^{(n)}$ и $f^{(n)}$ - N-мерные числовые векторы; N - некоторая целочисленная функция от n. Эта задача разрешима при любом n. Если для ее решения воспользоваться алгоритмом Гаусса, то через конечное число шагов мы придем к точному решению $v_*^{(n)}$; если же воспользоваться, например, алгоритмом простой итерации (для этого задачу (I.I.II) надо предварительно подвергнуть некоторому элементарному преобразованию), то в результате конечного числа шагов мы получим некоторый вектор $\tilde{v}^{(n)}$, близкий к вектору $v_*^{(n)}$. Набор элементарных операций, о котором говорилось выше, в обоих случаях содержит арифметические действия; в случае метода Гаусса в этот набор входит еще определение наибольшего числа из заданного конечного множества чисел.

5°. Для приложений наиболее интересен случай, когда X и X_n суть банаховы пространства. Предположим, еще, что f и $f^{(n)}$ также суть элементы некоторых банаховых пространств Y и Y_n соответственно. Эти предположения мы сохраним на протяжении всей книги.

§ 2. Источники погрешностей.

I°. Если задача (I.I.I) решается каким-нибудь методом, подходящим под схему § I, то, как кажется автору, речь может идти о четырех возможных источниках погрешностей. При этом мы считаем, что задача (I.I.I) поставлена точно. Тем самым мы пренебрегаем возможными погрешностями, связанными с постановкой упомянутой задачи как задачи естествознания или социально-научных

дисциплин (погрешности измерений, недостаточно точности основных гипотез и т.п.). По мнению автора, устранение или хотя бы уменьшение подобных погрешностей лежат вне пределов действия математических методов, и самое большее, на что здесь можно рассчитывать, это, при наличии достаточной информации о степени точности упомянутых выше факторов, воспользоваться методами теории возмущений и дать оценку погрешности, обусловленной неточностью этих факторов.

2°. Переходим к описанию источников погрешностей.

1. Погрешность аппроксимации. Переход от точно поставленной задачи (I.I.I) к аппроксимирующей задаче (I.I.6) с выбранным номером n приводит к тому, что элемент x_* пространства X заменяется элементом $x_*^{(n)} = P_n U_*^{(n)}$ того же пространства. Возникающую при этом погрешность $x_* - P_n U_*^{(n)}$ естественно оценить ее нормой

$$\rho_n = \| x_* - P_n U_*^{(n)} \| = \| x_* - x_*^{(n)} \| . \qquad (I.2.I)$$

Величину ρ_n будем называть погрешностью аппроксимации задачи (I.I.I).

Вопрос об оценке погрешности аппроксимации - один из важнейших в вычислительной математике, и ему посвящена обширная литература.

2. Погрешность искажения. Как правило, переход от точной задачи (I.I.I) к аппроксимирующей задаче (I.I.6) совершается с погрешностью. Элементы $f^{(n)}$ не заданы заранее, а вычисляются по данным точной задачи. Вычислительные операции почти всегда выполняются с той или иной погрешностью, и вместо элементов $f^{(n)}$ мы на самом деле получаем некоторые "искаженные" элементы $\tilde{f}^{(n)}$. Может случиться, что и соотношения τ_n также заменяются "искаженными" соотношениями $\tilde{\tau}_n$. Элементы $f^{(n)}$ и соотношения τ_n на самом деле остаются неизвестными; известными являются $\tilde{f}^{(n)}$ и $\tilde{\tau}_n$. В результате вместо последовательности аппроксимирующих задач (I.I.6) получится последовательность "искаженных аппроксимирующих задач"

$$n \geq 1 . \quad \{ z^{(n)} \in X_n , \ \tilde{z}^{(n)} \tilde{\tau}_n \tilde{f}^{(n)} \} . \qquad (I.2.2)$$

Допустим, что эти задачи разрешимы, и пусть $\tilde{z}_*^{(n)}$ их решения. Величину

$$\hat{\xi}_n = \| z_*^{(n)} - u_*^{(n)} \|_{Y_n} \qquad (I.2.3)$$

назовем " Y_n - погрешностью искажения". Больший интерес представляет величина

$$\xi_n = \| p_n z_*^{(n)} - p_n u_*^{(n)} \|_Y , \qquad (I.2.4)$$

которую мы назовем Y -погрешностью, или просто погрешностью искажения.

3. Искаженную задачу (I.2.2) предстоит решить. Будем считать (см. § I), что мы располагаем некоторым алгоритмом, который после выполнения конечного числа действий приводит, если эти действия выполнены без погрешности, либо к точному значению элемента $z_*^{(n)}$, либо к такому элементу $w^{(n)} \in Y_n$, что погрешность

$$\eta_n = \| p_n w^{(n)} - p_n z_*^{(n)} \|_Y \qquad (I.2.5)$$

можно оценить и можно добиться того, чтобы величина, оценивающая эту погрешность, была сколь угодно малой. Первый случай - частный по отношению ко второму, и мы будем рассматривать именно этот второй случай, допуская, что величина (I.2.5) может иногда равняться нулю. Эту величину мы назовем Y-погрешностью, или просто погрешностью алгоритма; можно ввести и " Y_n-погрешность алгоритма"

$$\hat{\eta}_n = \| w^{(n)} - z_*^{(n)} \|_{Y_n} . \qquad (I.2.6)$$

Отметим, что погрешность алгоритма можно рассматривать как погрешность аппроксимации для задачи (I.2.2).

4. Практически любой алгоритм содержит ряд операций, выполнение которых сопряжено с погрешностями. Таковы арифметические действия, операции вычисления значений функций по таблицам или по приближенным формулам и т.п. Названные погрешности приведут к тому, что вместо элемента $w^{(n)}$ мы получим другой элемент $\tilde{w}^{(n)} \in Y_n$. Величину

$$\hat{\zeta}_n = \| \tilde{w}^{(n)} - w^{(n)} \|_{Y_n} \qquad (I.2.7)$$

назовем " Y_n-погрешностью округления", а величину

$$\zeta_n = \| P_n \tilde{w}^{(n)} - P_n w^{(n)} \|_X \qquad (I.2.8)$$

- " X-погрешностью округления", или просто "погрешностью округления"

3°. В конечном счете вместо искомого элемента $x_* \in X$ мы вычислим элемент $\tilde{w}^{(n)} \in X_n$. За приближенное решение задачи (I.1.1) естественно принять элемент $P_n \tilde{w}^{(n)}$; его погрешность

$$\delta_n = \| x_* - P_n \tilde{w}^{(n)} \| \leq \rho_n + \xi_n + \eta_n + \zeta_n . \qquad (I.2.9)$$

4°. Приведем некоторые литературные указания. Вопросы оценки погрешности аппроксимации возникли почти одновременно с возникновением анализа. Достаточно указать, например, на оценки остаточного члена степенного ряда. Из многочисленных работ более близких лет, в которых изучается погрешность аппроксимации, мы отметим здесь монографии Вязова и Форсайта [I], Вайнико [IO]. Варга[I], Габдулхаева [3] , Гавурина[I], Канторовича и Акилова[I], Канторовича и Крылова[I], Красносельского и др. [I], Марчука[I], автора этой книги[3, 22] , Обэна[I], Оганесяна и Руховца[I], Прессдорфа и Зильбермана[I], Соболева [2] , Стренга и Фикса [I], Сьярле [I]. Понятие погрешности искажения было введено первоначально в работах автора [4, 5, 7] ; в первых двух исследована устойчивость классического метода Ритца по отношению к погрешностям искажения, в третьей получены некоторые общие теоремы об устойчивости вычислительных процессов, которые выше были названы свободными. Самый термин "искажение" ввела Л.Н.Довбыш в работе [I] , посвященной погрешностям искажения собственных чисел и собственных подпространств самосопряженных операторов. В работе Лсковой и Яковлева [I] исследована устойчивость относительно погрешностей искажения для метода Бубнова - Галеркина. Велиев [I - 9] исследовал устойчивость метода Бубнова - Галеркина для ряда нестационарных задач, линейных и нелинейных. Перечисленные здесь работы, опубликованные не позднее 1965 г., суммированы в книге автора [8], в которой кроме того содержатся некоторые результаты того же рода, но относящиеся к нелинейным задачам. Дальнейшие результаты по погрешностям искажения свободных вычислительных процессов, линейных и нелинейных, содержатся в работах автора [I2, 22] . Дем-

яновича [I] , Тинкера [I], Омодеи [I], Суворова [I]. Тот же вопрос для рекуррентных вычислительных процессов рассмотрен в работах автора [33, 34] .

Вопрос об оценке погрешности алгоритма в большинстве случаев достаточно прост и, насколько известно автору, специально не освещался.

Много работ посвящено погрешностям округления; эти работы относятся главным образом к различным методам решения линейных алгебраических систем. Отметим здесь работы Голуба [I] , Уилкинсона [I, 2] и Воеводина [I] . Некоторые другие задачи рассмотрены автором [33, 34].

В заключение отметим, что некоторая классификация погрешностей предложена в учебнике Березина и Жидкова [I] . Эти авторы различают три типа погрешностей: I) неустранимая погрешность, вытекающая из неточностей постановки задачи (I.I.I); 2) погрешность метода - она совпадает с определенной выше погрешностью аппроксимации; 3) вычислительная погрешность - она включает в себя введенные выше погрешности искажения, алгоритма и округления.

Соображения автора относительно неустранимой погрешности изложены (без применения самого термина) выше, в п.I° настоящего параграфа. Что касается вычислительной погрешности, то нам хотелось бы привести некоторые доводы в пользу нашей более подробной классификации. Погрешность искажения зависит лишь от того, с какой точностью вычислены объекты, которые являются данными в приближенной задаче (I.2.2); она не зависит от того, какой алгоритм применен для решения этой последней задачи и с какой точностью выполнены предписания этого алгоритма. От его выбора зависит погрешность алгоритма. Наконец, погрешность округления полностью определяется при выбранном алгоритме той точностью, с которой выполняются предписания алгоритма. Таким образом, вычислительная погрешность Березина и Жидкова распадается на три слагаемых, каждое из которых имеет свой источник и оценивается независимо.

Наша классификация дана в работе автора [33] .

§. Погрешности квадратурных формул.

Чтобы проиллюстрировать общие соображения § 2, мы проанализируем в настоящем параграфе погрешност. обычных квадратурных

формул.

Погрешность аппроксимации для этих формул изучалась с давних времен. Мы приведем здесь оценки этой погрешности для наиболее распространенных квадратурных формул. Через (a, b) обозначен интервал интегрирования, через n - число промежутков разбиения, через f - подынтегральная функция.

1) Формула прямоугольников

$$\rho_n \le \frac{(b-a)^2}{2n} \|f'\|_{C[a,b]} ;$$

2) формула средних прямоугольников

$$\rho_n \le \frac{(b-a)^3}{24n^2} \|f''\|_{C[a,b]} ;$$

3) формула трапеций

$$\rho_n \le \frac{(b-a)^3}{12n^2} \|f''\|_{C[a,b]} ;$$

4) формула Симпсона

$$\rho_n \le \frac{(b-a)^5}{180n^4} \|f^{(4)}\|_{C[a,b]} ;$$

5) формула Гаусса, $(a, b) = (-1, 1)$,

$$\rho_n \le \frac{2^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3} \|f^{(2n)}\|_{C[-1,1]} ;$$

6) формула Эйлера - Маклорена

$$\rho_n \le \frac{nh^{2m+1}B_{2m}}{(2m)!} \|f^{(2m)}\|_{C[a,b]} ;$$

в последней формуле m - произвольно выбранное натуральное число, от которого зависит точность формулы Эйлера - Маклорена, $h=(b-a)/n$, B_{2m} - числа Бернулли.

Отметим еще, что в книге Соболева [1] построены кубатурные формулы, которые для подынтегральных функций класса $W_2^{(m)}(\Omega)$ (Ω - область интегрирования) приводят к наименьшей или близкой к наименьшей погрешности аппроксимации.

Исследование остальных погрешностей мы предлагаем для класса квадратурных формул вида

$$\int_a^b f(t)\,dt \approx \sum_{k=1}^{n} c_k f(t_k), \qquad (1.3.1)$$

обладающих следующими свойствами: $C_k \geqslant 0$; формула (I.3.I) точна, если $f(t) = const$, так что

$$\sum_{k=1}^{n} C_k = b-a. \qquad (I.3.2)$$

На узлы t_k мы не накладываем никаких ограничений. Подынтегральную функцию предполагаем ограниченной: $|f(t)| \leqslant M = const$. Пусть коэффициенты C_k и ординаты $y_k = f(t_k)$ вычислены неточно, так что на самом деле нам известны их искаженные значения $\tilde{C}_k = C_k + \gamma_k$ и $\tilde{y}_k = y_k + \alpha_k$. Мы предполагаем при этом, что

$$|\alpha_k| \leqslant \alpha, \quad \sum_{k=1}^{n} |\gamma_k| \leqslant \gamma, \qquad (I.3.3)$$

где α и γ – малые числа, и что искаженные коэффициенты $\tilde{C}_k$ удовлетворяют условиям, сформулированным выше:

$$\tilde{C}_k \geqslant 0 \ , \ \sum_{k=1}^{n} \tilde{C}_k = b-a.$$

Погрешность искажения формулы (I.3.I) равна

$$\xi_n = \Big| \sum_{k=1}^{n} [(C_k+\gamma_k)(y_k+\alpha_k) - C_k y_k] \Big| =$$

$$= \Big| \sum_{k=1}^{n} (\gamma_k y_k + \tilde{C}_k \alpha_k) \Big| \leqslant M\gamma + \alpha .$$

Таким образом, малые искажения коэффициентов и ординат приводят к малому искажению интеграла, вычисленного по квадратурной формуле.

После того, как ординаты вычислены, алгоритм вычисления правой части формулы (I.3.I) содержит только конечное число действий сложения и умножения, поэтому для квадратурной формулы погрешность алгоритма равна нулю.

Оценим, наконец, погрешность округления. Интеграл $\int_a^b f(t)\,dt$ мы вычисляем по искаженной квадратурной формуле

$$\int_a^b f(t)\,dt \approx \sum_{k=1}^{n} \tilde{C}_k \tilde{y}_k . \qquad (I.3.4)$$

Допустим, что каждое из произведений $\tilde{C}_k \tilde{y}_k$ вычисляется с погрешностью, не превосходящей числа δ^{-p} , где δ – основание системы счисления, p – некоторое положительное число. Тогда погрешность округления формулы (I.3.4) не превосходит числа $n\delta^{-p}$.

Если желательно, чтобы эта погрешность была не больше заданного числа ε, то достаточно подчинить p неравенству

$$p \geq \frac{\ln(n/\varepsilon)}{\ln 3}. \qquad (I.3.5)$$

Нетрудно проанализировать также погрешности квадратурных формул типа формулы Эйлера – Маклорена, отличающихся наличием некоторого фиксированного числа отрицательных коэффициентов. Не вносит новых трудностей и исследование погрешностей кубатурных формул.

§ 4. Погрешности решения системы линейных алгебраических уравнений.

I°. Пусть дана система n линейных алгебраических уравнений с n неизвестными, которая в матричной форме записывается в виде

$$Ax = f. \qquad (I.4.1)$$

Искомый вектор x и данный вектор f будем рассматривать как элементы пространства R_n, а матрицу A – как оператор, действующий в том же пространстве. Допустим, что этот оператор обратим, так что $\|A^{-1}\| < \infty$.

Будем решать систему (I.4.I) по методу Гаусса с применением схемы единственного деления по главным элементам (см. Фаддеев и Фаддеева [I], § I6). Метод Гаусса состоит из прямого и обратного хода. Прямым ходом данная система преобразуется в систему с треугольной матрицей, главная диагональ которой состоит из единиц; обратный ход заключается в решении новой, рекуррентной системы. С точки зрения понятий, изложенных в § I, прямой ход можно рассматривать как преобразование задачи (I.I.I) в задачу (I.I.6); при этом $X_n = X = R_n$, $x^{(n)} = x$, $x \vee f := Ax = f$, $x^{(n)} \vee_n f^{(n)} := Lx = \tilde{f}$, где L – упомянутая выше треугольная матрица, а $\tilde{f}$ – столбец новых свободных членов, полученных в результате прямого хода. В данном случае погрешность аппроксимации равна нулю, но возникает погрешность искажения. Исследуем ее.

Действия, описанные ниже, требуют, чтобы делители были отличны от нуля, а это, в свою очередь, требует перенумерации уравнений и неизвестных. Будем считать, что перед каждым шагом такая перенумерация выполнена, и не будем этого особо оговаривать

вать.

Выпишем первое уравнение системы

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x^{(n)} = f_1 .$$

Разделим его на a_{11}. Если бы деление было выполнено точно, мы получили бы уравнение

$$x_1 + b_{12}^{(1)}x_2 + \cdots + b_{1n}^{(1)}x_n = \tilde{f}_1^{(1)} , \qquad (I.4.2)$$

где $b_{1k}^{(1)} = a_{1k}/a_{11}$, $\tilde{f}_1^{(1)} = f_1/a_{11}$. Однако, в общем случае деление совершается с погрешностью, и мы придем к уравнению вида

$$x_1 + (b_{12}^{(1)} + \gamma_{12})x_2 + \cdots + (b_{1n}^{(1)} + \gamma_{1n})x_n = \tilde{f}_1 + \delta_1 .$$

Следующий шаг заключается в исключении неизвестной x_1 из остальных $(n-1)$ уравнений. Это приведет нас к системе уравнений

$$\sum_{k=2}^{n}(b_{jk}^{(2)} + \gamma_{jk}^{(2)})x_k = \tilde{f}_j^{(2)} + \delta_j^{(2)} ; \quad 2 \le j \le n . \qquad (I.4.3)$$

Здесь через $b_{jk}^{(2)}$ и $f_j^{(2)}$ обозначены числа, которые получились бы если бы все описанные выше действия были выполнены точно. Продолжая тот же процесс исключения неизвестных, получим систему вида

$$\sum_{k=j}^{n}(b_{jk} + \gamma_{jk})z_k = \tilde{f}_j + \delta_j ; \quad 1 \le j \le n$$

или, короче

$$(L + \Gamma)z = \tilde{f} + \delta . \qquad (I.4.4)$$

В системе (I.4.4) неизвестный вектор обозначен через z вместо x, так как из-за погрешностей Γ и δ системы (I.4.1) и (I.4.4) не равносильны, вообще говоря, и их решения, как правило, не совпадают.

Оценим погрешность $\|z - x\|_{R_n} := \|z - x\|$

$$\|z - x\| = \|(L+\Gamma)^{-1}(\tilde{f} + \delta) - L^{-1}f\| \le$$
$$\le \|(L+\Gamma)^{-1}\tilde{f} - L^{-1}\tilde{f}\| + \|(L+\Gamma)^{-1}\delta\| .$$

Первый член справа оценивается так: если $\|L^{-1}\|\,\|\Gamma\| \le \beta < 1$, то

$$\|[(I + L^{-1}\Gamma)^{-1} - I]L^{-1}\tilde{f}\| \le \|(I + L^{-1}\Gamma)^{-1} - I\|\,\|x\| \le \frac{\|L^{-1}\| \cdot \|\Gamma\| \cdot \|x\|}{1 - \beta} .$$

Сходным образом получаем

$$\|(L+\Gamma)^{-1}\delta\| = \|(I+L^{-1}\Gamma)L^{-1}\delta\| \leqslant \frac{\|L^{-1}\|\cdot\|\delta\|}{1-\beta},$$

и окончательно мы приходим к следующей оценке погрешности искажения:

$$\|x_* - \tilde{x}_*\|_{R_N} \leqslant \frac{\|L^{-1}\|}{1-\beta}\left[\|\delta\| + \|\Gamma\|\cdot\|x\|\right]. \qquad (I.4.5)$$

2^0. Погрешности округления окажутся на обратном ходе. Если обозначить $L+\Gamma=\tilde{L}$, $f+\delta=\tilde{f}$, то нам предстоит решить рекуррентную систему

$$\begin{aligned}
&\tilde{x}_n = \tilde{f}_n\,,\\
&\tilde{x}_{n-1} - c_{n-1,n}\tilde{x}_n = \tilde{f}_{n-1}\,,\\
&\tilde{x}_{n-2} - c_{n-2,n-1}\tilde{x}_{n-1} - c_{n-2}\,, \qquad \text{и} \quad \tilde{x}_n = \tilde{f}_{n-2}\,, \qquad (I.4.6)\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&\tilde{x}_1 - c_{12}\tilde{x}_2 \;\ldots\; - c_{1n}\tilde{x}_n = \tilde{f}_1\,.
\end{aligned}$$

Первое уравнение не вносит погрешности округления. Из остальных уравнений (подробнее эти вопросы будут обсуждены в гл. IУ) мы вычислим, в результате ошибок округления, не величины $\tilde{x}_k$, $k = n-1,\ n-2,\ \ldots,\ 1$, а некоторые другие величины w_k. Пусть уже вычислены величины $w_n = \tilde{x}_n$, $w_{n-1}, \ldots, w_{k+1}$; каждая из них дает значение соответствующей величины $\tilde{x}_n, \tilde{x}_{n-1}, \ldots, \tilde{x}_{k+1}$ со своей погрешностью γ_i, так что $w_i = \tilde{x}_i + \gamma_i$, $k+1 \leqslant i \leqslant n$; отметим, что $\gamma_n = 0$. Следующее уравнение даст нам величину

$$w_k = \tilde{f}_k + \sum_{i=k+1}^{n} c_{ki}\, w_i + d_k\,;$$

здесь d_k – погрешность округления, допущенная при вычислении числа w_k. Из последнего равенства следует, что $w_k = \tilde{x}_k + \gamma_k$, где

$$\gamma_{k1} = \sum_{i=k+1}^{n} c_{ki}\,\gamma_i + d_k\,. \qquad (I.4.7)$$

Допустим для определенности, что $|d_k| \leqslant \varepsilon$, где ε – заранее заданное положительное число, которое оценивает, таким образом, абсолютную погрешность числа w_k. Решение системы (I.4.7), в которой γ_k рассматриваются как неизвестные, имеет очередную

оценку

$$|r_k| \leq C_0 \max_{j \geq k+1} |d_j| \leq C_0 \varepsilon ,$$

в которой C_0 есть норма обратной матрицы системы (I.4.6), рассматриваемой как оператор в пространстве n-мерных векторов, в котором норма определяется по формуле $\|a\| = \max_{1 \leq k \leq n} |a_k|$. Таким образом, если $r = (r_1, r_2, \ldots, r_n)$ есть вектор погрешности округления, то

$$\|r_n\|_{R_n} \leq C_0 \sqrt{n}\, \varepsilon . \qquad (I.4.8)$$

3°. Пусть $A = I - B$ и $\|B\|_{R_n \to R_n} = q < 1$. Тогда уравнение (I.4.1) можно решать итерациями по формуле

$$x^{(k)} = Bx^{(k-1)} + f \quad , \quad k = 1, 2, \ldots ; \qquad (I.4.9)$$

начальное приближение $x^{(0)}$ выбирается по произволу. Примем, что $x^{(0)} = 0$. В данном случае погрешности аппроксимации и искажения отсутствуют; погрешность алгоритма, если выполнены N итераций равна

$$\|x_* - x^{(N)}\|_{R_n} = \Big\| \sum_{k=N+1}^{\infty} B^k f \Big\|_{R_n} \leq \frac{q^{N+1}}{1-q} \|f\|_{R_n} .$$

Оценим погрешность округления. Как и в п.2°, ошибки округления приведут к тому, что вместо векторов $x^{(k)}$ мы получим некоторые другие векторы $w^{(k)}$. Обозначим через d_k результат ошибок округления на k-м шаге, так что

$$w^{(k)} = f + Bw^{(k-1)} + d_k . \qquad (I.4.10)$$

Допустим на этот раз, что вычисления производятся с заданной относительной погрешностью, так что $\|d_k\|_{R_n} \leq \varepsilon \|w^{(k)}\|_{R_n}$. Точные значения векторов $x^{(k)}$ суть

$$x^{(k)} = f + Bx^{(k-1)} .$$

Фактически мы получим последовательность векторов

$$w^{(0)} = 0 \ , \ w^{(1)} = f \ , \ w^{(2)} = f + Bx^{(1)} + d_2 = x^{(2)} + d_2$$

и т.д. Положим $r_0 = r_1 = 0$, $r_2 = d_2$. Далее, $w^{(3)} = f + Bw^{(2)} + d_3 = f + Bx^{(2)} + Bd_2 + d_3 = x^{(3)} + Br_2 + d_3 := x^{(3)} + r_3$.

Пусть для некоторого K оказалось $w^{(K-1)} = x^{(K-1)} + r_{K-1}$. Тогда

$$w^{(K)} = f + Bw^{(K-1)} + d_K = f + Bx^{(K-1)} + Br_{K-1} + d_K ,$$

что можно представить в виде $w^{(K)} = x^{(K)} + r_K$, где

$$r_K = B_{K-1} + d_K . \qquad (I.4.II)$$

Отсюда

$$\|r_K\| \le q\|r_{K-1}\| + \|d_K\| \le q\|r_{K-1}\| + \varepsilon\|r_K\| + \varepsilon\|x^{(K)}\| . \qquad (I.4.I2)$$

Но

$$\|x^{(K)}\| = \|f + Bf + \cdots + B^{K-1}f\| \le \frac{1}{1-q}\|f\| ,$$

и из (I.4.I2) находим

$$\|r_K\| \le q_1\|r_{K-1}\| + \varepsilon_1 ;\ q_1 = \frac{q}{1-\varepsilon} ,\ \varepsilon_1 = \frac{\varepsilon\|f\|}{(1-\varepsilon)(1-q)} .$$

Ясно, что $\|r_K\| \le \tau_K$, где τ_K есть решение разностного уравнения $\tau_K = q_1\tau_{K-1} + \varepsilon_1$, $\tau_0 = 0$; это решение есть

$$\tau_K = \frac{\varepsilon_1}{1-q_1}(1-q_1^K)$$

и погрешность округления после N итераций имеет оценку

$$\|r_N\| \le \frac{\varepsilon_1}{1-q_1}(1-q_1^N) < \frac{\varepsilon\|f\|}{(1-q_1)(1-q)(1-\varepsilon)} . \qquad (I.4.I3)$$

4°. Выше мы предполагали, что матрица A и столбец f заданы точно. Однако в практике применения приближенных методов мы чаще всего встречаемся с такой ситуацией, когда указанные величины известны с некоторыми погрешностями, так что решать приходится искаженную систему вида

$$(A + \Gamma)z = f + \delta .$$

Мы вправе считать, что нормы $\|\Gamma\|$ и $\|\delta\|$ достаточно малы; в частности, будем считать, что $\gamma = \|A^{-1}\| \cdot \|\Gamma\| < 1$. Оценим погрешность искажения $\rho = \|z - x\|_{Q_H}$. Имеем

$$\tilde{x} - x = [(A+\Gamma)^{-1} - A^{-1}]f + (A+\Gamma)^{-1}\delta .$$

Но $(A+\Gamma)^{-1} = (I + A^{-1}\Gamma)^{-1}A$ и $(A+\Gamma)^{-1} - A^{-1} =$ $= -(I + A^{-1}\Gamma)^{-1}A^{-1}\Gamma A^{-1}$, поэтому

$$\tilde{x} - x = -(I+A^{-1}\Gamma)^{-1}A^{-1}\Gamma x + (I+A^{-1}\Gamma)^{-1}A^{-1}\delta .$$

Далее, $\|(I+A^{-1}\Gamma)^{-1}\| \le (1-\tau)^{-1}$ и

$$\|\tilde{x} - x\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\Gamma\|} \left[\|\Gamma\| \|x\| + \|\delta\|\right] . \qquad (I.4.I4)$$

Отсюда легко вытекает неравенство

$$\frac{\|\tilde{x} - x\|}{\|x\|} \le \frac{P_A}{1 - P_A \frac{\|\Gamma\|}{\|A\|}} \left[\frac{\|\Gamma\|}{\|A\|} + \frac{\|\delta\|}{\|f\|}\right] ; \qquad (I.4.I5)$$

здесь $P_A = \|A\| \cdot \|A^{-1}\|$ – число обусловленности матрицы A. Формула (I.4.I5) хорошо известна (см., например, Гавурин [I]). Она дает оценку относительной погрешности решения через число обусловленности матрицы A и относительные погрешности искажения $\|\Gamma\| / \|A\|$ и $\|\delta\| / \|f\|$. Оценка (I.4.I5) очевидным образом переносится на произвольные уравнения вида (2.4.I), в которых A – произвольный оператор, действующий из одного банахова пространства в другое, при ус овии, что оба оператора A и A^{-1} ограничены. То же ве.о и для оценки (I.4.I4), с той важной оговоркой, что ограниченность оператора A здесь необязательна.

ГЛАВА II

Погрешность аппроксимации

На протяжении всей книги мы будем пользоваться следующими обозначениями и определениями. Норму в $W_2^{(s)}(\Omega)$ будем обозначать через $\|\cdot\|_{(s),\Omega}$ или, если это не может вызвать недоразумений, через $\|\cdot\|_{(s)}$; сокращения $\|\cdot\|_s$ или $\|\cdot\|_{s,\Omega}$ мы сохраним для нормы в $L_2(\Omega)$.

Степенью одночлена $x_1^{\alpha_1} x_2^{\alpha_2} \dots x_m^{\alpha_m}$ в R_m назовём мультииндекс $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)$; степенью многочлена в R_m назовём верхнюю грань степеней составляющих его одночленов. Мультииндекс, все составляющие которого равны одному и тому же числу α , будем обозначать через α . Таким образом, если $P(t)$ - многочлен в R_m , и по каждой из координат его степень не превосходит некоторого числа n , то степень $P(t)$ в R_m не превосходит мультииндекса n .

Буквой C без индексов будем обозначать постоянные, вообще говоря, различные, точные значения которых для нас несущественны.

§ 1. Классический метод Ритца. Оценки в энергетической норме.

1°. Напомним основы метода Ритца для линейных задач; подробно этот метод изложен в книге автора [3] . Пусть симметричный билинейный функционал $A(x,y)$ и соответствующий ему однородный квадратичный функционал $A(x) = A(x,x)$ определены на некотором линейном множестве M . Допустим, что функционал $A(x)$ положителен: это значит, что $A(x) > 0$, если x отличен от нулевого элемента множества M . Превратим M в предгильбертово пространство, введя скалярное произведение $[x,y] = A(x,y)$ и норму $|x| = \sqrt{[x,x]} = \sqrt{A(x)}$. Замкнув M в этой норме, получим гильбертово пространство, которое обозначим через H_A . Для упрощения последующих записей примем, что пространство H_A - вещественное. В приложениях часто встречается случай, когда M есть линейное плотное множество некоторого гильбертова /реже банахова/ пространства H , и $A(x,y) = (Ax,y)$, где A положительный /и симметричный - последняя оговорка необходима, если H - вещественное пространство/ оператор в H . В этом случае мы называем H_A энергетическим пространством оператора A , а скалярное произведение и норму в

H_A - соответственно энергетическим произведением и нормой.

Пусть $f(x)$ - линейный функционал, ограниченный в H_A и определенный на всем этом пространстве. Поставим задачу о минимуме функционала

$$\mathcal{F}(x) = A(x) - 2f(x). \qquad (\text{II.1.1})$$

Эта задача теоретически решается элементарно. По известной теореме Ф.Рисса, в H_A существует один и только один элемент x_*, такой, что $f(x) = [x, x_*]$, $\forall x \in H_A$. Теперь $\mathcal{F}(x) = |x|^2 - 2[x, x_*] = |x - x_*|^2 - |x_*|^2$, и ясно, что функционал $\mathcal{F}(x)$ достигает минимума при $x = x_*$.

2^0. Приближенно можно решать задачу (II.1.1), например, по классическому методу Ритца. Выбираем в H_A какое-нибудь конечномерное подпространство $H_A^{(n)}$ и ищем минимум функционала $\mathcal{F}(x)$, или, что равносильно, минимум нормы $|x - x_*|$, на этом подпространстве. Элемент $x_*^{(n)} \in H_A^{(n)}$, на котором последний минимум достигается, существует и единственен. Очевидно, $x_*^{(n)}$ есть ортогональная проекция в H_A элемента x_* - точного решения задачи (II.1.1) - на подпространство $H_A^{(n)}$. Сравнивая это с общей схемой § I гл.I, видим, что в данном случае $X = H_A$, $X_n = H_A^{(n)}$, P_n есть тождественный оператор, соотношения τ и τ_n определяются формулами

$$x\tau f := |x|^2 - 2f(x) = \min, \quad x \in H_A;$$

$$x^{(n)} \tau_n f^{(n)} := |x^{(n)}|^2 - 2f^{(n)}(x^{(n)}) = \min, \quad x^{(n)} \in H_A^{(n)}$$

и, наконец, $f^{(n)}$ есть сужение f на подпространство $H_A^{(n)}$. Отсюда ясно, что погрешность аппроксимации приближенного решения по Ритцу для задачи (II.1.1), равная $\rho_n = |x_* - x_*^{(n)}| = \min\limits_{x^{(n)} \in H_A^{(n)}} |x_* - x^{(n)}|$ совпадает с наилучшим приближением (в энергетической норме) точного решения x_* элементами из пространства $H_A^{(n)}$. Величина ρ_n зависит от характера энергетической нормы и, в ряде важных случаев, от гладкости элемента x_*; при исследовании последней полезно иметь в виду, что x_* есть решение уравнения Эйлера $\mathrm{grad}\,\mathcal{F} = 0$.

В ближайших параграфах настоящей главы мы применим высказанные выше общие соображения к краевым задачам для дифференциальных уравнений.

Разумеется, применять метод Ритца целесообразно только если есть уверенность, что $|x_*^{(n)} - x_*| \xrightarrow[n \to \infty]{} 0$. Это последнее соотно-

гение будет, очевидно, выполнено, если последовательность подпространств $\{H_A^{(n)}\}$ полна в H_A, иначе говоря, если по любому $\varepsilon>0$ можно найти такое n_0, что при любом $n \geq n_0$ существует элемент $x^{(n)} \in H_A^{(n)}$, удовлетворяющий неравенству $|x_* - x^{(n)}| < \varepsilon$.

В основополагающей работе Ритца [1] подпространства $H_A^{(n)}$ строятся так, что при возрастании n они расширяются. Более определенно, Ритц выбирает последовательность элементов $\{\varphi_n\} \in H_A$, обладающую полнотой в H_A; предполагается еще, что эти элементы, взятые в любом конечном числе, линейно независимы. За $H_A^{(n)}$ Ритц принимал подпространство с базисом $(\varphi_1, \varphi_2, \ldots, \varphi_n)$. Более поздние авторы назвали систему $\{\varphi_n\}$ координатной. Отметим, что Ритц не вводил энергетического пространства и требовал сходимости приближенных решений к точному в $C^{(1)}$ при подходящем выборе δ. Это ограничивало применение его метода и осложняло рассуждения. Энергетическая метрика была введена К.Фридрихсом; она была применена автором этой книги для исследования методов Ритца и Бубнова – Галеркина. Подробно об этом см. книгу автора [3].

Курант показал [1], что необязательно брать расширяющиеся подпространства $H_A^{(n)}$, и предложил такое видоизменение метода Ритца: пространства $H_A^{(n)}$ суть произвольные конечномерные подпространства энергетического пространства, подчиненные единственному требованию полноты их последовательности в H_A. Если выбрать в $H_A^{(n)}$ базис

$$(\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN}), \qquad (\text{II.I.2})$$

где $N = N(n)$ есть размерность $H_A^{(n)}$, то элементы φ_{nk} необязательно принадлежат подпространству $H_A^{(n')}$, $n'>n$; таким образом, может случиться, что $H_A^{(n)}$ и $H_A^{(n')}$ имеют общим только нулевой элемент.

В случае расширяющихся подпространств будем говорить о классическом методе Ритца, в общем случае – об обобщенном методе Ритца. Классический метод Ритца имеет то преимущество, что погрешность $|x_* - x_*^{(n)}|$ убывает с ростом n (см. книгу автора [3]); следует, однако, напомнить, что обобщенный метод Ритца является идейной основой метода конечных элементов – одного из самых распространенных в настоящее время методов решения краевых задач.

В заключение напомним, что если в $H_A^{(n)}$ выбран базис (II.I.2), то приближенное решение $x_*^{(n)}$ имеет вид

$$x_*^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk}, \qquad (\text{II.I.3})$$

причем коэффициенты $\alpha_k^{(n)}$ определяются из алгебраической системы

$$\sum_{k=1}^{N} [\varphi_{nk}, \varphi_{nj}]\alpha_k^{(n)} = f(\varphi_{nj}), \quad j=1,2,\dots,N. \quad (\text{II.I.4})$$

Последовательность базисов

$$(\varphi_{n1}, \varphi_{n2}, \dots, \varphi_{nN}), \quad n=1,2,\dots \quad (\text{II.I.5})$$

будем называть координатной системой обобщенного метода Ритца, а элементы φ_{nk} – координатными элементами этого метода.

§ 2. Некоторые обобщения теоремы Джексона на функции многих переменных.

1°. Наиболее важное, как нам кажется, обобщение теоремы Джексона на функции многих переменных получила Харрик [1 – 3]. В ее работах [1,2] рассмотрены более частные случаи; общая теорема доказана в работе [3]. Здесь мы сформулируем основной результат этой последней статьи.

Пусть Ω – ограниченная область в R_m, τ и $\delta<\tau$ – натуральные числа. Далее, пусть $\varphi(t)$ – функция класса $C^{(\nu)}$ в некоторой области $\Omega_1 \supset \bar{\Omega}$, причем ν достаточно велико, Ω_1 гомеоморфна шару, $\varphi|_{\partial\Omega}=0$, $grad\,\varphi|_{\partial\Omega}\neq 0$ и $\varphi(t)\neq 0$, $t\notin\partial\Omega$.

Теорема II.2.1 (Харрик [3]). Пусть $x\in C^{(\tau)}(\bar{\Omega})$ и $\mathcal{D}^\alpha x|_{\partial\Omega}=0$, $0\le|\alpha|\le\delta-1$. Тогда при любом натуральном n существует полином $P_n(t)$ степени $\le n$ в R_m, удовлетворяющий неравенству

$$\|x-\varphi^\delta P_n\|_{C^{(\bar\tau)}(\bar\Omega)} \le C\,\omega^{(\tau)}(x,\tfrac{1}{n})n^{-(\tau-\bar\tau)}; \quad 0\le\bar\tau\le\tau. \quad (\text{II.2.1})$$

Здесь $\omega^{(\tau)}$ – наибольший из модулей непрерывности производных порядка τ от функции $x(t)$ в метрике $C(\bar\Omega)$, а величина C не зависит ни от x, ни от n.

Доказательство теоремы II.2.1 основано на построении и исследовании операторов, аналогичных операторам Джексона. Заметим, что постоянную C в неравенстве (II.2.1) можно оценить.

2°. Теорема II.2.2. (Шапошникова [2]). Пусть область Ω и функция $\varphi(t)$ удовлетворяют условиям п.1°, $x\in W_2^{(\tau)}(\Omega)$ и $\mathcal{D}^\alpha x|_{\partial\Omega}=0$, $0\le|\alpha|\le\delta-1$. Тогда для любого натурального существует полином $P_n(t)$ степени $\le n$ в R_m, удовлетворяющий неравенству

$$\|x-\varphi^\delta P_n\|_{(\bar\tau),\Omega} \le C\omega_{L_2}^{(\tau)}(x,\tfrac{1}{n})n^{-(\tau-\bar\tau)}, \quad 0\le\bar\tau\le\tau, \quad (\text{II.2.2})$$

где C не зависит от x и n, а $\omega_{L_2}^{(\tau)}$ есть наибольший из L_2-модулей непрерывности производных $\mathcal{D}^\alpha x$, $|\alpha|=\tau$.

Доказательство теоремы II.2.2 основано на том, что основные

свойства операторов, введенных Харрик, сохраняются при замене $C^{(k)}(\bar{\Omega})$ на $W_2^{(k)}(\Omega)$. Постоянную C в неравенстве (П.2.2) также можно оценить.

5°. Следствие П.2.1. Пусть граница $\partial\Omega$ представляет собой $(m-1)$-мерную липшецеву поверхность, и $x \in W_2^{(\tau)}(\Omega)$. Тогда существует полином $Q_n(t)$ степени $\leqslant n$ в R_m, удовлетворяющий неравенству

$$\|x - Q_n\|_{(\bar{\tau}),\Omega} \leqslant C_1 \max_{|\alpha|=\tau} \|\mathcal{D}^\alpha x\|_{2,\Omega} n^{-(\tau-\bar{\tau})}, \quad 0 \leqslant \bar{\tau} \leqslant \tau - 1. \qquad (\text{П.2.3})$$

Если поверхность $\partial\Omega$ достаточно гладкая, так что существует функция $\varphi(t)$ со свойствами, описанными в п.1°, то достаточно положить в (П.2.2) $\delta = 0$, заметив при этом, что $\omega_{L_2}^{(\tau)}(x,\delta) \leqslant 2 \max_{|\alpha|=\tau} \|x^{(\alpha)}\|_{2,\Omega}$. В общем случае продолжим функцию $x(t)$ с сохранением класса на все R_m так, чтобы она обращалась в нуль вне некоторого шара $\Omega_0 = \{t : |t| \leqslant R\}$ – такое продолжение возможно (см. Кальдерон [1]). Пусть $x^*(t)$ – продолженная функция; тем же символом $x^*(t)$ обозначим ее сужение на Ω_0. Граница $\partial\Omega_0$ – сфера – есть бесконечно гладкая поверхность, и для функции $x^*(t)$, по сказанному выше, неравенство (П.2.3) справедливо: существует такой полином $Q_n(t)$ степени $\leqslant n$ в R_m, что

$$\|x^* - Q_n\|_{(\bar{\tau}),\Omega_0} \leqslant C \max_{|\alpha|=\tau} \|\mathcal{D}^\alpha x^*\|_{2,\Omega_0} n^{-(\tau-\bar{\tau})}. \qquad (\text{П.2.4})$$

Функция $x^*(t)$ продолжает функцию $x(t)$ с сохранением класса, поэтому существует такая постоянная $\varkappa$ ("постоянная продолжения"), что

$$|\alpha| \leqslant \tau, \ \|\mathcal{D}^\alpha x^*\|_{2,\Omega_0} \leqslant \varkappa \|\mathcal{D}^\alpha x\|_{2,\Omega};$$

кроме того, очевидно,

$$\|x^* - Q_n\|_{(\bar{\tau}),\Omega_0} \geqslant \|x - Q_n\|_{(\bar{\tau}),\Omega}.$$

Последние два соотношения позволяют заменить неравенство (П.2.4) неравенством (П.2.3) с постоянной $C_1 = \varkappa C$.

4°. Вернемся к теореме П.2.2. Пусть $m = 1$ и область Ω есть промежуток $(0, 1)$. В этом случае можно положить $\varphi(t) = t(1-t)$ и неравенство (П.2.2) дает следующее: если $x \in W_2^{(\tau)}(0,1) \cap \mathring{W}_2^{(\tau)}(0,1)$, то при любом натуральном n существует такой полином $P_n(t)$ степени $\leqslant n$, что

$$\|x - P_{n,\delta}(t)\|_{(\bar{\tau}),(0,1)} \leqslant C \omega_{L_2}^{(\tau)}\left(x, \frac{1}{n}\right) n^{-(\tau-\bar{\tau})};$$

$$0 \leqslant \bar{\tau} \leqslant \tau,\ P_{n,3}(t) = t^3(1-t)^3 P_n(t). \qquad (\text{П.2.5})$$

Точно так же из следствия П.2.1 вытекает, что если $x \in W_2^{(\tau)}(0,1)$, то при любом натуральном n существует такой полином $Q_n(t)$ степени $\leqslant n$, что справедливо неравенство

$$\|x - Q_n\|_{(\bar{\tau}),(0,1)} \leqslant C \|\mathcal{D}^{\tau} x\|_{2,(0,1)} \cdot n^{-(\tau-\bar{\tau})},\ 0 \leqslant \bar{\tau} \leqslant \tau - 1. \qquad (\text{П.2.6})$$

§ 3. Обыкновенные дифференциальные уравнения.

1°. Рассмотрим краевую задачу для уравнения второго порядка

$$Ax := -\frac{d}{dt}\left(p(t)\frac{dx}{dt}\right) + q(t)x(t) = f(t);\ 0 < t < 1, \qquad (\text{П.3.1})$$

$$x(0) = x(1) = 0.$$

Примем следующие допущения: $p \in C^{(1)}[0,1]$, $q \in L_\infty(0,1)$, $f \in L_2(0,1)$; $p_0 \leqslant p(t) \leqslant p_1$, $0 < p_0 \leqslant p_1 < \infty$, $q(t) \geqslant 0$. При этих допущениях оператор A – положительно определенный в $L_2(0,1)$; слабое решение задачи (П.3.1) совпадает с решением вариационной задачи

$$|x|^2 - 2\int_0^1 f(t)x(t)\,dt = \min;\ x(0) = x(1) = 0, \qquad (\text{П.3.2})$$

где

$$|x|^2 = \int_0^1 \left\{p(t)\left(\frac{dx}{dt}\right)^2 + q(t)\,x^2(t)\right\} dt. \qquad (\text{П.3.3})$$

Пространство H_A состоит из функций $x(t)$, для которых интеграл (П.3.3) конечен и которые обращаются в нуль при $t=0$ и $t=1$. Ниже нам будет полезна следующая оценка энергетической нормы: если $q(t) \leqslant q_1$, то

$$|x|^2 \leqslant (p_1 + q_1/\pi^2)\|x'\|^2 := c_0\|x'\|^2, \qquad (\text{П.3.4})$$

где $\|\cdot\|$ означает норму в $L_2(0,1)$. Доказательство очевидно и следует из того, что наименьшее собственное число оператора $-d^2/dt^2$ при краевых условиях $x(0) = x(1) = 0$ равно π^2.

2°. За $H_A^{(n)}$ примем n-мерное подпространство с базисом $\{\sin k\pi t\}$, $1 \leqslant k \leqslant n$. Если $x^{(n)}$ – произвольный элемент из $H_A^{(n)}$, то

$$\rho_n^2 = |x_* - x_*^{(n)}| \leqslant |x_* - x^{(n)}| \leqslant C_0\|x_*' - x^{(n)'}\|. \qquad (\text{П.3.5})$$

Допустим, что $x_* \in W_2^{(\tau)}(0,1)$, $\tau \geqslant 3$. Для этого достаточно, чтобы $p \in W_2^{(\tau-1)}(0,1)$, $q, f \in W_2^{(\tau-2)}(0,1)$.

Функция $x_*(t)$ допускает разложение в ряд Фурье

$$x_*(t) = \sum_{k=1}^{\infty} c_k \sin k\pi t.$$

За $x_*(t)$ примем отрезок этого ряда:

$$x^{(n)}=\sum_{k=1}^{n} a_k \sin k\pi t .$$

Оценим коэффициенты

$$a_k=2\int_0^1 x_*(t)\sin k\pi t\,dt.$$

Интегрируя по частям и принимая во внимание краевые условия (П.2.I), получим

$$a_k=-\frac{2}{k^3\pi^3}\int_0^1 x_*'''(t)\cos k\pi t\,dt+\frac{2}{k^3\pi^3}\,x_*''(t)\cos k\pi t\Big|_0^1 .$$

Отсюда

$$|a_k|\le\frac{C}{k^3}\left[\|x_*'''\|_{L_2}+\|x_*''\|_{L_\infty}\right].$$

Дальнейшее интегрирование по частям (возможное, если $\tau>3$), не улучшает последней оценки. Теперь

$$\|x_*'-x^{(n)\prime}\|^2=\frac{1}{2}\sum_{k=n+1}^{\infty}k^2\pi^2a_k^2\le C\left[\|x_*'''\|_{L_2}^2+\|x_*''\|_{L_\infty}\right]\times$$

$$\times\sum_{k=n+1}^{\infty}\frac{1}{k^4}<C\left[\|x_*'''\|_{L_2}^2+\|x_*''\|_{L_\infty}^2\right]\int_n^\infty\frac{dk}{k^4}=$$

$$=C\left[\|x_*'''\|_{L_2}+\|x_*''\|_{L_\infty}^2\right]\frac{1}{n^3}$$

и формула (П.3.5) приводит к следующей оценки погрешности аппроксимации:

$$\rho_n\le C\left[\|x_*'''\|_{L_2}+\|x_*''\|_{L_\infty}\right]n^{-3/2} . \qquad (П.3.6)$$

3°. По-прежнему пусть $x_*\subset W_2^{(\tau)}(0,I)$. При $\tau>3$ можно получить лучшую оценку аппроксимации, если за $H_A^{(n)}$ принять подпространство полиномов вида

$$Q_n(t)=t(1-t)\sum_{j=0}^{k}a_jt^j\ ;\ 0\le k\le n\ .$$

В силу неравенства (П.2.5) существует такой полином $Q_n(t)$, что

$$\|x_*-Q_n\|_{(\bar\tau)}\le C\omega_{L_2}^{(\tau)}(x_*,\frac{1}{n})n^{-(\tau-\bar\tau)},\quad 0\le\bar\tau\le\tau ,$$

поэтому, если $x_*^{(n)}$ - приближенное решение из подпространства $H_A^{(n)}$, то

$$\rho_n=|x_*-x_*^{(n)}|\le C\omega_{L_2}^{(\tau)}(x_*,\frac{1}{n})n^{-\tau+1} . \qquad (П.3.7)$$

4°. Погрешность аппроксимации для обыкновенных дифференциальных уравнений в равномерных метриках изучали Крылов [I], автор [2], Бертрам [I], Берш-Буган [I], Леман [I] Джишкариани [2,3] в соболевских пространствах ту же погрешность исследовал Ильин [2] и Джишкариани [3]. Вайникко [2] рассмотрел также задачу с

спектре.

§ 4. Уравнения эллиптического типа. Оценка в энергетической норме.

1°. Рассмотрим первую краевую задачу для эллиптического уравнения (или эллиптической системы уравнений) вида

$$\sum_{|\alpha|,|\beta|=\sigma} (-1)^{|\alpha|} \mathcal{D}^{\alpha}(A_{\alpha\beta}\mathcal{D}^{\beta}x) = f(t),\ x\in\Omega; \qquad (\text{П.4.1})$$

$$\mathcal{D}^{\gamma}x|_{\partial\Omega} = 0\ ,\quad 0\le|\gamma|\le\sigma-1\ . \qquad (\text{П.4.2})$$

Пусть $\partial\Omega$ удовлетворяет требованиям п.1° § 2: поверхность $\partial\Omega$ можно задать уравнением $\varphi(t)=0$, причем $\varphi\in C^{(\nu)}(\Omega_1)$, $\Omega_1\supset\Omega$ и Ω_1 гомеоморфна шару; $\varphi(t)\neq 0$, $t\notin\partial\Omega$ и $grad\,\varphi|_{\partial\Omega}\neq 0$. Допустим, что оператор задачи (П.4.1) – (П.4.2) – симметричный и положительно определенный в $L_2(\Omega)$. Энергетическая норма для этого оператора определяется формулой

$$|x|^2 = \int_{\Omega}\sum_{|\alpha|,|\beta|=\sigma} A_{\alpha\beta}\mathcal{D}^{\alpha}x\,\mathcal{D}^{\beta}x\,dt. \qquad (\text{П.4.3})$$

Будем считать, что коэффициенты $A_{\alpha\beta}$ суть функции от t, ограниченные и измеримые в $\bar{\Omega}$. Тогда, очевидно,

$$|x| \le C_0 \|x\|_{(\sigma),\Omega}. \qquad (\text{П.4.4})$$

Задачу (П.4.1) – (П.4.2) будем решать приближенно по методу Ритца, приняв за $H_A^{(n)}$ подпространство функций вида $\varphi^{\sigma}(t)P_n(t)$, где $P_n(t)$ – полиномом степени $\le n$ в R_m. В общем случае x_* – решение нашей задачи – принадлежит пространству $W_2^{(\sigma)}(\Omega)$; допустим, что на самом деле оно более гладкое: пусть $x_*\in W_2^{(\nu)}(\Omega)$, где ν достаточно большое число. Тогда существует такое произведение $\varphi^{\sigma}(t)\hat{P}_n(t)\in H_A^{(n)}$, что

$$|x_* - \varphi^{\sigma}\hat{P}_n| \le C\omega^{(\nu)}_{L_2}(x_*, \tfrac{1}{n})n^{-(\nu-\sigma)}\ .$$

Если $x_*^{(n)}$ – приближенное решение задачи (П.4.1) – (П.1.2) по Ритцу, то, как это вытекает из сказанного в § 1, $|x_*-x_*^{(n)}| \le |x_*-\varphi^{\sigma}\hat{P}_n| \le C_0\|x_*-\varphi^{\sigma}\hat{P}_n\|_{(\sigma)}$ и, следовательно,

$$\rho_n = |x_* - x_*^{(n)}| \le C\omega^{(\nu)}_{L_2}(x_*, \tfrac{1}{n})n^{-(\nu-\sigma)}\ . \qquad (\text{П.4.5})$$

2°. Пусть теперь дифференциальное уравнение (П.4.1) решается при естественных краевых условиях. По-прежнему будем считать, что оператор рассматриваемой задачи симметричный и положительно определенный в $L_2(\Omega)$, энергетическая норма удовлетворяет неравенству (П.4.4) и решение, которое мы по-прежнему обозначим через x_*, принадлежит к классу $W_2^{(\nu)}(\Omega)$. За $H_A^{(n)}$ примем

подпространство полиномов степени $\leqslant n$ в R_m. Используя следствие П.2.1 и повторяя рассуждения п.1° данного параграфа, мы придем к оценке

$$\| x_* - x_*^{(n)} \| \leqslant C \| x_* \|_{(\nu)} n^{-(\nu-s)} . \qquad (\text{П.4.6})$$

Ту же оценку (П.4.6) можно получить, если краевые условия имеют вид (П.4.2) при $0 \leqslant |\gamma| \leqslant \kappa-1$, $\kappa < s$, а остальные условия - естественные. В этом случае за $H_A^{(n)}$ следует принять подпространство функций $\vartheta^{\kappa}(t) P_n(t)$, где $P_n(t)$ полином степени $\leqslant n$ в R_m.

3°. Условия, при которых точное решение $x_* \in W_2^{(\nu)}(\Omega)$, см. Агмон, Дуглис, Ниренберг [1,2].

§ 5. Некоторые оценки погрешности производных. Литературные указания.

1°. Будем рассматривать задачу (П.4.1) - (П.4.2). В § 4 мы вывели оценку погрешности аппроксимации для приближенного решения этой задачи в энергетической метрике. По существу это оценка для производных, порядок которых равен половине порядка уравнения (П.4.1). Представляет значительный интерес получение аналогичных оценок для производных других порядков. Этому вопросу посвящен ряд работ, в которых рассматриваются координатные функции, описанные в п.2 § 4.

Ильин [1,3] рассмотрел основные краевые задачи для невырождающегося эллиптического уравнения второго порядка; для этих задач получены оценки нормы $\| x_* - x_*^{(n)} \|_{C(\bar{\Omega})}$ при условии, что известен порядок убывания некоторых величин A_n, определенным образом зависящих от приближенного решения $x_*^{(n)}$. В аналогичном предположении получены оценки погрешности аппроксимации в $C(\bar{\Omega})$ и в $C^{(1)}(\bar{\Omega})$ в случае первой краевой задачи для бигармонического уравнения. В [3] Ильин улучшил оценки для эллиптического уравнения второго порядка. Для того же уравнения Харрик [1] получала оценку величин A_n; аналогичная оценка для бигармонического уравнения дана в работе Харрик [2]. В статье Ильина [5] рассмотрены эллиптические уравнения любого четного порядка $2s$ и получены оценки погрешности аппроксимации в $W_p^{(\sigma)}(\Omega)$, $1 < p < \infty$, σ - не очень большое неотрицательно целое число.

2°. В статье Канторовича [3] (см. также монографию Канторовича и Акилова [1]) предложена некоторая общая схема исследования погрешности аппроксимации проекционных методов в банаховых

пространствах. В статье автора [II] предложено[1] некоторое видоизменение этой схемы, которое позволило получить оценку погрешности аппроксимации метода Ритца в $W_2^{(\tilde{\delta})}$, где $\tilde{\delta} > 2\delta$ – не очень большое натуральное число. Следует указать, что оценка этой последней статьи хуже оценок статьи Ильина [5], соответствующих значению $p = 2$. Другое видоизменение схемы Канторовича было использовано Шапошниковой [1], которая таким способом получила оценку погрешности аппроксимации для производных порядка $\tilde{\delta} < 2\delta$. Ее оценки хуже оценок Ильина [5] при $p = 2$ для $\tilde{\delta} > \delta$ и лучше для $\tilde{\delta} < \delta$. В статье Шапошниковой [3] получены оценки в некоторых банаховых пространствах; показано, в частности, что для $\tilde{\delta} \leq \delta$ оценки Ильина [5] можно улучшить и при $p \neq 2$.

3°. В статье автора [32] метод его работы [II] и работы Шапошниковой [1] перенесен на несамосопряженные эллиптические задачи, допускающие решение по методу Бубнова – Галеркина. Благодаря использованию уточненных марковских неравенств удалось улучшить результаты названных выше работ автора [II] и Шапошниковой [1]; новые оценки совпадают с оценками Ильина (при $p = 2$) для $\tilde{\delta} > \delta$ и с оценками Шапошниковой [1] при $\tilde{\delta} < \delta$.

Ниже, в § 6, будут изложены необходимые марковские неравенства, в §§ 7,8 – результаты работ автора [II, 32] и Шапошниковой [1] для самосопряженных эллиптических задач; в § 9 будут рассмотрены несамосопряженные эллиптические задачи.

§ 6. Некоторые марковские неравенства

1°. Пусть $P_n(t)$ – полином степени $\leq n$ от вещественной переменной t на промежутке $a \leq t \leq b$. Обозначим через $q_k(t)$ $k = 0, 1, 2, \ldots$, полиномы степени k, ортонормированные в $L_2(a,b)$; эти полиномы отличаются лишь постоянным множителем $[2/(b-a)]^{1/2}$ от нормированных полиномов Лежандра, преобразованных к промежутку (a,b). Справедливо тождество

$$P_n(t) = \sum_{k=0}^{n} a_k q_k(t); \quad a_k = const.$$

Отсюда

$$P_n^2(t) \leq \sum_{k=0}^{n} a_k^2 \sum_{k=0}^{n} q_k^2(t) = \|P_n\|^2_{L_2(a,b)} \sum_{k=0}^{n} q_k^2(t).$$

---

[1] Результаты статьи автора [II], вышедшей из печати только в 1970 г., были доложены в 1968 г. на симпозиуме в честь 60-летия акад. С.Л.Соболева.

Далее, $q_k(t)=[2/(b-a)]^{1/2} P_n(\tau)$, где $P_n(\tau)$ – полиномы Лежандра, ортонормированные на промежутке $(-1, +1)$, и $\tau = (b-a)^{-1}(2t-a-b)$. Так как $\max\limits_{|\tau|\le 1} P_n(\tau)=\sqrt{(2n+1)/2}$, то

$$\max_{t\in[a,b]} \sum_{k=0}^{n} q_k^2(t) = \frac{2}{b-a}\sum_{k=0}^{n} \frac{2n+1}{2}$$

и, следовательно,

$$\|P_n\|_{C[a,b]} \le Cn \|P_n\|_{L_2(a,b)} . \qquad (\text{II.6.1})$$

Это неравенство хорошо известно.

2°. Выведем марковское неравенство для полиномов от многих переменных в L_2 . В статье Хилле, Сеге и Тамаркина [1] (см. также Бари [1]) для полиномов от одной вещественной переменной установлено следующее неравенство: если $P_n(t)$ – такой полином степени n , то

$$\|P_n'\|_{L_p(a,b)} \le Cn^2 \|P_n\|_{L_p(a,b)}, \qquad (\text{II.6.2})$$

где C зависит только от p и от $b-a$. Мы будем пользоваться этим неравенством для значения $p=2$.

Неравенство (II.6.2) очевидным образом обобщается на полиномы степени $\le n$ в R_m , если отрезок (a,b) заменяется параллелепипедом. Действительно, можно считать, что ребра параллелепипеда параллельны координатным осям. Пусть $\alpha=(\alpha_1,\alpha_2,\dots,\alpha_m)$ – мультииндекс, причем $\alpha_k=\delta_{jk}$, где индекс j фиксирован, и пусть для точек j-го ребра параллелепипеда $a_j \le t_j \le b_j$. По неравенству (II.6.2)

$$\int_{a_j}^{b_j} |\mathcal{D}^{\alpha} P_n(t)|^2 dt_j \le Cn^4 \int_{a_j}^{b_j} |P_n(t)|^2 dt.$$

Интегрируя по остальным координатам, получим (Π – параллелепипед)

$$|\alpha|=1, \quad \|\mathcal{D}^{\alpha} P_n\|_{L_2(\Pi)} \le Cn^2 \|P_n\|_{L_2(\Pi)}. \qquad (\text{II.6.3})$$

Отсюда

$$\forall \alpha, \quad \|\mathcal{D}^{\alpha} P_n\|_{L_2(\Pi)} \le Cn^{2|\alpha|} \|P_n\|_{L_2(\Pi)}. \qquad (\text{II.6.4})$$

3°. Рассмотрим теперь в R_m некоторую конечную область Ω с границей, удовлетворяющей условиям п.1° § 2. Представим Ω в виде $\Omega = \Omega' \cup \Omega_\delta$; $\Omega' \cap \Omega_\delta = \varnothing$. Здесь Ω_δ – пограничная полоса ширины δ и δ – достаточно малое фиксированное число. Каждую точку замкнутой области $\bar{\Omega}'$ сделаем центром куба, ребра которого параллельны осям координат и который целиком вместе со своей границей лежит в Ω . По лемме Бореля выберем конечное

число таких кубов, образующих покрытие для Ω' ; пусть это будут кубы Q_1 , Q_2 , ..., Q_K . В силу неравенства (II.6.4)

$$\forall d,\ \|\mathcal{D}^d P_n\|_{2,Q_j} \leqslant Cn^{2|d|}\|P_n\|_{2,Q_j}.$$

Просуммируем это по j . Слева заменим полученную сумму меньшей величиной $\|\mathcal{D}^d P_n\|_{2,\Omega'}$, а справа – большей величиной $K\|P_n\|_{2,\Omega}$. В результате получим

$$\forall d,\ \|\mathcal{D}^d P_n\|_{2,\Omega'} \leqslant Cn^{2|d|}\|P_n\|_{2,\Omega'}. \qquad (\text{II.6.5})$$

4°. Теперь рассмотрим пограничную полосу Ω_δ . Ширину δ этой полосы выберем следующим образом. Пусть $t_0 \in \partial\Omega$. Построим сферу с центром в t_0 , обладающую тем свойством, что любая прямая, параллельная нормали к $\partial\Omega$ в точке t_0 , пересекает часть Γ_0 поверхности $\partial\Omega$, заключенную внутри упомянутой сферы, не более одного раза. Известно, (см., например, книгу автора [10]), что радиус d этой сферы можно выбрать независящим от t_0 . Участок Γ_0 границы $\partial\Omega$ можно задать уравнением вида $\xi_m = g(\xi')$, где $\xi' = (\xi_1, \xi_2, \ldots, \xi_{m-1})$, ось ξ_m направлена по нормали к $\partial\Omega$ в точке t_0 и начало новой системы координат совпадает с t_0 . Так как функция $\varphi(t)$ (см. п.I° § 2) достаточно гладкая, причем $\operatorname{grad}\varphi|_{\partial\Omega} \neq 0$, то функция $g(\xi')$ также достаточно гладкая, и можно найти столь малое $\delta > 0$, чтобы лежащая в Ω_δ замкнутая область

$$E_{t_0,\delta} = \{t : -\delta \leqslant \xi_i \leqslant \delta,\ 1 \leqslant i \leqslant m-1;$$

$$g(\xi') - \delta \leqslant \xi_m \leqslant g(\xi')\}$$

находилась внутри сферы радиуса d с центром в t_0 .

Введем координаты

$$\eta_i = \xi_i,\ 1 \leqslant i \leqslant m-1,\ \eta_m = -\xi_m + g(\xi').$$

В координатах η_1 , η_2 , ..., η_m область $E_{t_0,\delta}$ переходит в параллелепипед

$$\mathcal{F}_\delta = \{\eta : -\delta \leqslant \eta_i \leqslant \delta,\ 1 \leqslant i \leqslant m-1;\ 0 \leqslant \eta_m \leqslant \delta\}.$$

Очевидно, якобиан преобразования от координат ξ_i к координатам η_i по абсолютной величине равен единице, а производные первого порядка от x по t_i оцениваются через те же производные по η_j , $i, j = 1, 2, \ldots, m$, и наоборот.

Каждой точке $t_0 \in \partial\Omega$ можно привести в соответствие область $E_{t_0,\delta}$, и эти области образуют покрытие для Ω_δ. По лемме Бореля можно выбрать конечное число областей $E_{t_0,\delta}$, также образу-

ющих покрытие для Ω_δ. Для каждого из параллелепипедов $\mathcal{F}_\delta$ справедливо неравенство (П.6.4), из которого в свою очередь следует новое неравенство

$$\forall \alpha, \|\mathcal{D}^\alpha P_n\|_{2,\Omega_\delta} \le C \|P_n\|_{2,\Omega} n^{2|\alpha|}.$$

Сложив это с (П.6.5), получим искомое марковское неравенство

$$\forall \alpha, \|\mathcal{D}^\alpha P_n\|_{2,\Omega} \le C \|P_n\|_{2,\Omega} n^{2|\alpha|}. \qquad (П.6.6)$$

Нетрудно получить аналогичное неравенство и для $p \neq 2$.

5°. Докажем, что справедливо неравенство

$$\|\vartheta^{s-1} P_n\|_{2,\Omega} \le C n^2 \|\vartheta^s P_n\|_{2,\Omega}. \qquad (П.6.7)$$

Начнем со случая одного измерения. В этом случае можно свести дело к сравнению интегралов

$$\int_0^\delta P_n^2(t)dt, \quad \int_0^\delta t^2 P_n^2(t)dt,$$

где $\delta > 0$ – фиксированное число. Интегрируя по частям, получим

$$\int_0^\delta P_n^2(t)dt = \delta P_n^2(\delta) - 2\int_0^\delta t P_n(t) P_n'(t)dt.$$

Оценим правую часть. Прежде всего, по формуле (П.6.1),

$$\delta P_n^2(\delta) \le \frac{1}{\delta}\|t P_n\|^2_{C[0,\delta]} \le C n^2 \|t P_n\|^2_{2,(0,\delta)}.$$

Далее,

$$\left|2\int_0^\delta t P_n(t) P_n'(t)dt\right| \le \frac{1}{\varepsilon}\int_0^\delta t^2 P_n^2(t)dt + \varepsilon\int_0^\delta P_n'^2(t)dt \le$$

$$\le \frac{1}{\varepsilon}\int_0^\delta t^2 P_n^2(t)dt + C\varepsilon n^4 \int_0^\delta P_n^2(t)dt.$$

Положив $\varepsilon = (2Cn^4)^{-1}$, найдем искомую оценку:

$$\|P_n\|_{2,(0,\delta)} \le C n^2 \|t P_n\|_{2,(0,\delta)}. \qquad (П.6.8)$$

Обратимся к общему случаю. Внутри Ω оценка тривиальна: если Ω' – внутренняя подобласть и $|\vartheta(t)| \ge \beta$, $\forall t \in \Omega'$, то

$$\|\vartheta^{k-1} P_n\|_{2,\Omega'} \le \frac{1}{\beta}\|\vartheta^k P_n\|_{2,\Omega'}. \qquad (П.6.9)$$

Достаточно поэтому рассмотреть область вида $E_{t_0,\delta}$ и сравнить интегралы

$$\int_{E_{t_0,\delta}} \vartheta^{2k-2}(t) P_n^2(t)dt, \quad \int_{E_{t_0,\delta}} \vartheta^{2k} P_n^2(t)dt.$$

Сделаем замену переменных, перейдя от переменных t_i к переменным η_i (см. п.4°). Область $E_{t_0,\delta}$ перейдет в параллелепипед $\mathcal{F}_\delta$, при этом интеграл вида

$$\int_{E_{t_0},\delta} \varphi^{2s}(t) P_n^2(t) dt \qquad (\text{П.6.10})$$

оценится сверху и снизу через интеграл

$$\int_{J_\delta} \eta_m^{2s} P_n^2(t)\, d\eta \qquad (\text{П.6.11})$$

с постоянными, не зависящими от n. Выражение $\eta_m^{K-1} P_n(t)$ есть полином степени $\leqslant n+K-1$ от η_m, и по неравенству (П.3.6)

$$\int_0^\delta \eta_m^{2K-2} P_n^2(t) d\eta_m \leqslant Cn^4 \int_0^\delta \eta_m^{2K} P_n^2(t) d\eta_m .$$

Интегрируя это по η_i, $1 \leqslant i \leqslant m-1$, в пределах от $-\delta$ до $+\delta$ и пользуясь эквивалентностью интегралов (П.6.10) и (П.6.11), найдем, что

$$\int_{E_{t_0},\delta} \varphi^{2K-2}(t) P_n^2(t) dt \leqslant Cn^4 \int_{E_{t_0},\delta} \varphi^{2K}(t) P_n^2(t) dt .$$

Отсюда и из неравенства (П.3.9) вытекает оценка (П.3.7').

§ 7. Оценки аппроксимации старших производных

I°. Выше было упомянуто, что Л.В.Канторович предложил некоторую схему построения и исследования проекционных методов. Мы изложим здесь эту схему в условиях не самых общих, но достаточных для того, чтобы ее можно было применить к методу Ритца.

Рассмотрим уравнение

$$Ax = f , \qquad (\text{П.7.1})$$

где A – линейный оператор, взаимно однозначно отображающий банахово пространство X на банахово пространство Y, тогда оба оператора A и A^{-1} ограничены. Построим полную в X последовательность конечномерных пространств X_n; напомним, что это означает следующее:

$$\forall x \in \lambda ,\ \mathcal{E}_n(x) = \inf_{x^{(n)} \in X_n} \| x - x^{(n)} \|_n \underset{n\to\infty}{\longrightarrow} 0 . \qquad (\text{П.7.2})$$

Положим $Y_n = AX_n$; для каждого n выберем оператор Q_n, проектирующий Y на Y_n. Приближенное решение уравнения (П.7.1) будем строить как элемент $x^{(n)} \in X_n$, удовлетворяющий уравнению

$$Q_n A x^{(n)} = Q_n f . \qquad (\text{П.7.3})$$

Если x_* и $x_*^{(n)}$ – точные решения задач (П.7.1) и (П.7.3) соответственно, то, как доказано в работах Канторовича, цитированных выше,

$$\|x_* - x_*^{(n)}\|_X \leqslant (1 + \|A^{-1} Q_n A\|_X)\,\delta_n(x_*). \qquad (П.7.4)$$

Отсюда следует оценка

$$\|x_* - x_*^{(n)}\|_X \leqslant C\,\|Q_n\|_Y\,\delta_n(x_*). \qquad (П.7.5)$$

2^0. В статье [II] автор предложил некоторое видоизменение схемы Канторовича. Пусть $X \subset Y \subset Z$; здесь Y и Z – гильбертовы пространства, X – банахово пространство; знак $\subset$ означает плотное вложение. В уравнении (П.7.1) будем рассматривать A также как оператор в Z и будем считать, что он положительно определенный, так что

$$\forall x \in \mathcal{D}(A),\ (Ax, x)_Z \geqslant \gamma^2 \|x\|_Z^2;\ \gamma^2 = const > 0;$$

в то же время, в соответствии со сказанным в п. 1^0, будем считать, что A взаимно однозначно отображает X на Y.

Выберем полную в X последовательность подпространств $\{X_n\}$; пусть $(\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN})$ – базис в X_n. Положим $Y_n = AX_n$ и определим оператор Q_n, отображающий Y на Y_n, соотношением

$$\forall f \in Y,\ (f - Q_n f,\ \varphi_{nj})_Z = 0,\quad j = 1, 2, \ldots, N \qquad (П.7.6)$$

Докажем, что Q_n есть проектор из Y на Y_n. По определению, $Q_n : Y \to Y_n$, поэтому, если $f \in Y$, то $Q_n f \in Y_n$ и, следовательно

$$Q_n f = \sum_{k=1}^{N} a_k^{(n)} A\varphi_{nk}.$$

Далее, по тому же определению, $(Q_n f - f, \varphi_{nj}) = 0$, или

$$\sum_{k=1}^{N} a_k^{(n)} (A\varphi_{nk}, \varphi_{nj})_Z = (f, \varphi_{nj})_Z,\ 1 \leqslant j \leqslant N. \qquad (П.7.7)$$

Оператор A положительно определенный в Z, поэтому и матрица системы (П.7.7) положительно определенная; эта система разрешима единственным образом, и оператор Q_n определен для всех натуральных n. Остается доказать, что $Q_n^2 = Q_n$. Если $f \in Y$, то $Q_n f \in Y_n \subset Y$ и $Q_n^2 f \in Y_n$, или

$$Q_n f = \sum_{k=1}^{N} b_k^{(n)} A\varphi_{nk}.$$

Далее, $(Q_n^2 f - Q_n f, y_{nj}) = 0$, поэтому $(Q_n^2 f, y_{nj}) = (Q_n f, y_{nj}) = (f, y_{nj})$, или

$$\sum_{k=1}^{N} b_k^{(n)} (A y_{nk}, y_{nj}) = (f, y_{nj}) , \quad j = 1, 2, \ldots, N .$$

Система (П.7.7) разрешима единственным образом, поэтому $b_k^{(n)} = a_k^{(n)}$ и $Q_n^2 f = Q_n f$.

Заменив f на $Ax^{(n)}$ под знаком проектора Q_n в (П.7.6), получим систему уравнений Бубнова – Галеркина для уравнения (П.7.1); так как A – положительно определенный оператор в Z, то (П.7.6), или равносильная система (П.7.7), есть также и система Ритца для того же уравнения. Поскольку $Q_n f \in Y_n$, то

$$Q_n f = \sum_{k=1}^{N} a_k^{(n)} A y_{nk} , \quad a_k^{(n)} = const$$

и, следовательно,

$$\| Q_n f \|_Y^2 = \| \sum_{k=1}^{N} a_k^{(n)} A y_{nk} \|_Y^2 \leqslant \Lambda_n^{(N)} \sum_{k=1}^{N} | a_k^{(n)} |^2 , \qquad (П.7.8)$$

где $\Lambda_n^{(N)}$ – наибольшее собственное число матрицы скалярных произведений $(A y_{nk}, A y_{nj})_Y$; $j, k = 1, 2, \ldots, N$. Заметим, что $a_k^{(n)}$ суть коэффициенты Ритца приближенного решения $x_*^{(n)}$:

$$x_*^{(n)} = \sum_{k=1}^{N} a_k^{(n)} y_{nk} .$$

Обозначая, как обычно, через $[\cdot, \cdot]$ и $|\cdot|$ скалярное произведение и норму в энергетическом пространстве Z оператора A, получаем из последнего тождества

$$| x_*^{(n)} |^2 = \sum_{j,k=1}^{N} a_j^{(n)} \overline{a_k^{(n)}} [y_{nj}, y_{nk}] =$$

$$= \sum_{j,k=1}^{N} a_j^{(n)} \overline{a_k^{(n)}} (A y_{nj}, y_{nk})_Z \geqslant \lambda_n^{(1)} \sum_{k=1}^{r} | a_k^{(n)} |^2 ;$$

здесь $\lambda_n^{(1)}$ – наименьшее собственное число матрицы скалярных произведений $(A y_{nj}, y_{nk})_Z$; $j, k = 1, 2, \ldots, N$. Теперь (см. книгу автора [3])

$$\sum_{k=1}^{N} | a_k^{(n)} |^2 \leqslant \frac{| x_*^{(n)} |^2}{\lambda_1^{(n)}} \leqslant \frac{| x_* |^2}{\lambda_n^{(1)}} \leqslant \frac{\| f \|_Z^2}{r^2 \lambda_n^{(1)}} .$$

Из вложения $Y \subset Z$ вытекает существование такой постоянной C, что $\| f \|_Z \leqslant C \| f \|_Y$, и из (П.7.8) вытекает искомая оценка:

$$\|Q_n\|_{y\to y_n} \le C\sqrt{\Lambda_n^{(\sigma)}/\lambda_n^{(1)}}\,. \qquad (\text{П.7.9})$$

Если базис подпространства $H_A^{(n)}$ ортонормирован в Z_A, то $\lambda_n^{(1)} = 1$, и последняя оценка упрощается:

$$\|Q_n\|_{y\to y_n} \le C\sqrt{\Lambda_n^{(\sigma)}}\,. \qquad (\text{П.7.10})$$

Из формулы (П.7.5) получается оценка погрешности аппроксимации в X-норме:

$$\|x_* - x_*^{(n)}\|_X \le C\sqrt{\Lambda_n^{(\sigma)}}\,\mathcal{E}_n(x_*). \qquad (\text{П.7.11})$$

3°. Вернемся к задаче (П.4.1) – (П.4.2). Допустим, что система (П.4.1) равномерно эллиптична в Ω, а оператор рассматриваемой задачи положительно определен в $L_2(\Omega)$. Решение $u_*(t)$ будет сколь угодно гладким, если достаточно гладкими будут функции $\varphi(t)$, $f(t)$, $A_{\alpha\beta}(t)$. Будем считать, что $x_* \in W_2^{(\tau)}(\Omega)$, где τ достаточно большое число.

Рассмотрим сначала более простой случай, когда $\varphi(t)$ есть полином. Задачу (П.4.1) – (П.4.2) будем решать по Ритцу, взяв в качестве координатных функций полиномы $\varphi^0(t)P_n^{(j)}(t)$, где $P_n^{(j)}$ – полиномы степени $\le n$ в R_m, тогда приближенное решение имеет вид $x_*^{(n)}(t) = \varphi^0(t)\,P_n(t)$, где $P_n(t)$ – такой полином степени $\le n$ в R_m.

Введем в рассмотрение пространства $X = \mathring{W}_2^{(\tilde\sigma)}(\Omega)$, $Y = W_2^{(\ell)}(\Omega)$, $Z = L_2(\Omega)$; $\ell \ge 0$, $\tilde\sigma = \sigma + \ell \le \tau$. Волна сверху здесь означает подпространство пространства $W_2^{(\tilde\sigma)}(\Omega)$, определяемое условиями (П.4.2).

Координатные функции $\vartheta_{nj}(t)$ будем считать ортонормированными в $W_2^{(\sigma)}(\Omega)$ – тогда, как нетрудно проверить (подробнее см. книгу автора [8]) числа $\lambda_n^{(1)}$ положительно ограничены снизу и верна оценка (П.7.11).

Пусть N – число функций ϑ_{nj} при фиксированном n. Положим $\tau = (\tau_1, \tau_2, \ldots, \tau_N)$ и $\|\tau\|^2 = |\tau_1|^2 + |\tau_2|^2 + \ldots + |\tau_N|^2$. Тогда

$$\Lambda_n^{(\sigma)} = \max_{\|\tau\|=1} \|A\Phi_n(\cdot,\tau)\|_Y^2, \quad \text{где } \Phi_n(t,\tau) = \sum_{j=1}^{N} \tau_j \vartheta_{nj}(t).$$

Для любой функции $u \in \mathring{W}_2^{(\tilde\sigma)}(\Omega)$ справедливо неравенство

$\|Au\|_Y \leqslant C\|u\|_X$; в частности, $\|A\varphi_n\|_Y \leqslant C\|\varphi_n\|_X =$

$=C\|\varphi_n\|_{(\tilde{s}),\Omega}$. Но $\varphi_n(t,\tau)$ есть полином степени $\leqslant n + const$ в R_m, и по неравенству (II.6.3)

$$\|\varphi_n\|_{(\tilde{s}),\Omega} \leqslant Cn^{2(\tilde{s}-s)}\|\varphi_n\|_{(s),\Omega} \leqslant$$

$$\leqslant Cn^{2(\tilde{s}-s)}\sum_{k=1}^{N}|\tau_k|^2 = Cn^{2(\tilde{s}-s)}.$$

Теперь

$$\Lambda_n^{(N)} \leqslant Cn^{4(\tilde{s}-s)}. \qquad (II.7.12)$$

Остается оценить наилучшее приближение

$$\mathcal{E}_n(x_*) = \inf_{x^n \in X_n}\|x_* - x^{(n)}\|_X = \inf_{x^{(n)}\in X_n}\|x_* - x^{(n)}\|_{(\tilde{s}),\Omega}.$$

Функция $x^{(n)}$ есть полином степени $\leqslant n + const$ в R_m, а $x_* \in W_2^{(\tau)}(\Omega)$; по теореме II.2.2 существует такой полином $\bar{x}^{(n)} \in X_n$, что

$$\|x_* - \bar{x}^{(n)}\|_{(\tilde{s}),\Omega} \leqslant Cn^{-(\tau-\tilde{s})}\omega^{(\tau)}_{L_2}\left(x_*; \frac{1}{n}\right).$$

Это значит, что $\mathcal{E}_n(x_*) \leqslant Cn^{-(\tau-\tilde{s})}\omega^{(\tau)}_{L_2}\left(x_*, \frac{1}{n}\right)$.

Теперь из формул (II.7.5) и (II.7.12) следует:

$$\rho_n = \|x_* - \bar{x}^{(n)}_*\|_{(\tilde{s}),\Omega} \leqslant Cn^{-(\tau-\tilde{s})+2\tilde{s}}\omega^{(\tau)}_{L_2}\left(x_*, \frac{1}{n}\right); \qquad (II.7.13)$$

это и есть оценка погрешности аппроксимации для производных порядка $\tilde{s} \geqslant 2s$.

4°. Вернемся к общему случаю и будем считать, что функция $\varphi(t)$ достаточно гладкая, но не обязательно полином. Заново проведем оценку величины $\Lambda_n^{(N)} = \max_{|\tau|=1}\|A\varphi(\cdot,\tau)\|_Y$. Для любой функции $u \in \overset{\circ}{W}{}_2^{(\tilde{s})}(\Omega)$ справедливо неравенство $\|Au\|_{(\ell)} \leqslant$ $\leqslant C|u|_{(\tilde{s})}$; в частности, $\|A\varphi_n\|_{(\ell)} \leqslant C\|\varphi_n\|_{(\tilde{s})}$. Норму в $W_2^{(\tilde{s})}(\Omega)$ можно задать формулой

$$\|u\|_{(\delta)} = \sum_{\sigma,\tau} \|\mathcal{D}^{\sigma+\tau} u\|_2 ;$$

суммирование производится по всем мультииндексам σ и τ, таким, что $|\sigma| \geqslant \delta+\ell$ и $|\tau| \leqslant \delta$. Докажем неравенство

$$\|\mathcal{D}^{\sigma+\tau} \varphi_n\|_2 \leqslant C n^{2|\sigma|} \|\varphi_n\|_{(\delta)} . \qquad (\text{П.7.14})$$

Имеем $\varphi_n(t,\tau) = \vartheta^\delta(t) P_n(t,\tau)$, где $P_n(t,\tau)$ – полином степени $\leqslant \underline{n}$ в R_m. Построим полином $q_n(t)$ степени $\leqslant \underline{n}$ в R_m, такой, что

$$\forall \nu,\ |\nu| \leqslant |\sigma+\tau|,\ |\mathcal{D}^\nu [\vartheta^\delta(t) - q_n(t)] = O(n^{-\upsilon+\nu}) ;$$

это возможно в силу теоремы П.2.1. Будем считать, что υ достаточно велико. Далее,

$$\|\mathcal{D}^{\sigma+\tau} \varphi_n(\cdot,\tau)\|_2 \leqslant \|\mathcal{D}^{\sigma+\tau} (q_n P_n(\cdot,\tau)\|_2 +$$

$$+ \|\mathcal{D}^{\sigma+\tau} (\vartheta^\delta - q_n)) P_n(\cdot,\tau)\|_2 . \qquad (\text{П.7.15})$$

Выражение $\mathcal{D}^\tau (q_n(t) P_n(t,\tau))$ есть полином степени $\leqslant 2\underline{n} - \tau \leqslant 2n$ в R_m, и по неравенству (П.6.4)

$$\|\mathcal{D}^{\sigma+\tau} (q_n P_n(\cdot,\tau))\|_2 \leqslant C n^{2|\sigma|} \|\mathcal{D}^\tau (q_n P_n(\cdot,\tau))\|_2 . \qquad (\text{П.7.16})$$

$$\|\mathcal{D}^\tau (q_n P_n(\cdot,\tau)\|_2 \leqslant$$

$$\leqslant \|\mathcal{D}^\tau (\vartheta^\delta P_n(\cdot,\tau)\|_2 + \|\mathcal{D}^\tau [(\vartheta^\delta - q_n) P_n(\cdot,\tau)]\|_2 . \qquad (\text{П.7.17})$$

Оценим второе слагаемое:

$$\|\mathcal{D}^\tau [(\vartheta^\delta - q_n) P_n(\cdot,\tau)\|_2 = \|\sum_{\alpha \leqslant \tau} \mathcal{D}^\alpha (\vartheta^\delta - q_n) \mathcal{D}^{\tau-\alpha} P_n(\cdot,\tau)\|_2 \leqslant$$

$$\leqslant \frac{C}{n^{\upsilon-|\tau|}} \sum_{\alpha \leqslant \tau} \|\mathcal{D}^{\tau-\alpha} P_n(\cdot,\tau)\|_2 \leqslant C n^{-\upsilon+3|\tau|} \|P_n(\cdot,\tau)\|_2 \leqslant$$

$$\leqslant Cn^{-\sigma+3|\alpha|+2\delta}\|\mathcal{Y}^{\delta}P_n(\cdot,\tau)\|_2 \leqslant C\|\mathcal{Y}^{\delta}P_n(\cdot,\tau)\|_2 ;$$

здесь использованы неравенства (II.6.6) и (II.6.II), а также то, что σ достаточно велико. Соболевские теоремы вложения дают неравенство

$$\|\mathcal{Y}^{\delta}P_n(\cdot,\tau)\|_2 \leqslant C\|\mathcal{Y}^{\delta}P_n(\cdot,\tau)\|_{(\delta)},$$

а, следовательно, и неравенство

$$\|\mathcal{D}^{\alpha}[(\mathcal{Y}^{\delta}-q_m)P_n(\cdot,\tau)]\|_2 \leqslant C\|\mathcal{Y}^{\delta}P_n(\cdot,\tau)\|_{(\delta)}. \qquad (II.7.18)$$

Точно так же находим

$$\|\mathcal{D}^{\alpha+\sigma}[(\mathcal{Y}^{\delta}-q_k)P_n(\cdot,\tau)\|_2 \leqslant C\|\mathcal{Y}^{\delta}P_n(\cdot,\tau)\|_{(\delta)}. \qquad (II.7.19)$$

Из неравенств (II.7.I5–I9) очевидно следует (II.7.I4), из которого в свою очередь следует, что

$$\|\varphi_n\|^2_{(\delta)} \leqslant Cn^{4\delta-2\delta}\|\varphi_n\|^2_{(\delta)} \leqslant Cn^{4\delta-2\delta}.$$

В силу теоремы II.2.I, $\mathcal{E}_n(x_*) \leqslant Cn^{-\tau+\delta}\omega^{(\tau)}_{L_2}\left(x_*, \frac{1}{n}\right)$. Последние два неравенства вместе с формулой (II.7.I0) приводят к оценке (II.7.I3).

5°. Если уравнение (II.4.I) решать при естественных краевых условиях, то можно получить оценку, близкую к (II.7.I3). За координатные функции можно взять полиномы, при этом все предшествующие рассуждения в существенном сохраняются, но для наилучшего приближения получается (II.2.3), и окончательно

$$\rho_n = \|x_* - x_*^{(n)}\|_{(\delta)} \leqslant Cn^{-\tau+3\delta-2\delta}. \qquad (II.7.20)$$

§ 8. Оценка аппроксимации младших производных

I°. Такого рода оценки получила, в частности, Шапошникова [I], используя еще одно видоизменение схемы Канторовича. По-прежнему к пространствам X и Y добавляется еще одно пространство Z и по-прежнему пространства Y и Z гильбертовы, но, в отличие от § 7, рассматриваемые пространства связаны соотношением

$X \subset Z \subset Y$. Допустим, что существует такое сужение $\hat{A}$ оператора A задачи (П.7.1), что $\mathcal{D}(\hat{A}) \subset Z$, $R(\hat{A}) \subset Z$ и $\hat{A}$ – положительно определенный оператор в Z. Энергетическое пространство оператора $\hat{A}$ обозначим через $\hat{Z}$. Возможны два случая взаимного расположения пространств:

$$X \subset \hat{Z} \subset Z \subset Y, \qquad (П.8.1)$$

$$\hat{Z} \subset X \subset Z \subset Y. \qquad (П.8.2)$$

Введем в рассмотрение пространство Y^*, сопряженное с Y относительно скалярного умножения в Z; примем, что $Y^{**} = Y$, где Y^{**} – пространство, сопряженное с Y^* относительно того же скалярного умножения в Z.

2^0. Выберем последовательность конечномерных подпространств $\{X_n\}$, полную в $X \cap Y^*$, и пусть

$$(y_{n1}, y_{n2}, \ldots, y_{nr}), \quad n = 1, 2, \ldots \qquad (П.8.3)$$

– базис в X_n. Будем считать, что при любом n элементы (П.8.3) ортонормированы в $\hat{Z}$.

Обозначим через $\tilde{A}$ сужение оператора A на $Y^* \cap X$ и допустим, что $\tilde{A}$ – положительно определенный оператор из Y^* в Y (см. Бирман [1], а также книгу автора [8]). Допустим еще, что соответствующее энергетическое пространство совпадает с $\hat{Z}$ – энергетическим пространством оператора $\hat{A}$.

Положим $Y_n = AX_n$. Проекторы Q_n определим соотношениями

$$\forall f \in Y;\ Q_n f \in Y_n,\ (f - Q_n f, y_{nj}) = 0,\ j = 1, 2, \ldots, N; \qquad (П.8.4)$$

левая часть уравнения (П.8.4) рассматривается как значение функционала $(f - Q_n f) \in Y = Y^{**}$ на элементе $y_{nj} \in Y^*$. Как и в § 7, доказывается, что Q_n есть проектор из Y на Y_n.

Оценим норму $\|Q_n\|_Y$, предполагая, что выполнено соотношение (П.8.1). Приближенное решение имеет вид

$$x_*^{(n)} = \sum_{k=1}^{N} a_k^{(n)} y_{nk};$$

при этом $Ax_n^* = Q_n f$ и, следовательно,

$$\|Q_n f\|_Y^2 =$$

$$= \sum_{j,k=1}^{N} (A\varphi_{nk}, A\varphi_{nj})_Y a_k^{(n)} \bar{a}_{jn} \le \Lambda_n^{(N)} \sum_{k=1}^{N} |a_k^{(n)}|^2 ; \qquad (\text{П.8.5})$$

здесь, как и в § 7, $\Lambda_n^{(N)}$ есть наибольшее собственное число матрицы чисел $(A\varphi_{nk}, A\varphi_{nj})_Y$; $j,k = 1,2, \dots, n$. Элементы φ_{nk} ортонормированы в $\overset{\circ}{Z}$, поэтому

$$\sum_{k=1}^{N} |a_k^{(n)}|^2 = |x_*^{(n)}|^2 := \|x_*^{(n)}\|_{\overset{\circ}{Z}}^2 .$$

Нетрудно убедиться, что x_* есть решение задачи $\tilde{A}x = f$, а $x_*^{(n)}$ – приближенное решение этой задачи по Ритцу. В книге автора [3] показано, что $|x_*^{(n)}| \le |x_*|$; отсюда

$$\sum_{k=1}^{N} |a_k^{(n)}|^2 \le |x_*|^2 \le C \|x_*\|_X^2 = C \|\tilde{A}^{-1} f\|_X^2 \le C \|f\|_Y^2 .$$

Теперь из (П.8.5) следует, что $\|Q_n\|_Y \le C\sqrt{\Lambda_n^{(N)}}$ и

$$\|x_* - x_*^{(n)}\|_X \le C\sqrt{\Lambda_n^{(N)}}\, \mathcal{E}_n(x_*) . \qquad (\text{П.8.6})$$

3°. Рассмотрим теперь случай (П.8.2). Неравенство (П.8.5) остается справедливым. Для произвольных элементов $h \in Y$ и $g \in Y'$ имеет место неравенство $|(h,g)| \le \|h\|_Y \cdot \|g\|_{Y'}$. Используя уравнение (П.8.4) и ограниченность оператора $\tilde{A}^{-1}$, обратного положительно определенному оператору $\tilde{A}$, можно утверждать, что

$$\sum_{k=1}^{N} |a_k^{(n)}|^2 = |x_*^{(n)}|^2 \le |x_*|^2 = (\tilde{A}x_*, x_*) = (f, x_*) \le$$

$$\le \|f\|_Y \cdot \|x_*\|_{Y'} = \|f\|_Y \cdot \|\tilde{A}^{-1} f\|_{Y'} \le C \|f\|_Y^2 .$$

Теперь из (П.8.5) вытекает, что $\|Q_n f\|_Y \le C\sqrt{\Lambda_n^{(N)}}\, \|f\|_Y$ и, следовательно, $\|Q_n\|_Y \le C\sqrt{\Lambda_n^{(N)}}$. Отсюда окончательно

$$\|x_* - x_*^{(n)}\|_X \le C\sqrt{\Lambda_n^{(N)}}\, \mathcal{E}_n(x_*) . \qquad (\text{П.8.7})$$

4°. Обратимся к задаче (П.4.1-2); сохраним все сделанные в предшествующих параграфах допущения о данных и решении этой задачи. Возьмем целое число ℓ, $0 \le \ell \le 2s$, и обозначим $\tilde{s} = 2s - \ell$. Введем пространства $X = \overset{\circ}{W}{}_2^{(\tilde{s})}(\Omega)$, $Y = W_2^{(\ell)}(\Omega)$, $Z = L_2(\Omega)$

Тогда $X \subset Z \subset Y$; как и выше, нслик сверху означает, что функции из X удовлетворяют условиям (П.4.2). В данном случае $Z \sim \mathring{W}_2^{(3)}(\Omega)$, поэтому расположения пространств (П.8.I) и (П.8.2) соответствуют значениям $0<\ell<3$ и $3<\ell<23$. Чтобы доказать, что наша задача соответствует схеме п.2° настоящего параграфа, достаточно установить справедливость следующих утверждений: а) операторы A и A^{-1} ограничены, б) оператор $\tilde{A}$ положительно определенный, в) энергетическое пространство оператора $\tilde{A}$ совпадает с $\overset{\circ}{Z}$. Ограниченность оператора A очевидна. Далее, в рассматриваемом нами случае $Y^* = X$; отсюда вытекает, что $\tilde{A} = A$; нетрудно также убедиться, что A - симметричный и положительный оператор из $X = Y^*$ в Y , и что его энергетическое пространство совпадает с $\overset{\circ}{Z}$. Далее, если будет доказано, что оператор A^{-1} ограничен, то положительный оператор $\tilde{A}$ будет положительно определенным. Окончательно, нам будет достаточно доказать ограниченность оператора A^{-1}.

По определению, A есть расширение оператора $\hat{A}$, определяемое соотношением

$$(Ax,y)_Z = (x,\hat{A}^* y)_Z = (x,\hat{A}y)_Z ;$$
$$x \in \mathring{W}_2^{(3)}(\Omega) , \quad y \in W_2^{(23)}(\Omega). \qquad (П.8.8)$$

Оператор $\hat{A}$ положительно определенный в $Z = L_2(\Omega)$, поэтому он имеет только тривиальный нуль. Докажем, что у расширенного оператора A нет других нулей.

Пусть $Ax_0 = 0$. Тогда $(x_0, \hat{A}y)_Z = 0$. Это значит, что x_0, рассматриваемый как функционал в $Z^* = L_2(\Omega)$, аннулируется на множестве значений оператора $\hat{A}$, т.е. на всем пространстве $L_2(\Omega)$. Но тогда $x_0 = 0$, оператор A не имеет нетривиальных нулей и существует обратный оператор A^{-1}. Далее, оператор $\hat{A}$ очевидно нетеров, и его индекс равен нулю. Как доказано в статье автора [9] (см. также [I5]), оператор A также нетеров, и его индекс равен индексу оператора $\hat{A}$, т.е. равен нулю. Только-что было показано, что подпространство нулей оператора A имеет нулевую размерность, а тогда нулевую размерность и подпространство нулей сопряженного оператора A^*. Отсюда следует, что оператор A^{-1} определен на всем пространстве $W_2^{(3-23)}$. В то же время он замкнут как обратный нетерову (и, следовательно, замкнутому)

оператору A . Из сказанного следует, что оператор A^{-1} ограничен.

5°. Как и в § 7, имеем $x_*^{(n)}(t) = y^0(t)P_n(t)$, где $P_n(t)$ - полином степени $\leqslant n$ в R_m, и

$$\Lambda_n^{(N)} = \max_{\|\tau\|=1} \Big\| \sum_{k=1}^{N} \tau_k A y_{nk} \Big\|_{(-\ell)}^2 = \max_{\|\tau\|=1} \|A\Phi_n(\cdot,\tau)\|_{(-\ell)}^2 .$$

Оператор A ограничен, поэтому

$$\|A\Phi_n(\cdot,\tau)\|_{(-\ell)} \leqslant C\|\Phi_n(\cdot,\tau)\|_{(2\hat{s}-\ell)} = C\|\Phi_n(\cdot,\tau)\|_{(\tilde{s})} .$$

Рассмотрим сначала случай $\tilde{s} > s$, или $\ell < s$, что соответствует расположению (П.8.I). По неравенству (П.6.6)

$$\|\Phi_n(\cdot,\tau)\|_{(\tilde{s})} \leqslant Cn^{2(\tilde{s}-s)} \|\Phi_n(\cdot,\tau)\|_{(s)} \leqslant Cn^{2(\tilde{s}-s)} . \qquad (\text{П.8.9})$$

Далее, по теореме II.2.2 ,

$$\mathcal{E}_n(x_*) = \inf_{x^n \in X_n} \|x_* - x^{(n)}\|_{(\tilde{s})} \leqslant Cn^{-\tau+\tilde{s}} \omega_{L_2}^{(\tau)}\Big(x_*, \frac{1}{n}\Big) ,$$

и из формулы (П.8.6) следует

$$s < \tilde{s} \leqslant 2s, \quad \rho_n = \|x_* - x_*^{(n)}\|_{(\tilde{s})} \leqslant Cn^{-\tau+2\tilde{s}-2s} \omega_{L_2}^{(\tau)}\Big(x_*, \frac{1}{n}\Big) . \qquad (\text{П.8.I0})$$

Теперь рассмотрим случай $0 \leqslant \tilde{s} < s$. Неравенство (П.8.9) заменяется более простым:

$$\|\Phi_n(\cdot,\tau)\|_{(\tilde{s})} \leqslant C\|\Phi_n(\cdot,\tau)\|_{(s)} = C ,$$

отсюда следует оценка

$$0 \leqslant \tilde{s} < s, \quad \rho_n = \|x_* - x_*^{(n)}\|_{(\tilde{s})} \leqslant Cn^{-\tau+\tilde{s}} \omega_{L_2}^{(\tau)}\Big(x_*, \frac{1}{n}\Big) . \qquad (\text{П.8.II})$$

Оценка (П.8.10) точнее, а оценка (П.8.II) совпадает с соответствующей оценкой статьи Шапошниковой [I] .

Заметим, что при $p = 2$ и $\tilde{s} > s$ оценки Альпина [5] совпадают с соответствующими оценками §§ 7, 8 ; при $p = 2$ и $\tilde{s} \leqslant s$ оценки Альпина несколько хуже, чем соответствующие оценки § 4 ($\tilde{s} = s$)

и § 8 ($\hat{3}<3$). Шапошникова показала в [3], что указанные здесь оценки Ильина могут быть улучшены.

Некоторые вопросы оценки погрешности метода Ритца в общих банаховых пространствах исследованы в статье Шапошниковой [4].

§ 9. Метод Бубнова – Галеркина

1°. Основы метода Бубнова-Галеркина, а также более общего метода Галеркина-Петрова изложены в книгах автора [2] и [4]; будем считать, что эти основы читателю известны.

В данном параграфе мы будем заниматься оценкой погрешности аппроксимации метода Бубнова-Галеркина, впервые, по-видимому, установленная Красносельским [1] в весьма общей ситуации – для нелинейных уравнений. В разных вариантах эта оценка изложена с подробными доказательствами в книге Красносельского и др. [1]; там же приведены необходимые литературные указания. Здесь мы ограничимся тем, что приведем (без доказательства) названную оценку в простейшем случае. Пусть A_0 самосопряженный положительно определенный оператор в некотором гильбертовом пространстве H, $\mathcal{K}$ – линейный в H оператор, такой, что $\mathcal{D}(\mathcal{K})\supset\mathcal{D}(A_0)$ и $T=A_0^{-1}\mathcal{K}$ – оператор, вполне непрерывный в энергетическом пространстве. Пусть $A=A_0+\mathcal{K}$, и уравнение $Ax=0$ имеет только тривиальное решение. Тогда неоднородное уравнение $Ax=f$, $\forall f\in H$, имеет одно и только одно обобщенное решение (им является решение уравнения $x+Tx=A_0^{-1}f$). Обозначим это решение через x_*. В этих условиях (см. книгу автора [2]) при достаточно больших n существует одно и только одно приближенное решение $x_*^{(n)}$ по Бубнову – Галеркину, и $|x_*-x_*^{(n)}|\xrightarrow[n\to\infty]{}0$; $|\cdot|$ – норма в H_0. Упомянутая выше оценка погрешности имеет вид

$$\mathcal{E}_n(x_*)\le|x_*-x_*^{(n)}|\le C\,\mathcal{E}_n(x_*); \qquad (\text{П.9.1})$$

здесь $\mathcal{E}_n(x_*)$ имеет тот же смысл, что и в предшествующих параграфах настоящей главы.

Работы Вайникко [1, 3, 4, 9] содержат оценки аппроксимации для метода Бубнова-Галеркина в случае, когда координатные элементы суть собственные элементы самосопряженного оператора, сходного (т.е., имеющего общую область определения) с A_0. Работы того же автора [5, 6, 8] посвящены оценкам погрешности аппроксимации метода Бубнова – Галеркина в проблеме собственных значений

Оценкам погрешности аппроксимации для метода Бубнова-Галеркина посвящены работы Джишкариани [1, 4-7, 9] .

2°. Приведем некоторые результаты перечисленных выше работ. В статье Джишкариани [4] уравнение

$$A_0 x + \mathcal{K} x = f \qquad (II.9.2)$$

рассматривают в предположениях, упомянутых в начале данного параграфа и при дополнительном предположении, что в пространстве H_0 вполне непрерывен оператор $\mathcal{K}A_0^{-1}$. Пусть B - оператор, сходный с A_0 ; это значит, что B - самосопряженный положительно определенный оператор и $\mathcal{D}(B) = \mathcal{D}(A_0)$. Допустим еще, что B (а, следовательно, и A_0) имеет дискретный спектр; пусть ω_k и v_k - собственные числа и собственные элементы оператора B . Примем элементы v_k за координатные и обозначим через δ_n невязку приближенного решения $x_*^{(n)}$: $\delta_n = A_0 x_*^{(n)} + \mathcal{K} x_*^{(n)} - f$; допустим, что $\delta_n \in \mathcal{D}(B^\tau)$, $\tau \geq 0$; при $\tau = 0$ это допущение очевидно выполняется. Доказывается, что $\|\delta_n\|_H = o(1)$. В статье получены оценки

$$\|A_0(x_* - x_*^{(n)})\|_H = \frac{\|T_1\|_H \cdot \|B^\tau \delta_n\|}{\omega_{n+1}^\tau} ; \qquad (II.9.3)$$

$$|x_* - x_*^{(n)}| \leq \omega_{n+1}^{-\tau-\frac{1}{2}} \left[\|T\|_H \cdot \|T_1\|_H \cdot \|A_0^{-1} B\|_H\right]^{\frac{1}{2}} \|B^\tau \delta_n\| . \qquad (II.9.4)$$

Здесь $|\cdot|$ - норма в H_0 - энергетическом пространстве оператора A_0, I - тождественный оператор; $T = (I + A_0^{-1}\mathcal{K})^{-1}$, $T_1 = (I + \mathcal{K}A_0^{-1})^{-1}$.

Вайникко в [9] рассмотрел более общее уравнение вида $Ax = f$, где A оператор, действующий из одного банахова пространства в другое. Мы изложим здесь результаты упомянутой статьи, относящиеся к уравнению (II.9.2). По-прежнему пусть B - оператор, сходный с A , и пусть P_λ разложение единицы для оператора B . Обозначим $H_\lambda = P_\lambda H$. Приближенным решением уравнения (II.9.2) назовем на этот раз решение уравнения

$$P_\lambda (A x_\lambda - f) = 0 , \quad A = A_0 + \mathcal{K} . \qquad (II.9.5)$$

Допустим, что при некотором α_0 , $0 \leq \alpha_0 \leq 1$, оператор

$A^{\alpha_0}\mathcal{H}A^{\alpha_0}$ вполне непрерывен в H. Пусть T – замыкание этого оператора в H; допустим, что уравнение $x+Tx=0$ имеет только тривиальное решение. Пусть еще точное решение $x_*\in\mathcal{D}(B^{\alpha_1})$. Тогда существует такое λ_0, что при $\lambda\geqslant\lambda_0$ уравнение (П.9.6) имеет одно и только одно решение, и при $\alpha_0\leqslant\alpha\leqslant\alpha_1$ верна оценка

$$\mathcal{E}_\lambda^{(\alpha)}\leqslant\|B^\alpha(x_*-x_\lambda)\|\leqslant C\delta_\lambda^{(\alpha)};$$

$$\mathcal{E}_\lambda^{(\alpha)}=\|B^\alpha x_*-P_\lambda B^\alpha x_*\|=O(\lambda^{\alpha-\alpha_1}) \qquad (\text{П.9.6})$$

3°. Рассмотрим краевую задачу

$$\sum_{|\alpha|,|\beta|=0}^{\mathfrak{s}}(-1)^{|\alpha|}\mathcal{D}^\alpha(A_{\alpha\beta}\mathcal{D}^\beta x)=f(t),\ t\in\Omega;$$
$$\mathcal{D}^\gamma x|_{\partial\Omega}=0\ ,\ 0\leqslant|\gamma|\leqslant\mathfrak{s}-1. \qquad (\text{П.9.7})$$

допустим, что оператор этой задачи, который мы обозначим через A, можно представить в виде $A=A_0+\mathcal{H}$, где оператор A_0 положительно определен в $L_2(\Omega)$ и оператор $T=A_0^{-1}\mathcal{H}$ вполне непрерывен в H_0 – энергетическом пространстве оператора A_0. Допустим далее, что граница области Ω достаточно гладкая, а задача (П.9.7) имеет не более одного решения; тогда решение этой задачи существует (может быть, как обобщенное). Примем, что это решение $x\in W_2^{(\nu)}(\Omega)$, где ν достаточно велико. В соответствии с частным случаем видоизмененной схемы Канторовича, описанным в п.2° § 7, положим

$$X=\overset{\circ}{W}{}_2^{(\mathfrak{s})}(\Omega),\ Y=W_2^{(\ell)}(\Omega),\ Z=L_2(\Omega).$$

Далее, пусть $\ell\geqslant0$, $\mathfrak{s}=2\mathfrak{s}+\ell$ (см. п.3° § 7). Введем те же координатные функции, что и в п.2° § 7, и определим проекторы тем же соотношением (П.7.6). Тогда остаются в силе все рассуждения § 7, и мы находим, что оценка (П.7.10) справедлива и для приближенного решения задачи (П.9.7), полученного по методу Бубнова – Галеркина при указанных выше координатных функциях.

Точно также нетрудно убедиться, что для задачи (П.9.7) верны оценки младших производных, полученных в § 8.

§ 10. Разностные методы.

Потребность аппроксимации дозволенных методов довольно хоро-

шо освещена в монографической литературе, и мы ограничимся здесь самыми краткими указаниями. Более подробно изложение можно найти, например, в монографиях Вазова и Форсайта [I], Бабушки, Витасека, Прагера [I], Марчука [I], Годунова и Рябенького [I].

I°. Рассмотрим задачу Дирихле для уравнения Лапласа в двумерной области Ω с достаточно гладкой границей:

$$\Delta x = \frac{\partial^2 x}{\partial t_1^2} + \frac{\partial^2 x}{\partial t_2^2} = 0\ ,\ t=(t_1,t_2)\subset\Omega\ ,\ x|_{\partial\Omega}=\varphi. \quad (\text{II.IO.I})$$

Построим квадратную сетку с шагом h. Обозначим через Ω_1 замкнутое подмножество Ω, обладающее тем свойством, что если $t\in\Omega_1$, то при любых числовых значениях ξ, $|\xi|\leqslant 1$, будет $(t_1+\xi h, t_2)\in \bar{\Omega}$ и $(t_1, t_2+\xi h)\in\bar{\Omega}$. В точках множества Ω_1 уравнение Лапласа аппроксимируется разностным уравнением

$$x(t_1,t_2)=\frac{1}{4}\left[x(t_1+h,t_2)+x(t_1-h,t_2)+x(t_1,t_2+h)+x(t_1,t_2-h)\right]; \quad (\text{II.IO.2})$$

в точках множества $\Omega\setminus\Omega_1$ для функции x строится некоторая линейная интерполяционная формула, позволяющая имитировать краевое условие.

Узлы сетки, лежащие в Ω_1, называются внутренними, а узлы, лежащие в $\Omega\setminus\Omega_1$ - граничными. Во внутренних узлах запишем уравнение (II.IO.2), в граничных - упомянутую выше интерполяционную формулу. В результате получим некоторую линейно-алгебраическую систему (сеточную систему), о которой известно, что она имеет единственное решение x_h.

Доказывается (см. Вазов и Форсайт [I]), что если точное решение задачи (II.IO.I) $x_*\in C^{(4)}(\bar{\Omega})$ и x_h - функция, определенная на узлах сетки, лежащих в $\bar{\Omega}$, и совпадающая там с функцией x_*, то

$$\|x_h - x_{h_1}\|_{L_\infty} \leqslant \left(\frac{1}{12}M_4 \sup_{\Omega} Q + M_2\right)h^2. \quad (\text{II.IO.3})$$

Здесь Q - некоторая функция, непрерывная и неотрицательная в $\bar{\Omega}$, M_k - верхняя граница модулей частных производных от x_* порядка k.

Близкие результаты содержатся в той же монографии для общего эллиптического уравнения 2-го порядка в области Ω любой размерности и для более общих краевых условий.

В монографии Бабушки, Витасека и Прагера [1] даны оценки погрешности аппроксимации разностного метода для бигармонического уравнения на плоскости.

2°. Рябенький и Филиппов доказали весьма интересную общую теорему об оценке погрешности аппроксимации разностных методов. Эта теорема подробно доказывается в монографиях Годунова и Рябенького [1] и Марчука [1]. Рассмотрим вычислительную задачу вида

$$Lx = f, \quad lx = \varphi. \tag{II.10.4}$$

Здесь x и f – функции, искомая и заданная, определенные в области $\Omega \subset R_m$, φ – функция, определенная на $\partial\Omega$, L и l – линейные операторы. Пусть задача (II.10.4) заменена разностной задачей

$$L_h x^h = f^h, \quad l_h x^h = \varphi^h, \tag{II.10.5}$$

в которой x^h, f^h, φ^h – сеточные функции, принадлежащие, вообще говоря, различным пространствам сеточных функций, так что $x^h \in X_h$, $f^h \in \mathcal{F}_h$, $\varphi^h \in \Phi_h$.

Пусть $(\cdot)_h$ – оператор, превращающий функцию, заданную в Ω или на $\partial\Omega$, в сеточную функцию, заданную на соответствующем множестве узлов. Теорема Рябенького – Филиппова состоит в следующем:

Пусть справедливы аппроксимационные неравенства

$$\|(Lx)_h - L_h x^h\|_{X_h} \le Ch^K, \quad \|(lx)_h - l_h x^h\|_{\Phi_h} \le Ch^K,$$
$$\|(f)_h - f^h\|_{\mathcal{F}_h} \le Ch^K, \quad \|(\varphi)_h - \varphi^h\|_{\Phi_h} \le Ch^K \tag{II.10.6}$$

и пусть обратный оператор задачи (II.10.5) ограничен независимо от h, так что

$$\|x^h\|_{X_h} \le C\{\|f^h\|_{\mathcal{F}_h} + \|\varphi^h\|_{\Phi_h}\}. \tag{II.10.7}$$

Тогда справедлива оценка погрешности аппроксимации

$$\|(x_*)_h - x_*^h\| \le Ch^K. \tag{II.10.8}$$

§ 11. Метод коллокации.

1°. Этот метод был впервые предложен Канторовичем и его

статье [I]. Сущность метода такова. Пусть требуется приближенно решить линейное уравнение

$$Ax = f : \quad x \in X , \ y \in Y , \qquad (\text{II.II.I})$$

(требование линейности несущественно), в котором X и Y – банаховы пространства, A – оператор, действующий из X в Y, причем элементы пространства Y образуют подмножество множества функций, непрерывных на некотором компакте $\mathcal{K} \subset R_m$. Выберем некоторую последовательность конечномерных подпространств $\{X_n\}$, полную в X, и положим $\dim X_n = N(n) = N$, $Y_n = AX_n$.

Пусть существует оператор A^{-1}, определенный на всем пространстве Y, тогда $\dim Y_n = N$ и последовательность $\{Y_n\}$ полна в Y. Далее, если $\{\varphi_{nk}\}$, $k = 1, 2, \ldots, N$ – базис в X_n, и $\psi_{nk}(t) = A\varphi_{nk}$, $t \in \mathcal{K}$, то $\{\psi_{nk}\}$ есть базис в Y_n.

Выберем в $\mathcal{K}$ точки $t_j = t_j^{(n)}$, $j = 1, 2, \ldots, N$, называемые узлами коллокации. Приближенное решение ищем как элемент подпространства X_n,

$$x^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk} \qquad (\text{II.II.2})$$

и определяем коэффициенты $a_k^{(n)}$ из условия, чтобы уравнение (II.II.1) выполнялось в узлах коллокации:

$$\sum_{k=1}^{N} a_k^{(n)} \psi_{nk}(t_j) = f(t_j). \qquad (\text{II.II.3})$$

2^0. В ряде работ исследуется погрешность аппроксимации рассматриваемого здесь метода. Приведем здесь важнейшие результаты некоторых из этих работ. Сразу же отметим, что некоторые оценки упомянутой погрешности, а также дополнительная библиография по этому вопросу, приведены в книгах Канторовича и Акилова [I] Красносельского и др. [I].

В работах Карпиловской [I – 3] изучена погрешность аппроксимации коллокационного метода для обыкновенного дифференциального уравнения порядка 2 при условиях первой краевой задачи. Пусть уравнение задано на промежутке $a \leq t \leq b$, и пусть точное решение $x_* \in C^{(r,\alpha)}[a,b]$, $0 < \alpha \leq 1$. Приближенное решение $x_*^{(n)}$ строится в виде полинома степени $r+2s$, удовлетворяющего краевым условиям задачи. Пусть коэффициенты и свободный член дифференциального уравнения принадлежат классу $C^{(r,\alpha)}[a,b]$, и пусть задача

имеет не более одного решения. Если в качестве узлов коллокации выбраны узлы Чебышева или Гаусса, то соответственно

$$\| x_* - x_*^{(n)} \|_{C(S)} \leqslant C \frac{\ln n}{n^{r+\alpha}} \qquad (\text{II.II.4})$$

$$\| x_* - x_*^{(n)} \|_{C(S)} \leqslant \frac{C}{n^{r+\alpha-1/2}} . \qquad (\text{II.II.5})$$

В цитированных работах Карпиловской рассмотрена также первая краевая задача для уравнения

$$\Delta x + \lambda p(t) x = f(t) \qquad (\text{II.II.6})$$

в квадрате $R: 0 \leqslant t_1, t_2 \leqslant \pi$. За координатные функции взяты собственные функции той же задачи для уравнения Лапласа:

$$\varphi_{ik}(t) = \sin it_1 \sin kt_2 ; \quad 0 < i \leqslant m , \quad 0 < k \leqslant n ;$$

за узлы коллокации - точки $\left(\frac{2i-1}{2m+1}\pi , \frac{2k-1}{2n+1}\pi \right)$. Если λ отлично от собственных чисел рассматриваемой задачи, и $p, f \in Lip_\alpha(R)$, то ($x_*^{(m,n)}$ - приближенное решение)

$$\| \Delta(x_* - x_*^{(m,n)}) \|_{C(R)} \leqslant C \ln^2 m \ln^2 n \left(\frac{1}{m^2} + \frac{1}{n^2} \right). \qquad (\text{II.II.7})$$

В статьях Карпиловской [2, 3] рассмотрены и некоторые другие задачи.

3°. В работе Габдулхаева [1] рассматривается интегральное уравнение Фредгольма

$$Ax(t) := \mu x(t) - \frac{1}{2\pi} \int_0^{2\pi} K(t,s) x(s)\, ds = f(t) \qquad (\text{II.II.8})$$

в пространстве 2π-периодических функций с С-нормой. Приближенное решение строится в виде кусочно линейной функции с угловыми точками $t_j = t_j^{(n)} = 2\pi j / n$; эти же точки суть узлы коллокации. Доказывается следующее утверждение. Пусть μ отлично от собственных чисел уравнения (II.II.3), а n таково, что

$$\gamma := \omega_1 \left(K, \frac{2\pi}{n} \right) \| A^{-1} \| < 1. \qquad (\text{II.II.9})$$

Тогда алгебраическая система метода коллокации имеет единственное решение. При этом, если $x_*(t)$ - точное, а $x_*^{(n)}(t)$ - приближенное решение уравнения (II.II.8), то

$$\|x_* - x_*^{(n)}\|_C \le (1 + P_A/(1-\tau))\,\omega(x_*, \tfrac{2\pi}{n}). \qquad (\text{II.II.10})$$

Здесь $P_A = \|A\|\cdot\|A^{-1}\|$ – число обусловленности оператора A, а ω в двух последних формулах означает модуль непрерывности.

В той же статье Габдулхаева [I] содержатся и некоторые результаты, относящиеся к сингулярным одномерным интегральным уравнениям.

4°. В книге Зильбермана и Прессдорфа [I](см. также книгу Михлина и Прессдорфа [I]) метод коллокации применен к сингулярному одномерному интегральному уравнению вида

$$a(t)x(t) + \frac{b(t)}{\pi i}\int_\Gamma \frac{x(s)}{s-t}\,ds + Tx = f(t). \qquad (\text{II.II.II})$$

Здесь Γ – окружность $|t| = 1$, $a(t)$ и $b(t)$ непрерывны и символ не вырождается, так что $a^2(t) - b^2(t) \neq 0$, $t \in \Gamma$; T – оператор вполне непрерывный в $L_p(\Gamma)$, $1 < p < \infty$. Координатные функции – тригонометрические полиномы порядка n относительно $\theta = \arg t$, узлы коллокации совпадают с точками $t_j^{(n)} = \exp(\frac{2\pi i j}{n})$. Пусть

$$\tilde a(t) = a(t) + b(t),\ \tilde b(t) = a(t) - b(t);\ \operatorname{ind}\tilde a(t) = \operatorname{ind}\tilde b(t);$$
$$a(t),\ b(t) \in C^{(\tau,\mu)}(\Gamma);\quad T: L_p(\Gamma) \longrightarrow C^{(\tau,\nu)}(\Gamma).$$

Тогда

$$\|x_* - x_*^{(n)}\|_p = O(n^{-\tau-\lambda}),\ \lambda = \min(\mu,\nu). \qquad (\text{II.II.I2})$$

Некоторые результаты получены и для уравнений вида (II.II.II) с вырождающимся символом.

Прессдорф и Шмидт [I] рассмотрели метод коллокации для уравнения (II.II.II) при $T = 0$ с узлами $t_j^{(n)} = \exp(\frac{2\pi j}{n})$ и с кусочно линейными координатными функциями, угловые точки которых совпадают с узлами коллокации. Основной результат упомянутой статьи состоит в следующем: пусть коэффициенты $a(t)$ и $b(t)$ непрерывны. Для того, чтобы полученный здесь метод коллокации сходился в $L_2(\Gamma)$ необходимо и достаточно, чтобы

$$a(t) + \lambda b(t) \neq 0;\ \forall t \in \Gamma,\ \forall \lambda \in [-1, +1]. \qquad (\text{II.II.I3})$$

Если при этом $a, b \in \operatorname{Lip}_\mu(\Gamma)$, $f \in \operatorname{Lip}_\lambda(\Gamma)$, то

$$\|x_* - x_*^{(n)}\|_{L_2\rho_\nu(\Gamma)} = O(n^{-\nu}\ln n), \quad \nu = \min(\lambda, \mu). \text{ (II.II.I4)}$$

Некоторые результаты получены в цитированной статье и для разрывных символов.

5°. Ряд результатов по оцс ке погрешности аппроксимации для уравнений Фредгольма и одномерных сингулярных интегральных уравнений содержится в книге Габдулхаева [3].

ГЛАВА III

Погрешность искажения

§ I. Погрешность искажения линейного свободного вычислительного процесса

I°. Погрешность искажения вычислительных процессов исследована в ряде работ автора и его учеников; первая из работ автора этого направления [4] опубликована в 1960 г.; работы периода 1960 - 65 гг. суммированы в книге автора [8]; эти работы тесно связаны понятием, тогда же введенным автором, об устойчивости вычислительного процесса по отношению к искажению. В книге [8] содержатся и не публиковавшиеся ранее результаты, в основном касающиеся нелинейных вычислительных процессов; эти процессы будут рассмотрены ниже, в главах УП-IX. Ряд результатов был позднее получен для погрешности искажения метода конечных элементов (см. работы автора [I3, I4]); результаты этого плана будут рассмотрены в гл.У.

В данной главе мы изложим результаты, относящиеся к искажению линейных вычислительных процессов и полученные частично в упомянутых выше более старых работах, частично в более новой работе автора [33].

2°. Рассмотрим линейный свободный вычислительный процесс, состоящий в решении последовательности независимых уравнений

$$A_n x^{(n)} = f^{(n)}; \quad x^{(n)} \in X_n, \ f^{(n)} \in Y_n, \ n = 0,1,2,\dots \quad (\text{III.I.I})$$

Здесь X_n и Y_n – банаховы пространства, A_n – замкнутый линейный оператор, отображающий X_n на Y_n. Мы принимаем, что операторы A_n ограниченно обратимы, так что каждый из операторов A_n^{-1} существует, ограничен и определен на всем пространстве Y_n.

Искаженный вычислительный процесс, который естественно также считать линейным, имеет вид

$$(A_n + \Gamma_n)\tilde{x}^{(n)} = f^{(n)} + \delta^{(n)}; \quad n = 0,1,2,\dots . \quad (\text{III.I.2})$$

Можно считать, что оператор Γ_n – искажение оператора A_n в том или ином смысле мал по сравнению с A_n. Примем, что величина $\|A_n^{-1}\Gamma_n\|$ может быть сделана сколь угодно малой. Норму элемента $\delta^{(n)}$ также можно считать сколь угодно малой, однако в предлагаемой линейной теории это допущение не играет существенной роли

для последующих рассуждений.

Оценим погрешность искажения процесса (Ш.I.I). Зададим число β , $0<\beta<1$, и потребуем, чтобы $\|A_n^{-1}\Gamma_n\| \le \beta$. При таком условии оператор $A_n+\Gamma_n$ ограниченно обратим. Действительно (I_n – тождественный оператор в X_n) $A_n+\Gamma_n = A_n(I_n+A_n^{-1}\Gamma_n)$. Первый множитель ограниченно обратим по предположению, второй – по известной теореме Банаха, и

$$(A_n+\Gamma_n)^{-1} = (I_n+A_n^{-1}\Gamma_n)^{-1}A_n^{-1}. \qquad (Ш.I.3)$$

Решение уравнения (Ш.I.2) существует и единственно:

$$x_*^{(n)} = (I_n+A_n^{-1}\Gamma_n)^{-1}A_n^{-1}(f^{(n)}+\delta^{(n)}).$$

Решение уравнения (Ш.I.I) $x_*^{(n)} = A_n^{-1}f^{(n)}$, поэтому

$$\tilde{x}_*^{(n)} - x_*^{(n)} = \left[(I_n+A_n^{-1}\Gamma_n)^{-1}-I_n\right]x_*^{(n)} + (I_n+A_n^{-1}\Gamma_n)^{-1}A_n^{-1}\delta^{(n)}. \qquad (Ш.I.4)$$

Если I тождественный, а $I+\gamma$ ограниченно обратимый оператор в некотором банаховом пространстве, то $(I+\gamma)[(I+\gamma)^{-1}-I]=-\gamma$ и, следовательно $(I+\gamma)^{-1}-I = -(I+\gamma)^{-1}\gamma$. В частности,

$$(I_n+A_n^{-1}\Gamma_n)^{-1}-I_n = -(I_n+A_n^{-1}\Gamma_n)^{-1}A_n^{-1}\Gamma_n; \qquad (Ш.I.5)$$

замечая еще, что $\|(I_n+A_n^{-1}\Gamma_n)^{-1}\| \le (1-\beta)^{-1}$, мы приходим к искомой оценке погрешности искажения:

$$\hat{\xi}_n = \|\tilde{x}_*^{(n)} - x_*^{(n)}\| \le \frac{1}{1-\beta}\left[\|A_n^{-1}\Gamma_n x_*^{(n)}\| + \|A_n^{-1}\delta^{(n)}\|\right]. \qquad (Ш.I.6)$$

Отсюда вытекает несколько более простая и более удобная для применения оценка:

$$\hat{\xi}_n \le \frac{1}{1-\beta}\left[\|A_n^{-1}\Gamma_n\|\cdot\|x_*^{(n)}\| + \|A_n^{-1}\delta^{(n)}\|\right]. \qquad (Ш.I.7)$$

Формула (Ш.I.7) позволяет оценить погрешность искажения процесса (Ш.I.I), если известны оценки точных решений $x_*^{(n)}$ уравнения (Ш.I.I).

§ 2. Устойчивость свободного процесса относительно искажения

Представляет интерес установить условия, при которых погрешность искажения остается малой независимо от n , если

в том или ином смысле малы искажения Γ_n и $\delta^{(n)}$. Такие условия связаны с понятием устойчивости вычислительного процесса относительно искажения; это понятие сформулировано в книге автора [8] и несколько уточнено в его работах [17, 22].

Целесообразными оказываются два несколько различных определения устойчивости. Согласно первому из них мы называем процесс (III.1.1) устойчивым относительно погрешностей искажения в последовательности пар пространств (X_n, Y_n), если существуют такие положительные числа ρ, q, τ, что из неравенства $\|A_n^{-1}\Gamma_n\| \leqslant \tau$ следует однозначная разрешимость уравнения (III.1.2) при любом n и неравенство

$$\|\tilde{x}_*^{(n)} - x_*^{(n)}\| \leqslant \rho \|A_n^{-1}\Gamma_n\| + q\|\delta^{(n)}\|. \qquad (III.2.1)$$

Второе определение имеет смысл, если искажения Γ_n суть ограниченные операторы. В этом случае можно назвать процесс (III.1.1) устойчивым относительно искажения в последовательности пар пространств (X_n, Y_n), если существуют такие положительные постоянные ρ, q, τ, что из неравенства $\|\Gamma_n\| \leqslant \tau$ вытекает однозначная разрешимость уравнения (III.1.2) при любом n и неравенство

$$\|\tilde{x}_*^{(n)} - x_*^{(n)}\| \leqslant \rho \|\Gamma_n\| + q\|\delta^{(n)}\|. \qquad (III.2.2)$$

Ниже мы для краткости будем говорить, что вычислительный процесс устойчив (или неустойчив), если он устойчив (или неустойчив) относительно погрешностей искажения; там, где это не может вызвать недоразумений, мы будем опускать слова "в последовательности пар пространств (X_n, Y_n)".

2^0. Теорема III.2.1. Для того, чтобы вычислительный процесс (III.1.1) был устойчив в смысле первого определения, необходимо достаточно, чтобы

$$\|A_n^{-1}\| \leqslant C_1, \quad \|x_*^{(n)}\| \leqslant C_2; \quad C_1, C_2 = const. \qquad (III.2.3)$$

Достаточность условий (III.2.3) сразу вытекает из оценки (III.1.7): если указанные условия выполнены, то можно положить $\tau = \beta$, $0 < \beta < 1$, и из (III.1.6) следует неравенство (III.2.1) с постоянными

$$\rho = \frac{C_1}{1-\beta}, \quad q = \frac{C_1}{1-\beta}.$$

Необходимость. Пусть процесс (III.1.1) устойчив в смысле па

вого определения. Допустим, что $\Gamma_n = 0$, так что искажаются только свободные члены уравнений (Ш.1.1). Неравенство (Ш.2.1) принимает вид $\|y^{(n)}\| \leq q \|\delta^{(n)}\|$, где $y^{(n)} = z_*^{(n)} - x_*^{(n)}$. Далее, в данном случае $A_n z_*^{(n)} = f^{(n)} + \delta^{(n)}$, поэтому $A_n y^{(n)} = \delta^{(n)}$ и, следовательно, $\|A_n y^{(n)}\| \geq q^{-1} \|y^{(n)}\|$. Это означает, что нормы $\|A_n^{-1}\|$ ограничены независимо от n; можно положить $C_1 = q^{-1}$.

Чтобы доказать необходимость второго условия (Ш.2.3), допустим, что $\delta^{(n)} = 0$. Тогда, если только $\|A_n^{-1}\Gamma_n\| \leq \tau$, то

$$\|z_*^{(n)} - x_*^{(n)}\| \leq \rho \|A_n^{-1}\Gamma_n\|. \qquad (Ш.2.4)$$

Положим $\Gamma_n = \tau A_n$. Уравнение (Ш.1.2) принимает вид $(1+\tau) A_n x^{(n)} = f^{(n)}$. Отсюда $x_*^{(n)} = (1+\tau)^{-1} x_*^{(n)}$ и из (Ш.2.4) следует, что $\|x_*^{(n)}\| \leq \rho(1+\tau) = const$.

3°. Теорема Ш.2.2. (см. работы автора [7, 8]). Для того, чтобы процесс (Ш.1.1) был устойчив в смысле второго определения в последовательности (X_n, Y_n), необходимо и достаточно, чтобы были выполнены следующие условия: 1) $\|A_n^{-1}\| \leq C_1 = const$; 2) существует такая постоянная C_3, что каков бы ни был линейный оператор B_n с единичной нормой, действующий из X_n в Y_n, справедливо неравенство

$$\|A_n^{-1} B_n x_*^{(n)}\| \leq C_3. \qquad (Ш.2.5)$$

Необходимость условия $\|A_n^{-1}\| \leq C_1$ доказывается так же, как и в теореме Ш.2.1. Докажем необходимость условия (Ш.2.5). Пусть $\delta^{(n)} = 0$. Положим $\Gamma_n = \varepsilon B_n$, $\varepsilon = const < \tau$. Тогда $(A_n + \Gamma_n) z_*^{(n)} = f^{(n)}$, отсюда $z_*^{(n)} + A_n^{-1}\Gamma_n z_*^{(n)} = x_*^{(n)}$ и

$$z_*^{(n)} - x_*^{(n)} + A_n^{-1}\Gamma_n (z_*^{(n)} - x_*^{(n)}) = -A_n^{-1}\Gamma_n x_*^{(n)}. \qquad (Ш.2.6)$$

Далее,

$$\|A_n^{-1}\Gamma_n x_*^{(n)}\| = \varepsilon \|A_n^{-1} B_n x_*^{(n)}\| = \|\Gamma_n\| \cdot \|A_n^{-1} B_n x_*^{(n)}\|;$$

$$\|A_n^{-1}\Gamma_n (z_*^{(n)} - x_*^{(n)})\| = \varepsilon \|A_n^{-1} B_n (z_*^{(n)} - x_*^{(n)})\| \leq$$

$$\leqslant \varepsilon \|A_n^{-1}\| \cdot \|z_*^{(n)} - x_*^{(n)}\| \leqslant \varepsilon C_1 \|z_*^{(n)} - x_*^{(n)}\|,$$

и из (Ш.2.6) следует

$$\|z_*^{(n)} - x_*^{(n)}\| \geqslant \frac{\|\Gamma_n\| \cdot \|A_n^{-1} B_n x_*^{(n)}\|}{1+\varepsilon C_1}.$$

Сравнив это с неравенством устойчивости (Ш.2.2), которое в данном случае принимает вид $\|z_*^{(n)} - x_*^{(n)}\| \leqslant \rho \|\Gamma_n\|$, получаем

$$\|A_n^{-1} B_n x_*^{(n)}\| \leqslant \rho(1+\varepsilon C_1).$$

Полагая здесь $\varepsilon \to 0$, находим, что условие (Ш.2.5) выполнено со значением $C_3 = \rho$.

Теперь докажем достаточность условий теоремы. Положим $\tau = \beta/C_1$, $\beta = const$, $0<\beta<1$, и потребуем, чтобы $\|\Gamma_n\| \leqslant \tau$. Если $\Gamma_n = 0$, то $z_*^{(n)} = x_*^{(n)} + A_n^{-1}\delta^{(n)}$, $\|z_*^{(n)} - x_*^{(n)}\| \leqslant C_1 \|\delta^{(n)}\|$ и теорема доказана. Если же $\Gamma_n \neq 0$, то положим $B_n = \Gamma_n / \|\Gamma_n\|$ и $z_*^{(n)} = u^{(n)} + v^{(n)}$, где

$$(A_n + \Gamma_n) u^{(n)} = f^{(n)}, \quad (A_n + \Gamma_n) v^{(n)} = \delta^{(n)}. \qquad (Ш.2.7)$$

Имеем

$$\|v^{(n)}\| \leqslant \|(I_n + A_n^{-1}\Gamma_n)^{-1}\| \cdot \|A_n^{-1}\| \cdot \|\delta^{(n)}\| \leqslant \frac{C_1}{1-\beta} \|\delta^{(n)}\|. \qquad (Ш.2.8)$$

По формуле (Ш.I.5) находим из первого уравнения (Ш.2.7)

$$\|u^{(n)} - x_*^{(n)}\| = \|(I_n + A_n^{-1}\Gamma_n)^{-1} A_n^{-1} \Gamma_n x_*^{(n)}\| \leqslant \frac{\|A_n^{-1} B_n x_*^{(n)}\|}{1-\beta} \|\Gamma_n\| \leqslant \frac{C_2}{1-\beta} \|\Gamma_n\|;$$

неравенство устойчивости (Ш.2.2) выполняется при значениях постоянных $\rho = (1-\beta)^{-1} C_2$, $q = (1-\beta)^{-1} C_1$.

4°. Следствие Ш.2.I. Если нормы $\|x_*^{(n)}\|$ ограничены в совокупности, то для устойчивости процесса (Ш.I.I) в смысле второго определения необходимо и достаточно, чтобы $\|A_n^{-1}\| \leqslant C_1 = const$.

Отметим частный случай, обычно встречающийся при реализации приближенных методов. Пусть X_n суть подпространства пространства X, и пусть $x_*^{(n)} \underset{n\to\infty}{\longrightarrow} x_*$; в этом случае назовем вычисли-

тельный процесс (Ш.I.I) сходящимся. Для такого процесса величины $\|x_*^{(n)}\|$ ограничены, и для его устойчивости необходимо и достаточно, чтобы $\|A_n^{-1}\| \leqslant C_1 = const$.

Следствие Ш.2.2. (Суворов [I]). Для устойчивости процесса (Ш.I.I) в смысле второго определения необходимо и достаточно, чтобы

$$\|A_n^{-1}\| \leqslant C_1 \ , \ \|A_n^{-1}\| \cdot \|x_*^{(n)}\| \leqslant C_3 \ ; \ C_1, C_3 = const \ . \qquad (Ш.2.9)$$

Достаточность условий (Ш.2.9) и необходимость первого из них очевидны; докажем необходимость второго условия.

Пусть процесс устойчив. В силу теоремы Ш.2.2, $\|A_n^{-1} B_n x_*^{(n)}\| \leqslant C_2$, $\forall B_n$, $\|B_n\| = 1$. Допустим, что произведение $\|A_n^{-1}\| \cdot \|x_*^{(n)}\|$ неограничено. Перейдя к подпоследовательности, можно считать, что $\|A_n^{-1}\| \cdot \|x_*^{(n)}\| \xrightarrow[n \to \infty]{} \infty$. Для каждого n найдем такой элемент $y^{(n)}$, чтобы $\|y^{(n)}\| = I$ и $\|A_n^{-1} y^{(n)}\| \geqslant \frac{1}{2} \|A_n^{-1}\|$. Построим далее линейный функционал ℓ_n, определенный на X_n, со свойствами $\|\ell_n\| = 1$, $|\ell_n x_*^{(n)}| = \|x_*^{(n)}\|$. Положим $B_n x^{(n)} = \ell_n x^{(n)} \cdot y^{(n)}$; ясно, что $\|B_n\| = I$. Имеем $A_n^{-1} B_n x_*^{(n)} = \|x_*^{(n)}\| \cdot A_n^{-1} y^{(n)}$ и $\|A_n^{-1} B_n x_*^{(n)}\| \geqslant \frac{1}{2} \|A_n^{-1}\| \cdot \|x_*^{(n)}\| \xrightarrow[n \to \infty]{} \infty$, что противоречит предположению.

Из следствия Ш.2.2 сразу вытекает

Следствие Ш.2.3. Если нормы $\|A_n^{-1}\|$ положительно ограничены снизу, то для устойчивости процесса (Ш.I.I) в смысле второго определения необходимо и достаточно, чтобы $\|A_n^{-1}\| \leqslant C_1$, $\|x_*^{(n)}\| \leqslant C_2$; $C_1, C_2 = const$. Если еще процесс (Ш.I.I) - сходящийся, то для его устойчивости необходимо и достаточно, чтобы $\|A_n^{-1}\| \leqslant C_1$.

Следствие Ш.2.3 доказано в книге автора [8] независимо от следствия Ш.2.2.

5°. Если $\sup_n \|A_n^{-1}\| = \infty$, то процесс (Ш.I.I) неустойчив в последовательности (X_n, Y_n). Рассмотрим общий случай, когда $\sup_n \|A_n^{-1}\|_{Y_n \to X_n} \leqslant \infty$. Введем новые пространства $\bar{X}_n$ и $\bar{Y}_n$, которые совпадают соответственно с X_n и Y_n как множества элементов, но нормы в них определяются формулами

$$\|\cdot\|_{\bar{X}_n} = \frac{1}{\sqrt{\|A_n^{-1}\|}}\|\cdot\|_{X_n}\ ;\quad \|\cdot\|_{\bar{Y}_n} = \sqrt{\|A_n^{-1}\|}\cdot\|\cdot\|_{Y_n}. \qquad (\text{Ш.2.10})$$

Теорема Ш.2.3. Для того, чтобы процесс (Ш.I.I) был устойчив в смысле второго определения в последовательности $(\bar{X}_n, \bar{Y}_n)$, необходимо и достаточно выполнение неравенства

$$\|x_*^{(n)}\|_{\bar{X}_n} \le C = const. \qquad (\text{Ш.2.11})$$

Доказательство сразу вытекает из соотношений (Ш.2.9), написанных для пространств $\bar{X}_n$, $\bar{Y}_n$, и из очевидного тождества $\|A_n^{-1}\|_{\bar{Y}_n \to \bar{X}_n} = 1$.

Замечание Ш.2.I. Теорема Ш.2.3 остается в силе, если нормы в $\bar{X}_n$ и $\bar{Y}_n$ определить формулами

$$\|\cdot\|_{\bar{X}_n} = r(n)\|\cdot\|_{X_n};\ \|\cdot\|_{\bar{Y}_n} = \frac{1}{r(n)}\|\cdot\|_{Y_n}, \qquad (\text{Ш.2.12})$$

где $r(n)$ – любая положительная функция от n, такая, что

$$\left[\|A_n^{-1}\|_{Y_n \to X_n}\right]^{-1} \ge C r^2(n),\ C = const. \qquad (\text{Ш.2.13})$$

§ 3. Устойчивость процесса Ритца

I°. Метод Ритца коротко описан выше, в § I, гл.II. Обозначим соответственно через $a^{(n)}$ и $f^{(n)}$ векторы с составляющими $a_k^{(n)}$ и (f, y_{nk}), $k = 1, 2, \ldots, N$, а через M_n – матрицу элементов $[y_{nk}, y_{nj}]$; $k, j = 1, 2, \ldots, N$ – матрицу Ритца, и запишем систему (II.I.4) в виде

$$M_n a^{(n)} = f^{(n)};\quad n = 1, 2, \ldots . \qquad (\text{Ш.3.I})$$

Будем рассматривать $a^{(n)}$ и $f^{(n)}$ как элементы пространства R_n, а M_n как оператор в R_n. Если применить к данному случаю общую схему § I гл.I, то $X = H_A$, $X_n = R_n$, $x^{(n)} \quad a^{(n)}$; соотношение τ_n задается системой (II.I.4) или, что то же, системой (Ш.3.I). Наш оператор P_n, отображающий R_n в H_A, определяется формулой (II.1.3).

2°. С методом Ритца связаны два вычислительных процесса: процесс (Ш.3.I), приводящий к вычислению вектора коэффициентов $a^{(n)}$, и процесс вычисления приближенных решений $x_*^{(n)}$. Если обозначить через $a_*^{(n)}$ решение уравнения (Ш.3.I), то

$$x_*^{(n)} = P_n a_*^{(n)}\ ; \qquad (\text{Ш.3.2})$$

имея в виду применить общие теоремы § 2 настоящей главы, запишем этот процесс иначе. Будем рассматривать ρ_n как оператор из R_n в $X_n = H_A^{(n)}$, тогда существует обратный оператор $S_n = \rho_n^{-1}$: $S_n x^{(n)} = a^{(n)}$. Подставив это в (III.1.2), получим процесс, определяющий приближенные решения:

$$M_n S_n x^{(n)} = f^{(n)}. \qquad (III.3.3)$$

Ниже в данном параграфе мы исследуем устойчивость обоих указанных здесь процессов в смысле второго определения § 2 настоящей главы.

3°. Пусть в некотором гильбертовом пространстве задана последовательность вида (II.1.5) Пусть, далее, G_n – матрица Грама элементов $y_{n1}, y_{n2}, \ldots, y_{nN}$ и $\lambda_n^{(1)}$ – наименьшее собственное число этой матрицы. Последовательность (II.1.5) назовем сильно минимальной в H, если числа $\lambda_n^{(1)}$ ограничены снизу, положительным числом, не зависящим от n.

Понятие сильной минимальности было [illegible]лыкин [1] для последовательностей вида (II.1.6); некоторые свойства таких последовательностей рассмотрены в книге автора [8].

Теорема III.3.1. Для того, чтобы процесс (III.3.1) вычисления коэффициентов Ритца был устойчив в последовательности (R_n, R_n), необходимо и достаточно, чтобы координатная система была сильно минимальной в энергетическом пространстве рассматриваемой задачи.

Матрица Ритца совпадает с матрицей Грама (в энергетической метрике) координатных элементов $y_{n1}, y_{n2}, \ldots, y_{nN}$. Эта матрица – симметричная и положительно определенная, и ее собственные числа положительны; пусть $\lambda_n^{(1)}$ – наименьшее из них. Обозначим через A_n оператор, порожденный матрицей M_n и действующий в R_n. В данном случае элемент $x_*^{(n)}$ (§ 2 гл.I) следует отождествить с $a_*^{(n)}$ – решением системы (III.3.1). Очевидно, $\|A_n\| = 1/\lambda_n^{(1)}$. По теореме II.2.1, для устойчивости необходимо, чтобы $\|A_n^{-1}\| \leqslant C = const$, или $\lambda_n^{(1)} \geqslant 1/C_1$. Чтобы доказать достаточность условия теоремы, остается проверить (см. следствие II.2.1), что нормы $\|a_*^{(n)}\|$ ограничены. Процесс Ритца – сходящийся в H_A, поэтому $\|x_*^{(n)}\| \leqslant C = const$. Далее,

$$\|x_*^{(n)}\|^2 =$$

$$= \Big[\sum_{k=1}^{N} a_k^{(n)} \varphi_{nk} \,,\, \sum_{j=1}^{N} a_j^{(n)} \varphi_{nj}\Big] = (M_n a_*^{(n)}, a_*^{(n)}) \geqslant \frac{1}{\lambda_n^{(1)}} \|a^{(n)}\|^2 .$$

Отюда $\|a_*^{(n)}\| \leqslant \sqrt{C_1}\, C$. Теорема доказана.

Для классического метода Ритца теорема Ш.3.1 доказана в работе автора [4]; общий случай рассмотрен здесь впервые.

4°. Теорема Ш.3.2. Для того, чтобы процесс (Ш.3.3) вычисления приближенных решений по Ритцу был устойчив в последовательности пар пространств $(H_A^{(n)}, R_n)$, необходимо и достаточно, чтобы координатная система была сильно минимальна в H_A.

Процесс (Ш.3.3) сходящийся в H_A. По следствию Ш.2.1, для доказательства теоремы Ш.3.2 необходимо и достаточно установить что нормы $\|A_n^{-1}\|_{R_N \to H_A}$ ограничены. В данном случае следует отождествить оператор A_n с $M_n S_n = M_n \rho_n^{-1}$, так что $A_n^{-1} = \rho_n M_n^{-1}$. Имеем $x_*^{(n)} = \rho_n a_*^{(n)}$. Как мы видели выше, $\|x_*^{(n)}\| = (M_n a_*^{(n)}, a_*^{(n)})$ поэтому

$$\|\rho_n a_*^{(n)}\|^2 = (M_n a_*^{(n)}, a_*^{(n)}) = \|M_n^{1/2} a_*^{(n)}\|^2 .$$

Но $a_*^{(n)} = M_n^{-1} f^{(n)}$ и, следовательно,

$$\|\rho_n M_n^{-1} f^{(n)}\| = \|M_n^{-1/2} f^{(n)}\|_{R_N} ,$$

или

$$\|\rho_n M_n^{-1}\|_{R_N \to H_A^{(n)}} = \|M_n^{-1/2}\|_{R_N} = [\lambda_n^{(1)}]^{-1/2} .$$

Для того, чтобы последняя величина была ограничена, необходимо и достаточно, чтобы координатная система была сильно минимальна в энергетической норме. Теорема доказана.

Для классического метода Ритца достаточность условия теоремы Ш.3.2 доказана в статье Якубовской и Яковлева [1], а необходимость – в книге автора [8]. В общих случаях метод доказательства отличен от изложенного выше.

Замечание. Устойчивость вычислительных процессов, к которым приводит метод Ритца в задачах о спектрах несопряженного оператора (спектр предполагается дискретным), для классического метода Ритца исследована в статье автора [1]. Результаты этой статьи

с некоторыми изменениями изложены в книге автора [8].

5°. Исследуем устойчивость процесса Ритца в общем случае, когда $\inf_n \lambda_n^{(1)} \geq 0$. В последовательностях пар пространств (R_A, R_A) и $(H_A^{(n)}, R_A)$ эти процессы могут быть неустойчивыми (если $\inf_n \lambda_n^{(1)} = 0$), и мы введем вместо R_A новые пространства.

Пусть $\gamma(n)$ – такая положительная функция от n, что $\lambda_n^{(1)} \geq C_0^2 \gamma^2(n)$, $C_0 = const$. В частности, можно положить $\gamma^2(n) = \lambda_n^{(1)}$. Введем N-мерные гильбертовы пространства $\bar{X}_n$ и $\bar{Y}_n$, элементы которых суть N-компонентные числовые векторы, а нормы заданы формулами

$$\forall b \in R_A, \quad \|b\|_{\bar{X}_n} = \gamma(n)\|b\|_{R_A}, \quad \|b\|_{\bar{Y}_n} = \frac{1}{\gamma(n)}\|b\|_{R_A}. \qquad (\text{III.3.4})$$

Обозначим на этот раз через A_n оператор, порожденный матрицей M_n и действующий из $\bar{X}_n$ в $\bar{Y}_n$, и будем считать, что $a^{(n)} \in \bar{X}_n$, $f^{(n)} \in \bar{Y}_n$.

Теорема III.3.2. Процесс (III.3.1) устойчив в последовательности пар пространств $(\bar{X}_n, \bar{Y}_n)$.

Достаточно проверить, что выполнены неравенства

$$\|A_n^{-1}\|_{\bar{Y}_n \to \bar{X}_n} \leq C_1, \quad \|a_*^{(n)}\|_{\bar{X}_n} \leq C_2. \qquad (\text{III.3.5})$$

Оценим величину $\|A_n^{-1}\|_{\bar{Y}_n \to \bar{X}_n}$:

$$\|A_n^{-1}\|_{\bar{Y}_n \to \bar{X}_n} =$$

$$= \sup_{b \in R_A} \frac{\|A_n^{-1} b\|_{\bar{Y}_n}}{\|b\|_{\bar{X}_n}} = \gamma^2(n) \sup_{b \in R_A} \frac{\|M_n^{-1} b\|_{R_A}}{\|b\|_{R_A}} = \frac{\gamma^2(n)}{\lambda_n^{(1)}} \leq \frac{1}{C_0^2},$$

и первое из неравенств (III.3.5) установлено. Заметим, что если принять $\gamma(n) = \sqrt{\lambda_n^{(1)}}$, то получится $\|A_n^{-1}\|_{\bar{Y}_n \to \bar{X}_n} = 1$.

Пусть $v^{(n)}$ – произвольный элемент из $H_A^{(n)}$,

$$v^{(n)} = \sum_{k=1}^{n} b_k y_k.$$

Полагая $b = (b_1, b_2, \ldots, b_n)$, имеем

$$|v^{(n)}|^2 = (M_n b, b)_{R_A} \geq \lambda_n^{(1)} |b|^2_{R_A} \geq C_0^2 \|b\|^2_{\bar{X}_n}. \qquad (\text{III.3.6})$$

Метод Ритца сходится в H_A, поэтому $\|v^{(n)}\| \leq C = const$.

По неравенству (Ш.3.6) $\|a_*^{(n)}\| \leqslant C/c_0$ и второе неравенство (Ш.3.5) также установлено. Теорема доказана.

Замечание Ш.3.1. Пусть $\mu_n^{(1)}$ – наименьшее собственное число матрицы $\tilde{M}_n$ скалярных произведений $(\varphi_{nk}, \varphi_{nj})_H$; $j,k = 1,2,\ldots,N$. Достаточно подчинить $\gamma(n)$ неравенству $\mu_n^{(1)} \geqslant c^2\gamma^2(n)$, $c = const > 0$. Для доказательства достаточно заметить, что $\mu_n^{(1)} \leqslant c'\lambda_n^{(1)}$, $c' = const$. Действительно,

$$\mu_n^{(1)} = \inf_{b \in R_N} \frac{(\tilde{M}_n b, b)_{R_N}}{\|b\|^2_{R_N}} = \inf_{b \in R_N} \frac{\left\|\sum_{k=1}^{N} b_k \varphi_{nk}\right\|^2_H}{\|b\|^2_{R_N}},$$

$$\lambda_n^{(1)} = \inf_{b \in R_N} \frac{(M_n b, b)_{R_N}}{\|b\|^2_{R_N}} = \inf_{b \in R_N} \frac{\left|\sum_{k=1}^{N} b_k \varphi_{nk}\right|^2}{\|b\|^2_{R_N}}.$$

Пусть λ_0 – нижняя грань спектра оператора A, и пусть второй инфимум достигается при $b = \bar{b} = (\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_N)$. Тогда

$$\mu_n^{(1)} \leqslant \frac{\left\|\sum_{k=1}^{N} \bar{b}_k \varphi_{nk}\right\|^2_H}{\|\bar{b}\|^2_{R_N}} \leqslant \frac{1}{\lambda_0} \frac{\left|\sum_{k=1}^{N} \bar{b}_k \varphi_{nk}\right|^2}{\|\bar{b}\|^2_{R_N}} = \frac{\lambda_n^{(1)}}{\lambda_0},$$

и можно положить $C' = 1/\lambda_0$.

6°. Теорема Ш.3.4. Процесс (Ш.3.3) устойчив в последовательности пар пространств $(H_A^{(n)}, \bar{Y}_n)$.

Для доказательства этой теоремы мы воспользуемся приемом статьи Яковой и Яковлева [1].

Уравнение (Ш.3.1) запишем в виде

$$A_n a^{(n)} = f^{(n)}, \qquad (Ш.3.7)$$

где мы считаем, что $a^{(n)} \in \bar{X}_n$, $f^{(n)} \in \bar{Y}_n$, и обозначаем через A_n оператор, действующий из $\bar{X}_n$ в $\bar{Y}_n$ и порожденный матрицей M_n. Наряду с уравнением (Ш.3.7) рассмотрим искаженное уравнение

$$(A_n + \Gamma_n) C^{(n)} = f^{(n)} + \delta^{(n)}. \qquad (Ш.3.8)$$

Искаженное приближенное решение по Ритцу имеет вид

$$z_*^{(n)} = \sum_{k=1}^{N} C_k^{(n)} \varphi_{nk} = P_n C^{(n)}.$$

Отсюда

$$|z_*^{(n)} - x_*^{(n)}| = (M_n (c^{(n)} - a^{(n)}), c^{(n)} - a^{(n)})_{R_N} \leq \qquad (\text{Ш.3.9})$$

$$\leq \|A_n (c^{(n)} - a^{(n)})\|_{\bar{y}_n} \cdot \|c^{(n)} - a^{(n)}\|_{\bar{x}_n}.$$

В силу теоремы Ш.3.3 существуют такие числа $\rho, q, \tau > 0$, что если $\|\Gamma_n\|_{\bar{x}_n \to \bar{y}_n} \leq \tau$, то

$$\|c^{(n)} - a^{(n)}\|_{\bar{x}_n} \leq \rho \|\Gamma_n\|_{\bar{x}_n \to \bar{y}_n} + q \|\delta^{(n)}\|_{\bar{y}_n}. \qquad (\text{Ш.3.10})$$

Из (Ш.3.7) и (Ш.3.8) следует

$$(A_n + \Gamma_n)(c^{(n)} - a^{(n)}) = (I_n + \Gamma_n A_n^{-1}) A_n (c^{(n)} - a^{(n)}) = \delta^{(n)} - \Gamma_n a^{(n)},$$

I_n – тождественный оператор в $\bar{y}_n$. Как было доказано в п.5°, $\|A_n^{-1}\|_{\bar{y}_n \to \bar{x}_N} \leq 1/c_o^2$. Потребуем, чтобы $\|\Gamma_n\|_{\bar{x}_n \to \bar{y}_n} \leq \beta / c_o^2$, где β – какое-нибудь число из интервала $(0, 1)$. Тогда $\|(I_n + \Gamma_n A_n^{-1})\|_{\bar{y}_n} \leq (1-\beta)^{-1}$ и, так как нормы $\|a^{(n)}\|_{\bar{x}_n}$ ограничены, то

$$\|A_n (c^{(n)} - a^{(n)})\|_{\bar{y}_n} \leq \frac{1}{1-\beta} \left[\|\delta^{(n)}\|_{\bar{y}_n} + C' \|\Gamma_n\|_{\bar{x}_n \to \bar{y}_n} \right], \quad C' = const. \qquad (\text{Ш.3.11})$$

Из формул (Ш.3.9) – (Ш.3.11) очевидно следует, что

$$|z_*^{(n)} - x_*^{(n)}| \leq \rho_1 \|\Gamma_n\|_{\bar{x}_n \to \bar{y}_n} + q_1 \|\delta^{(n)}\|_{\bar{y}_n}; \qquad (\text{Ш.3.12})$$

за τ_1 можно принять число $\tau_1 = \min(\tau, \beta / c_o^2)$.

7°. Частными случаями теорем Ш.3.3 и Ш.3.4 являются сформулированные ниже, в гл.У, теоремы об устойчивости процесса конечных элементов.

8°. Пусть $\inf_n \lambda_n^{(1)} = 0$. Тогда процесс вычисления коэффициентов Ритца неустойчив в (R_N, R_N), а процесс вычисления приближенного решения неустойчив в (H_A, R_N). Это значит, что погрешность $\|c^{(n)} - a^{(n)}\|_{R_N}$ и $|z_*^{(n)} - x_*^{(n)}|$ могут быть велики, даже если малы погрешности $\|\Gamma_n\|_{R_N \to R_N}$ и $\|\delta^{(n)}\|_{R_N}$. Нетрудно, однако, убедиться, что оценка для $|z_*^{(n)} - x_*^{(n)}|$ в этом случае оказывается существенно меньшей, чем для величины $\|c^{(n)} - a^{(n)}\|_{R_N}$. Действитель-

но, из (III.3.10) и (III.3.12) следует

$$\|c^{(n)}-a^{(n)}\|_{R_E} \le \frac{1}{\gamma(n)}\left[\frac{p}{\gamma^2(n)}\|\Gamma_n\|_{R_E\to R_E}+\frac{q}{\gamma(n)}\|\delta^{(n)}\|_{R_E}\right] \le$$

$$\le \frac{\bar p}{\gamma(n)}\left[\frac{\|\Gamma_n\|_{R_E\to R_E}}{\gamma^2(n)}+\frac{\|\delta^{(n)}\|_{R_E}}{\gamma(n)}\right];\ \bar p=\max(p,q);$$

$$|z_k^{(n)}-\tau_k^{(n)}| \le \left[\frac{p'}{\gamma^2(n)}\|\Gamma_n\|_{R_E\to R_E}+\frac{q}{\gamma(n)}\|\delta^{(n)}\|_{R_E}\right] \le$$

$$\le \bar p'\left[\frac{\|\Gamma_n\|_{R_E\to z_E}}{\gamma^2(n)}+\frac{\|\delta^{(n)}\|_{R_E}}{\gamma(n)}\right];\ \bar p'=\max(p',q');$$

правая часть первого неравенства содержит малый делитель $\gamma(n)$, откуда и следует наше утверждение.

§ 4. Примеры.

1°. Пример I. Рассмотрим уравнение

$$(t+1)x(t)=1,\ 0\le t\le 1. \qquad \text{(III.4.1)}$$

Оператор умножения на $t+1$ положительно определенный в $L_2(0,1)$, и уравнение (III.4.1) равносильно задаче о минимуме функционала

$$\int_0^1 [(t+1)x^2(t)-2x(t)]dt. \qquad \text{(III.4.2)}$$

Эту последнюю задачу можно решать по методу Ритца. В качестве координатной системы возьмем последовательность t^n, $n=0,1,2,\dots$. По известной теореме Мюнца (см., например, Натансон [1]) эта система полна в $L_2(0,1)$, так как ряд обратных показателей $\sum_{n=1}^{\infty}\frac{1}{n}$ расходится. Из той же теоремы Мюнца легко вытекает (см. книгу автора [8]), что эта же система не сильно минимальна и процесс Ритца с такой координатной системой неустойчив в $L_2(0$. Посмотрим, как эта неустойчивость проявится в счете.

2°. При выбранной выше координатной системе приближенное решение имеет вид

$$x^{(n)}(t)=\sum_{k=1}^{n} a_k^{(n)} t^{k-1}, \qquad \text{(III.4.3)}$$

что приводит к нормальной системе Ритца

$$\sum_{k=1}^{n}\frac{2k-2j-3}{(k+j+1)(k+j+2)}a_k^{(n)}=\frac{1}{j+1},\ 0\le j\le n-1. \qquad \text{(III.4.4)}$$

Напишем расширенную матрицу Ритца 5-го порядка:

3	5	7	9	11	1
2	6	12	20	30	
5	7	9	11	13	1
6	12	20	30	42	2
7	9	11	13	15	1
12	20	30	42	56	3
9	11	13	15	17	1
20	30	42	56	72	4
11	13	15	17	19	1
30	42	56	72	90	5

(II.4.5.)

Как обычно, расширенные матрицы Ритца низшего порядка получаются усечением.

Системы Ритца порядка от 2 до 5 решались на ЭВМ БЭСМ-6. При этом оказалось, что для системы 5-го порядка наступает переполнение, и получить решение не удалось. Тем более не имело смысла решать системы Ритца более высоких порядков. Точные значения коэффициентов $a_k^{(n)}$, $n = 1, 2, 3, 4$ приведены в таблице 1. В таблице 2 приведены десятичные приближения тех же коэффициентов с шестью верными знаками. Приведены также десятичные приближения коэффициентов для $n = 5$, полученные следующим образом. Так как получить точные значения этих коэффициентов не удалось, то для $n = 5$ элементы расширенной матрицы Ритца были заменены их десятичными приближениями с 12 знаками; после этого были получены десятичные приближения коэффициентов, в которых мы сохранили 6 знаков.

Представляется интересным выяснить, насколько невязка приближение близко (в энергетической норме) к точному решению $x_*(t)$. В нашем примере энергетическая норма определяется формулой

$$\|x\|^2 = \int_0^1 (1+t)^2 x^2(t)\,dt \qquad (II.4.6)$$

и эквивалентна L_2-норме. Были вычислены значения $\|x_* - x_n\|$ для $n = 1, 2, 3, 4$. Результаты приведены в таблице 3.

Таблица I. Точные значения неискаженных коэффициентов Ритца

n \ k	I	2	3	4
I	2/3			
2	I2/I3	- 6/I3		
3	62/63	- 50/63	20/63	
4	320/321	- I00/I07	70/I07	- 70/32I

Таблица 2. Десятичные приближения неискаженных коэффициентов Ритца

	I	2	3	4	5
I	0,666667				
2	0,923077	-0,46I558			
3	0,98I270	-0,793651	0,3I7460		
4	0,996885	-0,934579	0,654206	-0,2I8069	
5	0,992859	-0,864970	0,567892	0,200555	-0,199993

Таблица 3. Погрешность аппроксимации приближенных решений

n	I	2	3	4	5
	0,I627	0,0220	0,005I	0,0008	0,0016

Таблица 3 показывает, что приближенные решения, вычисленные без погрешности довольно хорошо сходятся к точному решению в энергетической метрике. В классическом методе Ритца погрешность аппроксимации в энергетической метрике убывает с ростом n. Последнее число в таблице 5 показывает, что при $n = 5$ коэффициенты Ритца оказались вычисленными недостаточно точно.

В заключение настоящего пункта приведем таблицу значений искомой функции точных и приближенных, соответствующих значениям $n = 3, 4, 5$. Все значения даны с шестью верными знаками после запятой.

Таблица 4. Точные и приближенные неискаженные значения искомой функции

t	точные значения	приближенные значения		
		$n = 3$	$n = 4$	$n = 5$
0,0	I,000000	0,984310	0,995807	0,994871
0,I	0,909091	0,908003	0,909744	0,910432
0,2	0,833333	0,828072	0,835098	0,833902
0,3	0,769231	0,774517	0,773706	0,776198
0,4	0,714286	0,717339	0,721801	0,708751
0,5	0,666667	0,666538	0,678546	0,663192
0,6	0,625000	0,622114	0,641773	0,630532
0,7	0,588235	0,584066	0,606823	0,609512
0,8	0,555556	0,552595	0,576560	0,593628
0,9	0,526316	0,527100	0,541364	0,566283
I.0	0,500000	0,508182	0,499133	0,493396

Как мы видим, приближенные значения решений в фиксированных точках приближаются к точным значениям довольно медленно и не монотонно. Это связано с тем, что метод Ритца сходится, вообще говоря, только в энергетической метрике, которая в данном случае эквивалентна метрике L_2.

3°. Покажем расширенную матрицу Ритца, заменив все ее элементы их десятичными приближениями с четырьмя верными знаками после запятой. Приведем таблицу искаженных коэффициентов Ритца, вычисленных с шестью верными знаками после запятой.

Таблица 5. Искаженные коэффициенты Ритца

n \ k	I	2	3	4	5
3	0,98310	-0,794960	0,517048		
4	0,995807	-0,923202	0,627642	-0,201115	
5	0,994871	-0,882381	0,377335	0,258577	-0,255153

Сравнивая эти с таблицей 2, видим, что искажение коэффициентов сказывается уже, как правило, во втором или третьем знаке после запятой, тогда как в расширенной матрице Ритца и каковые

касалось только пятых знаков. Это связано с неустойчивостью процесса Ритца в данном примере в последовательности пространств (R_n, R_n).

3°. Рассмотрим теперь процесс решения задачи (III.4.I) по Ритцу, когда за координатные функции взяты полиномы Лежандра, умноженные на множители, по порядку роста совпадающие с нормирующими множителями:

$$\varphi_n(t) = [\sqrt{n}]\, P_{n-1}(2t-1).$$

Функции $\varphi_n(t)$ сильно минимальны в $L_2(0,1)$ (см. книгу автора [8]) а, следовательно, и в энергетической метрике задачи (III.4.I). При $n \leqslant 6$ получены точные (в виде простых дробей) значения Ритца; в таблице 6 приведены десятичные значения этих коэффициентов с шестью верными знаками после запятой. При взгляде на эту таблицу можно догадаться, что коэффициенты Ритца $a_k^{(n)}$ при возрастании n стремятся к некоторым пределам, а при возрастании k довольно быстро убывают. В книге автора [6] показано, что если координатная система классического метода Ритца сильно минимальна в энергетической норме, то существуют пределы $a_k = \lim_{n\to\infty} a_k^{(n)}$ и последовательность $\{a_k\} \in l_2$.

Расширенные матрицы были подвергнуты искажению, такому же как в п.3°: их элементы были заменены десятичными приближениями с четырьмя верными знаками, и полученные системы были решены с возможно большей точностью. Значения искаженных коэффициентов Ритца с шестью верными знаками приведены в таблице 7. Сравнение с таблицей 6 показывает, что искажения коэффициентов Ритца проявляются только в пятых и шестых знаках и, следовательно, являются величинами того же порядка, что и искажения элементов расширенной матрицы Ритца.

Таблица 6. Десятичные приближения неискаженных коэффициентов Ритца; координатные функции – полиномы Лежандра

	1	2	3	4	5	6
5	0,693146	-0,238316	0,054516	-0,005432		
6	0,693147	-0,238324	0,054565	-0,005615	0,001070	
7	0,693147	-0,238325	0,054567	-0,005620	0,001101	-0,000204
8	0,693147	-0,238325	0,054567	-0,005620	0,001102	-0,000210

Значения $a_6^{(\infty)} = 0{,}000038$

Таблица 7. Десятичные приближения искаженных коэффициентов Ритца; координатные функции - полиномы Лежандра

	I	2	3	4	5	6
3	0,693I46	-0,238314	0,054516	-0,005452		
4	0,693146	-0,238320	0,054564	-0,005615	0,001070	
5	0,693146	-0,238321	0,054566	-0,005620	0,001101	-0,000204
6	0,693146	-0,238321	0,054566	-0,005620	0,001102	-0,000210

Искаженные значения $a_5^{(n)} = 0{,}000038$.

5°. В пп. 5° и 6° настоящего параграфа мы вкратце опишем результаты вычислений по двум примерам, подробно рассмотренным в книге автора [8]. Будем решать по методу Ритца задачу, положительно определенную в $L_2(0, 1)$:

$$0 \leqslant t \leqslant 1, \quad -\frac{d^2x}{dt^2} = \frac{1}{1+t}; \quad x(0) = x'(1) = 0; \qquad (\text{III}.4.6)$$

В качестве координатной системы возьмем последовательность $\varphi_n(t) = t^n$, $n = 1,2,\ldots$; опираясь на теорему Мюнца, легко доказать, что эта система полна, но не сильно минимальна в энергетической метрике задачи (III.4.6).

Первые восемь приближений можно получить, решая линейную алгебраическую систему с расширенной матрицей (III.4.7)

I	I	I	I	I	I	I	I	I - $\ln 2$
I	4/3	3/2	8/5	5/3	I2/7	7/4	I6/9	$\ln 2$ - I/2
I	3/2	9/5	2	I5/7	9/4	7/3	I2/5	5/6 - $\ln 2$
I	8/5	2	I6/7	5/2	8/3	I4/5	32/II	$\ln 2$ - 7/I2
I	5/3	I5/7	5/2	25/9	3	35/II	I0/3	47/6 - $\ln 2$
I	I2/7	9/4	8/3	3	36/II	7/2	48/I3	$\ln 2$ - 37/60
I	7/4	7/3	I4/5	35/II	7/2	49/I3	4	3I9/420 - $\ln 2$
I	I6/9	I2/5	32/II	I0/3	48/I3	4	64/5	$\ln 2$ - 533/840

а также все системы, получаемые из этой усечением. Коэффициенты Ритца - решения упомянутых систем - обозначим через $a_k^{(n)}$. Решим эти точно, без погрешностей, а затем вместо полученных таким образом искаженных значений коэффициентов Ритца их десятичными приближениями с четырьмя верными знаками (таблица 8)

Таблица 8. Десятичные приближения неискаженных коэффициентов Ритца; координатные функции - степени t.

	I	2	3	4	5	6	7	8
I	0,3069							
2	0,6480	-0,34II						
3	0,6869	-0,4579	0,0788					
4	0,6922	-0,4899	0,I3I2	-0,0267				
5	0,6930	-0,4977	0,I547	-0,054I	0,0II0			
6	0,693I	-0,4995	0,I63I	-0,0708	0,0260	-0,0050		
7	0,693I	-0,4999	0,I657	-0,0786	0,0378	-0,0I36	0,0025	
8	0,693I	-0,5000	0,I664	0,0817	0,0446	0,02I8	0,0075	0,00I3

Искажим теперь расширенную матрицу (Ш.4.7), заменив в ней свободные члены их десятичными приближениями с четырьмя верными знаками после запятой. Искаженные коэффициенты Ритца, которые получим, решая систему с искаженной расширенной матрицей, а также соответствующие усеченные системы, обозначим через $b_k^{(n)}$. Полученные системы опять решим точно, а затем заменим коэффициенты $b_k^{(n)}$ их десятичными приближениями с четырьмя верными знаками (таблица

Таблица 9. Десятичные приближения искаженных коэффициентов Ритца

	I	2	3	4	5	6	7	8
I	0,3069							
2	0,6483	-0,34I4						
3	0,6883	-0,46I4	0,0800					
4	0,6974	-0,5I60	0,I7I0	-0,0455				
5	0,7I27	-0,6690	0,6300	-0,58I0	0,2I42			
6	0,7635	-I,428	4,I72	-7,665	6,590	-2,I25		
7	0,9I4I	-4,595	25,28	-7I,00	I0I,6	-7I,79	I9,9I	
8	I,627	-24,54	204,8	-8I9,I	I747	-2047	I243	-305,7

Сравнение таблиц 8 и 9 показывает, что погрешность искажения коэффициентов Ритца в данном случае весьма сильно возрастает вместе с n.

6°. Рассмотрим еще задачу

$$-1<t<1,\quad -\frac{d}{dt}\left[(2+t)\frac{dx}{dt}\right]=1,\quad x(-1)=x(1)=0,$$

(Ш.4.8

положительно определенную в $L_2(0, 1)$. В качестве координатной системы выберем последовательность интегралов от нормированных полиномов Лежандра

$$\varphi_n(t) = \sqrt{\frac{2n+1}{2}} \int_{-1}^{t} P_n(t)\,dt, \quad n = 1, 2, \ldots \qquad (\text{III}.4.9)$$

эта система сильно минимальна в энергетическом пространстве задачи (III.4.8) (см. книгу автора [8]). Как и в предшествующем примере, были составлены и точно решены системы Ритца от 1-го до 8-го порядка. В таблице 10 приведены десятичные приближения этих решений с четырьмя верными знаками после запятой. Далее системы Ритца были искажены: все элементы расширенных матриц были заменены их десятичными приближениями с четырьмя верными знаками после запятой. В этом случае получение точных решений оказалось затруднительным, поэтому, чтобы по возможности избежать ошибок округления, вычисления производились с большим числом десятичных знаков - с семью знаками для систем до 4-го порядка включительно и с 12 знаками для систем порядка от 5-го по 8-й.

В таблице 11 приведены искаженные значения коэффициентов Ритца с четырьмя верными знаками после запятой. Сравнение двух последних таблиц показывает, что в данном случае искажения коэффициентов Ритца суть величины того же порядка, что и искажения элементов расширенной матрицы Ритца.

Таблица 10. Десятичные приближения неискаженных коэффициентов Ритца.

	1	2	3	4	5	6	7	8
1	-0,4082							
2	-0,4374	0,1129						
3	-0,4395	0,1213	-0,0308					
4	-0,4397	0,1219	-0,0330	0,0083				
5	-0,4397	0,1219	-0,0332	0,0089	-0,0022			
6	-0,4397	0,1220	-0,0332	0,0090	-0,0024	0,0006		
7	-0,4397	0,1220	-0,0332	0,0090	-0,0024	0,0006	-0,0002	
8	-0,4397	0,1220	-0,0332	0,0090	-0,0024	0,0006	-0,0002	0,0001

7°. Численный пример, в котором отмечается явление неустойчивости, дан в работах автора [17] и [22].

Таблица II. Десятичные приближения искаженных коэффициентов Ритца.

	I	2	3	4	5	6	7	8
I	-0,4083							
2	-0,4374	0,II29						
3	-0,4396	0,I2I3	-0,0308					
4	-0,4397	0,I2I9	-0,0330	0,0083				
5	-0,4398	0,I220	-0,0332	0,0089	-0,0022			
6	-0,4398	0,I220	-0,0332	0,0089	-0,0024	0,0006		
7	-0,4398	0,I220	-0,0332	0,0089	-0,0024	0,0007	-0,0002	
8	-0,4398	0,I220	-0,0332	0,0089	-0,0024	0,0007	-0,0002	0,0001

§ 5. Об устойчивости процесса Бубнова - Галеркина

I°. В сепарабельном гильбертовом пространстве, которое для упрощения обозначений считаем вещественным, рассмотрим уравнение

$$Ax := A_0 x + \mathcal{K}x = f, \qquad (\text{III}.5.1$$

где A_0 - положительно определенный самосопряженный оператор и произведение $T = A_0^{-1}\mathcal{K}$ вполне непрерывно в энергетическом пространстве H_0 оператора A_0. В работе автора I948 г. (подробнее см. книгу автора [3]) установлены следующие теоремы:

а) Уравнение (3.5.1) эквивалентно уравнению

$$x + Tx = A^{-1}f := \tilde{f}, \qquad (\text{III}.5.2$$

рассматриваемому в H_0, в следующем смысле: любое решение уравнения (III.5.I) удовлетворяет уравнению (III.5.2), а любое решение уравнения (III.5.2) является обобщенным решением уравнения (III.5.

б) При заданной координатной системе метод Бубнова - Галеркина, примененный к уравнениям (III.5.I) и (III.5.2), приводит к одной и той же алгебраической системе.

в) Если координатная система полна в H_0, то метод Бубнова - Галеркина для уравнения (III.5.I) (а, следовательно, и для уравнения (III.5.2) сходится в H_0.

2°. Из сказанного в п.I° следует, что устойчивость метода Бубнова - Галеркина достаточно исследовать для уравнения (III.5. Как и в случае метода Ритца, мы рассмотрим два вычислительных процесса: для вычисления вектора коэффициентов Бубнова - Галер

кина и для вычисления приближенного решения по Бубнову - Галеркину. Устойчивость этих процессов исследована в работе Жоховой и Яковлева [1] для более общего метода, часто называемого методом Петрова - Галеркина; доказано, что для устойчивости названных процессов достаточно, чтобы координатная система была сильно минимальна в H_0. В статье Вайникко [7] доказано, что условие сильной минимальности и необходимо; там же упрощено доказательство достаточности. В данном параграфе мы изложим содержание статьи Вайникко [7], ограничившись при этом более простым методом Бубнова - Галеркина.

3°. Пусть для построения приближенного по Бубнову - Галеркину решения уравнения (Ш.5.2) использована координатная система $\{\varphi_n\} \in H_0$, подчиненная обычным условиям; в частности, при любом n элементы $\varphi_1, \varphi_2, \ldots, \varphi_n$ предполагаются линейно независимыми. Систему $\{\varphi_n\}$ можно ортонормировать в H_0; пусть $\{\omega_n\}$ - система, полученная ортогонализацией системы $\{\varphi_n\}$. Обе последовательности связаны рекуррентными соотношениями вида

$$\omega_k = \sum_{j=1}^{k} C_{kj}\varphi_j, \quad \varphi_k = \sum_{j=1}^{k} \gamma_{kj}\omega_k. \qquad (Ш.5.3)$$

Обозначим через C_n матрицу первых n соотношений (Ш.5.3).

Лемма Ш.5.1. Если последовательность $\{\varphi_n\}$ сильно минимальна в H_0, то нормы $\|C_n\|_{R_n \to R_n}$ ограничены; в противном случае они монотонно стремятся к бесконечности вместе с n.

Имеем $[\omega_k, \omega_j] = \delta_{kj}$; используя представление (Ш.5.3), получаем

$$C_n M_n C_n^* = I_n, \qquad (Ш.5.4)$$

где M_n имеет тот же смысл, что и в § 3, звездочка означает сопряженную матрицу, I_n - единичная матрица в R_n. Из (Ш.5.4) следует, что $M_n^{-1} = C_n^* C_n$ и, следовательно, $\|C_n\| = \sqrt{\|M_n^{-1}\|} = [\lambda_n^{(1)}]^{-1/2}$. Дальнейшее очевидно.

Теорема Ш.5.1. Для устойчивости процесса вычисления приближенного по Бубнову - Галеркину решения уравнения (Ш.5.2) в последовательности пар пространств $(H_0^{(n)}, R_n)$ необходимо и достаточно, чтобы координатная система была сильно минимальна в H_0.

Через $H_0^{(n)}$ обозначено подпространство H_0 с базисом (φ_i).

$\varphi_2, \ldots, \varphi_n$).

Обозначим соответственно неискаженный и искаженный вектор коэффициентов Бубнова – Галеркина через $a_*^{(n)}$ и $b_*^{(n)}$, неискаженное и искаженное приближенное решение – через $x_*^{(n)}$ и $z_*^{(n)}$. Далее, положим

$$f^{(n)}=((f,\varphi_1),(f,\varphi_2),\ldots,(f,\varphi_n)),\ \varphi^{(n)}=$$
$$=(\varphi_1,\varphi_2,\ldots,\varphi_n),\ \alpha^{(n)}=(C_n^*)^{-1}a_*^{(n)},\ \beta^{(n)}=(C_n^*)^{-1}b_*^{(n)}.$$

Наконец, через B_n обозначим матрицу элементов $[\varphi_k+T\varphi_k,\ \varphi_j]$; $j,k=\overline{1,2,\ldots,n}$. В этих обозначениях имеем

$$x_*^{(n)}=\sum_{k=1}^{n}a_k^{(n)}\varphi_k=a^{(n)}\cdot\varphi^{(n)}.$$

Далее, $a_*^{(n)}=C_n^*\alpha^{(n)}$, отсюда

$$x_*^{(n)}=\sum_{k=1}^{n}\sum_{j=k}^{n}c_{jk}\alpha_j^{(n)}\varphi_k=\sum_{j=1}^{n}\alpha_j^{(n)}\sum_{k=1}^{j}c_{jk}\varphi_k=\sum_{j=1}^{n}\alpha_j^{(n)}\omega_j.$$

Аналогично

$$z_*^{(n)}=\sum_{j=1}^{n}\beta_j^{(n)}\omega_j$$

и, следовательно,

$$z_*^{(n)}-x_*^{(n)}=\sum_{j=1}^{n}(\beta_j^{(n)}-\alpha_j^{(n)})\omega_j. \qquad (\text{Ш}.5.5)$$

Системы уравнений Бубнова – Галеркина, искаженная и неискаженная, имеют соответственно вид

$$(B_n+\Gamma_n)b^{(n)}=f^{(n)}+\delta^{(n)};\ B_n a^{(n)}=f^{(n)},$$

или

$$(C_nB_nC_n^*+C_n\Gamma_nC_n^*)\beta^{(n)}=C_n(f^{(n)}+\delta^{(n)}),$$

$$C_nB_nC_n^*\alpha^{(n)}=C_nf^{(n)}.$$

Отсюда

$$(C_nB_nC_n^*+C_n\Gamma_nC_n^*)(\beta^{(n)}-\alpha^{(n)})=C_n\delta^{(n)}-C_n\Gamma_nC_n^*\alpha^{(n)}. \qquad (\text{Ш}.5.6$$

Обозначим через Π_n ортопроектор из H_0 на $H_0^{(n)}$. Докажем, что существует такая постоянная $\tau > 0$, что

$$\forall n,\ \forall v^{(n)} \in (I+T)H_0^{(n)};\ |\Pi_n v^{(n)}| \geq \tau |v^{(n)}|. \qquad (\text{II.5.7})$$

Допустим противное. Тогда можно построить такую последовательность $v^{(n)} \in (I+T)H_0^{(n)}$, $n = 1, 2, \ldots$, что $|v^{(n)}| = 1$ и $|\Pi_n v^{(n)}| \xrightarrow[n\to\infty]{} 0$. Пусть $v^{(n)} = (I+T)w^{(n)}$, тогда $w^{(n)} \in H_0^{(n)}$. При этом $|w^{(n)}| \leq |I+T|$ и $|\Pi_n (I+T) w^{(n)}| = |w^{(n)} + \Pi_n T w^{(n)}| \xrightarrow[n\to\infty]{} 0$.

Последовательность $\{w^{(n)}\}$ ограничена и потому слабо компактна в H_0. Выделим из $\{w^{(n)}\}$ подпоследовательность, которую опять обозначим через $\{w^{(n)}\}$ и которая слабо сходится в H_0 к некоторому элементу $w^{(0)}$. Тогда $Tw^{(n)} \longrightarrow Tw^{(0)}$ в H_0. Далее

$$\Pi_n T w^{(n)} = \Pi_n T w^{(0)} + \Pi_n T (w^{(n)} - w^{(0)}).$$

Первое слагаемое справа стремится к $Tw^{(0)}$ в норме H_0, потому что последовательность подпространств $\{H_0^{(n)}\}$ полна в H_0; второе слагаемое стремится к нулю в той же норме. Таким образом, $\Pi_n T w^{(n)} \longrightarrow T w^{(0)}$. Теперь из соотношения $|w^{(n)} + \Pi_n T w^{(n)}| \to 0$ следует, что $(I+T)w^{(0)} = 0$. Отсюда вытекает, что

$$\lim v^{(n)} = \lim (I+T) w^{(n)} = (I+T) w^{(0)} = 0,$$

что противоречит условию $|v^{(n)}| = 1$.

Докажем теперь, что при достаточно больших n имеет место неравенство

$$\forall c^{(n)} \in R_n,\ \frac{1}{\mu} \|c^{(n)}\| \leq |C_n B_n C_n^* c^{(n)}| \leq$$
$$\leq |I+T| \cdot \|c^{(n)}\|,\quad \mu = \tau^{-1} |(I+T)^{-1}|. \qquad (\text{II.5.8})$$

Как легко видеть, $C_n B_n C_n^*$ есть матрица скалярных произведений $[\omega_k + T\omega_k, \omega_j]$; $j, k = 1, 2, \ldots, n$. Отсюда

$$\|C_n B_n C_n^* c^{(n)}\|^2 = \sum_{j=1}^{n} \Big| \sum_{k=1}^{n} (\omega_k + T\omega_k, \omega_j) c_k^{(n)} \Big|^2 =$$
$$= \sum_{j=1}^{n} \Big| \Big((I+T) \sum_{k=1}^{n} c_k^{(n)} \omega_k, \omega_j \Big) \Big|^2 = |\Pi_n (I+T) \sum_{k=1}^{n} c_k^{(n)} \omega_k|^2 \leq$$

$$\leq \|I+T\|^2 \cdot \|\sum_{k=1}^{n} c_k^{(n)} \omega_k\|^2 = \|I+T\|^2 \cdot \|c^{(n)}\| .$$

С другой стороны, неравенство (Ш.5.7) дает

$$\|\Pi_n (I+T) \sum_{k=1}^{n} c_k^{(n)} \omega_k\|^2 \geq \tau^2 \|(I+T) \sum_{k=1}^{n} c_k^{(n)} \omega_k\|^2 \geq \frac{1}{\mu^2} \|\sum_{k=1}^{n} c_k^{(n)} \omega_k\|^2 = \frac{1}{\mu^2} \|c^{(n)}\|^2 .$$

4°. Переходим к доказательству утверждений теоремы. Пусть координатная система $\{\varphi_n\}$ сильно минимальна в H_0. По лемме Ш.5.1 $\|C_n\| = \|C_n^*\| \leq c_0 = const$. Положим $\tau = (2\mu c_0^2)^{-1}$, и пусть $\|\Gamma_n\| \leq \tau$. Тогда $\|C_n \Gamma_n C_n^*\| \leq 1/2\mu$. В то же время, как показывает левое неравенство (Ш.5.8), $\|C_n B_n C_n^*\| \geq 1/\mu$, матрица в левой части уравнения (Ш.5.6) обратима. При этом $\|[C_n(B_n+\Gamma_n)C_n^*]^{-1}\| \leq 2\mu$; из соотношений (Ш.5.6) и из факта сходимости метода Бубнова - Галеркина следует

$$\|x_*^{(n)} - x_n^{(n)}\| = \|\beta^{(n)} - a^{(n)}\| \leq p\|\Gamma_n\| + q\|\delta^{(n)}\|, \qquad (Ш.5.9)$$

где p и q некоторые постоянные. Достаточность условия теоремы доказана.

Пусть теперь координатная система не сильно минимальна в H_0. По лемме Ш.5.1 $\|C_n\| \xrightarrow[n\to\infty]{} \infty$. Выберем $\delta^{(n)}$ так, чтобы $\|C_n \delta^{(n)}\| \geq \frac{1}{2}\|C_n\| \cdot \|\delta^{(n)}\|$. Правое неравенство (Ш.5.8) показывает, что при $\Gamma_n = 0$

$$\|\beta^{(n)} - a^{(n)}\| \geq \frac{\|C_n\|}{2\|I+T\|} \|\delta^{(n)}\|$$

и, так как коэффициент при $\|\delta^{(n)}\|$ бесконечно возрастает, то рассматриваемый процесс неустойчив. Этим доказана необходимость условий теоремы.

5°. Теорема Ш.5.2. Если координатная система сильно минимальна в H_0, то процесс вычисления коэффициентов Бубнова - Галеркина устойчив в последовательности пар пространств (R_n, R_n).

Доказательство очень просто: так как $b^{(n)} - a^{(n)} = C_n^*(\beta^{(n)} - a^{(n)})$, а $\|C_n^*\| \leq c_0$, то по неравенству (Ш.5.9)

$$\|b^{(n)} - a^{(n)}\| \leq c_0 p\|\Gamma_n\| + c_0 q\|\delta^{(n)}\| .$$

Замечание. В работах М.А.Белиева [I - 9] исследована устойчивость метода Бубнова - Галеркина (точнее, Фадо - Галеркина) для довольно широкого класса нестационарных операторных уравнений. Часть результатов Белиева изложена (и частично улучшена) в книге автора [8]; некоторые его результаты усилены в статье Тополянского и Запрудского [I]. Отметим еще, что в работах А.В. Маматова [I, 2] исследована устойчивость метода Фадо - Галеркина для задач, в которых квазилинейное параболическое уравнение решается при краевом условии (вообще говоря, нелинейном), содержащим производную по времени от искомой функции. Атакишиева [I] и Запрудский [I] изучали устойчивость метода Бубнова - Галеркина в банаховых пространствах.

§ 6. Погрешность искажения рекуррентного вычислительного процесса.

I°. Рассмотрим бесконечный рекуррентный вычислительный процесс

$$n=0,1,2,\ldots;\ \sum_{k=0}^{n} A_{nk}x^{(k)}=f^{(n)}. \qquad (\text{III}.6.1)$$

Примем, что $x^{(n)}\in X_n$, $f^{(n)}\in Y_n$, где X_n и Y_n – банаховы пространства, и что A_{nk} – линейные операторы, действующие из X_k в Y_n. Будем считать, что каждый из операторов A_{nn} ограниченно обратим, а произведения $A_{nn}^{-1}A_{nk}$, $0\leq k\leq n-1$, ограничены. При этих предположениях существует последовательность элементов $x^{(n)}\in X_n$, удовлетворяющих бесконечной системе (III.6.1).

Пусть каждый из операторов A_{nk} вычислен с погрешностью Γ_{nk}, а каждый из элементов $f^{(n)}$ – с погрешностью $\delta^{(n)}$, так что вместо системы (III.6.1) на самом деле решается система

$$n\geq 0;\ \sum_{k=0}^{n}(A_{nk}+\Gamma_{nk})\tilde{x}^{(k)}=f^{(n)}+\delta^{(n)}. \qquad (\text{III}.6.2)$$

Допустим, что погрешности Γ_{nk} можно сделать сколь угодно малыми в следующем смысле: каково бы ни было число $\varepsilon>0$, можно добиться того, чтобы

$$\forall n\geq 0,\ \sum_{k=0}^{n}\|A_{nn}^{-1}\Gamma_{nk}\|\leq\varepsilon. \qquad (\text{III}.6.3)$$

2°. Оценим погрешность искажения $\xi_n=\|\tilde{x}^{(n)}-x^{(n)}\|$. Уравнения (III.6.1) и (III.6.2) преобразуем к виду

$$A_{nn}x^{(n)} = g^{(n)} := f^{(n)} - \sum_{k=0}^{n-1} A_{nk}x^{(n)}; \qquad (\text{III.6.4})$$

$$(A_{nn} + \Gamma_{nn})z^{(n)} = g^{(n)} + \varpi^{(n)};$$

$$\varpi^{(n)} := \delta^{(n)} - \sum_{k=0}^{n-1} A_{nk}(z^{(k)} - x^{(k)}) - \sum_{k=0}^{n-1} \Gamma_{nk}z^{(k)}. \qquad (\text{III.6.5})$$

Переход к этим новым уравнениям преобразует рекуррентный процесс в эквивалентный свободный, и для оценки погрешности искажения можно воспользоваться неравенством (III.I.7), которое в данном случае означает следующее: если $\|A_{nn}^{-1}\Gamma_{nn}\| \le \beta$, $0 < \beta < 1$, то

$$\hat{\xi}_n \le \frac{1}{1-\beta}\left[\|A_{nn}^{-1}\Gamma_{nn}x_*^{(n)}\| + \|A_{nn}^{-1}\varpi^{(n)}\|\right] \le$$

$$\le \frac{1}{1-\beta}\Big\{\|A_{nn}^{-1}\Gamma_{nn}\|\cdot\|x_*^{(n)}\| + \sum_{k=0}^{n-1}\|A_{nn}^{-1}\Gamma_{nk}\|\cdot\|x_*^{(k)}\| +$$

$$+ \sum_{k=0}^{n-1}\left[\|A_{nn}^{-1}A_{nk}\| + \|A_{nn}^{-1}\Gamma_{nk}\|\right]\hat{\xi}_k + \|A_{nn}^{-1}\|\cdot\|\delta^{(n)}\|\Big\} = \qquad (\text{III.6.6})$$

$$= \frac{1}{1-\beta}\Big\{\sum_{k=0}^{n}\|A_{nn}^{-1}\Gamma_{nk}\|\cdot\|x_*^{(k)}\| + \|A_{nn}^{-1}\|\cdot\|\delta^{(n)}\| + \sum_{k=0}^{n-1}\left[\|A_{nn}^{-1}A_{nk}\| + \|A_{nn}^{-1}\Gamma_{nk}\|\right]\hat{\xi}_k$$

Формула (III.6.6) позволяет последовательно вычислить оценки для $\hat{\xi}_0$, $\hat{\xi}_1$, $\hat{\xi}_2$, ..., если известны оценки для $\|x_*^{(n)}\|$, $n = 0, 1, 2, \ldots$

§ 7. Устойчивость рекуррентного вычислительного процесса

I°. Процесс (III.6.I) назовем устойчивым относительно искажения в последовательности пар пространств (X_n, Y_n), если существуют такие постоянные ρ_1, ρ_2, $\rho_3 > 0$, что из неравенства

$$\forall n \ge 0,\ \sum_{k=0}^{n}\|A_{nn}^{-1}\Gamma_{nk}\| \le \rho_3 \qquad (\text{III.7.I})$$

вытекает следующая оценка погрешности искажения

$$\hat{\xi}_n \le \rho_1 \max_{j\le n} \sum_{k=0}^{j} \|A_{jj}^{-1}\Gamma_{jk}\| + \rho_2 \max_{j\le n} \|A_{jj}^{-1}\delta^{(j)}\|. \qquad (\text{III}.7.2)$$

Это определение соответствует первому из определений устойчивости свободного вычислительного процесса.

Замечание. В предшествующем параграфе было отмечено, что любой рекуррентный вычислительный процесс легко преобразуется в свободный. Можно было бы определить устойчивость рекуррентного процесса как устойчивость соответствующего свободного процесса, и тогда были бы применимы утверждения § 2 настоящей главы. Однако, условия этих утверждений сказываются в этом случае громоздкими и трудно проверяемыми, почему мы и предпочитаем дать для рекуррентного процесса независимое определение устойчивости.

Теорема III.7.I. Пусть $h^{(n)} \in Y_n$, $n = 0, I, 2, \ldots$, и пусть $t^{(n)}$, $n = 0, I, 2, \ldots$, есть последовательность решений бесконечной рекуррентной системы

$$\forall n \ge 0, \quad \sum_{k=0}^{n} A_{nk} t^{(k)} = h^{(n)}. \qquad (\text{III}.7.3)$$

Если выполнены неравенства

$$\|t^{(n)}\| \le C_0 \max_{j\le n} \|A_{jj}^{-1} h^{(j)}\|, \quad C_0 = const \qquad (\text{III}.7.4)$$

и

$$\|x_*^{(n)}\| \le C_1 = const, \qquad (\text{III}.7.5)$$

где $x_*^{(n)}$, $n = 0, I, 2, \ldots$ - последовательность решений системы (III.6.I), то вычислительный процесс (III.6.I) устойчив.

Уравнение (I 3.2) преобразуем к виду

$$\sum_{k=0}^{n} A_{nk}(\tilde{x}^{(k)} - x^{(k)}) = \delta^{(n)} - \sum_{k=0}^{n} \Gamma_{nk}(\tilde{x}^{(k)} - x^{(k)}) - \sum_{k=0}^{n} \Gamma_{nk} x^{(k)}.$$

По неравенствам (III.7 4) и (III.7.5)

$$\hat{\xi}_n \le C_0 \Big\{ \max_{j\le n} \|A_{jj}^{-1}\delta^{(j)}\| + \max_{j\le n} \sum_{k=0}^{j} \|A_{jj}^{-1}\Gamma_{jk}\| \times \qquad (\text{III}.7.6)$$

$$\times \max_{j\le n} \hat{\xi}_j + C_1 \max_{j\le n} \sum_{k=0}^{j} \|A_{jj}^{-1}\Gamma_{jk}\| \Big\}.$$

Обозначим

$$\sigma_n = \max_{j\le n} \|A_{jj}^{-1}\delta^{(j)}\| + C_1 \max_{j\le n} \sum_{k=0}^{j} \|A_{jj}^{-1}\Gamma_{jk}\| . \qquad (\text{Ш}.7.7)$$

Очевидно, σ_n не убывает с возрастанием n. Возьмем ρ_3 столь малым, чтобы $\beta = C_0\rho_3 < 1$. Тогда

$$\hat{\xi}_n \le \beta \max_{j\le n} \hat{\xi}_j + \sigma_n . \qquad (\text{Ш}.7.8)$$

Докажем, что

$$\forall n \ge 0 , \quad \hat{\xi}_n \le \frac{\sigma_n}{1-\beta} . \qquad (\text{Ш}.7.9)$$

Неравенство (Ш.7.9) верно при $n = 0$ – в этом случае (Ш.7.8) принимает вид $\hat{\xi}_0 \le \beta\hat{\xi}_0 + \sigma_0$. Пусть неравенство (Ш.7.9) верно для $n \le N-1$; докажем, что оно тогда верно и для $n = N$. Может случиться, что

$$\max_{j\le N} \hat{\xi}_j = \hat{\xi}_N , \qquad (\text{Ш}.7.10)$$

тогда по (Ш.7.8), $\hat{\xi}_N \le \beta\hat{\xi}_N + \sigma_N$, и неравенство (Ш.7.9) доказано. Если же равенство (Ш.7.10) неверно, то $\max_{j\le N} \hat{\xi}_j = \hat{\xi}_K$, где K – одно из чисел $0, 1, \dots, N-1$, и тогда

$$\hat{\xi}_N < \hat{\xi}_K \le \frac{\sigma_K}{1-\beta} \le \frac{\sigma_N}{1-\beta} ,$$

что и требовалось доказать. Ясно, что теперь неравенство (Ш.7.2) справедливо при значениях постоянных

$$\rho_1 = \frac{C_1}{1-\beta} , \quad \rho_2 = \frac{1}{1-\beta} , \quad \rho_3 = \frac{\beta}{C_0} ; \quad \forall \beta \in (0,1). \qquad (\text{Ш}.7.11)$$

2^0. Нетрудно указать простое условие, при котором неравенство (Ш.7.4) справедливо: достаточно, чтобы

$$\sum_{k=0}^{n-1} \|A_{nn}^{-1}A_{nk}\| \le q = const < 1 . \qquad (\text{Ш}.7.12)$$

Действительно, из уравнения (Ш.7.3) следует, что

$$\|t^{(n)}\| \le q \max_{k\le n-1} \|t^{(k)}\| + \|A_{nn}^{-1}h^{(n)}\| . \qquad (\text{Ш}.7.13)$$

Пусть $\max_{k\le n} \|t^{(k)}\| = \|t^{(l)}\|$, $l \le n$. Если $l = n$, то из (Ш.7.13 находим $\|t^{(n)}\| \le q\|t^{(n)}\| + \|A_{nn}^{-1}h^{(n)}\|$, или $\|t^{(n)}\| \le (1-q)^{-1}\|A_{nn}^{-1}h^{(n)}\|$. Если же $l < n$, то, заменив в (Ш.7.13 n на l, получаем $\|t^{(l)}\| \le q\|t^{(l)}\| + \|A_{ll}^{-1}h^{(l)}\|$ и, следова-

тельно,

$$\|t^{(n)}\| < \|t^{(l)}\| \leq (1-q)^{-1} \|A_{ll}^{-1} h^{(l)}\| \leq (1-q)^{-1} \max_{j \neq n} \|A_{jj}^{-1} h^{(j)}\|.$$

3°. Теорема Ш.7.2. Рекуррентный процесс (Ш.6.I) устойчив, если выполнены условия (Ш.7.5) и (Ш.7.I2).

Доказательство следует из неравенства (Ш.6.6). Допуская, что

$$\sum_{k=0}^{n} \|A_{nn}^{-1} \Gamma_{nk}\| \leq \beta, \quad 0 < \beta < 1,$$

получаем из (Ш.6.6)

$$\hat{\xi}_n \leq \frac{1}{1-\beta} \left[C_1 \sum_{k=0}^{n} \|A_{nn}^{-1} \Gamma_{nk}\| + \|A_{nn}^{-1} \delta^{(n)}\| + (q+\beta) \max_{k \leq n} \hat{\xi}_k \right].$$

Выберем β столь малым, чтобы $q_1 := (q+\beta)(1-\beta)^{-1} < 1$. Имеем

$$\hat{\xi}_n \leq q_1 \max_{k<n} \hat{\xi}_k + \frac{1}{1-\beta} \left[C_1 \sum_{k=0}^{n} \|A_{nn}^{-1} \Gamma_{nn}\| + \|A_{nn}^{-1} \delta^{(n)}\| \right]. \quad (Ш.7.I4)$$

Уравнения (Ш.7.I4) и (Ш.7.8) с точностью до обозначений совпадают, и аналогично формуле (Ш.7.9) получаем

$$\hat{\xi}_n \leq \rho_1' \sum_{k=0}^{n} \|A_{nn}^{-1} \Gamma_{nk}\| + \rho_2' \|A_{nn}^{-1} \delta^{(n)}\|. \quad (Ш.7.I5)$$

Замечание 2. Из замечания I (см. п.I°) ясно, что теорема Ш.7.2 является частным случаем теоремы Ш.7.I. Мы считаем целесообразным выделить этот частный случай, потому что оценка (Ш.7.I5) несколько лучше по порядку, чем это требуется формулой (Ш.7.2), входящей в определение устойчивости.

§ 8. Погрешность искажения и устойчивость метода наискорейшего спуска.

I°. Результаты §§ 6,7 применим к методу наискорейшего спуска, предложенному в свое время Дж. Темплем [I] и, независимо от него, Канторовичем [2](см. также Канторович и Акилов [1]), который установил также оценку скорости сходимости метода (см. ниже, формула (Ш.8.3)); в статье Темпля установлен только самый факт сходимости метода. Коротко напомним основные черты метода наискорейшего спуска.

Пусть A - ограниченный самосопряженный положительно определенный оператор, действующий в некотором гильбертовом простран-

стве H, и пусть требуется решить уравнение

$$Ax = f. \qquad (\text{III}.8.1)$$

Зададим произвольный элемент $x^{(0)} \in H$, который будем рассматривать как нулевое приближение к искомому решению x_* уравнения (III.8.1). Пусть построено $(n-1)$-е приближение $x^{(n-1)}$, тогда следующее приближение $x^{(n)}$ строится по формулам

$$y^{(n-1)} = Ax^{(n-1)} - f, \quad \sigma_{n-1} = \frac{\|y^{(n-1)}\|^2}{(Ay^{(n-1)}, y^{(n-1)})}, \quad x^{(n)} = x^{(n-1)} - \sigma_{n-1} y^{(n-1)}. \qquad (\text{III}.8.2)$$

Как показано в статье Канторовича [2], $x^{(n)} \underset{n\to\infty}{\Longrightarrow} x_*$ с оценкой

$$\|x_* - x^{(n)}\| \leqslant \frac{\|y_1\|}{m} \left(\frac{M-m}{M+m}\right)^n, \qquad (\text{III}.8.3)$$

где M и m суть соответственно верхняя и нижняя грани оператора A:

$$M = \sup_{\|f\|=1} (Af, f), \quad m = \inf_{\|f\|=1} (Af, f).$$

Таким образом, метод наискорейшего спуска связан с бесконечным рекуррентным процессом

$$x^{(n)} - B_n x^{(n-1)} = f^{(n)}, \qquad (\text{III}.8.4)$$

где

$$B_n = I - \sigma_{n-1} A, \quad f^{(n)} = \sigma_{n-1} f. \qquad (\text{III}.8.5)$$

2°. Исследуем погрешность искажения и устойчивость процесса (III.8.4). Очевидно, $M^{-1} \leqslant \sigma_{n-1} \leqslant m^{-1}$; верхняя и нижняя грани оператора B_n суть, следовательно, $1 - \sigma_{n-1} m$ и $1 - \sigma_{n-1} M$. Далее,

$$1 - \sigma_{n-1} M \geqslant -\frac{M-m}{m}, \quad 1 - \sigma_{n-1} m \leqslant \frac{M-m}{M}.$$

Отсюда

$$\|B_n\| \leqslant \max\left(\frac{M-m}{m}, \frac{M-m}{M}\right) = \frac{M-m}{m}. \qquad (\text{III}.8.6)$$

В обозначениях § 6 $A_{nn} = I$, $A_{n,n-1} = -B_n$, $A_{nk} = 0$, $k \leqslant n-2$. Обозначим через ϑ_n погрешность вычисления величины σ_{n-1}, тогда $\Gamma_{n,n-1} = \vartheta_n A$ и $\Gamma_{nk} = 0$ при $k \neq n-1$; кроме того $\delta^{(n)} = \vartheta_n f$.

Обратимся к формуле (III.6.6). Условие $\|A_{nn}\Gamma_{nn}\| \leqslant \beta$, $0 < \beta < 1$,

здесь выполняется тривиально, потому что $\Gamma_{nn} = 0$. Можно считать $\beta = 0$, и формула (Ш.6.5) принимает вид

$$\hat{\xi}_n \leqslant \left(\frac{M-m}{m} + |\vartheta_n| M\right)\hat{\xi}_{n-1} + |\vartheta_n|(c_0 M + \|f\|). \qquad (Ш.8.7)$$

Здесь c_0 – какая-нибудь верхняя граница значений $\|x_*^{(n)}\|$; такая конечная граница существует, потому что процесс наискорейшего спуска сходится. Эту границу нетрудно вычислить, исходя из неравенства (Ш.8.3).

Проанализируем формулу (Ш.8.7). Пусть погрешность ϑ_n достаточно мала, $|\vartheta_n| \leqslant \delta$. Если при этом $M < 2m$, то

$$\frac{M-m}{m} + |\vartheta_n| M \leqslant \frac{M-m}{m} + \varepsilon M := q$$

и при ε достаточно малом будет $q < 1$. Теперь

$$\hat{\xi}_n \leqslant q\hat{\xi}_{n-1} + \varepsilon a, \quad a = c_0 M + \|f\|. \qquad (Ш.8.8)$$

Заметим еще, что

$$\hat{\xi}_0 = 0 \qquad (Ш.8.9)$$

потому что элемент $x^{(0)}$ не вычисляется, а задается, и естественно считать, что $z^{(0)} = x^{(0)}$. Очевидно, что $\xi_n \leqslant \tau_n$, где τ_n – решение разностного уравнения

$$\tau_n = q\tau_{n-1} + \varepsilon a, \ \tau_0 = 0. \qquad (Ш.8.10)$$

Решение этого уравнения есть

$$\tau_n = \frac{\varepsilon a}{1-q}(1-q^n),$$

и оценка погрешности искажения имеет вид

$$\hat{\xi}_n \leqslant \frac{\varepsilon a}{1-q}(1-q^n) < \frac{\varepsilon a}{1-q}. \qquad (Ш.8.11)$$

Таким образом, если: 1) $M < 2m$ и 2) ε – верхняя граница погрешности величины δ_{n-1} – достаточно мала, то погрешность искажения метода наискорейшего спуска есть величина порядка $O(\varepsilon)$, и эта оценка – равномерная относительно n .

Если хотя бы одно из условий 1) , 2) нарушено, то может случиться, что погрешность искажения будет неограниченно возрастать вместе с n .

3°. Близкие результаты дает и анализ устойчивости. Условие (Ш. .7) очевидно выполнено, а условие (Ш.7.12) в данном случае означает, что $\|B_n\| \leqslant q < 1$; это условие будет выполнено, если $M < 2m$ – тогда можно положить $q = (M-m)/m$. По теореме

(Ш.7.2) процесс наискорейшего спуска устойчив, если $M < 2m$. Тот же результат дает и теорема Ш.7.1.

4^0. Если $M \geqslant 2m$, то для оценки погрешности можно воспользоваться непосредственно формулой (Ш.8.7). Допустим, что величина ϑ_n ограничена, тогда $(M-m)/m + \vartheta_n M \leqslant \tau = const$ и

$$\hat{\xi}_n \leqslant \tau \hat{\xi}_{n-1} + |\vartheta_n| a, \quad \hat{\xi}_0 = 0. \tag{Ш.8.12}$$

Решение этого неравенства есть

$$\hat{\xi}_n \leqslant a \sum_{k=1}^{n} |\vartheta_k| \tau^{n-k} \tag{Ш.8.13}$$

Зафиксируем натуральное число N. Если желательно, чтобы было $\hat{\xi}_N \leqslant \xi$, где ξ малое число, то достаточно выполнять σ_{k-1}, $1 \leqslant k \leqslant N$, с такой точностью, чтобы

$$|\vartheta_k| \leqslant \frac{\xi}{N a \tau^{N-k}}. \tag{Ш.8.14}$$

§ 9. Об устойчивости метода коллокации.

1^0. Пусть немасштабированная алгебраическая система метода коллокации имеет вид

$$M_n a^{(n)} = f^{(n)}, \tag{Ш.9.1}$$

где M_n – матрица чисел $\psi_{kn}(t_j^{(n)}) = (A\varphi_{kn})(t_j^{(n)})$, $a^{(n)} = (a_1^{(n)}, a_2^{(n)}, \ldots, a_N^{(n)})$, $f^{(n)} = (f(t_1^{(n)}), f(t_2^{(n)}), \ldots, f(t_N^{(n)}))$.

Пусть $\mathcal{K}$ – компакт в m-мерном евклидовом пространстве R_m. Точки $t_j^{(n)}$ – узлы коллокации – выберем в вершинах некоторой параллелепипедальной сетки, лежащих в $\mathcal{K}$. Ребра параллелепипедов сетки, параллельные k-й координатной оси, пусть имеют длину $h_k^{(n)}$, $k = 1, 2, \ldots, m$; допустим, что $c_1 h_n \leqslant h_k^{(n)} \leqslant c_2 h_n$, где c_1 и c_2 – постоянные и $h_n \xrightarrow[n\to\infty]{} 0$.

Будем рассматривать $a^{(n)}$ как элемент пространства R_N, а $f^{(n)}$ – как элемент пространства Y_N, также N-мерного, в котором норма определяется так:

$$\|f^{(n)}\|_{Y_N} = h_n^{1/2} \|f^{(n)}\|_{l_N}, \tag{Ш.9.2}$$

l_N – вещественная постоянная. Оператор, порожденный матрицей M_n и действующий из R_N в Y_N, обозначим через A_n; систему (Ш.9.1) можно теперь записать в такой операторной форме:

$$A_n a^{(n)} = f^{(n)}; \qquad (\text{III.9.3})$$

соответствующую искаженную систему запишем в обычной форме

$$(A_n + \Gamma_n) b^{(n)} = f^{(n)} + \delta^{(n)}. \qquad (\text{III.9.4})$$

Оператор, порожденный матрицей M_n и действующий в R_N, обозначим тем же символом M_n.

2°. Пусть $s_n^{(k)}$, $k = 1, 2, \ldots, N$ – сингулярные числа матрицы M_n, т.е., собственные числа неотрицательной матрицы $M_n^* M_n$. Расположим их в порядке возрастания: $s_n^{(1)} \le s_n^{(2)} \le \ldots \le s_n^{(N)}$.

Теорема III.9.1. Если

$$s_n^{(1)} \ge C_0 h_n^{-\delta}, \quad C_0 = const > 0 \qquad (\text{III.9.5})$$

и $\|a^{(n)}\| \le C_1 = const$, то вычислительный процесс метода коллокации (III.9.3) устойчив в смысле второго определения (§ 2, гл. III) в последовательности пар пространств (R_N, Y_{N3}). Если же

$$s_n^{(1)} : r(h_n) h_n^{-\delta}, \quad r(h_n)_{h_n \to 0} \to 0, \qquad (\text{III.9.6})$$

то в той же последовательности процесс (III.9.3) неустойчив.

По следствию III.2.2, для устойчивости процесса (III.9.3) в упомянутом смысле необходимо и достаточно, чтобы: 1) $\|A_n^{-1}\| \le C_1$, 2) $\|A_n^{-1}\| \cdot \|a^{(n)}\|_{R_N} \le C_2$, где C_1 и C_2 – постоянные. Оценим величину

$$\|A_n^{-1}\| = \sup_{g^{(n)} \in Y_{N3}} \frac{\|A_n^{-1} g^{(n)}\|_{R_N}}{\|g^{(n)}\|_{Y_{N3}}} =$$

$$= h_n^{-3/2} \sup_{g^{(n)} \in R_N} \frac{\|M_n^{-1} g^{(n)}\|_{R_N}}{\|g^{(n)}\|_{R_N}} = h_n^{-3/2} \|M_n^{-1}\|.$$

Из соотношения $M_n^{-1} (M_n^*)^{-1} = (M_n^* M_n)^{-1}$ следует, что сингулярные числа матрицы M_n^{-1} суть $1/s_n^{(N)}, 1/s_n^{(N-1)}, \ldots, 1/s_n^{(1)}$. Отсюда $\|M_n^{-1}\| = \|(M_n^* M_n)^{-1}\|^{1/2} = 1/\sqrt{s_n^{(1)}}$ и окончательно

$$\|A_n^{-1}\| = \frac{1}{\sqrt{h_n^3 s_n^{(1)}}}. \qquad (\text{III.9.7})$$

Если верно неравенство (III.9.6), то $\|A_n^{-1}\| \ge [r(h_n)]^{-1/2} h_n \xrightarrow{} \infty$, и процесс (III.9.3) неустойчив.

Пусть оправедливо неравенство (Ш.9.5). Тогда $\|A_n^{-1}\| \leqslant C_0^{-1/2}$, и условие I) выполнено. Условие 2) выполнено по предположению, и процесс коллокации устойчив.

Отметим частный случай. Прежде всего заметим следующее: так как $f \in C(\mathcal{K})$, то тем более $f \in L_2(\mathcal{K})$, а тогда

$$h_n^m \sum_{j=1}^{N} f_j^2 \leqslant C = const$$

или, что то же, $\|f^{(n)}\|_{y_{Nm}} \leqslant C$. Пусть теперь неравенство (Ш.9. выполнено при $\delta = m$, тогда норма $\|A_n\|$ ограничена постоянной $C^{-1/2}$. Как только что было отмечено, нормы $\|f^{(n)}\|_{y_{Nm}}$ также ограничены, а тогда ограничены и нормы $\|a^{(n)}\|_{R_N} \leqslant \|A_n^{-1}\| \cdot \|f^{(n)}\|_{y_{Nm}}$.

По следствию Ш.2.2, процесс (Ш.9.3) устойчив. Таким образом, при $\delta = m$ условие (Ш.9.5) необходимо и достаточно для устойчивости процесса коллокации (Ш.9.3).

3°. Выше (см. § 3, гл.Ш) было дано определение сильной минимальности последовательности систем

$$(y_{n1}, y_{n2}, \ldots, y_{nN}); \quad N = N(n), \; n = 1, 2, \ldots . \qquad (Ш.9.8)$$

Введем еще одно определение: систему последовательностей (Ш.9. назовем почти ортонормированной, если с...ственные числа матриц ...рама этих систем положительно ограничены оверху и снизу. Очевидно, почти ортонормированная последовательность также и сильно минимальна. Это определение обобщает известное определение Талдыкина [I]. Системы, которые мы называем почти ортонормированными, Талдыкин называет регулярными.

Теорема Ш.9.2. Пусть X - сепарабельное гильбертово пространство, и последовательность (Ш.9.8) почти ортонормирована в X, а процесс (Ш.9.3) устойчив в смысле второго определения (§ 2, гл.Ш). Тогда процесс вычисления приближенного решения по методу коллокации устойчив в смысле того же определения в последовательности пар пространств (X_N, y_{N3}). Если последовательность (Ш.9.8) сильно минимальна в X и процесс (Ш.9.3) неустойчив, то процесс вычисления приближенного решения также неустойчив.

Если последовательность (Ш.9.8) почти ортонормирована в X то

$$\|x_*^{(n)} - z_*^{(n)}\|_X^2 = (M_n(a^{(n)} - b^{(n)}), a^{(n)} - b^{(n)}) \leqslant \Lambda_0 \|a^{(n)} - b^{(n)}\|_{R_N}^2 ;$$

здесь M_n - матрица Грама элементов (Ш.9.8) при фиксированном n, Λ_0 - верхняя граница собств ных чисел этих матриц при всех n, $x_*^{(n)}$ и $z_*^{(n)}$ - соответственно неискаженное и искаженное приближенные решения, с которым векторы $a^{(n)}$ и $b^{(n)}$ связаны соотношениями $x_*^{(n)} = \sum_{k=1}^{n} a_k^{(n)} y_{nk}$, $z_*^{(n)} = \sum_{k=1}^{n} b_k^{(n)} y_{nk}$. Так как процесс вычисления коэффициентов устойчив, то существуют такие постоянные $\rho, q, \tau > 0$, что при $\|\Gamma_n\| \le \tau$ будет

$$\|x_*^{(n)} - z_*^{(n)}\| \le \sqrt{\Lambda_0}\,(\rho\|\Gamma_n\|_{R_K \to Y_{K3}} + q\|\delta^{(n)}\|_{Y_{K3}});$$

последнее неравенство означает, что процесс вычисления приближенного решения устойчив.

Если последовательность (Ш.9.8) сильно минимальна в X, то, обозначая положительную нижнюю границу собственных чисел матрицы M_n через λ_0, имеем

$$\|x_*^{(n)} - z_*^{(n)}\| \ge \sqrt{\lambda_0}\,\|a^{(n)} - b^{(n)}\|_{R_K}.$$

По предположению, процесс (Ш.9.8) неустойчив, поэтому при любых, фиксированных независимо от n, нормах $\|\Gamma_n\|$ и $\|\delta^{(n)}\|$ нормы $\|a^{(n)} - b^{(n)}\|$ неограничены, а тогда неограничены и нормы $\|a_*^{(n)} - z_*^{(n)}\|$, и процесс вычисления приближенных решений неустойчив.

4°. Некоторые оценки наименьшего сингулярного числа $\delta_n^{(1)}$ можно получить, исследуя определитель $\mathcal{D}_n = \det M_n$. Ограничимся случаем $m = 1$, $K = [0, 1]$, и пусть узлы $t_j^{(n)} = j/n$, $j = 1, 2, \ldots, n$. Допустим, что функции $\psi_{nk}(t) = (Ay_{nk})(t)$ удовлетворяют условию Липшица с фиксированным показателем γ и с постоянной, которая ограничивается некоторой степенью K:

$$|\psi_{nk}(t') - \psi_{nk}(t'')| \le C K^{\beta} |t' - t''|^{\gamma}; \quad \beta = \text{const}; \qquad (\text{Ш.9.9})$$

предположим еще, что

$$|\psi_{nk}(t)| \le C K^{-\alpha}, \quad \alpha = \text{const} \ge 0 \qquad (\text{Ш.9.10})$$

Имеем (для упрощения обозначений пишем ψ_k вместо ψ_{nk} и t_j вместо $t_j^{(n)}$)

$$\mathcal{D}_n = \begin{vmatrix} \psi_1(t_1) & \psi_2(t_1) \dots \psi_n(t_1) \\ \psi_1(t_2) & \psi_2(t_2) \dots \psi_n(t_2) \\ \dots\dots\dots\dots \\ \psi_1(t_n) & \psi_2(t_n) \dots \psi_n(t_n) \end{vmatrix}. \qquad \text{(Ш.9.1)}$$

Из первой строки определителя $\mathcal{D}_n$ вычтем вторую, из второй – третью и т.д.; из $(n-1)$-й строки вычтем n-ю. Элементы первых $n-1$ строк имеют вид $\psi_k(t_j) - \psi_k(t_{j+1})$ и оцениваются, по равенотну (Ш.9.9), следующим образом:

$$|\psi_k(t_j) - \psi_k(t_{j+1})| \leqslant C n^{-\gamma} k^{\beta};$$

элементы г следней строки им ют оценку Ck^{α}. По теореме Адама (см., например, Смирнов [1], стр. 64) получаем

$$|\mathcal{D}_n| \leqslant \frac{C^n}{n^{\gamma(n-1)}} \Big[\sum_{k=1}^{n} k^{2\alpha}\Big]^{1/2} \Big[\sum_{k=1}^{n} k^{2\beta}\Big]^{\frac{n-1}{2}}.$$

Так как $\alpha \geqslant 0$, то $k^{2\alpha} \leqslant \int_k^{k+1} u^{2\alpha} du$ и

$$\sum_{k=1}^{n} k^{2\alpha} < \int_1^{n+1} u^{2\alpha} du < (2\alpha+1)^{-1} (n+1)^{2\alpha+1};$$

аналогично

$$\sum_{k=1}^{n} k^{2\beta} < (2\beta+1)^{-1} (n+1)^{2\beta+1}.$$

Теперь

$$|\mathcal{D}_n| \leqslant \frac{C_1^n}{n^{\gamma(n-1)}} (n+1)^{\frac{2\alpha+1}{2} + \frac{(2\beta+1)(n+1)}{2}}; \quad C_1 = const,$$

откуда легко получается и несколько более простая оценка

$$|\mathcal{D}_n| \leqslant C_2^n n^{(\beta+\frac{1}{2}-\gamma)n}; \quad C_2 = const. \qquad \text{(Ш.9.}$$

Далее, $\mathcal{D}_n^2 = Det(M_n^* M_n) = \nu_n^{(1)} \nu_n^{(2)} \dots \nu_n^{(n)} \geqslant [\nu_n^{(n)}]^n$;

откуда

$$\delta_n \leqslant \mathcal{D}_n \leqslant C_3^n n^{(2\beta+2\gamma)(1-\frac{1}{n})\frac{2\alpha+1}{n}} \leqslant C_3^n n^{\gamma(\beta-\gamma)+1}; \quad C_3 = const. \qquad \text{(Ш.9.}$$

Теперь из теоремы (III.9.I) вытекает, что если $m=\delta=1$ и $\beta>\gamma$, то процесс (III.9.3) неустойчив. В частности, если $\gamma=1$, то процесс (III.9.3) неустойчив при $\beta>1$.

5°. Теорема Адамара часто дает грубую оценку определителя. Рассмотрим пример, в котором эту оценку можно улучшить.

Пусть A – тождественный оператор. Задача состоит тогда в интерполяции функции $f(t)$ линейной комбинацией вида

$$\sum_{k=1}^{n} a_k^{(n)} \varphi_{nk}(t)$$

на заданном компакте $\mathcal{K}$. Пусть $m=1$, $\mathcal{K}=[0,1]$ и $\varphi_{nk}(t)=\varphi_k(t)=t^{k-1}$, $k=1,2,\dots$. Положим еще $t_j=t_j^{(n)}=j/n$, $j=0,1,2,\dots,n-1$. Определитель коллокационной системы

$$\mathcal{D}_n=\begin{vmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & t_1 & t_2 & \dots & t_{n-1} \\ 1 & t_1^2 & t_2^2 & \dots & t_{n-1}^2 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & t_1^{n-1} & t_2^{n-1} & \dots & t_{n-1}^{n-1} \end{vmatrix} = t_1 t_2 \dots t_{n-1} \begin{vmatrix} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_{n-1} \\ \dots & \dots & \dots & \dots \\ t_1^{n-2} & t_2^{n-2} & \dots & t_{n-1}^{n-2} \end{vmatrix}$$

равен определителю Вандермонда чисел $t_1, t_2, \dots t_{n-1}$, умноженному на $t_1 t_2 \dots t_{n-1} = \frac{(n-1)!}{n^{n-1}}$. Таким образом,

$$\mathcal{D}_n = \frac{(n-1)!}{n^{n-1}} \prod_{1\le k<j\le n} (t_j - t_k) < 1.$$

В таком случае $\delta_n^{(1)} < 1 = h \cdot h^{-1}$; $h = 1/n$, $r(h) = h \to 0$

и коллокационный процесс неустойчив в последовательности пар пространств $(R_n, \mathcal{Y}_{n2})$.

6°. В заключение параграфа коротко изложим содержание работы Вайнико [8], в которой, по-видимому, впервые рассмотрен вопрос об устойчивости метода коллокации. В этой работе рассматривается обыкновенное дифференциальное уравнение

$$x^{(m)}(t) - \sum_{k=1}^{m-1} c_k(t) x^{(k)}(t) = f(t) \qquad \text{(III.9.14)}$$

при краевых условиях

$$\sum_{k=0}^{m-1} [\alpha_{jk} x^{(k)}(a) + \beta_{jk} x^{(k)}(b)] = 0; \qquad \text{(III.9.15)}$$

$$j = 0, 1, \dots, m-1; \quad \alpha_{jk}, \beta_{jk} = \text{const}.$$

Предположим, что соответствующая однородная задача имеет только тривиальное решение. Тогда существует последовательность полиномов степени $m+k$

$$p_k(t)=\sum_{l=0}^{m+k} c_{kl}t^l \; ; \; k=0,1,2,\dots,$$

удовлетворяющая краевым условиям (II.9.15). Эту последовательность примем за координатную систему метода коллокации. За узлы коллокации при данном n примем корни $t_j^{(n)}$ полинома $\omega_n(t)$, где $\{\omega_n(t)\}$ - система полиномов, ортонормированная и полная в метрике $L_2(a,b;\rho(t))$, причем вес $\rho(t)$ есть функция, неотрицательная и суммируемая на (a,b), такая, что

$$\int_a^b [\rho(t)]^{-1}dt<\infty.$$

Таким образом, приближенное решение задачи ищется как функция вида

$$x_n(t)=\sum_{k=0}^{n-1} a_k^{(n)}p_k(t); \qquad (II.9.16)$$

коэффициенты $a_k^{(n)}$ определяются из системы уравнений

$$\sum_{k=0}^{m}\left[p_k^{(m)}(t_j^{(n)})-\sum_{i=0}^{m-1}c_i(t_j^{(n)})p_k^{(i)}(t_j^{(n)})\right]a_k^{(n)}=f(t_j^{(n)}); \; j=1,2,\dots,n. \qquad (II.9.17)$$

В предположении, что коэффициенты $c_k(t)$ и свободный член $f(t)$ непрерывны, доказывается, что система (II.9.17) однозначно разрешима при достаточно больших n; приближенные решения $x_n(t)$ сходятся к точному решению $x_*(t)$ в метрике $L_2^{(m)}(a,b;\rho(t))$. Погрешность аппроксимации имеет оценку

$$\|x_*^{(m)}-x_n^{(m)}\|_{L_2(a,b;\rho)}\le C\mathcal{E}_n(x_*^{(m)}), \qquad (II.9.18)$$

где $\mathcal{E}_n(u)$ есть наилучшее приближение функции $u(t)$ полиномами степени не выше n в метрике $L_2(a,b;\rho)$.

Обозначим через H_n подпространство пространства $L_2(a,b;\rho)$, порожденное полиномами вида (II.9.16) с произвольными коэффициентами $a_k^{(n)}$. Доказывается, что процесс вычисления приближенных решений $x_n(t)$ устойчив в паре пространств (H_n,R_n) тогда и только тогда, когда последовательность $\{p_k^{(m)}(t)\}$ сильно минимальна в $L_2(a,b;\rho)$.

ГЛАВА IУ

Погрешности алгоритма и округления

§ I. Число обусловленности.

I°. Пусть A - ограниченный и ограниченно обратимый оператор, действующий из банахова пространства X в такое же пространство Y. Числом обусловленности оператора A называется произведение

$$P_A = \|A\| \cdot \|A^{-1}\|. \qquad (IУ.I.I)$$

Очевидно, $P_A \geqslant 1$.

Если $X = Y = R_n$, то A есть числовая матрица порядка $n \times n$. Если эта матрица положительно определенная, то $\|A\| = \lambda_n^{(n)}$, $\|A^{-1}\| = 1/\lambda_n^{(1)}$, где $\lambda_n^{(1)}$ и $\lambda_n^{(n)}$ суть собственные числа матрицы A, соответственно наименьшее и наибольшее; в этом случае число обусловленности равно

$$P_A = \lambda_n^{(n)} / \lambda_n^{(1)}. \qquad (IУ.I.2)$$

Если A - произвольная особенная матрица в R_n, то матрица $A^* A$ положительно определенная. Пусть $\delta_n^{(n)}$ и $\delta_n^{(1)}$ - наибольшее и наименьшее собственные числа последней матрицы, иначе говоря, $\delta_n^{(n)}$ и $\delta_n^{(1)}$ суть наибольшее и наименьшее сингулярные числа матрицы A. Тогда

$$P_A = \sqrt{\delta_n^{(n)} / \delta_n^{(1)}}. \qquad (IУ.I.3)$$

2°. Число обусловленности оператора имеет большое значение в некоторых вопросах, касающихся оценки погрешности; один такой случай отмечен выше, в § 4 гл.I. Укажем еще один важный случай. Пусть A - ограниченный и ограниченно обратимый оператор, действующий в некотором сепарабельном гильбертовом пространстве H. Уравнение

$$Ax = f \qquad (IУ.I.4.)$$

будем приближенно решать по методу наименьших квадратов: каждому натуральному n сопоставим другое натуральное N и выделим из H подпространство $H^{(N)}$ измерений N; будем, далее, искать такой элемент $x_*^{(n)} \in H^{(N)}$, чтобы $\|Ax_*^{(n)} - f\| = \min$. Существование такого элемента очевидно. Оценим (см. статью автора и Р.К.Радевой [1]) погрешность аппроксимации $\rho_n = \|x_* - x_*^{(n)}\|$. Имеем

$$x_* - x_*^{(n)} = A^{-1} A (x_* - x_*^{(n)}) = A^{-1} (f - A x_*^{(n)})$$

и, следовательно, $\rho_n \leqslant \|A^{-1}\| \cdot \|f - Ax_*^{(n)}\|$. Если $x^{(n)}$ – произвольный элемент из $H^{(n)}$, то

$$\|f - Ax_*^{(n)}\| \leqslant \|f - Ax^{(n)}\| = \|A(x_* - x^{(n)})\| \leqslant \|A\| \cdot \|x_* - x^{(n)}\|; \quad \rho_n \leqslant P_A \|x_* - x^{(n)}\|.$$

Взяв справа нижнюю грань по всевозможным $x^{(n)} \subset H^{(n)}$ и обозначив

$$\forall x \in H, \; \mathcal{E}_n(x) = \inf_{x^{(n)} \in H^{(n)}} \|x - x^{(n)}\|,$$

получим интересующую нас оценку:

$$\rho_n = \|x_* - x_*^{(n)}\| \leqslant P_A \mathcal{E}_n(x_*). \qquad (IУ.I.5)$$

§ 2. Погрешность алгоритма в итерационном процессе.

I°. Как хорошо известно, если A – самосопряженный оператор в гильбертовом пространстве H и его нижняя и верхняя грани удовлетворяют неравенствам $0 < m \leqslant M < \infty$, то задачу (IУ.I.4) можно легко преобразовать так, чтобы она решалась итерациями. Достаточно, именно, заменить уравнение (IУ.I.4) таким:

$$(I - B)x = \frac{2}{M+m} f \; ; \quad B = I - \frac{2}{M+m} A \, . \qquad (IУ.2.I)$$

Нижняя и верхняя грани оператора B равны соответственно $+(M-m)/(M+m)$, поэтому

$$\|B\| = \frac{M-m}{M+m} = \frac{P_A - 1}{P_A + 1} < 1 \qquad (IУ.2.2)$$

и процесс итерации для уравнения (IУ.2.2) сходятся как прогрессия с знаменателем $(P_A - 1)/(P_A + 1)$. Ясно, что сходимость тем быстрее, чем меньше число обусловленности; если P_A достаточно велико, то сходимость итераций будет сколь угодно медленной.

2°. Высказанные в п.I° соображения применим к классическому методу Ритца. В сепарабельном гильбертовом пространстве H_0 рассмотрим задачу о минимуме функционала $|x|^2 - 2lx$, где $|\cdot|$ – норма в H_0 , а l – линейный функционал, определенный на всем пространстве H_0 и ограниченный в нем. Выберем в H_0 координатную систему $\{\varphi_n\}$ и обозначим через $H_0^{(n)}$ подпространство пространства H_0 с базисом $(\varphi_1, \varphi_2, \ldots, \varphi_n)$. Приближенное решение $x_*^{(n)}$ нашей вариационной задачи построим как элемент подпространства $H_0^{(n)}$, реализующий минимум функционала $|x|^2 - 2lx$ на этом подпространстве. Очевидно,

$$x_*^{(n)} = \sum_{k=1}^{n} a_k^{(n)} y_k ;$$

коэффициенты $a_k^{(n)}$ определяются из алгебраической системы

$$\sum_{k=1}^{n} [y_k, y_j] a_k^{(n)} = l y_j \; ; \; j = 1, 2, \ldots, n ,$$

в которой квадратные скобки означают скалярное произведение в H_0. Последнюю систему запишем в виде векторного уравнения

$$M_n a^{(n)} = f^{(n)} ; \qquad (IV.2.3)$$

смысл обозначений очевиден. Обозначим через $\lambda_n^{(1)}$ и $\lambda_n^{(n)}$ наименьшее и наибольшее собственные числа матрицы Ритца M_n и положим $B_n = I_n - 2M_n/(\lambda_n^{(1)} + \lambda_n^{(n)})$.

Уравнение (IV.2.3) заменим равносильным

$$a^{(n)} - B_n a^{(n)} = \frac{2 f^{(n)}}{\lambda_n^{(1)} + \lambda_n^{(n)}} , \qquad (IV.2.4)$$

которое разрешимо итерациями, так как $\|B_n\| = (P_n - 1)/(P_n + 1)$ $P_n = P_{M_n}$. Если начальное приближение равно нулю и если ограничиться N итерациями, то соответствующая погрешность алгоритма для уравнения (IV.2.4) имеет оценку

$$\frac{2\|f^{(n)}\|}{\lambda_n^{(1)} + \lambda^{(n)}} \cdot \frac{\|B\|^{N+1}}{1 - \|B\|} = \frac{\|f^{(n)}\|}{\lambda_n^{(1)}} \|B\|^{N+1} . \qquad (IV.2.5)$$

Если $\lambda_{n\,n\to\infty}^{(1)} \to 0$ или $\lambda_n^{(n)} \underset{n\to\infty}{\to} \infty$, то $P_n \to \infty$, $\|B_n\| \to 1$ и оценка погрешности алгоритма ухудшается. Желательно поэтому использовать такую координатную систему, для которой $\lambda_0 \le \lambda_n^{(1)} \le \lambda_n^{(n)} \le \Lambda_0$, где λ_0 и Λ_0 – положительные постоянные; иначе говоря, целесообразно в качестве координатной брать почти ортонормированную систему в H_0. В этом случае можно упростить определение матрицы B_n, именно, можно положить $B_n = I_n - 2M_n/(\Lambda_0 + \lambda_0)$ и вместо (IV.2.4) составить уравнение

$$a^{(n)} - B_n a^{(n)} = \frac{2 f^{(n)}}{\Lambda_0 + \lambda_0} , \qquad (IV.2.5')$$

также равносильное уравнению (IV.2.3). При таком определении B_n верхняя грань матрицы B_n соответственно равна

$$1 - \frac{2\lambda_n^{(1)}}{\Lambda_0 + \lambda_0} \ge 1 - \frac{2\lambda_0}{\Lambda_0 + \lambda_0} = \frac{\Lambda_0 - \lambda_0}{\Lambda_0 + \lambda_0}$$

и

$$1-\frac{2\lambda_n^{(n)}}{\lambda_0+\lambda_0}\le 1-\frac{2\lambda_0}{\lambda_0+\lambda_0}=\frac{\lambda_0-\lambda_0}{\lambda_0+\lambda_0}.$$

Отсюда $\|B_n\|\le(\lambda_0-\lambda_0)/(\lambda_0+\lambda_0)$; правая часть этого нер венства постоянна и меньше единицы. Для погрешности алгоритм после N итераций получаем оценку

$$\frac{\|f^{(n)}\|}{\lambda_0}\left(\frac{\lambda_0-\lambda_0}{\lambda_0+\lambda_0}\right)^{N+1}. \qquad (IУ.2.6$$

Докажем, что $\|f^{(n)}\|\le C$, где C не зависит от n . Почти тонормированная система сильно минимальна; в книге автора [8] доказано (см. теорему 7.2), что если координатная система сильно минимальна, то последовательность векторов $a^{(n)}$, дополненн нулями, имеет предел в l_2 . Отсюда вытекает, что последовательность норм $\|a^{(n)}\|_{R_n}$ ограничена. Пусть $\|a^{(n)}\|_{R_n}\le c_0=const$. Тогда из уравнения (IУ.2.3) получаем

$$\|f^{(n)}\|\le\|M_n\|\cdot\|a^{(n)}\|=\lambda_n^{(n)}\cdot\|a^{(n)}\|\le\lambda_0 c_0.$$

Теперь из (IУ.2.6) следует оценка погрешности алгоритма, не висящая от n :

$$\frac{\lambda_0 c_0}{\lambda_0}\left(\frac{\lambda_0-\lambda_0}{\lambda_0+\lambda_0}\right)^{N+1}. \qquad (IУ.2.7$$

Соображения и результаты настоящего параграфа легко перенося на обобщенный метод Ритца.

§ 3. Погрешность округления в рекуррентном процессе

1°. Обратимся к системе (Ш.6.2); будем считать, что она может быть как конечной, так и бесконечной. Для дальнейшего ва что операторы $A_{nk}+\Gamma_{nk}$ и свободные члены $f^{(n)}+\delta^{(n)}$ известны точ Для краткости обозначим

$$A_{nk}+\Gamma_{nk}=\hat{A}_{nk},\quad f^{(n)}+\delta^{(n)}=\hat{f}_n, \qquad (IУ.3.1$$

тогда система (Ш.6.2) запишется в виде

$$\forall n\ge 0,\ \sum_{k=0}^{r}\hat{A}_{nk}z^{(n)}=\hat{f}_n. \qquad (IУ.3.2$$

Нетрудно видеть, что если нормы $\|A_{nn}^{-1}\Gamma_{nk}\|$ достаточно малы, то операторы $\hat{A}_{nn}$ ограниченно обратимы, а произведения $\hat{A}_{nn}^{-1}\hat{A}_{nk}$ ограничены, и что если справедливо неравенство (Ш.7.1), то справедливо и аналогичное неравенство

$$n \geq 0, \quad \|\bar{z}^{(n)}\| \leq \hat{C}_0 \max_{j \leq n} \|\hat{A}_{jj}^{-1} \hat{f}^{(j)}\|, \quad \hat{C}_0 - const. \tag{IУ.3.3}$$

Точное решение системы (IУ.3.2) есть

$$n \geq 0, \quad \bar{z}^{(n)} = -\sum_{k=0}^{n-1} \hat{A}_{nn}^{-1} \hat{A}_{nk} \bar{z}^{(k)} + \hat{A}_{nn}^{-1} \hat{f}^{(n)}. \tag{IУ.3.4}$$

Вычисление правой части равенства (IУ.3.4) обычно связано с некоторой погрешностью. Пусть $v^{(1)}$, $v^{(2)}$, ..., $v^{(n-1)}$ – некоторые точно известные элементы, и пусть $\bar{z}^{(n)}$ – точное значение выражения

$$-\sum_{k=0}^{n-1} \hat{A}_{nn}^{-1} \hat{A}_{nk} v^{(k)} + \hat{A}_{nn}^{-1} \hat{f}^{(n)}; \tag{IУ.3.5}$$

примем, что в результате ошибок округления мы, вычисляя величину (IУ.3.5), получим не элемент $\bar{z}^{(n)}$, а некоторый другой элемент $v^{(n)} = \bar{z}^{(n)} + d^{(n)}$. Здесь $d^{(n)}$ – результирующая ошибка округления, о которой мы допустим, что она удовлетворяет неравенству

$$\|d^{(n)}\| \leq \varepsilon \|v^{(n)}\|, \tag{IУ.3.6}$$

где ε – заранее заданное положительное число, которое может быть сколь угодно малым. Мы принимаем, следовательно, что элемент $\bar{z}^{(n)}$ вычисляется с относительной погрешностью, не превосходящей ε.

2°. Обозначим $\hat{A}_{nn}^{-1} \hat{A}_{nk} = a_{nk}$, $\hat{A}_{nn}^{-1} \hat{f}^{(n)} = b^{(n)}$. В этих обозначениях система (IУ.3.4) и неравенство (IУ.3.3) примут вид

$$n \geq 0, \quad \bar{z}^{(n)} = \sum_{k=0}^{n-1} a_{nk} \bar{z}^{(k)} + b^{(n)} \tag{IУ.3.7}$$

и

$$n \geq 0, \quad \|\bar{z}^{(n)}\| \leq \hat{C}_0 \max_{j \leq n} \|b^{(j)}\|. \tag{IУ.3.8}$$

Решая систему (IУ.3.7) и учитывая ошибки округления, получим последовательность элементов

$$v^{(0)} = \bar{z}^{(0)} + d^{(0)} = b^{(0)} + d^{(0)} =: \bar{z}^{(0)} + r^{(0)},$$

$$v^{(1)} = a_{10} v^{(0)} + b^{(1)} + d^{(1)} = \bar{z}^{(1)} + a_{10} d^{(0)} + d^{(1)} =: \bar{z}^{(1)} + r^{(1)}$$

и т.д. Общие формулы имеют вид

$$v^{(n)} = \sum_{k=0}^{n-1} a_{nk} v^{(k)} + b^{(n)} + d^{(n)}, \qquad \text{(IУ.3.9)}$$

$$v^{(n)} = z^{(n)} + \gamma^{(n)}, \qquad \text{(IУ.3.I0)}$$

$$\gamma^{(n)} = \sum_{k=0}^{n-1} a_{nk} \gamma^{(k)} + d^{(n)}. \qquad \text{(IУ.3.II)}$$

Система (IУ.3.I0) с точностью до обозначений совпадает с системой (IУ.3.7); из (IУ.3.8) следует тогда

$$\|\gamma^{(n)}\| \leqslant \hat{C}_0 \max_{j\leqslant n} \|d^{(j)}\| \leqslant \hat{C}_0 \varepsilon \max_{j\leqslant n} \|v^{(j)}\| \leqslant \qquad \text{(IУ.3.I2)}$$

$$\leqslant \hat{C}_0 \varepsilon \left[\max_{j\leqslant n} \|z^{(j)}\| + \max_{j\leqslant n} \|\gamma^{(j)}\|\right].$$

Допустим, что нормы $\|z^{(n)}\|$ ограничены независимо от n; это будет, например, если выполнены условия теоремы Ш.7.I, а погрешности Γ_{nk} и $\delta^{(n)}$ достаточно малы. Пусть $\|z^{(n)}\| \leqslant \hat{C}_1$, тогда

$$\|\gamma^{(n)}\| \leqslant \hat{C}_0 \hat{C}_1 \varepsilon + \hat{C}_0 \varepsilon \max_{j\leqslant n} \|\gamma^{(j)}\|. \qquad \text{(IУ.3.I3)}$$

Пусть $k < n$. Из (IУ.3.I3) следует:

$$\|\gamma^{(k)}\| \leqslant \hat{C}_0 \hat{C}_1 \varepsilon + \hat{C}_0 \varepsilon \max_{j\leqslant k} \|\gamma^{(j)}\| \leqslant \hat{C}_0 \hat{C}_1 \varepsilon + \hat{C}_0 \varepsilon \max_{j\leqslant n} \|\gamma^{(n)}\|.$$

Допустим, что $\max_{j\leqslant k} \|\gamma^{(j)}\|$ достигается при $j = \ell$. Тогда

$$\|\gamma^{(\ell)}\| \leqslant \hat{C}_0 \hat{C}_1 \varepsilon + \hat{C}_0 \varepsilon \max_{j\leqslant n} \|\gamma^{(\ell)}\| \quad \text{или} \quad \|\gamma^{(\ell)}\| \leqslant \hat{C}_0 \hat{C}_1 \varepsilon / (1 - \hat{C}_0 \varepsilon)$$

и, тем более,

$$\|\gamma^{(n)}\| \leqslant \frac{\hat{C}_0 \hat{C}_1 \varepsilon}{1 - \hat{C}_0 \varepsilon}. \qquad \text{(IУ.3.I4)}$$

Таким образом, при сделанных выше предположениях ошибки округления ограничены и стремятся к нулю вместе с ε равномерно по n.

3°. Близкие результаты получаются и в том случае, когда вычисления ведутся с заданной абсолютной погрешностью. По-прежнему считаем, что при вычислении элемента (IУ.3.5) получается элемент $v^{(n)} = z^{(n)} + d^{(n)}$, но на этот раз $\|d^{(n)}\| \leqslant \varepsilon$; число ε по-прежнему считаем сколь угодно малым. Соотношения (IУ.3.7) – (IУ.3.II) остаются в силе; цепочка неравенств (IУ.3.I2) упрощается

ется и приводит к нужному результату:

$$\|r^{(n)}\| \leqslant \hat{C}_0 \max_{j \leqslant n} \|d^{(j)}\| \leqslant \hat{C}_0 \varepsilon ; \qquad (\text{IV.3.15})$$

требование ограниченности норм $\|x^{(n)}\|$ в данном случае не необходимо.

Замечание 1. Оценки (IV.3.14) и (IV.3.15) можно получить, предполагая, что операторы $\hat{A}_{nk}$ удовлетворяют условию (III.7.11).

Замечание 2. Во многих работах трактуются погрешности округления, возникающие при решении линейных алгебраических систем. Отметим некоторые, весьма немногие, из этих работ. Погрешностям округления прямых методов посвящены книги Уилкинсона [1, 2] и Воеводина [1]. В последней книге много внимания уделено вероятностным оценкам. Из работ, в которых изучаются ошибки округления итерационных процессов для линейных алгебраических систем, отметим более старую статью Голуба [1] и недавнюю работу Божняковского [1]. Результаты статьи Голуба мы кратко изложим в § 5 настоящей главы.

§ 4. Погрешность округления метода наискорейшего спуска.

1°. Как было сказано в § 8 гл.III, метод наискорейшего спуска сводится к рекуррентному вычислительному процессу; операторы B_n и свободные члены $f^{(n)}$ определяются формулами (III.8.5) и (III.8.12).

Погрешность округления $r^{(n)} = v^{(n)} - \tilde{x}^{(n)}$ в общем случае определяется рекуррентным соотношением (IV.3.11). Для определенности допустим, что вычисления производятся с заданной абсолютной погрешностью, так что $\|d^{(n)}\| < \varepsilon$. Для метода наискорейшего спуска

$$A_{nn} = I ; \quad A_{n,n-1} = -B_n = -(I - \sigma_{n-1} A) ; \quad A_{nk} = 0$$

$$k \leqslant n-2 ; \quad \Gamma_{nk} = 0 , \; k \neq n-1 ; \quad \Gamma_{n,n-1} = -\vartheta_n A$$

(см. § 8 гл.III). Отсюда

$$\hat{A}_{nn} = I ; \quad \hat{A}_{n,n-1} = -I + (\sigma_{n-1} + \vartheta_n) A ; \quad \hat{A}_{nk} = 0,$$

$$k \leqslant n-2 ; \quad a_{n,n-1} = I - (\sigma_{n-1} + \vartheta_n) A ; \quad a_{nk} = 0 , \; k \neq n-1.$$

Уравнение (IV.7.11) в данном случае имеет вид

$$\forall n \geqslant 1 , \; r^{(n)} = \left[I - (\sigma_{n-1} + \vartheta_n) A \right] r^{(n-1)} + d^{(n)} . \qquad (\text{IV.4.1})$$

Оценим величину $\mu_n := \|a_{n,n-1}\|$. Допустим, что σ_{n-1} при любом n вычисляется с малой относительной погрешностью ε', так что $|\vartheta_n| \le \varepsilon' \sigma_{n-1}$. Оператор $a_{n,n-1}$ – ограниченный самосопряженный; его нижняя и верхняя грани суть соответственно

$$1-(\sigma_{n-1}+\vartheta_n)M \ge 1-\left(\frac{1}{m}+\frac{\varepsilon}{m}\right)M = -\frac{M-m}{m}-\varepsilon'\frac{M}{m}$$

и

$$1-(\sigma_{n-1}+\vartheta_n)m \le 1-\left(\frac{1}{M}-\frac{\varepsilon'}{M}\right)m = \frac{M-m}{M}+\varepsilon'\frac{m}{M}.$$

Отсюда

$$\mu_n \le \frac{M-m}{m}+\varepsilon'\frac{M}{m} := \mu. \qquad (IУ.4.2)$$

Из этого соотношения вытекает неравенство

$$\forall n \ge 1,\ \|r^{(n)}\| \le \mu \|r^{(n-1)}\| + \varepsilon; \qquad (IУ.4.3)$$

при этом отметим, что $r^{(0)} = 0$, так как $r^{(0)}$ есть погрешность точно заданного элемента x_0. Решение системы неравенств (IУ.4.3) есть

$$\forall n \ge 1,\ \|r^{(n)}\| \le \frac{\mu^n - 1}{\mu - 1}\varepsilon. \qquad (IУ.4.4)$$

Если $M < 2m$, то при достаточно малом ε' будет $\mu < 1$ и

$$\|r^{(n)}\| < \frac{\varepsilon}{1-\mu}. \qquad (IУ.4.5)$$

В этом случае погрешность округления ограничена независимо от n и имеет порядок $O(\varepsilon)$. Если же $M \ge 2m$, то не исключено, что с возрастанием n погрешность округления может неограниченно возрастать. Тот же по существу результат получается, если округление производится с заданной относительной погрешностью (см. статью автора [33]).

2°. Рассмотрим пример. В пространстве $L_2(0, 1)$ рассмотрим уравнение

$$(at+1)x(t) = 1;\quad a = const > 0 \qquad (IУ.4.6)$$

с очевидным решением $x = (at+1)^{-1}$. Оператор умножения на функцию $at+1$ – положительно определенный и ограниченный, и уравнение (IУ.4.6) можно решать методом наискорейшего спуска; за начальное приближение возьмем $x_0 = 0$. Вычисления проведем для значений $a = 1/2$, 1, 3, 10. В нашем примере $m = 1$, $\frac{M-m}{M+m} = \frac{a}{a+2}$ и,

следовательно, $\frac{M-m}{M+m} = \frac{a}{a+2}$ и $\frac{M-m}{m} = a$.

Наименьшее число итераций, необходимое для получения теоретической погрешности, меньшей, чем 10^{-4}, было подсчитано по формуле (III.8.3), значения этого числа приведены в таблице I; результаты вычислений приведены в таблице 2. Они, как и следовало ожидать, вполне удовлетворительны для значения $a = 1/2$, при котором $M<2m$; они оказались удовлетворительными также при $a = 1$, когда $M = 2m$. Большие значения a приводят к следующему результату: при не очень большом числе итераций точность значений функции $x(t)$ повышается, затем эта точность резко падает; далее наступает переполнение и вычисление последующих итераций становится невозможным. Эти явления находятся в соответствии с результатами данного параграфа и § 8 гл.III.

Таблица I. Число необходимых итераций

$a = \frac{M-m}{m}$	1/2	1	3	10
$\frac{M-m}{M+m}$	1/5	1/3	1/2	5/6
число необходимых итераций	6	9	19	51

Таблица 2. Точные и приближенные значения искомой функции

узлы	$a = 1/2$		$a = 1$		$a = 3$		
	точное решение	результат 6 итераций	точное решение	результат 9 итераций	точное решение	результат 10 итераций	результат 15 итераций
0,0	1,0000	0,9920	1,0000	0,9986	1,0000	0,9525	-21609
0,1	0,9524	0,9484	0,9091	0,9061	0,7692	0,7601	-16654
0,2	0,9091	0,9075	0,8333	0,8331	0,6250	0,6240	- 5187
0,3	0,8696	0,8691	0,7692	0,7692	0,5263	0,5264	639,9
0,4	0,8333	0,8333	0,7143	0,7143	0,4545	0,4546	440,7
0,5	0,8000	0,8000	0,6667	0,6667	0,4000	0,4000	33,54
0,6	0,7692	0,7692	0,6250	0,6250	0,3571	0,3571	- 48,32
0,7	0,7407	0,7410	0,5882	0,5882	0,3226	0,3226	- 1792
0,8	0,7143	0,7152	0,5556	0,5556	0,2942	0,2942	17528
0,9	0,6897	0,6919	0,5263	0,5263	0,2703	0,2707	165169
1,0	0,6667	0,6711	0,5000	0,5019	0,2500	0,2527	455127

Продолжение таблицы 2

...- лы	$\lambda = 10$		
	точное решение	результат 8 итераций	результат 10 итераций
0,0	1,0000	0,9238	3983
0.1	0,5000	0,5016	-39 ,8
0,2	0,3333	0,3333	23,97
0,3	0,2500	0,2556	-79,02
0,4	0,2000	0,2052	-7152
0,5	0,1667	0,1667	-13,63
0,6	0,1429	0,1398	8091
0,7	0,1250	0,1286	4940
0,8	0,1111	0, 119	-3715
0,9	0,1000	0,0995	2581
1,0	0,0909	0,0930	-14112

§ 5. Погрешность округления метода Ричардсона для линейных алгебраических систем .

В качестве второго примера, к которому могут быть применены соображения § 3 данной главы, мы коротко изложим здесь содержание части статьи Колуба [1].

В пространстве R_m рассмотрим векторное уравнение

$$(I-B)x = f. \qquad (IV.5.$$

Здесь I – единичная матрица в R_m, B – симметричная матрица порядка $m \times m$, спектральный радиус которой $\rho < 1$; существует, следовательно, ограниченный оператор $(I-B)^{-1}$. Примем, что матрица B и вектор f заданы точно, без погрешности. По методу Ричардсона решение уравнения (IV.5.1) отыскивается как предел векторов $x^{(n)}$, которые вычисляются по двухшаговому рекуррентному процессу

$$x^{(n)} - \omega B x^{(n-1)} + (\omega-1) x^{(n-2)} = \omega f ; \qquad (IV.5.$$

$$\omega = \frac{2}{1+\sqrt{1-\rho^2}} ; \qquad (IV.5.$$

начальные приближения $x^{(0)}$ и $x^{(1)}$ задаются произвольно.

Рассмотрим несколько более общий рекуррентный процесс

$$t^{(n)}-\omega Bt^{(n-1)}+(\omega-1)t^{(n-2)}=h^{(n)} \qquad (IУ.5.4)$$

и докажем, что для этого процесса справедливо соотношение (Ш.7.4). С этой целью построим решения конечной системы

$$k\le n,\ t^{(k)}-\omega Bt^{(k-1)}+(\omega-1)t^{(k-2)}=h^{(k)}; \qquad (IУ.5.5)$$

$t^{(0)}$ и $t^{(1)}$ считаем заданными произвольно.

Матрицу B можно представить в виде $B=Q\Lambda Q^{*}$, где $\Lambda=(\lambda_1,\lambda_2,\dots,\lambda_m)$ есть диагональная матрица собственных чисел матрицы B, Q – ортогональная матрица, составленная из собственных векторов матрицы B, Q^{*} – матрица, сопряженная с Q. В статье Голуба [I] показывается, что решение системы (IУ.5.5) можно представить в виде

$$k\le n,\ t^{(k)}=\sum_{j=1}^{k-1}QS_{k-j}Q^{*}h^{(j)}; \qquad (IУ.5.6)$$

S_ν – диагональные матрицы, элементы которых, лежащие на главной диагонали, определяются следующим образом: если $\theta_j=\arccos(\lambda_j/\rho)$ и $\omega-1=\tau^2$, то

$$(S_\nu)_{jj}=\begin{cases}\tau^{\nu-1}\dfrac{\sin\nu\theta_j}{\sin\theta_j}, & \theta_j\neq 0,\pi;\\ \nu\tau^{\nu-1}, & \theta_j=0;\\ (-1)^{\nu-1}\nu\tau^{\nu-1}, & \theta_j=\pi.\end{cases}$$

Отсюда

$$\|QS_{k-j}Q^{*}h^{(j)}\|\le(k-j)\tau^{k-j-1}\|h^{(j)}\|;$$

по формуле (IУ.5.6)

$$k\le n,\ \|t^{(k)}\|\le\max_{j\le n}\|h^{(j)}\|\sum_{j=1}^{k}(k-j)\tau^{k-j-1}. \qquad (IУ.5.7)$$

Так как $0<\tau<1$, то сумма в (IУ.5.7) ограничена. Но тогда условие (Ш.7.4) выполнено, и можно оценить погрешность округления для процесса Ричардсона (IУ.5.2).

Решая бесконечную систему (IУ.5.2), мы из-за погрешностей округления получим вместо последовательности $\{x^{(n)}\}$ некоторую другую последовательность $v^{(n)}$, которая точно удовлетворяет другой бесконечной системе

$$v^{(n)} - \omega B v^{(n-1)} + (\omega-1) v^{(n-2)} = \omega f + \delta^{(n)}. \qquad \text{(IV.5.8)}$$

Допуская, например, что вычисления ведутся с фиксированной абсолютной погрешностью ε, имеем $\|\delta^{(n)}\| \le \varepsilon$. Из результатов § 3 данной главы вытекает, что

$$\|\gamma^{(n)}\| := \|v^{(n)} - x^{(n)}\| \le C\varepsilon, \quad C = const. \qquad \text{(IV.5.9)}$$

Постоянную C легко определить. Вычитая (IV.5.2) из (IV.5.8) получаем

$$\gamma^{(n)} - \omega B \gamma^{(n-1)} + (\omega-1)\gamma^{(n-2)} = \delta^{(n)}.$$

Сравнивая это с (IV.5.5), находим по неравенству (IV.5.8)

$$\|\gamma^{(n)}\| \le \max_{j\le n} \|\delta^{(j)}\| \sum_{j=1}^{n} (n-j) r^{n-j-1} \le \left[\frac{1-r^n}{(1-r)^2} - \frac{r^{n-1}}{1-r}\right] \max_{j\le n} \|\delta^{(j)}\|. \qquad \text{(IV.5.10)}$$

Несколько усилив последнее неравенство, получим

$$\|\gamma^{(n)}\| \le \frac{\varepsilon}{(1-r)^2}$$

и, следовательно,

$$C = \frac{1}{(1-r)^2}. \qquad \text{(IV.5.11)}$$

Нетрудно получить аналогичные оценки в предположении, что вычисления выполняются с заданной относительной погрешностью.

Значительная часть статьи Голуба [1] посвящена вероятностным оценкам погрешности округления для метода Ричардсона. Мы не будем здесь на этом останавливаться.

ГЛАВА У.

Погрешности метода конечных элементов

§ I. Обзор некоторых результатов (погрешность аппроксимации)

1°. Метод конечных элементов (МКЭ) в настоящее время является одним из самых распространенных и часто применяемых методов приближенного решения дифференциальных уравнений, особенно уравнений эллиптического типа. Этому методу посвящена обширная литература, как монографическая, так и журнальная. Мы отметим здесь только монографии Варга[I], автора этой книги [22], Обэна [I], Оганесяна и Руховца [I], Стренга и Фикса [I, 2], Сьярле [I]. Как кажется автору, в литературе, посвященной теории МКЭ, изучаются в основном две проблемы этой теории: техника МКЭ для различных классов задач механики, математической физики и т.п. и погрешности аппроксимации вычислительных процессов МКЭ (по терминологии настоящей книги). Небольшое количество работ посвящено другим погрешностям МКЭ; так, в некоторых работах автора исследована устойчивость МКЭ относительно искажения (см., например, [22]); в некоторых работах, из которых мы здесь отметим работу Руховца [I], исследовано число обусловленности матрицы МКЭ. В настоящем параграфе мы дадим весьма краткий обзор результатов, относящихся к погрешностям МКЭ. Мы не будем добиваться полноты обзора и остановимся только на результатах, которые кажутся нам наиболее важными.

2°. Для оценки погрешности аппроксимации МКЭ основную роль играет теорема об аппроксимации произвольной функции заданного функционального пространства линейными агрегатами координатных функций МКЭ. Если сетка "регулярная", т.е., любой геометрический элемент получается из некоторой фиксированной геометрической фигуры преобразованием масштаба и сдвигом на целочисленный вектор, то координатные функции МКЭ обычно можно получать такими же преобразованиями независимых переменных из некоторого конечного набора функций, удовлетворяющих определенным требованиям гладкости и имеющих компактный носитель. Мы называем эти функции "исходными".

В общем случае упомянутые выше линейные агрегаты могут не аппроксимировать функции заданного класса. Для того, чтобы такая аппроксимация имела место, необходимо и достаточно, чтобы исходные функции удовлетворяли некоторому конечному числу соот-

ношений. Смысл этих соотношений состоит в том, что некоторая линейные комбинации исходных функций со сдвинутыми аргументами предоставляют собой полиномы (см. Стрэнг и Фикс [I], Стрэнг [I] автор [I2, 22], Обэн [I]). Поясним это на примерах исходных функций, рассмотренных автором и Обэном.

3°. Пусть t - точка евклидова пространства R_m размерности m. В упомянутых выше работах автора [I2, 22] определяется понятие системы исходных функций размерности m. Это набор функций $\omega_{qs}(t)$, где $t \in R_m$, q - мультииндекс размерности m, который принимает конечное число значений, зависящее от натурального параметра s; функции ω_{qs} удовлетворяют следующим условиям: $\omega_{qs} \in C^{(s-1)}(R_m) \cap W_p^{(s)}(R_m)$, где p какое-нибудь число из промежутка $1 \le p \le \infty$; $\operatorname{supp} \omega_{qs} \subset \{t \in R_m : \underline{0} \le t \le \underline{2}\}$, где символ $\underline{\alpha}$ означает вектор, все составляющие которого в выбранной системе координат равны числу α; наконец, $\mathcal{D}^{(\alpha)} \omega_{qs}(\underline{1}) = \delta_{\alpha q}$; $0 \le |\alpha| \le s-1$. Координатные функции МКЭ определяются формулой

$$\varphi_{qjh} = \omega_{qs}\left(\frac{t}{h} - j\right); \quad j \in Z^m, \quad h > 0. \qquad (\text{У.I.I})$$

Доказывается следующее утверждение: множество линейных агрегатов функций (У.I.I) плотно в $W_p^{(s)}(R_m)$ тогда и только тогда, когда исходные функции удовлетворяют соотношениям, которые автор называет фундаментальными:

$$0 \le |\gamma| \le s, \quad \sum_{|q|=0}^{s-1} \sum_{t \in I} \frac{(\underline{1}-i)^{\gamma-q}}{(\gamma-q)!} \omega_{qs}(t+i) = \frac{t^\gamma}{\gamma!}; \qquad (\text{У.I.})$$

$$\underline{0} \le t \le \underline{1};$$

здесь I - совокупность радиусов-векторов вершин единичного куба $\underline{0} \le t \le \underline{1}$. Может случиться, что фундаментальные соотношения выполнены для $0 \le |\gamma| \le \tilde{s}$ при $\tilde{s} > s$. В этом случае мы говорим об усиленных фундаментальных соотношениях.

Пусть $\tilde{s} \ge s$, и фундаментальные соотношения верны при $0 \le |\gamma| \le \tilde{s}$, и пусть $x(t)$ - произвольная функция класса $W_p^{(\tilde{s}+1)}(R_m)$. Тогда при любом $h > 0$ существует такой линейный агрегат $x^h(t)$ функций (У.I.I), что

$$\forall s \le s, \quad \|x - x^h\|_{W_p^{(s)}(R_m)} \le C h^{\tilde{s}-s+1} \max_{|\alpha|=\tilde{s}+1} \|\mathcal{D}^\alpha x\|_{L_p(R_m)}. \qquad (\text{У.I.})$$

Заметим, что оценка (У.I.3) остается справедливой, если заменить $W_p^{(s)}$ и $W_p^{(s+1)}$ на $C^{(s)}$ и $C^{(s+1)}$.

Близкие результаты, но только для $p = 2$, содержатся у Фикса и Стрэнга [I] и Стрэнга [I]. Для $\tilde{s} = s$ неравенство (У.3.I) дано в статье Демьяновича и автора [I].

Рассмотрим простейший случай, когда $m = I$, и пусть исходная система состоит из s функций $\omega_{qs}(t)$, $0 \le q \le s-1$. Наибольшее значение $\tilde{s}$, для которого могут быть выполнены усиленные фундаментальные соотношения, равно $2s - I$. Соответствующие исходные функции определяются однозначно: они равны нулю вне отрезка $0 < t < 2$; на каждом из отрезков $0 < t < I$ и $I < t < 2$ они суть полиномы степени $2s - I$. Если положить $\omega_{qs}(t) = \sigma_{qs}(t+1)$, то

$$-1 < t < 0, \quad \sigma_{qs}(t) = \frac{(1+t)^s t^q}{q!} \sum_{k=0}^{s-q-1} a_k^{(s)} t^k ;$$

$$0 < t < 1, \quad \sigma_{qs}(t) = \frac{(1-t)^s t^q}{q!} \sum_{k=0}^{s-q-1} (-1)^k a_k^{(s)} t^k ; \qquad (У.I.4)$$

$$a_k^{(s)} = \binom{s+k+1}{k}.$$

Можно отметить, что сумма в первой из формул (У.I.4) есть отрезок ряда Тейлора функции $(I - t)^{-s}$.

4°. Из неравенства (У.I.3) легко вытекает оценка погрешности аппроксимации МКЭ в энергетической метрике. Рассмотрим эллиптическое уравнение (скалярное или векторное - безразлично) в конечной области $\Omega \subset R_m$ при таких краевых условиях, при которых оператор задачи положительно определен в $L_2(\Omega)$. Относительно области Ω предположим следующее: если все краевые условия - естественные, то Ω может быть произвольной областью с липшицевой границей; если среди краевых условий есть главные, то предположим следующее: существует такая последовательность $h \to 0$, что при любом h из этой последовательности область Ω может быть исчерпана кубами сетки с длиной ребра $2h$.

Пусть исходная система ω_{qs} имеет размерность m, и пусть она удовлетворяет усиленным фундаментальным соотношениям при $0 \le r \le \tilde{s}$, $\tilde{s} > s$. Наконец, допустим, что энергетическая норма

задачи эквивалентна норме $W_2^{(s+1)}$. Тогда

$$|x_* - x_*^{(h)}| \le C h^{\bar{s}-s+1} \max_{|\alpha|=s+1} \|\mathcal{D}^\alpha x_*\|_{2,\Omega}; \qquad (У.I.5)$$

здесь x_* – точное решение задачи, $x_*^{(h)}$ – ее приближенное решение по МКЭ.

Если среди краевых условий есть главные, а Ω имеет произвольную липшицеву границу, то удается установить значительно худшую оценку: $|x_* - x_*^{(h)}| = O(\sqrt{h})$.

5°. Представляет интерес оценка постоянной C в неравенстве (У.I.3). Такие оценки получены для ряда случаев в работах автора [24, 31]). Здесь отметим случай, который кажется нам более важным. Пусть $m = I$, исходные функции определены формулами (У.I.4) и в формуле (У.I.3) $W_p^{(s)}(R_m)$ и $W_p^{(\bar{s}+1)}(R_m)$ заменены на $C^{(s)}(R_m)$ и $C^{(\bar{s}+1)}(R_m)$; пусть еще $\bar{s} = 2s-1$. Постоянная C в неравенстве (У.I.3) зависит от s и $\bar{s}$; обозначим ее через $C(s, \bar{s})$. Доказывается справедливость оценки

$$C(s,\bar{s}) < \frac{1}{(2s-\bar{s})!} + \frac{13e^{3/2} - 8e^{3/2}}{5} \frac{2^{2s-2} s\mu(s,\bar{s})}{(s+1)!} < \qquad (У.I.6)$$

$$< \frac{1}{(2s-\bar{s})!} + \frac{8{,}54\, s\mu(s,\bar{s}) \cdot 2^{2s-2}}{(s+1)!};$$

здесь $\mu(s, \bar{s})$ – функция двух целочисленных аргументов, определяемая формулой

$$\mu(s,\bar{s}) = \begin{cases} (s-1)^2 (s-2)^2 \dots (s-\frac{\bar{s}-1}{2})^2, & \bar{s} \text{ – нечетное}, \\ (s-1)^2 (s-2)^2 \dots (s-\frac{\bar{s}}{2}+1)^2 (s-\frac{\bar{s}}{2}), & \bar{s} \text{ четное}. \end{cases} \qquad (У.I.7)$$

6°. Обэн ([I], гл.У) также рассматривал кубическую сетку в R_m; его система исходных функций состоит из единственной функции $\mu(t) \in W_2^{(s)}(R_m)$, имеющей компактный носитель. Пусть $x(t)$ – произвольная функция класса $W_2^{(s+1)}(R_m)$. Доказывается, что существует функция вида

$$x^h(t) = \sum_{j \in Z^m} a_j^{(s)} \mu\left(\frac{t}{h} - j\right), \qquad (У.I.8)$$

удовлетворяющая неравенству

$$\|x-x^h\|_{(s)} \le Ch^{s+1-\bar{s}}\|x\|_{(s+1)}, \qquad (У.I.9)$$

если исходная функция $\mu(t)$ удовлетворяет фундаментальным соотношениям (по терминологии Обэна, "критерию сходимости")

$$\sum_{k\in\bar{Z}^m}\frac{k^j}{j!}\mu(x-k)=\sum_{0\le\rho\le j}\beta^{j-\rho}\frac{x^\rho}{\rho!};\ |j|\le s,\ \int_{R_m}\mu(t)dt=1. \qquad (У.I.10)$$

Здесь $\bar{Z}^m$ – множество всех m-мерных целочисленных векторов, $\beta^{j-\rho}$ – некоторые постоянные. Условие (У.I.10) не только достаточно, но и необходимо. Отметим, что довольно близкие результаты содержатся в несколько более раннем препринте Фикса и Стрэнга [I].

Приведем еще один пример из книги Обэна.

Пусть $\pi_1(t)$ – характеристическая функция интервала $0<t<1$, а $\pi_k(t)$, $k=2,3,\dots$, последовательные свертки:

$$\pi_k(t)=\int_{-\infty}^{+\infty}\pi_1(\tau)\pi_{k-1}(t-\tau)d\tau=\int_0^1\pi_{k-1}(t-\tau)\,d\tau. \qquad (У.I.11)$$

Очевидно, $\pi_{s+2}\in W_2^{(s)}(R_1)$; доказывается, что функция $\pi_{s+1}(x)$ удовлетворяет соотношениям (У.I.10), при этом

$$\beta^j=\sum_{j_1+\dots+j_{s+1}=j}\frac{(-1)^j}{(j_1+1)!\dots(j_{s+1}+1)!}. \qquad (У.I.12)$$

7°. В своей статье [4] Шапошникова рассмотрела ореди других вопрос об оценке погрешности производных порядка δ решения первой краевой задачи для уравнения

$$\sum_{|\alpha|,|\beta|=0}(-1)^\alpha\mathcal{D}^\alpha(A_{\alpha\beta}\mathcal{D}^\beta x)=f(t) \qquad (У.I.13)$$

в m-мерном кубе Ω. Как обычно, предполагается, что оператор задачи положительно определенный в $L_2(\Omega)$, уравнение (У.I.13) в Ω не вырождается и коэффициенты $A_{\alpha\beta}$ – достаточно гладкие. Допустим, что точное решение задачи $x_*\in W_\rho^{(\nu)}(\Omega)$, где $1<\rho<\infty$ и $\nu>\delta$; через $x_*^{(h)}(t)$ обозначим приближенное решение, полученное по МКЭ. Доказывается, что справедлива следующая оценка погрешности аппроксимации:

$$\|x_* - x_*^{(h)}\|_{W_p^{(s)}(\Omega)} \leq \begin{cases} ch^{\tau-s+\frac{m}{2}-\frac{m}{p}}\,\omega_{L_p}^{(\tau)}(x_*,h), & p\geq 2; \\ ch^{\tau-s+\frac{m}{2}-\frac{m}{p}}\,\omega_{L_p}^{(\tau)}(x_*,h), & 1<p<2. \end{cases} \qquad (У.I.I4)$$

Здесь $\omega_{L_p}^{(\tau)}(x_*,h)$ – наибольший L_p-модуль непрерывности производных порядка τ решения $x_*(t)$.

Джиккариани [I3] получил оценку

$$\|x_* - x_*^{(h)}\|_{(s)} \leq Ch^{\tau-\tilde{\tau}}\,\|x_*\|_{(\tilde{\tau})} \qquad (У.I.I5)$$

в предположении, что $\tau > \tilde{\tau} > s$ и что краевые условия могут быть произвольными – требуется только, чтобы оператор данной задачи был положительно определенным в $L_2(\Omega)$.

8°. Небольшое число работ посвящено оценке погрешности аппроксимации МКЭ для вырождающихся дифференциальных уравнений. Не приводя результатов, отметим работу Гусмана и Оганесяна [I], в которой рассмотрено слабо вырождающееся эллиптическое уравнение в прямоугольнике на двумерной плоскости, и работу автора [I4] (см. также [22]), в которой рассмотрено линейное обыкновенное дифференциальное уравнение второго порядка с любым показателем вырождения.

§ 2. Обзор некоторых результатов; погрешность искажения и число обусловленности МКЭ.

I°. Сформулируем полученные нами (см. работы автора [I2, I3, 22]) результаты, которые относятся к устойчивости МКЭ относительно искажения и к оценке числа обусловленности матрицы МКЭ; о значении числа обусловленности для оценок погрешностей сказано выше, в гл. IУ.

Рассмотрим эллиптическое дифференциальное уравнение

$$\sum_{|\alpha|,|\beta|=0}^{s} (-1)^{\alpha} D^{\alpha}(A_{\alpha\beta} D^{\beta} x) = f(t); \qquad (У.2.I)$$

будем считать, что t меняется в конечной области $\Omega \subset R_m$ с липшицевой границей. Краевые условия будем считать такими, что оператор задачи (У.2.I) при этих условиях – положительно определенный в $L_2(\Omega)$. Примем еще, что $f \in L_2(\Omega)$ и что дифференциальный оператор (У.2.I) не вырождается в $\bar{\Omega}$. Наконец, примем, что $x(t)$ есть K-мерный вектор, $K \geq 1$.

Нашу краевую задачу будем приближенно решать по МКЭ при координатных функциях, описанных в п.3° § I. Устойчивость и число обусловленности исследуем в таких предположениях: либо речь идет о первой краевой задаче, тогда область Ω может быть произвольной липшицевой, либо Ω – куб, тогда краевые условия могут быть произвольными.

Рассмотрим сначала первую краевую задачу. Пусть h – шаг сетки. Рассмотрим открытые кубы сетки с ребром h. Радиусы-векторы их вершин имеют вид jh, где $j \in \mathbb{Z}^m$ есть целочисленный вектор, который будем называть номером соответствующей вершины. У каждого куба сетки есть вершина с наименьшим номером j. Обозначим через $\mathcal{J}^h$ множество таких наименьших номеров, соответствующих кубам сетки, лежащим внутри Ω. Приближенное решение задачи будем искать в виде

$$x^h(t) = \sum_q \sum_{j \in \mathcal{J}^h} a_{qj}^{(h)} \varphi_{qjh}(t), \qquad (У.2.2)$$

где φ_{qjh} – функции (У.I.I). Обозначим через $a^{(h)}$ вектор коэффициентов $a_{qj}^{(h)}$ и через $\mathcal{N}$ его размерность. Этот вектор определяется из алгебраической системы

$$A_h a^{(h)} = f^{(h)}, \qquad (У.2.3)$$

где A_h – матрица порядка $\mathcal{N} \times \mathcal{N}$, составленная из скалярных произведений $[\varphi_{qjh}, \varphi_{q',j',h}]$, а $f^{(h)}$ – вектор с составляющими $(f, \varphi_{q',j',h})$.

Наряду с обычным евклидовым пространством $\mathcal{N}$ измерений $R_{\mathcal{N}}$ рассмотрим два $\mathcal{N}$-мерных гильбертовых пространства $\bar{X}_{\mathcal{N}}$ и $\bar{Y}_{\mathcal{N}}$ с нормами

$$\|a\|_{\bar{X}_{\mathcal{N}}} = h^{m/2} \|a\|_{R_{\mathcal{N}}}, \quad \|a\|_{\bar{Y}_{\mathcal{N}}} = h^{-m/2} \|a\|_{R_{\mathcal{N}}}. \qquad (У.2.4)$$

Пусть $\{h_\kappa\}$ – последовательность значений h, стремящаяся к нулю, и $\{\mathcal{N}_\kappa\}$ соответствующая последовательность размерностей векторов коэффициентов. Доказывается, что процесс (У.2.3) определения вектора коэффициентов $a^{(h)}$ устойчив (в смысле второго определения) в последовательности пар пространств $(\bar{X}_{\mathcal{N}_\kappa}, \bar{Y}_{\mathcal{N}_\kappa})$. Существуют, следовательно, такие постоянные $p, q, \tau > 0$, что если $\|\Gamma_{h_\kappa}\| \leq \tau$, то искаженное уравнение

$$(A_{h_\kappa} + \Gamma_{h_\kappa}) b^{(h_\kappa)} = f^{(h_\kappa)} + \delta^{(h_\kappa)}$$

имеет единственное решение и

$$\|b^{(h_k)} - a^{(h_k)}\|_{\bar{x}_{h_k}} \leq \rho \|\Gamma_{h_k}\|_{\bar{x}_{h_k} \to \bar{y}_{h_k}} + q \|\delta^{(h_k)}\|_{\bar{y}_{h_k}}. \qquad (У.2.5$$

Далее, пусть

$$z^{(h_k)}(t) = \sum_{q} \sum_{j \in \bar{y}_{h_k}} b_{qj}^{(h_k)} \varphi_{qj|h_k}$$

– искаженное приближенное решение. Доказывается, что существуют такие постоянные $\rho', q', r' > 0$, что если $\|\Gamma_{h_k}\| \leq r'$, то

$$|z^{(h_k)} - x^{(h_k)}| \leq \rho' \|\Gamma_{h_k}\| + q' \|\delta^{(h_k)}\|; \qquad (У.2.6$$

иначе говоря, доказывается устойчивость процесса вычисления приближенного решения в соответствующей последовательности пар пространств.

Для той же первой краевой задачи доказывается, что число обусловленности матрицы A_h имеет порядок $O(h^{-2})$; см. работу автора [I, 22].

Те же результаты получаются, если область Ω – куб, а краевые условия – естественные. Эти результаты верны и для процесса вычисления собственных чисел по МКЭ; см. книгу автора [22].

2°. Руховец [I] предложил некоторое простое видоизменение функционала энергии для уравнения (У.2.I) и соответствующих естественных краевых условий, в результате которого точное решение задачи не меняется, а оценка $O(h^{-2})$ для числа обусловленности матрицы МКЭ оправедлива в случае любой области с достаточно гладкой границей. Как доказал Чистов [I], для функционала Руховца оправедливы приведенные в § I данной главы теоремы об устойчивости процесса МКЭ, если граница области Ω достаточно гладкая.

3°. Первой работой, в которой исследовалась устойчивость процесса МКЭ, была статья Демьяновича [I]. В этой статье рассмотрена несамосопряженная эллиптическая задача

$$-\sum_{i,k=1}^{m} \frac{\partial}{\partial t_i}\left(a_{ik} \frac{\partial x}{\partial t_k}\right) + \sum_{i=1}^{m} a_i \frac{\partial x}{\partial t_i} + ax = f(t), \; t \in \Omega. \qquad (У.2.7)$$

На область Ω накладываются некоторые ограничения, которые выполняются, например, если $\partial\Omega \in C^{(2)}$. Вводится кусочно полиномиальная исходная функция; с ее помощью строятся координатные функции МКЭ. Вводится в рассмотрение пространство сеточных фу

ций с узлами, лежащими внутри Ω ; на этом пространстве вводятся две гильбертовы метрики, $(,)_0$ и $(,)_1$ которые в пределе, при $h \to 0$, переходят соответственно в метрику L_2 и в энергетическую метрику главной части оператора задачи. Норма произвольной матрицы A определяется формулой

$$\|A\| = \sup \frac{|(Av, v')_0|}{\|v\|_1 \cdot \|v'\|_1} ; \qquad (У.2.8)$$

здесь v, v' – произвольные элементы упомянутого выше пространства сеточных функций. Пусть Γ_h, $\delta^{(h)}$, $z^{(h)} - x^{(h)}$ суть соответственно искажения матрицы МКЭ, столбца свободных членов и приближенного решения; само приближенное решение обозначим через $x^{(h)}$. Доказывается существование таких постоянных $\rho, q, \gamma > 0$, что если $\|\Gamma_h\| < \gamma$, то

$$\|z^{(h)} - x^{(h)}\|_1 \le \rho \|\Gamma_h\| + q \|\delta^{(h)}\|_0 . \qquad (У.2.9)$$

В той же статье исследуется устойчивость процесса МКЭ для уравнения (У.2.7) в предположении, что искажение Γ_h принадлежит некоторому специальному пространству матриц. Идеи работы Демьяновича [I] развиты в последующих работах того же автора [2,3].

§ 3. МКЭ с одной кусочно полиномиальной исходной функцией

I°. В ближайших параграфах данной главы мы подробнее исследуем погрешности МКЭ в том случае, когда исходная система содержит только одну функцию, удовлетворяющую фундаментальным соотношениям (У.I.IO). Для простоты ограничимся более частным случаем, когда размерность $m = I$, а исходная функция есть кусочно полиномиальная функция $\pi_k(t)$ (формула (У.I.II)). Погрешность аппроксимации в этом случае исследована в работах Стренга и Фикса [I] и Обэна [I]; см. также п.6° § I данной главы, и в этом плане мы ограничемся тем, что оценим постоянную C в неравенстве (У.I.9). Этому будет посвящен настоящий параграф. В § 4 мы рассмотрим погрешность искажения, а в § 5 – число обусловленности матрицы МКЭ для той же исходной функции $\pi_k(t)$. В § 6 мы коротко рассмотрим погрешности МКЭ для симплициальных сеток.

2°. Начнем с вывода оценки производных полиномов $\pi_k(t)$. Имеем

$$\pi_0(t) = \begin{cases} 1, & t \in (0,1), \\ 0, & t \notin (0,1) \end{cases}$$

и

$$\pi_k(t)=\int_0^1 \pi_{k-1}(t-\tau)d\tau; \qquad (У.3.1$$

отсюда, очевидно,

$$0\leqslant \pi_{k-1}(t)\leqslant 1. \qquad (У.3.2$$

Замена $t-\xi=\tau$ дает

$$\pi_k(t)=\int_{t-1}^{t}\pi_{k-1}(\xi)d\xi$$

и, следовательно,

$$k\geqslant 1,\ \pi_k^{(s)}(t)=\pi_{k-1}^{(s-1)}(t)-\pi_{k-1}^{(s-1)}(t-1). \qquad (У.3.3$$

В частности, при $s=1$, $|\pi_k'(t)|\leqslant \pi_{k-1}(t)+\pi_{k-1}(t-1)\leqslant 2$, и индукции

$$|\pi_k^{(s)}(t)|\leqslant 2^s. \qquad (У.3.4$$

С целью оценить постоянную в формуле (У.I.9), мы воспроизведем здесь вывод этой формулы, данный в книге Обэна [I], дополнив его оценками констант, вводимых в ходе рассуждений. Начнем с построения аппроксимирующей функции $x^{(h)}(t)$. Наряду с исходной функцией $\mu(t)=\pi_k(t)$ введем в рассмотрение еще одну функцию $\lambda(t)$ со следующими свойствами: носитель этой функции компактен, она ограничена и измерима и

$$\int_a^{+\infty}\lambda(t)dt=1,\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\mu(t)\lambda(\tau)(t-\tau)^k\,dt\,d\tau=\begin{cases}1, & k=0\\ 0, & 0<k\leqslant s\end{cases}. \qquad (У.3.$$

Пусть $x\in W_2^{(s+1)}(R_1)$, причем $\operatorname{supp} x$ компактен. На каждом интервале $(jh, (j+1)h)$ разложим $x(t)$ по формуле Тейлора в круг точки jh и введем обозначения

$$X_{jh}^{s}(x)(t)=\sum_{q=0}^{s-1}\frac{h^q}{q!}x^{(q)}(jh)\left(\frac{t}{h}-j\right)^q,$$

$$Y_{jh}^{s}(x)(t)=\int_{jh}^{t}\frac{(t-\tau)^{s-1}}{(s-1)!}x(\tau)d\tau, \qquad (У.3.$$

тогда $x(t)=X_{jh}^{s}(x)(t)+Y_{jh}^{s}(x^{(s)})(t)$.

Вводится оператор, который мы обозначим через g_h и который действует по правилу

$$(g_h x)(t) = \sum_{\rho \in Z} x_h^\rho \mu\left(\frac{t}{h} - \rho\right); \quad \mu(t) = \pi_k(t);$$

$$x_h^\rho = \frac{1}{h}\int_{-\infty}^{+\infty} x(t)\lambda\left(\frac{t}{h} - \rho\right) dt.$$

Доказывается, что этот оператор оставляет инвариантными усочно полиномиальные функции степени k.

Положим теперь $x^{(h)}(t) = (g_h x)(t)$ и оценим разность

$$\|\mathcal{D}^j(x(t) - x^{(h)}(t))\|_{L_2}^2 = \sum_{i \in Z} U_{ih}^2; \quad \mathcal{D}^j = \frac{d^j}{dt^j};$$

$$U_{ih}^2 = \int_{ih}^{(i+1)h} \left[\mathcal{D}^j(I - g_h)(y_{ih}^s(x^{(s)})(t))\right]^2 dt \leq \qquad (У.3.7)$$

$$\leq 2\int_{ih}^{(i+1)h} \left[y_{ih}^{s-j}(x^{(s)})(t)\right]^2 dt + 2\int_{ih}^{(i+1)h} \left[\mathcal{D}^j g_h(y_{ih}^s(x^{(s)})(t)\right]^2 dt.$$

Первый член справа оценим по неравенству Коши – Буняковского:

$$\int_{ih}^{(i+1)h} \left[y_{ih}^{s-j}(x^{(s)})(t)\right]^2 dt \leq$$

$$\leq \frac{h^{2(s-j)}}{[(s-j-1)!]^2(2s-2j-1)(2s-2j)} \int_{ih}^{(i+1)h} [x^{(s)}(t)]^2 dt. \qquad (У.3.8)$$

Займемся вторым интегралом. Временно положим $y_{ih}^s(x^{(s)})(t) = v(t)$. По определению оператора g_h имеем

$$\mathcal{D}^j g_h(y_{ih}^s(x^{(s)})(t) = (\mathcal{D}^j g_h v)(t) = h^{-j}\sum_{\rho \in Z} v_h^\rho \mu^{(j)}\left(\frac{t}{h} - \rho\right),$$

$$v_h^\rho = \frac{1}{h}\int_{(a+\rho)h}^{(b+\rho)h} \lambda\left(\frac{\tau}{h} - \rho\right) v(\tau) d\tau,$$

где $(a, b) \supset \operatorname{supp} \lambda(t)$. Далее,

$$(\mathcal{D}^j g_h v)^2 \leq h^{-2j}\sum_{\rho \in Z} (v_h^\rho)^2 \sum_{\rho \in Z} [\pi_k^{(j)}\left(\frac{t}{h} - \rho\right)]^2. \qquad (У.3.9)$$

Каждая из производных $\pi_k^{(j)}\left(\frac{t}{h} - \rho\right)$, отличных от нуля (их число не превосходит $k+1$), не превосходит 2^j, поэтому

$$(\mathcal{D}^j g_h v)^2 \leq (k+1) 2^{2j} h^{-2j} \sum_{\rho \in Z} (v_h^\rho)^2. \qquad (У.3.10)$$

Суммировать в (У.3.10) достаточно по тем ρ, для которых $0 \leq$

$\leq \frac{t}{h} - \rho \leq \kappa+1$; число тех ρ не превосходит $\kappa+1$

Пусть $|\lambda(t)| \leq c_0 = const$. Тогда

$$(v_h^\rho)^2 \leq \frac{1}{h} c_0^2 (b-a) \int_{(i+\rho)h}^{(i+\rho)h} v^2(t)\,dt .$$

Интегрируя неравенство (У.3.10) по интервалу $(ih,(i+1)h)$, получаем

$$\int_{ih}^{(i+1)h} (\mathcal{D}^j q_h v)^2 dt \leq c_0^2 (b-a) 2^{2j} h^{-2j} \sum \int_{(a+\rho)h}^{(b+\rho)h} v^2(t)\,dt . \qquad (У.3.11)$$

Найдем пределы изменения ρ , когда $ih \leq t \leq (i+1)h$. Число ρ должно удовлетворять неравенствам $i-\rho \geq 0$, $i+1-\rho \leq \kappa+1$, которые означают, что $\left(\frac{t}{h}-\rho\right) \in [0,\kappa+1] = \operatorname{supp} \pi_\kappa(t)$. Отсюда следует, что в (У.3.11) $\rho = i-\kappa,\ i-\kappa+1,\ldots,i$, и это неравенство можно заменить следующим:

$$\int_{ih}^{(i+1)h} (\mathcal{D} q_h v)^2 dt \leq c_0^2 (b-a)(\kappa+1) 2^{2j} h^{-2j} \int_{(a+i-\kappa)h}^{(b+i)h} v^2(t)\,dt .$$

Вспомним, что $v = y_{ih}^s(x^{(s)})$. Это приведет нас к неравенству

$$\int_{ih}^{(i+1)h} (\mathcal{D}^j q_h y_{ih}^s(x^{(s)}))^2 dt \leq$$

$$\leq c_0^2 (b-a)(\kappa+1) 2^{2j} h^{-2j} \int_{(a+i-\kappa)h}^{(b+i)h} (y_{ih}^s(x^{(s)}))^2 dt .$$

Аналогично неравенству (У.3.8) нетрудно получить неравенство

$$\int_{(a+i-\kappa)h}^{(b+i)h} (\mathcal{D}^j q_h y_{ih}^s(x))^2 dt \leq \frac{h^{2(s-j)} c_0^2 (b-a) 2^{2j} (b^{2s}-(a-\kappa)^{2s})}{2s(2s-1)((s-1)!)^2} \int_{ih}^{(i+1)h} (x^{(s)}(t))^2 dt .$$

Отсюда и из (.3.9) найдем

$$U_{ih}^2 \leq C_1^2 h^{2(s-j)} \int_{ih}^{(i+1)h} (x^{(s)})^2 dt , \qquad (У.3.12)$$

где

$$C_1^2 = [(s-j-1)!]^{-2} [(s-j)(2s-2j-1)]^{-1} + \frac{c_0^2 (b-a)(\kappa+1) 2^{2j} (b^{2s}-(a-\kappa)^{2s})}{s(2s-1)((s-1)!)^2} . \qquad (У.3.13)$$

Наконец, просуммировав неравенство (У.3.I2) по i, получим

$$\int_{-\infty}^{+\infty}[\mathfrak{D}^{j}(x(t)-x^{(h)}(t))]^{2}dt \le C_1^2 h^{2(s-j)}\int_{-\infty}^{+\infty}(x^{(s)}(t))^{2}dt; \quad (У.3.I4)$$

величина C_1 определяется формулой (У.3.I3).

Допустим, что энергетическая норма нашей задачи имеет оценку

$$|x|^2 \le C_2^2 \sum_{j=0}^{K} \|x^{(j)}\|_2^2. \quad (У.3.I5)$$

Обозначим через x_* точное решение задачи, а через $x_*^{(h)}$ приближенное решение по МКЭ; допустим еще, что $x_* \in W_2^{(s)}$. Тогда, полагая $h < I$, находим

$$|x_* - x_*^{(h)}|^2 \le C_2^2 \sum_{j=0}^{K} \|\mathfrak{D}^{j}(x_* - x_*^{(h)})\|^2 \le \quad (У.3.I6)$$

$$\le (C_1 C_2)^2 \frac{h^{2(s-K)}}{1-h^2}\|x_*^{(s)}\|_2^2;$$

постоянные C_1 и C_2 определяются формулами (У.3.I3) и (У.3.I5).

§ 4. Погрешность вложения

I°. Рассмотрим обыкновенное дифференциальное уравнение

$$\sum_{k=0}^{s}(-1)^k \frac{d^k}{dt^k}\left[p_k(t)\frac{d^k x}{dt^k}\right] = f(t),\ 0<t<1. \quad (У.4.I)$$

Коэффициенты $p_k(t)$ предполагаем измеримыми и ограниченными, причем коэффициент $p_s(t)$ положительно ограничен снизу; краевые условия таковы, что оператор задачи положительно определен в $L_2(0, I)$.

Приближенное решение задачи будем искать в виде

$$x^{(h)}(t) = \sum_{i\in Z} a_j^{(h)} \pi_s\left(\frac{t}{h} - j\right),\ h = \frac{1}{(s+1)n},$$

где n – произвольно выбранное натуральное число. Нас интересуют только значения $t \in (0, I)$, и достаточно суммировать лишь по тем j, для которых

$$\frac{t}{h} - j \in \operatorname{supp}\pi_s = [0, s+1]. \quad (У.4.2)$$

Пусть $t/h = j_0 + \tau$, где j_0 целое число и $0 \le \tau < 1$. Тогда $t/h - j = j_0 - j + \tau$, и условие (У.4.2) будет выполнено, если $0 \le j_0 - j \le s$, или $j_0 - s \le j \le j_0$. Далее, j_0 меняется в пределах $0 \le j_0 \le (s+1)n - 1$, отсюда получаются пределы изменения j: $-s \le j \le (s+1)n - 1$, и мы приходим к следующему выражению

для $x^{(h)}$:

$$x^{(h)} = \sum_{j=-3}^{n(3+1)} a_j^{(h)} \Pi_3\left(\frac{t}{h} - j\right). \qquad (У.4.3)$$

Система уравнений МКЭ имеет вид:

$$\sum_{j=-3}^{n(3+1)} a_j^{(h)} [\varphi_{3jh}, \varphi_{3j'h}] = (f, \varphi_{3j'h}); \quad \varphi_{3jh} = \Pi_3\left(\frac{t}{h} - j\right); \qquad (У.4.4)$$

как обычно, здесь квадратными скобками обозначено скалярное произведение в энергетической метрике, а круглыми скобками – скалярное произведение в $L_2(0, 1)$.

Обозначим через $N = n(3+1)+3$ порядок системы (У.4.4), через A_h – матрицу этой системы, через $a^{(h)}$ и $f^{(h)}$ – векторы с составляющими $a_j^{(h)}$ и (f, φ_{3jh}) соответственно. Система (У.4.4) записывается в виде

$$A_h a^{(h)} = f^{(h)}. \qquad (У.4.5)$$

2°. Нормы $\|x^{(h)}\|_2$ и $\|a^{(h)}\|_2$ связаны двусторонней оценкой; выведем ее. Имеем

$$\|x^{(h)}\|_2^2 = \sum_{l=1}^{(3+1)n} \int_{(l-1)h}^{lh} [x^{(h)}]^2 dt.$$

На промежутке $((l-1)h, lh)$ отличны от тождественного нуля только функции $\varphi_{3jh}(t)$, $j = l-3-1, l-3, \ldots, l-1$, поэтому

$$\int_{(l-1)h}^{lh} [x^{(h)}]^2 dt = \int_{(l-1)h}^{lh} \left[\sum_{j=l-3-1}^{l-1} a_j^{(h)} \Pi_3\left(\frac{t}{h} - j\right)\right]^2 dt = h \int_0^1 \left[\sum_{p=0}^{3} a_{l-1-p-3}^{(h)} \Pi_3(z+p)\right]^2 dz. \qquad (У.4.6)$$

Последний интеграл есть неотрицательная квадратичная форма от переменных $a_{l-1-3}^{(h)}, a_{l-3}^{(h)}, \ldots, a_{l-1}^{(h)}$. Докажем, что эта форма положительно определенная. Пусть она равна нулю. Тогда при $0 \le z \le 1$

$$\sum_{p=0}^{3} a_{l-1-p-3}^{(h)} \Pi_3(z+p) \equiv 0.$$

Надо доказать, что коэффициенты $a_p := a_{l-1-p-3}^{(h)}$ равны нулю, т.е. что функции $\Pi_3(z+p)$, $p = 0, 1, 2, \ldots, 3$ линейно независимы. Это утверждение очевидно справедливо при $3 = 0$ и $3 = 1$. Допустим, что оно справедливо для некоторого индекса $3 - 1$ и д

кажем, что оно справедливо и для индекса δ . Итак, пусть

$$\sum_{\rho=0}^{\delta} d_\rho \Pi_\delta(z+\rho)=0\,,\quad 0\le z\le 1\,.$$

По определению функций Π_δ

$$0=\sum_{\rho=0}^{\delta} d_\rho \int_{-\infty}^{+\infty} \Pi_{\delta-1}(y)\Pi_1(z+\rho-y)\,dy=\sum_{\rho=0}^{\delta} d_\rho \int_{z+\rho-1}^{z+\rho} \Pi_{\delta-1}(y)\,dy\,.$$

Продифференцируем по z :

$$\sum_{\rho=-1}^{\delta} \beta_\rho \Pi_{\delta-1}(z+\rho)=0\,,$$

где $\beta_{-1}=-d_0$; $\beta_\delta=d_\delta$; $\beta_\rho=d_\rho-d_{\rho+1}$, $0\le\rho\le\delta-1$.

Так как z меняется в пределах $(0,\ 1)$, то $\Pi_{\delta-1}(z-1)=0$, $\Pi_{\delta-1}(z+\delta)=0$, и члены с индексами $\rho=-1$ и $\rho=\delta$ исчезают. Мы приходим к тождеству

$$\sum_{\rho=0}^{\delta-1} \beta_\rho \Pi_{\delta-1}(z+\rho)=0\,.$$

По индукционному предположению, $\beta_\rho=0$, $0\le\rho\le\delta-1$, или $d_0=d_1=\ldots=d_\delta:=d$. Теперь

$$d\sum_{\rho=0}^{\delta} \Pi_\delta(z+\rho)\equiv 0\,;$$

так как $\Pi_\delta(z+\rho)\not\equiv 0$ и $\Pi_\delta(z+\rho)\ge 0$, то $d=0$, и наше утверждение доказано. Из этого утверждения вытекает существование положительных и не зависящих от ℓ и h величин C_1 и C_2 , таких, что справедливы неравенства

$$\frac{C_1^2}{\delta+1}\sum_{j=\ell-\delta-1}^{\ell-1}(a_j^{(h)})^2\le\int_0^1\Big[\sum_{j=\ell-\delta-1}^{\ell-1}a_j^{(h)}\Pi_\delta\Big(\frac{t}{h}-j\Big)\Big]^2dt\le\frac{C_2^2}{\delta+1}\sum_{j=\ell-\delta-1}^{\ell-1}(a_j^{(h)})^2\,,$$

или, если воспользоваться формулой (У.4.6),

$$\frac{C_1^2 h}{\delta+1}\sum_{j=\ell-\delta-1}^{\ell-1}(a_j^{(h)})\le\int_{(\ell-1)h}^{\ell h}[x^{(h)}(t)]^2\,dt\le\frac{C_2^2 h}{\delta+1}\sum_{j=\ell-\delta-1}^{\ell-1}(a_j^{(h)})^2\,.$$

Просуммировав это по ℓ , получим искомое неравенство

$$C_1\sqrt{h}\,\|a^{(h)}\|_{R_N}\le\|x^{(h)}\|_2\le C_2\sqrt{h}\,\|a^{(h)}\|_{R_N}\,. \qquad (У.4.7)$$

3°. Введем в рассмотрение N-мерные гильбертовы пространства X_N и Y_N с нормами

$$\forall a\in R_N\,,\ \|a\|_{X_N}=\sqrt{h}\,\|a\|_{R_N}\,,\ \|a\|_{Y_N}=\sqrt{h^{-1}}\,\|a\|_{R_N}\,.$$

зти пространства – сопряженные относительно скалярного умнож в R_γ.

Обозначим через A_h оператор, действующий из $\bar{X}_\gamma$ в $\bar{Y}_\gamma$ и рожденный матрицей M_h. Будем рассматривать $a^{(h)}$ и $f^{(h)}$ как эл менты пространств $\bar{X}_h$ и $\bar{Y}_h$ соответственно. Тогда уравнению (У.4.5) можно придать вид

$$A_h a^{(h)} = f^{(h)}. \qquad (У.4.8$$

Теорема У.4.I. Вычислительный процесс (У.4.8) устойчив в смысле второго определения в последовательности пар простран $(\bar{X}_\gamma, \bar{Y}_\gamma)$.

Доказательство этой теоремы дословно совпадает с доказат ством теоремы I § 2 главы IX книги автора [22], и мы его зде приводить не будем.

Обозначим через $H^{(h)}$ подпространство энергетического прос ранства задачи настоящего параграфа, образованное функциями да

$$x^{(h)}(t) = \sum_{j=-3}^{n(3+1)-1} d_j \pi_3\left(\frac{t}{h} - j\right) \qquad (У.4.9$$

с произвольными коэффициентами d_j. Оператор (У.4.9), перевод вектор $d = (d_{-3}, d_{-3+1}, \ldots, d_{n(3+1)-1})$ в функцию $x^{(h)} \in H^{(h)}$, обоз чим через Π_h. Приближ ое решение $x^{(h)}(t)$ можно определить из внения

$$A_h \Pi_h^{-1} x^{(h)} = f^{(h)}. \qquad (У.4.I$$

Теорема У.4.2. Вычислительный процесс (У.4.I0) устойчив смысле второго определения в последовательности пар простран $(H^{(h)}, R_\gamma)$.

Эту теорему мы здесь также не будем доказывать: ее доказа тельство совпадает с доказательством теоремы 2 § 2 главы IX книги автора [22].

Так же, в цитирова ной здесь книге [22], доказывается, чт число обусловленности матрицы M_h (уравнение (У.4.5)) имеет порядок $O(h^{-2})$.

§ 5. Симплициальные сетки; оценка аппроксимации.

I°. Симплициальные сетки имеют то важное преимущество пер кубическими, что они, вообще говоря, позволяют лучше аппрокси мировать произвольно заданную область. Из большого количества работ, посвященных методу конечных элементов с симплициальным

етками, мы отметим здесь лишь немногие из тех, в которых изуча-
тся погрешность аппроксимации или искажения; укажем только, .
то симплициальные сетки систематически используются в книгах
танесяна и Руховца [I], Стренга и Фикса [2] и Сьярле [I].
раткий обзор работ по МКЭ на симплициальных сетках за время
о 1973 г. дан в статье Корнеева [I].

Зламал [I] построил приближенное решение двумерного эллип-
тического уравнения четвертого порядка в виде функции, которая
на каждом треугольнике сетки представляет собой полином 5-й
тепени. Метод и результаты этой работы были обобщены Женишеком
[I], который рассмотрел кусочно полиномиальные функции степени
$\delta + I$ и обосновал метод для значений $\delta = I, 2$.

Дальнейшее обобщение дано в работе Брэмбла и Зламала [I], в
которой рассматриваются эллиптические уравнения любого порядка
в двумерной области. В каждом из треугольников сетки выделяется
совокупность узловых точек, одна из которых лежит в центре тя-
жести треугольника, а остальные выбираются в вершинах и, по оп-
ределенному правилу, на сторонах треугольника. В узлах задаются
значения полинома и некоторых его производных; максимальные по-
рядки этих производных определяются так, чтобы число условий ра-
вилось $(2\delta + I)(4\delta + 3)$ – числу коэффициентов полинома степени
$\delta + I$ от двух переменных. Доказываются следующие утверждения:

I) Существует единственный полином $P_\delta(t_1, t_2)$ степени
$4\delta + I$, удовлетворяющий упомянутым выше условиям. Заметим, что
в работе Корнеева и Пономарева [I] дан явный вид полиномов
$P_\delta(t_1, t_2)$.

2) Пусть Ω – замкнутая многоугольная область, разбитая на
замкнутые же треугольники T_i. Допустим, что узлы, расположен-
ные на сторонах или в вершинах треугольников, выбраны одинако-
выми для всех тех треугольников, которым рассматриваемый узел
принадлежит. Тогда функция $\psi_\delta(t_1, t_2)$, которая на каждом из
треугольников T_i совпадает с соответствующим полиномом $P_\delta(t_1, t_2)$
принадлежит при подходящем выборе δ к классу $C^{(m)}$, где m – по-
ловина порядка данного дифференциального уравнения.

3) Пусть дана функция $x \in W_2^{(k)}(T)$, где $2\delta+2 \le k \le 4\delta+2$, и
T – некоторый треугольник с выбранными на нем узлами. Зададим
на этих узлах значения функции x и ее соответствующих производ-
ных и построим по этим данным полином $P_\delta(t_1, t_2)$ так, как это

указано выше. Тогда справедлива оценка

$$\|x-p_s\|_{(\ell)} \leq \frac{Ch^{k-\ell}}{(\sin\alpha)^{s+\ell}}\|x\|_{(k)}, \quad \ell \leq k-1. \qquad (У.5.1$$

Здесь α – наименьший угол и h – наибольшая сторона треугольника T. Оценка (У.5.I) установлена для всех $k \geq 2$, кроме $k=$ для $k = 3$ оценка (У.5.I) получена Чистовым [I].

2°. В своих статьях [2,3] Женишек рассматривает тетраэдральную сетку в трехмерном пространстве. По аналогии с результатами, полученными для треугольной сетки на плоскости, Женишек делает следующее предположение: простейший полином в m-мерном симплексе, который позволяет строить кусочно полиномиальные раз непрерывно дифференцируемые функции, имеет степень $n = 5$. В цитированных работах Женишека излагаются результаты, полученные им для $m = 3$ и $s = I$, $s = 2$; степени соответствующих полиномов суть 9 и I7. Построение этих полиномов требует соответственно выполнения 220 и II40 условий. Хотя автор и приводит связанные с этими полиномами оценки аппроксимации, однако, как он замечает сам его полиномы практически не применимы.

3°. В статье Бергера и Маркса [I] рассмотрена аппроксимация функций класса $C^{(2)}(\bar{\Omega})$ кусочно полиномиальными функциями 5-й степени на сетке из конгруэнтных прямоугольных треугольников. Область Ω предполагается многоугольной; в каждом из треугольников сетки интерполирующий полином 5-й степени определяется значениями интерполируемой функции и ее первых производных в вершинах треугольника. Погрешность аппроксимации в $C(\bar{\Omega})$ и $C^{(1)}(\bar{\Omega})$ равна соответственно $O(h^2)$ и $O(h)$.

4°. В работе Чистова [I] рассматривается задача аппроксимации по МКЭ функции $x \in W_p^{(3)}(\Omega)$ (Ω – конечная область в R_2 на сетке, образованной неналегающими правильными треугольниками с длиной стороны h. На такие сетки распространено определение исходных функций, а также способ построения координатных функций из исходных; носители координатных функций суть правильные шестиугольники с центрами в вершинах треугольников сетки. Выведены фундаментальные соотношения, аналогичные соотношениям гл. III и дающие необходимые и достаточные условия полноты системы координатных функций в $W_p^{(3)}(\Omega)$. Для $s = 2$ построена система кусочно полиномиальных исходных функций 5-й степени. Эти функции зависят от существенно меньшего числа параметров, чем упомяну-

ные в п. I° полиномы Женишека, поэтому они приводят к с тебрическим системам МКЭ более низкого порядка.

§ 6. Теорема об устойчивости.

I°. Устойчивость МКЭ на симплициальных сетках исследована в работе Чистова [2] для первой краевой задачи. Рассмотрим формально самосопряженное эллиптическое уравнение

$$\sum_{|\alpha|,|\beta|=0}^{s} (-1)^{\alpha} \mathcal{D}^{\alpha}(A_{\alpha\beta}\mathcal{D}^{\beta}x) = f(t),\ f \in L_2(\Omega) \qquad (У.6.1)$$

при краевом условии

$$\mathcal{D}^{\gamma}x\big|_{\partial\Omega} = 0,\quad 0 \le |\gamma| \le s-1. \qquad (У.6.2)$$

Будем считать, что оператор задачи (У. .1) – (У.6.2) положительно определенный в $L_2(\Omega)$ и что Ω – конечная область в R_m.

Пусть Ω^h – многогранник, $\Omega^h \subset \bar{\Omega}$, и пусть в Ω^h выбрано конечное множество t_k узлов, занумерованных в произвольном порядке, $k = 1, 2, \ldots, L$. Разобьем Ω^h на симплексы S_τ, $\tau = 1, 2, \ldots, \tau_0$ с вершинами в узлах t_k. Потребуем, чтобы S_τ удовлетворяли "условиям совместимости" (см. Оден [1], п. 7.3). Обозначим через $l_{\tau i k}$ длины ребер симплекса S_τ, сходящихся к вершине с номером k. Пусть $P_{\tau k}$ – параллелепипед, построенный на этих ребрах, и $\rho_{\tau k}$ – параллелепипед с ребрами единичной длины, построенный на тех же направлениях.

Если Q множество в R_m, и $|Q|$ – объем (мера) этого множества, то, как известно,

$$|S_\tau| = \frac{|P_{\tau k}|}{m!} = \frac{|\rho_{\tau k}|}{m!}\prod_{i=1}^{m} l_{\tau k i}. \qquad (У.6.3)$$

Введем обозначения: $h_1 = \max h_{\tau k i}$, $h_2 = \min h_{\tau k i}$, $\vartheta = \min |\rho_\tau|$. Допустим, что

$$h_1/h_2 \le C_1,\ \vartheta \ge C_2;\ C_1, C_2 = const. \qquad (У.6.4)$$

С каждым узлом t_k свяжем координатные функции $\Phi_{ik}(t)$; здесь i – мультииндекс размерности m, пробегающий множество $0 \le |i| \le \tilde{s}$, причем число $\tilde{s}$ одинаково для всех k. Будем считать, что функции Φ_{ik} обладают следующими свойствами:

а) $\Phi_{ik} \subset C^{(\tilde{s}-1)}(R_m) \cap W_2^{(\tilde{s})}(R_m)$;

б) $supp\,\Phi_{ik} \subset Z_k$, где Z_k – многогранник, образован-

ный симплексами с общей вершиной t_k ;

в) $\Phi_{ik}^{(h)}(t_k)=\delta_{di}$, $0\le|d|$, $|i|\le\tilde{s}$;

г) Существует такая последовательность $h_n\to 0$, что последовательность соответствующих конечных систем координатных функций $\{\Phi_{ik}\}$ полна в энергетическом пространстве H_A нашей задачи.

д) Пусть S_0 - симплекс с вершинами (0,0,...,0), (I,0,..., 0),..., (0,...,0,I). Существует такое афинное преобразование симплекса S_0 на S_τ: $t=\omega_\tau(\tau)$, что если $\vartheta_{ik\tau}(\tau)=\Phi_{ik}(\omega_\tau(\tau))$, то наименьшее собственное число матрицы скалярных произведений $\left(\vartheta_{ik\tau}\,,\,\vartheta_{i'k'\tau}\right)_{L_2(S_\tau)}$ ограничено снизу положительным числом λ_0, которое может зависеть только от области Ω и от постоянных $C_1,..C_n$ из неравенств (У.6.4).

Из условия д) вытекает, что процесс МКЭ сходится при $h_n\to 0$ в норме H_A . Тем более он сходится в более слабой норме $L_2(\Omega)$, поэтому, если x_* есть точное решение, а

$$x_*^h=\sum_{|i|=0}^{\tilde{s}}\sum_{k=1}^{L}a_{ik}^{(h)}\Phi_{ik}(t)$$

есть приближенное решение задачи (У.6.I-2), то существует такая постоянная C_0 , что

$$\|x_*^{(h)}\|_{2,\Omega}\le C_0\|x_*\|_{2,\Omega}\,. \qquad (У.6.5)$$

Рассмотрим подпространство $H_A^{(h)}$ энергетического пространства H_A, образованное функциями вида

$$x^{(h)}(t)=\sum_{|i|=0}^{\tilde{s}}\sum_{k=1}^{L}d_{ik}\Phi_{ik}(t). \qquad (У.6.6)$$

Оценим норму вектора d с составляющими d_{ik} через норму $\|x^{(h)}\|$:

$$\|x^{(h)}\|_{2,\Omega}^2=\sum_{\tau=1}^{\tau_0}\int_{S_\tau}\Big[\sum_{|i|=0}^{\tilde{s}}\sum_{k=1}^{L}d_{ik}\Phi_{ik}(t)\Big]^2dt= \qquad (У.6.7)$$

$$=\int_{S_0}\sum_{\tau=1}^{\tau_0}\Big[\sum_{|i|=0}^{\tilde{s}}\sum_{k=1}^{L}d_{ik}\vartheta_{ik\tau}(\tau)\Big]^2J_\tau\,d\tau\,.$$

Здесь J_τ - якобиан преобразования симплекса S_τ в симплекс S_0; как хорошо известно $|J_\tau|=\Omega^m|S_\tau|$. Далее, из соотношений (У.6.3-4) следует

$$\|x^{(h)}\|_{2,\Omega^h} \geq \frac{2^m \vartheta}{m!} h_2^m (M_h d, d)_{R_N} \geq \frac{2^m \vartheta \lambda_0}{m!} h_2^m \|d\|^2_{\dots N}; \quad \text{(У.6.8)}$$

здесь N – число составляющих вектора d. Так как $\|x^{(h)}\|_{2,\Omega^h} \leq$ $\leq \|x^{(h)}\|_{2,\Omega}$, то отсюда и из (У.6.5) следует, что величина $h_2^m \|a^{(h)}\|^2_{R_N}$ ограничена. Обозначим теперь через $\bar{X}_N$ и $\bar{Y}_N$ гильбертовы N-мерные пространства с нормами

$$\|\cdot\|_{\bar{X}_N} = h^{m/2} \|\cdot\|_{R_N}, \quad \|\cdot\|_{\bar{Y}_N} = h^{-m/2} \|\cdot\|_{R_N},$$

где h любое число из последовательности $\{h_n\}$.

На основании замечания III.3.4 можно утверждать, что процесс вычисления коэффициентов $a_{ik}^{(h_n)}$ устойчив относительно погрешностей искажения в последовательности $(\bar{X}_N, \bar{Y}_N)$, а процесс вычисления приближенного решения $x_*^{(h)}$ устойчив в последовательности $(A_A^{(h_n)}, \bar{Y}_N)$.

Аналогичные теоремы можно доказать при произвольных краевых условиях, оставляя их оператор (У.6.I) положительно определенным, если обычный функционал энергии заменить функционалом, который ввел Руховец [I].

ГЛАВА VI
Погрешности приближенного решения интегральных уравнений.

Проблема, формулированная в названии данной главы, по-видимому, пока еще исследована не вполне достаточно. Перечислим те результаты, которые кажутся нам более важными.

Канторович (см. Канторович и Крылов [I], гл.II) исследовал погрешность аппроксимации метода механических квадратур и метода замены ядра вырожденным для уравнений Фредгольма. Большое количество работ, посвященных приближенному решению Фредгольмовых уравнений, а также одномерных сингулярных уравнений, принадлежит Габдулхаеву (подробная библиография приведена в его книге [3]). Некоторые результаты по приближенному решению интегральных уравнений фредгольмовых и сингулярных одномерных, приведены в книге автора и Смолицкого [I], глава VI. В книге Гохберга и Фельдмана [I] с большой полнотой изложено применение проекционных методов к "уравнениям в свертках", т.е., к уравнениям, являющимся обобщением известных уравнений Винера – Хопфа. Некоторые новые подходы к приближенному решению уравнений Фредгольма указаны в книге Анселона [I]. Приближенному решению одномерных сингулярных интегральных уравнений посвящена книга Пресдорфа и Зильбермана [I]. Ряд результатов по приближенному решению сингулярных интегральных уравнений, одномерных и многомерных, содержится в книге автора – Пресдорфа [I]. Приближенному решению многомерных сингулярных интегральных уравнений посвящены работы автора и Радевой [I], автора [25]. Отметим еще сборники статей под редакцией Делвса и Уолша [I], Андерссена, де-Хуга, Лиаоса [I], посвященных приближенному решению интегральных уравнений, а также весьма интересную статью Пресдорфа и Шмидта [I], в которой получено необходимое и достаточное условие сходимости одного варианта метода коллокации, связанного с МКЭ, и дана оценка соответствующей погрешности аппроксимации. В работах автора [18-21,23] развит метод приближенного вычисления резольвенты Фредгольма, основанный на кусочно-многоэлементной аппроксимации ядра.

Для упрощения записи, а отчасти и выкладок, мы будем рассматривать ниже скалярные уравнения в R_1 и будем считать, что промежуток интегрирования есть отрезок $[0, 1]$. Распространение результатов на более общие уравнения Фредгольма очевидно. Исклю

чаниям, пожалуй, являются аппроксимация ядра по МКЭ в том случае, когда множество интегрирования не допускает взаимно-однозначного отображения на куб; этот случай рассмотрен в статье автора [20].

§ I. Применение метода механических квадратур к уравнениям Фредгольма

I°. Рассмотрим уравнение

$$x(t)-\int_0^1 \mathcal{K}(t,\tau)\,x(\tau)\,d\tau = f(t)\,. \qquad (\text{VI.I.I})$$

Допустим, что $f\in C^{(s)}[0,1]$, $\mathcal{K}\in C^{(s,s)}([0,1]\times[0,1])$ и что число I не характеристическое для ядра $\mathcal{K}(t,\tau)$. Условие на ядро надо понимать так: в квадрате $[0,1]\times[0,1]$ существуют и непрерывны все производные вида

$$\frac{\partial^{k+l}\mathcal{K}}{\partial t^k\,\partial\tau^l};\quad 0\le k,l\le s.$$

Из наших условий вытекает, что уравнение (VI.I.I) имеет одно и только одно решение $x_*\in C^{(s)}[0,1]$.

Будем приближенно решать уравнение (VI.I.I) по методу механических квадратур, заменяя интеграл по квадратурной формуле вида

$$\int_0^1 y(\tau)\,d\tau = \sum_{k=1}^{n} C_k^{(n)}\,y(\tau_k^{(n)})+\rho_n. \qquad (\text{VI.I.2})$$

Примем, что эта формула обладает следующими свойствами:

$$C_k^{(n)}\ge 0,\ \sum_{k=1}^{n} C_k^{(n)} = 1\,; \qquad (\text{VI.I.3})$$

если $y\in C^{(s)}[0,1]$, то

$$\rho_n \le C_s\,h^s\,\|y^{(s)}\|_C\,;\ h=\max_k(\tau_{k+1}^{(n)}-\tau_k^{(n)});\ C_s=\text{const}. \qquad (\text{VI.I.4})$$

Заменяя в (VI.I.I) интеграл по формуле (VI.I.2) и отбрасывая остаточный член, получим новое уравнение

$$y^{(n)}(t) = \sum_{k=1}^{n} C_k^{(n)}\,\mathcal{K}(t,\tau_k)\,y^{(n)}(\tau_k)+f(t).\quad \tau_k=\tau_k^{(n)} \qquad (\text{VI.I.5})$$

Решение $y_*^{(n)}(t)$ которого будем рассматривать как наиболее решение уравнения (VI.I.I). Чтобы определить $y_*^{(n)}(t)$, положим в (VI.I.I) $t=\tau_j$, $1\le j\le n$. Это даст нам алгебраическую систему

$$y_j^{(n)} - \sum_{k=1}^{n} c_k^{(n)} \mathcal{K}(\tau_j, \tau_k) y_k^{(n)} = f_j^{(n)}, \quad 1 \leqslant j \leqslant n, \qquad \text{(VI.I.6)}$$

где $y_j^{(n)} = y^{(n)}(\tau_j)$, $f_j^{(n)} = f(\tau_j)$. Если система (VI.I.5) разрешима единственным образом, то тем же свойством обладает уравнение (VI.I.5). Очевидно и обратное; очевидно также, что любое решение уравнения (VI.I.5) принадлежит к классу $C^{(s)}[0,$

Операторы

$$\int_0^1 \mathcal{K}(t,\tau)x(\tau)d\tau, \quad \sum_{k=1}^{n} c_k^{(n)} \mathcal{K}(t,\tau_k)x(\tau_k) \qquad \text{(VI.I.}$$

будем рассматривать как операторы в $C^{(s)}[0, 1]$; оценим норму их разности. Эта норма равна

$$\sup_{\|x\|=1} \sum_{j=0}^{s} \left\| \int_0^1 \frac{\partial^j \mathcal{K}(t,\tau)}{\partial t^j} x(\tau)d\tau - \sum_{k=1}^{n} c_k^{(n)} \frac{\partial^j \mathcal{K}(t,\tau_k)}{\partial t^j} x(\tau_k) \right\|_C .$$

По оценке (VI.I.4) это не превосходит величины

$$\sup_{\|x\|=1} Ch^s \sum_{j=0}^{s} \left\| \frac{\partial^s}{\partial \tau^s} \frac{\partial^j \mathcal{K}(t,\tau)}{\partial t^j} x(\tau)d\tau \right\|_C \leqslant Ch^s; \quad C = const.$$

Таким образом, при достаточно малом h разность операторов (VI.I.7) сколь угодно мала по норме $C^{(s)}$.

Интегральные операторы в (VI.I.7) обозначим соответственно через $\mathcal{K}$ и $\mathcal{K}_n$, а их разность через σ_n. Уравнение (VI.I.5) можно записать в виде

$$y^{(n)}(t) - (\mathcal{K}y^{(n)})(t) + (\sigma_n y^{(n)})(t) = f(t).$$

Отсюда

$$y^{(n)}(t) + \left((I-\mathcal{K})^{-1}\sigma_n y^{(n)}\right)(t) = (I-\mathcal{K})^{-1} f(t) = x_*(t)$$

и, следовательно,

$$\|x_* - y_*^{(n)}\|_{C^{(s)}} \leqslant \|(I-\mathcal{K})^{-1}\|_{C^{(s)}} \cdot \|\sigma_n\|_{C^{(s)}} \cdot \|y_*^{(n)}\|_{C^{(s)}}.$$

При достаточно малом h будет $\|(I-\mathcal{K})^{-1}\|_{C^{(s)}} \cdot \|\sigma_n\|_{C^{(s)}} \leqslant 1/2$ и

$$\|y_*^{(n)}\|_{C^{(s)}} - \|x_*\|_{C^{(s)}} \leqslant \frac{1}{2}\|y_*^{(n)}\|_{C^{(s)}}; \quad \|y_*^{(n)}\|_{C^{(s)}} \leqslant 2\|x_*\|_{C^{(s)}}.$$

Теперь,

$$\|x_* - y_*^{(n)}\| \leqslant Ch^s \|x_*\|_{C^{(s)}}, \quad C = const. \qquad \text{(VI.I.}$$

постоянную C нетрудно уточнить:

$$C = 2C_3 \|(I-\mathcal{K})^{-1}\|_{C(S)}. \qquad (\text{VI.I.9})$$

Формулы (VI.I.8) и (VI.I.9) дают оценку погрешности аппроксимации метода механических квадратур для уравнения Фредгольма.

2°. Рассмотрим случай, когда ядро и свободный член уравнения (VI.I.I) только непрерывны. Для произвольной функции $u(t)$ определим ее модуль непрерывности $\omega(u,h)$ и допустим, что $|u(t')-u(t'')| \leqslant M_u r(h)$, где $|t'-t''| \leqslant h$, $r(h) \xrightarrow[h\to 0]{} 0$, причем M_u зависит от u, но не от h. Наименьшее возможное значение M_u обозначим через $M(u)$:

$$\omega(u,h) \leqslant M(u) r(h). \qquad (\text{VI.I.IO})$$

Допустим, что модули непрерывности ядра $\mathcal{K}(t,\tau)$ и свободного члена $f(t)$ удовлетворяют неравенству (VI.I.IO). Тогда тому же неравенству удовлетворяет и решение уравнения (VI.I.I).

Введем пространство S функций, непрерывных на отрезке $[0,1]$ с модулем непрерывности, удовлетворяющим неравенству (VI.I.II) и с нормой

$$\|\cdot\|_S = \|\cdot\|_C + M(\cdot). \qquad (\text{VI.I.II})$$

Очевидно, $f, x_k \in S$. Докажем, что операторы (VI.I.7) ограничены в S. Обозначим $(\mathcal{K}x)(t) = y(t)$. Имеем $\|y\|_C \leqslant \|\mathcal{K}\|_C \cdot \|x\|_C$. Далее,

$$\omega(y,h) \leqslant \omega(\mathcal{K},h)\|x\|_C \leqslant M(\mathcal{K}) r(h) \|x\|_C. \qquad (\text{VI.I.I2})$$

Отсюда $M(y) \leqslant M(\mathcal{K})\|x\|_C$ и

$$\|y\|_S \leqslant [\|\mathcal{K}\|_C + M(\mathcal{K})]\cdot\|x\|_C \leqslant [\|\mathcal{K}\|_C + M(\mathcal{K})]\cdot\|x\|_S. \qquad (\text{VI.I.I3})$$

Таким образом, $\|\mathcal{K}\|_{S\to S} \leqslant \|\mathcal{K}\|_C + M(\mathcal{K})$. Теперь рассмотрим оператор $\mathcal{K}_n$. Обозначим $\psi_n(t) = (\mathcal{K}_n x)(t)$. Тогда

$$|\psi_n(t)| \leqslant \sum_{k=1}^{n} C_k^{(n)} \|\mathcal{K}\|_C \cdot \|x\|_C = \|\mathcal{K}\|_C \cdot \|x\|_C;$$

$$\omega(\psi_n,h) \leqslant \sum_{k=1}^{n} C_k^{(n)} \omega(\mathcal{K},h)\|x\|_C = \omega(\mathcal{K},h)\|x\|_C \leqslant M(\mathcal{K})\|x\|_C r(h).$$

Отсюда $M(\psi_n) \leqslant M(\mathcal{K})\|x\|_C$ и $\|\mathcal{K}_n\|_{S\to S} \leqslant \|\mathcal{K}\|_C + M(\mathcal{K})$.

Допустим, что для функций $y(t)$, удовлетворяющих условию

(VI.I.I0), остаточный член квадратурной формулы (VI.I.4) имеет оценку

$$\rho_n \le C_0 r(h)[\|y\|_C + M y];\ C_0 = const,\ r(h) \xrightarrow[h\to 0]{} 0 . \qquad (VI.1.14)$$

Дополнительно предположим, что

$$M_t(\mathcal{K}) := \frac{1}{r(h)} \sup_{|t''-t'|\le h} |\mathcal{K}(t'',\tau) - \mathcal{K}(t',\tau)| \le C_1 ;$$

$$M_{t,\tau}(\mathcal{K}) := \frac{1}{r^2(h)} \sup_{\substack{|t''-t'|\le h \\ |\tau''-\tau'|\le h}} |\mathcal{K}(t'',\tau'') - \mathcal{K}(t'',\tau') -$$

$$- \mathcal{K}(t',\tau'') + \mathcal{K}(t',\tau')| \le C_1 ;\ C_1 = const .$$

Оценим $\|\sigma_n\|_{S\to S}$. Имеем $\|\sigma_n x\|_S = \|\sigma_n x\|_C + M(\sigma_n x)$. По оценке (VI.I.I4)

$$\|\sigma_n x\|_C = \|\mathcal{K}x - \mathcal{K}_n x\|_C \le C_0 r(h)[\|\mathcal{K}\|_C \cdot \|x\|_C + M(\mathcal{K}x) \le$$

$$\le 2C_0 \|\mathcal{K}\|_C \cdot \|x\|_C r(h) \le C_2 \|x\|_C r(h);\ C_2 = const . \qquad (VI.1.15)$$

Далее, по той же оценке (VI.I.I4)

$$\omega(\sigma_n x, h) = \sup_{|t''-t|\le h} \Big| \int_0^1 [\mathcal{K}(t'',\tau) - \mathcal{K}(t',\tau)] x(\tau) d\tau -$$

$$- \sum_{k=1}^{n} C_k^{(n)} [\mathcal{K}(t'',\tau_k) - \mathcal{K}(t',\tau_k)] x(\tau_k) \Big| \le$$

$$\le C_0 r(h)[\max_{\tau} |\mathcal{K}(t'',\tau) - \mathcal{K}(t',\tau)| \cdot \|x\|_C +$$

$$+ M_\tau(\mathcal{K}(t'',\tau) - \mathcal{K}(t',\tau))\|x\|_C] \le 2C_0 C_1 r^2(h)\|x\|_C .$$

Отсюда $M(\sigma_n x) \le 2C_0 C_1 r(h)\|x\|_C$ и $\|\sigma_n x\|_S \le C r(h)\|x\|_C \le C r(h)\|x\|_S$, $C = const$. Таким образом, $\|\sigma_n\|_{S\to S} \le C r(h)$ и, при h достаточно малом, будет $\|(I-\mathcal{K})\|_{S\to S}\|\sigma_n\|_{S\to S} \le \frac{1}{2}$. Поступая далее, как в конце п.I°, получим

$$\|x_n - y_n^{(n)}\| \le C r(h),\ C = const. \qquad (VI.1.16)$$

3°. Рассмотрим погрешность искажения для уравнений Фредгольма. Для определенности остановимся на случае п.I°. Обозначим

через A_n матрицу системы (VI.1.6): $A_n = I_n - \mathcal{K}^{(n)}$, где $\mathcal{K}^{(n)}$ – матрица элементов $C_k^{(n)} \mathcal{K}(\tau_j, \tau_k)$, I_n – единичная матрица порядка $n \times n$. Будем рассматривать $\bar{y}^{(n)} = (y_1^{(n)}, y_2^{(n)}, \dots, y_n^{(n)})$ и $\bar{f}^{(n)} = (f_1^{(n)}, f_2^{(n)}, \dots, f_n^{(n)})$ как векторы в пространстве Y_n (см. § 9 гл. II). Пусть $\delta_n^{(1)}$ – наименьшее сингулярное число матрицы A_n. Так как $y_k^{(n)}$ суть значения $y_*^{(n)}$ – решения уравнения (VI.1.5) в точках τ_k, а $y_*^{(n)} \xrightarrow{C} x_*$ и тем более $y_*^{(n)} \xrightarrow{L_2} x_*$, то нормы $\|\bar{y}^{(n)}\|_{Y_n}$ ограничены в совокупности и по следствию II.2.2 для устойчивости процесса (VI.1.6) в последовательности (Y_n, Y_n) необходимо и достаточно, чтобы $\|A_n^{-1}\|_{Y_n \to Y_n} \le C = const$.

Но

$$\|A_n^{-1}\|_{Y_n \to Y_n} = \|A_n^{-1}\|_{R_n \to R_n}$$

и для устойчивости процесса (VI.1.6) в упомянутом здесь смысле необходимо и достаточно, чтобы $\delta_n^{(1)}$ было положительно ограничено снизу. Легко убедиться, что это же условие необходимо и достаточно для устойчивости процесса вычисления приближенного решения в соответствующей последовательности пар подпространств.

4°. В общем случае можно оценить погрешности искажения и округления для метода механических квадратур, используя соответствующие оценки для линейных алгебраических систем (см. § 4 гл. I).

§ 2. Уравнения Фредгольма, разрешимые методом итераций

1°. Пусть в уравнении (VI.1.1) ядро удовлетворяет неравенству

$$\max_{0 \le t, \tau \le 1} |\mathcal{K}(t, \tau)| = q < 1. \qquad (VI.2.1)$$

Тогда решение этого уравнения можно построить по методу простой итерации:

$$x_* = f(t) + \sum_{k=1}^{\infty} \int_0^1 \mathcal{K}_n(t, \tau) f(\tau) d\tau; \qquad (VI.2.2)$$

$\mathcal{K}_n$ – итерированные ядра. Ряд (VI.2.2) сходится в $C[0, 1]$ как прогрессия со знаменателем q, точнее, n-й член этого ряда не превосходит величины $q^n \int_0^1 |f(\tau)| d\tau$. Удобнее производить вычисления не по формуле (VI.2.2), а по обычной схеме итераций:

$$x_0(t) = f(t); \quad n \ge 1, \quad x_n(t) = f(t) + \int_0^1 \mathcal{K}(t, \tau) x_{n-1}(\tau) d\tau. \qquad (VI.2.3)$$

Однако, если ядро $\mathcal{K}(t,\tau)$ имеет более или менее сложную структуру, то вычисление интегралов, входящих в формулу (УI.2.3), становиться затруднительным. Поэтому мы прибегнем к следующему приему (см. работы автора [I8, 22], где этот прием использован в более общей ситуации).

Введем функцию

$$\omega(t)=\begin{cases} t, & 0\le t\le 1,\\ 2-t, & 1<t\le 2,\\ 0, & t\notin[0,2]\end{cases} \qquad (\text{УI.2.4})$$

и обозначим

$$\mathcal{K}^h(t,\tau)=\sum_{j,k=-1}^{2n-1}\mathcal{K}(t_{j+1},t_{k+1})\omega\left(\frac{t}{h}-j\right)\omega\left(\frac{\tau}{h}-k\right). \qquad (\text{УI.2.5})$$

Здесь n – произвольно выбранное натуральное число, $h=1/2n$, $t_j=(j+1)h$. Если $\mathcal{K}\in C^2(Q)$, $Q=[0,1]\times[0,1]$, то (см. работу автора [22], гл.III, § I)

$$\|\mathcal{K}-\mathcal{K}^h\|_{C(Q)}\le C\max_{|\alpha|=2}\|\mathcal{D}^\alpha\mathcal{K}\|_{C(Q)}h^2.$$

Постоянную C в последнем неравенстве можно вычислить по формуле (9.I4) статьи автора [24]; см. также книгу автора [3] Кар.IУ, § 9). Заметим, что в нашем случае следует в упомянутой формуле (9.I4) положить $\delta=1$, а тогда легко получается $C=1$, так что

$$\|\mathcal{K}-\mathcal{K}^h\|_{C(Q)}\le 10\max_{|\alpha|=2}\|\mathcal{D}^\alpha\mathcal{K}\|_{C(Q)}h^2. \qquad (\text{УI.2.6})$$

Возьмем n столь большим, чтобы правая часть оценки (УI.2.6) была меньше, чем $1-q_1$. Тогда, очевидно, $\|\mathcal{K}^h\|_{C(Q)}:=q_1<1$. Заменим в уравнении (УI.I.I) ядро $\mathcal{K}$ на $\mathcal{K}^h$, а свободный член $f(t)$ – его конечноэлементной аппроксимацией

$$f^h(t)=\sum_{j=-1}^{2n-1}f(t_{j+1})\omega\left(\frac{t}{h}-j\right).$$

Разность $f(t)-f^h(t)$ есть разность ординат кривой $y=f(t)$ и вписанной в нее ломаной $y=f^h(t)$. Если $f\in C^{(2)}[0,1]$, то эту разность можно вычислить по формуле

$$t_{j+1}\le t\le t_{j+2},\quad f(t)-f^h(t)=-\frac{1}{h}\int_{t_{j+1}}^{t_{j+2}}G(t-t_{j+1},\tau-t_{j+1})f''(\tau)\,d\tau,$$

где

$$G(t,\tau)=\begin{cases}t(h-\tau), & 0\le t\le\tau,\\ \tau(h-t), & 0\le\tau\le t.\end{cases}$$

Отсюда

$$\|f-f^h\|_C \le h^2\|f''\|_C. \qquad (\text{VI}.2.7)$$

2°. Оценим погрешность аппроксимации от такой замены. Новое интегральное уравнение имеет вид

$$x^{(h)}(t)-\int_0^1 \mathcal{K}^h(t,\tau)x^h(\tau)d\tau=f^{(h)}(t). \qquad (\text{VI}.2.8)$$

Сравнивая это с общей схемой п.4° § 4 гл.I, видим, что в данном случае $\Gamma=\mathcal{K}-\mathcal{K}^h$, где $\mathcal{K}$ и $\mathcal{K}^h$ – интегральные операторы с ядрами $\mathcal{K}(t,\tau)$ и $\mathcal{K}^h(t,\tau)$ соответственно.

По формуле (I.4.14)

$$\|x-x^{(h)}\|\le\frac{\|(I-\mathcal{K})^{-1}\|}{1-\|(I-\mathcal{K})^{-1}\|\cdot\|\mathcal{K}-\mathcal{K}^h\|}\left[\|\mathcal{K}-\mathcal{K}^h\|\cdot\|x\|+\|\delta\|\right];$$

все нормы взяты в $C[0,1]$ или, соответственно, в $C([0,1]\times[0,1])$. Используя неравенства (VI.2.6) и (VI.2.7), а также неравенство $\|(I-\mathcal{K})^{-1}\|\le(1-q)^{-1}$, получим оценку погрешности аппроксимации:

$$\|x-x^{(h)}\|\le$$

$$\le\left[1-10q\max_{|\alpha|=2}\|\mathcal{D}^\alpha\mathcal{K}\|h^2\right]^{-1}\left[10\max_{|\alpha|=2}\|\mathcal{D}^\alpha\mathcal{K}\|\cdot\|x\|+\|f''\|\right]h^2; \qquad (\text{VI}.2.9)$$

можно еще воспользоваться тем, что

$$\|x\|\le\frac{\|f\|}{1-q}. \qquad (\text{VI}.2.10)$$

3°. Решим уравнение (VI.2.8) по методу итераций и оценим погрешность алгоритма для этого уравнения. В данном случае итерации сходятся как прогрессия со знаменателем q_1. Если выполнить точно N итераций, то погрешность алгоритма будет иметь оценку

$$\|x^{(h)}-x_N^{(h)}\|\le$$

$$\le\sum_{k=N+1}^{\infty}\left\|\int_0^1\mathcal{K}_k^h(t,\tau)f^h(\tau)\,d\tau\right\|\le\left[\|f\|+h^2\|f''\|\right]\frac{q_1^{N+1}}{1-q_1}. \qquad (\text{VI}.2.11)$$

4°. Допустим, что числа $\mathcal{K}(t_j, t_k)$ и $f(t_j)$ вычислены с погрешностей, тогда мы имеем точную формулу

$$x_k^{(h)}(t) = f^{(h)}(t) + \int_0^1 \mathcal{K}^h(t,\tau)\, x_{k-1}^{(h)}(\tau)\, d\tau. \qquad (\text{VI.2.12})$$

Нетрудно видеть, что

$$x_{k-1}^{(h)}(t) = \sum_{j=-1}^{2n-1} a_{j,k-1}^{(h)}\, \omega\left(\frac{t}{h} - j\right).$$

Подставив это, а также значения $f^{(h)}(t)$ и $\mathcal{K}^h(t,\tau)$, в (VI.2.12) и приравнивая коэффициенты при $\omega(\frac{t}{h} - j)$, получим

$$a_{jk}^{(h)} = f(t_{j+1}) + \sum_{l,m=-1}^{2n-1} a_{m,k-1}\, \mathcal{K}(t_{j+1}, t_{l+1}) \int_0^1 \omega\left(\frac{\tau}{h} - l\right) \omega\left(\frac{\tau}{h} - m\right) d\tau.$$

Интеграл в последнем тождестве просто вычисляется: обозначив через hw_{lm}, имеем $w_{lm} = w_{ml}$ и

$$w_{lm} = \begin{cases} 0, & |m-l| \geqslant 2 \\ 1/6, & m-l = \pm 1 \\ 2/3, & m = l,\ 0 \leqslant m \leqslant 2n-2 \\ 1/3, & m = l = -1;\ m = l = 2n-1. \end{cases} \qquad (\text{VI.2.1})$$

В результате мы приходим к системе рекуррентных соотношений для коэффициентов $a_{jk}^{(h)}$:

$$a_{jk}^{(h)} = f(t_{j+1}) + h \sum_{l,m=-1}^{2n-1} a_{m,k-1}^{(h)}\, \mathcal{K}(t_{j+1}, t_{l+1})\, w_{lm}; \qquad (\text{VI.2.1})$$

систему (VI.2.14) надо решать при начальном условии $a_{j0}^{h} = f(t$

Если мы введем в пространстве $2n$-компонентных векторов норму по формуле $\|a\| = \max |a_i|$, то норма оператора в правой части равенства (VI.2.14) равна

$$B = h \max_{l} \sum_{j,m=-1}^{2n-1} |\mathcal{K}(t_{j+1}, t_{l+1})|\, w_{lm}. \qquad (\text{VI.2.1})$$

Если же принять значения чисел w_{lm} (формула (VI.2.13)), то найдем, что

$$\sum_{j,m=-1}^{2n-1} w_{lm} |\mathcal{K}(t_{j+1}, t_{l+1})| \leqslant \frac{1}{h} \max |\mathcal{K}(t,\tau)| = q \cdot 2n.$$

Отсюда $B \le h \cdot 2nq = q$, и при фиксированном индексе j итерационный процесс (VI.2.14) сходится. Если ограничиться N итерациями, то погрешность алгоритма для процесса (VI.2.14) оценивается величиной

$$\frac{q^{N+1}}{1-q}\|f(t)\|_C. \qquad (VI.2.16)$$

5°. В нашей задаче можно не выделять погрешность искажения и считать, что остальные погрешности связаны с округлением. Погрешность округления легко оценить по способу § 3 гл. IV; так как итерационный процесс (VI.2.14) сходится, как геометрическая прогрессия, то при бесконечном возрастании N погрешность округления остается ограниченной.

§ 3. Погрешность резольвенты Фредгольма

1°. Результаты настоящего параграфа содержатся в статье автора [18], ряд обобщений – в его статьях [19-21]; материал этих статей изложен в книге автора [22]. Ввиду этого мы здесь ограничимся только постановкой задачи и формулировкой результатов, и притом для простейшего класса интегральных уравнений и простейшей аппроксимации ядра. Подробное доказательство мы проведем только для оценки погрешности алгоритма, которая уточняет соответствующую оценку из [18] и [22].

2°. Рассмотрим интегральное уравнение Фредгольма

$$x(t) - \lambda \int_0^1 K(t,\tau)x(\tau)\,d\tau = f(t) \qquad (VI.3.1)$$

с ядром $K \in C^{(2)}$; λ – произвольный, вообще говоря, комплексный параметр. Требование на ядро можно ослабить – достаточно считать его непрерывным, но при этом некоторые последующие оценки окажутся хуже. Будем считать еще, что $|K(t,\tau)| < 1$ – этого всегда можно добиться подходящим изменением параметра.

Напомним основные факты, относящиеся к резольвенте Фредгольма.

1. Резольвента Фредгольма $\Gamma(t,\tau;\lambda)$ для ядра $K(t,\tau)$ имеет вид

$$\Gamma(t,\tau;\lambda) = \frac{\mathcal{D}(t,\tau;\lambda)}{\mathcal{D}(\lambda)}, \qquad (VI.3.2)$$

где "определитель Фредгольма" $\mathcal{D}(\lambda)$ и "первый минор Фредгольма" $\mathcal{D}(t,\tau;\lambda)$ – целые функции от λ.

2. Если $\mathcal{D}(\lambda) \neq 0$, то уравнение (VI.3.1) имеет одно и то...

ко ог) решение

$$x(t)=f(t)+\lambda\int_0^1\Gamma(t,\tau;\lambda)f(\tau)d\tau. \qquad (VI.3$$

3. *Коэффициенты рядов*

$$\mathcal{D}(\lambda)=\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}c_k\lambda^k \qquad (VI.3$$

и

$$\mathcal{D}(t,\tau;\lambda)=\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}B_k(t,\tau)\lambda^k \qquad (VI.3$$

удовлетворяют рекуррентным соотношениям

$$c_0=1,\ c_k=\int_0^1 B_{k-1}(\tau,\tau)d\tau; \qquad (VI.3$$

$$B_0(t,\tau)=\mathcal{K}(t,\tau),$$

$$B_k(t,\tau)=c_k\mathcal{K}(t,\tau)-k\int_0^1\mathcal{K}(t,\tau')B_{k-1}(\tau',\tau)d\tau'. \qquad (VI.3$$

4. Справедлива формула, выражающая коэффициенты $B_k(t$ непосредственно через ядро:

$$k>0,\ B_k(t,\tau)=\int_0^1\cdots\int_0^1\begin{vmatrix}\mathcal{K}(t,\tau) & \mathcal{K}(t,\tau_1)\ldots\mathcal{K}(t,\tau_k)\\ \mathcal{K}(\tau_1,\tau) & \mathcal{K}(\tau_1,\tau_1)\ldots\mathcal{K}(\tau_1,\tau_k)\\ \cdots & \cdots\\ \mathcal{K}(\tau_k,\tau) & \mathcal{K}(\tau_k,\tau_1)\ldots\mathcal{K}(\tau_k,\tau_k)\end{vmatrix}d\tau_1\ldots d\tau_k. \qquad (VI.3$$

3°. Заменим ядро $\mathcal{K}(t,\tau)$ его аппроксимацией (VI.2.5 силу неравенст (VI.2.6) при достаточно малом h будет $|\mathcal{K}^h(t,\tau)|<1$. Обозначим через C_k^h и $B_k^h(t,\tau)$ коэффициен рядов Неймана ядра $\mathcal{K}^h(t,\tau)$. Нетрудно убедиться, что

$$B_k^h(t,\tau)=\sum_{j,l=-1}^{2k-1}b_{jlh}^{(k)}\omega\left(\frac{t}{h}-j\right)\omega\left(\frac{\tau}{h}-l\right), \qquad (VI.3$$

а рекуррентные соотношения (VI.3.6), (VI.3.7) переходят в думные:

$$C_0^h=1,\quad C_k^h=h\sum_{j,l=-1}^{2k-1}b_{jlh}^{(k-1)}w_{jl}. \qquad (VI.3$$

$$b^{(0)}_{jlh} = \mathcal{K}((j+1)h, (l+1)h); \quad \kappa > 0,$$

$$b^{(\kappa)}_{jlh} = c^h_\kappa b^{(0)}_{jlh} - \kappa h \sum_{i,s=-1}^{2n-1} b^{(0)}_{jsh}\, b^{(\kappa-1)}_{slh}\, \sigma_{is}. \qquad (\text{УІ.3.11})$$

Значения w_{ij} даны в формуле (УІ.2.13). Соотношения (УІ.3.10-11) позволяют вычислить резольвенту Фредгольма для ядра $\mathcal{K}^h(t,\tau)$.

4°. В обозначениях § I гл.I $\mathfrak{X}$ есть совокупность рядов Фредгольма $\mathfrak{D}(\lambda)$ и $\mathfrak{D}(t,\tau;\lambda)$, f - ядро $\mathcal{K}(t,\tau)$, соотношение τ определено равенствами (УІ.3.4-7). Далее, $f^{(n)}$ есть ядро $\mathcal{K}^h(t,\tau)$, $\mathfrak{X}^{(n)}$ - совокупность рядов Фредгольма этого ядра, $t_n = \tau$. Погрешность аппроксимации оценивается формулами

$$|\mathfrak{D}(t,\tau;\lambda) - \mathfrak{D}^h(t,\tau;\lambda)| \le Ch^2 [P_3(|\lambda|^2) + |\lambda|^2 \exp(e|\lambda|^2)]^{\frac{1}{2}},$$

$$|\mathfrak{D}(\lambda) - \mathfrak{D}^h(\lambda)| \le Ch^2 [P_3(|\lambda|^2) + |\lambda|^2 \exp(e|\lambda|^2)]^{\frac{1}{2}}; \qquad (\text{УІ.3.12})$$

P_3 - некоторый полином 3-й степени.

Погрешность искажения в данном случае определяется погрешностями вычисления значений $\mathcal{K}(t_{j+1}, t_{l+1})$, т.е., значений коэффициентов $b^{(0)}_{jl}$. Поэтому нет нужды особо выделять погрешность искажения, и можно оценить суммарно погрешности искажения и округления для процесса (УІ.3.10-11).

Оценим погрешность алгоритма. Прежде всего заметим, что ядро $\mathcal{K}^h(t,\tau)$ - вырожденное:

$$\mathcal{K}^h(t,\tau) = \sum_{j,l=-1}^{2n-1} \mathcal{K}(t_{j+1}, t_{l+1}) \omega\left(\frac{t}{h} - j\right) \omega\left(\frac{\tau}{h} - l\right) = \sum_{j=-1}^{2n-1} a_j(t)\, b_j(\tau),$$

где

$$a_j(t) = \omega\left(\frac{t}{h} - j\right), \quad b_j(\tau) = \sum_{l=-1}^{2n-1} \mathcal{K}(t_{j+1}, t_{l+1}) \omega\left(\frac{\tau}{h} - l\right).$$

Отсюда следует, что ряды $\mathfrak{D}^h(t,\tau;\lambda)$ и $\mathfrak{D}^h(\lambda)$ суть полиномы от λ степени не выше $2n$, так что $B^h_\kappa(t,\tau) \equiv 0$ и $C^h_\kappa = 0, \kappa > 2n$, и соотношениями (УІ.3.10-11) достаточно воспользоваться $2n$ раз. Допустим теперь, что мы ограничились вычислением коэффициентов B^h_κ и C^h_κ для $\kappa \le \mathcal{N}$, $\mathcal{N} < 2n$. Погрешность для $\mathfrak{D}^h(t,\tau;\lambda)$ при этом не превосходит величины

$$\eta_{N,h} = \sum_{k=N+1}^{2n} \frac{|B_k^h(t,\tau)|}{k!} |\lambda|^k \leq \sum_{k=N+1}^{2n} \frac{(k+1)^{(k+1)/2}}{k!} |\lambda|^k.$$

Оценим общий член последней суммы. По формуле Стирлинга

$$k! = \sqrt{2\pi k}\left(\frac{k}{e}\right)^k e^{\theta/12k}, \quad 0 < \theta < 1, \quad \text{Отсюда}$$

$$\frac{|\lambda|^k (k+1)^{(k+1)/2}}{k!} < \frac{\sqrt{(k+1)!}\, e^{k/2}}{\sqrt{2\pi(k+1)}\, k!} |\lambda|^k = \frac{\sqrt{e}}{\sqrt{2\pi}} \frac{(|\lambda|\sqrt{e})^k}{\sqrt{k!}} (k+1)^{1/4}$$

и, следовательно,

$$\eta_{N,h} \leq \frac{\sqrt{e}}{\sqrt{2\pi}} \sum_{k=N+1}^{2n} \frac{(|\lambda|\sqrt{e})^k}{\sqrt{k!}} (k+1)^{1/4} < \frac{\sqrt{e}}{\sqrt{2\pi}} \sum_{k=N+1}^{2n} \frac{k^{-1/4}(|\lambda|^2 e)^{k/2}}{\sqrt{(k-1)!}} <$$

$$< \frac{\sqrt{e}}{\sqrt{2\pi}} \frac{1}{\sqrt{N}} \sum_{k=N+1}^{2n} \frac{(|\lambda|^2 e)^{k/2}}{\sqrt{(k-1)!}} =$$

$$= \frac{\sqrt{e}}{\sqrt{2\pi}} \frac{1}{\sqrt{N+1}} \frac{(|\lambda|^2 e)^{(N+1)/2}}{\sqrt{N!}} \sum_{k=N+1}^{2n} \frac{(|\lambda|^2 e)^{(k-N-1)/2}}{\sqrt{(N+1)(N+2)\ldots(k-1)}} \leq$$

$$\leq \frac{\sqrt{e}}{\sqrt{2\pi(N+1)}} \frac{(|\lambda|^2 e)^{(N+1)/2}}{\sqrt{N!}} \sqrt{2n-N} \left[\sum_{\nu=0}^{\infty} \frac{(|\lambda|^2 e)^{\nu}}{(N+1)\ldots(N+\nu)}\right]^{1/2}.$$

Если $N = 0$, то знаменатель в последней сумме надо заменить единицей. Заметим, что общий член этой суммы равен

$$\frac{(|\lambda|^2 e)^{\nu}}{\nu!} \frac{\nu!}{(N+1)\ldots(N+\nu)} < \frac{(|\lambda|^2 e)^{\nu}}{\nu!} \frac{1}{N+1}$$

и упомянутая сумма имеет оценку

$$1 + \sum_{\nu=1}^{\infty} \frac{(|\lambda|^2 e)^{\nu}}{\nu!} \cdot \frac{1}{N+1} < 1 + \frac{\exp(|\lambda|^2 e)}{N+1}.$$

Окончательно, погрешность алгоритма для $\mathcal{D}(t, \tau; \lambda)$ имеет оценку

$$\eta_{N,h} < \frac{\sqrt{2(2n-N)}}{\sqrt{\pi(N+1)}} \frac{(|\lambda|^2 e)^{(N+1)/2}}{\sqrt{N!}} \sqrt{1 + \frac{\exp(|\lambda|^2 e)}{N+1}}. \qquad (\text{VI.3.13})$$

Та же величина оценивает погрешность алгоритма и для $\mathcal{D}(\lambda)$. Формула (VI.3.13) точнее аналогичной формулы, полученной в [18, 22].

5°. Обозначим

$$\frac{b_{jlh}^{(k)}}{k!} = \beta_{jlh}^{(k)}; \quad \frac{c_k^h}{k!} = \gamma_k^{(h)}. \qquad (\text{У}.3.14)$$

погрешности округления величин $\beta_{jlh}^{(k)}$ и $\gamma_k^{(h)}$ обозначим соответственно через $\delta_{jlh}^{(k)}$ и $\delta_k^{(h)}$. Пусть еще погрешности округления правых частей формул (У1.3.10-11), поделенных на $k!$, суть соответственно $\tilde{\varepsilon}_k^{(h)}$ и $\varepsilon_{jlh}^{(k)}$. Положим, наконец, $\delta_{kh} = \max_{jl} |\delta_{jlh}^{(k)}|$ и $\varepsilon_{kh} = \max_{jl} [|\varepsilon_{jlh}^{(k)}| + |\tilde{\varepsilon}_k^{(h)}| + \beta_{jlh}^{(k)}| / k]$. Приняв во внимание, что (см. формулу (У1.2.1

$$\sum_{j,l=-1}^{2n-1} w_{jl} \leqslant 2n = 1/h,$$

легко получить неравенство

$$\delta_{kh} \leqslant M\left(1+\frac{1}{k}\right)\delta_{k-1,h} + \varepsilon_{kh}, \quad M = \max|K(t,\tau)|. \qquad (\text{У}\ 3.15)$$

Пусть вычисления производятся с такой точностью, что $\varepsilon_{kh} \leqslant \varepsilon = const$. Тогда можно получить (см. книгу автора [22]) следующую оценку (C_0 – некоторая постоянная, $\bar{C}$ – постоянная Эйлера)

$$\delta_{kh} \leqslant e^{\bar{C}}\,\delta_{0h}\,k\,(C_0M)^k\,\frac{\left(\frac{3C_0M}{2}\right)^k - 1}{3C_0M/2 - 1}. \qquad (\text{У1.3.16})$$

Если ядро мало, так что $C_0M < 2/3$, то δ_{kh} ограничена; если при этом и λ невелико, то ограничена и погрешность первого минора Фредгольма. Нетрудно убедиться, что при этих же условиях ограничена и погрешность определителя Фредгольма.

6°. Более оптимистическими оказываются вероятностные оценки погрешности округления. Допустим, что в суммах (У1.3.10-11), поделенных на k, ошибки округления частично сокращаются. При некоторых естественных предположениях (см. книгу автора [22]) это приводит к замене неравенства (У1.3.15) более слабым неравенством

$$\delta_{kh} \leqslant M h^{\alpha}\left(1+\frac{1}{k}\right)\delta_{k-1,h} + \varepsilon_{kh}, \quad \alpha = const > 0,$$

что в свою очередь приводит к оценке

$$\delta_{kh} \leqslant e^{\bar{C}}\,\delta_{0h}\,k\,(C_0Mh^{\alpha})^k + \frac{\left(\frac{3C_0Mh^{\alpha}}{2}\right)^k - 1}{3C_0Mh^{\alpha}/2 - 1}, \qquad (\text{У1.3.17})$$

которая выполняется с вероятностью, не меньшей, чем $1 - Ch^{3-2\alpha}$,

$C = const$; таким образом, постоянную α можно выбирать по произволу из промежутка $(0, 3/2)$.

4. Сведение к алгебраической системе

1°. Рассмотрим уравнение (VI.1.1) в следующих предположениях: а) $\mathcal{K} \in L_2([0,1] \times [0,1])$; б) единица не есть характеристическое число ядра $\mathcal{K}(t,\tau)$; в) $f \in L_2(0, 1)$. Пусть построено такое вырожденное ядро

$$\mathcal{K}^{(n)}(t,\tau) = \sum_{k=1}^{n} a_k(t)\beta_k(\tau), \qquad (VI.4.1)$$

что

$$\int_0^1\int_0^1 [\mathcal{K}(t,\tau)-\mathcal{K}^{(n)}(t,\tau)]^2 dt\,d\tau := \gamma_n \xrightarrow[n\to\infty]{} 0 . \qquad (VI.4.2)$$

В качестве $\mathcal{K}^{(n)}(t,\tau)$ можно взять, например, кусочно-постоянную аппроксимацию ядра $\mathcal{K}(t,\tau)$, если только это последнее обладает необходимой гладкостью; отметим, что числовые эксперименты подтвердили целесообразность такого выбора. Другие способы выбора ядра $\mathcal{K}^{(n)}(t,\tau)$ указаны в книге автора и Смолицкого [1]. Будем считать, что свободный член уравнения $f(t)$ также заменен другим свободным членом $f^{(n)}(t)$, близким к $f(t)$ в норме $L_2(0, 1)$, так что $\|f - f^{(n)}\|_2 \xrightarrow[n\to\infty]{} 0$.

Таким образом, уравнение (VI.1.1) мы заменили новым уравнением

$$x^{(n)}(t) - \int_0^1 \mathcal{K}^{(n)}(t,\tau)x^{(n)}(\tau)d\tau = f^{(n)}(t); \qquad (VI.4.3)$$

оценим погрешность аппроксимации такой замены. Пусть $\mathcal{K}$ и $\mathcal{K}^{(n)}$ – операторы Фредгольма с ядрами, соответственно равными $\mathcal{K}(t,\tau)$ и $\mathcal{K}^{(n)}(t,\tau)$. Положим $R = (I - \mathcal{K})^{-1}$, $R^{(n)} = (I - \mathcal{K}^{(n)})^{-1}$. Тогда $x_* = Rf$, $x_*^{(n)} = R^{(n)}f^{(n)}$ и

$$\|x_* - x^{(n)}\| = \|R(f - f^{(n)}) + (R - R^{(n)})f^{(n)}\| \le$$

$$\le \|R\|\gamma_n' + \|R - R^{(n)}\|(\|f\| + \gamma_n'); \quad \gamma_n' = \|f - f^{(n)}\|. \qquad (VI.4.4)$$

Далее, положим $\mathcal{K}^{(n)} - \mathcal{K} = \Delta^{(n)}$. Тогда $R^{(n)} = (I - \mathcal{K} + \Delta^{(n)})^{-1} = R(I + \Delta^{(n)}R)^{-1}$; отсюда $R - R^{(n)} = -R\Delta^{(n)}R(I + \Delta^{(n)}R)^{-1}$; если n достаточно велико, то $\|\Delta^{(n)}R\| \le \gamma_n\|R\| < 1$ и

$$\|R-R^{(n)}\| \le \frac{\gamma_n \|R\|^2}{1-\gamma_n\|R\|}.$$

Подставив это в (VI.4.4), получим следующую оценку погрешности аппроксимации

$$\|x_*-x_*^{(n)}\| \le \|R\|\gamma_n + \frac{\gamma_n\|R\|^2}{1-\gamma_n\|R\|}(\|f\|+\gamma_n') \qquad (VI.4.5)$$

2°. Обычным способом решение уравнения (VI.4.3) сводится к решению некоторой линейной алгебраической системы n порядка n; при этом возникает погрешность искажения, которую можно оценить по формуле (I.4.I4). Погрешности алгоритма и округления можно оценить по способу § 4 гл.I; оценки зависят от метода, которым система решается.

3°. Уравнение Фредгольма можно свести к линейной алгебраической системе по методу Бубнова - Галеркина. В этом случае погрешность аппроксимации можно оценить по формуле (II.8.I). Далее, если координатная система сильно минимальна в $L_2(0, I)$, то процесс вычисления коэффициентов устойчив (в смысле второго определения) в последовательности (R_n, R_n), а процесс вычисления приближенного решения устойчив в последовательности $(H^{(n)}, R_n)$; смысл обозначений очевиден.

4°. Еще один способ сведения к линейной алгебраической системе дает метод наименьших квадратов. Мы не будем здесь останавливаться на сущности метода - об этом достаточно полно сказано в книге автора [3], и сделаем только два замечания общего характера.

I) Пусть метод наименьших квадратов применен к уравнению $Ax = f$, где A - линейный оператор, действующий в некотором гильбертовом пространстве H, и что этот оператор ограниченно обратим. Если приближенное решение $x_*^{(n)}$ уже построено, то его погрешность оценивается очень просто [3]:

$$\|x_*-x_*^{(n)}\| \le \|A^{-1}\|\cdot\|f-Ax_*^{(n)}\|. \qquad (VI.4.6)$$

Последняя формула дает суммарную погрешность приближенного решения.

2) Пусть еще оператор A ограничен в H. Тогда можно оценить погрешность аппроксимации метода наименьших квадратов по форму-

ле (IУ.I.4).

Для уравнений Фредгольма обе упомянутые здесь формулы очевидно справедливы.

Сравнительно просто решается вопрос об устойчивости относительно искажения. В книге автора [3] приведены следующие простые факты. Если оператор A замкнут в гильбертовом пространстве H и имеет ограниченный обратный, и если $f \in \mathcal{D}(A^*)$, то уравнения $Ax = f$ и $A^*Ax = A^*f$ равносильны. Оператор A^*A положительно определен, его энергетическое пространство совпадает с множеством значений оператора A, а энергетическая норма $|x| = \|Ax\|$. Если A - ограниченный оператор, то, очевидно, эта последняя норма эквивалентна норме пространства H. Из сказанного сразу вытекает следующее утверждение, верное, если уравнение Фредгольма разрешимо единственным образом.

Для того чтобы процесс наименьших квадратов был устойчив относительно погрешностей искажения, необходимо и достаточно, чтобы координатная система была сильно минимальна в $L_2(0, I)$.

Система уравнений метода наименьших квадратов - алгебраическая линейная, для погрешностей алгоритма и округления этой системы справедливо все сказанное по этому поводу в § 4 гл.I.

Можно еще отметить, что если координатная система почти ортонормирована в $L_2(0, I)$, то число обусловленности матрицы метода наименьших квадратов ограничено независимо от n.

Сказанное в пп. 3^0 и 4^0 настоящего параграфа полностью относится к уравнениям второго рода с вполне непрерывным оператором в сепарабельном гильбертовом пространстве.

5^0. К линейной алгебраической системе можно приближенно свести уравнение Фредгольма и по методу коллокации. В книге Пудучаева [3] доказывается, что оптимальная по порядку оценка погрешности аппроксимации достигается при использовании подходящим образом подобранных сплайнов в качестве координатных функций. Если ядро и свободный член уравнения (УI.I.I) принадлежат к классу $C^{(k)}$, а их k-е модули непрерывности мажорируются одной и той же функцией $\omega(\delta)$, то погрешность аппроксимации коллокационного метода имеет оценку $O(n^{-k}\omega(n^{-1}))$. В той же книге даны и некоторые другие результаты, близкие к только-что сформулированному. По поводу устойчивости процесса коллокации относительно погрешностей искажения можно повторить то, что было

сказано в общем случае в § 9 гл.Ш. В частности, если выбрана последовательность координатных функций $\varphi_k(t)$, $1 \le k < \infty$, и $|\varphi_k(t)| \le Ck^{\alpha}$, $\alpha < 1$, то процесс коллокации неустойчив.

При фиксированном порядке приближения и при заданных координатных функциях и узлах коллокации мы имеем дело с вполне определенной алгебраической системой. Для этой системы можно оценить погрешности искажения, алгоритма и округления так, как это описано в § 4 гл.I.

§ 5. Сингулярные интегральные уравнения. Метод наименьших квадратов

1°. Будем рассматривать одномерное сингулярное интегральное уравнение вида

$$Ax := a(t)x(t) + b(t)(Sx)(t) + (Tx)(t) = f(t), \qquad (\text{VI.5.1})$$

где S - оператор Коши

$$(Sx)(t) = \frac{1}{\pi i}\int_{\Gamma} \frac{x(\tau)}{\tau - t}\,d\tau, \qquad (\text{VI.5.2})$$

в следующих предположениях: Γ - замкнутый плоский контур класса $C^{(1,\lambda)}$, ограничивающий некоторую область $\mathcal{D}$, односвязную или многосвязную, $a(t)$ и $b(t)$ - функции (скалярные или матричные), непрерывные на Γ, T - оператор, вполне непрерывный в $L_2(\Gamma)$, индекс уравнения (VI.5.1) равен нулю и это уравнение имеет не более одного решения, так что оно разрешимо при любой (соответственно скалярной или векторной) функции $f(t)$. При перечисленных условиях операторы A и A^{-1} ограничены в $L_2(\Gamma)$.

2°. Уравнение (VI.5.1) можно решать по методу наименьших квадратов, и к этому уравнению в полной мере относится то, что было сказано в п.4° предшествующего параграфа.

Отметим некоторые координатные системы.

1) Пусть $\Gamma = \bigcup_{j=0}^{q} \Gamma_j$, где Γ_j - замкнутые кривые; предположим, что Γ_0 есть граница некоторой конечной области $\mathcal{D}$ и что Γ_0 ограничивает область $\mathcal{D}$ извне. Внутри Γ_j, $1 \le j \le q$, выберем точки t_j. Обозначим через $\rho_0, \rho_1, \dots, \rho_q$ числа, удовлетворяющие неравенствам

$$0 < \rho_0 < \min_{\tau\in\Gamma_0} |\tau - t_0|, \qquad \rho_j > \max_{\tau\in\Gamma_j} |\tau - t_j|, \quad 1 \le j \le q.$$

Тогда система функций

$$\left(\frac{t}{\rho_0}\right)^{n+1},\ \left(\frac{\bar{t}}{\rho_0}\right)^{n+1},\ \left(\frac{\rho_1}{t-t_j}\right)^{n},\ \left(\frac{\rho_j}{\bar{t}-\bar{t}_j}\right)^{n},\quad n=1,2,\ldots,\quad 1\le j\le q$$

сильно минимальна в $L_q(\Gamma)$ (ом. книгу автора [3]).

2) Пусть Γ – вещественная ось, и пусть вполне непрерывный оператор T – сглаживающий, так что $T: W_2^{(k)}(R_1) \to W_2^{(k+1)}(R_1)$. Взять в качестве координатных взять функции

$$\varphi_k^{(1)}(t) = \sqrt{\frac{2}{1+t^2}}\, T_k\left(\frac{2t}{1+t^2}\right),\quad k=0,1,2,\ldots,$$

$$\varphi_k^{(2)}(t) = \frac{\sqrt{2}\,(1-t^2)}{(1+t^2)^{3/2}}\, U_{k-1}\left(\frac{2t}{1+t^2}\right),\quad k=1,2,\ldots,$$

где T_k и U_k – полиномы Чебышева соответственно первого и второго рода, то общую оценку (I.I.5) можно уточнить: при $f\in W_2^{(2)}$ и $a(t), b(t)\in W_2^{(2)}(R_1)$ будет

$$\|x_* - x_*^{(n)}\|_{2,2_1} = O(n^{-3}).$$

При этом процесс наименьших квадратов устойчив относительно погрешностей искажения, а число обусловленности матрицы метода наименьших квадратов ограничено (ом. книгу автора и Прессдорф [1]).

3°. Для простоты допустим, что Γ – окружность $|t|=1$ и что уравнение (VI 5.1) – скалярное. Допустим, что решение уравнения (VI.5.1) $x_*\in W_2^{(2s)}$; для этого достаточно, например, чтобы вполне непрерывный оператор T был сглаживающим (или чтобы он действовал из $L_2(\Gamma)$ в $W_2^{(2s)}(\Gamma)$), а $f(t), a(t), b(t)\in W_2^{(2s)}$. В качестве координатных возьмем функции, которые получаются кусочно полиномиальных, степени $2s-1$, функций МКЭ (см. книгу автора [22], гл.II, § 6). В этом случае, как нетрудно доказать, погрешность аппроксимации имеет оценку $O(n^{-2s})$. Повторяя, с некоторыми естественными упрощениями, рассуждения книги автора [22], гл.IX, § 2, легко доказать, что при выборе координатных функций, указанном в данном пункте, справедливо следующее утверждение: процесс наименьших квадратов устойчив относительно погрешностей искажения в последовательности пар пространств (X_n, Y_n), где X_n и Y_n – n-мерные пространства с нормами

$$\forall a \in R_n,\ \|a\|_{\tilde{X}_n} = h^{\frac{1}{2}} \|a\|_{R_n},\ \|\gamma\|_{\tilde{Y}_n} = h^{\frac{1}{2}} \|a\|_{R_n}.$$

Точно так же доказывается, что процесс построения приближенного решения по методу наименьших квадратов устойчив в последовательности $(H_n, \tilde{Y}_n)$, где H_n есть n-мерное подпространство пространства $L_2(\Gamma)$, натянутое на соответствующие координатные функции. Наконец, рассуждения § 5 гл.IX книги автора [22] показывают, что число обусловленности матрицы метода наименьших квадратов при том же выборе координатных функций с точностью до n .

Распространение результатов данного пункта на системы одномерных сингулярных уравнений не встречает затруднений.

4°. Отметим результаты статьи [12] Джишкариани, в которой исследуются одномерные сингулярные интегральные уравнения на разомкнутом контуре. В цитируемой статье Джишкариани ограничивается случаем уравнений

$$a x(t) - \frac{b}{\pi i} \int_{-1}^{1} \frac{x(\tau)}{\tau - t} d\tau + \frac{1}{\pi} \int_{-1}^{1} K(t,\tau) x(\tau) d\tau = f(t), \qquad (\text{VI.5.3})$$

где a, b – постоянные, $a^2 + b^2 > 0$. Уравнение рассматривается в пространстве $H = L_2(-1,+1;\rho)$; вес $\rho(t)$ выбирается в зависимости от индекса оператора (VI.5.3). Интегральный оператор

$$\int_{-1}^{1} K(t,\tau) x(\tau) d\tau$$

предполагается вполне непрерывным в H ; предполагается также, что уравнение (VI.5.3) имеет в H единственное решение. Приближенное решение этого уравнения строится по методу Бубнова – Галеркина, в качестве координатных функций выбираются полиномы Чебышева или Якоби. Доказывается, что алгебраическая система Бубнова – Галеркина достаточно высокого порядка однозначно разрешима и что приближенные решения сходятся к точному решению в H . Устанавливается устойчивость процесса.

5°. Перейдем к рассмотрению многомерных сингулярных уравнений. Рассмотрим сингулярное (скалярное или векторное, безразлично) интегральное уравнение

$$Ax := a(t)x(t) + \int_{\Gamma} K(t,\tau) x(\tau) d\tau + (Tx)(t) = f(t). \qquad (\text{VI.5.4})$$

Здесь Γ – достаточно гладкое многообразие m измерений без края, $K(t,\tau)$ – сингулярное ядро, T – оператор, вполне непрерывный в том банаховом пространстве, в котором сингулярный оператор ограничен; для определенности будем говорить о пространстве $L_2(\Gamma)$. Примем следующие допущения: 1) символ $\Phi(t,\tau)$ уравнения (VI.5.1) – достаточно гладкая функция своих аргументов; 2) однородное уравнение равен нулю; 3) уравнение (VI.5.4) имеет в $L_2(\Gamma)$ не более одного решения – тогда оно разрешимо для любой правой части из $L_2(\Gamma)$. Операторы A и A' ограничены. Повторяя многократно использованные выше рассуждения, мы придем к следующим результатам.

Уравнение (VI.5.4) допускает применение метода наименьших квадратов; погрешность аппроксимации оценивается по формуле (IV.1.4). Если координатная система сильно минимальна в L_2, то соответствующий вычислительный процесс устойчив относительно погрешностей искажения; если эта система еще и почти ортонормирована, то число обусловленности алгебраической системы метода наименьших квадратов ограничено независимо от порядка этой системы. Таким образом, если эту систему решать по методу простых итераций, то эти итерации сходятся, как прогрессия с фиксированным знаменателем, меньшим единицы, и погрешности округления при возрастании порядка системы остаются ограниченными.

В статьях автора и Рудевой [1] и автора [25] приведены примеры координатных систем, для которых общая оценка (IV.1.4) может быть уточнена.

§ 6. Методы механических квадратур и коллокации для одномерных сингулярных уравнений.

1°. Приложениям методов механических квадратур к одномерным сингулярным интегральным уравнениям посвящен ряд работ Габдулхаева и некоторых других авторов. Довольно полная библиография приведена в статье Габдулхаева и Душкова [1]; мы ограничимся здесь кратким изложением статьи Габдулхаева [2].

Рассмотрим уравнение (VI.5.1), в котором вполне непрерывный оператор T – интегральный:

$$(Tx)(t)=\int_{\Gamma} K(t,\tau)x(\tau)d\tau. \qquad (VI.6$$

Примем, что Γ – окружность $|\tau|=1$, и что функции $a(t)$, $b(t)$, $f(t)$, $K(t,\tau)$ принадлежат к классу $C^{(\nu,\alpha)}$, где

$\nu > 0$, $0 < \alpha < 1$. Допустим далее, что индекс рассматриваемого уравнения равен нулю и что это уравнение имеет не более одного решения. Приближенное решение будем искать в виде тригонометрического многочлена

$$x_n(t) = \sum_{k=-n}^{n} a_k t^k. \qquad (\text{VI.6.2})$$

Из элементарных свойств оператора Коши вытекает, что

$$(Ax^{(n)})(t) =$$

$$[a(t)+b(t)]\sum_{k=0}^{n} a_k t^k + [a(t)-b(t)]\sum_{k=-1}^{-n} a_k t^k + \int_\Gamma \tau(t,\tau)\sum_{k=-n}^{n} a_k \tau^k d\tau.$$

Заменим здесь интеграл его приближенным значением по формуле прямоугольников, соответствующей разбиению окружности Γ на $2n+1$ равных дуг с точками деления $t_j = \exp[2\pi ij/(2n+1)]$, и потребуем, чтобы полученная таким образом функция от t совпала с $f(t)$ в точках t_j. Это приведет нас к системе линейных алгебраических уравнений

$$[a(t_j)+b(t_j)]\sum_{k=0}^{n} a_k t_j^k + [a(t_j)-b(t_j)]\sum_{k=-1}^{-n} a_k t_j^k +$$

$$+\sum_{k=-n}^{n} a_k \sum_{\rho=0}^{2n} \mathcal{K}(t_j, t_\rho)\, t_\rho^k (t_{\rho+1}-t_\rho) = f(t_j);\ j=0,1,\dots,2n. \qquad (\text{VI.6.3})$$

В статье Габдулхаева [...] эта система трактуется как с... метода механических квадратур для уравнения (VI.5.1). Может быть, стоит отметить, что метод, примененный Габдулхаевым, представляет собой некоторый синтез методов коллокации и механических квадратур: он переходит в метод коллокации при $\mathcal{K}(t,\tau)\equiv 0$ и в метод механических квадратур при $b(t)\equiv 0$.

Основной результат рассматриваемой статьи содержится в следующей теореме:

Теорема VI.6.1. Пусть β – любое число, заключенное в интервале $0 < \beta < \alpha$. Существуют такие положительные постоянные A_1, B_1, A_2, B_2, что если

$$(A_1 \ln n + B_1)\, n^{-\nu-\alpha-\beta} < 1, \qquad (\text{VI.6.4})$$

то система (УІ.6.3) имеет единственное решение. Погрешность проксимации приближенного решения (УІ.6.2) оценивается по формуле

$$\|x_* - x_*^{(n)}\|_{H_\beta} \le (A_2 \ln n + B_2) n^{-\tau-\alpha+\beta}; \qquad (\text{УІ.6.}$$

H_β – пространство функций, заданных на Γ, с нормой

$$\|x\|_{H_\beta} = \max_{t\in\Gamma} |x(t)| + \sup_{t',t''\in\Gamma} \frac{|x(t') - x(t'')|}{|t' - t''|^\beta}.$$

Доказательство этой теоремы использует общую теорию приближенных методов, построенную Канторовичем (см. Канторович и Акилов [1]).

2°. Рассмотрим теперь уравнение (УІ.5.1) в следующих предположениях: $\Gamma = 0$; $a(t)$, $b(t)$ непрерывны на Γ ; индекс уравнения равен нулю. Заметим, что результаты статьи Габдулхаева [2] здесь неприменимы, потому что $\tau = \alpha = 0$. Данный случай рассмотрен в статье Преслорфа и Шмидта [1]. В качестве узлов коллокации взяты точки $t_k = t_k^{(n)} = \exp(2\pi i k/n)$, а в качестве координатных функций – кусочно линейные функции, широко употребляемые в МКЭ и определяемые формулой

$$\varphi_k^{(n)}(t) = \begin{cases} \dfrac{t - t_{k-1}}{t_k - t_{k-1}}; & \dfrac{2\pi(k-1)}{n} \le \arg t \le \dfrac{2\pi k}{n}; \\ \dfrac{t_{k+1} - t}{t_{k+1} - t_k}, & \dfrac{2\pi k}{n} \le \arg t \le \dfrac{2\pi(k+1)}{n}; \\ 0 \text{ в остальных случаях;} \end{cases} \qquad (\text{УІ.6.}$$

Приближенное решение имеет вид

$$x^{(n)}(t) = \sum_{k=0}^{n-1} a_k^{(n)} \varphi_k^{(n)}(t); \qquad (\text{УІ.6.}$$

коэффициенты $a_k^{(n)}$ определяются из системы

$$a(t_j) a_j^{(n)} + b(t_j) \sum_{k=0}^{n-1} a_k^{(n)} (S\varphi_k^{(n)})(t_j) = f(t_j); \quad j = 1, 2, \ldots, n. \qquad (\text{УІ.6.}$$

Основные результаты статьи Преслорфа и Шмидта дают следующие теоремы:

Теорема УІ.6.2. Пусть коэффициенты $a(t)$, $b(t)$ непрерывны, а правая часть $x_* \in L_2(\Gamma)$. Пусть выполнено условие

$$\forall\lambda\in[-1,1],\ \forall\, t\in\Gamma;\ a(t)+\lambda b(t)\neq 0. \qquad (VI.6.9)$$

тогда при достаточно больших n система (VI.6.8) разрешима единственным образом и $\|x_* - x_*^{(n)}\|_{2,\Gamma} \xrightarrow[n\to\infty]{} 0$.

Теорема VI.6.3. Если $a(t)$, $b(t)\in H_\mu$, $0<\mu<1$, и $f\in H_\nu$, $0<\nu<1$, то справедлива оценка

$$\|x_* - x_*^{(n)}\|_{2,\Gamma} \leqslant c n^{-\sigma} \ln n,\quad \sigma = \min(\mu,\nu).$$

Теорема VI.6.4. Пусть $a(t)$, $b(t)\in H_\mu$, $1/2 < \mu < 1$. Если при любой непрерывной функции $f(t)$ система (VI.6.8) разрешима при достаточно большом n, и $\|x_* - x_*^{(n)}\|_{2,\Gamma} \xrightarrow[n\to\infty]{} 0$, то справедливо неравенство (VI.6.9).

3°. При фиксированном n можно получить оценки для погрешностей искажения, алгоритма и округления так, как об этом сказано в § 4 гл.I.

4°. Дыликорилин [14] применил метод коллокации к уравнениям вида (VI.5.8) при ограничениях, перечисленных в п.3° § 5 данной главы.

Доказывается, что если за узлы коллокации взять корни полиномов Якоби, выбранных подходящим образом, то алгебраические системы метода коллокации, порядок которых достаточно велик, однозначно разрешимы, и приближенные решения сходятся к точным в соответствующей метрике.

ГЛАВА УП

Нелинейные вычислительные процессы

§ I. О погрешности аппроксимации метода Ритца

В п.п. I^0 - 4^0 настоящего параграфа мы коротко изложим постановку задачи о минимуме функционала и метод Ритца для ее решения; подробно эти вещи изложены в книге автора [8], гл.IX. В п.5 будет рассмотрена погрешность аппроксимации процесса Ритца.

I^0. Пусть B - рефлексивное банахово пространство и $\mathcal{F}$ - функционал, определенный на линейном множестве $\mathcal{D}(\mathcal{F})$, плотном в B. Ограничимся случаем, когда функционал $\mathcal{F}$ непрерывно дифференцируем на любой конечномерной гиперплоскости, лежащей в B. Конечномерной гиперплоскостью в банаховом пространстве B мы называем совокупность элементов вида

$$x_0+\sum_{k=1}^{n} a_k x_k , \qquad (\text{УП.I.I})$$

где n - фиксированное натуральное число, называемое размерностью гиперплоскости; a_k, $1\le k\le n$ - произвольные числа, x_k, $0<k\le n$ - фиксированные элементы пространства B. Говоря, что $\mathcal{F}$ обладает той или иной гладкостью на гиперплоскости (УП.I.I), мы имеем под этим, что такой гладкостью обладает функция переменных $a_1, a_2, \dots, a_n$, равная $\mathcal{F}(x_0+a_1x_1+\dots+a_nx_n)$.

Если функционал $\mathcal{F}$ имеет локальный минимум в точке $x_0\in\mathcal{D}(\mathcal{F})$, то

$$(grad\,\mathcal{F})(x_0)=0. \qquad (\text{УП.I.2})$$

Мы будем обозначать $grad\,\mathcal{F}=P$; очевидно, P действует из B в B^*.

Функционал $\mathcal{F}$ называется выпуклым, если

$$\forall x,\ y\in\mathcal{D}(\mathcal{F}),\ \xi\in(0,1)\quad \xi\mathcal{F}(x)+\mathcal{F}(y)-2\mathcal{F}\left(\frac{x+y}{2}\right)\ge 0; \qquad (\text{УП.I.3})$$

он называется существенно выпуклым, если в соотношении (УП.I.3) знак равенства имеет место только при $x=y$. Далее, функционал $\mathcal{F}$ называется положительным, если $\mathcal{F}(x)\to+\infty$ тогда и только тогда, когда $\|x\|\to\infty$. Функционал $\mathcal{F}$ называется полунепрерывным снизу (сверху) в точке x_0, если $\varliminf_{x\to x_0}\mathcal{F}(x)\ge\mathcal{F}(x_0)$ (соответственно, $\varlimsup_{x\to x_0}\mathcal{F}(x)\le\mathcal{F}(x_0)$). Этот же функционал называется слабо полунепрерывным снизу (сверху), если упомянутые

соотношения выполняются каждый раз, когда $x \longrightarrow x_0$; символ $\longrightarrow$ означает слабую сходимость.

Теорема УП.1.1. Если функционал $\mathcal{F}$ возрастающий, существенно выпуклый и слабо полунепрерывный снизу, то он ограничен снизу и его нижняя грань достигается в единственной точке, к которой слабо сходится любая минимизирующая последовательность.

Эта теорема доказана в работе Гельмана [1].

2°. В данном пункте излагаются результаты работы Бирмана [1]

Пусть линейный оператор A действует из рефлексивного пространства B в сопряженное пространство B^*, тогда сопряженный оператор A^* также действует из B в B^*. Будем говорить, что оператор A : 1) симметричен, если $\mathcal{D}(A) \subset \mathcal{D}(A^*)$ и $Ax = A^*x$, $x \in \mathcal{D}(A)$; 2) самосопряжен, если он симметричен и $\mathcal{D}(A) = \mathcal{D}(A^*)$; 3) положителен, если он симметричен и $(Ax, x) > 0$ при $x \neq 0$; 4) положительно определен, если он симметричен и $(Ax, x) \geqslant \gamma^2 \|x\|$, где γ – положительная постоянная. Если A – положительный (в частности, положительно определенный) оператор, то с ним можно связать энергетическое пространство B_A так же, как в случае, когда пространство B само гильбертово (см. книгу автора [3]). Именно, за B_A мы принимаем гильбертово пространство, получаемое замыканием множества $\mathcal{D}(A)$ в норме, порожденной скалярным произведением $[x, y]_A = [x, y] = (Ax, y)$; здесь (Ax, y) есть значение функционала $Ax \in B^*$ на элементе $y \in B$. Норму в B_A будем обозначать через $|\cdot|$, так что $|x| = \sqrt{[x, x]}$.

Если A положительно определенный оператор, то B_A ограниченно вкладывается в B ; для положительного, но не положительно определенного оператора это утверждение неверно

3°. Пусть оператор $P = grad\, \mathcal{F}$ имеет производную Гато P'_u со следующими свойствами: а) эта производная существует при любом $u \in \mathcal{D}(P)$; б) $\mathcal{D}(P'_u) \supset \mathcal{D}(P)$; в) оператор P'_u – положительно определенный при любом $u \in \mathcal{D}(P)$, так что

$$\forall x \in \mathcal{D}(P'_u), (P'_u x, x) \geqslant \gamma^2 \|x\|^2 , \qquad (УП.1.4)$$

причем величина γ не зависит от u . Доказывается, что в этих условиях функционал $\mathcal{F}$ ограничен снизу, и любая минимизирующая последовательность сходится в метрике B к пределу, который не зависит от выбора минимизирующей последовательности. Этот предел

будем рассматривать как обобщенное решение задачи о минимуме функционала $\mathcal{F}$.

Представляет интерес случай, когда P_u удовлетворяет не только неравенству (УП.I 1), но и более сильному неравенству

$$\forall x \in \mathcal{D}(P_u'),\ (P_u' x, x) \geqslant d^2 (P_{u_0}' x, x),\ d = const > 0, \quad (\text{УП.I.}$$

где u_0 – некоторый фиксированный элемент из $\mathcal{D}(P_u')$. Обозначим через U_0 энергетическое пространство оператора P_{u_0}' и через $[\,,\,]_0$, $|\cdot|_0$ скалярное произведение и норму в этом пространстве. Можно доказать, что в этом случае обобщенное решение принадлежит пространству B_0 .

4°. Коротко опишем метод Ритца для задачи о минимуме функционала $\mathcal{F}$. Напомним, что частным случаем метода Ритца является МКЭ. Зададим последовательность конечных последовательностей координатных элементов

$$\{\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN}\};\ n = 1, 2, \ldots;\ N = N(n), \quad (\text{УП.I.6})$$

подчиненных условиям: а) $\varphi_{nk} \in \mathcal{D}(\mathcal{F}) \cap B_0$; б) при данном n элементы φ_{nk} линейно независимы; в) последовательность (УП.I. полна в B_0 . Обозначим через B_n подпространство пространства B_0 с базисом $\{\varphi_n\}$, $k = 1, 2, \ldots, N$, и будем искать минимум $\mathcal{F}$ на B_n . Решение $x_*^{(n)}$ этой новой задачи – приближенное решение данной задачи по Ритцу – существует и единственно, если $\mathcal{F}$ удовлетворяет условиям пп. 2°, 3°; это приближенное решение имеет вид

$$x_*^{(n)} = \sum_{k=1}^{n} a_k^{(n)} \varphi_{nk}, \quad (\text{УП.I.7})$$

где $a_k^{(n)}$ – численные коэффициенты, определяемые из нелинейной системы Ритца

$$\frac{\partial}{\partial a_j^{(n)}} \mathcal{F}\Big(\sum_{k=1}^{N} a_k^{(n)} \varphi_{nk}\Big) = 0\ ,\ 1 \leqslant j \leqslant N. \quad (\text{УП.I.8})$$

Будем рассматривать только то решение системы (УП.I.8), которое доставляет величине $\mathcal{F}\Big(\sum_{k=1}^{N} a_k^{(n)} \varphi_{nk}\Big)$ абсолютный минимум (эта оговорка не нужна, если функционал $\mathcal{F}$ существенно выпуклый – в этом случае система (УП.I.8) имеет единственное решение).

Если последовательность подпространств B_n расширяющаяся, то (см. книгу автора [8]) последовательность $\{x_*^{(n)}\}$ приближенных

решений по Ритцу - минимизирующая для функционала $\mathcal{F}$, если этот последний - возрастающий и полунепрерывный сверху. Нетрудно показать, что это утверждение верно и тогда, когда подпространства B_n не расширяющиеся (так будет, например, в случае МКЭ), но их последовательность полна в B_0. В цитированной выше книге автора [8], § 7°, доказывается (без предположения о том, что подпространства B_n расширяющиеся), что из любой последовательности приближенных решений по Ритцу можно выделить минимизирующую подпоследовательность. Допустим, что последовательность всех приближенных решений по Ритцу не минимизирующая. Тогда существует такое число $\varepsilon_0 > 0$ и такая подпоследовательность $\{x_*^{(n_j)}\}$ приближенных решений по Ритцу, что $\mathcal{F}(x_*^{(n_j)}) \geqslant d + \varepsilon_0$, $d = \inf \mathcal{F}(x)$.

Но из этой подпоследовательности можно выделить минимизирующую последовательность, что невозможно. Полученное противоречие доказывает наше утверждение.

Из доказанного утверждения следует, что о сходимости метода Ритца можно повторить все сказанное в п.3° о сходимости минимизирующей последовательности.

5°. Для некоторого класса функционалов можно получить оценку погрешности аппроксимации процесса Ритца. Допустим, что производная P'_u оператора $P = \operatorname{grad} \mathcal{F}$ удовлетворяет, кроме неравенств (УП.I.4) и (УП.I.5), еще и неравенству

$$\forall x \in \mathcal{D}(P'_u),\ (P'_u x, x) \leqslant \beta^2 (P'_{u_0} x, x),\quad \beta = \text{const} > 0. \quad \text{(УП.I.9)}$$

Воспользуемся равенством

$$\mathcal{F}(x_* + y) - \mathcal{F}(x_*) = \int_0^1 \tau^{-1} d\tau \int_0^1 (P_{x_* + \tau y}\, \tau y, \tau y)\, dt.$$

Применив это равенство (УП.I.9) и положив $x_* + y = x$, получим

$$\mathcal{F}(x) - \mathcal{F}(x_*) \leqslant \beta^2 \int_0^1 \tau^{-1} d\tau \int_0^1 (P'_{u_0}(\tau(x - x_*)), \tau(x - x_*))\, dt = \frac{1}{2}\beta^2 \|x - x_*\|_0^2. \quad \text{(УП.I.10)}$$

Пусть $y^{(n)}$ – элемент подпространства B_n, на котором достигается наилучшее приближение элемента x_*:

$$\mathcal{E}_n(x_*) = \inf_{x^{(n)} \in B_n} \|x_* - x^{(n)}\|_0 = \|x_* - y^{(n)}\|_0.$$

Полагая в (УП.I.10) $x = y^{(n)}$, находим

$$\mathcal{F}(y^{(n)}) - \mathcal{F}(x_*) \leq \frac{1}{2}\beta^2 |y^{(n)} - x_*|_0^2 = \frac{1}{2}\beta^2 \mathcal{E}_n^2(x_*). \qquad (\text{УП.I.I})$$

Если $x_*^{(n)}$ – приближенное решение по Ритцу, то $\mathcal{F}(x_*^{(n)}) \leq \mathcal{F}(y$ по неравенствам (УП.I.5) и (УП.I.9) получаем

$$\frac{1}{2}\alpha^2 |x_* - x_*^{(n)}|_0^2 \leq \mathcal{F}(x_*^{(n)}) - \mathcal{F}(x_*) \leq \mathcal{F}(y^{(n)}) - \mathcal{F}(x_*) \leq \frac{1}{2}\beta^2 \mathcal{E}_n^2(x_*),$$

что дает следующую оценку погрешности аппроксимации:

$$|x_* - x_*^{(n)}|_0 \leq \frac{\beta}{\alpha}\mathcal{E}_n(x_*). \qquad (\text{УП.I.I})$$

Если пространство B – функциональное, то оценка порядка величины $\mathcal{E}_n(x_*)$ как функции от n существенно зависит от гладкости решения x_*. По этому тому см., например, книги Мо [I], Лещиненко и Уральцевой [I]. Для метода конечных элем оценки величины $\mathcal{E}_n(x_*)$ в зависимости от гладкости функции см. книги автора [I7, 22], Обэна[I], Стренга и Фикса[I], Съя ле[I].

§ 2. Погрешность аппроксимации и устойчивость вычислительного процесса

I°. Пусть даны две последовательности банаховых пространс X_n и Y_n, $n = 0, I, 2, \ldots$, и пусть при любом n оператор (в общем говоря, нелинейный) A_n действует из X_n в Y_n, определ на всем пространстве X_n (это требование можно ослабить), уд влетворяет равенству $A_n(0) = 0$, имеет обратный оператор A который в свою очередь определен на всем пространстве Y_n и у влетворяет следующему условию Липшица: если $\|y_1\|, \|y_2\| \leq t$,

$$\|A_n^{-1}(y_1) - A_n^{-1}(y_2)\| \leq C_n(t)\|y_1 - y_2\|, \qquad (\text{УП.2.I})$$

где $C_n(t)$ не зависит от y_1 и y_2.

Здесь и ниже, там, где это не может вызвать недоразумений мы не ставим обозначения пространства при знаке нормы.

Рассмотрим свободный вычислительный процесс, состоящий в шении последовательности независимых уравнений

$$A_n(x^{(n)}) = f^{(n)}; \quad n = 0, 1, 2, \ldots, \qquad (\text{УП.2.2})$$

и соответствующий испорченный процесс

$$A_n(x^{(n)}) + \Gamma_n(x^{(n)}) = f^{(n)} + \delta^{(n)}. \qquad (\text{УП.2.3})$$

Допустим, что искажения Γ_n и $\delta^{(n)}$ зависят от параметра $\delta > 0$ так, что

$$\|\delta^{(n)}\| \leq \delta \qquad (\text{УП.2.4})$$

и существуют числовые функции $\varphi_n(t,\delta)$ и $\psi_n(t,\delta)$ положительных числовых переменных t и δ со следующими свойствами: $\lim_{\delta\to 0} \varphi_n(t,\delta) = \lim_{\delta\to 0} \psi_n(t,\delta) = 0$; если x, x', $x'' \in X_n$ и нормы этих элементов не превосходят t, то

$$\|\Gamma_n(x)\| \leq \varphi_n(t,\delta), \qquad (\text{УП.2.5})$$

$$\|\Gamma_n(x') - \Gamma_n(x'')\| \leq \psi_n(t,\delta)\|x' - x''\|. \qquad (\text{УП.2.6})$$

Уравнение (УП.2.8) равносильно следующему

$$x^{(n)} = A_n^{-1}(-\Gamma_n(x^{(n)}) + f^{(n)} + \delta^{(n)}). \qquad (\text{УП.2.7})$$

Зафиксируем n и обозначим $\|f^{(n)}\| = T_n$. Докажем, что оператор в правой части уравнения (УП.2.7) есть оператор сжатия в шаре $S_n = \{x^{(n)} \in X_n : \|x^{(n)}\| \leq 2T_n C_n(T_n)\}$, если только δ достаточно мало. Имеем

$$\|-\Gamma_n(x^{(n)}) + f^{(n)} + \delta^{(n)}\| \leq \varphi_n(2T_n C_n(T_n), \delta) + T_n + \delta,$$

что не превосходит $2T_n$ при достаточно малом δ. Теперь по неравенству (УП.2.1) $\|A_n^{-1}(-\Gamma_n(x^{(n)}) + f^{(n)} + \delta^{(n)})\| \leq 2T_n C_n(T_n)$ и оператор (УП.2.7) отображает шар S_n в себя. Далее, если $x^{(n)} \in S_n$, то $\|x^{(n)}\| \leq 2T_n C_n(T_n)$, и по неравенств и (УП.2.1) и (УП.2.6)

$$\|A_n^{-1}(-\Gamma_n(x^{(n)}) + f^{(n)} + \delta^{(n)}) - A_n^{-1}(-\Gamma_n(z^{(n)}) + f^{(n)} + \delta^{(n)})\| \leq$$

$$\leq C_n(2T_n C_n(T_n))\|\Gamma_n(x^{(n)}) - \Gamma_n(z^{(n)})\| \leq$$

$$\leq C_n(2T_n C_n(T_n))\psi_n(2T_n C_n(T_n), \delta)\|x^{(n)} - z^{(n)}\|;$$

если δ достаточно мало, то множитель при норме $\|z^{(n)}-\zeta^{(n)}\|$ меньше единицы, и наше утверждение доказано. В силу принципа сжатых отображений уравнение (УП.2.3) имеет в шаре S_n ровно одно решение. Если $z_*^{(n)}$ – это решение, то

$$\|z_*^{(n)}-x_*^{(n)}\| = \|A_n^{-1}(-\Gamma_n(z_*^{(n)})+f^{(n)}+\delta^{(n)})-A_n^{-1}(f^{(n)})\| \leq$$

$$\leq C_n(2T_n)\|\Gamma_n(z_*^{(n)})-\delta^{(n)}\| \leq C_n(2T_n)[\varphi_n(2T_nC_n(T_n),\delta)+\delta]. \quad (УП.2.$$

Формула (УП.2.8) дает оценку погрешности искажения для процесса (УП.2.2) в довольно общих условиях; при фиксированном n эта оценка стремится к нулю вместе с δ. Представляют интерес случаи, когда указанное стремление происходит равномерно относительно n. Этим мы займемся в следующем пункте.

2°. В книге автора [8] дано некоторое определение устойчивости сеточного нелинейного вычислительного процесса и получена теорема о достаточных условиях устойчивости процесса Ритца для некоторого класса нелинейных уравнений. Несколько более общее определение устойчивости нелинейного сеточного процесса дал Таккер [1]. Сомодей в своей диссертации [1] доказал некоторые теоремы об устойчивости по Таккеру и применил их для исследования устойчивости МКЭ.

Мы дадим здесь другое определение устойчивости, естественным образом обобщающее определения гл.III.

Допустим, что в некотором шаре $\|x^{(n)}\| \leq \bar{t}$ уравнение (УП.2.2) однозначно разрешимо при любом n. Относительно искаженного процесса (УП.2.3) допустим, что справедливы неравенства (УП.2.4) –6' причем функции φ и ψ_n не зависят от n; таким образом если $x, x', x''-X$, и нормы этих элементов не превосходят числа t, то

$$\|\Gamma_n(x)\| \leq \varphi(t,\delta), \quad (УП.2.9$$

$$\|\Gamma_n(x')-\Gamma_n(x)\| \leq \psi(t,\delta)\|x'-x''\|, \quad (УП.2.1$$

$$\lim_{\delta\to 0}\varphi(t,\delta) = \lim_{\delta\to 0}\psi(t,\delta) = 0. \quad (УП.2.1$$

Назовем процесс (УП.2.2) устойчивым в последовательности пар пространств (X_n, U_n), если для любого $\varepsilon > 0$ существует такое

$\delta > 0$, ч при выполнении неравенств (УП.2.4), (УП.2 3–II) уравнение (УП.2.3) однозначно разреши в некоторой окрестности точки $x_*^{(n)}$, причем радиус этой окрестности не зависит от n, и справедливо неравенство

$$\|z^{(n)} - x^{(n)}\| < \varepsilon. \qquad \text{(УП.2.12)}$$

Теорема УП.2.I Пусть $\|f^{(n)}\| \le T = const$, $A_n(0) = 0$, оператор A_n определен в шаре $\|f^{(n)}\| \le T$ и удовлетворяет в этом шаре условию Липшица

$$\|A_n^{-1}(f_1) - A_n^{-1}(f_2)\| \le C\|f_1 - f_2\| \qquad \text{(УП.2.13)}$$

с постоянной C, которая не зависит от n. Пусть еще искажения $\delta^{(n)}$ и Γ_n удовлетворяют неравенствам (УП. 4), (УП.2.9–II). Тогда процесс (УП.2.2) устойчив в последовательности (U_n Y_n).

Для доказательства достаточно заметить, что в условиях теоремы УП.2.I радиус шара S_n, в котором разрешимо искаженное уравнение (УП.2.3), не зависит от n, а что правая часть формулы (УП.2.8) также не зависит от n и стремится к нулю вместе с δ.

Замечание УП.2.I. Условие $\|f^{(n)}\| \le T$ можно отбросить, если оператор A_n^{-1} определен на всем пространстве Y_n и удовлетворяет условию Липшица (УП.2.I3), в котором C не зависит от n.

§ 3. Об устойчивости процессов Ритца и конечных элементов

I°. В книге автора [8] § 76, доказано, что классический процесс Ритца для задачи $\mathcal{F}(x) = \min$ устойчив, если: I) оператор $P = grad\,\mathcal{F}$ имеет производную Гато P'_u, удовлетворяющую неравенствам (УП.I.4-5); 2) координатная система сильно минимальна в энергетическом пространстве оператора P'_{u_0}. Доказательство проведено в предположении, что искажения Γ_n удовлетворяют требованиям, несколько более ограничительным, чем требования (УП.2.9–I0). Однако, легко провести доказательство, предполагая выполненными только последние требования. Мы не будем останавливаться на этом подробнее; ниже мы докажем более общую теорему, аналогичную теореме Ш.3.2 для линейных задач.

2°. Пусть функционал $\mathcal{F}$ определенный в банаховом пространстве B, обладает следующими свойствами: оператор $P = grad\,\mathcal{F}$ имеет в любой точке u производную Гато P'_u, удовлетворяющую неравенствам (УП.I.4-5); 2) $P(0) = 0$. Поставим задачу о минимуме функционала $\mathcal{F}(x) - (f, x)$, где f - фиксированный эле-

мент пространства B^*. Пусть координатная система имеет вид (УП.1.6). Составим матрицу M_n чисел $[\varphi_{nj}, \varphi_{nk}]$; $j,k = 1,2,\dots,$ где квадратные скобки означают скалярное произведение в B_0 – энергетическом пространстве оператора P_{u_0} (формула (УП.1.5)). Пусть $\lambda_n^{(1)}$ – наименьшее собственное число матрицы M_n, и пусть $\gamma(n)$ – положительная функция натурального числа n, удовлетворяющая неравенству

$$\lambda_1^{(n)} > C\gamma^2(n), \quad C = const. \qquad (УП.3.1)$$

Введем N-мерные гильбертовы пространства $\overline{X}_n$, $\overline{Y}_n$ с нормами

$$\forall b \in R_N, \quad \|b\|_{\overline{X}_N} = \gamma(n)\|b\|_{R_N}, \quad \|b\|_{\overline{Y}_N} = \frac{1}{\gamma(n)}\|b\|_{R_N}. \qquad (УП.3.2)$$

Примем, что координатная система удовлетворяет обычным условиям: 1) элементы $\varphi_{n1}, \varphi_{n2}, \dots, \varphi_{nN}$ образуют базис в натянутом на них подпространстве $B_0^{(n)}$ пространства B_0; 2) последовательность $\{B_0^{(n)}\}$ полна в B_0. Добавим к этому еще одно условие, более ограничительное, чем в линейном случае: 3) $\varphi_{nk} \in \mathcal{D}(P)$. Как всегда в методе Ритца, будем искать приближенное решение задачи о минимуме как элемент подпространства $B_0^{(n)}$:

$$x^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk}.$$

Условие 3) позволяет записать уравнения Ритца в форме Бубнова Галеркина:

$$\left(P\left(\sum_{k=1}^{N} a_k^{(n)} \varphi_{nk}\right), \varphi_{nj}\right) = (f, \varphi_{nj}), \quad 1 \le j \le N. \qquad (УП.3.3)$$

Обозначим $a^{(n)} = (a_1^{(n)}, a_2^{(n)}, \dots, a_N^{(n)})$, $f^{(n)} = ((f, \varphi_{n1}), (f, \varphi_{n2}), (f, \varphi_{nN}))$. Положим еще

$$\left(P\left(\sum_{k=1}^{N} a_k^{(n)} \varphi_{nk}\right), \varphi_{nj}\right) = P_j^{(n)}(a^{(n)})$$

$$P_n(a^{(n)}) = (P_1^{(n)}(a^{(n)}), P_2^{(n)}(a^{(n)}), \dots, P_N^{(n)}(a^{(n)})).$$

Тогда систему Ритца можно записать в виде

$$P_n(a^{(n)}) = f^{(n)}; \qquad (УП.3.4)$$

будем рассматривать $a^{(n)}$ и $f^{(n)}$ как элементы пространства $\overline{X}_n$ и

$\overline{Y}_n$ соответственно.

Составление оператора P_n и вектора $f^{(n)}$ сопряжено с погрешностями вычислений. Пусть в результате вычислений мы вместо P_n и $f^{(n)}$ получили искаженные объекты $P_n + \Gamma_n$ и $f^{(n)} + \delta^{(n)}$. Тогда вместо системы (УП.3.4) мы на самом деле будем решать систему

$$(P_n + \Gamma_n)c^{(n)} = f^{(n)} + \delta^{(n)}. \qquad (УП.3.5)$$

3°. **Теорема УП.3.1.** В условиях данного параграфа процесс (УП.3.3) устойчив в $(\overline{X}_n, \overline{Y}_n)$.

Прежде всего покажем, что нормы $\|a^{(n)}\|_{\overline{X}_n}$ ограничены независимо от n. Так как

$$P(x_*^{(n)}) = P(x_*^{(n)}) - P(0) = \int_0^1 P'_{tx_*^{(n)}} x_*^{(n)} dt,$$

то уравнениям Ритца можно придать вид

$$\int_0^1 (P'_{tx_*^{(n)}} x_*^{(n)}, \varphi_{nj}) dt = (f, \varphi_{nj}), \quad 1 \le j \le N.$$

Умножая это на $a_j^{(n)}$ и складывая, найдем

$$\int_0^1 (P'_{tx_*^{(n)}} x_*^{(n)}, x_*^{(n)}) dt = (f, x_*^{(n)}).$$

Слева заменим подинтегральную функцию по неравенству (УП.1.5) меньшей величиной $d^2 (P'_{u_0} x_*^{(n)}, x_*^{(n)})$, а справа заменим $f, x_*^{(n)}$ большей величиной $\|f\| \cdot \|x_*^{(n)}\|$:

$$d^2 |x_*^{(n)}|^2 \le \|f\| \cdot \|x_*^{(n)}\|,$$

$|\,|$ – норма в B_2. По неравенству (УП.1.4) $\|x_*^{(n)}\| \le \gamma^{-1} |x_*^{(n)}|$, поэтому

$$|x_*^{(n)}| \le \frac{\|f\|}{d^2 \gamma}. \qquad (УП.3.6)$$

Далее,

$$|x_*^{(n)}|^2 = (A_n a^{(n)}, a^{(n)})_{R_N} \ge \lambda_n^{(n)} \|a^{(n)}\|^2_{R_N} \ge$$

$$\ge C \gamma^2(n) \|a^{(n)}\|^2_{R_N} = C \|c^{(n)}\|^2_{\overline{X}_n}$$

и, следовательно,

$$\|x^{(n)}\|_{\overline{X}_n} \le \frac{1}{\sqrt{C}}\,|x_*^{(n)}| \le \frac{\|f\|}{d^2\gamma\sqrt{C}}. \qquad (\text{УП.3.7})$$

4°. Дальнейшее основано на результатах Лангенбаха [I], ко рые мы здесь коротко опишем. Пусть P оператор, вообще говоря нелинейный, определенный на линейном множестве $\mathcal{D}(P)$, плотно в банаховом пространстве B, и действующий из B в B^*. Пусть $P(0)=0$. Предположим, что производная Гато P'_u определен при любом $u\in\mathcal{D}(P)$ и $\mathcal{D}(P'_u)\supset\mathcal{D}(P)$. Допустим далее, что эта производная удовлетворяет неравенствам (УП.I.4-5). Оператор P совпадает на $\mathcal{D}(P)$ с градиентом функционала

$$\mathcal{F}(x)=\int_0^1 (P(tx),x)\,dt$$

и, следовательно, решение (если оно существует) уравнения

$$Px=f,\quad f\in B^* \qquad (\text{УП.3.8})$$

реализует минимум функционала

$$\mathcal{F}(x)-(f,x)=\int_0^1 (P(tx),x)\,dt-(f,x). \qquad (\text{УП.3.9})$$

Эта последняя задача имеет обобщенное решение при любом $f\in B$ (см. § I данной главы); обозначим его через $x=\tilde{P}^{-1}f$. Нетрудно доказать, что различным x соответствуют различные f (см. Лангенбах I); существует, следовательно, обратный к $\tilde{P}^{-1}$ оператор $(\tilde{P}^{-1})^{-1}:=\tilde{P}$, и $\tilde{P}x=f$. Если $x\in\mathcal{D}(P)$, то $Px=f$. Это значит, что $\tilde{P}$ есть расширение оператора P. Доказывается, что $\tilde{P}^{-1}$ удовлетворяет условию Липшица

$$|\tilde{P}^{-1}f_1-\tilde{P}^{-1}f_2|\le\frac{1}{d\gamma}\|f_1-f_2\|_{B^*}. \qquad (\text{УП.3.I0})$$

Рассмотрим теперь уравнение

$$\tilde{P}(z)+\Gamma(z)=f, \qquad (\text{УП.3.II})$$

где Γ удовлетворяет условию

$$\|\Gamma(z_1)-\Gamma(z_2)\|\le\varkappa|z_1-z_2|. \qquad (\text{УП.3.I2})$$

Доказывается, что при $\varkappa<d\gamma$ уравнение (УП.3.I0) имеет единственное решение; если x_* и z_* – решения уравнений (УП.3

τ (УП.3.II) соответственно, то $|z_\varepsilon - x_*|_{\mathcal{Z}} \to 0$.

5°. Вернемся к теореме УП.3.I. Воспользуемся результатами, изложенными в п.4°. С этой целью отождествим $\widetilde{P}$ с P_n, Γ - с Γ_n, B - с $\overline{X}_N$; тогда B^* следует отождествить с $\overline{Y}_n$ и B_0 - с R_N. Далее, отождествим x с $a^{(n)}$, f - с $f^{(n)}$ и $\mathcal{Z}$ с $C^{(n)}$.

Оценим значения величин γ, α, $\varkappa$. Прежде всего,

$$P'_{n,a^{(n)}} b^{(n)} = (P'_{1,a^{(n)}} b^{(n)}, P'_{2,a^{(n)}} b^{(n)}, \ldots, P'_{N,a^{(n)}} b^{(n)}).$$

Имеем

$$P'_{j,a^{(n)}} b^{(n)} = \frac{d}{dt}\left(P(x^{(n)} + t y^{(n)}), y_{nj}\right)\Big|_{t=0};$$

здесь

$$x^{(n)} = \sum_{k=1}^{N} a_k^{(n)} y_{nk}, \quad y^{(n)} = \sum_{k=1}^{N} b_k^{(n)} y_{nk}.$$

Далее

$$\frac{d}{dt}\left(P(x^{(n)} + t y^{(n)}), y_{nj}\right)\Big|_{t=0} =$$

$$= \left(\frac{d}{dt} P(x^{(n)} + t y^{(n)})\Big|_{t=0}, y_{nj}\right) = \left(\overset{\circ}{P}'_{x^{(n)}} y^{(n)}, y_{nj}\right).$$

Умножая это на $b_j^{(n)}$ и складывая, получим

$$\left(P'_{n,a^{(n)}} b^{(n)}, b^{(n)}\right)_{R_N} = \left(P'_{x^{(n)}} y^{(n)}, y^{(n)}\right). \qquad (\text{УП.3.I3})$$

По неравенству (УП.I.5)

$$\left(P'_{n,a^{(n)}} b^{(n)}, b^{(n)}\right)_{R_N} \geqslant \alpha^2 \left(P'_{u_0} y^{(n)}, y^{(n)}\right) = \alpha^2 |y^{(n)}|^2. \qquad (\text{УП.3.I4})$$

Воспользуемся теперь неравенством (УП.3.7):

$$\left(P'_{n,a^{(n)}} b^{(n)}, b^{(n)}\right)_{R_N} \geqslant C\alpha^2 \|b^{(n)}\|^2_{\overline{X}_N}. \qquad (\text{УП.3.I5})$$

Неравенство (УП.3.I4) является аналогом неравенства (УП.I.4); отсюда видно, что в утверждениях п.4° можно, применительно к нашему случаю, заменить γ на $\alpha\sqrt{C}$; важно отметить, что последняя величина не зависит от n. Далее, по формулам (УП.3.I3-I4)

$$(P'_{n,a^{(n)}} b^{(n)}, b^{(n)})_{R_N} = (P'_{x^{(n)}} y^{(n)}, y^{(n)}) \geqslant$$

$$\geqslant d^2 (P'_0 y^{(n)}, y^{(n)}) = d^2 (P'_{n,0} b^{(n)}, b^{(n)}).$$

(УП.3.16)

Мы приняли здесь $u_0 = 0$ что, очевидно, не уменьшает общности. Неравенство (УП.3.16) есть аналог неравенства (УП.1.5) с той же постоянной d.

Из результатов, перечисленных в п.4°, вытекает теперь, что операторы $P'_{n,a^{(n)}}$ удовлетворяют условию Липшица с постоянной C, которая не зависит от n и от $f^{(n)}$.

Рассмотрим теперь уравнение Гитца (УП.3.4), искаженное уравнение Ритца (УП.3.5) и вспомогательное уравнение

$$P_n g^{(n)} = f^{(n)} + \delta^{(n)}. \qquad (УП.3.17)$$

Так как P_n^{-1} удовлетворяет условию Липшица с постоянной C, то

$$\| g^{(n)} - a^{(n)} \|_{\bar{X}_N} \leqslant C \| \delta^{(n)} \|_{\bar{Y}_N}, \quad \delta \to 0. \qquad (УП.3.18)$$

Далее, Γ_n удовлетворяет условию Липшица (УП.3.12), в котором следует положить $\varkappa = \psi(t, \delta)$. Как было доказано в п.3°, нормы $\| a^{(n)} \|_{\bar{X}_N}$ ограничены, поэтому можно считать число t фиксированным; в силу сказанного в конце п.4°, $\| c^{(n)} - g^{(n)} \|_{\bar{X}_N} \xrightarrow{\delta \to 0} 0$. Вместе с неравенством (УП.3.18) это дает

$$\| c^{(n)} - a^{(n)} \|_{\bar{X}_N} \xrightarrow{\delta \to 0} 0 \qquad (УП.3.19)$$

а это означает, что вычислительный процесс (УП.3.3) устойчив в последовательности $(\bar{X}_N, \bar{Y}_N)$.

Замечание УП.3.3. Если координатная система сильно минимальна в энергетической норме оператора P'_0, то можно положить $\gamma(n) = 1$. Отсюда следует, что в рассматриваемом случае процесс (УП.3.3) устойчив в последовательности (R_N, R_N).

6°. Теорема УП.3.2. Если координатная система почти ортонормирована в B_0, то процесс вычисления приближенного решения по Ритцу в условиях настоящего параграфа устойчив в последова-

тельности ($B_0^{(n)}$, R_N).

Доказательство очень просто. Если $z_*^{(n)}$ – искаженное приближенное решение по Ритцу, то

$$|z_*^{(n)} - x_*^{(n)}| = (M_n(c^{(n)} - a^{(n)}), c^{(n)} - a^{(n)})_{R_N} \leq$$

$$\leq \Lambda_0 \|c^{(n)} - a^{(n)}\|^2_{R_N},$$

где Λ_0 – верхняя граница собственных чисел матрицы M_n. Почти ортонормированная система сильно минимальна, и процесс вычисления коэффициентов Ритца устойчив в (R_N, R_N), и при достаточно малом δ будет $\|c^{(n)} - a^{(n)}\|_{R_N} \leq \varepsilon$. Но тогда $|z_*^{(n)} - x_*^{(n)}| \leq \sqrt{\Lambda_0}\delta$, и теорема доказана.

7°. Обратимся к МКЭ: для определенности остановимся на варианте этого метода, развитом в книге автора [22]. В этом случае система координатных функций не сильно минимальна в энергетической норме: в частности, в случае первой краевой задачи, а также в случае кубической области и более произвольных краевых условий, наименьшее собственное число матрицы Грама элементов (УП.1.6) имеет порядок $O(h^m)$, где h – шаг сетки и m размерность евклидова пространства задачи (см. [22], гл.IX, §§ 2,3[1]).

Справедлива теорема, которая является частным случаем теоремы УП.3.1:

Теорема УП.3.3. Пусть достаточно гладкий функционал $\mathcal{F}(x)$ определен на функциях пространства $B = \mathring{W}_p^{(s)}(\Omega)$, где Ω – ограниченная область в R_m. Пусть операторы $P = grad\,\mathcal{F}$ и P'_u определены на множестве, плотном в B, причем P'_u удовлетворяет неравенствам (УП.1.4-5). Пусть $h = 1/2n$ – шаг сетки, а $N = N(n)$ – число координатных функций, входящих в выражение приближенного решения. Вычислительный процесс МКЭ устойчив в последовательности пар N-мерных гильбертовых пространств (X_N, Y_N) с нормами

$$\forall a^{(n)} \in R_N,\ \|a^{(n)}\|_{X_N} = h^{m/2}\|a^{(n)}\|_{R_N},\ \|a^{(n)}\|_{Y_N} = h^{-m/2}\|a^{(n)}\|_{R_N}.$$

Аналогичная теорема верна и для задачи с естественными краевыми условиями, если Ω есть куб в R_m.

[1] По недосмотру автора в этом месте напечатано h^{-m} вместо h^m.

§ 4. Погрешности искажения и округления рекуррентного вычислительного процесса

I°. Рассмотрим нелинейный рекуррентный процесс

$$A_n(x^{(0)}, x^{(1)}, \dots, x^{(n)}) = f^{(n)};\ n = 0, 1, 2, \dots, \quad (\text{УП.4.I})$$

где $x^{(k)} \in X_k$, $f^{(n)} \in Y_n$, а X_n, Y_n – банаховы пространства. Искаженный процесс запишем в виде

$$A_n(z^{(0)}, z^{(1)}, \dots, z^{(n)}) + \Gamma_n(z^{(0)}, z^{(1)}, \dots, z^{(n)}) = f^{(n)} + \delta^{(n)};\ n = 0, 1, 2, \dots \quad (\text{УП.4.2})$$

Примем следующие допущения: I) если элементы $x^{(0)}, x^{(1)}, \dots, x^{(n)}$ фиксированы и их нормы не превосходят некоторого числа $\bar{t}$, то n-е уравнение (УП.4.I) имеет не более одного решения в шаре $\|x^{(n)}\| \le \bar{t}$; 2) существует такое число $\bar{\Gamma} = const$, что если $\|f^{(n)}\| \le \bar{t}$ при любом n, то бесконечная система (УП.4.I) разрешима и $\|x^{(n)}\| \le \bar{\Gamma}$.

Нетрудно указать свободный вычислительный процесс, эквивалентный рекуррентному процессу (УП.4.I). Построим две новые последовательности банаховых пространств $\bar{X}_n$ и $\bar{Y}_n$, $n = 0, 1, 2, \dots$. Элементами пространства $\bar{X}_n$ являются конечные последовательности $\bar{x}^{(n)} = (x^{(0)}, x^{(1)}, \dots, x^{(n)})$, $x^{(k)} \in X_k$; норму в $\bar{X}_n$ определим формулой

$$\|\bar{x}^{(n)}\|_{\bar{X}_n} = \max_{k \le n} \|x^{(k)}\|_{X_k} \quad (\text{УП.4.3})$$

(допустимы и другие определения нормы); аналогично определяется $\bar{Y}_n$. Определим еще оператор $\bar{A}_n$, действующий из $\bar{X}_n$ в $\bar{Y}_n$ по формуле

$$\bar{A}_n(\bar{x}^{(n)}) = \{A_k(x^{(0)}, \dots, x^{(k)})\}_{k=0}^{n}; \quad (\text{УП.4.4})$$

аналогично определим оператор $\bar{\Gamma}_n$. Уравнения (УП.4.I) и (УП.4.2) соответственно равносильны уравнениям свободных вычислительных процессов

$$\forall n \ge 0,\ \bar{A}_n(\bar{x}^{(n)}) = \bar{f}^{(n)}; \quad (\text{УП.4.5})$$

$$\forall n \ge 0,\ \bar{A}_n(\bar{z}^{(n)}) + \bar{\Gamma}_n(\bar{z}^{(n)}) = \bar{f}^{(n)} + \bar{\delta}^{(n)}. \quad (\text{УП.4.6})$$

Мы дадим здесь определение устойчивости нелинейного рекуррентного процесса, отличное от аналогичного определения для линейного процесса, данного в гл.III. Рекуррентный процесс (УП.4.1) назовем устойчивым, если устойчив соответствующий свободный процесс (УП.4.5).

2^0. Условия устойчивости рекуррентного процесса получаются из результатов § 2. Эти условия таковы: 1) $A_n(0)=0$; 2) оператор $\bar{A}_n^{-1}$ определен в шаре $\|\bar{f}^{(n)}\|\le T$ и удовлетворяет условию Липшица $\|\bar{A}_n^{-1}(\bar{f}_1)-\bar{A}_n^{-1}(\bar{f}_2)\| \le C\|\bar{f}_1-\bar{f}_2\|$ с постоянной, не зависящей от n ; 3) $\|\bar{\delta}^{(n)}\|\le\delta$; 4) если $\bar{x}, \bar{x}', \bar{x}''\in\bar{X}_n$ и нормы этих элементов не превосходят t , то

$$\|\bar{\Gamma}_n(\bar{x})\|\le\varphi(t,\delta);\ \|\bar{\Gamma}_n(\bar{x})-\bar{\Gamma}_n(\bar{x}'')\|\le\varphi(t,\delta)\|\bar{x}'-\bar{x}''\|.$$

Переформулируем эти условия в терминах операторов A_n и Γ_n. Условие 1) означает, что

$$\forall n>0,\quad A_n(0,0,\dots,0)=0. \qquad (\text{УП.4.7})$$

Далее, $\bar{A}_n^{-1}$ есть оператор, определяющий решение конечной системы

$$A_\kappa(x^{(0)}, x^{(1)},\dots,x^{(\kappa)})=f^{(\kappa)},\ \kappa=0,1,\dots,n.$$

Это решение имеет вид

$$\bar{x}^{(\kappa)}=B_\kappa(f^{(0)}, f^{(1)},\dots,f^{(\kappa)}),\ \kappa=0,1,\dots,n.$$

Отсюда следует, что $\bar{A}_n^{-1}=\{B_0, B_1,\dots,B_n\}$, и условие 2) принимает вид

$$\forall n\ge0,\ \|B_n(f_1^{(0)}, f_1^{(1)},\dots,f_1^{(n)})-B_n(f_2^{(0)}, f_2^{(1)},\dots,f_2^{(n)})\|<$$

$$<C\max_\kappa\|f_1^{(n)}-f_2^{(n)}\|. \qquad (\text{УП.4.8})$$

Условие 3) означает, что

$$\|\delta^{(n)}\|<\delta. \qquad (\text{УП.4.9})$$

Наконец, условия 4) сводятся к следующим:

$$\|\Gamma_n(x^{(0)}, x^{(1)}\dots,x^{(n)})\|\le\varphi(t,\delta) \qquad (\text{УП.4.10а})$$

$$\|\Gamma_n(x_1^{(0)}, x_1^{(1)}, \dots, x_1^{(n)}) - \Gamma_n(x_2^{(0)}, x_2^{(1)}, \dots, x_2^{(n)})\| \le$$

$$\le \psi(t,0) \max_{0\le k\le n} \|x_1^{(k)} - x_2^{(k)}\|. \qquad (\text{УП.4.}$$

На основании теоремы УП.2.I можно утверждать, что нелинейный рекуррентный процесс (УП.4.I) устойчив в смысле определения настоящего параграфа, если $\|f^{(n)}\| \le T = const$ и выполнены условия (УП.4.7–I0).

Если процесс (УП.4.I) устойчив, то при малых δ погрешности искажения ограничена независимо от n.

3°. Отметим случай, когда погрешность округления рекуррентного нелинейного процесса также остается ограниченной. Обозначим $A_n + \Gamma_n = \tilde{A}_n$, $f^{(n)} + \delta^{(n)} = \tilde{f}^{(n)}$, и пусть точное решение искаженной системы (УП.4.2) имеет вид

$$\forall n \ge 0, \quad \tilde{z}^{(n)} = \tilde{B}_n(\tilde{z}^{(0)}, \tilde{z}^{(1)}, \dots, \tilde{z}^{(n-1)}, \tilde{f}^{(n)}). \qquad (\text{УП.4.}$$

Допустим, что операторы $\tilde{B}_n$ удовлетворяют условию Липшица

$$\|\tilde{B}_n(\tilde{z}_1^{(0)}, \tilde{z}_1^{(1)}, \dots, \tilde{z}_1^{(n-1)}, \tilde{f}^{(n)}) - \tilde{B}_n(\tilde{z}_2^{(0)}, \tilde{z}_2^{(1)}, \dots, \tilde{z}_2^{(n-1)}, \tilde{f}^{(n)})\| \le$$

$$\le q \max_{0\le k\le n} \|\tilde{z}_1^{(k)} - \tilde{z}_2^{(k)}\|, \quad q = const < 1. \qquad (\text{УП.4.}$$

Пусть $v^{(0)}, v^{(1)}, \dots, v^{(n-1)}$ – точно известные элементы, $\tilde{z}^{(n)}$ есть точное значение выражения $\tilde{B}_n(v^{(0)}, v^{(1)}, \dots, v^{(k-1)}, \dots)$. Допустим, что в результате погрешностей вычислений мы вместо $\tilde{z}^{(n)}$ получим элемент $v^{(n)} = \tilde{z}^{(n)} + d^{(n)}$, будем считать, что вычислительная погрешность округления ограничена числом $\varepsilon > 0$, так что $\|d^{(n)}\| \le \varepsilon \|v^{(n)}\|$. По неравенству (УП.4.I2) имеем

$$\forall n \ge 0, \quad \|v^{(n)} - \tilde{z}^{(n)}\| \le q \max_{0\le k\le n-1} \|v^{(k)} - \tilde{z}^{(k)}\| + \varepsilon \|v^{(n)}\|.$$

Допустим, что $\|\tilde{z}^{(n)}\| \le const$, получаем из последнего неравенства

$$\hat{z}_n := \|v^{(n)} - \tilde{z}^{(n)}\| \le q' \max_{0\le k\le n-1} \hat{z}_k + C\varepsilon, \quad q = \frac{q}{\varepsilon}. \qquad (\text{УП.4.I3}$$

Пусть $\max_{0\le k\le n}\hat{\xi}^{(k)}=\hat{\xi}^{(l)}$, $0\le l\le n$. Неравенство (УП.4.13) дает при $n=l$ неравенство $\hat{\xi}_l\le q'\hat{\xi}_l+C\varepsilon$ или, если $q'<1$, $\hat{\xi}_l\le C\varepsilon/(1-q')$. Отсюда следует искомая оценка погрешности округления

$$\hat{\xi}_n\le\frac{C\varepsilon}{1-q'}. \qquad (\text{УП.4.14})$$

§ 5. Метод Ньютона - Канторовича

1°. В настоящем параграфе мы исследуем погрешности искажения и округления двух методов Ньютона - Канторовича: модифицированного и основного. Эти методы приводят к рекуррентным процессам более специального вида, нежели рассмотренные в § 4. Устойчивость таких процессов можно исследовать проще и в более общих предположениях. С рассмотрения процессов такого вида мы и начнем.

Пусть задан рекуррентный процесс вида

$$A_n x^{(n)}+Q_n(x^{(n-1)})=f^{(n)}, \qquad (\text{УП.5.1})$$

где A_n линейные, а Q_n – нелинейные операторы. Для простоты допустим что все эти операторы действуют в одном и том же банаховом пространстве X. Искаженный процесс запишем в виде

$$A_n\tilde{x}^{(n)}+\Gamma_{1n}\tilde{x}^{(n)}+Q(\tilde{x}^{(n-1)})+\Gamma_{2n}(\tilde{x}^{(n-1)})=f^{(n)}+\delta^{(n)}. \qquad (\text{УП.5.2})$$

Примем следующие допущения. 1) Операторы A_n^{-1} равномерно ограничены: $\|A_n^{-1}\|\le\alpha=const$. 2) При любом n оператор $A_n^{-1}Q_n$ есть оператор сжатия, так что

$$\|A_n^{-1}Q_n(x)-A_n^{-1}Q_n(y)\|\le q\|x-y\|,\ q=const<1. \qquad (\text{УП.5.3})$$

3) Справедливы следующие неравенства:

$$\|\delta^{(n)}\|\le\delta; \qquad (\text{УП.5.4})$$

если $\|x\|,\ \|x'\|,\ \|x''\|\le t$, то ($k=1,2$)

$$\|A_n^{-1}\Gamma_{kn}(x)\|\le\varphi(t,\delta), \qquad (\text{УП.5.5})$$

$$\|A_n^{-1}\Gamma_{kn}(x')-A_n^{-1}\Gamma_{kn}(x'')\|\le\psi(t,\delta)\|x'-x''\|; \qquad (\text{УП.5.6})$$

Функции φ и Ψ обладают теми же свойствами, что и в предшествующих параграфах.

Имеем

$$x^{(n)}=A_n^{-1}f^{(n)}-A_n^{-1}Q_n(x^{(n-1)}),$$

$$z^{(n)}=A_n^{-1}f^{(n)}+A_n^{-1}\delta^{(n)}-A_n^{-1}Q_n(z^{(n-1)})-$$

$$-A_n^{-1}\Gamma_{1n}z^{(n)}-A_n^{-1}\Gamma_{2n}(z^{(n)}).$$

Отсюда

$$z^{(n)}-x^{(n)}=$$

$$=A_n^{-1}\delta^{(n)}-[A_n^{-1}Q_n(z^{(n-1)})-A_n^{-1}Q_n(x^{(n-1)})]-A_n^{-1}\Gamma_{1n}(z^{(n)})-A_n^{-1}\Gamma_{2n}(z^{(n-1)}).$$

Как и в предшествующих параграфах, доказывается, что если нормы $\|x^{(n)}\|$ ограничены в совокупности, $\|x^{(n)}\|\leqslant t$, то уравнение (УП.5.2) имеет в шаре $\|z\|\leqslant 2t$ одно и только одно решение. Оценим разность $z^{(n)}-x^{(n)}$, где $x^{(n)}$ и $z^{(n)}$ – решения уравнений (УП.5.1) и (УП.5.2) соответственно. Имеем

$$\|z^{(n)}-x^{(n)}\|\leqslant q\|z^{(n-1)}-x^{(n-1)}\|+a\delta+2\varphi(2t,\delta),$$

или, обозначая $\|z^{(n)}-x^{(n)}\|=\hat{\xi}_n$, $a\delta+2\varphi(2t,\delta)=q$,

$$\hat{\xi}_n\leqslant q\hat{\xi}_{n-1}+q.$$

Ясно, что $\hat{\xi}_n\leqslant\tau_n$, где τ_n – решение системы $\tau_0=\hat{\xi}_0$, $\tau_n=q\tau_{n-1}+q$, $n=1,2,\ldots$. Это решение есть $\tau_n=q^n\hat{\xi}_0+q(1-q)^{-1}$ и, следовательно

$$\|z^{(n)}-x^{(n)}\|\leqslant q^n\|z^{(0)}-x^{(0)}\|+(1-q)^{-1}[a\delta+2\varphi(2t,\delta)]. \qquad (УП.5.7)$$

Неравенство (УП.5.7) дает оценку погрешности искажения для процесса (УП.5.1); из этого неравенства вытекает также устойчи-

вость этого процесса при условиях (УП.5.3–6).

Замечание УП.5.1. Джикариани [8, 10, 11] рассмотрел частный случай процесса (УП.5.1), который получается применением классического метода Ритца к уравнению $Ax+Qx=f$, где A – самосопряженный положительно определенный оператор с дискретным спектром, действующий в сепарабельном гильбертовом пространстве H, а Q – нелинейный непрерывный потенциальный оператор, имеющий положительно определенный дифференциал Фреше. Предполагается, что Q определен на всем пространстве H и $Q(0)=0$. Наконец, $f\in H$. Принимаем, что рассматриваемое уравнение имеет единственное решение.

В качестве координатных элементов выбираются собственные элементы оператора B, сходного (см. книгу автора [8]) с A. Доказывается, что невязка приближенного решения по Ритцу, $x_*^{(n)}$, стремится к нулю:

$$\varepsilon^{(n)} := \|Ax_*^{(n)}-Qx_*^{(n)}-f\|_n \longrightarrow 0.$$

Доказывается, что в метрике H погрешность аппроксимации имеет оценку (ω_{n+1} есть $(n+1)$-е собственное число оператора B)

$$\|x_*-x_*^{(n)}\|_H = O\left(\frac{\|\delta^{(n)}\|}{\omega_{n+1}}\right),$$

а в энергетической норме оператора A – оценку

$$|x_*-x_*^{(n)}| = O\left(\frac{\|\delta^{(n)}\|}{\sqrt{\omega_{n+1}}}\right).$$

Устойчивость процесса доказана в том же предположении, что координатные элементы суть собственные элементы оператора, соответствующие

2^0. Перехожим к ошибкам округления. Обозначим $A_n+\Gamma_{1n}=\tilde{A}_n$, $Q_n+\Gamma_{2n}=\tilde{Q}_n$ – оба эти оператора точно известны, а оператора $\tilde{A}_n^{-1}$ существует. Допустим, что оператор $\tilde{A}_n^{-1}\tilde{Q}_n$ есть оператор сжатия в некотором шаре, содержащем $z^{(n)}$, с коэффициентом сжатия $\tilde{q}_n$, не зависящим от n. Уравнение (УП.5.2) можно привести к виду

$$z^{(n)} = a_n(z^{(n-1)}) + b^{(n)},$$

$$a_n = \tilde{A}_n^{-1}\tilde{Q}_n, \quad b^{(n)} = \tilde{A}_n^{-1}(\ldots) + \tilde{\varepsilon}^{(n)}. \qquad (\ldots)$$

Уравнение (УП.5.8) разрешимо итерациями, которые можно записать так:

$$m=0,1,2,\dots;\quad y^{(m)}=a_n(y^{(m-1)})+b^{(n)}, \qquad (УП.5.9)$$

где $y^{(m)}$ есть m-е приближение к решению $z^{(n)}$. Допустим, что, вычисляя правую часть формулы (УП.5.9), мы совершаем ошибку округления, относительная величина которой не превосходит числа $\varepsilon>0$. Пусть, далее, известное нам округленное значение $(m-1)$-го приближения $y^{(m-1)}$ есть $v^{(m-1)}$. Тогда в результате вычислений мы вместо элемента $y^{(m)}$ получим некоторый элемент $v^{(m)}$, удовлетворяющий соотношению

$$v^{(m)}=a_n(v^{(m-1)})+b^{(n)}+d^{(n)}, \qquad (УП.5.10)$$

$$\|d^{(m)}\|\le\varepsilon\|v^{(m)}\|.$$

Вычитая отсюда равенство (УП.5.9), получаем

$$v^{(m)}-y^{(m)}=a_n(v^{(m-1)})-a_n(y^{(m-1)})+d^{(m)}$$

и

$$\gamma_m:=\|v^{(m)}-y^{(m)}\|\le\tilde{q}_1\gamma_{m-1}+\varepsilon\gamma_m+\varepsilon\|y^{(m-1)}\|.$$

Допустим, что нормы $\|b^{(n)}\|$ ограничены в совокупности: $\|b^{(n)}\|\le C$. Тогда нормы $\|y^{(m)}\|$ также ограничены; пусть $\|y^{(m-1)}\|\le c_0$. В этом случае

$$\gamma_m\le\frac{\tilde{q}}{1-\varepsilon}\gamma_{m-1}+\frac{c_0\varepsilon}{1-\varepsilon}. \qquad (УП.5.11)$$

Обозначим $\tilde{q}/(1-\varepsilon)=\tau$, $c_0\varepsilon/(1-\varepsilon)=\lambda$, так что $\gamma_m\le\tau\gamma_{m-1}+\lambda$. Заметим, что $v^{(0)}=y^{(0)}$; так как $y^{(0)}$ – начальное значение – не вычисляется, а задается, $\gamma_0=0$. Теперь из (УП.5.11) следует

$$\gamma_m\le\frac{1-\tau^m}{1-\tau}\frac{c_0\varepsilon}{1-\varepsilon}. \qquad (УП.5.12)$$

Если ε столь мало, что $\tau<1$, то оценка (УП.5.12) дает далее

$$\gamma_m\le\frac{c_0\varepsilon}{(1-\tau)(1-\varepsilon)}; \qquad (УП.5.13)$$

ошибка округления ограничена независимо от m и n и равномерно

стремится по этим индексам к нулю при $\varepsilon \to 0$. Если ε не очень мало, так что $\nu > 1$, то оценка (УП.5.12) стремится к бесконечности вместе с m.

3°. Модифицированный метод Ньютона – Канторовича применяется к решению нелинейных задач вида $P(x) = 0$. Достаточно полное описание этого метода, условий и скорость его сходимости дены в книге Канторовича и Акилова [1]; здесь мы отметим только, что самый метод состоит в применении рекуррентного процесса

$$x^{(n)} = x^{(n-1)} - A^{-1} P(x^{(n-1)}); \quad A = P'_{x^{(0)}}; \qquad (УП.5.14)$$

$x^{(0)}$ – начальное приближение к искомому решению. Нам будет удобнее записать этот процесс в виде

$$Ax^{(n)} + Q(x^{(n-1)}) = 0; \quad Q = A - P. \qquad (УП.5.15)$$

Заметим, что оператор A не задан, а вычисляется, естественно, с некоторой погрешностью.

Проверим выполнение условий п.1°. Условие $\|A_n^{-1}\| \leqslant const$ вытекает из того, что оператор $A_n = A = P'_{x^{(0)}}$ не зависит от n, а что ограниченность оператора $[P'_{x^{(0)}}]$ принадлежит к числу условий сходимости модифицированного метода Ньютона – Канторовича. Условие (УП.5.3) будет выполнено, если P имеет непрерывную первую производную; в теореме о сходимости метода требуется существование второй производной. Мы предполагаем также, что выполнены условия (УП.5.4-6), характеризующие искажение процесса. По показанному в пп.1°, 2°, вычислительный процесс, связанный с модифицированным методом Ньютона – Канторовича, устойчив; если уравнения модифицированного процесса решать и вручную с достаточно малыми относительными погрешностями, то погрешность округления остается ограниченной.

4°. В книге Канторовича и Акилова [1] доказано следующее утверждение. Пусть $x^{(0)}$ – начальное приближение, а $x^{(1)}, x^{(2)}, \ldots, x^{(n)}, \ldots$ – приближения, построенные по основному методу Ньютона – Канторовича. Если в точке $x^{(0)}$ выполнены условия сходимости модифицированного метода Ньютона – Канторовича, то они выполнены и в точках $x^{(1)}$ $x^{(2)}, \ldots$ Отсюда нетрудно получить, что для вычислительного процесса, связанного с основным методом Ньютона – Канторовича, справедливы заключения, сделанные в конце п.3° для модифицированного метода.

ГЛАВА VIII

Некоторые односторонние вариационные задачи.

В обширной литературе по задачам, связанным с вариационными неравенствами, видное место занимают так называемые односторонние вариационные задачи. Обычная постановка этих задач такова: на рефлексивном банаховом пространстве B задан полунепрерывный снизу возрастающий функционал $J(x)$ и замкнутое непустое выпуклое множество M. Ставится задача: найти такой элемент $x_* \in M$, чтобы

$$\forall x \in M,\ J(x_*) \leqslant J(x).$$

Условие $x \in M$ называется односторонним ограничением, а множество M называется множеством ограничений.

В книге Лионса [1] доказано, что такая задача имеет решение оно единственно, если функционал J – строго выпуклый. В книге Дюво и Лионса [1] рассмотрен ряд задач физики и механики, которые приводятся к вариационным неравенствам и, в частности, к односторонним вариационным задачам. В книге Гловинского, Лионса и Тремольера [1] излагается ряд методов приближенного решения таких задач и даны приложения этих методов к большому числу более конкретных задач. Для многих приближенных методов даны доказательства сходимости. Однако, по мнению авторов этой книги, – получение оценок погрешности приближенных методов – задача более деликатная.

В настоящей главе будут изложены результаты работ автора [26-30] по оценке погрешностей для односторонних вариационных задач; в основном речь будет идти о некотором классе таких задач, подробное описание которого дано ниже, в § 1. Будут изложены также (в § 4), но без доказательств, некоторые результаты работ Демьяновича [4] и Тихтина [1]. В конце главы будет дан обзор работ последнего времени по оценкам погрешности аппроксимации для вариационных неравенств; этот обзор написан Тихтиным.

§ 1. Постановка задачи и ее приближенное решение.

1°. Пусть требуется решить задачу о минимуме функционала

$$J(x) = \|x\| \qquad (\forall x \in B) \qquad (1.1)$$

при одностороннем ограничении

$$\|x - g\| \geqslant d = \text{const}. \qquad (1.2)$$

где $\mathcal{F}$ - квадратичная форма, определенная и положительная на некотором линейном множестве. Прежде чем переходить к описанию остальных данных задачи, заметим, что на упомянутом множестве можно построить энергетическое пространство H со скалярным произведением $[u,v]=\mathcal{F}(u,v)$ и нормой $|u| = [\mathcal{F}(u)]^{1/2}$; здесь $\mathcal{F}(u,v)$ - билинейная форма, соответствующая квадратичной форме $\mathcal{F}(u)$. Примем, что l - линейный функционал, ограниченный в энергетической норме, и что $\|\cdot\|$ - полунорма, определенная (но необязательно всюду конечная) на всем пространстве H.

Введем еще в рассмотрение гильбертово пространство $\tilde{H}$, по отношению к которому H является подпространством (может быть, несобственным). Скалярное произведение и норму в $\tilde{H}$ будем обозначать теми же символами, что и в H. Будем считать, что g - заданный элемент пространства $\tilde{H}$ и что полунорма $\|\cdot\|$ определена (хотя, разумеется, необязательно всюду конечна) на всех элементах этого пространства. Необходимо потребовать, чтобы множество (УШ.I.2) не было пустым - для этого необходимо и достаточно, чтобы $\|g\|<\infty$ и чтобы расстояние $\rho(g,H)$, вычисленное в метрике $\|\cdot\|$, не превосходило числа d.

2^0. Последний вопрос рассмотрим подробнее. Вычислим величину $\varkappa = \rho(g,H)$. Если $\varkappa>d$, то, как только-что было отмечено, множество

$$M=\{x\in H:\|x-g\|\leqslant d\} \qquad \text{(УШ.I.3)}$$

пусто, и задача (УШ.I.I-2) не имеет решения. Пусть $\varkappa<d$, тогда существует такой элемент $x'\in H$, что $\|x'-g\|<d$, и множество M не пусто. Положим $g-x'=h$, $x-x'=y$, тогда множество ограничений определяется соотношением $\|y-h\|\leqslant d$, в котором $\|h\|<d$.

Пусть теперь $\varkappa=d$. Может случиться, что $\inf\limits_{x\in H}\|x-g\|$ не достигается, тогда $\forall x\in H$, $\|x-g\|>d$, и множество M пусто. Если же инфимум достигается, то множество M не пусто, и задача (УШ.I.I-2) имеет решение (см. ниже, п.3^0). Однако, это решение неустойчиво относительно малых изменений d. Действительно, если заменить d на $d'<d$, то $\varkappa=\rho>d'$, новое множество

$$M'=\{x\in H,\|x-g\|<d'\}$$

пусто, и новая односторонняя вариационная задача не имеет
шения.

Ниже, рассматривая множество вида (УШ.I.3), мы всегда с
считать, что $\beta := \|g\| < d$.

3°. По теореме Рисса существует единственный элемент x_*
удовлетворяющий тождеству $\forall x \in H$, $\ell x = [x, \omega_0]$; этот элем
решающий задачу о минимуме функционала (УШ.I.I) при отсутст
ограничений, будем считать известным.

Задача (УШ.I.I-2) равносильна следующей:

$$|x - x_0| = \min, \quad \|x - g\| \le d. \qquad (\text{УШ.I}$$

Множество ограничений M, определяемое неравенством (УШ
(или (УШ.I.3)) - выпуклое; мы принимаем полунорму $\|\cdot\|$ так
что это множество замкнуто в метрике H . В силу результатов
изложенных в книге Лионса [I], гл.I, задача (УШ.I.4) имеет
и только одно решение, которое мы обозначим через x_* . Если
$\|x_0 - g\| \le d$, то, очевидно, $x_* = x_0$; если же $\|x_0 - g\| > d$
x_* решает задачу

$$|x_* - x_0| = \min, \quad \|x_* - g\| = d. \qquad (\text{УШ.I}$$

Докажем это. Рассмотрим прямолинейный отрезок Λ , соединяющ
точки x_0 и x_* : $\Lambda = \{x \in H : x = \lambda x_* + (1-\lambda) x_0 \; ; 0 \le \lambda \le 1\}$. Установим
Λ направление от x_* к x_0 и в соответствии с этим назовем x_*
левым, а x_0 - правым концом отрезка Λ . Ясно, что левый конец
отрезка Λ принадлежит множеству M, правый - множеству $H \setminus$
Введем на Λ расстояние, порожденное нормой $|\cdot|$, и разделим
отрезок пополам. Один из двух отрезков деления таков, что ег
левый конец принадлежит M , а правый принадлежит $H \setminus M$. Э
отрезок разделим пополам, и т.д. В результате получится посл
довательность вложенных отрезков с длинами, стремящимися к
лю, и обладающих тем свойством, что их левые концы принадлеж
M , а правые - $H \setminus M$. Эти отрезки содержат ровно одну общую
точку; обозначим ее через y. Она - предельная для множеств
так как M замкнуто, то $y \in M$, и $\|y - g\| \le d$. Одновременно точка
предельная для $M \setminus H$: если $\{y_k\}$ - последовательность правых
концов рассмотренных нами вложенных отрезков, то $|y_n - y| \to 0$,
одномерном отрезке полунорма $\|\cdot\|$ непрерывна, поэтому $\|y - y_n\|$
но $\|g - y_n\| > d$ к, следовательно, $\|g - y\| \ge d$. Окончательно

$|||y-g|||=d$.

Допустим теперь, что $|||x_*-g|||<d$. Тогда на отрезке Λ точка y – промежуточная между x_* и x_0, поэтому $|x_0-y|<|x_0-x_*|=\min\limits_{x\in M}|x_0-x|$, а это нелепо. Отсюда следует, что $|||x_*-g|||=d$, и наше утверждение доказано.

Множество $\{x\in H: |||x-g|||=d\}$ назовем границей множества M и обозначим через ∂M. Можно, следовательно, считать, что при $|||x_0-g|||>d$ элемент x_* решает задачу

$$|x-x_0|=\min,\quad x\in\partial M. \qquad (\text{УШ.I.6})$$

Ниже всюду принимается, что $|||x_0-g|||>d$.

4°. Задачу (УШ.I.6) будем решать приближенно, сводя ее к конечномерной задаче. Выберем натуральное число n и N-мерное ($N=N(n)$) подпространство H_n пространства H; потребуем, чтобы на элементах подпространства H_n полунорма $|||\cdot|||$ была конечной. Поставим одностороннюю вариационную задачу

$$x^{(n)}\in H_n\cap M,\quad |x_0-x^{(n)}|=\min. \qquad (\text{УШ.I.7})$$

Задачу (УШ.I.7) можно несколько упростить. Пусть x_{0n} – ортогональная проекция элемента x_0 на H_n, тогда

$$|x_0-x^{(n)}|^2=|x_{0n}-x^{(n)}|^2+|x_0-x_{0n}|^2.$$

Второе слагаемое справа – постоянное, поэтому задача (УШ.I.7) равносильна следующей:

$$|x_{0n}-x^{(n)}|^2=\min,\quad x^{(n)}\in H_n\cap M. \qquad (\text{УШ.I.}\ldots)$$

Преимуществом этой новой формы нашей задачи является то, что как данный элемент x_{0n}, так и искомый элемент $x^{(n)}$ принадлежат одному и тому же пространству H_n.

Пересечение подпространства H_n и замкнутого выпуклого множества M есть замкнутое выпуклое множество. Потребуем еще, чтобы оно было непустым, тогда задача (УШ.I.8) будет иметь одно и только одно решение, которое будем рассматривать как приближенное решение задачи (УШ.I.6). Заметим, что если $x_0\notin M$, то при N достаточно большом будет $x_{0n}\notin M$ и, по доказанному в п.I°, $x_*^{(n)}\in\partial M$.

5°. Известны многие способы решения задач типа (УШ.I.8); здесь ограничимся ссылкой на гл.II книги Гловинского, Лионса и Тремольера [I]. Укажем еще один прием, который может оказаться небесполезным: см. статью автора [27].

Пусть $\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN}$ – базис в H_n. Допустим, что существует такая монотонная функция $\omega: R_1^+ \longrightarrow R_1^+$, что $\omega(\| x^{(n)} - q \|)$, где

$$x^{(n)} = \sum_{k=1}^{N} a_k \varphi_{nk}$$

есть достаточно гладкая функция коэффициентов a_k. Положим

$$x_*^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk};$$

обозначим через $a^{(n)}$ вектор $(a_1^{(n)}, a_2^{(n)}, \ldots, a_N^{(n)})$, через M_n – матрицу чисел $[\varphi_{nk}, \varphi_{nj}]$; $j, k = 1, 2, \ldots, N$. Положим еще $c_j^{(n)} = [u_{on}, \varphi_{nj}]$, $c^{(n)} = (c_1^{(n)}, c_2^{(n)}, \ldots, c_N^{(n)})$ и $\Phi_n(a^{(n)}) = \omega(\| x_*^{(n)} - q \|)$. По способу множителей Лагранжа задача (Уш.I.8) сводится к системе уравнений относительно коэффициентов $a_j^{(n)}$:

$$\frac{\partial}{\partial a_j^{(n)}}(M_n a^{(n)}, a^{(n)}) - 2\lambda \frac{\partial \Phi_n(a^{(n)})}{\partial a_j^{(n)}} = 2c_j^{(n)}, \quad 1 \le j \le N,$$

или, короче,

$$M_n a^{(n)} - \lambda \operatorname{grad} \Phi_n(a^{(n)}) = c^{(n)}. \qquad (УШ.I.9)$$

Будем рассматривать λ как независимую переменную и $a_j^{(n)}$ как функции от λ. Дифференцируя (УШ.I.9) по λ, получим систему обыкновенных дифференциальных уравнений

$$(M_n - \lambda \Psi_n)\frac{da^{(n)}}{d\lambda} - \operatorname{grad} \Phi_n(a^{(n)}) = 0, \qquad (УШ.I.I$$

которую надо решать при начальном условии

$$c^{(n)}|_{\lambda=0} = M_n^{-1} c^{(n)}; \qquad (Уш.I.I$$

в уравнении Ш.I.I0) Ψ_n означает матрицу вторых производных функции Φ_n.

Задачу (УШ.I.9–II) будем решать каким-нибудь разностным методом. Пусть h – шаг соответствующей сетки. Положим $\lambda_k = \pm kh$, $k = 1, 2, \ldots$ пределим вектор $a^{(n)}(k)$, соответствую

некоторому λ_k, вычислим величину $\left\| \sum_{j=1}^{N} a_j^{(n)}(\kappa) \vartheta_{nj} - g \right\|$. При малых κ она близка к $\|u_0 - g\|$ и, следовательно, больше, чем d. Процесс решения задачи (УШ.I.10-II) следует приостановить при том значении κ, положительном или отрицательном, при котором в пределах принятой точности будет выполняться равенство

$$\left\| \sum_{j=1}^{N} a_j^{(n)}(\kappa) \vartheta_{nj} - g \right\| = d. \qquad (УШ.I.12)$$

Вычисление величины $a_j^{(n)}$ упрощается, если полунорма $\|\cdot\|$ определяется равенством $\|u\|^2 = b(u,u)$, где $b(u,v)$ - симметричная билинейная форма, определенная на всем пространстве $\widetilde{H}$. Если в этом случае положить $\omega(t) = t^2$, то система (УШ.I.9) примет вид

$$(\mathcal{M}_n - \lambda \mathcal{F}_n) a^{(n)} = c^{(n)} - f^{(n)}, \qquad (УШ.I.13)$$

где $\mathcal{F}_n$ - матрица чисел $b(\vartheta_{nj}, \vartheta_{nk})$; $j,k = 1, 2, \ldots, N$, а $f^{(n)}$ - вектор с составляющими $f_k^{(n)} = b(g, \vartheta_{nk})$. Матрица $\mathcal{M}_n$ положительно определенная и потому невырожденная; систему (УШ.I.13) можно решить, обратив матрицу $\mathcal{M}_n - \lambda \mathcal{F}_n$ по методу Фаддеева (Фаддеев и Фаддеева [1]; см. также статью Вайнико и автора [1]). Определив коэффициенты $a_j^{(n)}$, мы затем найдем λ из уравнения $b(x^{(n)} - g, x^{(n)} - g) = d^2$.

§ 2. Погрешность аппроксимации.

I°. Пусть B - банахово пространство с нормой $\|\cdot\|$, обладающее следующими свойствами:

1) Γ непрерывно вкладывается в $\widetilde{H}$; при этом

$$\forall x \in B, \quad |x| \leqslant \|x\|, \quad \|x\| \leqslant \|x\|' \qquad (УШ.2.1)$$

2) элементы x_0 и x_* содержатся в B;

3) при любом n подпространство $H_n \subset B$, и последовательность $\{H_n\}$ полна в B.

Если в дальнейшем $\|x_0\|'$ конечна, что мы ниже всегда предполагаем, то условиям 1)-2) можно удовлетворить, положив, например,

$$\|x\|^q = |x|^q + \|x\|'^q, \quad 1 \leqslant q \leqslant \infty; \qquad (УШ.2.2)$$

укажем класс случаев, когда условие 3) также выполнено. Допустим, например, что $\widetilde{H} \sim W_2^{(k)}(\Omega)$, где k натуральное число и Ω конечная область евклидова пространства, удовлетворяющая усло-

вию конуса, и пусть для некоторых $\delta \geqslant \kappa$, $\rho \geqslant 1$ справедливо неравенство $\|\cdot\| \leqslant C \|\cdot\|_{W_\rho^{(\delta)}(\Omega)}$, $C = const$. В качестве H_n возьмем подпространства, натянутые на координатные функции МКЭ (см. книгу автора [22]), которые в свою очередь получены преобразованием независимых переменных из соответствующих функций; эти последние можно выбрать так, чтобы последовательность $\{H_n\}$ была полной в $W_\rho^{(\delta)}(\Omega)$. Пусть еще ρ таково, что $W_\rho^{(\delta)}(\Omega)$ ограниченно вкладывается в $W_2^{(\kappa)}(\Omega)$. Тогда $H_n \subset B$ при любом n и последовательность $\{H_n\}$ полна в B .

Ниже мы будем пользоваться равенством (УШ.2.2) при $q = 1$.

2°. Лемма УШ.2.1. Пусть M – выпуклое замкнутое в норме $\|\cdot\|$ непустое множество. Пусть, далее, x_* – решение односторонней вариационной задачи

$$\|x_0 - x_*\| = \min, \quad x_0 \in H \setminus M, \quad x_* \in M, \qquad (УШ.2.3)$$

а x – произвольный элемент множества M . Тогда

$$\|x_* - x\|^2 \leqslant d'^2 - d^2, \qquad (УШ.2.4)$$

$$d = \|x_0 - x_*\| \quad d' = \|x_0 - x\|. \qquad (УШ.2.5)$$

В пространстве H проведем двумерную плоскость P через точки x_0 , x_* , x . Гильбертова метрика пространства H индуцирует на P евклидову метрику. Построим треугольник с вершинами x_0 , x_1 , x_* и обозначим через ψ угол при вершине x_* . Докажем, что $\psi \geqslant \pi/2$. Допустим, что $\psi < \pi/2$. Через точку x_0 проведем в P прямую, наклонную к прямой $\overline{x_0 x_*}$, под углом, меньшим, чем $\pi/2 - \psi$, и пересекающую отрезок $\overline{x_* x}$; Пусть y – точка пересечения (рис.1). В треугольнике с вершинами x_0 , x_* , y угол при вершине y тупой, поэтому $\|x_0 - y\| < d$. Множество M выпуклое, а x , $x_* \in M$, поэтому $y \in M$. Теперь неравенство $\|x_0 - y\| < d$ означает, что x_* не есть решение задачи (УШ.2.3), вопреки предположению. Таким образом, $\psi \geqslant \pi/2$.

Теперь по теореме косинусов,

$$d'^2 = \|x_* - x\|^2 + d^2 - 2d\|x_* - x\| \cos\psi,$$

но $\cos\Psi \le 0$, и $d'^2 \ge |x_* - x|^2 + d^2$. Неравенство (УШ.2.4) доказано.

3°. Лемма УШ.2.2. Пусть выполнены условия I) - 3) п.I° настоящего параграфа, множество ограничений M определено формулой (УШ.I.I), причем $\beta = |||g||| < d$. Пусть x - некоторая точка на ∂M, так что $|||x-g||| = d$. Допустим, что существует точка $x^{(n)} \in H_n$, удовлетворяющая неравенству $\|x - x^{(n)}\| \le \gamma_n$, где $\gamma_n < (d-\beta)/2$, и $\|\cdot\| = |\cdot| + |||\cdot|||$. Тогда существует элемент $y^{(n)} \in M \cap H_n$, такой, что $\|x - y^{(n)}\| \le C_x \gamma_n$, где C_x может зависеть от x, но не зависит от n.

Положим $y^{(n)} = ax^{(n)}$, $0 < a < 1$; очевидно, $y^{(n)} \in H_n$. Выясним, при каких a справедливо включение $y^{(n)} \in M$. Имеем

$$|||y^{(n)} - g||| \le a\,|||x^{(n)} - g||| + (1-a)\beta,$$

$$|||x^{(n)} - g||| \le |||x^{(n)} - x||| + |||x - g||| \le \gamma_n + d.$$

Отсюда $|||y^{(n)} - g||| \le a(\gamma_n + d) + (1-a)\beta$. Если $a(\gamma_n + d) + (1-a)\beta = d$, или

$$a = \frac{d-\beta}{d-\beta+\gamma_n} \qquad \text{(УШ.2.6)}$$

то $y^{(n)} \in M$. При этом

$$\|x - y^{(n)}\| \le \|x - x^{(n)}\| + (1-a)\|x^{(n)}\| \le \gamma_n + \frac{\gamma_n \|x^{(n)}\|}{d-\beta+\gamma_n}.$$

Далее, $\|x^{(n)}\| \le \|x\| + \|x - x^{(n)}\| \le \|x\| + \gamma_n$, следовательно,

$$\|x - y^{(n)}\| \le \gamma_n \left[1 + \frac{\|x\| + \gamma_n}{d-\beta+\gamma_n}\right]. \qquad \text{(УШ.2.7)}$$

Отсюда вытекает более простая оценка

$$\|x - y^{(n)}\| \le C_x \gamma_n, \quad C_x = 2\left[1 + \frac{\|x\|}{d-\beta}\right]. \qquad \text{(УШ.2.8)}$$

Лемма доказана. Аналогично можно получить оценки

$$|x - y^{(n)}| \le C'_x \gamma_n, \quad |||x - y^{(n)}||| \le C''_x \gamma_n, \qquad \text{(УШ.2.9)}$$

где

$$C'_x = 2\left[1 + \frac{|x|}{d-\beta}\right], \quad C''_x = 2\left[1 + \frac{|||x|||}{d-\beta}\right]. \qquad \text{(УШ.2.10)}$$

Теорема УШ.2.1. Если пространство B удовлетворяет условиям I) - 3) п.I°, а x_0, x_*, $x_*^{(n)}$ имеют тот же смысл, что и выше, то

$$\|x_* - x_*^{(n)}\| = O\left(\sqrt{e_n(x_*)}\right), \qquad (\text{УШ.2.11})$$

где $e_n(x)$ есть наилучшее приближение элемента $x \in B$ элементами подпространства H_n в норме $\|\cdot\|$.

Заметим, что по предположению последовательность $\{H_n\}$ полна в B, поэтому $e_n(x) \xrightarrow[n\to\infty]{} 0$. В частности, $\theta_n(x_*) \xrightarrow[n\to\infty]{} 0$.

Как было доказано в п.4° § I, $x_*^{(n)} \subset \partial M$, если n достаточно велико.

По определению наилучшего приближения, при любом заданном $\varepsilon > 0$ существует такой элемент $x^{(n)} \in H_n$, что $\|x_* - x^{(n)}\| \leqslant (1+\varepsilon)e_n(x_*)$. По лемме УШ.2. существует элемент $y^{(n)} \in M_n := M \cap H_n$, что $\|x_* - y^{(n)}\| \leqslant C_* e_n(x_*)$, $C_* = (1+\varepsilon)C_{x_*}$.

По лемме УШ.2 I, $\|x_* - x_*^{(n)}\|^2 \leqslant d_n^2 - d^2$, $d_n^2 = \|x_0 - x_*^{(n)}\|^2$.

Но $x_*^{(n)}$ есть решение задачи УШ.2.7), поэтому $d_n \leqslant d_n' = \|x_0 - y^{(n)}\|$ и

$$\|x_* - x_*^{(n)}\|^2 \leqslant d_n'^2 - d^2 = (d_n' + d)(d_n' - d).$$

Кроме того,

$$d_n' = \|x_0 - y^{(n)}\| \leqslant \|x_0 - x_*\| + \|x_* - y^{(n)}\| \leqslant d + C_* e_n(x_*).$$

Отсюда

$$\|x_* - x_*^{(n)}\|^2 \leqslant [2d + C_* e_n(x_*)] C_* e_n(x_*). \qquad (\text{УШ.2.12})$$

Теорема доказана.

§ 3. Случаи улучшения оценки погрешности аппроксимации.

Существуют случаи, когда оценка (УШ.2.11) может быть существенно улучшена. В настоящем параграфе приведены два таких случая, описанных в статьях автора [27] и [28].

1°. В п.I заменим точку x на $x_*^{(n)}$. По теореме косинусов

$$\|x_* - x_*^{(n)}\|^2 = \|x_0 - x_*^{(n)}\| - d)^2 + 4d\,\|x_0 - x_*^{(n)}\| \sin^2 \frac{\theta_n}{2};$$

через θ_n здесь обозначен угол при вершине x_0 в треугольнике $(x_0, x_*, x_*^{(n)})$. Величина $\|x_0 - x_*^{(n)}\| \xrightarrow[n\to\infty]{} d$, поэтому при n достаточно большом $\|x_0 - x_*^{(n)}\| \leqslant 2d$. Обозначим еще через $x^{(n)}$ элемент подпространства H_n, на котором достигается равенство

$|x_* - x^{(n)}| = e_n(x_*)$, тогда

$$d \leq |x_0 - x_*^{(n)}| \leq |x_0 - x^{(n)}| \leq |x_c - x| + |x_* - x^{(n)}| = d + e_n(x_*);$$

отсюда $0 \leq |x_0 - x_*^{(n)}| - d \leq e_n(x_*)$ и, следовательно,

$$|x_* - x_*^{(n)}|^2 \leq e_n^2(x_*) + 2d^2\theta_n^2 , \qquad (\text{Ш.3.1})$$

и дело сводится к оценке угла θ_n . Допустим, что многообразие ∂M достаточно гладкое, тогда вектор $x_0 - x_*$ направлен по нормали к ∂M . Приведем доказательство этого простого утверждения. Выберем на ∂M произвольную точку y и проведем через точки x_0 , x_* , y двумерную плоскость ω (рис.2), которая пересечет ∂M по достаточно гладкой кривой $\mathcal{L}$. Введем в ω декартовы координаты (ξ_1 , ξ_2); за ось ξ_1 примем прямую проходящую через точки x_0 и x_* и направленную от x_0 к x_* ; ту же ось примем за ось полярных координат, центр которых поместим в x_0 . Уравнение касательной к кривой $\mathcal{L}$ в точке x_* можно написать в виде (ξ_1 - абсцисса переменной точки x)

$$\frac{\xi_1 - \xi_{*1}}{d\xi_{*1}} = \frac{\xi_2}{d\xi_{*2}} .$$

Пусть $\rho = \rho(\theta)$ есть уравнение кривой $\mathcal{L}$ в полярных координатах вблизи точки x_* , тогда

$$d\xi_{*1} = d(\rho\cos\theta) = \cos\theta d\rho - \rho\sin\theta d\theta .$$

Но в точке x_* $d\rho = 0$, потому что в этой точке ρ достигает минимума; кроме того, в этой же точке $\sin\theta = 0$. Таким образом $d\xi_{*1} = 0$ и уравнение касательной к $\mathcal{L}$ в точке x_* есть $\xi_1 = \xi_{*1}$, и прямая $\overline{x_* x_0}$ направлена по нормали к ∂M . Как было только-что отмечено, в введенных нами полярных координатах можно записать уравнение кривой $\mathcal{L}$ вблизи точки x_* в виде $\rho = \rho(\theta)$, где $\rho(\theta)$ - достаточно гладкая функция от θ , причем $\rho(0) = d$ $\rho'(0) = 0$. Наложим на многообразие ∂M следующее требование Г :

Существуют такие положительные числа K_0 , K_1 , $\bar{K}$, что

$$K_0 \leq \frac{1}{2}\rho''(0) \leq K_1 ; \quad \frac{1}{6}|\rho'''(\theta)| \leq \bar{K} ; \qquad (\text{Ш.3.2})$$

эти числа могут зависеть от x_0, но не зависят от выбора двумерного сечения, проходящего через точки x_0 и x_*; второе неравенство (УШ.3.2) должно выполняться при достаточно малых θ.

Обозначим $\rho(\theta) - d = l(\theta)$. Разлагая $\rho(\theta)$ по формуле Тейлора, находим $l(\theta) = K\theta^2 + \rho'''(\tau)\theta^3/6$; $K = \rho''(0)/2$.

Отсюда $\theta^2 = l(\theta)/K - \rho'''(\tau)\theta^3/K$ и, следовательно,

$$\frac{1}{K_1} l(\theta) - \frac{\bar{K}}{K_0}|\theta|^3 \le \theta^2 \le \frac{1}{K_0} l(\theta) + \frac{\bar{K}}{K_0}|\theta|^3.$$

Пусть $|\theta| \le K_0/2\bar{K}$. Из последнего неравенства находим

$$\frac{2}{3K_1} l(\theta) \le \theta^2 \le \frac{2}{K_0} l(\theta). \qquad (УШ.3.3)$$

По теореме косинусов

$$|x_* - y|^2 = d^2 + \rho^2(\theta) - 2d\rho(\theta)\cos\theta =$$

$$= l^2(\theta) + 4d(d + l(\theta))\sin^2\frac{\theta}{2};$$

отсюда и из неравенства (УШ.3.3) вытекает неравенство

$$\beta_1\theta \le |y - x_*| \le \beta_2\theta \qquad (УШ.3.4)$$

с постоянными β_1 и β_2, которые не зависят от выбора описанного выше двумерного сечения.

Теорема УШ.3.1. Пусть выполнено требование Т, а также предположения п.1° § 2. Тогда

$$|x_* - x_*^{(n)}| \le C e_n(x_*), \quad C = const. \qquad (УШ.3.5)$$

Существует такой элемент $x^{(n)} \in M_n$, что $|x_* - x^{(n)}| \le (1 + \varepsilon) e_n(x_*)$. По лемме УШ.2.2 найдется элемент $y^{(n)} \in M_n$, удовлетворяющий неравенству $|x_* - y^{(n)}| \le C_* e_n(x_*)$. Воспользуемся обозначениями рис.3; все расстояния и углы измеряются в плоскости π. По построению приближенного решения, $\rho_n \le \rho_n'$. Обозначим $l_n = \rho_n - d$ и $l_n' = \rho_n' - d$, тогда $l_n < l_n'$.

в силу неравенства (УШ.3.5)

$$\theta_n^2 \leqslant \frac{2}{l_n} < \frac{2}{K_0} l_n' \leqslant \frac{3K_1}{K_0} \theta_n'^2 ,$$

или $|\theta_n| < |\theta_n'| \sqrt{3K_1/K_0}$. Теперь по неравенству (УШ.3.4),

$$|x_* - x_*^{(n)}| \leqslant \beta_2 \theta_n < \beta_2 \theta_n' \sqrt{3K_1/K_0} \leqslant$$

$$\leqslant \frac{\beta_2}{\beta_1} \sqrt{3K_1/K_0}\, |x_* - y^{(n)}| \leqslant \frac{\beta}{\beta_1} \sqrt{3K_1/K_0}\, C_* e_n(x_0) .$$

Таким образом, если выполнено требование Т, то погрешность аппроксимации приближенного решения есть величина порядка $O(e_n(x_0))$.

2°. В настоящем пункте мы укажем некоторый простой класс полунорм, удовлетворяющих требованию Т. Именно, пусть $\|x\|^2 = b(x) = b(x,x)$, где $b(x,y)$ – симметричная билинейная форма, заданная на H, которой соответствует неотрицательная квадратичная форма $b(x)$. Для упрощения выкладок будем в данном пункте считать, что $q = 0$, $\hat{H} = H$.

Произвольную двумерную плоскость, проходящую через точки x_0 и x_*, можно построить так. В пространстве H возьмем такую точку y, что $|x_0 - y| = d$ и $[x_0 - x \;\; x_0 - y] = 0$, и положим $e_1 = (x_* - x_0)/d$, $e_2 = (y - x_0)/d$. Двумерное множество, определяемое уравнением $x - x_0 = \xi e_1 + \eta e_2$, где ξ и η – произвольные вещественные числа, есть плоскость, проходящая через точки x_0 , x_* , y ; элементы вида $x - x_0$, соответствующие числам $\xi = 0$ и $\eta = 0$, ортогональны, поэтому ξ и η можно рассматривать как декартовы координаты на нашей плоскости; начальному координат соответствует точка x_0.

Введем полярные координаты по обычным формулам $\xi = \rho \cos\theta$, $\eta = \rho \sin\theta$. Отметим полярные координаты точек x_0 , x_* , y , именно, $x_0 : \rho = 0$; $x_* : \rho = d$, $\theta = 0$; $u : \rho = d$, $\theta = \pi/2$. В полярных координатах $x = x_0 + \rho v_\theta / d$, где $v_\theta = (x_* - x_0)\cos\theta + (y - x_0)\sin\theta$.

Если $\omega(x,y)$ – симметричная билинейная форма, то

$$\omega(x) := \omega(x,x) = d^2 \rho^2 \omega(v_\theta) + 2d\rho\omega(x_0, v_\theta) + \omega(x_0).$$

В частности, многообразие ∂M определяется уравнением

$$d^{-2}\rho^2 b(v_\theta)+2d^{-1}\rho\, b(x_0,v_\theta)+b(x_0)-d=0. \qquad (УШ.3.$$

Нетрудно видеть, что $b(v_0)>0$. Действительно, если $b(v_0)=0$, то $\|x_0-x_*\|=0$ и $\|x_0\|=\|x_*\|=d$; это означает, что x_0 есть решение односторонней вариационной задачи, вопреки предположению.

Итак, $b(v_0)>0$. Но тогда при достаточно малых θ и $b(v_\theta)>0$. Уравнение (УШ.3.7) – квадратное относительно ρ, из его корней один равен $\rho(\theta)$, а другой больше, чем $\rho(\theta)$. Отсюда вытекают неравенства

$$-b(x_0,v_0)-b(x_0,x_0-x_*)>0, \qquad (УШ.3.$$

$$b(x_0)-d^2>0. \qquad (УШ.3.$$

Положив в (УШ.3.7) $\theta=0$, $\rho=d$, получим

$$b(x_0)-d^2=-[b(v_0)+2b(x_0,v_0)]=b(x_0+x_*,x_0-x_*),$$

и уравнение (УШ.3.7) можно записать в виде

$$d^{-2}\rho^2 b(v_\theta)-2d^{-1}\rho b(x_0-v_\theta)+b(x_0+x_*,x_0-x_*)=0, \qquad (УШ.3.$$

а неравенство (УШ.3.9) – в виде

$$b(x_0+x_*,\,x_0-x_*)>0. \qquad (УШ.3.$$

Вычислим теперь величину $\rho''(0)$. Продифференцируем тождество (УШ.3.7) два раза по θ. Положив $\theta=0$, получим

$$\rho''(0)=d\,\frac{b(x_*,\,x_0-x_*)+b(x_0-v_0)}{b(x_0+x_*,\,x_0-x_*)}. \qquad (УШ.3.$$

Докажем, что знаменатель дроби (УШ.3.12) положителен. При $\theta=0$ уравнение (УШ.3.10) принимает вид

$$d^{-2}\rho^2 b(x_0-x_*)-2d^{-1}\rho\, b(x_0,x_0-x_*)+b(x_0+x_*,x_0-x_*)=0; \qquad (УШ.3.$$

его корни суть

$$\frac{d}{b(x_0-x_*)}\left\{b(x_0,x_0-x_*)-\left[b^2(x_0,x_0-x_*)-b(x_0-x_*)b(x_0+x_*,x_0-x_*)\right]^{1/2}\right\}=$$

$$=\frac{d\left[b(x_0,x_0-x_*)-b(x_*,x_0-x_*)\right]}{b(x_0-x_*)}. \qquad (\text{УШ.3.14})$$

Если в (УШ.3.14) выбрать знак минус, то получим меньший корень уравнения (УШ.3.13) $\rho \quad d$; отсюда следует, что $(v_*, x_0-x_*)>0$.

Из формулы (УШ.3.12) вытекает теперь, что $\rho''(\cdot)\geqslant d$. С другой стороны, так как форма $b(x)$ задана на всем пространстве H, существует такая постоянная $\bar{c}>0$, что $b(x_0-y)\leqslant \bar{c}|x_0-y|^2=\bar{c}d^2$, и мы приходим к оценке

$$K_0:=\frac{d}{2}\leqslant\frac{1}{2}\rho''(0)\leqslant\frac{d}{2}+\frac{\bar{c}d^3}{2b(x_*,x_0-x_*)}:=K_1, \qquad (\text{УШ.3.15})$$

в которой K_0 и K_1 не зависят от выбора элемента y.

Докажем теперь, что при малых θ третья производная $\rho'''(\theta)$ ограничена по модулю независимо от ρ и v. Меньший корень уравнения (УШ.3.11) равен

$$\rho(\theta)=\frac{d}{b(v_\theta)}\left[b(x_0,-v_\theta)-\right.$$

$$\left.-\sqrt{b^2(x_0,-v_\theta)-b(x_0+x_*,x_0-x_*)b(v_\theta)}\right]. \qquad (\text{УШ.3.16})$$

Как мы видели, $b(v_\theta)>0$ при малых θ. Из формулы (УШ.3.16) видно, что любая производная от $\rho(\theta)$ будет ограничена, если ограничено одно лишь подкоренное выражение. Выше было установлено, что при $\theta=0$ это равно $b^2(x_*,x_0,x_*)>0$. Далее,

$$v_\theta=x_0-x_*-2(x_0-x_*)\sin^2\frac{\theta}{2}+(x_0-y)\sin\theta:=x_0-x_*+\eta;$$

заметим, что $|\eta|\leqslant d(|\theta|+\frac{\theta^2}{2})$ и потому при малых θ будет $|\eta|\leqslant\beta|\theta|$, где β не зависит от θ от y. Подкоренное

выражения в (УШ.3.16) равно

$$b^2(x_0, x_0 - x_*) + b^2(x_0, \eta) b(x_0, x_0 - x_*) + b^2(x_0, \eta) -$$

$$-[2b(x_0 - x_*, \eta) + b(\eta)] b(x_0 + x_*, x_0 - x_*) =$$

$$= b^2(x_0, x_0 - x_*) + O(|\theta|),$$

и при достаточно малых δ оно ограничено снизу положительным числом, не зависящим от θ и ψ.

Заметим еще, что требование Т выполнено, если для любого элемента ψ линия $\mathcal{L}$ – прямая (или ее часть); в этом случае $\rho(\theta) = d/\cos\theta$ и можно положить $K_0 = K_1 = a/2$ и, при $|\theta| \leqslant \pi$ $K = \pi d/\theta$.

3°. Как и выше, возьмем произвольную точку $\psi \in \partial M$ и проведем в H двумерную плоскость P через точки x_0, x_*, ψ (см. рис. 2) и введем на этой плоскости полярные координаты ρ и θ. Отметим, что полунорма $\|\cdot\|'$ конечна на всех элементах плоскости P; кроме того, как было показано в § 2, $\Psi \geqslant \pi/2$. Пусть по-прежнему $L = \partial M \cap P$ и $\rho = \rho(\theta)$ есть уравнение кривой L.

Функция $\rho(\theta)$ достигает минимума при $\theta = 0$, поэтому $\rho(\theta)$ возрастает при возрастании $|\theta|$. Из выпуклости кривой L следует, что с возрастанием θ угол Ψ также возрастает.

Сделаем теперь следующее допущение: существуют такие положительные постоянные θ_0, K, $\delta < \pi/2$, что при $|\theta| \leqslant \theta_0$ справедливы неравенства $|\rho'(\theta)| \leqslant K$ и $\delta + \pi/2 \leqslant \Psi$. Эти постоянные могут зависеть от x_0, но не от выбора двумерной плоскости P, проходящей через точки x_0 и x_*.

Из неравенства $\Psi \leqslant \delta + \pi$ следует, что x_* – "коническая" точка множества ∂M: вектор $x_0 - x_*$ образует угол, не меньший чем $\delta + \pi/2$, с любой секущей множества ∂M, проходящей через точку x_*.

Покажем, что при сформулированном здесь допущении верна оценка (УШ.3.5) для погрешности приближенного решения.

Пусть $|\theta| \leqslant \theta_0$. Обозначим $\ell(\theta) = \rho(\theta) - d$. Имеем

$$\ell(\theta) = \rho(\theta) - d = \rho(\theta) - \rho(0) = \int_0^\theta \rho'(\vartheta)\, d\vartheta \leqslant K|\theta|, \qquad (\text{УШ.3.}$$

поэтому, если x – точка с полярными координатами $(\rho(\theta), \theta)$, то

$$|x_* - x|^2 = \rho^2(\theta) + d^2 - 2\rho(\theta)d\cdot\cos\theta =$$

$$= \ell^2(\theta) + 4\rho(\theta)d\cdot\sin^2\frac{\theta}{2} \le C_1^2\theta^2, \quad C_1 = \mathrm{const}. \qquad \text{(УШ.3.18)}$$

С другой стороны, обозначая $\lambda = |x_* - x|$, находим

$$\lambda \ge |\theta|\,\frac{\lambda\sin\psi}{\rho(\theta)}; \quad |\theta| \le \frac{\theta_0}{\sin\theta_0}\,\frac{\lambda}{d} =: C_2|x_* - x|. \qquad \text{(УШ.3.19)}$$

Кроме того,

$$\lambda = \frac{\rho(\theta)\sin|\theta|}{\sin\psi} \ge \frac{\rho(\theta)\sin|\theta|}{\cos\delta}$$

и, следовательно,

$$\ell^2(\theta) + 4\rho(\theta)d\cdot\sin^2\frac{\theta}{2} \ge \frac{\rho^2(\theta)\sin^2\theta}{\cos^2\delta}. \qquad \text{(УШ.3.20)}$$

Докажем, что при достаточно малом θ_0 справедливо неравенство

$$4\rho(\theta)d\cdot\sin^2\frac{\theta}{2} \le \sigma\rho^2(\theta)\frac{\sin^2\theta}{\cos^2\delta}, \qquad \text{(УШ.3.21)}$$

в котором σ – постоянная, $0 < \sigma < 1$. Неравенство (УШ.3.21) равносильно следующему: $d \le \sigma\rho(\theta)\cos^2\frac{\theta}{2}\cos^2\delta$, и достаточно, чтобы $d \le \sigma\rho(\theta)\cos^2\frac{\theta_0}{2}\cos^2\delta$. Если взять $C_* < 2\delta$, то последнее неравенство будет выполнено при $\sigma = \cos^2\delta/\cos^2\frac{\theta_0}{2}$. Теперь

$$\ell^2(\theta) \ge (1-\sigma)\frac{\rho^2(\theta)\sin^2\theta}{\cos^2\delta} > (1-\sigma)d\,\frac{\sin^2\theta}{\cos^2\delta},$$

и окончательно

$$\ell(\theta) \ge C_3|\theta|, \quad C_3 = \frac{d\sqrt{1-\sigma}}{\theta_0}\frac{\sin\theta_0}{\cos\delta}. \qquad \text{(УШ.3.22)}$$

Пусть $\bar{x}^{(n)}$ элемент подпространства H_n, на котором достигается наилучшее приближение элемента x_*, т.е. что $e_n(x_*) = \|x_* - \bar{x}^{(n)}\|$. По лемме УШ.2. существует элемент $y^{(n)} \in M \cap H_n$, такой, что $\|x_* - y^{(n)}\| \le C e_n(x_*)$. Воспользуемся обозначениями рис.3, в котором на этот раз будем считать $x^{(n)} = y^{(n)}$; все углы и расстояния измеряются в метрике L_1. При построении

приближенного реше... $\rho'_n > \rho_n$. Обозначим $l_n = \rho_n - d$, $l'_n = \rho'_n - d$, так что $l_n < l'_n$. По формулам (УШ.3.18), (УШ.3.22), (УШ.3.17) и (УШ.3.19) получаем

$$|x_* - x_*^{(n)}| \le C_1|\theta_n| \le \frac{C_1}{C_3} l_n < \frac{C_1}{C_3} l'_n \le \frac{C_1 K}{C_3}|\theta'_n| \le$$

$$\le \frac{C_1 C_2 K}{C_3}|x_* - y^{(n)}| \le C e_n(x_*); \quad C = \frac{C_1 C_2 K}{C_3},$$

и наше утверждение доказано.

§ 4. Оценка погрешности аппроксимации для более общей задачи

1°. В работе Тихтина [1] исследована погрешность аппроксимации более общей задачи, сформулированной в начале настоящей главы. Напомним постановку этой задачи. В гильбертовом банаховом пространстве B с нормой $\|\cdot\|$ заданы непустое выпуклое замкнутое множество M и выпуклый слабо полунепрерывный снизу функционал γ; если множество M неограниченное, то дополнительно потребуем, чтобы $\gamma(x) \xrightarrow[\|x\| \to \infty]{} +\infty$. Ставится задача о нахождении элемента $x_* \in M$, реализующего минимум функционала $\gamma(x)$ на множестве M. Как было сказано в начале главы, решение этой задачи существует и единственно, если функционал γ строго выпуклый.

2°. Приближенное решение $x^{(n)}$ этой более общей задачи можно построить так же, как в § 2: выберем полную в B последовательность конечномерных подпространств $\{B_n\}$ и примем за $x^{(n)}$ решение задачи о минимуме функционала γ на множестве $M \cap B_n$. Полнота последовательности $\{B_n\}$ в цитируемой статье Тихтина постулируется в следующем смысле. Пусть $P_M(x)$ функционал Минковского для множества M. Этот функционал определяется следующим образом. Допустим, что L – произвольное линейное пространство и M' – выпуклое множество в L. Точка $\xi \in M'$ принадлежит ядру множества M' (терминология книги Колмогорова и Фомина [1]; в книге Балакришнана [1] она называется алгебраически внутренней точкой множества M'), если для любого $\eta \in L$ существует такое число $a > 0$, что $(\xi + t\eta) \in M'$, если $|t| < a$. Пусть ξ_0 – точка ядра множества M'. Функционал Минковского множества M' определяется формулой (см. цитированную выше книгу

$$\forall x\in L,\ P_M(x)=\inf_{\lambda>0}\left\{\lambda:\frac{x-\xi_0}{\lambda}\in M\right\}. \qquad (\text{УШ.}4.1)$$

Ниже буде. считать, что $\xi_0=0$ – это, очевидно, не ограничивает общности.

Положим теперь

$$\forall x\in B,\ \varepsilon_n(x)=\inf_{x^{(n)}\in B_n}\left[\|x-x^{(n)}\|+|P_M(x)-P_M(x_n)|\right]. \qquad (\text{УШ.}4.2)$$

Будем счит ть, что на элемен..ах подпространств B_n функционал ..инковского конечен. По определению, при этому в работе Тихтина [1], последовательность $\{B_n\}$ полна в точке $x\in B$, если $\lim_{n\to\infty}\varepsilon_n(x)=0$. Отсюда видно, для такой полноты необходимо, чтобы $P_M(x)<\infty$.

3°. Введем в рассмотрение функцию

$$q(\alpha)=\inf_{y\in M_\alpha}\left[J(y)-J(x_*)\right],$$
$$M_\alpha=\{y\in M;\ \|y-x_*\|=\alpha\}, \qquad (\text{УШ.}4.3)$$

которую назовем минорантой функционала J на множестве M.

Оценка погрешности аппроксимации дается следующей теоремой:

Теорема УШ.4.1. Пусть выполнены следующие условия: а) на всех элементах всех подпространств B_n функционал Минковского конечен; б) в некоторой окрестности (в смысле нормы пространства B) точного решения x_* функционал $J(x)$ конечен и удовлетворяет условию Липшица; в) при малых значениях числа $\alpha\geq 0$ функция $q(\alpha)$ строго монотонна; г) последовательность подпространств $\{B_n\}$ полна в B в указанном выше смысле. Тогда имеет место оценка погрешности аппроксимации

$$\|x_*-x_*^{(n)}\|\leq q^{-1}(\beta\varepsilon_n(x_*));\quad \beta=\beta\max(1,\|x_*\|), \qquad (\text{УШ.}4.4)$$

β – постоянная Липшица для функционала J в окрестности точки x_*.

Замечание УШ.4.1. В работе Тихтина [1] доказано, что J удовлетворяет условию Липшица, если M имеет внутреннюю точку

Замечание УШ.4.2. При доказательстве теоремы УШ.4.1 играет

важную роль лемма, которая является обобщением леммы УШ.2. :

Лемма УШ.4.1. Пусть $x \in M$ и $x^{(n)} \in B_n$. Существует точка $y^{(n)} \in B_n \cap M$, такая, что

$$\|x - y^{(n)}\| \le \|x - x^{(n)}\| + \|x\| \cdot |P_M(x) - P_M(x^{(n)})|. \qquad (УШ.4.5)$$

Отметим важный частный случай, рассмотренный в цит. рованной работе Бахтина. Пусть B есть гильбертово пространство и $\gamma(x) = |x - x_0|^2$, $x_0 \notin M$; причимаем, что M произвольное непустое выпуклое замкнутое множество в B. Тогда справедлива

Теорема УШ.4.2. Если функционал Минковского P_M конечен на всех элементах пространства B_n и последовательность $\{B_n\}$ полна в точке x_* в помянутом выше смысле, то

$$\|x_* - x_*^{(n)}\| = O(\sqrt{E_n(x_*)}). \qquad (УШ.4.6)$$

4°. В настоящем пункте мы изложим, также без доказательств, некоторые из результатов работ Демьяновича [4]. В ней рассмотрена задача п.3 § 1 о минимуме расстояния от заданной точки $x_0 \in H$, где H - гильбертово пространство, до непустого выпуклого замкнутого в H множества M. Пусть H_n - подпространства пространства H; $x_*, x_*^{(n)}$ и $x^{(n)}$ имеют тот же смысл, что и во всем предшествующем. Обозначим

$$\forall x \in H,\ E_n(x) = \inf_{x^{(n)} \in H_n} |x - x^{(n)}|; \qquad (УШ.4.7)$$

кроме того, пусть P_n -ортопроектор из H на H_n. Доказывается следующая теорема:

Теорема УШ.4.3. Пусть множество $M_n = M \cap H_n \neq \emptyset$, $P_n x_* \in M_n$. Тогда имеет место оценка

$$|x_* - x_*^{(n)}| \le E_n(x_*) + [E_n^2(x_*) + E_n(x_*) E_n(x_0)]^{1/2}. \qquad (УШ.4.8)$$

Заметим, что если $E_n(x_*)$ и $E_n(x_0)$ суть величины одного порядка, то из (УШ.4.8) следует, что

$$|x_* - x_*^{(n)}| = O(E_n(x_*)) \qquad (УШ.4.9)$$

К сожалению, условие $P_n x_* \in M_n$ трудно проверяемо.

§ 5. Об устойчивости в точке решения односторонней вариационной задачи

1°. Здесь рассматривать задачу о минимальном расстоянии в

гильбертовом пространстве H от фиксированной точки x_0 этого пространства до непустого выпуклого замкнутого множества M :

$$|x - x_0| = \min, \quad x \in M. \qquad (\text{УШ.5.I})$$

Поставим вопрос (см. работу автора [29]) об устойчивости точного решения задачи (УШ.5.I) относительно малых изменений ее данных: элемента x_0 , нормы $|\cdot|$ пространства H и множества M . Этот вопрос, который кажется нам интересным и сам по себе, поможет нам потом (см. § 6) разобраться в погрешности искажения приближенного решения задачи (УШ.5.I). Эту последнюю мы будем ниже называть простейшей односторонней вариационной задачей.

2°. Пусть норма $|\cdot|$ и множество M остаются невозмущенными, а элемент x_0 заменен элементом $y_0 \in H$, причем $|x_0 - y_0| \le \delta$; через δ здесь и ниже в данном параграфе обозначается малое положительное число. Решение возмущенной таким образом задачи (УШ.5.I) обозначим через y_* . Так как, по обычному предположению, $x_0 \notin M$, то при достаточно малом δ будет также $y_0 \notin M$, а тогда $x_* \neq x_0$, $y_* \in y_0$, $x_* \in \partial M$, $y_* \in \partial M$.

Рассмотрим четырехугольник рис.49 Его четыре вершины определяют трехмерное (в частном случае – двумерное) пространство E , все элементы которого принадлежат H . Из доказательства леммы УШ.2.I вытекает, что углы четырехугольника в вершинах x_* и y_* – тупые или прямые, поэтому, если через точки x_* и y_* провести в E плоскости (прямые, если пространство E двумерное), нормальные к отрезку, соединяющему точки x_* и y_* , то x_0 и y_0 будут лежать вне или на границе слоя, ограниченного проведенными плоскостями. Отсюда следует, что

$$|x_* - y_*| \le |x_0 - y_0| \le \delta, \qquad (\text{УШ.5.2})$$

т.е. решение простейшей задачи устойчиво относительно малых возмущений элемента x_0 .

3°. В пространстве H заменим норму $|\cdot|$ другой гильбертовой нормой $|\cdot|_1$ так, чтобы

$$\forall x \in H, \quad (1-\delta)|x| \le |x|_1 \le (1+\delta)|x|. \qquad (\text{УШ.5.3})$$

Пространство H , снабженное метрикой $|\cdot|_1$, обозначим через H_1,

а решение возмущенной задачи $|x-x_0|_1 = \min$, $x \in M$ – через z_*. Оценим величину $|x_*-z_*|$. Через точки x_0, z_* проведем в H двумерную плоскость P и введем на ней евклидову метрику, индуцированную гильбертовой нормой $|\cdot|$. В обозначениях рис.5 имеем

$$|x_0-z_*|^2 = d^2 + |z_*-x_*|^2 - 2d|z_*-x_*|\cos\psi .$$

Очевидно, $\psi \geqslant \pi/2$ – в противном случае x_* не была бы точкой минимума, поэтому

$$|z_*-x_*|^2 \leqslant |x_0-z_*|^2 - d^2 . \qquad (\text{УШ.5.}$$

Аналогично, если на плоскости P ввести евклидову метрику, индуцированную гильбертовой нормой $|\cdot|_1$, и обозначить $|x_0-z_*|_1 = d_1$, то получим

$$|x_*-z_*|_1^2 \leqslant |x_0-z_*|_1^2 - d_1^2 . \qquad (\text{УШ.5.}$$

Если $d_1 \leqslant d$, то из (УШ.5.9) следует

$$|x_*-z_*|^2 \leqslant \frac{|x_0-z_*|_1^2}{(1-\delta)^2} - d^2 = \frac{d_1^2}{(1-\delta)^2} - d^2 \leqslant \frac{d^2(2-\delta)}{1-\delta}\delta ,$$

или

$$|x_*-z_*| \leqslant \frac{\sqrt{2-\delta}\,d}{1-\delta}\sqrt{\delta} ; \qquad (\text{УШ.5.6})$$

если же $d_1 > d$, то по неравенству (УШ.5.5)

$$|x_*-z_*|_1^2 \leqslant (1+\delta)d^2 - d_1^2 \leqslant (2+\delta)d^2\delta ,$$

откуда

$$|x_*-z_*| \leqslant \frac{|x_*-z_*|_1}{1-\delta} \leqslant \frac{\sqrt{2+\delta}\,d}{1-\delta}\sqrt{\delta} . \qquad (\text{УШ.5.7})$$

Из неравенств (УШ.5.6) и (УШ.5.7) следует, что

$$|x_*-z_*| \leqslant C_\delta\sqrt{\delta} , \quad C_\delta = \frac{\sqrt{2+\delta}}{1-\delta}d , \qquad (\text{УШ.5.8})$$

и, значит, решение простейшей задачи устойчиво относительно малых возмущений метрики пространства H.

4^0. Допустим теперь, что элемент x_0 и метрика пространства

H оставлены неизменными, а множество M заменено близким к нему множеством M', которое мы также предполагаем непустым, выпуклым и замкнутым в норме H. Близость множеств M и M' мы здесь определяем следующим требованием: существует такое биективное преобразование $S: M' \to M$, при котором

$$\forall x \in M,\ |x - S^{-1}x| \leqslant \delta(|x|+1). \qquad (\text{Ш.5.9})$$

Отметим одно следствие из неравенства (Ш.5.9). Пусть $x' \in M'$, тогда $Sx' \in M$ и, по неравенству (Ш.5.9)

$$|x' - Sx'| = |Sx' - S^{-1}Sx'| \leqslant \delta(|Sx'|+1).$$

Полагая в том же неравенства (Ш.5.9) $x = Sx'$, получаем $|Sx'| \leqslant (1-\delta)^{-1}(|x'|+1)$ и, следовательно,

$$\forall x' \in M',\ |x' - Sx'| \leqslant \frac{\delta}{1-\delta}(|x'|+1). \qquad (\text{Ш.5.10})$$

Наряду с задачей (Ш.5.1) рассмотрим задачу

$$|x' - x_0| = \min,\ x' \in M'; \qquad (\text{Ш.5.11})$$

ее решение обозначим через x'_*. Пусть, как и выше, $x_0 \notin M$. Докажем, что при достаточно малом δ будет $x_0 \notin M'$. Действительно, если $x_0 \in M'$, то $Sx_0 \in M$ и $|Sx_0 - x_0| \geqslant d$; как и выше, здесь $d = |x_* - x_0|$. С другой стороны, по неравенству (Ш.5.10) $|x_0 - Sx_0| \leqslant \delta(1-\delta)^{-1}(|x_0|+1)$, что противоречит предшествующему неравенству, если δ достаточно мало.

Обозначим $d' = \min |x' - x_0| = |x'_* - x_0|$. По неравенству (Ш.5.9)

$$d' \leqslant |x_* - S^{-1}x_*| \leqslant |x_0 - x_*| + |x_* - S^{-1}x_*| \leqslant d + \delta(|x_*|+1)$$

и, так как $S^{-1}x_* \in M'$, то $d' \leqslant |x_0 - S^{-1}x_*| \leqslant d + \delta(|x_*|+1)$.

Аналогично, так как $x'_* \in M'$, то $Sx'_* \in M$, и по неравенству (Ш.5.10)

$$d \leqslant |x_0 - Sx'_*| \leqslant |x_0 - x'_*| + |x' - Sx_*| \leqslant d' + \frac{\delta}{1-\delta}(|x'_*|+1).$$

Таким образом,

$$-\delta(|x_*|+1) \leqslant d - d' \leqslant \frac{\delta}{1-\delta}(|x'|+1). \qquad (\text{Ш.5.12})$$

Оценим разность $x_* - x'_*$. Имеем

$$|x_* - x'_*| \leqslant |x_* - S^{-1}x_*| + |S^{-1}x_* - x'_*| \leqslant \delta(|x_*|+1) + |S^{-1}x_* - x'_*|. \quad (УШ.5.$$

Точка $S^{-1}x_* \in M'$. По лемме УШ.2.I

$$|S^{-1}x_* - x'_*|^2 \leqslant |x_0 - S^{-1}x_*|^2 - d'^2 =$$

$$= [|x_0 - S^{-1}x_*| + d][|x_0 - S^{-1}x_*| - d']$$

и по неравенству (УШ.5.I2)

$$|S^{-1}x_* - x'_*|^2 \leqslant [\delta(1-\delta)^{-1}(|x'_*|+1) + \delta(|x_*|+1)] \times \quad (УШ.5.I$$

$$\times [2d + \delta(1-\delta)^{-1}(|x'_*|+1) + \delta(|x_*|+1)].$$

Обозначим $|S^{-1}x_* - x'_*| = \varepsilon$. Из неравенства (УШ.5.I3) получаем $|x'_*| \leqslant (1+\delta)(|x_*|+1) + \varepsilon$. Подставив это в (УШ.5.I4), приходим к квадратному неравенству для ε вида

$$(1-a\delta^2)\varepsilon^2 - 2b\delta\varepsilon - c\delta \leqslant 0; \quad (УШ.5.I$$

здесь a, b, c суть функции от δ и от $|x_*|$, ограниченные и положительные при фиксированной $|x_*|$ и при достаточно малых δ. Из (УШ.5.I5) следует, что $\varepsilon \leqslant C\delta^{1/2}$, $C = const$ и, в силу неравенства (УШ.5.I3),

$$|x'_* - x_*| \leqslant C'\sqrt{\varepsilon}, \quad C' = const. \quad (УШ.5.I6$$

Неравенство (УШ.5.I6) показывает, что решение задачи (УШ.5.I) устойчиво относительно возмущений множества M, малых в смысле неравенства (УШ.5.3).

5°. Исследуем устойчивость решения задачи (УШ.I.I–). В силу результатов п.2° § I достаточно ограничиться случаем, когда $\beta = |||\eta||| < d$. Устойчивость названной задачи мы исследуем в двух случаях, когда множество M' изменяется так, что это изменение естественно считать малым, но оно может не описаться соотношением вида неравенства (УШ.5.9). В настоящем пункте мы рассмотрим случай, когда M' – возмущенное множество – опреде-

лается формулой

$$M' = \{x \in H : \|x - g\| \leq d'\}, \qquad \text{(УШ.5.17)}$$

в которой d' – число, близкое к d. Заметим, что если $g = 0$ (к этому очевидным образом сводится более общий случай $g \in H$), то мы оказываемся в условиях п.4°: за S можно принять преобразование $\forall x' \in M'$, $Sx' = dx'/d'$. Ниже рассмотрен общий случай $g \in H$.

Пусть сначала $d' = d - \delta$. Обозначим

$$M_- = \{x \in H : \|x - g\| \leq d - \delta\}. \qquad \text{(УШ.5.18)}$$

ложим еще

$$\forall x \in M,\ Qx = (1-\sigma)x = x_-,\ \sigma > 0 \qquad \text{(УШ.5.19)}$$

и покажем, что при σ достаточно малом Q есть инъекция $M \to M_-$. Имеем, действительно,

$$\|x - x_-\| = \|(1-\sigma)(x-g) - \sigma g\| \leq (1-\sigma)d + \sigma\beta = d - \sigma(d-\beta)$$

достаточно принять $\sigma = \delta(d-\beta)^{-1}$; отметим еще, что $\sigma = c\delta$, $c = (d-\beta)^{-1} = const$ и что справедливо равенство

$$|x - Qx| = c\delta|x|. \qquad \text{(УШ.5.20)}$$

Наряду с задачей (УШ.1.1-2) рассмотрим задачу $|x - x_0| = \min$, $x \in M_-$, где M_- определено формулой (УШ.5.18). Решение этих задач соответственно обозначим через x_* и y_*. Пусть еще $d = |x_0 - x_*|$ и $d_- = |x_0 - y_*|$. Так как $M_- \subset M$, то $d \leq d_-$; кроме того, $x_* \in M$ и поэтому $Qx_* \subset M_-$. Отсюда, в силу неравенства (УШ.5.20),

$$d \leq d_- \leq |x_0 - Qx_*| \leq |x_0 - x_*| + |x_* - Qx_*| = d + c\delta|x_*|. \qquad \text{(УШ.5.21)}$$

Оценим разность $x_* - y_*$. Имеем

$$|x_* - y_*| \leq |x_* - Qx_*| + |Qx_* - y_*| = c\delta|x_*| + |Qx_* - y_*|. \qquad \text{(УШ.5.22)}$$

Точка $Qx_* \in M_-$; по лемме УШ.2.1 $|Qx_* - y_*|^2 \leq |x_0 - Qx_*|^2 - d_-^2$. Теперь по неравенству (УШ.5.21) $|Qx_* - y_*|^2 \leq c\delta|x_*|(2d + c\delta|x_*|)$. Подставив это в (УШ.5.22), получим

$$|x_* - y_*| \leq c|x_*|\delta + \sqrt{c|x_*|(2d + c\delta|x_*|)}\sqrt{\delta}\ ; \qquad \text{(УШ.5.23)}$$

если элемент x_* фиксирован, то отсюда получается более простое неравенство

$$|x_* - y_*| \le C\sqrt{\delta}\,, \quad C = const\,, \qquad \text{(УШ.5.24)}$$

которое доказывает устойчивость решения задачи (УШ.I.I-2) по отношению к малому уменьшению числа α, если $\|g\| < \alpha$.

Пусть теперь α заменено на $\alpha+\delta$. Обозначим

$$M_+ = \{x \in H : \|x - g\| \le \alpha + \delta\}\,. \qquad \text{(УШ.5.25)}$$

Если z_* – решение задачи

$$|x - x_0| = min\,, \quad x \in M_+\,, \qquad \text{(УШ.5.26)}$$

и $d_+ = |x_0 - z_*|$, то справедливы соотношения, аналогичные соотношениям (УШ.5.2I) и (УШ.5.23):

$$d_* \le d \le d_+ + C\delta\,|z_*|\,, \qquad \text{(УШ.5.27)}$$

$$|x_* - z_*| \le C\,|z_*|\,\delta + \sqrt{C\,|z_*|\,(2d + C\delta\,|z_*|)}\,\sqrt{\delta}\,. \qquad \text{(УШ.5.28)}$$

В неравенстве (УШ.5.28) заменим левую часть меньшей величиной $|z_*| - |x_0|$. Теперь легко получить для $|z_*|$ квадратное неравенство

$$|z_*|^2 - 2|z_*|\,\frac{|x_*|(1-C\delta) + 2Cd\delta}{1 - 2C\delta - C^2\delta^2} + \frac{|x_0|^2}{1 - 2C\delta - C^2\delta^2} \le 0\,,$$

из которого следует, что

$$|z_*| \le (1 - 2C\delta - C^2\delta^2)^{-1}\,[(1 - C\delta)\,|x_*| + 2Cd\delta +$$

$$+ \sqrt{2C^2\delta^2\,|x_*|^2 + 2Cd\delta\,|x_*| + 4C^2\delta^2 d^2}\,]\,.$$

При малых δ последний радикал есть величина порядка $O(|x_*|\delta + \sqrt{|x_*|\delta} + \delta)$ и, следовательно, $|z_*| \le |x_*| + O(\sqrt{|x_*|\delta})$. Подставив это в (УШ.5.28), убедимся, что при фиксированном x_*

$$|x_* - z_*| = O(\sqrt{\delta})\,; \qquad \text{(УШ.5.29)}$$

тем самым доказана устойчивость решения рассматриваемой задачи относительно малого увеличения числа α, если $\|g^m\| < \alpha$.

6°. Принимая по-прежнему, что $\|g\|<d$, допустим, что g заменено элементом g', таким, что $\|g-g'\|\le\delta$; остальные данные не изменены. Имеем новую задачу

$$|x-x_0|=\min,\ x\in M',\ M'=\{x\in H:\|x-g'\|\le d\}. \qquad (УШ.5.30)$$

При этом

$$\|x-g\|-\|g-g'\|\le\|x-g'\|\le\|x-g\|+\|g-g'\|,$$

или

$$\|x-g\|-\delta\le\|x-g'\|\le\|x-g\|+\delta. \qquad (УШ.5.31)$$

Обозначим, как и выше,

$$M_\pm=\{x\in H,\ \|x-g\|\le d\pm\delta\}. \qquad (УШ.5.32)$$

Пусть $x\in M_-$. Тогда $\|x-g'\|\le\|x-g\|+\delta\le d$, т.е. $x\in M'$. Таким образом, $M_-\subset M'$. Точно так же доказывается, что $M'\subset M_+$. Окончательно, неравенство (УШ.5.31) означает, что

$$M_-\subset M'\subset M_+. \qquad (УШ.5.33)$$

При расширении множества минимум уменьшается, поэтому (смысл обозначений очевиден)

$$d_+\le d'\le d_-,\ d_+\le d\le d_-. \qquad (УШ.5.34)$$

Из формул (УШ.5.21) и (УШ.5.27) следует, что $d_--d_+=O(\delta)$, а тогда тот же порядок имеют и разности d_--d', d_--d, $d'-d_+$, $d-d_+$. Пусть, далее, x'_* и z_* – минимизирующие элементы, соответствующие множествам M' и M_+. Так как $M'\subset M_+$, то $x'_*\in M_+$; по лемме УШ.2.1, $|x'_*-z_*|^2\le d'^2-d_+^2$ и величина $|x'_*-z_*|=O(\sqrt{\delta})$. Точно так же $x_*\in M\subset M_+$, и $|x_*-z_*|=O(\sqrt{\delta})$. Теперь

$$|x_*-x'_*|\le|x_*-z_*|+|z_*-x'_*|=O(\sqrt{\delta}); \qquad (УШ.5.35)$$

решение задачи (УШ.I.I-2) оказывается устойчивым по отношению к малым (в метрике $\|\cdot\|$) возмущениям элемента g.

7°. Пусть одновременно мало возмущены все параметры задачи (УШ.I.I-2): элемент x_0 заменен на x_1, норма $|\cdot|$ – на $|\cdot|_1$, число d заменено на $d=d\pm\delta$ и, наконец, элемент g заменен близким,

в метрике $\|\cdot\|$, элементом q'. По-прежнему пусть x_* означает решение невозмущенной задачи (УШ.I.I-2); через x_{1*}, x_{2*}, x_{3*}, x_{4*} обозначим решения следующих возмущенных задач: x_{1*} соответствует элементу x_1 при остальных невозмущенных параметрах, x_{2*} соответствует элементу x_1 и норме $|\cdot|_1$ при невозмущенных d и q, x_{3*} – решение, полученное при невозмущенном q и остальных возмущенных параметрах и, наконец, x_{4*} соответствует возмущенным значениям всех четырех параметров. По-прежнему положим $d=|x_0-x_*|$; пусть, кроме того, $d_i=|x_i-x_{i*}|$, $i=1, 2, 3, 4$. Оценим разность x_*-x_{4*}. Имеем

$$|x_*-x_{4*}|\le\sum_{i=1}^{4}|x_{i-1,*}-x_{i*}|,\quad x_{0*}=x_*;$$

в силу неравенств (УШ.5.2-3)

$$|x_*-x_{4*}|\le\delta+\sum_{i=2}^{4}|x_{i-1,*}-x_{i*}|. \qquad (УШ.5.$$

Как видно из формул (УШ.5.6), (УШ.5.29) и (УШ.5.35), оценки норм в (УШ.5.36) определяются через оценки величин d_i в зависимости от d. Найдем эти последние оценки. Очевидно,

$$d_1\le|x_1-x_0|+|x_0-x_*|+|x_*-x_{1*}|\le d+2\delta; \qquad (УШ.5.$$

по формулам (УШ.5.3), (УШ.5.8) и (У.5.37)

$$d_2=|x_1-x_{2*}|\le(1-\delta)^{-1}|x_1-x_2|_1\le(1-\delta)^{-1}(|x_1-x_{1*}|_1+|x_{1*}-x_{2*}|_1)\le \qquad (УШ.5.$$

$$\le(1+\delta)(1-\delta)^{-1}(d+2\delta+C_6(d+2\delta)\sqrt{\delta})=d+O(\sqrt{\delta}).$$

Оценим величину d_3. По формуле (УШ.5.27), в случае $d'=d+\delta$ будет $d_3=d_2+O(\delta)=d+O(\sqrt{\delta})$; то же соотношение имеет место и для $d'=d-\delta$ – это следует из формулы (УШ.5.21). Теперь из (УШ.5.34) вытекает, что $d_4=d_3+O(\sqrt{\delta})=d+O(\sqrt{\delta})$.

Легко проверить, что все слагаемые в (УШ.5.36) суть величины порядка $O(\sqrt{\delta})$; это означает, что решение x_* устойчиво относительно малых возмущений всех параметров, перечисленных в начале настоящего пункта.

Аналогичными рассуждениями доказывается устойчивость решения задачи (УШ.5.I) относительно одновременных малых возмуще

элемента x_0, нормы $|\cdot|$ и множества M, если последнее возмущение соответствует условиям п. 4°.

§ 6. Оценка погрешности искажения

1°. Пусть $a(x,y)$ симметричная билинейная форма, определенная на некотором линейном множестве $\mathcal{L}$, и соответствующая ей квадратичная форма $a(x) = a(x,x)$ положительна, - это значит, что $a(x) \geqslant 0$ и $a(x) = 0 \Rightarrow x = 0$. Введя скалярное произведение $[x,y] = a(x,y)$ и норму $|x| = \sqrt{a(x)}$, мы превратим множество $\mathcal{L}$ в предгильбертово пространство; замкнув последнее в указанной норме, получим гильбертово пространство, которое обозначим через H. Примем, что H сепарабельно; для упрощения последующих записей примем еще, что оно вещественно.

2°. Пусть $f(x)$ - линейный ограниченный в H функционал, определенный на всем пространстве, и $M \subset H$ - непустое выпуклое множество, замкнутое в норме H. Задача

$$a(x) - 2f(x) = |x|^2 - 2f(x) = \min, \quad x \in M \qquad (\text{УШ.6.1})$$

имеет одно и только одно решение: она равносильна задаче

$$|x - x_0| = \min, \quad x \in M, \qquad (\text{УШ.6.2})$$

где x_0 - элемент пространства H, определяемый тождеством

$$\forall x \in H, \; f(x) = [x, x_0]; \qquad (\text{УШ.6.3})$$

существование и единственность элемента x_0 вытекают из известной теоремы Ф.Рисса.

3°. Приближенное решение задачи (УШ.3.1) можно строить, например, по методу Ритца (в книге Гловинского, Лионса и Тремольера [1] вместо термина "метод Ритца" используется термин "внутренние методы"). Зададим полную в H последовательность конечномерных подпространств $\{H_n\}$, и пусть $(y_{n1}, y_{n2}, \ldots, y_{nN})$, $N = N(n)$ - базис подпространства H_n. Задачу (УШ.6.1) заменим следующей

$$|x^{(n)}|^2 - 2f(x^{(n)}) = \min, \quad x^{(n)} \in H_n \cap M, \qquad (\text{УШ.6.4})$$

если, что то же,

$$|x^{(n)} - x_0| = \min , \quad x^{(n)} \in H_n \cap M . \qquad (\text{Ш}.6.$$

Если пересечение $H_n \cap M$ непусто, то задача (Ш.6.4) имеет одно и только одно решение.

4°. Пусть

$$x = \sum_{k=1}^{N} a_k^{(n)} \varphi_{kn} , \qquad (\text{Ш}.6.$$

тогда

$$|x^{(n)}|^2 = \sum_{j,k=1}^{N} [\varphi_{jn}, \varphi_{kn}] a_j^{(n)} a_k^{(n)} . \qquad (\text{Ш}.6.$$

Скалярные произведения $[\varphi_{jn}, \varphi_{kn}]$ вычисляются с некоторыми погрешностями, которые можно считать сколь угодно малыми. Обозначим погрешность вычисления величины $[\varphi_{jn}, \varphi_{kn}]$ через $\gamma_{jk}^{(n)}$ и положим $[\varphi_{jn}, \varphi_{kn}] + \gamma_{jk}^{(n)} = d_{jk}^{(n)}$; числа $d_{jk}^{(n)}$ нам известны. Обозначим соответственно через M_n и Γ_n матрицы элементов $[\varphi_{j\cdot}, \varphi$ и $\gamma_{jk}^{(n)}$; $j, k = 1, 2, \ldots, N$, и через $\|M_n\|, \|\Gamma_n\|$ нормы этих матриц как операторов в R_N.

Матрица M_n симметрична, и мы вправе считать, что матрица Γ_n также симметрична. Можно считать также, что матрица $M_n + \Gamma_n$ положительно определенная – это будет всегда, когда Γ_n достаточно мала по отношению к M_n.

Если $x^{(n)}$ – произвольный элемент из H_n,

$$x^{(n)} = \sum_{k=1}^{N} c_k^{(n)} \varphi_{kn} ,$$

то по формуле (Ш.6.7)

$$|x^{(n)}|^2 = (M_n c^{(n)}, c^{(n)}),$$

где $c^{(n)} = (c_1^{(n)}, c_2^{(n)}, \ldots, c_N^{(n)})$. Искаженная матрица $\tilde{M}_n := M_n + \Gamma_n$ порождает в H_n новую норму

$$|x^{(n)}|_1^2 := (\tilde{M}_n c^{(n)}, c^{(n)}) = |x^{(n)}|^2 + (\Gamma_n c^{(n)}, c^{(n)}).$$

Примем, что матрица Γ_n мала по сравнению с M_n в следующем смысле:

$$\forall c^{(n)} \in R_N , \quad \frac{|(\Gamma_n c^{(n)}, c^{(n)})|}{(M_n c^{(n)}, c^{(n)})} \leq \delta_1 , \qquad (\text{Ш}.$$

где δ_1 – малое число. Заметим, что условие (УШ.6.8) выполнено, если

$$\|\Gamma_n\| \cdot \|M_n^{-1}\| \leqslant \delta_1 . \qquad (УШ.6.9)$$

Действительно, обозначим $M_n^{1/2} c^{(n)} = \rho^{(n)}$, тогда отношение (УШ.6.8) принимает вид

$$\frac{|(\Gamma_n M_n^{-1/2} \rho^{(n)}, M_n^{-1/2} \rho^{(n)})|}{\|\rho^{(n)}\|^2} \leqslant \|\Gamma_n\| \cdot \|M_n^{-1/2}\|^2 = \|\Gamma_n\| \cdot \|M_n^{-1}\|.$$

Если соблюдено условие (УШ.6.8), то

$$(1-\delta)|x^{(n)}| \leqslant |x^{(n)}|_1 \leqslant (1+\delta)|x^{(n)}| , \qquad (УШ.6.10)$$

где $\delta = max(\sqrt{1+\delta_1}-1 , 1-\sqrt{1-\delta_1}) = \delta_1/(1+\sqrt{1-\delta_1})$

5°. Положим $x_0 = x_{0n} + x_{1n}$, где $x_{0n} \in H_n$ и $x_{1n} \perp H_n$. Задача (УШ.6.5) равносильна такой задаче:

$$|x^{(n)} - x_{0n}| = min , \; x^{(n)} \in H_n \cap M ; \qquad (УШ.6.11)$$

элемент x_{0n} можно построить как проекцию x_0 на H_n , т.е., как решение экстремальной задачи

$$|x_0 - \sum_{k=1}^{N} b_k y_{kn}|^2 = min$$

что приводит к системе уравнений

$$\sum_{k=1}^{N} b_k [y_{in}, y_{jn}] = [x_0, y_{jn}] = f(y_{jn}), \; 1 \leqslant j \leqslant N . \qquad (УШ.6.12)$$

Эту систему можно записать в виде

$$[x_{0n}, y_{jn}] = f(y_{jn}), \; 1 \leqslant j \leqslant n ;$$

таким образом, x_{0n} есть элемент, даваемый теоремой Ф.Рисса, если функционал f задан на H_n.

При составлении системы (УШ.6.12) будут допущены погрешности: вместо $[y_{kn}, y_{jn}]$ получатся коэффициенты $[y_{kn}, y_{jn}] + \alpha_{kj}^{(n)} = d_{kj}$; кроме того, числа $f(y_{jn})$ также будут вычислены с погрешностью, пусть последние равны $\delta_j^{(n)}$. Вместо системы (УШ.6.12) мы на самом деле получим искаженную систему

$$\widetilde{M}_n \widetilde{b}^{(n)} = f^{(n)} + \delta^{(n)} . \qquad (УШ.6.13)$$

Здесь $\tilde{b}^{(n)}$ – искаженный вектор коэффициентов,

$$\tilde{b}^{(n)} = (\tilde{b}_1^{(n)}, \tilde{b}_2^{(n)}, \dots, \tilde{b}_N^{(n)}); \qquad \delta^{(n)} = (\delta_1^{(n)}, \delta_2^{(n)}, \dots, \delta_N^{(n)}). \qquad \text{(УШ.6.}$$

Допустим, что линейный вычислительный процесс определени элемента x_0 по методу Ритца устойчив в последовательности (R_N); напомним, что для этого необходимо и достаточно, чтобы координатная система была сильно минимальна в H . Тогда суще вуют такие положительные постоянные ρ , q , τ , что если $\|\Gamma_n\|$ то

$$|y_{on} - x_{on}| \le \rho \|\Gamma_n\| + q \|\delta^{(n)}\|;$$

здесь y_{on} – искаженное значение элемента x_{on} . Пусть $\|\Gamma_n\|$ $\|\delta^{(n)}\|$ достаточно малы. Тогда $|y_{on} - x_{on}| < \delta$, где δ – малое чис не зависящее от n . Одновременно, в силу необходимых условий тойчивости, нормы $\|A_n^{-1}\|$ равномерно ограничены, и неравенства (УШ.6.9-10) также выполнены. Таким образом, искажение сводите тому, что при каждом n возмущаются элемент x_{on} и норма в пр пространстве H_n , и эти возмущения малы равномерно относитель n . По доказанному в § 5, искажение элемента x_{on} приводит к искажению решения на $O(\|\Gamma_n\| + \|\delta^{(n)}\|)$, а искажение нормы – к и кажению решения на $O(\sqrt{\|\Gamma_n\| \cdot \|A_n^{-1}\|}) = O(\sqrt{\|\Gamma_n\|})$. Окончате но, если $y_*^{(n)}$ – искаженное решение n-й приближенной задачи

$$\|y_*^{(n)} - x_*^{(n)}\| = O(\sqrt{\|\Gamma_n\|} + \|\delta^{(n)}\|. \qquad \text{(УШ.6.}$$

Как показывает неравенство (УШ.6.15), с целью избежать больш искажений следует вычислять матрицу Ритца точнее, чем свобод члены системы Ритца – так, чтобы $\|\Gamma_n\| = O(\|\delta^{(n)}\|^2)$.

Можно исследовать устойчивость процесса вычисления коэф циентов $a_k^{(n)}$ не в последовательности (R_N , R_N), а в последов тельности ($\bar{X}_N$, $\bar{Y}_N$) (см. формулу Ш.3.4). Сформулируем результ упомянутый процесс устойчив и при достаточно малых нормах $\|\Gamma_n\|_{\bar{X}_N \to \bar{Y}_N}$ верна оценка

$$\|y_*^{(n)} - x_*^{(n)}\|_{\bar{Y}_N} = O(\sqrt{\|\Gamma_n\|_{\bar{X}_N \to \bar{Y}_N}} + \|\delta^{(n)}\|_{\bar{Y}_N}). \qquad \text{(УШ.3.}$$

Эта оценка прагедна, в частности, для метода конечных элементов.

Дополнение к главе УШ

В настоящем дополнен и, написанном В.Б.Тихтиным излагаются результаты некоторых работ последних лет, посвященных оценкам аппроксимации для решений односторонних вариационных задач, отличных, как правило, от простейшей задачи, изученной в данной главе. Приближенные решения в этих работах обычно строятся по МКЭ.

В Дополнении приняты несколько иные термины и обозначения, чем в основном тексте гл.УШ. Соболевские пространства $W_2^{(k)}$ и $\overset{\circ}{W}{}_2^{(k)}$ обозначаются соответственно через H^k и $\overset{\circ}{H}{}_0^k$. Множество ограничений обозначается через $\mathcal{K}$. Если ограничительное условие имеет вид $u(x) \geqslant \chi(x)$ или $u(x) \geqslant \chi(x)$, то множество ограничений называется препятствием; различаются препятствия в области и на границе, в зависимости от того, где выполняется упомянутое неравенство. Далее, x означает точку евклидова пространства; функции, заданные в области евклидова пространства, чаще всего обозначаются буквами u и v. Задачи, как правило, формулируются в терминах вариационных неравенств.

I°. В статье Балка [I] рассматривается плоская задача о препятствии в области. Пусть $\Omega \subset R_2$ – выпуклая область с границей класса $C^{(2)}$, и (по повторяющимся индексам производится суммирование от I до 2)

$$a(u,v)=\int_\Omega \left(a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_j}+cuv\right)dx,$$

где

$$a_{ij}(x)\in C^{(1)}(\bar\Omega),\ a_{ij}=a_{ji},\ c(x)\in L_\infty(\Omega);$$
$$a_{ij}\xi_i\xi_j \geqslant d|\xi|^2,\ d>0,\ c(x)\geqslant\lambda>0.$$

Для $f\in L_2(\Omega)$ рассматривается задача об отыскании такой функции $u\in\mathcal{K}$, что

$$a(u,v-u)\geqslant\int_\Omega f(v-u)dx,\ \forall v\in\mathcal{K}, \qquad (\text{УШ.I})$$

где

$$\mathcal{K}=\left\{v\in H_0^1(\Omega): v(x)\geqslant\chi(x)\ \text{ почти всюду в }\Omega\right\};$$

$\chi \in H^2(\Omega)$ - заданная функция, $\chi|_{\partial\Omega}=0$.

При перечисленных предположениях решение задачи принадлежит к классу $H^2(\Omega)$ (см. Брезис и Стампаккья [1]).

Для приближенного решения задачи (УШ.1) применяется МКЭ с кусочно линейными координатными функциями. Пусть Ω_h - вписанный в Ω многоугольник, длины сторон которого не превосходят h. Этот многоугольник разбивается на треугольники со сторонами $\leq h$. Через V_h обозначим пространство функций, заданных на Ω_h, непрерывных, кусочно линейных и равных нулю на $\Omega \setminus \Omega_h$. Для всякой функции $v \in C(\bar{\Omega})$ равной нулю на $\partial\Omega$, обозначим через v_I ее интерполянт, т.е., функцию из V_h, совпадающую с v в вершинах триангуляции. Приближенное решение задачи (УШ.1) определяется как решение задачи

$$u_h \in K_h;\ a(u_h, v_h - u_h) \geq \int_\Omega f(v_h - u_h)dx;\ \forall v_h \in K_h, \qquad (УШ.2)$$

где $K_h = \{v_h \in V_h : v_h \geq \chi_I$ в узлах триангуляции$\}$. При некоторых "условиях регулярности" описанного здесь варианта МКЭ устанавливается оценка

$$\|u - u_h\|_{H^1(\Omega)} \leq ch\,\|u\|_{H^2(\Omega)}. \qquad (УШ.3)$$

В работе [2] Фалк обобщил результаты работы [1] на случай невыпуклой области $\Omega \subset R_2$. Предположим, что $\partial\Omega \in C^{(2)}$ и имеет лишь конечное число точек, в которых кривизна меняет знак. Тогда оценка (УШ.3) по-прежнему имеет место.

2^0. В статье Скарпини и Вивальди [1] рассмотрена плоская задача с препятствием на границе. Пусть $\Omega \in R_2$ - выпуклая ограниченная область с границей класса $C^{(2)}$, $a(u, v)$ - билинейная форма, такая же, как в п.1^0, $f \in L_2(\Omega)$, $\psi \in H^2(\Omega)$. Рассматривается задача (УШ.1), но с другим препятствием, именно,

$$K = \{v \in H^1(\Omega) : v \geq \psi \text{ на } \partial\Omega\}. \qquad (УШ.4)$$

Известно (см. Брезис [1]), что при указанных условиях решение задачи (УШ.1), (УШ.4) $u \in H^2(\Omega)$.

Для приближенного решения задачи применен метод, близкий к методу п.1^0. Вписан в Ω многоугольник Ω_h, длины сторон которого не превосходят h. Разобьем Ω_h на треугольники. Обозначим через $I_0 = \{1, 2, \ldots, N_0\}$ индексы внутренних узлов, через $I_1 =$

$= \{N_0+1, \ldots, N\}$ - индексы граничных узлов; сами узлы обозначим через x_i ; положим еще $I = I_0 \cup I_1$. Далее, пусть Γ_i - часть $\partial\Omega$ между узлами x_i и x_{i+1} , Σ_i - часть Ω , ограниченная отрезком $[x_i, x_{i+1}]$ и кривой Γ_i .

Если T_i - треугольник со стороной $[x_i, x_{i+1}]$, $i \in I_1$, то объединение $T_i \cup \Sigma_i$ рассматривается как криволинейный элемент.

Через $\varphi_i^h(x)$, $x \in \Omega$, $i \in I$, обозначена непрерывная в $\bar{\Omega}$ функция, линейная на каждом треугольнике или криволинейном элементе, равная единице в узле x_i и нулю во всех остальных узлах. Пусть

$$H_h^1(\Omega) = \{v_h : v_h = \sum_{i \in I} v_i^h \varphi_i^h(x),\ v_i^h \in R_1,\ \forall i \in I\}. \qquad (\text{Ш}.5)$$

Приближенная задача имеет вид (Ш.2), но $\mathcal{K}_h$ определяется формулой

$$\mathcal{K}_h = \{v_h \in H_h^1 : v_h(x_i) \geq \psi(x_i),\ \forall i \in I\}. \qquad (\text{Ш}.6)$$

При условии регулярности использованного МКЭ верна оценка

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^{3/4} |u|_{H^2(\Omega)}. \qquad (\text{Ш}.7)$$

3°. В статье Брецци, Хагера и Рагьяра [1] рассматривается конечно-элементная аппроксимация задачи с препятствием в области или на границе для некоторых нелинейных форм.

Задача с препятствием в области ставится в форме

$$u \in \mathcal{K};$$

$$\int_\Omega [\nabla u \cdot \nabla(v-u) + u(v-u)]\,dx \geq \int_\Omega f(v-u)\,dx;\ \forall v \in \mathcal{K}; \qquad (\text{Ш}.8)$$

$$\mathcal{K} = \{v \in H^1(\Omega) : v \geq \psi \ \text{ в } \Omega\ ,\ v|_{\partial\Omega} = g\}.$$

Здесь Ω - выпуклая область класса $C^{(1,1)}$; $g, \psi \in H^2(\Omega)$, $\psi|_{\partial\Omega} \leq g$, $f \in L_2(\Omega)$.

Пусть Ω_h - вписанный в Ω многоугольник со сторонами, по длине не превосходящими h , и пусть этот многоугольник регулярным образом триангулирован. Для каждого треугольника триангуляции T_i , прилегающего к $\partial\Omega_h$, построим треугольник T_i' зеркальное отражение треугольника T_i в его стороне, принадлежащей

$\partial\Omega_h$. Пусть $R_{i_\nu} = UT_i'$. Будем считать h настолько малым, что $(\Omega_h \cup R_h) \subset \Omega$. Обозначим через $\hat{\Omega}$ некоторую область, содержащую $\Omega_h \cup R_h$ при любом достаточно малом h и через V_h – множество непрерывных кусочно линейных на $\Omega_h \cup R_h$ функций. Продолжим решение $u \in H^2(\Omega)$ задачи (УШ.7) до функции $\tilde{u} \in H^2(\hat{\Omega})$. Пусть $\tilde{u}_I$, ψ_I – интерполянты функций $\tilde{u}$ и ψ соответственно на $\Omega_h \cup R_h$. Множество $\mathcal{K}_h$ определяется следующим образом:

$$\mathcal{K}_h = \{v_h \in V_h : v_h \geqslant \psi_I \text{ в узлах на } \Omega_h, \quad v_h = \tilde{u}_I \text{ на } \Omega \setminus \Omega_h\}.$$

Приближенное решение u_h определяется согласно (УШ.7) с заменой $\mathcal{K}$ на $\mathcal{K}_h$. Устанавливается оптимальная по порядку оценка погрешности аппроксимации:

$$\|u - u_h\|_{H^1(\Omega)} = O(h). \qquad \text{(УШ.9)}$$

Наряду с кусочно линейными координатными функциями для приближенного решения задачи (УШ.7) применяются также кусочно квадратичные функции. В этом случае, в предположении, что точное решение $u \in H^{5/2-\varepsilon}(\Omega)$, $\varepsilon > 0$, устанавливается оценка

$$\|u - u_h\|_{H^1(\Omega)} = O(h^{3/2-\varepsilon}). \qquad \text{(УШ.10)}$$

Задача о вращательном движении на границе также ставится в форме (УШ.8), но $\mathcal{K}$ определяется формулой $\mathcal{K} = \{v \in H(\Omega), v|_{\partial\Omega} \geqslant g\}$.

Пространство V_h определяется следующим образом. Если $v_h \in V_h$, то v_h непрерывна и кусочно линейна на Ω_h ; при этом, если точка $Q \in \Omega \setminus \Omega_h$, то $v_h(Q) = v_h(P)$, где P – проекция Q на $\partial\Omega_h$.

Положим в данном случае

$$\mathcal{K}_h = \{v_h \in V_h : v_h \geqslant g \text{ в узлах на } \partial\Omega\}.$$

Кроме того решение строится как решение задачи (УШ.8), в которой $\mathcal{K}$ заменено на $\mathcal{K}_h$. В предположении, что точное решение $u \in H^2(\Omega)$, а $u, g \in W_\infty^1$ вблизи $\partial\Omega$ и свободная граница состоит из конечного числа точек, устанавливается оптимальная по порядку оценка погрешности аппроксимации:

$$\|u - u_h\|_{H^1(\Omega)} = O(h). \qquad \text{(УШ.11)}$$

Таким образом, при более сильных ограничениях на решение получена более высокая по порядку оценка погрешности сравнительно с работой Скарпини и Ливальди [ ].

4°. В статье Ниче [1] рассматривается задача о минимуме функционала

$$I(v)=\int_\Omega |\nabla v|^2 dx - 2\int_\Omega fv\,dx \qquad (\text{ш.12})$$

на множестве $\mathcal{K}=\{v : v\in H_0^1(\Omega),\ v\le\chi$ почти всюду в $\Omega\}$, где $\Omega\in R_2$ – область с достаточно гладкой границей, $\chi\in W_\infty^{(2)}(\Omega)$ и χ, $f\in L_\infty(\Omega)$ – заданные функции.

Производится регулярная триангуляция с параметром h, области Ω и строится пространство S_h непрерывных кусочно линейных функций, равных нулю на $\partial\Omega$. Пусть $\mathcal{K}_h=\mathcal{K}\cap S_h$. Для задачи (ш.12) приближенная задача состоит в минимизации функционала $I(v)$ на множестве $\mathcal{K}_h$. Пусть u и u_h соответственно точное и приближенное решение. Если $u\in W_\infty^{(2)}(\Omega)$, то устанавливается следующая оценка погрешности аппроксимации:

$$\|u-u_h\|_{\infty(\Omega)} \le Ch^2|\ln h|\cdot\|u\|_{W_\infty^{(2)}(\Omega)}, \qquad (\text{ш.13})$$

где $C>0$ не зависит от h.

5°. В статье Мительмана [1] рассматривается задача о приближении в области для некоторого неквадратичного функционала. А именно, пусть $\Omega\subset R_n$, $\mathcal{K}=\{v\in W_2^{(1)}(\Omega): v(x)\ge\psi(x)$ в Ω, $v|_{\partial\Omega}=f\}$, где ψ и f заданные функции, причем $\psi|_{\partial\Omega}\le f$. Рассматривается задача о минимизации функционала

$$\int_\Omega \mathcal{F}(\nabla u)\,dx \qquad (\text{ш.14})$$

на множестве $\mathcal{K}$. В интеграле (ш.14) $\mathcal{F}\in C^{(2)}(R_n)$.

Обозначим $a(p)=\nabla\mathcal{F}(p)=(a_1(p),a_2(p),\dots,a_n(p))$, $p\in R_n$ и допустим, что $(a(p)-a(q))\cdot(p-q)\ge C\max(|p|,|q|)|p-q|$ $\forall p, q\in R_n$, $|\cdot|$ – евклидова норма в R_n. Задача (ш.14) равносильна задаче

$u\in\mathcal{K}$;

$$a(u,v-u):=\int_\Omega a_i(\nabla u)\frac{\partial}{\partial x_i}(v-u)dx \geq 0\ ,\quad \forall v\in\mathcal{K}\ ;$$

эта последняя аппрокс мируется по МКЭ с кусочно линейными
диатными функциями. Пусть $\partial\Omega\in C^{(2)}$. Обозначим через $T_1, T_2,$
T_m симплексы разбиения и через Ω_h – объединение $\bigcup_{i=1}^{m} T_i$.
h обозначен параметр, которого не превосходят длины ребер
плексов T_i . Предполагается, что внутренности симплексов T_i
пересекаются, и что каждая грань любого симплекса либо есть
грань другого симплекса, либо ее вершины лежат на $\partial\Omega$.
V_h обозначено пространство функций, непрерывных на Ω_h и ли
ных на каждом симплексе T_i . Обозначим еще через $\mathcal{N}=\mathcal{N}_h$ мно
во узлов в $\bar{\Omega}_h$. Пусть

$$\mathcal{K}_h=\{v_h\in V_h : v_h(x)\geq\psi(x),\ x\in\Omega\cap\mathcal{N};\ v_h(x)=f(x),\ x\in\partial\Omega\cap\mathcal{N}\}$$

Приближенное решение определяется как решение следующей за

$$u_h\in\mathcal{K}_h,$$

$$\int_{\Omega_h} a_i(\nabla u_h)\cdot\frac{\partial}{\partial x_i}(v_h-u_h)dx\geq 0\ ,\quad \forall v_h\in\mathcal{K}_h\ ;$$

в предположении, что $u\in H^{(2)}(\Omega)\cap W_\infty^{(1)}(\Omega)$, уста
ется оценка погрешности

$$\|u-u_h\|_{H^1(\Omega_h)}=O(h). \qquad (\ldots$$

6°. В работе Ломинского и Марроко [1] рассмотрена сл
щая задача. Пусть Ω – выпуклая область в R_n и $p>1$. Треб
найти функцию $u\in\mathring{W}_p^{(1)}(\Omega)$, такую, что

$$J(u)=\inf J(v),\quad v\in\mathring{W}_p^{(1)}(\Omega),$$

$$J(v)=\frac{1}{p}\int_\Omega|\nabla u|^p dx-\int_\Omega fv\,dx;\ f\in W_{p'}^{(-1)}(\Omega),\ \frac{1}{p}+\frac{1}{p'}=1.$$

Пусть V_h – семейство конечномерных подпространств в $\mathring{W}$
Приближенное решение задачи (...16) определим как решение
чи

$$J(u_h)=\inf_{v_h\in V_h} J(v_h).$$

задачи (УШ.16) и (УШ.17) имеют единственные решения u и u_h соответственно (см., например, Фикс [2]).

Пусть $E_h(u) = \inf_{v_h \in V_h} \| v_h - u \|_{\mathring{W}_\rho^{(1)}(\Omega)}$. Устанавливаются следующие оценки погрешности аппроксимации:

$$\| u_h - u \|_{\mathring{W}_\rho^{(1)}(\Omega)} \leqslant C [E_h(u)]^{1/(\rho-1)}, \quad \rho \geqslant 2 ; \qquad (\text{УШ.18})$$

$$\| u_h - u \|_{\mathring{W}_\rho^{(1)}(\Omega)} \leqslant C [E_h(u)]^{1/(3-\rho)}, \quad 1 < \rho < 2 ,$$

где C не зависит от h.

Оценки (УШ.18) улучшены в работе Тюленина [2] :

$$\| u_h - u \|_{\mathring{W}_\rho^{(1)}(\Omega)} \leqslant C [E_h(u)]^{2/\rho}, \quad \rho \geqslant 2 ; \qquad (\text{УШ.19})$$

$$\| u_h - u \|_{\mathring{W}_\rho^{(1)}(\Omega)} \leqslant C [E_h(u)]^{\rho/2}, \quad 1 < \rho < 2 .$$

7°. Пусть Ω - единичный квадрат на плоскости: $\Omega = [0,1] \times [0,1]$; в Ω заданы функции u_0, v_0, причем $u_0 \leqslant v_0$ почти всюду в Ω, $u_0|_{\partial\Omega} \leqslant 0 \leqslant v_0|_{\partial\Omega}$, $\rho > 2$

$$\mathcal{J}(u) = \frac{1}{\rho} \sum_{i=1}^{2} \int_\Omega \left| \frac{\partial u}{\partial x_i} \right|^\rho dx - \int_\Omega f u \, dx. \qquad (\text{УШ.20})$$

В работе Скарпини [1] рассматривается, в частности, проблема оценки приближенного решения для задачи минимизации функционала (УШ.20) на множестве

$$\mathcal{K} = \left\{ v \in \mathring{W}_\rho^{(1)}(\Omega) : u_0(x) \leqslant v(x) \leqslant v_0(x) \quad \text{почти всюду в } \Omega \right\} .$$

Пусть u - точное решение. Если $f \in L_{\rho'}(\Omega)$; $u_0, v_0 \in W_\rho^{(1+\sigma)}(\Omega)$, $\sigma = 1/(\rho - 1)$, то $u \in W_\rho^{(1+\sigma)}(\Omega)$ (см. Эль Колли [1], Коско [1]). Пусть u_h - приближенное решение, полученное по МКЭ при кусочно линейных координатных функциях. Доказывается, что

$$\| u_0 - u_h \|_{\mathring{W}_\rho^{(1)}(\Omega)} \leqslant C h^{\sigma^2}, \qquad (\text{УШ.21})$$

где C не зависит от h.

8°. Статья Эллиотта [1] посвящена следующей задаче. Пусть $\Omega \subset R_2$ - выпуклый многоугольник. Требуется найти такую функцию $u \in H^1(\Omega)$, что

$$\mathcal{J}(u) := \alpha\int_\Omega u^2 dx + \beta\int_\Omega |\nabla u|^2 + j(u) - \int_\Omega f u dx = \min, \qquad (\text{УШ.22})$$

или, что то же,

$$\alpha(u, v-u) + \beta a(u, v-u) + j(v) - j(u) \geq f(v-u);$$

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v dx, \quad (u,v) = \int_\Omega uv dx, \quad f \in L_2(\Omega). \qquad (\text{УШ.23})$$

Здесь α, β – положительные постоянные,

$$j(v) = \int_\Omega \Phi(v(x)) dx,$$

а функция Φ определяется формулой

$$\Phi(\lambda) = \frac{1}{2} d_2 (\lambda - u_0)_+^2 + \frac{1}{2} d_1 (\lambda - u_0)_-^2 + L(\lambda - u_0),$$

где u_0, d_1, d_2, L – положительные постоянные, $\lambda_+ = \max(\lambda, 0)$, $\lambda_- = \min(\lambda, 0)$, $\lambda \in R_1$.

Существование и единственность решения задачи (УШ.22) следуют из общих теорем (см. Лионс [1]).

Задача (УШ.22) аппроксимируется по МКЭ с кусочно линейными координатными функциями. Область Ω покрывается семейством τ^h треугольников, h – наибольшая из сторон треугольников. Пусть V_h – пространство непрерывных кусочно линейных функций, равных нулю на $\partial\Omega$. Через $v_I \in V_h$ обозначается интерполянт функции v по узлам триангуляции. Обозначим еще

$$j_h(v) = \int_\Omega \{\Phi(v)\}_I dx = \frac{1}{6} \sum_{T \in \tau^h} A(T) \sum_{i=1}^{3} \Phi(v_i^T),$$

где $A(T)$ есть площадь треугольника T, а $v_i^{(T)}$ есть значение функции v в i-й вершине треугольника T. Определяется дискретное скалярное произведение

$$(w, v)_h = \sum_{T \in \tau^h} \frac{1}{3} A(T) \sum_{i=1}^{3} w_i^T v_i^T.$$

Приближенная задача ставится так: найти функцию $u_h \in V_h$, удовлетворяющую дискретному вариационному неравенству

$$\alpha(u_h, v_h - u_h)_h + \beta a(u_h, v_h - u_h) + j_h(v_h) - j_h(u_h) \geq (f, v_h - u_h); \qquad (\text{УШ.24})$$

$$\forall v \in V_h.$$

Решение задачи (УШ.24) существует и единственно. При некоторых предположениях о свойствах точного решения устанавливается следующая оценка погрешности аппроксимации:

$$\|u - u_h\|_{H^m(\Omega)} \leq Ch, \qquad \text{(УШ. 5)}$$

где C не зависит от h.

ГЛАВА IX

Упруго-пластическое состояние по Сен-Венану - Мизесу и Хаару - Карману

Как хорошо известно, впервые математическая теория пластического состояния была предложена Сен-Венаном в 1870 г. Эта теория давала описание пластического состояния в условиях плоской деформации; на трехмерный случай ее распространил М.Леви в том же 1870 г. В 1913 г. Мизес предложил описывать трехмерное пластическое состояние условием, близким к условию Сен-Венана и в то же время значительно более простым. Как выяснилось потом, условие Мизеса имеет простую механическую трактовку, и для плоской деформации уравнения Сен-Венана и Мизеса совпадают.

В 1909 г. Хаар и Карман опубликовали вариационный принцип, позволяющий описывать напряженное состояние в упруго-пластической среде, т.е. в среде, часть которой находится в упругом состоянии, а остальная часть - в пластическом. Принцип Хаара - Кармана опирается на условие пластичности Сен-Венана - Леви, но аналогичный принцип очень легко сформулировать, исходя из условия Мизеса.

Впоследствии выяснилось, что явление пластического состояния значительно сложнее и многообразнее, и что уравнения Сен-Венана - Леви и Мизеса описывают только так называемое пластическое течение. В этом плане следует прежде всего отметить работы Хенки, Прандтля, Прагера, Ходжа, Хилла, Ильюшина, Качанова и ряда других авторов.

В настоящей главе мы рассмотрим некоторые из задач, к которым приводит принцип Хаара - Кармана. Как мы увидим ниже, эти задачи подходят под схему (III.I-2). В § I мы для облегчения ссылок формулируем уравнения Сен-Венана - Леви и Мизеса и вариационный принцип Хаара - Кармана. В § 2 рассматривается задача упруго-пластического кручения цилиндрических стержней, в § 3 исследуется плоская деформация, а в § 4 - трехмерная задача упруго-пластического состояния. Мы остановимся в этих параграфах исключительно на вопросах устойчивости точных решений перечисленных задач с дополнительно малыми изменениями, а также и на оценках погрешности при аппроксимации и искажении.

§ I. Уравнения Сен-Венана - Леви и Мизеса. Вариационный принцип Хаара - Кармана.

I°. Тензор напряжений будем обозначать так:

$$T = \begin{pmatrix} \tau_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \tau_{yy} & \tau_{yz} \\ \tau_{xz} & \tau_{yz} & \tau_{z} \end{pmatrix}. \qquad (IX.1.1)$$

Если $\vec{K} = (X, Y, Z)$ есть вектор объемных сил, то в состоянии равновесия тензор напряжений удовлетворяет уравнению равновесия

$$div\, T + \vec{K} = 0 \qquad (IX.1.2)$$

или, в более подробной записи

$$\begin{aligned} \tau_{xx/x} + \tau_{xy/y} + \tau_{xz/z} + X &= 0, \\ \tau_{xy/x} + \tau_{yy/y} + \tau_{yz/z} + Y &= 0, \\ \tau_{xz/x} + \tau_{yz/y} + \tau_{zz/z} + Z &= 0, \end{aligned} \qquad (IX.1.3)$$

индекс за косой черточкой означает дифференцирование по соответствующему направлению.

Вектор смещений будем обозначать через $U = (U_x, U_y, U_z)$, вектор скоростей смещений – через V, так что $V = \dot{U} = (U_{x/t}, U_{y/t}, U_{z/t})$. Будем считать деформации и их скорости малыми, тогда тензор скоростей деформации можно записать так:

$$\frac{1}{2}\begin{pmatrix} 2V_{x/x} & V_{x/y} + V_{y/x} & V_{x/z} + V_{z/x} \\ V_{x/y} + V_{y/x} & 2V_{y/y} & V_{y/z} + V_{z/y} \\ V_{x/z} + V_{z/x} & V_{y/z} + V_{z/y} & 2V_{z/z} \end{pmatrix}. \qquad (IX.1.4)$$

Главные нормальные напряжения обозначим через $\tau_1 \geqslant \tau_2 \geqslant \tau_3$; они суть корни уравнения

$$\begin{vmatrix} \tau_{xx} - \tau & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \tau_{yy} - \tau & \tau_{yz} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} - \tau \end{vmatrix} = 0. \qquad (IX.1.5)$$

Наибольшее касательное напряжение равно

$$\tau_m = \frac{\tau_1 - \tau_3}{2}. \qquad (IX.1.6)$$

2°. Сен-Венан [1] сформулировал условия пластичности и написал уравнения пластического состояния для случая плоской деформации; в этом случае напряжения и смещения не зависят от одной из координат, скажем от z, и $u_z = \tau_{xz} = \tau_{yz} = 0$. Одно из главных нормальных напряжений равно τ_{zz}, а остальные два равны

$$\frac{1}{2}\left[\tau_{xx} + \tau_{yy} \pm \sqrt{(\tau_{xx} - \tau_{yy})^2 + 4\tau_{xy}^2}\right]. \qquad (IX.1.7)$$

Опираясь на опыты Треска, Сен-Венан формулирует следующие гипотезы пластического состояния:

I. Наибольшее касательное напряжение в пластическом состоянии есть величина постоянная, характерная для данной среды. Сен-Венан обозначает эту постоянную буквой $\mathcal{K}$. Мы будем называть эту гипотезу "условием пластичности Сен-Венана".

II. Направление наибольшей скорости скольжения совпадает с направлением наибольшего касательного напряжения.

III. Вещество в пластическом состоянии несжимаемо.

Кроме этих явно сформулированных гипотез, Сен-Венан пользуется еще двумя неформулированными гипотезами:

IV. τ_{zz} лежит между главными напряжениями (IX.1.7).

V. Производные по координатам от смещений и скоростей смещений малы.

Перечисленные гипотезы приводят к следующим уравнениям пластического состояния в плоском случае:

$$\tau_{xx/x} + \tau_{xy/y} + X = 0, \quad \tau_{xy/x} + \tau_{yy/y} + Y = 0; \qquad (IX.1.8)$$

$$(\tau_{xx} - \tau_{yy})^2 + 4\tau_{xy}^2 = 4\mathcal{K}^2; \qquad (IX.1.9)$$

$$(v_{y/y} - v_{x/x}) : (v_{y/x} + v_{x/y}) = (\tau_{yy} - \tau_{xx}) : 2\tau_{xy}; \qquad (IX.1.10)$$

$$v_{x/x} + v_{y/y} = 0. \qquad (IX.1.11)$$

Уравнения (IX.I.8) суть общие уравнения равновесия механики сплошной среды, уравнение (IX.I.9) есть условие пластичности Сен-Венана, уравнения (IX.I.IO) и (IX.I.II) соответствуют гипотезам IУ и П.

3°. Уравнения теории пластичности в трехмерном случае получил Леви [1] на основании условия пластичности Сен-Венана. Леви сохраняет гипотезы I, Ш и У Сен-Венана. Гипотеза IУ становится ненужной, гипотеза П заменяет на следующей:

Па. Девиатор напряжений и тензор скоростей деформаций пропорциональны.

Гипотезы Леви приводят к следующим уравнениям пластического состояния:

$$(\upsilon_{x/x}-\upsilon_{y/y}):(\tau_{xx}-\tau_{yy})=(\upsilon_{y/y}-\upsilon_{z/z}):(\tau_{yy}-\tau_{zz})= \tag{IX.I.I2}$$

$$=(\upsilon_{x/y}+\upsilon_{y/x}):2\tau_{xy}\quad=(\upsilon_{y/z}+\upsilon_{z/y}):2\tau_{yz}\quad=$$

$$=(\upsilon_{z/x}+\upsilon_{xz}):2\tau_{xz};$$

$$\upsilon_{x/x}+\upsilon_{y/y}+\upsilon_{z/z}=0. \tag{IX.I.I3}$$

Условие пластичности дает уравнение 6

$$\mathcal{K}^6-(a^2+3b)\mathcal{K}^4/2+(a^2+2b)^2\mathcal{K}^2/16-(4a^3c+a^2b^2+$$

$$+18abc+4b^3-27c^2)/64=0. \tag{IX.I.I4}$$

Здесь a, b, c - коэффициенты уравнения

$$\tau^3+a\tau^2-b\tau-c:=\begin{vmatrix}\tau_{xx}-\tau & \tau_{xy} & \tau_{xz}\\ \tau_{xy} & \tau_{yy}-\tau & \tau_{yz}\\ \tau_{xz} & \tau_{yz} & \tau_{zz}-\tau\end{vmatrix}=0,$$

корни которого суть главные нормальные напряжения.

К уравнениям (IX.I.I2-I4) следует добавить общие уравнения равновесия (IX.I.3).

Заметим, что в плоском случае из уравнений Леви вытекает использованная Сен-Венаном гипотеза IУ. В самом деле, в плоском

случае $v_x = 0$, $\tau_{xz} = \tau_{yz} = 0$, уравнение (IX.I.18) дает $v_{x/x} = -v_{y/y}$ и из первого уравнения (IX.I.12) следует тогда $\tau_x = (\tau_{xx} + \tau_{yy})/2$. Гипотеза IУ вытекает из (IX.I.7) и их последнего неравенства.

Замечание. Уравнение (IX.I.I4) дано в книге автора "Основные уравнения математической теории пластичности", изд. АН СССР 1934. В статье Леви [I] дано другое условие, по всей видимости ошибочное, так как из него не вытекает условие Сен-Венана для плоской задачи. В сборнике "Теория пластичности" [I] дан перевод другой статьи Леви, в которой дана еще одна запись условия пластичности. Автору кажется, что эта запись также ошибочна, потому что размерности слагаемых в этом условии неодинаковы.

4°. Коротко изложим теорию Мизеса [I]. Существенным ее отличием является иная формулировка условия пластичности. Пусть τ_1 , τ_2 , τ_3 – главные нормальные напряжения. Положим

$$\tau_1 = \frac{\tau_2 - \tau_3}{2} , \quad \tau_2 = \frac{\tau_3 - \tau_1}{2} , \quad \tau_3 = \frac{\tau_1 - \tau_2}{2} ; \qquad \text{(IX.I.}$$

так что

$$\tau_1 + \tau_2 + \tau_3 = 0 . \qquad \text{(IX.I.}$$

Мизес принимает следующую гипотезу, более общую, чем гипотеза Сен-Венана: в пластическом состоянии напряжения остаются на границе упругости, в пространстве координат τ_1 , τ_2 , τ_3 предел упругости представляется в виде замкнутой кривой, лежащей в плоскости (IX.I.I6) и обходящей начало координат.

По Сен-Венану $|\tau_i| \leqslant \mathcal{K}$, $i = 1, 2, 3$. Эти неравенства определяют в пространстве (τ_1, τ_2, τ_3) куб, который пересекается с плоскостью (IX.I.I6) по правильному шестиугольнику. Мизес предложил заменить куб Сен-Венана сферой

$$\tau_1^2 + \tau_2^2 + \tau_3^2 = 2\mathcal{K}'^2 , \quad \mathcal{K}' = \text{const} . \qquad \text{(IX.I.}$$

Уравнение (IX.I.I7) легко преобразуется к виду

$$\tau_i^2 := \frac{1}{6}\left[(\tau_{xx} - \tau_{yy})^2 + (\tau_{yy} - \tau_{zz})^2 + (\tau_{zz} - \tau_{xx})^2 + \right.$$

$$\left. + 6(\tau_{xy}^2 + \tau_{yz}^2 + \tau_{xz}^2)\right] = \frac{4}{3}\mathcal{K}'^2 . \qquad \text{(IX.I.}$$

Величина τ_i называется интенсивностью касательных напряжений.

В условиях плоской задачи $\tau_i^2 = \frac{1}{4}(\tau_x - \tau_{yy})^2 + \tau_{xy}^2$;

условие пластичности Сен-Венана будет получаться как частный случай условия Мизеса, если положить

$$k' = \frac{\sqrt{3}}{2} k \; ; \qquad (IX.I.19)$$

при этом условие Мизеса немного упрощается:

$$\tau_i = k \; . \qquad (IX.I.20)$$

Остальные уравнения теории пластического состояния по Мизесу совпадают с соответствующими уравнениями Леви.

5°. Вариационный принцип Хаара и Кармана был опубликован в статье [I] этих знаменитых авторов в 1909 г., до появления статьи [I] Мизеса; естественно, что этот принцип был основан на условии пластичности Сен-Венана. Однако, принцип Хаара - Кармана легко формулируется в терминах условия пластичности Мизеса, и в этом случае названный принцип утверждает следующее.

Пусть среда, заполняющая область Ω трехмерного (или двумерного) пространства, находится в упруго-пластическом состоянии. Для простоты допустим, что объёмные силы отсутствуют, так что вектор $\overset{0}{k} = 0$ и тензор напряжений T удовлетворяет однородному уравнению

$$div\, T = 0. \qquad (IX.I.21)$$

Пусть $W(T)$ - плотность энергии упругой деформации, выраженная в терминах тензора напряжений T . Тогда в упруго-пластическом состоянии этот тензор реализует минимум интеграла

$$J(T) = \int_\Omega W(T)\, d\omega \qquad (IX.I.22)$$

на множестве симметричных тензоров T , удовлетворяющих следующим условиям: а) интеграл (IX.I.22) конечен; б) T удовлетворяет уравнениям (IX.I.21) и известным краевым условиям Коши, означающим, что известны внешние силы, приложенные к границе области Ω ; в) T удовлетворяет условию ограничения

$$\tau_i \leqslant k \; , \qquad (IX.I.23)$$

где τ_i - интенсивность касательных напряжений тензора T.

В статье Хаара и Кармана [I] доказывается некоторая общая

теорема, из которой в частности следует: если T решает сформулированную выше одностороннюю вариационную задачу, то в каждой точке области Ω удовлетворяется либо уравнение Эйлера для функционала (IX.1.22), либо уравнение $\tau_i = \mathcal{H}$ (т.е., условие пластичности).

Следует, однако, иметь в виду, что в пластической зоне должны выполняться также уравнения, содержащие скорости деформаций - так, в трехмерной задаче должны быть удовлетворены уравнения (IX.1.12-13). Поэтому из существования решения вариационной задачи Хаара - Кармана еще не вытекает автоматически существование решения задачи об упруго-пластическом состоянии.

Принцип Хаара - Кармана тесно связан с известным принципом Кастильяно о теории упругости. Напомним этот последний принцип: если тензор напряжений реализует минимум функционала (IX.1.22) при условиях а) и б), то этот тензор решает задачу теории упругости при краевых условиях, упомянутых в б). Таким образом, формально можно получить принцип Хаара - Кармана из принципа Кастильяно, если потребовать, чтобы функционал Кастильяно (IX.1.22) достигал минимума на множестве тензоров напряжений, удовлетворяющих не только условиям а), б), но и условию в), т.е., неравенству (IX.1.23).

§ 2. Кручение стержня.

1°. Пусть нормальное сечение G скручиваемого стержня есть конечная и, в общем случае, многосвязная область плоскости (x, y). Если стержень весь находится в упругом состоянии, то отличные от тождественного нуля напряжения $\tau_x := \tau_{xz}$ и $\tau_y := \tau_{yz}$ удовлетворяют уравнениям

$$\tau_{x/x} + \tau_{y/y} = 0, \quad (x,y) \in G; \tag{IX.2.1}$$

$$\tau_x \cos(\nu, x) + \tau_y \cos(\nu, y) = 0, \ (x,y) \in \partial G; \tag{IX.2.2}$$

$$\iint_G (x\tau_y - y\tau_x)\, dx dy = \mathcal{M}; \tag{IX.2.3}$$

$$\Delta\tau_x = \Delta\tau_y = 0. \tag{IX.2.4}$$

Здесь ν - нормаль к ∂G, $\mathcal{M}$ - момент внешних сил, приложенных к основанию стержня. Если стержень находится упруго-пластическом состоянии, то уравнения (IX.2.I-3) сохраняют силу; уравнения (IX.2.4) остаются оправедливыми в упругой зоне. Приняв, что напряжения в пластической зоне удовлетворяют условию Сен-Венана (уравнение (IX.I.I4)) или уравнению Мизеса (IX.1.20), мы должны заменить в пластической зоне уравнения (IX.2.4) одним уравнением

$$\tau_x^2 + \tau_y^2 = \mathcal{K}^2. \qquad \text{(IX.2.5)}$$

Введем предгильбертово пространство элементы которого суть векторы $\tau = (\tau_x, \tau_y)$, удовлетворяющие уравнению (IX.2.I) и краевому условию (IX.2.2); норму в этом пространстве зададим формулой

$$|\tau|^2 = \iint_G (\tau_x^2 + \tau_y^2)\, dx dy. \qquad \text{(IX.2.6)}$$

Пополним наше пространство в метрике (IX.2.6), получим гильбертово пространство, которое обозначим через $\widetilde{H}$. Об элементах этого пространства будем говорить, что они удовлетворяют (в обобщенном смысле) уравнениям (IX.2.I-2).

Если стержень находится в упругом состоянии, то в силу принципа Кастильяно задача кручения равносильна задаче о минимуме функционала (IX.2.6) на множестве векторов, удовлетворяющих уравнениям (IX.2 I-3) - иначе говоря, на множестве элементов $\widetilde{H}$, удовлетворяющих уравнению (IX.2.3). Пусть $g = -(g_x, g_y)$ - какой-нибудь вектор из $\widetilde{H}$, удовлетворяющий этому уравнению. Положим $g + \tau = \sigma = (\sigma_x, \sigma_y)$. Тогда σ удовлетворяет однородному уравнению

$$\iint_G (x\sigma_y - y\sigma_x)\, dx dy = 0. \qquad \text{(IX.2.7)}$$

Обозначим через H подпространство тех элементов из $\widetilde{H}$, которые удовлетворяют последнему уравнению. В силу принципа Хавара - Кармана решение задачи упруго-пластического кручения заменяется решением следующей односторонней вариационной задачи:

$$|\sigma - g| = \min, \quad \sigma \in M;$$

$$M = \{\sigma \in H,\ \|\sigma - g\| \leqslant \mathcal{K}\}, \qquad \text{(IX.2.8)}$$

$$\|\tau\| = \sup_{(x,y)\in G} \operatorname{ess} \sqrt{\tau_x^2 + \tau_y^2}\,.$$

2°. Исследуем множество M, определенное в формулах (IX.2.8). При больших по модулю значениях $\mathcal{M}$ оно будет пустым. Действительно,

$$|\mathcal{M}| \leqslant \iint_G \rho\sqrt{\tau_x^2+\tau_y^2}\,dxdy\,; \quad \rho^2 = x^2 + y^2\,;$$

в упруго-пластическом состоянии $\tau_x^2 + \tau_y^2 \leqslant \mathcal{K}^2$, поэтому

$$|\mathcal{M}| \leqslant \mathcal{K}\iint_G \rho\,dxdy. \qquad (IX.2.9)$$

Правая часть последнего неравенства есть некоторая постоянная: если в уравнении (IX.2.3) $|\mathcal{M}|$ больше этой постоянной, то M пусто.

Назовем момент $\mathcal{M}$ допустимым, если ему соответствует непустое множество M. Изменив в случае необходимости ориентацию осей координат, можно считать, что $\mathcal{M} \geqslant 0$. Обозначим точную верхнюю границу допустимых моментов через $\overline{\mathcal{M}}$ и докажем, что множество допустимых моментов есть либо отрезок $0 \leqslant \mathcal{M} \leqslant \overline{\mathcal{M}}$, либо полуинтервал $0 \leqslant \mathcal{M} < \overline{\mathcal{M}}$. Достаточно доказать, что последний полуинтервал содержится в множестве допустимых моментов.

В любой левой полуокрестности точки $\overline{\mathcal{M}}$ содержится хотя бы один допустимый элемент $\mathcal{M}'$. Ему соответствует по крайней мере один вектор $\tau' = (\tau'_x, \tau'_y)$, такой, что

$$\iint_G (x\tau'_y - y\tau'_x)\,dxdy = \mathcal{M}'\,; \quad \|\tau'\| \leqslant \mathcal{K}.$$

Пусть $0 \leqslant \mathcal{M}'' < \mathcal{M}'$. Положим $\tau'' = \mathcal{M}''\tau'/\mathcal{M}'$. Тогда $\tau'' \in \tilde{H}$ и

$$\iint_G (x\tau''_y - y\tau''_x)\,dxdy = \mathcal{M}'', \quad \|\tau''\| = \frac{\mathcal{M}''}{\mathcal{M}'}\|\tau'\| < \mathcal{K}.$$

Последнее соотношение означает, что множество M'', которое соответствует моменту $\mathcal{M}''$, содержит элемент τ'' и потому не пусто. Этим и доказано, что полуинтервал $0 \leqslant \mathcal{M} < \overline{\mathcal{M}}$ содержится в множестве допустимых моментов.

Докажем теперь, что множество M, определенное формулами (IX.2.8), замкнуто в норме $\|\cdot\|$. Пусть $\tau_n = (\tau_{nx}, \tau_{ny})$ – последовательность векторов из M, и пусть $\|\tau_n - \tau_0\| \xrightarrow[n\to\infty]{} 0$.

Докажем, что $\tau_0 \in M$. Из сходящейся в H последовательности $\{\tau_n\}$ можно выделить подпоследовательность $\{\tau_{n_k}\}$, сходящуюся к τ_0 почти всюду в G. Так как $\|\tau_{n_k}\| = \sup \sqrt{\tau_{n_k x}^2 + \tau_{n_k y}^2} \le \mathcal{K}$, то и $\sup \sqrt{\tau_{ox}^2 + \tau_{oy}^2} \le \mathcal{K}$ почти всюду в Ω. Это и означает, что $\tau_0 \in M$ и, следовательно, множество M замкнуто.

Из известных общих теорем (см. § I гл.УШ) следует, что задача (IX.2.8) имеет одно и только одно решение.

3°. Задача (IX.2.8) есть частный случай задачи (УШ.I.I-2), в которой, в частности, следует положить $d = \mathcal{K}$. Если $0 \le m < \bar{m}$, то, как мы видели в п.2°, в множестве M существует элемент, полунорма которого меньше $\mathcal{K}$. Примем этот элемент за g. На основании результатов § 5 гл.УШ можно утверждать следующее.

Решение задачи (IX.2.8) устойчиво относительно малых возмущений вектора g и постоянной пластичности $\mathcal{K}$, если крутящий момент лежит в полуинтервале $0 \le m < \bar{m}$.

Пусть момент m заменен моментом m', мало отличающимся от m и по-прежнему лежащим в полуинтервале $[0, \bar{m})$. Если моменту m соответствует вектор g, $\|g\| < \mathcal{K}$, то моменту m' можно привести в соответствие вектор $g' = m'g/m$; если m и m' достаточно близки, то $\|g'\| < \mathcal{K}$ и величина $\|g - g'\|$ достаточно мала. Из сказанного вытекает следующее утверждение:

Решение задачи (IX.2.8) устойчиво относительно малых возмущений крутящего момента, не выводящих из полуинтервала $[0, \bar{m})$.

Ниже мы покажем, что при том же условии $0 \le m < \bar{m}$ решение задачи (IX.2.8) устойчиво относительно малых изменений области G.

4°. Для последующих ссылок приведем некоторые известные факты теории упругого кручения стержней.

Пусть область G - $(n+1)$-связная. Тогда ∂G состоит из $n+1$ связных компонент, которые мы обозначим через ∂G_j, $j = 0, 1, \ldots, n$; через ∂G_0 обозначим ту компоненту, которая ограничивает область G извне. Обозначим еще через G_j, $j = 1, 2, \ldots, n$, ту область, которая извне ограничена контуром ∂G_j. Пусть $G_0 = G \cup G_1 \cup \ldots \cup G_n$; G_0 есть конечная односвязная область, ограниченная контуром ∂G_0. Можно определить G_0 как наименьшую односвязную область, содержащую область G.

Уравнение (IX.2.1) позволяет ввести так называемую функцию напряжений Прандтля

$$\tau_x = \frac{\partial u}{\partial y}, \quad \tau_y = -\frac{\partial u}{\partial x}. \qquad \text{(IX.2.1}$$

Из краевого условия (IX.2.2) следует, что на границе области функция Прандтля удовлетворяет условию

$$u|_{\partial G_j} = A_j = const. \qquad \text{(IX.2.1}$$

При изменении функции Прандтля на постоянное слагаемое напряжения не меняются, поэтому можно положить $A_0 = 0$.

Как видно из уравнений (IX.2.4), в упругой области функция Прандтля удовлетворяет уравнению Пуассона

$$-\Delta u = C_0, \quad C_0 = const. \qquad \text{(IX.2.1}$$

Известна формула, связывающая крутящий момент M с функцией Прандтля $u(x,y)$, именно

$$M = 2\iint_G u\,dx\,dy + 2\sum_{j=1}^{n} A_j |G_j|. \qquad \text{(IX.2.1}$$

Продолжим функцию u на всю область G_0, положив $u(x,y) = A_j$, $(x,y) \in G_j$. Тогда формула (IX.2.13) принимает более простой вид

$$M = 2\iint_{G_0} u\,dx\,dy. \qquad \text{(IX.2.1}$$

5°. Задачу кручения можно свести к вариационной задаче, которая по форме отлична от задачи (IX.2.8). Если скручиваемый стержень находится в упругом состоянии, то задачу (IX.2.8) можно записать в терминах функции Прандтля следующим образом:

$$\iint_G |\nabla u|^2\,dx\,dy = \min,$$

$$u = \begin{cases} 0 \text{ на } \partial G_0 \\ A_j = const \text{ на } \partial G_j, \ 1 \le j \le n, \end{cases} \qquad \text{(IX.2.1}$$

$$2\iint_G u\,dx\,dy + 2\sum_{j=1}^{n} A_j |G_j| = M.$$

Это задача на условный минимум; по теореме Эйлера существует такая постоянная C_0, что

$$\delta\left\{\iint_G [|\nabla u|^2 - 2C_0 u]\,dx\,dy - 2C_0 \sum_{j=1}^{N} A_j |G_j|\right\} = 0, \qquad (IX.2.16)$$

δ – знак вариации; функция u должна удовлетворять краевому условию из (IX.2.15). Если эту функцию продолжить на G_0, как об этом сказано в п.4°, то уравнение (IX.2.16) примет вид

$$\delta \iint_{G_0} [|\nabla u|^2 - 2C_0 u]\,dx\,dy = 0. \qquad (IX.2.17)$$

Обозначим через $\overline{H}_0$ подпространство пространства $\overset{\circ}{W}{}_2^{(1)}(G_0)$, образованное функциями, которые остаются постоянными в каждой из внутренних областей G_j. Норму в $\overset{\circ}{W}{}_2^{(1)}(G_0)$ зададим формулой

$$|u|^2 = \iint_{G_0} |\nabla u|^2 dx\,dy;$$

соответственно скалярное произведение равно интегралу

$$[u,v] = \iint_{G_0} \nabla u \cdot \nabla v\,dx\,dy.$$

Вариационная задача (IX.2.15) приводится к виду

$$|u - u_0| = \min, \quad u \in \overline{H}_0,$$

где u_0, в соответствии с теоремой Ф.Рисса, определяется тождеством

$$\forall u \in \overline{H}_0, \quad [u, u_0] = C_0 \iint_{G_0} u\,dx\,dy.$$

Применяя принцип Хаара – Кармана, мы сведем задачу упруго-пластического кручения к следующей односторонней вариационной задаче:

$$|u - u_0| = \min, \quad u \in \overline{M}; \qquad (IX.2.18)$$

$$\overline{M} = \{u \in \overline{H}_0 : \|u\| \le K\};$$

$$\|u\| = \sup_G |\nabla u|$$

Задача (IX.2.18) сформулирована в статье Эллен Лаушон [1]. При любом конечном значении C_0 эта задача имеет одно и только одно решение, которое мы обозначим через $u_* = u_*(x, y, C_0)$. В статье Лаушон [1] доказывается, что решение задачи (IX.2.18) удовлетворяет всем дифференциальным уравнениям и краевым условиям задачи упругопластического кручения.

Если задан крутящий момент M, то C_0 следует определять из уравнения

$$2\iint_{G_0} u_*(x, y, C_0)\,dx\,dy = M. \qquad \text{(IX.2.}$$

Каждому конечному значению C_0 соответствует по формуле (IX.2.19) одно и только одно допустимое значение M. Обратно, как видно из результатов пп. 2^0 – 4^0, каждому допустимому значению M из полуинтервала $0 \leqslant M < \overline{M}$ соответствует одна и только одна функция Прандтля, которая в упругой области удовлетворяет уравнению вида (IX.2.12), иначе говоря, каждому допустимому значению M из полуинтервала $[0, \overline{M})$ соответствует одно и только одно конечное значение C_0, и при любом значении $M \in [0, \overline{M})$ уравнение (IX.2.19) имеет одно и только одно решение.

6^0. Займемся вопросом об устойчивости решения задачи (IX.2.18) относительно малых возмущений области G. Наряду с задачей (IX.2.18) рассмотрим такую же задачу для стержня из того же материала и с нормальным сечением G', близким к G. Близость областей G и G' определим следующим образом. Обозначим через ξ, η и x, y декартовы координаты в областях G' и G соответственно и допустим, что существует преобразование координат S, которое взаимно однозначно переводит G' в G_* и G_0' в G_0 (здесь G_* – наименьшая односвязная область, содержащая область G'), причем якобиева матрица преобразования S имеет вид $I + O(\varepsilon)$, где I – единичная матрица, а ε – малое положительное число. Тогда, если $\mathcal{J}$ – якобиан преобразования S, то $\mathcal{J} = 1 + O(\varepsilon)$.

Задачу кручения для области G' можно записать так:

$$\iint_{G_0'} \left[\left(\frac{\partial u'}{\partial \xi}\right)^2 + \left(\frac{\partial u'}{\partial \eta}\right)^2 - 2C_0' u'\right] d\xi\,d\eta = \min; \quad u' \in \overline{H}_0',$$

$$\sup_{G'} \left[\left(\frac{\partial u'}{\partial \xi}\right)^2 + \left(\frac{\partial u'}{\partial \eta}\right)^2\right] \leqslant K^2; \qquad \text{IX.2.2}$$

здесь $\overline{H}_0'$ – подпространство пространства $W_2^{(1)}(G_0')$, образованное

функциями, постоянными в каждой из связных компонент множества $G_0 \setminus G'$.

В (IX.2.20) выполним преобразование S; заметим, что оно переводит $\bar{H}'_0$ в $\bar{H}_0$. Обозначая $\nabla' = \left(\frac{\partial}{\partial\xi}, \frac{\partial}{\partial\eta}\right)$, получаем

$$\iint_{G_0} \{(u_x)^2 |\nabla x|^2 + 2u_x u_y (\nabla x, \nabla y) + (u_y)^2 |\nabla y|^2 - 2c_0^2 u\} J\, dx\, dy = \min;$$

$$u(x,y) = u'(\xi,\eta); \quad u \in \bar{H}_0;$$

$$\sup_G \operatorname{ess} \{(u_x)^2 |\nabla x|^2 + 2u_x u_y (\nabla x, \nabla y) + (u_y)^2 |\nabla y|^2\} \leqslant \mathcal{K}^2. \qquad \text{(IX.2.21)}$$

В пространстве $\bar{H}_0$ введем новую норму $|\cdot|_1$ (и соответственно новое скалярное произведение $[\,,\,]_1$), положив

$$|u|_1^2 = \iint_{G_0} \{(u_x)^2 |\nabla x|^2 + 2u_x u_y (\nabla x, \nabla y) + (u_y)^2 |\nabla y|^2\} J\, dx\, dy. \qquad \text{(IX.2.22)}$$

Функционал $l_1 u = \iint_{G_0} c_0^2 J u\, dx\, dy$ ограничен в норме (IX.2.22), поэтому $l_1 u = [u, u_1]_1$, где u_1 – некоторый элемент из $\bar{H}_0$. Наконец, обозначим через $\bar{M}'$ множество функций из $\bar{H}_0$, удовлетворяющих последнему из соотношений (IX.2.21) и через $\|u\|_1^2$ – левую часть этого соотношения. Задача (IX.2.21) принимает вид

$$|u - u_1|_1 = \min, \quad u \in \bar{M}';$$

$$\bar{M}' = \{u \in \bar{H}_0 : \|u\|_1 \leqslant \mathcal{K}\}. \qquad \text{(IX.2.23)}$$

7°. При переходе от задачи (IX.2.18) к задаче (IX.2.23) меняются все параметры, определяющие задачу: элемент f_0, норма $|\cdot|$ и множество $\bar{M}$. Покажем, что все эти изменения малы в смысле определения § 5 гл. VIII.

Прежде всего докажем, что нормы $|\cdot|$ и $|\cdot|_1$ связаны неравенством вида (VIII.5.3). Выражение под знаком интеграла в (IX.2.22) есть квадратичная форма относительно u_x и u_y; нетрудно видеть, что ее собственные числа имеют вид $\lambda_k = 1 + O(\varepsilon)$, $k = 1, 2$. В п. 6° было отмечено, что $J = 1 + O(\varepsilon)$. Из сказанного следует существование такой постоянной δ_1, зависящей только от ε, что $\lim_{\varepsilon \to 0} \delta_1 = 0$ и $(1 - \delta_1)^2 \leqslant J_0 \lambda_k \leqslant (1 + \delta_1)^2$; $k = 1, 2$. Тогда

$$(1-\delta_1)|u| \leqslant |u|_1 \leqslant (1+\delta)|u|.$$

Теперь оценим величину $|u_0 - u_1|$. Рассмотрим разность

$$[u, u_1] - [u_1, u_1]_1 =$$

$$= \iint_{G_0} \left\{ u_{/x} u_{1/x}(1-\gamma|\nabla' x|^2) + u_{/y} u_{1/y}(1-\gamma|\nabla' y|^2) - (u_{/x} u_{1/y} + u_{/y} u_{1/x})\gamma(\nabla' x, \nabla' y) \right\} dx dy$$

Из последнего тождества находим $[u,u_1]-[u,u_1]_1 = |u| \cdot |u_1| \cdot O(\varepsilon)$ и, следовательно,

$$[u, u_0 - u_1] = [u, u_0] - [u, u_1]_1 + |u| \cdot |u_1| \cdot O(\varepsilon).$$

Далее, принимая, что $c_0 - c_0' = O(\varepsilon)$, находим

$$[u, u_0]\ [u, u_1]_1 = c_0 \iint_{G_0} u\, dx dy - c_0' \iint_{G_0} u\gamma\, dx dy =$$

$$= (c_0 - c_0') \iint_{G_0} u\, dx dy + c_0' \iint_{G_0} u(1-\gamma)\, dx dy = |u| \cdot O(\varepsilon)$$

и $[u, u_0 - u_1] = |u|(1+|u_1|)O(\varepsilon)$. Полагая здесь $u = u_0 - u_1$, находим

$$|u_0 - u_1| \leqslant C\varepsilon(1+|u_1|), \quad C = const. \qquad (IX.2.24)$$

Отсюда $|u_1| - |u_0| \leqslant C\varepsilon\,(1+|u_1|)$ и $|u_1| \leqslant |u_0| + O(\varepsilon)$. Теперь то же неравенство (IX.2.24) дает $|u_0 - u_1| = O(\varepsilon)$; существует, следовательно, такая величина δ_2, зависящая только от ε, что $\lim_{\varepsilon \to 0} \delta_2 = 0$ и $|u_0 - u_1| \leqslant \delta_2$.

8°. Исследуя влияние замены множества $\bar{M}$ на $\bar{M}'$, примем пока, что $G' \subset G$. Введенные выше пространства $\bar{H}_0$ и $\bar{H}_0'$ будем теперь обозначать через $H_0(G)$ и $H_0(G')$. Функции класса $\bar{H}_0(G')$ доопределим в $G_0 \setminus G_0'$, положив их там равными нулю; после этого $H_0(G')$ станет подпространством пространства $H_0(G)$.

Выше мы определили S как преобразование координат; расширим это определение, перенеся его также на функции, заданные в соответствующих областях, так что если функция $u(\xi, \eta)$ определена в G_0', то $(Su)(x,y) = u(\xi(x,y), \eta(x,y))$, где $(x,y) = S(\xi, \eta)$. При таком определении S переводит $\bar{M}'$ в $\bar{M}$; докажем, что при этом выполняется неравенство (VII.5.9). Пусть $u \in \bar{M}$. При этом

$$S^{-1}u=\begin{cases} u(\xi(x,y),\eta(x,y)) , & \xi(x,y)\subset G_0' , \\ 0 , & \xi(x,y)\in G_0\setminus G_0' \end{cases}$$

и

$$|u-S^{-1}u|=\iint\limits_{G_0\setminus G_0'}|\nabla u|^2dxdy+\iint\limits_{G_0'}[\nabla(u(\xi,\eta)-(S^{-1}u)(\xi,\eta)]^2d\xi d\eta .$$

Первый интеграл справа, в силу условия пластичности Сен-Венана, имеет порядок $O(\varepsilon)$. Второй интеграл равен

$$\iint\limits_{G_0'}[(u_\xi)^2[(1-\xi_x)^2+\xi_y^2]+2u_\xi u_\eta[(1-\xi_x)\eta_y+\xi_y(1-\eta_x)]+$$
$$+(u_\eta)^2[(\eta_x)^2+(1-\eta_y)^2]]dxdy .$$

Якобиева матрица преобразования S имеет вид $I+O(\varepsilon)$; отсюда следует, что последний интеграл оценивается величиной $|u|^2O(\varepsilon^2)$. Окончательно, $|u-S^{-1}u|^2=O(\varepsilon)+|u|^2O(\varepsilon^2)$, $|u-S^{-1}u|\le O(\sqrt{\varepsilon})+|u|O(\varepsilon)=(1+|u|)O(\sqrt{\varepsilon})$. Это значит, что существует величина δ_3, зависящая только от ε, такая, что $\lim\limits_{\varepsilon\to 0}\delta_3=0$ и $|u-S^{-1}u|\le\delta_3(1+|u|)$.

Из доказанного в настоящем пункте следует, что задача упругопластического кручения устойчива относительно малых возмущений области сечения стержня, если возмущенная область лежит внутри первоначальной.

9°. Пусть область G' получена малым возмущением области G, но G' не вкладывается в G. Построим область $\tilde{G}$, которая может быть получена малым возмущением каждой из областей G и G', и притом такую, что как G, так и G' вкладывается в $\tilde{G}$. По доказанному в п.8°, функции Прандля u и u', соответствующие сечениям G и G', мало отличаются от $\tilde{u}$ - функции Прандтля для сечения $\tilde{G}$. Но тогда u и u' мало различаются между собой.

Таким образом, решение задачи упруго-пластического кручения устойчиво относительно малых возмущений сечения стержня, если малость возмущения понимается в смысле п.6° настоящего параграфа.

10°. На решение задачи упруго-пластического кручения естественным образом распространяются результаты §§ 2 и 5 гл.V. Таким образом если приближенное значение $\tilde{t}^{(0)}$ функции Прандтля

$u_*(x,y)$ определено по методу Ритца (в частности, по методу конечных элементов) как элемент соответствующего подпространства H_n, а $u_*(x,y)$ есть точное значение функции Прандтля, то

$$|u_* - u_*^{(n)}| = O\left(\sqrt{e_n(u_*)}\right); \qquad (IX.2.25)$$

$$e_n(u) = \inf_{u^{(n)} \in H_n} \left[|u - u^{(n)}| + |||u - u^{(n)}|||\right],$$

причем норма $|\cdot|$ и полунорма $|||\cdot|||$ соответственно определяются формулами

$$|u|^2 = \iint_G |\nabla u|^2 dx dy, \quad |||u||| = \sup_G |\nabla u|.$$

Погрешность искажения приближенного решения $u_*^{(n)}(x,y)$ имеет вид $O(\sqrt{\|\Gamma_n\|} + \|\delta^{(n)}\|)$, где Γ_n - искажение матрицы Ритца, $\delta^{(n)}$ есть искажение столбца свободных членов системы Ритца.

II°. В литературе много внимания уделяется приближенному решению задачи (IX.2.18). В случае односвязной области этот вопрос обстоятельно изложен в книге Гловинского, Лионса и Тремольера [I] ; некоторые результаты содержатся в более ранней книге Черноусько и Баничука [I] . Случаю многосвязного сечения посвящена работа Гловинского и Ланшон [1] . Оценки погрешности приближения в этих работах отсутствуют. Упруго-пластическое кручение стержней с односвязным сечением изучалось как в точной, так и в приближенной постановке в ряде работ и независимо от вариационного принципа. Первой в этом ряду была знаменитая работа Надаи (см. Надаи [1]). Интересные результаты были получены в этом направлении Галиным [I - 2] . Отметим еще работу Перлина [I] .

§ 3. Плоская задача

I°. Введем следующие обозначения: σ - постоянная Пуассона упругой среды,

$$T = \begin{pmatrix} \tau_{xx} & \tau_{xy} \\ \tau_{xy} & \tau_{yy} \end{pmatrix} \qquad (IX.3.1)$$

- тензор напряжений,

$$\tau_m = \left[\frac{1}{4}(\tau_{xx} - \tau_{yy})^2 + \tau_{xy}^2\right]^{1/2} \qquad (IX.3.2)$$

- максимальное касательное напряжение, X и Y - составляющие объемных сил, X_ν и Y_ν - составляющие поверхностных усилий. Далее, G - конечная область в плоскости (x,y), $\partial G = \bigcup_{j=0}^{\nu} \partial G_j$

граница области G, ∂G_j – ее связные компоненты. Тензор напряжений удовлетворяет внутри G уравнениям равновесия

$$\tau_{xx,x}+\tau_{xy,y}+X=0\ ,\ \tau_{xy,x}+\tau_{yy,y}+Y=0, \qquad (IX.3.3)$$

а на ∂G - краевым условиям

$$\tau_{xx}\cos(\nu,x)+\tau_{xy}\cos(\nu,y)=X_\nu,\ \tau_{xy}\cos(\nu,x)+\tau_{yy}\cos(\nu,y)=Y_\nu. \qquad (IX.3.4)$$

Если среда находится в упругом состоянии, то по принципу Кастильяно тензор упругих напряжений, который мы обозначим через $\hat{\sigma}$, сообщает минимальное значение интегралу

$$\iint_G [(1-2\sigma)(\tau_{xx}+\tau_{yy})^2+4\tau_m^2]\,dxdy \qquad (IX.3.5)$$

на множестве тензоров, удовлетворяющих уравнениям (IX.3.3-4). По условию пластичности Сен-Венана, в пластическом состоянии в условиях плоской задачи выполняется равенство

$$\tau_m^2=\frac{1}{4}[(\tau_{xx}-\tau_{yy})^2+4\tau_{xy}^2]=\mathcal{K}^2, \qquad (IX.3.6)$$

а тогда, по принципу Хаара - Кармана, тензор T_* упруго-пластических напряжений решает одностороннюю задачу о минимуме интеграла (IX.3.5) на множестве тензоров, удовлетворяющих уравнениям (IX.3.3-4) и, кроме того, неравенству

$$(\tau_{xx}-\tau_{yy})^2+4\tau_{xy}^2\le 4\mathcal{K}^2. \qquad (IX.3.7)$$

2°. Сформулированную здесь одностороннюю вариационную задачу сведем к задаче вида (УШ.I.1-2). Введем гильбертово пространство $\tilde{H}$ тензоров вида (IX.3.1), составляющие которых квадратично суммируемы в G, с нормой

$$|T|=\left\{\iint_G [(1-2\sigma)(\tau_{xx}+\tau_{yy})^2+4\tau_{xy}^2]\,dxdy\right\}^{1/2} \qquad (IX.3.8)$$

и с соответствующим скалярным произведением, а также подпространство H этого пространства, образованное тензорами, удовлетворяющими однородным уравнениям равновесия

$$\tau_{xx,x}+\tau_{xy,y}=0\ ,\ \tau_{xy,x}+\tau_{yy,y}=0 \qquad (IX.3.9)$$

и однородным краевым условиям

$$\tau_{xx}\cos(\nu,x)+\tau_{xy}\cos(\nu,y)=0,\ \tau_{xy}\cos(\nu,x)+\tau_{yy}\cos(\nu,y)=0. \qquad (IX.3.10)$$

Аналогично тому, как это сделано в книге автора [5] для трехмерной задачи теории упругости, можно доказать, что пространство $\tilde{H} \ominus H$ состоит из тензоров, которые удовлетворяют соотношениям (IX.3.3-4) при подходящих X, Y, X_y, Y_y и уравнениям плоской задачи теории упругости; в частности, $\hat{g} \in \tilde{H} \ominus H$.

Обозначим через $\|\cdot\|$ полунорму

$$\|T\| = \operatorname*{supess}_{(x,y)\in G} \tau_m . \qquad (IX.3.11)$$

Положим $\bar{g} = \hat{g} + V$, где V – произвольный тензор из H. Если $\inf_{V\in H} \|\bar{g}\| > \mathcal{K}$, то односторонняя вариационная задача п. I° неразрешима (см. гл. VII, § I, п. 2°). Будем считать, что $\inf_{V\in H} \|\bar{g}\| < \mathcal{K}$. Тогда найдется такой тензор $V_0 \in H$, что $\|g\| < \mathcal{K}$, где $g = \hat{g} + V_0$. Обозначая $T = V - g$, $V \in H$, мы приведем задачу п. I° к виду

$$|V - g| = \min, \quad V \in \mathcal{M};$$
$$\mathcal{M} = \{V \in H : \|V - g\| \le \mathcal{K}\}. \qquad (IX.3.12)$$

Последнюю задачу можно слегка упростить. Имеем $V - g = (V - V_0) - \hat{g}$, слагаемые справа ортогональны в метрике (IX.3.8) и $|V - g|^2 = |V - V_0|^2 + |\hat{g}|^2$. Постоянная величина $|\hat{g}|^2$ не влияет на минимизирующий элемент V, и ее можно отбросить. В результате задача (IX.3.12) заменяется следующей:

$$|V - V_0| = \min, \quad V \in \mathcal{M};$$
$$\mathcal{M} = \{V \in H : \|V - g\| \le \mathcal{K}\}. \qquad (IX.3.13)$$

Аналогично тому, как это сделано в п. 2° §2, можно доказать, что множество $\mathcal{M}$ задачи (IX.3.12) замкнуто в норме $|\cdot|$.

Если граница области достаточно гладкая, то основные задачи линейной теории можно свести к интегральным уравнениям Фредгольма или сингулярным; подробно об этом см. книги Мусхелишвили [1], автора [1,6], Купрадзе и др. [1]. Несложный анализ этих уравнений показывает, что при подходящем выборе необходимых матриц тензор напряжений, решающий вторую краевую задачу теории упругости, устойчив в метрике C по отношению к малым изменениям объемных и поверхностных сил, а также области, занятой упругой средой.

Опираясь на этот факт, мы сделаем следующее допущение: при

подходящем выборе метрики для объемных сил и при подходящем определении малого возмущения области G тензор напряжений $\hat{g}$ устойчив в метрике $C(\bar{G})$ относительно малых возмущений перечисленных параметров; очевидно, при этих же условиях тензор $\hat{g}$ устойчив в метрике (IX.3.8).

При малых возмущениях тензора $\hat{g}$ и постоянном тензоре V_0 тензор $g = \hat{g} + V_0$ оказывается также мало возмущенным; неравенство $\|g\| < \varkappa$ при этом сохраняется. Отсюда следует, что тензор g устойчив как в метрике $C(\bar{G})$, так и в метрике (IX.3.8) относительно малых возмущений внешних сил и области.

Если малое возмущение испытывают только внешние усилия, а область G остается неизменной, то элемент V_0 и метрика (IX.3.8) остаются неизменными, меняется только множество M. Из результатов пп. 5°–7° § 5 гл. I вытекает тогда, что тензор напряжений, решающий плоскую задачу упруго-пластического состояния, устойчив относительно малых возмущений внешних усилий, а также относительно малых возмущений постоянной пластичности $\varkappa$.

3°. Рассмотрим вопрос об устойчивости решения плоской задачи упруго-пластического состояния при малых возмущениях области G. Будем предполагать, что эти возмущения во всяком случае удовлетворяют требованиям, сформулированным в § 2. Временно предположим еще, что возмущенная область $G' \subset G$.

Объекты, связанные с областью G', будем обозначать теми же символами, что и аналогичные объекты, связанные с областью G, но со штрихом сверху. Новая задача равносильна следующей вариационной задаче:

$$\|V' - V_0\| = \min, \quad V' \in M'; \quad M' = \{V' \in H_0, \ \|V' - g'\| \leqslant \varkappa\}. \qquad (IX.3.14)$$

Введем в рассмотрение преобразование S § 2. Объекты, полученные в результате этого преобразования, будем отмечать индексом 1 снизу. В частности,

$$\|T\|_1^2 = \iint_G \left[(1-2\sigma)(\tau_{xx} + \tau_{yy})^2 + \tau_m^2 \right] J \, dx\,dy. \qquad (IX.3.15)$$

где J - якобиан преобразования S. Из свойств этого преобразования и из формулы (IX.3.15) легко вытекает, что норма $\|\cdot\|_1$ удовлетворяет неравенствам вида (VII.5.3) и (VII.5.5). Тензор $\hat{g}$ по предположению устойчив относительно малых возмущений области G;

отсюда теперь легко заключить, что такой же устойчивостью обладают тензоры q и V_0. Из результатов § 5 гл. УШ следует, что тензор напряжений, решающий плоскую задачу теории упруго-пластического состояния, устойчив относительно малых возмущений области, если возмущенная область $G' \subset G$. Последнее требование несущественно - от него можно освободиться так же, как в п.9° § 2.

4°. Для плоской задачи упруго-пластического состояния верны оценки погрешностей аппроксимации и искажения, аналогичные оценкам п.10° § 2. Мы не будем здесь останавливаться на этом подробнее.

5°. Поставим вопрос: при каких обстоятельствах можно утверждать, что решение плоской задачи упруго-пластического состояния, полученное из принципа Хаара - Кармана, является на самом деле решением этой задачи. По самой постановке задачи, тензор T упругих напряжений, полученный на основе этого принципа, удовлетворяет уравнениям равновесия (IX.3.9)(мы принимаем для простоты, что объемные силы отсутствуют). Далее, как показано в работе Хаара и Кармана [1], в упругой зоне этот тензор удовлетворяет уравнению Эйлера, которое в данном случае сводится к дифференциальному уравнению

$$\Delta(\tau_{xx} + \tau_{xy}) = 0 \qquad (IX.3.16)$$

и к краевому условию (если границы области Ω и упругой зоны имеют общую часть)

$$T_\nu = k_\nu ; \qquad (IX.3.17)$$

здесь ν - внешняя нормаль к ∂G, k_ν - вектор напряжений, действующих на "элементарную площадку" границы ∂G. В силу результатов той же статьи [1] Хаара и Кармана, в пластической зоне тензор T удовлетворяет условию пластичности Сен-Венана (IX.1.9).

Таким образом, в упругой зоне три составляющих тензора T удовлетворяют системе трех дифференциальных уравнений (IX.3.9)и (IX.3.16); как известно, если T удовлетворяет этой системе,то можно найти такой вектор смещений, что соответствующий тензор деформаций вместе с тензором T удовлетворяет уравнениям закона Гука.

В пластической зоне тензор T удовлетворяет системе трех

уравнений (IX.3.9) и (IX.1.9). Этот тензор будет решать плоскую задачу упруго-пластического состояния, если в пластической зоне существует вектор смещений, который удовлетворяет уравнениям (IX.1.10-11) и вместе со своей производной по времени непрерывно переходит в вектор упругих смещений при переходе через общую границу упругой и пластической зон (см.Соболев [1]). Можно указать простое достаточное условие, при котором последнее требование выполнено, так что принцип Хаара-Кармана приводит к решению плоской задачи упруго-пластического состояния. Это будет, если данные задачи не зависят от времени. В этом случае тензор напряжений, а также вектор смещений в упругой зоне не зависят от времени: на границе раздела зон упругости и пластичности вектор скоростей смещений равен нулю. Но тогда можно удовлетворить уравнениям (IX.1.10-11), положив вектор скоростей смещений U тождественно равным нулю в пластической зоне.

Что касается вектора смещений в пластической зоне, то он не определяется из уравнений Сен-Венана и его можно подчинить только двум требованиям: он непрерывно переходит в вектор упругих смещений на общей границе обеих зон; он не зависит от времени.

Заметим, что высказанные здесь соображения в полной мере относятся и к трехмерной задаче упруго-пластического состояния (см.ниже, § 4).

6°. В заключение данного параграфа заметим следующее. Полунорма (IX.3.11) не является непрерывно дифференцируемой, и построение приближенного решения наталкивается на некоторые трудности. Можно упростить задачу, заменив полунорму (IX.3.11) следующей

$$\|T\|_\rho = \frac{1}{|G|^{1/2\rho}}\|\tau_{in}\|_{L_\rho} = \tag{IX.3.18}$$

$$= \frac{1}{|G|^{1/2\rho}}\left\{\iint_G \left[\frac{1}{4}(\tau_{xx}-\tau_{yy})^2 + \tau_{xy}^2\right]^\rho dxdy\right\}^{1/2\rho},$$

а неравенство $\|V-\bar{q}\| \le \mathcal{K}$ – неравенством

$$\|V-\bar{q}\|_\rho \le \mathcal{K}. \tag{IX.3.19}$$

В формуле (IX.3.18) ρ достаточно большое натуральное число; очевидно, новая полунорма непрерывно дифференцируема. Обозначим множество тензоров V, удовлетворяющих неравенству (IX.3.19), через $\mathcal{M}_\rho$. Так как $\|\cdot\|_\rho \le \|\cdot\|$, то множество $\mathcal{M}_\rho$ не пусто, если не пусто множество $\mathcal{M}$. Далее, очевидно, что множество $\mathcal{M}_\rho$ выпук-

лое. Докажем, что оно замкнуто в норме $\|\cdot\|_p$. Пусть $\{V_n\} \subset {}^1l_p$ - последовательность сходящаяся в упомянутой норме (формула (IX.3.8)) к некоторому тензору V_0. Тогда $f_n \to f_0$ и $g_n \to g_0$ в $L_2(G)$, где

$$f_n = V_{nxx} - V_{nyy}, \quad g_n = V_{nxy},$$
$$f_0 = V_{0xx} - V_{0yy}, \quad g_0 = V_{0xy}.$$

Положим еще $g_{xx} - g_{yy} = \tilde{f} - \tilde{\tilde{f}}$, тогда в $L_2(G)$

$$f_n - \tilde{f} \to f_0 - \tilde{f}, \quad g_n - \tilde{g} \to g_0 - \tilde{g}. \qquad (IX.3.20)$$

Отсюда вытекает существование такой подпоследовательности $\{V_{n_k}\}$, что те же соотношения (IX.3.22) для этой подпоследовательности выполняются почти всюду в G. По теореме Лагу (см., напр мер, Вулих [I])

$$\|V_0 - V_0\|_p = \frac{1}{|G|^{1/p}} \left\{ \iint_G \left[\frac{1}{4} (f_0 - \tilde{f})^2 + (g_0 - \tilde{g})^2 \right]^{p} dx dy \right\}^{1/2p} \leq$$

$$\leq \frac{1}{|G|^{1/p}} \sup_k \operatorname{ess} \left\{ \iint_G \left[\frac{1}{4} (f_{n_k} - \tilde{f})^2 + (g_{n_k} - \tilde{g})^2 \right]^{p} dx dy \right\}^{1/2p} =$$

$$= \sup_k \|V_{n_k} - g\|_p \leq \mathcal{K}.$$

Это означает, что $V_0 \in {}^1l_p$ и, следовательно, множество замкнуто.

Аналогично можно упростить задачи § 2 и § 4 (см.ниже).

§ 4. Трехмерная задача

I°. Пусть известны векторы $\vec{\mathcal{K}}$ - объемных и $\vec{\mathcal{K}}_\nu$ - поверхностных сил, действующих на упруго-пластическую среду, которая занимает конечную область Ω трехмерного пространства. Тензор напряжений удовлетворяет в Ω дифференциальному уравнению

$$\operatorname{div} T + \vec{\mathcal{K}} = 0, \qquad (IX.4.1)$$

а на границе этой области - краевому условию

$$T \cdot \nu|_{\partial\Omega} = \vec{\mathcal{K}}_\nu; \qquad (IX.4.2)$$

ν - внешняя норма к $\partial\Omega$.

Если рассматриваемая среда находится в упругом состоянии, то по принципу Кастильяно тензор T сообщает минимальное значение функционалу

$$\iiint_{\Omega}\left[\frac{1-2\sigma}{6(1+\sigma)}(\tau_{xx}+\tau_{yy}+\tau_{zz})^2+\tau_i^2\right]dxdyd\ldots \qquad (IX.4.3)$$

на множестве тензоров, удовлетворяющих уравнениям (IX.4.1-2); через τ_i обозначена интенсивность касательных напряжений. Применяя принцип Хаара-Кармана, будем считать, что тензор напряжений в упруго-пластическом состоянии сообщает минимальное значение функционалу (IX.4.3) на множестве тензоров, которые удовлетворяют, кроме уравнений (IX.4.1-2), еще и неравенству (IX.1.23).

2°. Только-что сформулированную одностороннюю вариационную задачу преобразуем к обычному виду. Примем, что тензор напряжений квадратично суммируем в Ω. На множестве таких тензоров введем норму

$$\|T\|^2=\iiint_{\Omega}\left[\frac{1-2\sigma}{1+\sigma}(\tau_{xx}+\tau_{yy}+\tau_{zz})^2+\tau_i^2\right]dxdydz, \qquad (IX.4.4)$$

и обозначим полученное таким образом гильбертово пространство через $\tilde{H}$. Выделим множество H элементов пространства $\tilde{H}$, которые удовлетворяют однородным уравнениям

$$div\,T=0\,,\quad T\cdot\nu|_{\partial\Omega}=0. \qquad (IX.4.5)$$

Как доказано в книге автора [3] H есть подпространство в $\tilde{H}$. Ортогональное к H подпространство состоит из тензоров упругих напряжений: это значит, что если $T\in\tilde{H}\ominus H$, то T удовлетворяет уравнениям (IX.4.1-2) при подходящем выборе векторов $\mathcal{K}$ и $\mathcal{K}_\nu$ и существует такой вектор смещений $U=(U_x, U_y, U_z)$, что

$$\tau_{xx}=\lambda\,div\,U+2\mu U_{xx},\quad \tau_{xy}=\mu(U_{xy}+U_{yx})$$

и т.п. Здесь λ и μ – постоянные Ляме для рассматриваемой упругой среды.

Пусть $\hat{T}$ – тензор напряжений, отвечающий для области Ω задачу теории упругости при заданных $\mathcal{K}$ и $\mathcal{K}_\nu$; составляющие этого тензора пусть будут $\hat{T}_{xx}$, $\hat{T}_{xy}$, …, $\hat{T}_{zz}$. Ясно,

что $\hat{q} \in \tilde{H} \ominus H$. Положив $T = V - \hat{q}$, мы сведем вариационную задачу п. 1^0 к следующей:

$$|V-\hat{q}| = \min, \quad V \in M = \{V \in H : \|V-\hat{q}\| \leq \mathcal{K}\}; \qquad (IX.4.6)$$

$$\|T\| = \tau_s .$$

Как обычно, задача (IX.4.6) неразрешима, если $\varkappa = \inf_{V \in H} \|V-\hat{q}\| > \mathcal{K}$; она разрешима единственным образом, если $\varkappa < \mathcal{K}$; неразрешима или имеет неустойчивое решение, если $\varkappa = \mathcal{K}$. Будем считать, что $\varkappa < \mathcal{K}$. Тогда существует такой тензор $V_0 \in H$, что $\|q\| < \mathcal{K}$, $q = V_0 + \hat{q}$. Обозначим $V = V' - V_0$. Тогда $V - \hat{q} = V' - q$. Далее, $|V-\hat{q}|^2 = |V'-\hat{q}|^2 = |V'-V_0|^2 + |\hat{q}|^2$; заменив обозначение V' на V, мы приведем задачу (IX.4.6) к виду

$$|V-V_0| = \min, \quad V \in M;$$
$$M = \{V \in H : \|V-q\| \leq \mathcal{K}\}. \qquad (IX.4.7)$$

Примем в пущение, аналогичное тому, которое было принято при анализе плоской задачи: тензор упругих напряжений $\hat{q}$ устойчив в метриках C и (IX.4.4) относительно малых возмущений векторов k и $\mathcal{F}_V$ и области Ω. Тогда (ср. п. 3^0 § 2) относительно тех же возмущений устойчив и тензор q. Повторив рассуждения п.3^0 § 2, мы убедимся, что решение трехмерной задачи упруго-пластического состояния устойчиво относительно перечисленных выше малых возмущений, а также относительно малых возмущений постоянной пластичности $\mathcal{K}$.

3^0. Легко убедиться, что общая оценка погрешности аппроксимации (VII.2.11) верна для задачи настоящего параграфа: если T_* – точное решение этой задачи а $T_*^{(n)}$ – его приближенное решение, полученное заменой пространства H его конечномерным подпространством H_n, то

$$|T_* - T_*^{(n)}| = O(\sqrt{\varepsilon_n(T_*)}); \quad \varepsilon_n(T_*) = \inf_{T_n \in H_n} \|T_* - T_n\|; \qquad (IX.4.8)$$

$$\|T\| = |T| + \|T\| .$$

Однако так же верна и общая оценка погрешности искажения (VII.6.1^0).

4^0. В заключение отметим некоторые работы, в которых ис-

следуются вариационные неравенства, относящиеся к задачам упруго-пластического состояния. В работе [1] Гаевский рассмотрел две задачи упруго-пластического состояния: о плоском напряженном состоянии и о кручении цилиндрического стержня; обе задачи исследованы в вариационной постановке для случая односвязной области. Каждая из этих задач сведена к уравнению с оператором, удовлетворяющим условию теоремы Браудера о неподвижной точке. Лангенбах [2] получил аналогичный результат для трехмерной задачи. Оценки устойчивости для решений некоторых вариационных неравенств дал Гаевский [3] ; в качестве примера им рассмотрено, в частности, прямолинейное течение вязко-пластической жидкости в трубе. В работе [4] того же автора принцип Хаара-Кармана используется в рамках теории пластичности Прандтля-Рейсса.

ЛИТЕРАТУРА

Агмон С., Дуглис А., Ниренберг Л. I. Оценки решений эллиптических уравнений вблизи границы. Перев. с англ. 1962, М., с.205.

Атакишиева Р.Х. I. Устойчивость метода Бубнова-Галеркина для уравнения с вполне непрерывным оператором в банаховом пространстве. Докл.АН Азерб.ССР, 1966, т.22, № I, 7-II.

Бабушка И., Витасек Э., Прагер М. I. Численные процессы решения дифференциальных уравнений. Перев. с англ. 1969, М., с. 368.

Балакришнан А.В. I Прикладной функциональный анализ. Перев. с англ. 1980, М., с.383.

Бари Н.К. I. Обобщение неравенств С.Н.Бернштейна и А.А.Маркова. Изв. АН СССР, серия матем., 1954, I8, № 2, 159-176.

Березин И.С., Жидков Н.П. I. Методы вычислений, т.I, 1966, М., с.632.

Бирман М.Ш. I. О методе Фридрихса расширения положительно определенного оператора до самосопряженного. Зап.Ленингр.Горн. ин-та, 1956, 33, № 3, 132-136.

Вазов В., Форсайт Дж.I. Разностные методы решения дифференциальных уравнений в частных производных. Перев. с англ., 1963, М., с.487.

Вайникко Г.М. I. Оценки погрешности метода Галеркина для линейного дифференциального уравнения. Уч.зап.Тартуск. ун-та, 1962, вып. I29, Труды по матем. и мех., III, 394-416.

- - - 2. Оценка погрешности метода Ритца для линейного однородного уравнения. Уч. зап. Тартуск. ун-та, 1962, вып.I29. Труды по матем. и мех., III, 417-427.

- - - 3. Некоторые оценки погрешности метода Бубнова-Галеркина. I. Асимптотические оценки. Уч.зап. Тартуск. ун-та, 1963, вып. 150. Труды по матем. и мех., IУ, 188-201.

- - - 4. Некоторые оценки метода Бубнова-Галеркина. II. Оценки n-го порядка. Уч. зап. Тартуск. ун-та, 1964, вып.150. Труды по матем. и мех., I., с. 2-5.

- - - 5. Асимптотические оценки погрешности проекционных методов в проблеме собственных значений. ЖВиМФ, 1964, 4, № , 405-415.

- - - 6. Оценки погрешности метода Бубнова-Галеркина в проблеме собственных значений. ЖВМиМ., 1965, 5, № 4, 587-607.

- - - 7. Необходимые и достаточные условия устойчивости метода Галеркина - Петрова. Уч. зап. Тартусск. ун-та, 1965, вып. 177, 141-147.

- - - 8. О сходимости и устойчивости метода коллокации. Диф. ур-ия, 1965, 1, № 2, 244 254.

- - - 9. О быстроте сходимости некоторых приближенных методов типа Бубнова - Галеркина в проблеме собственных значений. Изв. ВУЗ. Математика, 1966, № 2, 55-45.

- - -10. О сходных операторах. ДАН СССР, 1968, 179, № 5, 1029-1031.

- - -11. Анализ дискретизационных методов. Тарту, 1976, с.181.

Вайникко Г.М., Михлин С.Г. 1. Резольвента Фредгольма и обращение матрицы, линейно зависящей от параметра. Уч.зап.Тартусск. ун-та, 1978, вып. 448. Труды по матем. и мех., XXI, 94-98.

Варга Р.С. 1. Функциональный анализ и теория аппроксимации в численном анализе. Перев. с англ. 1974, М., с. 126.

Велиев М.А. 1. Исследование устойчивости метода Бубнова - Галеркина для нестационарных задач. ДАН СССР 1964, 157, № 1, 16-19.

- - - 2. Об устойчивости метода Бубнова - Галеркина для нестационарных задач. В сборн. "Вопросы вычисл. матем. и вычисл техн.", изд. АН Азерб.ССР, 1964, вып. 3.

- - - 3. Об устойчивости метода Бубнова - Галеркина для уравнения 2-го порядка в гильбертовом пространстве. Сиб. матем ж., 1968. IX, № 3, 783-789.

- - - 4. Об устойчивости метода Бубнова - Галеркина для уравнений первого порядка с переменными коэффициентами в гильбертовом пространстве. Диф. ур-ия, 1969, 5, № 3, 473-487.

- - - 5. Некоторые достаточные условия устойчивости метода Бубнова - Галеркина для уравнений второго порядка в гильбертовом пространстве. Уч.зап. Азерб. ун-та, серия физ.-мат.наук, 1972, № 2, 17-24.

- - - 6. К устойчивости метода Бубнова - Галеркина для одного класса вырождающихся параболических уравнений. Науч.труды

Азерб. ун-та; Вопросы прикл. матем. и киберн., № 1979, № 1, 30-34.

– – – 7. Об устойчивости метода конечных элементов для параболических уравнений. Докл. АН Азерб.ССР, 1981, 37, № 4, 3-6.

– – – 8. К устойчивости метода Бубнова – Галеркина для линейных уравнений в гильбертовом пространстве. Докл. АН Азерб. ССР, 1981, 37, № 5, 3-6.

– – – 9. Устойчивоть метода Бубнова – Галеркина для линейных уравнений второго порядка с переменными коэффициентами. Докл. АН Азерб.ССР, 1981, 37, № 7, 8-11.

Воеводин В.В. 1. Ошибки округления и устойчивость в прямых методах линейной алгебры. 1969, М., с.153.

Вулих Б.З. 1. Краткий курс теории функций вещественной переменной. 1973, М., с.350.

Габдулхаев Б.Г. 1. Об одном прямом методе решения интегральных уравнений. Изв. ВУЗ, Математика, 1965, № 3, 51-60.

– – – 2. Приближенное решение сингулярных интегральных уравнений методом механических квадратур. ДАН СССР, 1968, 179, № 2, 260-263.

– – – 3. Оптимальные аппроксимации решений линейных задач. 1980, Казань, с.231.

Габдулхаев Б.Г., Душков П.Н. 1. Метод механических квадратур для сингулярных интегральных уравнений. Изв.ВУЗ, Математика, 1974, № 12 (151), 3-14.

Гавурин М.К. 1. Лекции по методам вычислений. 1971, М., с.248.

Галин Л.А. Упруго-пластическое кручение отерженей полигонального сечения. Прикл.матем. и мех., 1944, 8, № 4, 307-322.

– – – 2. Упруго-пластическое кручение призматических стержней. Прикл. матем. и мех., 1949, 13, № 3, 287-296.

Гельман И.В.К задаче о минимума нелинейного функционала. Уч. зап. Лен. Пед. ин-та им. Герцена, 1958, 166, 255-263.

Гловински Р., Лионс Ж.-Л., Тремольер Р. 1. Численное исследование вариационных неравенств. Перев. с франц., 1979, М., с.574.

Годунов С.К., Рябенький В.С. 1. Введение в теорию разностных схем. 1962, М., с.340.

Головкин К.К. I. О приближении функций в произвольных нормах. Труды матем.ин-та им.Стеклова АН СССР, 1964, 70, 26-37

Гохберг И.Ц., Фельдман И.А. I. Уравнения в свертках и проекционные методы их решения. 1971, М., с.352.

Гурса Э. I. Курс математического анализа, т.Ш, ч.П. Перев. с франц., 1934, М., с.318.

Гусман Ю.А., Оганесян Л.А. I.Оценка сходимости конечноразностных схем для вырожденных эллиптических уравнений. ЖВМиМФ 1965, 5, № 2, 351-357.

Демьянович Ю.К. I. Устойчивость метода сеток для эллиптических задач. ДАН СССР, 1965, 164, № I, 20-23.

- - - 2. Об устойчивости проекционных методов. Зап.научн. семин.ЛОМИ АН СССР, 1971, 23, 5-15.

- - - 3. Об устойчивости и длительности вычислений в вариационно-разностном методе. Зап.научн.семин.ЛОМИ АН СССР, 1978, 80, 5-29.

- - - 4.Об оценках скорости сходимости проекционных методов решения некоторых вариационных неравенств. Деп. в ВИНИТИ, 1980, № 1247-80 ДЕП, с.47

Демьянович Ю.К., Михлин С.Г. I.О точной аппроксимации функций соболевских пространств. Зап.научн.семин.ЛОМИ АН СССР, 1973, 35, 6-11.

Джишкариани А.В. I. Некоторые оценки погрешности метода Ритца - Галеркина для обыкновенных дифференциальных уравнений. Труды Тбил. матем. ин-та, 1959, 25, 285-306.

- - - 2. Оценки погрешности метода Ритца для неоднородного дифференциального уравнения. Сообщ. АН Груз.ССР, 1960, 25, № 3, 257-262.

- - - 3. О скорости сходимости приближенного метода Ритца. ЖВМиМФ, 1963, т.3, № 4, 654-660.

- - - 4. О скорости сходимости метода Бубнова - Галеркина. ЖВМиМФ, 1964, 4, № 2, 343-348.

- - - 5. О скорости сходимости обобщенных методов Ритца - Бубнова - Галеркина. Сообщ. АН Груз. ССР, 1967, 46, № 3, 553-560.

- - - 6. О методах Ритца и Бубнова - Галеркина. Сообщ. АН Груз.ССР, 1967, 47, № I, 11-16.

- - - 7. О методе Бубнова - Галеркина. ЖВиМФ, 1967, 7, № 6, 1398-1402.

- - - 8. О методе Ритца для одного нелинейного уравнения. Сообщ. АН Груз. ССР, 1968, 51, № 1, 19-24.

- - - 9. О методах наименьших квадратов и Бубнова - Галеркина. ЖВиМФ, 1968, 8, № 5, 1110-1116.

- - - 10. О приближенном методе Ритца для одного нелинейного уравнения. Труды Тбил. матем. ин-та, 1969, 36, 29-46.

- - - 11. Устойчивость метода Ритца для одного нелинейного уравнения. ЖВиМФ, 1970, 10, № 4, 841-847.

- - - 12. К решению сингулярных интегральных уравнений приближенными проекционными методами. ЖВиМФ, 1979, 19, № 5, 1149-1161.

- - - 13. О сходимости невязки в методе конечных элементов. ДАН СССР, 1980, 254, № 5, 1052-1055.

- - - 14. К решению сингулярных интегральных уравнений коллокационными методами. ЖВиМФ, 1981, 21, № 2, 355-362.

Дочбеш Л.Н. 1. Устойчивость метода Ритца для задач спектральной теории операторов. Труды матем. ин-та им. Стеклова АН СССР, 1965, 84, 78-92.

Дюво Г., Лионо Ж.-Л. 1. Неравенства в механике и физике. Перев. с франц. 1980. М., с.383.

Запрудский Я.М.1. Стійкість методу Гальоркіна - Петрова у банаховому просторі. Докл.АН УССР, серия А, 1972, № 4, 309-312.

Ильин В.П. 1. О сходимости вариационных процессов. ДАН СССР. 1951, 81, № 2, 137-140.

- - - 2. Оценка погрешности в методе Ритца для обыкновенных дифференциальных уравнений. Труды матем.ин-та им.Стеклова АН СССР, 1959, 53, 43-63.

- - - 3. Некоторые неравенства в функциональных пространствах и их применение к исследованию сходимости вариационных процессов. Труды матем.ин-та им.Стеклова АН СССР, 1959, 53, 64-127.

- - - 4. Свойства некоторых классов дифференцируемых функций многих переменных, заданных в n-мерной области. Труды матем.ин-та им.Стеклова АН СССР, 1962, 66, 227-363.

- - - 5. Некоторые замечания о сходимости последователь-

ности функций полиномиального типа в пространствах $W_p^{(\alpha)}(G)$. Зап. научн.семин.ЛОМИ АН СССР, 1968, II, 73-96.

Канторович Л.В. I. Об одном методе приближенного решения дифференциальных уравнений в частных производных. ДАН СССР, 1934, 2, 532-536.

- - - 2. Об одном эффективном методе решения экстремальных задач для квадратичных функционалов. ДАН СССР, 1945, 48, № 7, 455-460.

- - - 3. Функциональный анализ и прикладная математика. УМН, 1948, 3, № 6(28), 89-185.

Канторович Л.В., Акилов Г.П. I. Функциональный анализ, изд. 2-е, М., 1977. с.741.

Канторович Л.В., Крылов В.И. I. Приближенные методы высшего анализа, изд. 3-е, Л.-М., 1949 с. 695.

Каршиловская Э.Б. I. О сходимости интерполяционного метода для обыкновенных дифференциальных уравнений. УМН, 1953, 8, № 3 (55), III-118.

- - - 2. О сходимости метода коллокации. ДАН СССР, 1963, 151, № 4, 736-769.

- - - 3. О сходимости метода коллокации для некоторых граничных задач математической физики. Сиб.матем.ж., 1963, 4, № 3, 632-640.

Колмогоров А.Н., Фомин С.В. I. Элементы теории функций и функционального анализа. Изд. 5-е. 1981, М., с. 543.

Корнеев В.Г. I. Некоторые вопросы построения и исследования схем метода конечных элементов. Числ.методы мех. сплошн. среды, 1974, № 5, 59-87.

Корнеев В.Г., Пономарев С.Е. I. Применение криволинейных конечных элементов в схемах решения линейных эллиптических уравнений порядка 2m. Числ.методы мех.сплошн.среды, 1974, 5, № 5

Красносельский М.А. I. Сходимость метода Галеркина для нелинейных уравнений. ДАН СССР, 1950, 73, № 6, 1121-1124.

Красносельский М.А. и др. I. Приближенное решение операторных уравнений. 1969, М., с. 455.

Крылов Н.М. и. Избранные труды. т.3, 1961, Киев, с.393.

Купрадзе и др. I. Трехмерные задачи математической теории упругости и термоупругости. 1971, М., с.333.

Ладыженская О.А., Уральцева Н.Н. I. Линейные и квазилинейные уравнения эллиптического типа, изд. 2-е, 1973, М., с.576.

Лионо Ж.-Л. I. Оптимальное управление системами, описываемыми уравнениями в частных производных. Перев. с франц., 1972, М., с.414.

- - - 2. Некоторые методы решения нелинейных краевых задач. Перев. с франц., 1972, М., с. 587.

Мамаков А.З. I. Применение метода Галеркина к некоторому квазилинейному уравнению параболического типа. Вестн.ЛГУ, 1981, № 13, 37-45.

- - - 2. О погрешности и устойчивости метода Бубнова - Галеркина для квазилинейных параболических задач с производной по времени в граничном условии. Деп. в ВИНИТИ, 1981, № 146-81 ДЕП, с.36.

Марчук Г.И. I. Методы вычислительной математики, 1973, Новосибирск, с. 352.

Михлин С.Г. I. Интегральные уравнения и их приложения, М.-Л., 1949, с. 280.

- - - 2. Проблема минимума квадратичного функционала. 1952, М., с.216.

- - - 3. Вариационные методы в математической физике, изд. 2-е, 1970, М., с.512.

- - - 4. Об устойчивости метода Ритца. ДАН СССР, 1960, 135, № 1, 16-19.

- - - 5. Некоторые условия устойчивости метода Ритца. Вестн. ЛГУ, 1961, № 19, 40-51.

- - - 6. Многомерные сингулярные интегралы и интегральные уравнения. 1962, М., с. 254.

- - - 7. Об устойчивости некоторых вычислительных процессов. ДАН СССР, 1964, 157, № 2, 271-273.

- - - 8. Численная реализация вариационных методов. 1966, М., с. 432.

- - - 9. О функциях Коссера. В сборн.: "Проблемы матем. анализа", изд. ЛГУ, 1966, 59-69.

- - - 10. Курс математической физики. 1968, М., с. 575.

- - - 11. Неравенства типа А.А.Маркова и полиномиальные приближения по Ритцу. В книге: "Дифференциальные уравнения с частными производными" 1970, М., 158-169.

- - - I2. О вариационно-разностном методе для многомерных краевых задач. Зап.научн.семин.ЛОМИ АН СССР, 1971, 23,99-114.

- - - I3. О вариационно-разностном методе для одномерных краевых задач. ДАН СССР, 1971, 198, № I, 39-41.

- - - I4. О сеточной аппроксимации вырождающихся одномерных дифференциальных уравнений второго порядка. Вестн.ЛГУ,1973, № I, 52-67.

- - - I5. Спектр пучка операторов теории упругости. УМН, 1973, 28, № 5(171), 43-82.

- - - I6. О числе обусловленности вариационно-сеточной матрицы. Вестн. ЛГУ, 1973, № I3, I62-I64.

- - - I7. Вариационно-сеточная аппроксимация. Зап.научн. семин.ЛОМИ АН СССР, 1974, 48, 32-I38.

- - - I8. Об одном методе приближенного решения интегральных уравнений. Вестн.ЛГУ, 1974, № I3, 26-33.

- - - I9. О новом методе приближенного решения интегральных уравнений. Вестн.ЛГУ, 1976, № I, 35-44.

- - - 20. Замечания о резольвентном методе. Зап.научн.семин.ЛОМИ АН СССР, 1976, № 56, 5-13.

- - - 2I. О приближенном решении интегральных уравнений со слабой особенностью. Вестн. ЛГУ, 1976, № I3, 3I-36.

- - - 23. Об одном варианте резольвентного метода. Вестн. ЛГУ, 1978, № I, 54-59. — ? 22

- - - 24. О постоянных множителях в оценках погрешности вариационно-сеточной аппроксимации Зап.Научн.семин.ЛОМИ АН СССР, 1978, 80, 125-166.

- - - 25. О методе наименьших квадратов для многомерных сингулярных интегральных уравнений. В сборн. "Комплексн.анализ и его прилож.". К 70-летию акад.И.Н.Векуа, 1978, М., 401-408

- - - 26. О приближенном решении односторонних вариационных задач. УМН, 1979, 34, № 4(208), с.166.

- - - 27. О приближенном решении односторонних вариационных задач. Изв. ЛГУ, Математика, 1980. № 2(213), 45-58.

- - - 28. О погрешности приближенного решения односторонних вариационных задач. Вестн. ЛГУ, 1980, № I3, 46-50.

- - - 29. О устойчивости решений односторонних вариационных задач; приложения к теории пластичности. Зап. науч. се-

мин. ЛОМИ АН СССР, 1980, 102, 68-101.

- - - 30. Погрешность искажения в простейшей односторонней вариационной задаче. Вестн.ЛГУ, 1981, № 13, 45-50.

- - - 32. О погрешности метода Бубнова - Галеркина для эллиптических краевых задач. Зап.научн.семин.ЛОМИ АН СССР, 1981, 111, 137-144.

- - - 33. О погрешностях вычислительных процессов I. Изв. ВУЗ. Математика, 1981, № 7(230), 62-71.

- - - 34. О погрешностях вычислительных процессов. II. Изв. ВУЗ. Математика, 1981, № 8(231), 32-38.

Михлин С.Г., Радова Р.К., 1. О приближенном решении сингулярных интегральных уравнений. Изв. ВУЗ, Математика, 1974, № 5 (144). 158-162.

Михлин С.Г., Смолицкий Х.Л. 1. Приближенные методы решения дифференциальных и интегральных уравнений. 1965, М., с.383.

Мусхелишвили Н.И. 1. Некоторые задачи теории упругости, изд.3-е. 1940, М., с.335.

Надаи А. 1. Пластичность. Перев. с англ., 1936, М.-Л., с.280.

Натансон И.П. 1. Конструктивная теория функций. 1949, М.-Л., с.688.

Оден Д.-П. 1. Приближенное решение эллиптических краевых задач. 1977, М., с.383.

Отнесян Л.А., Руховец Л.А. 1. Вариационно-разностные методы решения эллиптических уравнений. 1979, Ереван, с.335.

Оден Дж. 1. Конечные элементы в нелинейной механике сплошных сред. Перев. с англ., 1976, М., с.464.

Перлин П.И. 1. Упруго-пластическое кручение стержней овального поперечного сечения. Инж.сборник, 1961, 31, 202-205.

Руховец Л.А. 1. К вопросу о построении вариационно-разностных схем для эллиптических уравнений. ЖВМиМФ, 1972, 12, № 3, 787-796.

Смирнов В.И. 1. Курс высшей математики, т.Ш, ч.I. 1949, М.-Л., с.335.

Соболев С.Л. 2. Введение в теорию кубатурных формул. 1974, М., с.808.

Стренг Г. Фикс Дж. 2. Теория метода конечных элементов. Перев. с англ., 1977, М., с.349.

Суворов С.Г. I. К вопросу об устойчивости вычислительного процесса. Вестн.ЛГУ, 1972, № 7, 105-116.

Сьярле Ф. I. Метод конечных элементов для эллиптических задач. Перев. с англ., 1980, М., с.512.

Талденкин А.Т. I. Системы элементов гильбертова пространства и ряды по ним. Матем.сборн., 1951, 29(71), № I, 79-120.

Теория пластичности. Сборник статей. I. Перев. с англ., франц. и нем., 1948, М., с.452.

Тополянский Д.Б., Запрудский Я.М.I. Исследование устойчивости метода Бубнова-Галеркина для некоторых операторных уравнений с переменными коэффициентами. Укр.матем.ж., 1974, 26, №5, 621-634.

Толтин В.Б. I. Об оценке погрешности приближенных решений односторонних вариационных задач. Вестн. ЛГУ, 1981, № 13, 65-69.

- - - 2. О скорости сходимости приближенных методов решений односторонних вариационных задач. I.Вестн.ЛГУ, 1982, № 13, 111-118.

Уилкинсон Дж.Х. 2. Алгебраическая проблема собственных значений. Перев. с англ., 1970, М., с.564.

Фаддеев Д.К., Фаддеева В.Н. I. Вычислительные методы линейной алгебры, изд. 2-е, 1963, М.-Л., с.734.

Харрик И.Ю.I.О приближении функций, обращающихся на границе в нуль, функциями особого вида. Матем.сборн., 1955, 37, 353-384.

- - - 2. О приближении функций, обращающихся в нуль вместе с градиентом на границе области, функциями особого вида. Матем. Сборн., 1959, 47(89), № 2, 177-208.

- - - 3. О приближении функций, обращающихся в нуль на границе области вместе с частными производными, функциями особого вида. Сиб.матем.ж., 1968, 4, № 2, 408-425.

Черноусько Ф.Л., Баничук Н.I. I. Вариационные задачи механики и управления. 1973, М., с.238.

Чистов К.Е. I. Вариационно-сеточная аппроксимация на регулярной треугольной сетке. Изв. ВУЗ, Математика, 1981, № 11, 58-63.

- - 2. Устойчивость вариационно-сеточного процесса на симплектических сетках. Изв.ВУЗ, Математика, 1982, № 4, 81-85.

Шапошникова Т.О. 1. Некоторые оценки сходимости вариационных методов. Вестн. ЛГУ, 1971, № 19, 64-69.

- - - 2. Априорные оценки погрешности некоторых вариационных методов. Диссертация, ЛГУ, 1973, с.93.

- - - 3. Априорные оценки погрешности некоторых вариационных методов в соболевских пространствах. В сборн.: Вариационно-разностные методы в матем. физ. СО АН СССР, 1976, Новосибирск, 182-194.

- - - 4. Априорные оценки погрешности вариационных методов в банаховых пространствах. ЖВМиМФ, 1977, 17, № 5, 1144-1152.

Яскова Г.Н., Яковлева М.Н. 1. Некоторые условия устойчивости метода Петрова - Галеркина. Труды Матем. ин-та им. Стеклова АН СССР, 1962, 66, 182-189.

Agmon S., Douglis A., Nierenberg L. 2. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. Comm. on pure and appl. Math., 1964, XVII, 35-92.

Andersen R.S., de Hoog F.R., Lucas M.A. (ed.). 1. The application and numerical solution of integral equations. Alphen aan den Rijn, 1980, pp.259.

Anselon 1. Collectively compact operator approximation theory and applications to integral equations. Englewood Cliffs, 1971, pp.138.

Berger G., Marx M., 1. Interpolation über ein Dreieoksnets. Wiss. Zeitschr d.Univ. Halle, 1976, 25, N 1, 43-48.

Bertram G. 1. Verschärfung einer Fehlerabschung zum Ritz-Galerkinschen Verfahren von Kryloff für Randwertaufgaben. Numerische Math., 1959, 1, N 3, 135-141.

Börsch-Supan W. 1. Bemerkungen zur Fehlerabschätzung beim Ritz - Galerkin Verfahren nach Krylow Numerische Math., 1960, 2, N 2, 79-83.

Bramble J.H., Zlamal M. 1. Triangle elements in the finite element method. Math. of Comput., 1970, 24, N 112, 809-820.

Brezis H. 1. Problèmes unilateraux. J. de Math. pure et appliquée. 1972, 51, 1-168.

Brezis H., Stampacchia G. 1. Sur la régularité de la solution d'inéquation elliptiques. Bull. Soc. Math. de France, 1968, 96, 153-180.

Brezzi F., Hager W.W., Raviart P.A. 1. Error estimates for the finite element solution of variational inequalities. Part I, Primal theory, Numerische Math., 1977, 28, 431-443.

Calderon A.P. 1. Lebesgue spaces of differentiable functions and distributions. Partial differential equations Amer. Math.Soc.Proc. of symposia in pure mathematics, 1961, IV, 33-49

Courant R. 1. Variational methods for the solution of equilibrium and vibrations. Bull.Amer.Math.Soc., 1943, 49, N 1, 1-23.

Delves L.M., Walsh J. (ed.). 1. Numerical solution of integral equations. Oxford, 1974, pp.339.

El Kolli A. 1. Régularité de la solution et majoration de l'erreur d'approximation pour un problème de Dirichlet non linéaire. Compt. Rend Acad.Sci.Franç.Ser A, 1973, 277, 735-737.

Elliott C.M. 1. On the finite element approximation of an elliptic variational inequality arising from an implicite time discretisation of the Stefan problem. IMA Journ. of numer. analysis. 1981, 1, N 1, 115-126.

Falk R.S. 1. Error estimates for the approximation of a class of variational inequalities. Math.Comput., 1974, 28, 963-971.

- - - 2. Approximation of an elliptic boundary value problem with unilateral constraintes. RAIRO, Anal.Numer., 1975, 9, 5-12.

Gajewski H. 1. Zur Lösung einer Klasse konvexer Minimalprobleme der Plastizitätstheorie. Zeitschr. f. angew. Math. und Mech., 1969, 49, h. 1/2, 83-89.

- - - 2. Zur numerischen Stabilität des Ritzschen Verfahrens bei nichtlinearen Gleichungen. Math.Nachr., 1970, 43, N 1-6, 2[illegible]-311

- - - 3. Stabilitätsaussagen für einige Aufgaben mit freier Grenze. Zeitschrift f. angew.Math. und Mech., 19[illegible], 57, 439-447.

- - - 4. Ein Verfahren zur Ermittelung elastisch-plastischen Spannungsfelder. Zeit. f. Spannungs- und Dehnungsanalyse, herausg. von K.Schröder, Berlin. 1968, V 15-[illegible].

[illegible]ominski [illegible], Larchon [illegible]. 1. Torsion élasto-plastique d'une barre cylindrique de section multi-connexe. Journ. de mécanique, 19[illegible]3. 12, 151-171.

Glowinski R., Marocco A. 1. Sur l'approximation par éléments finis d'ordre un, et la résolution par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. RAIRO, Anal.numér., R-2, 1975, 41-76.

Golub G.M. 1. Bounds for roun-off errors in the Richardson second order method. Nordisk Tidscriftt for Infomationsbehandling (BIT), 1962, 2, H. 4, 212-223.

Haar A., v.-Karman T. 1, Zur Theorie der Spannungszustande in plastischen und sandertigen Zuständen. Nachr. d. Königlich. Gesellsch. d. Wissensch. zu Göttingen, 1909, H. 2, 204-218.

Hencky H. 1. Über enige statisch bestimmte Fälle des Gleichgewichts in plastischen Körper. Zeitschr. f. angew. Math. und Mech., 1921, 3, H. 4, 241-251.

Hille E., Szegö G., Tamarkin J.D. 1. On some generalieations of a theorem of A.Markoff. Duke Math.J., 3, 1937, 729-739.

Lanchon H. 1. Solution du problème de torsion d'une barre cylindrique de section quelconque. Comptes Rend. de l'Acad. d. Sci.Franç, ser.A, 1969, 269, 791-794.

Langenbach A. 1. Die Linearisierung nichtlinearen Gleichungen. Math.Nachr., 1962, 24, 33-51.

- - - 2. Monotone Potentialoperatoren. 1976, Berlin, SS. 358.

Lehman N.J. 1. Eine Fehlerabschätzung zum Ritzschen Verfahren für inhomogene Randwertaufgaben. Numerische Math., 1960, 2, N 2, 60-66.

Lévy M. 1. Mémoire sur les équations générales des mouvements intérieurs des corps solides ductiles au delà des limites ou l'élasticité pourrait les ramener à leur premier état. Comptes Rend de l'Acad. d.Sci.Franç., 1870, 70, 1323-1325.

Michlin S.G. 22 Approximation auf dem kubischen Gitter. Berlin, 1976, SS.197

- - - 31. Konstanten in einige Ungleichungen der Analysis. Leipzig , 1981, SS.119.

Michlin S.G., Prößdorf S. 1. Singuläre Integraloperatoren. Berlin 1980. SS.514.

v.-Mises R. 1. Mechanik der festen Körper in plastisch-deformablen Zustand. Nachr. v. d. Königlich. Gesellsch. d.

Wissensch. zu Göttingen, 1913, H. 4, S.582.

- - - 2. Bemerkungen zur Formulierung des mathematischen Problems der Plastizitätstheorie. Zeitschr. f. angew.Math. und Mech., 1925, IV, H. 2, 147-149.

Mittelman H.D. 1. On the approximate solution of nonlinear variational inequalities. Numerische. Math., 1978, 29, 451-462.

Morry B.Jr. 1. Multiple integrals in the calculus of variation N.Y., 1966, pp.506.

. Mosco U. 1. Implicit variational problems and quasi variational inequalities. Lecture Notes in Math., 1976, 543, 83-136.

Nitsche J. 1. L-convergence of finite element approximation. Lecture Notes in Math., 1977, 606, 261-274.

Omodei B.J. 1. Stability of the Rayleigh - Ritz - Galerkin procedure for elliptic boundary value problem. A thesis submitted to the Australian National Univ. 1975, pp.138.

Prößdorf S., Schmidt G. 1. A finite element collocation method for singular integral equations. 1. Math.Nachr., 1981, 100, 33-60.

Prößdorf S., Silberman B. 1. Projektionsverfahren und die näherungsweise Lösung singulärer Gleichungen. Leipzig , 1977, SS.226.

Ritz W. 1. Über eine neue Methode zur Lösung gewisser Variationsprobleme der Mathematischen Physik. J. f. d. reine und angew. Math., 1908, 135, H.1.

de-Saint-Venant B. 1. Sur l'établissement des équations des mouvements intérieurs opérés dans les corps ductiles au delà des limites où l'élasticité pourrait les ramener à leur premier état. Comptes Rend. de l'Acad. d. Sci Franç., 1870, 70, 473-480.

Scarpini F. . Some nonlinear complementary systems algorithms and applications to unilateral boundary value problems. Ann.Math.pura ed appl. 1980, 123, 185-202.

Scarpini F., Vivaldi M.A. Error estimates for the approximation of some unilateral problem . RAIRO, Anal.numer , 1977, 11, N 2, 197-208.

Sobolev S. 1. The problem of propagation of plastical state Труды Сейсмологического ин-та АН СССР, № 49 1935, pp. 15

Strang G. 1. Approximation in the finite element method

Numerisch Math., 1972, 19, 91-98.

Strang G., Fix G. 1. A Fourier analysis of the finite element method. (Препринт; год издания не указан).

Tucker T.S. Stability of nonlinear computing schemes. SIAM. J. Numer. anal., 1969, 6, 74-81.

Temple G. The general theory of relaxation methods applied to linear systems. Proc. of Royal Soc., ser.A, 1939, 16[illegible], 476-500.

Wilkinson J.H. 1. Rounding errors in algebraic processes. Nat. Phys. Laboratory. Notes on appl. science. Englewood Cliffe, 1963, pp. 161.

Woźniakowski H. 1. Round off error analysis of iterations for large linear systems. Numerische Math., 1978, 30, fasc.3, 301-314.

Zenisek A. 1. Interpolation polynomials on the triangle. Numerische Math., 1970, 15, 283

- - - 2. Hermite interpolation on symplexes in the finite element method. Proc. EQUADIFF-III, Brno 1972, p.271.

- - - 3. Polynomial approximations on tetrahedrons in the finite element method. J. Approx. Theory, 1973, 7, N 4, 334-351.

Zlamal M. 1. On the finite elemen. method. Numerische Math. 1968, 12, 394-408.

ს.გ.მიხლინი

რიცხვითი პროცესების ცდომილებანი

რეზიუმე

წიგნი ეძღვნება რიცხვითი მათემატიკისა და რიცხვითი პროცესების საკმარისად ფართო კლასის საკითხებს, რომლებიც მოცემული ამოცანის ამოხსნის სიზუსტესთანაა დაკავშირებული. წყაროები იმ ცდომილებებისა, რომლებიც დაკავშირებულია ამ პროცესებთან, განიხილება იმ დაშვებით, რომ ამოცანა დასმულია ზუსტად. ავტორის აზრით, ზოგ შემთხვევაში ცდომილების ზუსტად ოთხი წყარო გვაქვს: მეთოდის ცდომილება, რომელიც განპირობებულია იმის შედეგად, რომ მოცემული ამოცანა შეცვლილია ახალი განსხვავებული ამოცანით; დამრგვალების ცდომილება, რომელიც დაკავშირებულია იმ ფაქტთან, რომ განსხვავებული ამოცანა ამოხსნილია გარკვეული ცდომილებით; ალგორითმული ცდომილება, რომელიც წარმოიშობა იმ შემთხვევაში, როდესაც ალგორითმი, განსაზღვრული განსხვავებული ამოცანის ამოსახსნელად, გვაძლევს მხოლოდ მიახლოებით ამონახსნს ბიჯების სასრული რიცხვით; განხორციელების ცდომილება, რომელიც წარმოიშობა ხსენებული ალგორითმის ინსტრუქციების შესრულების ცდომილების გამო. მოცემულია ყველა ამ ცდომილების შეფასება რიცხვითი პროცესების ფართო კლასისათვის.

წიგნი შეიცავს 9 თავს. თავი I ეძღვნება ამოცანების დასმას და ორი უმარტივესი მაგალითის განხილვას. წრფივი რიცხვითი პროცესების ცდომილებანი განიხილება II – VI თავებში, ხოლო არაწრფივი – VII - IX თავებში.

S.G.Mikhlin

Errors of numerical processes

Summary

This book deals with a sufficiently wide class of problems of the numerical mathematics and the numerical processes leading to the construction of the solution of a given problem. The sources of the errors related with these processes are investigated under the assumption that the given problem is posed exactly. The author thinks that in many cases there are exactly four sources of errors: approximation error, resulting from the change of the given problem by a new simplified problem; distortion error which is connected with the fact that the data of the simplified problem are calculated with some errors; algorithm error, which arise in the cases when the algorithm, used for solve the simplified problem, gives only an approximate solution after a finite number of steps; round-off error, which arise because of the errors of the fulfilment of the instructions of the mentioned algorithm. Estimates of all these errors are given for a wide class of numerical processes.

The book contains 9 chapters. The Chapter I contains the statement of the problem and the investigation of two simpler examples. Errors of linear numerical processes are investigated in the Chapt. II-VI and the errors of non-linear ones are considered in the Chapt. VII-IX.

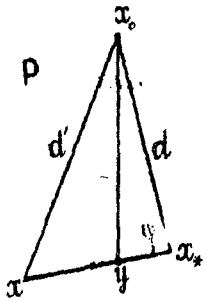

Рис. 1

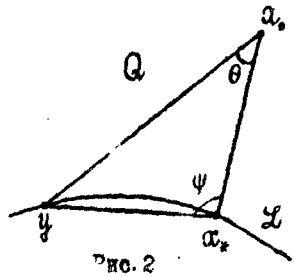

Рис. 2

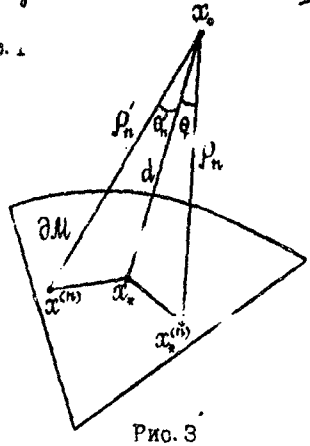

Рис. 3

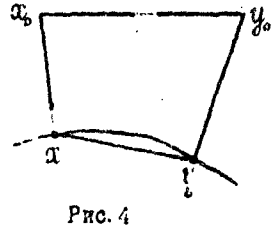

Рис. 4

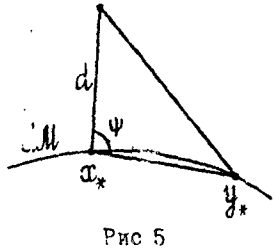

Рис 5

ОГЛАВЛЕНИЕ

УДК 519.6

Погрешности вычислительных процессов. Михлин С.Г. Институт прикладной математики имени акад. И.Н.Векуа Тбилисского государственного университета, 1983, с. 260.

В книге рассматривается довольно широкий класс задач вычислительной математики и вычислительные процессы, которые приводят к построению решения данной задачи, анализируются источники погрешностей и оцениваются самые погрешности в предположении, что данная задача поставлена точно. Как кажется автору, во весьма многих случаях существует точно четыре источника погрешности: погрешность аппроксимации, проистекающая от замены данной задачи другой, упрощенной задачей; погрешность искажения, связанная с тем, что данные упрощенной задачи вычисляются неточно; погрешность алгоритма, возникающая в тех случаях, когда алгоритм, примененный для решения упрощенной задачи, дает после конечного числа шагов лишь приближенное её решение; погрешность округления, возникающая из-за того, что предписания упомянутого алгоритма выполняются неточно. Для широкого круга вычислительных процессов даны оценки всех перечисленных погрешностей.

Книга содержит 9 глав. В главе I дана постановка проблемы и исследованы два более простых примера. В главах II-VI исследуются погрешности линейных вычислительных процессов, в главах VII-IX тот же вопрос исследуется для нелинейных процессов.

Книга содержит расширенное изложение докладов, сделанных автором в Институте прикладной математики имени академика И.Н.Векуа в 1982 году.

ს.ლომინ კრივოლის და მისლინი

გამოთვლათა პროცესების ღრმულდება

(რუსულ ენაზე)

თბილისის უნივერსიტეტის გამომცემლობა

თბილისი, 1983

Редактор издательства Н.Девдариани
Подписано в печать 16.05.83
Печатных листов 16,5
Учётно-изд.листов 13,31
Заказ 48 УЭ 03803 Тираж 290
Цена 1 руб. 30 коп.

36,56

თსუ გამოყენებითი მათემატიკის ინსტიტუტის როტაპრინტი
380043, თბილისი, 43.
Ротапринт Института прикладной математики ТГУ
380043, Тбилиси, 43