Сибирский математический журнал
Ноябрь—декабрь, 2013. Том 54, № 6

УДК 519.632.6+532.582.33

ОСОБЕННОСТИ РЕАЛИЗАЦИИ НЕНАСЫЩАЕМОГО ЧИСЛЕННОГО МЕТОДА ДЛЯ ВНЕШНЕЙ ОСЕСИММЕТРИЧНОЙ ЗАДАЧИ НЕЙМАНА

В. Н. Белых

Аннотация. На основе фундаментальных идей К. И. Бабенко построен принципиально новый — *ненасыщаемый* — метод численного решения спектральной задачи для оператора внешней осесимметричной задачи Неймана для уравнения Лапласа. Дана оценка уклонения первого собственного числа дискретизованной задачи от собственного числа оператора Неймана. Точнее, результатом ненасыщаемой дискретизации спектральной задачи Неймана является алгебраическая задача с «хорошей» матрицей, т. е. с матрицей, наследующей спектральные свойства оператора Неймана, и потому в ее спектральном портрете отсутствуют «паразитические» собственные числа, если только погрешность дискретизации достаточно мала. При этом оценка погрешности первого собственного числа содержит эффективно вычисляемые параметры, что в случае C^∞-гладких данных составляет основание для гарантированного (доказательного) успеха.

Ключевые слова: уравнение Лапласа, осесимметричная задача Неймана, спектральная задача, ненасыщаемый численный метод, экспоненциальная сходимость.

1. Преамбула

Внешняя задача Неймана для уравнения Лапласа возникает во многих разделах гидродинамики в качестве важного промежуточного этапа. При этом трудности ее численного исследования столь разнообразны, что далеко не каждый метод способен представить реальный практический интерес. Связано это с тем, что от качественности численной реализации этой эллиптической задачи порой зависит корректность исследования самой гидродинамической задачи в целом (например, в случае задач, описываемых уравнениями пограничного слоя [1]).

Теория эллиптических задач для уравнения Лапласа хорошо разработана, но ряд практически важных вопросов в ней до сих пор остаются нерешенными. Так, до настоящего времени имеется существенный пробел в прецизионном численном решении этих задач в гладких трехмерных областях. В первую очередь это связано с отсутствием «хорошей» теории конструктивного приближения функций на гладких многообразиях, гомеоморфных двумерной сфере. Именно по этой причине проблема построения численного ответа гарантированного качества в гладких трехмерных областях до сих пор представляется

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (коды проектов 11–01–00147–а, 12–01–00061–а).

очень трудной вычислительной проблемой. Разрешение ее, как и предсказывает теория, относится к сфере, в гораздо бо́льшей степени интеллектуальной: на первый план выдвигается проблема конструирования ненасыщаемых вычислительных методов [2]. Существующие на этом пути трудности усугубляются еще и отсутствием тех прозрений здравого смысла, которые могли бы удовлетворить реальным запросам вычислительной практики. Сравнительно просто, но не тривиально указанная проблематика разрешается в случае C^∞-гладких осесимметричных областей достаточно произвольной формы [3].

Отыскание численного решения эллиптической краевой задачи осуществляется в два этапа: сначала бесконечномерная задача сводится к некоторой аппроксимирующей ее конечномерной задаче, а затем указывается эффективный алгоритм численного решения полученной системы линейных алгебраических уравнений. При этом многие численные методы конструируются так, чтобы получить в итоге не наилучшую дискретизацию эллиптической задачи, а просто решаемую систему линейных уравнений [4]. Традиционные методы численного решения эллиптических задач (конечно-разностные, конечных элементов, квадратур) теряют на первом этапе бо́льшую часть информации, содержащейся в таблице m чисел, которая возникает при их дискретизациях. Эти методы не используют экстраординарную гладкость решений, и их погрешность имеет степенной порядок убывания вида m^{-r}, где $r > 0$ — целое число. В действительности решение эллиптической задачи может обладать более высоким $l > r$ запасом гладкости решений, но это обстоятельство никак не способно повлиять на численный результат, если метод имеет главный член погрешности m^{-r}, т. е. *насыщаем* [4]. Идеальным здесь было бы положение, когда каждой теореме существования решения (в форме шаудеровских оценок) соответствовала бы теорема (оценка) сходимости приближенного решения в нормах, отвечающих гладкости отыскиваемого решения задачи [5]. Но практика далека еще от такого положения, хотя и имеется некоторый прогресс, связанный с открытием принципиально новых — *ненасыщаемых* — вычислительных методов [2–4], которые до недавнего времени не входили в арсенал средств вычислителей-практиков.

Основное свойство ненасыщаемых численных методов — отсутствие главного члена погрешности — и как результат — способность автоматически подстраиваться к имеющимся экстраординарным «запасам» гладкости решений задач. Это означает, в свою очередь, что с ростом запаса гладкости (при прочих равных условиях) скорость убывания погрешности метода к нулю возрастает. Пика эффективности — *экспоненциальной сходимости* — ненасыщаемые методы достигают на классах бесконечно гладких решений задач (скорость сходимости сравнивается с числом m свободных параметров у функций, из которых конструируются приближения). Последнее создает необходимые предпосылки к чрезвычайно эффективному построению численного решения эллиптической краевой задачи, а в случае «хорошей» обусловленности дискретизации — к его отысканию с экспоненциальной точностью. При этом не возникает проблем, связанных с решением линейных алгебраических систем «большой» размерности, а стало быть, и коллизий между точностью компьютерного числового ответа и объемом перерабатываемой битовой информации, если решение задачи достаточно гладкое, например, бесконечно дифференцируемое [3].

В [3] построена ненасыщаемая хорошо обусловленная дискретизация эллиптической краевой задачи и указан итерационный метод численного решения полученной системы линейных алгебраических уравнений с несимметричной полностью заполненной квадратной матрицей A. Понятно, что в отборе метода численного решения такой системы линейных уравнений нужно быть весьма осмотрительным, с тем чтобы не лишить метод практической целесообразности — допустимого по точности ответа. Одним из факторов, обеспечивающих эффективное функционирование итерационного метода, является удачный выбор его итерационных параметров. При этом сам выбор не является волюнтаристским: его содержательность всякий раз должна подкрепляться «скрывающейся» в спектральном портрете матрицы A информацией, всецело определяемой способом дискретизации рассматриваемой задачи. При этом наличие у A кратных или достаточно близких собственных значений может серьезно сказаться на скорости сходимости итерационного метода, ибо поведение степеней A^k определяется исключительно структурой спектра матрицы A. Поэтому наличие у A жордановых клеток — это худший случай, который может здесь представиться (отсутствие у A указанных патологий отнюдь не очевидно ввиду несамосопряженности рассматриваемой в [3] задачи).

Ограниченность объема статьи [3] не позволила в деталях обсудить эти вопросы. Предлагаемое дополнение восполняет этот пробел. Именно, в нем выявлен субстрат тех общих представлений, который в итоге и обеспечил построение эффективного итерационного метода решения системы линейных алгебраических уравнений, метода, способного автоматически, исходя из складывающейся ситуации, отбирать итерационные параметры. Иначе говоря, ненасыщаемая дискретизация оператора K рассматриваемой в [3] эллиптической задачи приводит в итоге к алгебраической задаче с «хорошей» матрицей A, т. е. с матрицей, наследующей спектральные свойства оператора K. Поэтому в спектральном портрете матрицы A уже невозможны патологии, вроде указанных выше, если они отсутствуют у оператора K. Показано также, что если первое собственное число λ оператора K «хорошо» отделено от остальной части спектра, то указанный в [3] итерационный процесс численно устойчив и эффективно сходится. При этом число λ вычисляется приближенно с точностью, которая диктуется самой ненасыщаемой дискретизацией. Это позволяет всякий раз, исходя из конкретной ситуации, автоматически отбирать параметры итерационного метода. В результате ненасыщаемая дискретизация эллиптической краевой задачи с компактным оператором K снабжает нас элегантным вычислительным средством, способным в дискретизованной форме наследовать как дифференциальные, так и спектральные характеристики оператора K. Последнее служит серьезным основанием для отыскания компьютерного числового ответа с гарантированной (доказательной) точностью, если решение эллиптической краевой задачи достаточно гладкое, например, C^∞-гладкое.

2. Дискретизация спектральной задачи

Пусть $C \equiv C[0, 2\pi]$ — класс вещественных 2π-периодических непрерывных функций с чебышевской нормой $\|\cdot\|$, а $C_+ \equiv C_+[0, 2\pi] \subset C$ — множество четных функций в нем; класс k раз непрерывно дифференцируемых функций, производная k-го порядка которых удовлетворяют условию Гёльдера с показателем $0 < \alpha < 1$, обозначим через $C_+^{k+\alpha} \subset C_+$ $(k \geq 0)$. Пусть $\mathscr{T}^m \subset C_+$ — подпространство тригонометрических многочленов порядка не выше m, $m \geq 0$ целое, $Q_m : C_+ \to \mathscr{T}^m$ — проектор, а $e_m(g) = \inf\limits_{H_m \in \mathscr{T}^m} \|g - H_m\|$ — наилучшее (чебышевское) приближение функции $g \in C_+$.

Классический способ представления решения внешней осесимметричной задачи Неймана для уравнения Лапласа основан на понятии гармонического потенциала и допускает равносильную формулировку в терминах решения граничного интегрального уравнения. Роль оператора задачи в нем полноценно исполняет интегральный оператор

$$K[\psi](s) = \int_0^\pi k(s,\sigma)\psi(\sigma)\,d\sigma, \quad s \in [0,\pi],\ \psi \in C_+, \tag{1}$$

— прямое значение нормальной производной потенциала простого слоя на достаточно гладкой (в частности, C^∞-гладкой) замкнутой поверхности вращения [6]. Ядро $k(s,\sigma)$ оператора K слабо сингулярно [6, 7]. Имея слабую особенность, оператор $K : C_+ \to C_+$ обладает более сильными свойствами непрерывности, чем сами функции, на которые он действует: он компактен в C_+. При этом его норма в C_+ вычисляется по формуле

$$\|K\| \equiv \max_{0 \le s \le \pi} \int_0^\pi |k(s,\sigma)|\,d\sigma.$$

Производные функции $w(s) = -(K\psi)(s)$ в случае $\psi \in C_+^\alpha$ существуют [6, с. 129] и удовлетворяют условию Гёльдера с показателем $0 < \beta < \alpha$. Явный вид оператора K на C^∞-гладкой замкнутой поверхности вращения указан в [3].

Для оператора $K : C_+ \to C_+$ рассмотрим следующую спектральную задачу:

$$K\psi = \lambda\psi, \quad \|\psi\| = 1. \tag{2}$$

Формальный подход к ее дискретизации будет осуществляться по следующей схеме. Пусть $w_0(s), w_1(s), \dots, w_m(s)$ — базис в $\mathscr{T}^m$. Тогда

$$Q_m g = \sum_{k=0}^m g_k w_k(s) \quad \text{и} \quad \|Q_m\| = \max_{0\le s\le\pi} \sum_{k=0}^m |w_k(s)| \text{ — норма проектора } Q_m \text{ в } C_+.$$

Функции $g \in C_+$ соответствует система чисел $g_0, g_1, \dots, g_m$ — координат в $\mathscr{T}^m$.

Обозначив $\varrho_m = \psi - Q_m\psi$ и использовав в (1) равенство $\psi = Q_m\psi + \varrho_m$, получим

$$KQ_m\psi = \lambda\psi + \rho(\psi), \quad \text{где } \rho(\psi) = -K[\psi - Q_m\psi]. \tag{3}$$

В выборе дискретизации спектральной задачи (2) мы вовсе не свободны: базис в $\mathscr{T}^m$ задается, исходя из условий близости операторов $KQ_m : C_+ \to C_+$ к оператору K при $m \to \infty$ в равномерной (операторной) топологии.

Зададим базис по узлам

$$s_i = 2\pi i/(2m+1), \quad 0 \le i \le m,$$

и линейному отображению

$$J : C_+ \to \mathbb{R}^m, \quad Jg = (g(s_0), \dots, g(s_m)), \quad |Jg|_\infty = \max_{0\le k\le m} |g_k|.$$

Построим семейство $w_0(s), w_1(s), \dots, w_m(s)$ фундаментальных многочленов лагранжевой интерполяции, образующее базис в $\mathscr{T}^m \subset C_+$, $w_k(s_j) = \delta_{kj}$ $(0 \le k, j \le m)$. Затем рассмотрим интерполяционный тригонометрический многочлен в форме Лагранжа [3]:

$$(Q_m g)(s) \equiv Q_m(s; Jg) = \sum_{k=0}^m g(s_k) w_k(s). \tag{4}$$

Ясно, что $Q_m w_k \equiv w_k$ и $JQ_m g \equiv Jg$. Многочлен $Q_m(s; Jg)$ соответствует проектору $Q_m : C_+ \to \mathscr{T}^m$, $\|Q_m\|$ — его константа Лебега. Порядок роста величин $\|Q_m\|$ с ростом m зависит от выбора конкретного базиса. Оптимальный выбор базиса осуществляется на основе неравенства Лебега [4]

$$\|g(s) - Q_m(s; Jg)\| \le (1 + \|Q_m\|)e_m(g),$$

причем таким образом, что $\|Q_m\| \le 3 + 2\pi^{-1} \ln m$. Важно, чтобы норма проектора $Q_m : C_+ \to \mathscr{T}^m$ была минимально возможной.

Известно, что для тригонометрического многочлена Лагранжа $Q_m(s; Jg)$ и любого $1 < p < \infty$ справедливо следующее предельное равенство [8]:

$$\lim_{m\to\infty} \|g(s) - Q_m(s; Jg)\|_p = 0 \quad \forall g \in C[0, 2\pi].$$

Здесь $\|\cdot\|_p$ обозначает норму в $L_p[0, 2\pi]$:

$$\|g\|_p \equiv \left(\int\limits_0^{2\pi} |g(t)|^p \, dt \right)^{1/p}.$$

Применив к обеим частям равенства (3) оператор J и обозначив $u = J\psi$, получим соотношение

$$Au = \lambda u + \delta, \quad u = J\psi, \quad \delta = -JK[\psi - Q_m\psi], \quad u, \delta \in \mathbb{R}^m, \ \psi \in C_+. \tag{5}$$

Здесь $A = (a_{jk})$ — матрица с элементами $a_{jk} = K[w_k](s_j)$, $0 \le k \le m$, $0 \le j \le m$.

Матрица A определяет линейный оператор $A : \mathbb{R}^m \to \mathbb{R}^m$, который и предлагается рассматривать в качестве дискретизации оператора K. Матрица A полностью заполнена, что резко контрастирует с ситуацией в конечно-разностных методах, где матрицы дискретизации обычно сильно разрежены.

Число λ в соотношении (5) — это в точности искомое собственное значение задачи (2), а компоненты вектора $u = J\psi \in \mathbb{R}^m$ — точные значения соответствующей λ собственной функции $\psi(s)$ в узлах интерполяции s_j, $0 \le j \le m$.

Отбрасывая в (5) погрешность $\delta \in \mathbb{R}^m$ и обозначая приближенное значение вектора $u = J\psi \in \mathbb{R}^m$ через $\bar{\psi} \in \mathbb{R}^m$, получим искомую дискретизацию задачи (2):

$$A\bar{\psi} = \mu\bar{\psi}, \quad |\bar{\psi}|_\infty = 1. \tag{6}$$

Здесь μ — собственное число, а $\bar{\psi} \in \mathbb{R}^m$ — собственный вектор матрицы A соответственно, причем компоненты $\bar{\psi}$ — приближенные значения в узлах s_j, $0 \le j \le m$, собственной функции ψ задачи (2). Число μ и полином $Q_m(s; \bar{\psi})$ — искомые приближения к собственному числу λ и собственной функции ψ спектральной задачи (2). Несмотря на то, что собственные значения вещественных несимметричных матриц A могут оказаться, вообще говоря, комплексными, нас будет интересовать случай вещественных μ (см. [3]).

3. Сходимость операторов KQ_m к оператору K. Основной результат

К. И. Бабенко первым [4] оценил важность ненасыщаемых дискретизаций в спектральных задачах, чем был достигнут определенный успех в понимании роли и уровня влияния экстраординарных «запасов» гладкости собственных функций на величину погрешности и особенность функционирования компьютерных вычислений.

Оценим возмущение, вносимое в собственное число λ спектральной задачи (2) отбрасываемой в (5) погрешностью дискретизации — вектором $\delta = -JK[\psi - Q_m\psi]$, $\delta \in \mathbb{R}^m$. С этой целью выясним, как изменяются собственные числа оператора K при малых (в равномерной норме) его возмущениях. Точнее, пусть оператор K имеет собственное число λ. Тогда выясним, имеется ли у его дискретизации K_m с матрицей A собственное число μ, близкое к λ.

Общий метод исследования задач такого рода обычно зиждется на теории регулярных возмущений линейных ограниченных операторов в банаховом пространстве [9]. Согласно этой теории возмущение спектра оператора K связывается с равномерной (по норме) сходимостью последовательности приближающих его операторов K_m.

Установление факта сходимости K_m по норме обычно встречает затруднение принципиального характера, связанное с выявлением условий, обеспечивающих выполнение предельного соотношения

$$\|K - K_m\| = \sup_{\|g\|\le 1} \|Kg - K_m g\| \to 0 \quad \text{при } m \to \infty.$$

Здесь супремум берется по единичному шару, т. е. по некомпактному в C_+ множеству.

Специфика решаемой задачи состоит в следующем. Если заранее ничего не известно о спектральных свойствах оператора K, то сталкиваемся с трудной вычислительной проблемой [10], в разрешении которой можно надеяться разве что на удачу, а не на гарантированный успех. В рассматриваемой нами ситуации эту трудность удалось преодолеть, ориентируясь на ключевые свойства исходного интегрального оператора K. Этот оператор компактен, ненулевые точки его спектра, т. е. вещественные простые полюсы резольвенты $R(\zeta, K) \equiv (K - \zeta I)^{-1}$, содержатся в промежутке $(-1, 1]$, а его собственные функции непрерывны [6].

Указанные свойства оператора K в сочетании с введенным С. Л. Соболевым [11, 12] весьма важным понятием близости операторов, отличным от близости по норме, дали возможность эффективно разрешить рассматриваемую проблему.

Покажем, что выбор (4) интерполяционного проектора Q_m обеспечивает равномерную по норме сходимость операторов KQ_m к компактному оператору K задачи (2).

Теорема 1. *Последовательность операторов $\{KQ_m\}$ сходится при $m \to \infty$ равномерно к компактному оператору K задачи* (2):

$$\|K - KQ_m\| = \sup_{\|g\|\le 1} \|Kg - KQ_m g\| \to 0 \quad \text{при } m \to \infty. \tag{7}$$

Доказательство. В силу линейности оператора K, определенного равенством (1), а также с учетом (4) имеем

$$K[Q_m g](s) = \sum_{k=0}^{m} g(s_k) a_k(s), \quad \text{где } a_k(s) = K[w_k](s). \tag{8}$$

Далее,

$$\|KQ_m g\| \le \max_{0\le s\le \pi} \sum_{k=0}^{m} |g(s_k)||a_k(s)| \le |Jg|_\infty \|KQ_m\|,$$

где $\|KQ_m\| = \max\limits_{0\le s\le\pi} \sum\limits_{k=0}^{m} |a_k(s)|$. Интегральный оператор $K : C_+ \to C_+$, имея слабую особенность, обладает более сильными свойствами непрерывности, чем сами функции, на которые K действует [6, с. 129]. При этом существует [7, с. 65] число $q > 1$ такое, что

$$|(KQ_m g)(s)| \le \left(\int\limits_0^{\pi} |k(s,\sigma)|^q \, d\sigma\right)^{1/q} \|Q_m g\|_p, \quad \frac{1}{p} + \frac{1}{q} = 1 \quad \forall s \in [0,\pi] \; \forall g \in C_+. \tag{9}$$

Отсюда, из соотношения (8) и теоремы Банаха — Штейнгауза в применении к последовательности Q_m имеем

$$\|KQ_m\| \le Q\|Q_m\|_p < C < \infty, \quad Q = \max_{0\le s\le\pi} \left(\int\limits_0^{\pi} |k(s,\sigma)|^q \, d\sigma\right)^{1/q} < \infty, \tag{10}$$

$$|(KQ_m g)(s) - (KQ_m g)(t)| \le \left(\int\limits_0^{\pi} |k(s,\sigma) - k(t,\sigma)|^q \, d\sigma\right)^{1/q} \|Q_m g\|_p \quad \forall s,t \in [0,\pi], \tag{11}$$

где постоянная C не зависит от m.

Таким образом, нормы операторов KQ_m равномерно ограничены в пространстве C_+. Совершенно похожим способом получаем, что

$$\|KQ_m g - Kg\| \le Q\|Q_m g - g\|_p \to 0 \quad \text{при } m \to \infty \quad \forall g \in C_+. \tag{12}$$

Следовательно, операторы KQ_m сильно сходятся к оператору K.

В силу известного критерия компактности [13, с. 162] интегрального оператора K из неравенств (10) и (11) следует, что операторы KQ_m равномерно вполне непрерывны в единичном шаре $G = \{g \in C_+ \mid \|g\| \le 1\}$ [11]. Таким образом, замыкание объединения $\bigcup\limits_{m\ge 0} KQ_m(G)$ в силу классической теоремы Арцела компактно в C_+.

Предположим, что равенство (7) не выполняется, т. е. существуют $\varepsilon > 0$ и некоторая последовательность $\{m_j\} \to \infty$ такие, что $\|K - KQ_{m_j}\| > \varepsilon$. Поскольку KQ_m сходятся к K сильно и замыкание множества $\bigcup\limits_{m\ge 0} KQ_m(G)$ компактно в C_+, найдется такая подпоследовательность $\{n_i\} \subseteq \{m_j\}$, что

$$\|K - KQ_{n_i}\| \to 0 \quad \text{при } i \to \infty;$$

противоречие сделанному ранее допущению.

Теорема 1 доказана.

Следствие 1. *Пусть $r(K) = \lim\limits_{n\to\infty} \|K^n\|^{1/n}$ — спектральный радиус компактного оператора K и $\|K - KQ_m\| \to 0$ при $m \to \infty$. Тогда $r(K)$ является непрерывной функцией K: $r(K) = \lim\limits_{m\to\infty} \|KQ_m\|$.*

Доказательство. Известно, что $r(K)$ — полунепрерывная сверху функция [9, с. 264]. Поэтому $\overline{\lim\limits_{m\to\infty}} \|KQ_m\| \le \|K\|$. Из (12) в силу равномерной ограниченности последовательности норм $\|KQ_m\|$ следует, что $\|K\| \le \underline{\lim\limits_{m\to\infty}} \|KQ_m\|$.

Следствие 2. *Пусть λ_0 принадлежит спектру $\sigma(K)$ компактного оператора K, являясь простым полюсом его резольвенты $R(\zeta, K) \equiv (K - \zeta I)^{-1}$. Тогда для всякого $\varepsilon > 0$ существует такое m_0, что $d(\lambda_0, \sigma(KQ_m)) < \varepsilon$ для всех $m \geq m_0$ (здесь d — евклидово расстояние на комплексной плоскости ζ).*

Доказательство. Предположим противное: существуют $\varepsilon > 0$ и подпоследовательность $\{KQ_{m_j}\}$, для которых выполняются условия:

1) $d(\lambda_0, \sigma(KQ_{m_j})) \geq \varepsilon$;

2) $R(\zeta, KQ_{m_j})$ аналитична в открытом круге $|\zeta - \lambda_0| < 2\varepsilon$;

3) если $0 < |\zeta - \lambda_0| < 2\varepsilon$, то $\zeta \notin \sigma(K)$.

Пусть $\psi \neq 0$ и $K\psi = \lambda_0\psi$. В силу условия 2 и соотношения

$$\lim_{m\to\infty} \|K - KQ_m\| = 0$$

получаем

$$0 = \frac{1}{2\pi i} \lim_{j\to\infty} \int\limits_{|\zeta-\lambda_0|=\varepsilon} R(\zeta, KQ_{m_j})\, d\zeta = \frac{1}{2\pi i} \int\limits_{|\zeta-\lambda_0|=\varepsilon} R(\zeta, K)\, d\zeta$$

$$= \frac{1}{2\pi i} \int\limits_{|\zeta-\lambda_0|=\varepsilon} \frac{1}{\zeta - \lambda_0}\, d\zeta = 1.$$

Это противоречие доказывает следствие 2.

Обратим внимание на то, что спектральная задача (6), с которой мы имеем дело, несамосопряженная. Первое ее собственное число μ, как известно, вещественное [3] и согласно следствию 2 простое. Интересно выявить факторы, влияющие на величину погрешности в определении первого собственного числа оператора K. Иными словами, интересно выяснить, насколько близки спектры $\sigma(K)$ и $\sigma(KQ_m)$ двух несамосопряженных операторов, если близки сами операторы K и KQ_m (например, по норме).

Идея решения вопросов такого рода заимствована из [14–16]. Реализовать ее удалось, используя теорию регулярных возмущений линейных ограниченных операторов в банаховом пространстве [9]. Возмущение, вносимое ненасыщаемой дискретизацией в оператор K, зависит от того, насколько близки резольвенты операторов K и KQ_m вблизи первого собственного числа оператора K. Тем самым задача формально сводится к чисто технологической ее переформулировке в проблему устойчивости спектрального портрета оператора K при ненасыщаемых его возмущениях операторами $K_m \equiv KQ_m$.

Теорема (С. Д. Алгазин [15]). *Пусть K и K_m — ограниченные в банаховом пространстве* **B** *операторы и $\|K - K_m\| \to 0$ при $m \to \infty$. Пусть λ принадлежит точечному спектру оператора K, являясь простым полюсом его резольвенты $R(\zeta, K) \equiv (K - \zeta I)^{-1}$. Если внутри замкнутого контура Γ_λ, содержащего λ внутри себя, нет других собственных значений оператора K и при этом выполнено условие*

$$\sup_{\zeta\in\Gamma_\lambda} r(R(\zeta, K)(K_m - \zeta I) - I) < 1,$$

то внутри контура Γ_λ находится ровно одно собственное значение μ оператора K_m, причем выполняется неравенство

$$|\lambda - \mu| \leq \varrho \frac{\|R(\zeta_0, K)\| \|K - K_m\|}{1 - \|R(\zeta_0, K\| \|K - K_m\|}, \quad \varrho = \max_{\zeta\in\Gamma_\lambda} |\lambda - \zeta|, \tag{13}$$

где $\zeta_0 \in \Gamma$ — *точка, в которой достигается максимум* $\|R(\zeta_0, K)\| = \max\limits_{\zeta\in\Gamma_\lambda} \|R(\zeta, K)\|$.

Охарактеризуем первое собственное число матрицы A рассматриваемой дискретизации (6). Точнее, покажем, что результатом ненасыщаемой дискретизации спектральной задачи (2) является алгебраическая задача (6) с «хорошей» матрицей A, наследующей спектральные свойства оператора K. Поэтому в спектральной структуре матрицы A патологии невозможны, если только они отсутствуют у самого K. Собственное число λ задачи (2) вычисляется приближенно, причем с той точностью, которая диктуется ненасыщаемой дискретизацией.

Теорема 2. *Пусть внутри гладкого замкнутого контура* Γ_λ *находится ровно одно собственное значение* λ *оператора* K. *Тогда внутри* Γ_λ *имеется ровно одно собственное число* μ *матрицы* A *задачи* (6). *Справедлива также следующая оценка погрешности:*

$$|\mu - \lambda| \le c(m+1)^{1/2}|\delta|_\infty, \quad \text{где } |\delta|_\infty \le \|K\|(1 + \|Q_m\|)e_m(\psi),$$

и положительная постоянная c *вычисляется эффективно.*

Доказательство. Простое собственное число λ оператора K вещественно. Пусть выполнены соотношения

$$Au = \lambda u + \delta, \quad u = J\psi \in \mathbb{R}^m, \ \delta = -JK(\psi - Q_m\psi) \in \mathbb{R}^m.$$

Матрица $A : \mathbb{R}^m \to \mathbb{R}^m$ здесь та же, что и в (5). Вектор $u = J\psi \in \mathbb{R}^m$ совпадает с точными значениями в узлах s_j, $0 \le j \le m$, собственной функции ψ задачи (2). Вектор $\delta \in \mathbb{R}^m$ — погрешность дискретизации.

Норму матрицы, подчиненную евклидовой норме вектора, обозначим через $\|\cdot\|_2$. Известны неравенства $|u|_\infty \le \|u\|_2 \le \sqrt{m+1}|u|_\infty$. Отметим, что оператор KQ_m действует в том же пространстве C_+, что и оператор K. Поэтому в силу теоремы 1 и теоремы Алгазина внутри контура Γ_λ находится единственное собственное число оператора KQ_m.

Покажем, что и собственное число μ спектральной задачи (6) единственно, т. е. у матрицы A отсутствуют «паразитические» собственные числа, если только погрешность дискретизации достаточно мала.

Для доказательства воспользуемся следующим приемом [15]. Введем матрицу

$$B = A - (u,u)^{-1}\delta \otimes u, \quad (u,u) = \|u\|_2^2 = \sum_{i=0}^{m} u_i^2, \tag{14}$$

где $\delta \otimes u$ — кронекерово произведение векторов из $\mathbb{R}^m$.

Нетрудно убедиться, что $Bu = \lambda u$, т. е. λ — собственное число, а u — собственный вектор матрицы B. Из определения (14) следует, что

$$\begin{aligned}\|A - B\|_2 &= \sup_{\|g\|_2\le 1} \|(A-B)g\|_2 = \sup_{\|g\|_2\le 1} \|(u,u)^{-1}(\delta\otimes u)g\|_2 \\ &= \sup_{\|g\|_2\le 1} \|(u,u)^{-1}(u,g)\delta\|_2 \le \sup_{|g\|_2\le 1} |(u,g)(u,u)^{-1}|\|\delta\|_2 \le \|\delta\|_2 \le \sqrt{m+1}|\delta|_\infty.\end{aligned}$$

Учитывая, что $\delta = -JK[\psi - Q_m\psi]$, в силу условия $\psi \in C_+^{1+\beta}$ $(0 < \beta < \alpha)$ [6, с. 129] из классической теоремы Джексона заключаем, что при $m \to \infty$

$$\|A - B\|_2 \le \sqrt{m+1}|JK[\psi - Q_m\psi]|_\infty \le |\sqrt{m+1}\|K\|(1 + \|Q_m\|)e_m(\psi) \to 0. \tag{15}$$

Обозначив $R(\zeta) \equiv R(\zeta, A) = (A - \zeta I)^{-1}$, получаем

$$
\begin{aligned}
R(\zeta)(B - \zeta I) - I &= R(\zeta)(A - (u,u)^{-1}\delta \otimes u - \zeta I) - I \\
&= R(\zeta)((A - \zeta I) - (u,u)^{-1}\delta \otimes u) - I = R(\zeta)(A - \zeta I) - (u,u)^{-1}R(\zeta)\delta \otimes u - I \\
&= -(u,u)^{-1}R(\zeta)\delta \otimes u.
\end{aligned}
$$

Используя определение матричной нормы $\|\cdot\|_2$ в $\mathbb{R}^m$, находим

$$
\begin{aligned}
\|R(\zeta)(B - \zeta I) - I\|_2 &= \sup_{\|g\|_2 \le 1} \|R(\zeta)(B - \zeta I)g - Ig\|_2 \\
&= \sup_{\|g\|_2 \le 1} \|(u,u)^{-1}R(\zeta)(\delta \otimes u)g\|_2 = \sup_{\|g\|_2 \le 1} \|(u,g)(u,u)^{-1}R(\zeta)\delta\|_2 \\
&\le \|R(\zeta)\delta\|_2 \sup_{|g\|_2 \le 1} |(u,g)(u,u)^{-1}| \le \|R(\zeta)\delta\|_2 \le \|R(\zeta)\|_2\|\delta\|_2 \\
&\le \sqrt{m+1}\|R(\zeta)\|_2|\delta|_\infty \to 0 \quad \text{при } m \to \infty, \qquad (16)
\end{aligned}
$$

поскольку $|\delta|_\infty \le \|K\|(1 + \|Q_m\|)e_m(\psi)$ и $\psi \in C_+^{1+\beta}$.

Повторно применяя теорему Алгазина, заключаем из оценок (15) и (16), что контур Γ_λ (содержащий внутри себя единственное собственное значение оператора KQ_m) отличных от λ собственных чисел матрицы B не содержит. Иначе говоря, если погрешность δ в (5) достаточно мала, то внутри Γ_λ отличные от λ собственные числа матрицы B отсутствуют. Следовательно, у матрицы A имеется внутри Γ_λ единственное вещественное собственное число μ [3].

Чтобы оценить погрешность приближенного вычисления собственного числа λ спектральной задачи (2), воспользуемся оценкой (13). Действительно, из (15) следует, что

$$
|\lambda - \mu| \le c\|\delta\|_2 \le c\sqrt{m+1}|\delta|_\infty, \quad c = \varrho\|R(\zeta_0, A)\|_2(1 - \|R(\zeta_0, A\|_2\|\delta\|_2)^{-1}.
$$

Здесь

$$
\varrho = \max_{\zeta \in \Gamma_\lambda} |\lambda - \zeta|, \quad \|R(\zeta_0, A)\|_2 = \|(A - \zeta_0 I)^{-1}\|_2 = \max_{\zeta \in \Gamma_\lambda} \|R(\zeta, A)\|_2.
$$

Теорема 2 доказана.

Следствие 1. *Если* $d \equiv d(\lambda, \sigma(K)) > 0$ — *расстояние от собственного числа* λ *оператора* K *до остальной части его спектра* $\sigma(K)$, $\|K - KQ_m\| < d/2$ *и* $\varrho = \max_{\zeta \in \Gamma_\lambda} |\lambda - \zeta|$, *то для приближенного вычисления* λ *справедлива следующая оценка погрешности:*

$$
|\lambda - \mu| \le c_1(d)\|K - KQ_m\|, \quad \text{где } c_1(d) = 0.5d\varrho\left(1 - \|K - KQ_m\|\frac{2}{d}\right)^{-1}.
$$

Следствие 2. *Матрица* A *задачи* (6) *имеет собственный вектор* $\bar{\psi}$ *такой, что*

$$
|\bar{\psi} - J\psi|_\infty \le c_2(m+1)^{1/2}|\delta|_\infty, \quad \text{где } |\delta|_\infty \le \|K\|(1 + \|Q_m\|)e_m(\psi),
$$

и положительная постоянная c_2 *эффективно вычисляется.*

Из теорем 1 и 2 следует, что скорость сходимости последовательности приближенных решений $\mu \equiv \mu_m$ и $\bar{\psi} \equiv \bar{\psi}_m$ спектральной задачи (6) с ростом параметра $m \ge m_0$ определяется исключительно гладкостью собственной функции ψ задачи (2). При этом возмущение, вносимое ненасыщаемой дискретизацией в спектральную проблему (2), зависит от близости резольвент операторов

K и KQ_m вблизи первого собственного числа оператора K. Более того, если простое собственное число λ оператора K хорошо отделено от остальных его собственных чисел, то построенный в [3] итерационный процесс будет численно устойчив.

Все преимущества рассмотренного численного метода сохраняются лишь при условии, что точность приближенной реализации интегрального оператора K имеет тот же порядок, что и величина погрешности $|\delta|_\infty = |JK[\psi - Q_m\psi]|_\infty$ при $m \geq m_0$. Реализовать же это условие возможно лишь при использовании ненасыщаемых квадратурных формул [3]. В результате элементы $a_{jk} = K[w_k](s_j)$ матрицы A вычисляются приближенно с наперед заданной точностью $(1 + 2\|Q_m\|)e_m(\psi)$ с помощью ненасыщаемых квадратурных формул. При экстраординарной гладкости собственной функции ψ числовые характеристики $e_m(\psi)$ достаточно быстро убывают к нулю с ростом параметра m, а в случае $\psi \in C_+^\infty$ — экспоненциально быстро [3, 17]. Это обстоятельство принципиально отличает построенный ненасыщаемый численный метод решения проблемы (2) от численных методов с главным членом погрешности, т. е. насыщаемых.

4. Заключение

Элементы матрицы $A = (a_{ik})$, $a_{ik} = K[w_k](s_i)$, $0 \leq k \leq m$, $0 \leq i \leq m$, спектральной задачи (6) определены всегда с точностью, которая не может быть меньше $|\delta|_\infty$. Поэтому можно изменить элементы a_{ik} матрицы A на величину порядка $|\delta|_\infty$. Иначе говоря, задача абсолютно точного определения собственных чисел матрицы A лишена всякого смысла. В общей ситуации имеет смысл лишь задача о «почти собственных значениях» матрицы A при величине невязки порядка $|\delta|_\infty$. Любой метод численного решения алгебраической задачи на собственные значения есть метод вычисления «почти собственных значений», и потому важно уметь адаптировать его к классу корректности рассматриваемой задачи [10].

Для численного решения спектральных задач имеется много конкурирующих методов. Это прежде всего проекционные методы Ритца, Бубнова — Галеркина, а также конечно-разностные методы и методы конечных элементов. Об эффективности и о точности, которую дают эти методы, известно немало [4, 18]. Однако ряд важных обстоятельств при их конструировании не учитывается, что значительно снижает их эффективность. Чаще всего отыскиваемые решения спектральных задач обладают большой гладкостью либо даже аналитичны, являясь при этом элементами функциональных компактов с известными асимптотиками их m-поперечников и ε-энтропии [4]. Естественно, что метод, в основу которого положен рациональный способ приближения искомого элемента, даст численный ответ, близкий к оптимально возможному. Приведенные факты свидетельствуют, что без высококачественной дискретизации, учитывающей как имеющиеся запасы гладкости решений (первый этап), так и саму специфику организации ее компьютерной реализации (второй этап), рассчитывать в приближенном решении спектральной задачи можно лишь на удачу, а не на гарантированный успех.

Из сказанного явствует, насколько важно осуществлять ненасыщаемые дискретизации спектральных задач. Между тем действительность оказалась не столь замечательна, как об этом старательно напоминает теория: не воздав должное ошибкам округления, легко ошибиться и в оценке достоверности компьютерного числового ответа.

Что касается ошибок округления, то, по большому счету, вопрос, что чем управляет (ошибки округления процессом вычислений или процесс вычислений ошибками округления), при правильно поставленной вычислительной проблеме абсолютно надуман: компьютерных вычислений без ошибок округления (в арифметике вещественных чисел) не существует, а ошибки округления без компьютерных вычислений просто отсутствуют. Именно поэтому так важно уметь адаптировать численный метод к выполнению операций над числами конечной разрядности, т. е. к ошибкам округления, которые в спектральных задачах линейной алгебры всегда проявляются в форме чрезвычайной и даже грозной реальности [10]. В этой связи вопрос о корректных с точки зрения компьютерных вычислений постановках спектральных задач для конечномерных операторов (матриц) детально исследован С. К. Годуновым [10], чем был внесен весьма ощутимый концептуальный вклад в выработку новых представлений и формулировку компьютерных постановок спектральных задач линейной алгебры. Им же указаны примеры патологических матриц, для которых спектральная проблема плохо поставлена. Построению численных алгоритмов, обеспечивающих гарантированную точность в условиях приближенных вычислений, посвящены работы [3, 10, 19].

Автор благодарен проф. В. Л. Васкевичу, тщательно просмотревшему статью и сделавшему ряд весьма ценных уточнений и замечаний.

ЛИТЕРАТУРА

1. *Бэтчелор Дж.* Введение в динамику жидкости. М.: Мир, 1973.

2. *Бабенко К. И.* Несколько замечаний о дискретизации эллиптических задач // Докл. АН СССР. 1975. Т. 221, № 1. С. 11–14.

3. *Белых В. Н.* Ненасыщаемый численный метод решения внешней осесимметричной задачи Неймана для уравнения Лапласа // Сиб. мат. журн. 2011. Т. 52, № 6. С. 1234–1252.

4. *Бабенко К. И.* Основы численного анализа. М.: Наука, 1986. (2-е изд. М.; Ижевск: РХД, 2002).

5. *Агмон С., Дуглис А., Ниренберг Л.* Оценки решений эллиптических уравнений вблизи границы. М.: Изд-во иностр. лит., 1962.

6. *Гюнтер Н. М.* Теория потенциала и ее применение к основным задачам математической физики. М.: Гостехтеоретиздат, 1953.

7. *Михлин С. Г.* Многомерные сингулярные интегралы и интегральные уравнения. М.: Физматгиз, 1962.

8. *Marcinkiewicz I.* Sur l'interpolation // Studia Math. 1936. V. 6. P. 1–17.

9. *Като Т.* Теория возмущений линейных операторов. М.: Мир, 1972.

10. *Годунов С. К.* Лекции по современным аспектам линейной алгебры. Новосибирск: Науч. кн., 2002. (Университетская серия; Т. 12).

11. *Соболев С. Л.* Некоторые замечания о численном решении интегральных уравнений // Изв. АН СССР. Сер. мат. 1956. Т. 20, № 4. С. 413–436.

12. *Соболев С. Л.* Замыкание вычислительных алгоритмов и некоторые его применения. М.: АН СССР, 1955.

13. *Функциональный анализ*, 2 изд., перераб. и доп. / Под ред. С. Г. Крейна. М.: Наука, 1972. (Сер. «Справочная математическая библиотека»).

14. *Алгазин С. Д., Бабенко К. И.* Об одном численном алгоритме решения задачи на собственные значения для линейных дифференциальных операторов // Докл. АН СССР. 1979. Т. 244, № 5. С. 1049–1053.

15. *Алгазин С. Д.* О локализации собственных значений замкнутых линейных операторов // Сиб. мат. журн. 1983. Т. 24, № 2. С. 3–8.

16. *Алгазин С. Д.* Численные алгоритмы классической математической физики. М.: Диалог-МИФИ, 2010.

17. *Белых В. Н.* О свойствах наилучших приближений C^{∞}-гладких функций на отрезке вещественной оси (к феномену ненасыщаемости численных методов) // Сиб. мат. журн. 2005. Т. 46, № 3. С. 483–499.
18. *Голуб Дж., Ван Лоун Ч.* Матричные вычисления. М.: Мир, 1999.
19. *Годунов С. К., Антонов А. Г., Кирилюк О. П., Костин В. И.* Гарантированная точность решения систем линейных уравнений в евклидовых пространствах. Новосибирск: Наука, 1992.

Статья поступила 1 ноября 2012 г.

Белых Владимир Никитич
Институт математики им. С. Л. Соболева СО РАН,
пр. Академика Коптюга, 4, Новосибирск 630090
belykh@math.nsc.ru